Introduction to
Survey Data Analysis

JULY 2011

Afsaneh Yazdani

Preface

Learning from Data

Four-step process by which we can learn from data:

1. Defining the Problem
2. Collecting the Data
3. Summarizing the Data
4. Analyzing Data, Interpreting the Analyses, and Communicating the results

Preface

Survey

Is a systematic method for gathering information from (a sample of) entities for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the entities are member.

Preface

Survey

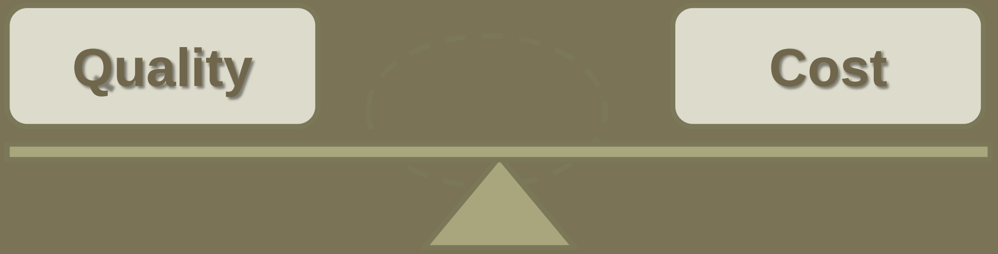
Is a systematic method for gathering information from (a sample of) entities for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the entities are member.

Describe non-observed on the basis of observed

Preface

Survey

Is a systematic method for gathering information from (a sample of) entities for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the entities are member.



Quality

Cost

Preface

There are two parallel aspects of surveys:

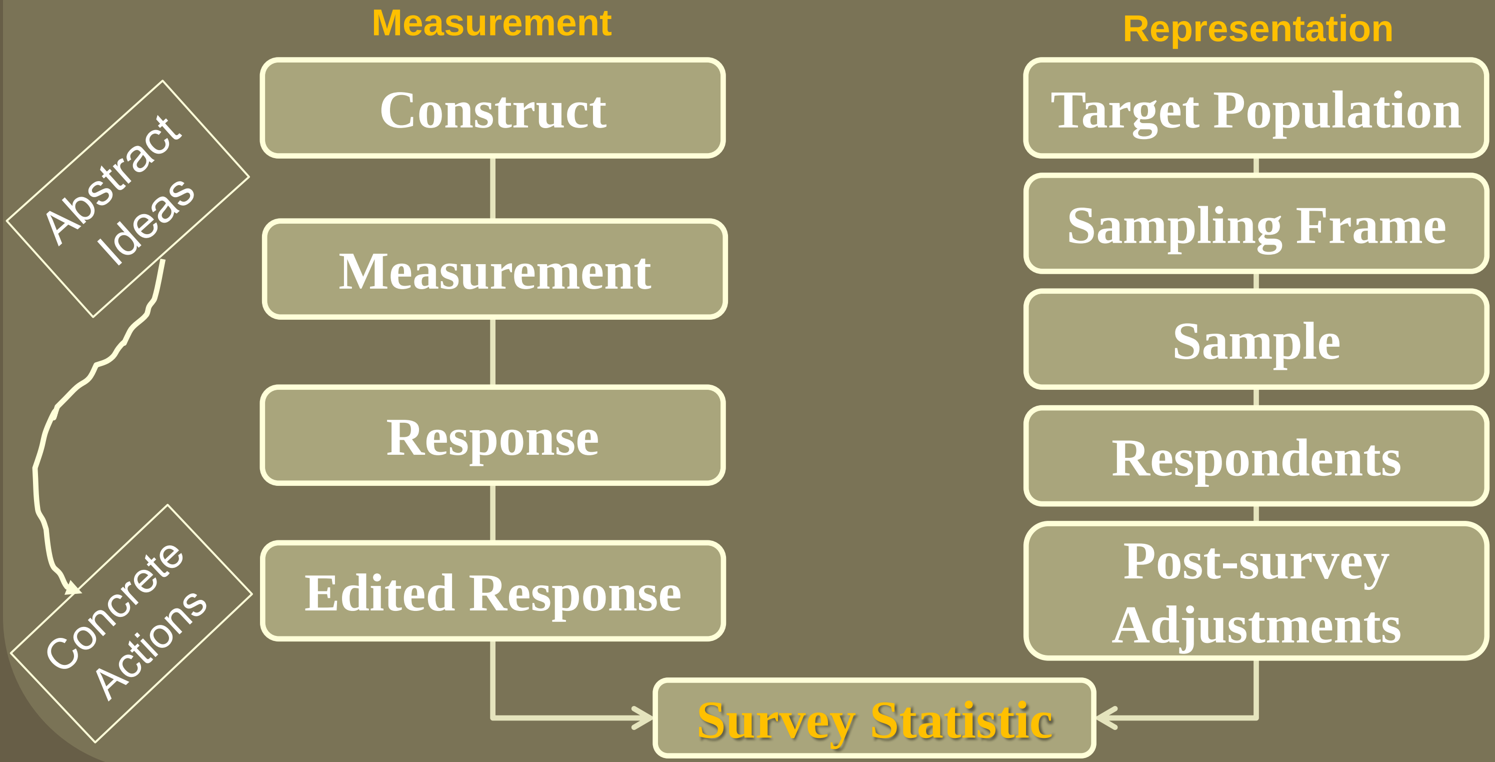
“The measurement of constructs”

&

“Descriptions of population attributes”

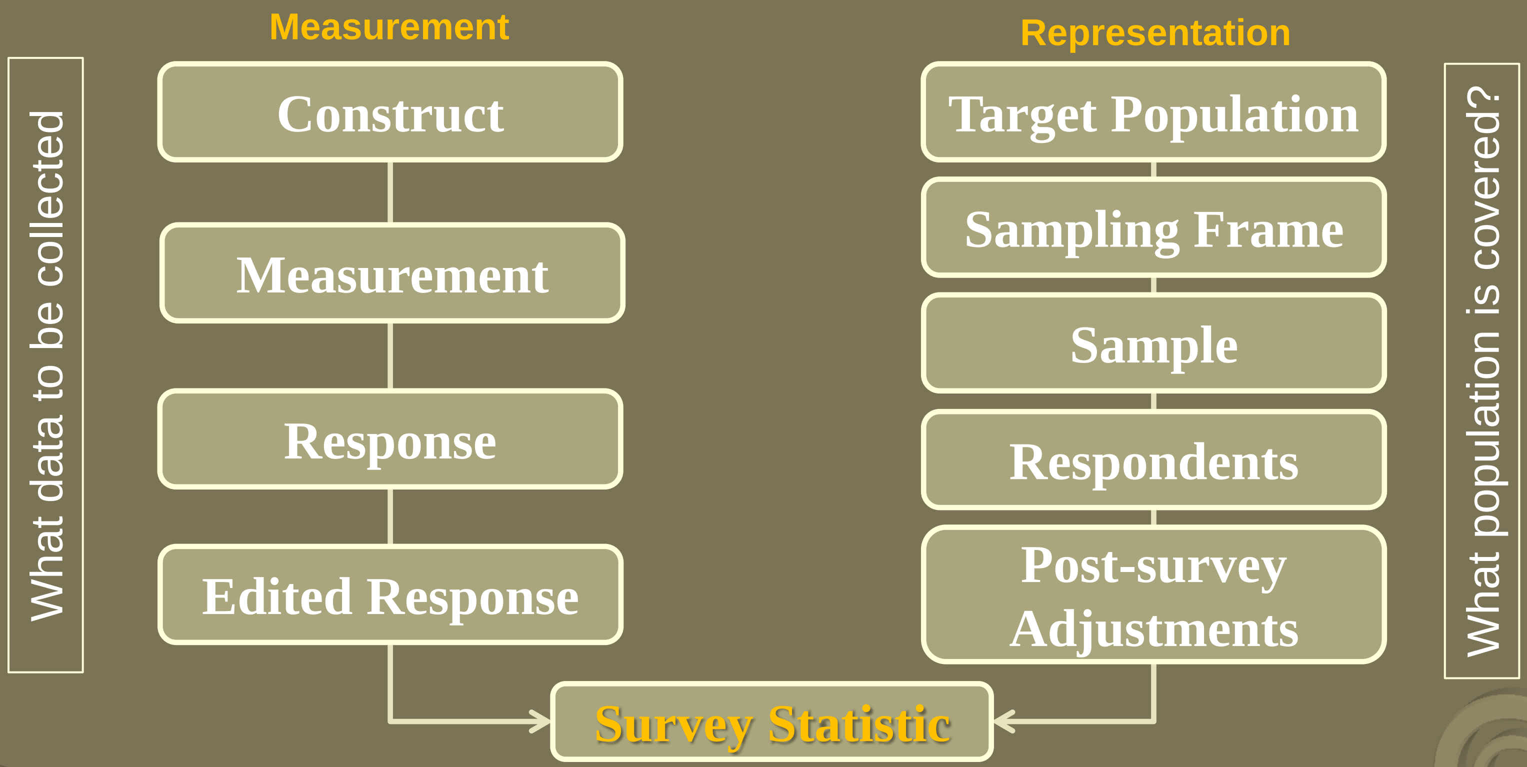
Preface

Life Cycle of a Survey from a Design Perspective



Preface

Life Cycle of a Survey from a Design Perspective



Measurement

What data to be collected

Construct

Measurement

Response

Edited Response

Representation

What population is covered?

Target Population

Sampling Frame

Sample

Respondents

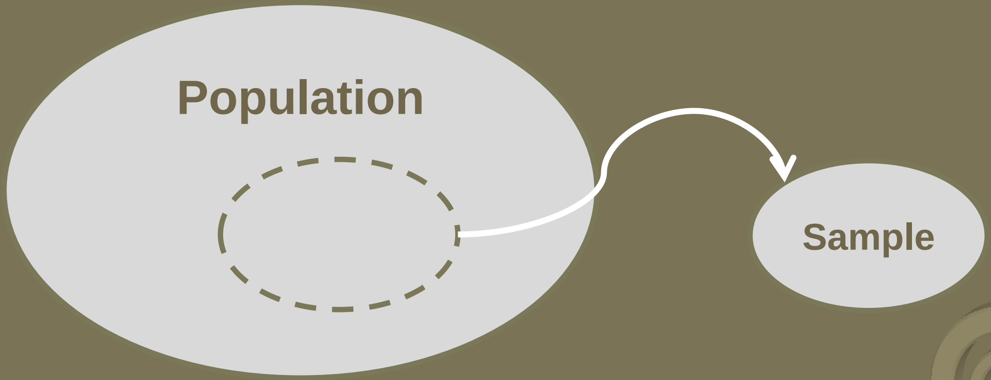
Post-survey Adjustments

Survey Statistic

Preface

Using Surveys to gather data

The manner in which the sample is selected from the population (sampling design) must be determined, so that the sample accurately reflects the population as a whole (representative sample)



Sample Design

We consider sample designs that satisfy following requirements:

- Probability Sampling
- Frame is Available

Sample Design

We consider sample designs that satisfy following requirements:

- Probability Sampling
- Frame is Available

each element of a population has a known (nonzero) probability of being included in the sample. This is the basis for applying statistical theory in the derivation of the properties of the survey estimators for a given design.

Sample Design

We consider sample designs that satisfy following requirements:

- Probability Sampling

- Frame is Available

a sampling frame that lists suitable sampling units that encompass all elements of the population

Type of Sample Design

Simple Random Sampling (SRS):

The simplest sample design which requires that each element have an equal probability of being included in the sample and that the list of all population elements be available

Type of Sample Design

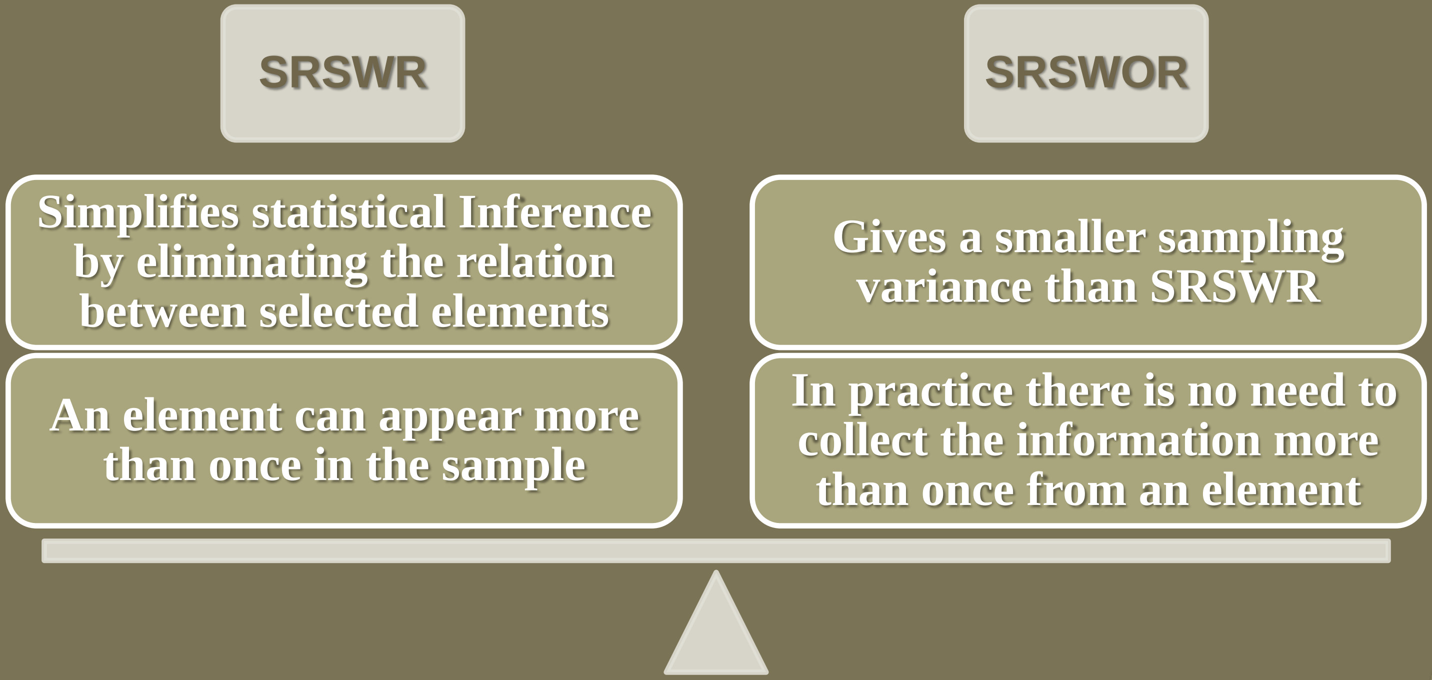
Simple Random Sampling (SRS):

The simplest sample design which requires that each element have an equal probability of being included in the sample and that the list of all population elements be available

Selection of Sample Units can be carried out with (without) replacement SRSWR (SRSWOR).

Type of Sample Design

Comparison of SRSWR and SRSWOR



SRSWR

Simplifies statistical Inference by eliminating the relation between selected elements

An element can appear more than once in the sample

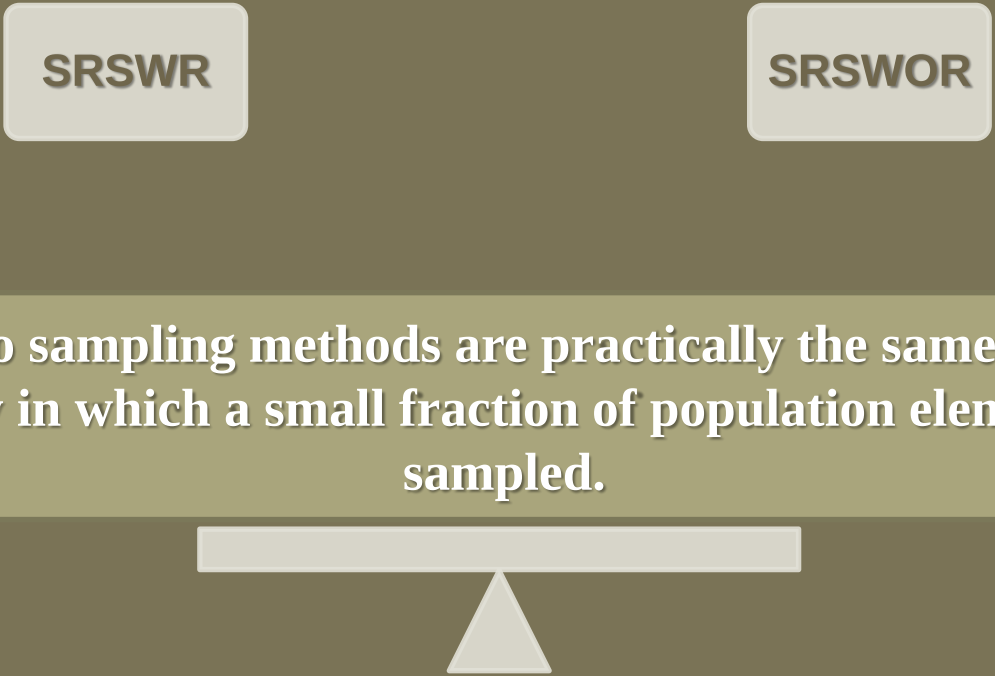
SRSWOR

Gives a smaller sampling variance than SRSWR

In practice there is no need to collect the information more than once from an element

Type of Sample Design

Comparison of SRSWR and SRSWOR

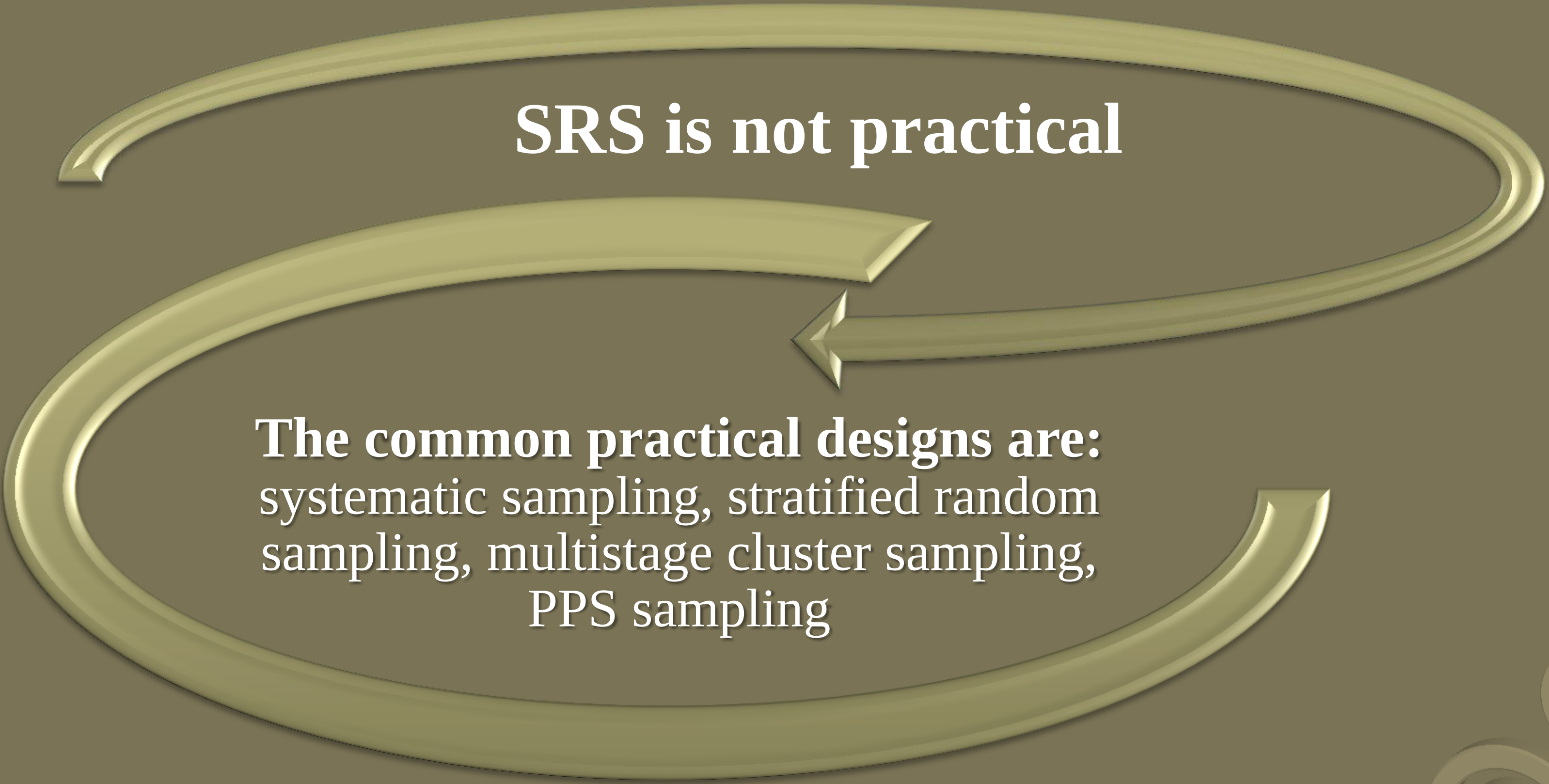


SRSWR

SRSWOR

These two sampling methods are practically the same in a large survey in which a small fraction of population elements is sampled.

Type of Sample Design



SRS is not practical

The common practical designs are: systematic sampling, stratified random sampling, multistage cluster sampling, PPS sampling

Type of Sample Design

Practical Methods Deviate from SRS

1. Inclusion probabilities for the elements may be unequal
2. Sampling unit can be different from the population element of interest

Type of Sample Design

Practical Methods Deviate from SRS

1. Inclusion probabilities for the elements may be unequal
2. Sampling unit can be different from the population element of interest

Complicate the usual methods of estimation and variance calculation
If proper methods of analysis are not used, can lead to a bias in estimation and statistical tests

Type of Sample Design

Systematic Sampling

Commonly used as an alternative to SRS because of its simplicity. It selects every k-th element after a random start (between 1 and k).

Assigns each element in a population the same probability of being selected (when $N = nk$, or 'N' is large)

Type of Sample Design

Systematic Sampling

Systematic sampling can give an unrealistic estimate, when the elements in the frame are listed in a cyclical manner with respect to survey variables and the selection interval coincides with the listing cycle.

Type of Sample Design

Systematic Sampling

Even when the listing is randomly ordered, unlike SRS, different sets of elements may have unequal inclusion probabilities. This complicates the variance calculation.

Type of Sample Design

Repeated Systematic Sampling

Instead of taking a systematic sample in one pass through the list, several smaller systematic samples are selected, going down the list several times with a new starting point in each pass.

Type of Sample Design

Repeated Systematic Sampling

Instead of taking a systematic sample in one pass through the list, several smaller systematic samples are selected, going down the list several times with a new starting point in each pass.

Guards against possible periodicity in the frame
Allows variance estimation directly from the data

Type of Sample Design

Stratified Random Sampling

Classifies the population elements into strata and samples separately from each stratum

It is used mostly because:

The sampling variance can be reduced if strata are internally homogeneous.

Type of Sample Design

Stratified Random Sampling

Sample allocation across the strata:

Proportionate

sampling fraction is uniform across the strata

Disproportionate

e.g. a higher sampling fraction is applied to a smaller stratum to select a sufficient number of subjects for comparative studies

Type of Sample Design

Stratified Random Sampling

Estimation is more complicated

Weighted Statistics should be used.

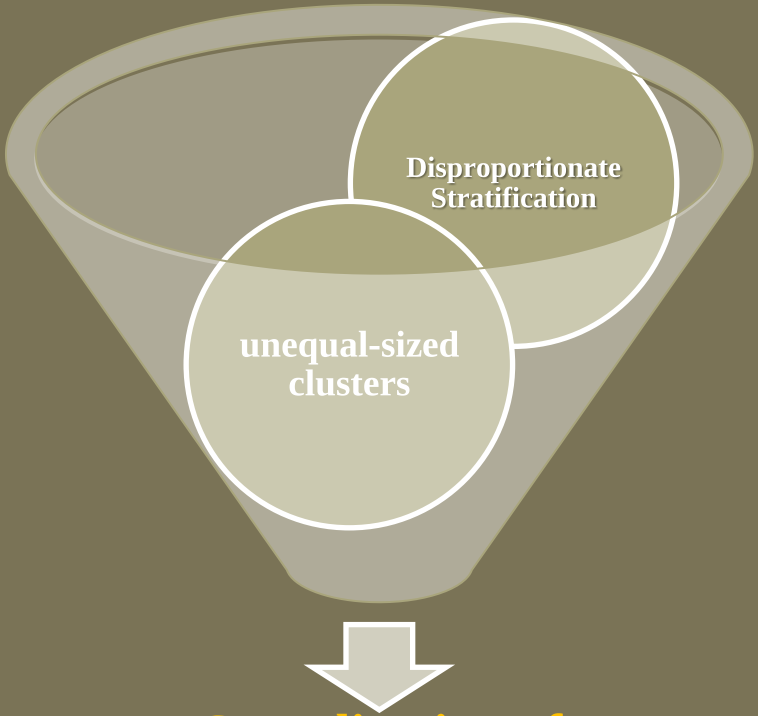
Type of Sample Design

Cluster Sampling

often a practical approach to surveys because it samples by groups (clusters) of elements rather than by individual elements.

- It simplifies the task of constructing sampling frames
- It reduces the survey costs.

Type of Sample Design



Complication of the estimation process

Type of Sample Design

One method to draw a self-weighting sample of elements in one-stage cluster sampling of unequal-sized clusters is to sample clusters with probability proportional to the size of clusters (PPS sampling)

mplica n of
the estimation process

Type of Sample Design

PPS Sampling

This requires that the true size of clusters be known.

Because the true sizes usually are unknown at the time of the survey, the selection probability is instead made proportional to the estimated size (PPES sampling).

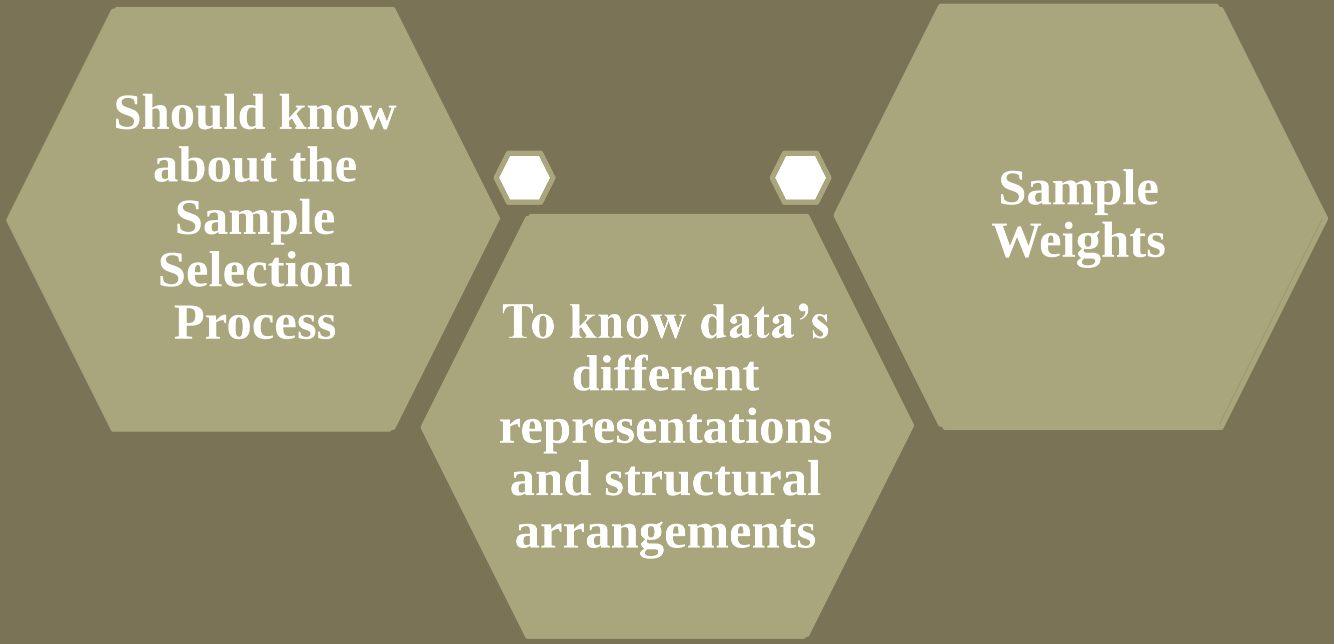
Type of Sample Design

Important Consequence of PPES Sampling

- The expected sample size will vary from one primary sampling unit (PSU) to another, i.e. the sample size is not fixed
- The denominator in the calculation of a sample mean, is a random variable
- The sample mean becomes a ratio of two random variables
- A ratio variable, requires special strategies for variance estimation

Nature of Survey Data

Inference from sample to population



Should know about the Sample Selection Process

To know data’s different representations and structural arrangements

Sample Weights

Nature of Survey Data

Sample Weights

Are used to reflect the differing probabilities of selection of the sample elements. The development of sample weights requires:

1. Keeping track of selection probabilities
2. Correcting for differential response rates
3. Adjusting the sample distribution by demographic variables to known population distributions (post-stratification adjustment)

Nature of Survey Data

Sample Weights

We may feel secure in the exclusion of the weights when a self-weighting design is used

But in practice, however, the self-weighting feature is destroyed by nonresponse and possible errors in the sampling frame(s)

Nature of Survey Data

Sample Weights

Two methods employed in an attempt to reduce the bias are “post-stratification” and “nonresponse” adjustments.

Nature of Survey Data

Sample Weights

Two methods employed in an attempt to reduce the bias are "post-stratification" and "nonresponse" adjustments.

Post-stratification involves assigning weights to bring the sample proportion and population proportion in demographic subgroups into agreement.

Nature of Survey Data

Sample Weights

Two methods employed in an attempt to reduce the bias are "post-stratification" and "nonresponse" adjustments.

Nonresponse adjustment inflates the weights for those who participate in the survey to account for the non-respondents with similar characteristics.

Nature of Survey Data

Sample Weights

Two methods employed in an attempt to reduce the bias are “post-stratification” and “nonresponse” adjustments.

Meanwhile, the weights are adjusted for frame under-coverage, and also over-coverage (non-eligible units)

Nature of Survey Data

Sample design

The sampling design specifies the probability of selection of each potential sample, and a proper estimator is chosen to reflect the design.

Sample Design = F (Sample Space, Probability, Estimator)

Nature of Survey Data

Sample design

Affects the estimation of standard errors, hence must also be incorporated into the analysis

SRSWR	SRSWOR
• $\widehat{V(\bar{y})} = \frac{s^2}{n}$	• $\widehat{V(\bar{y})} = \frac{s^2}{n}(1 - \frac{n}{N})$

Nature of Survey Data

Sample design

Affects the estimation of standard errors, hence must also be incorporated into the analysis

SRSWR	SRSWOR
• $\widehat{V(\bar{y})} = \frac{s^2}{n}$	• $\widehat{V(\bar{y})} = \frac{s^2}{n}(1 - \frac{n}{N})$

When n/N is small are the same

Nature of Survey Data

Sample design

Affects the estimation of standard errors, hence must also be incorporated into the analysis

$\widehat{V(\overline{y})}$ **in Stratified Sampling** $\leq \widehat{V(\overline{y})}$ **SRS** $\leq \widehat{V(\overline{y})}$ **Cluster Sampling**

Nature of Survey Data

Sample design

Affects the estimation of standard errors, hence must also be incorporated into the analysis

$\widehat{V(\overline{y})}$ **in Stratified Sampling** $\leq$ $\widehat{V(\overline{y})}$ **SRS** $\leq$ $\widehat{V(\overline{y})}$ **Cluster Sampling**

Neyman, or Optimal Allocation

Positive Intra-correlation Coefficient

Complexity of Analyzing Survey Data

Two essential aspects of survey data analysis are:

- Adjusting for the differential representation of sample observations

- Assessing the loss or gain in precision resulting from the complexity of the sample selection design.

Complexity of Analyzing Survey Data

Adjusting for Differential Representation: The Weight

Expansion Weight: The reciprocal of the selection probability

Developing the Weight by Post-stratification

Adjustment Factor: $\dfrac{\text{Population Distribution}}{\text{Sample Distribution}}$

We should check for extremely large values of weight, it happens when the sample size in a post-stratum is very small, and so not reliable. If so, post-stratum should be collapsed.

Complexity of Analyzing Survey Data

Assessing the Loss or Gain in Precision: The Design Effect

- Ratio comparing the variance of some statistic from any particular design to that of SRSWR/SRSWOR
- Used to assess the loss or gain in precision of sample estimates from the design used, compared to a SRSWR design

Complexity of Analyzing Survey Data

Assessing the Loss or Gain in Precision: The Design Effect

- Ratio comparing the variance of some statistic from any particular design to that of SRSWR/SRSWOR
- Used to assess the loss or gain in precision of sample estimates from the design used, compared to a SRSWR design

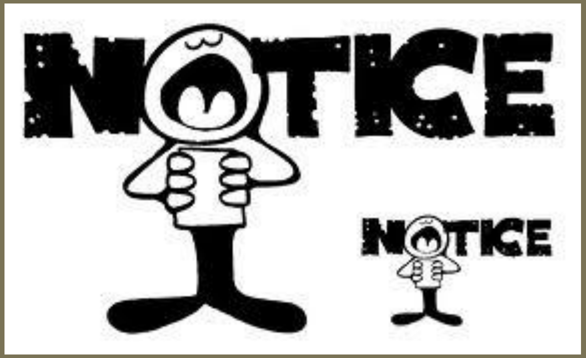
Design Effect less than ‘1’: Fewer observations needed

Design Effect more than ‘1’: More observations needed

Complexity of Analyzing Survey Data

Assessing the Loss or Gain in Precision: The Design Effect

In complex surveys the design effect is usually calculated based on the variance of the weighted statistic under SRSWOR design.



Strategies for Variance Estimation

The estimation of the variance of a survey statistic is complicated:

- By the complexity of the sample design, as seen in the previous chapters
- By the form of the statistic

Strategies for Variance Estimation

Variance Estimation Methods:

- Replicated Sampling
- Balanced Repeated Application
- Jackknife-repeated Replication (JRR)
- Bootstrap Method
- Taylor Series Method

Strategies for Variance Estimation

Replicated Sampling:

- Selecting a set of replicated subsamples, each subsample be drawn independently using an identical sample selection design.
- An estimate is made in each subsample by the identical process.
- Sampling variance of the overall estimate can be estimated from the variability of these independent subsamples' estimates

Preparing For Survey Data Analysis

Data Requirements for Survey Analysis:

It is necessary that data set include the **weights** and the identification of **sampling units** (PSU, USU) and **strata**.

Preparing For Survey Data Analysis

Importance of Preliminary Analysis:

Survey data analysis begins with a preliminary exploration to see whether the data are suitable for a meaningful analysis.

- Examine whether there is a sufficient number of observations available in the various subgroups to support the proposed analysis, check number of observations with missing value, or extreme values (based on unweighted tabulation)

Preparing For Survey Data Analysis

Importance of Preliminary Analysis:

Handling the missing data:

- Excluding the observations with missing value (tends to underestimate the variance)
- Adjusting the weight to compensate the missing value, assuming that the there is no systematic pattern among the subjects with missing values
- To impute the missing values by some reasonable method

Preparing For Survey Data Analysis

Importance of Preliminary Analysis:

Handling the missing data:

- Mean imputation for continuous variables
- Hot deck imputation
- Regression imputation
- Multiple imputation.

Preparing For Survey Data Analysis

Importance of Preliminary Analysis:

Prior to analysis, it is also necessary to examine whether each of the PSUs has a sufficient number of observations. It is possible that some PSUs may contain only a few observations, or even none, because of nonresponse and exclusion of missing values.

Preparing For Survey Data Analysis

Importance of Preliminary Analysis:

Handling PSUs with insufficient observations:

- Combine with adjacent PSU in the same stratum
- Combine a stratum with a single PSU with adjacent stratum

Preparing For Survey Data Analysis

Importance of Preliminary Analysis:

Handling PSUs with insufficient observations:

- Combine with adjacent PSU in the same stratum
- Combine a stratum with a single PSU with adjacent stratum

collapsing too many PSUs and strata destroys the original sample design

Preparing For Survey Data Analysis

Preliminary Analysis:

Explore the basic distributions of key variables.

Based on summary statistics, one may learn about interesting patterns and distributions

Investigate the existence of relations

Conducting Survey Data Analysis

1- Determine the Clustering, the Stratification variables, the Weights

2- Define the software to use

3- Conduct the Analysis

- Conducting Descriptive Analysis
- Conducting Tests
- Conducting Contingency Table Analysis
- Conducting Linear Regression Analysis