Introduction to Sampling Design

- **Basic concepts for sampling**
- **Non-Probability and Probability Sampling**
- **Steps for sample survey under probability sampling**

Cenker Burak METİN

Head of Survey and Questionnaire Design Group

Basic concepts for sampling

Sampling is a means of selecting a subset of units from a population for the purpose of collecting information for those units to draw inferences about the population as a whole.

- A perfect sample would be a "scaled-down" version of the population, mirroring every characteristic of the whole population. (almost impossible)
- A good sample would be **representative**

Survey statisticians want to give results as close as possible to the true values. It depends on

- ✓ The type of population (the distribution of target variable)
- ✓ The sample size
- ✓ The method of sample selection
- ✓ The estimation procedure

Basic concepts for sampling

Observation unit: An object on which a measurement is taken. This is the basic unit of observation, sometimes called as an element

Target population: The complete collection of observations we want to study.

Sampled population: The collection of all possible observation units that might have been chosen in a sample; the population from which the sample was taken.

Parameter: Certain value that is calculated from entire population and mostly unknown

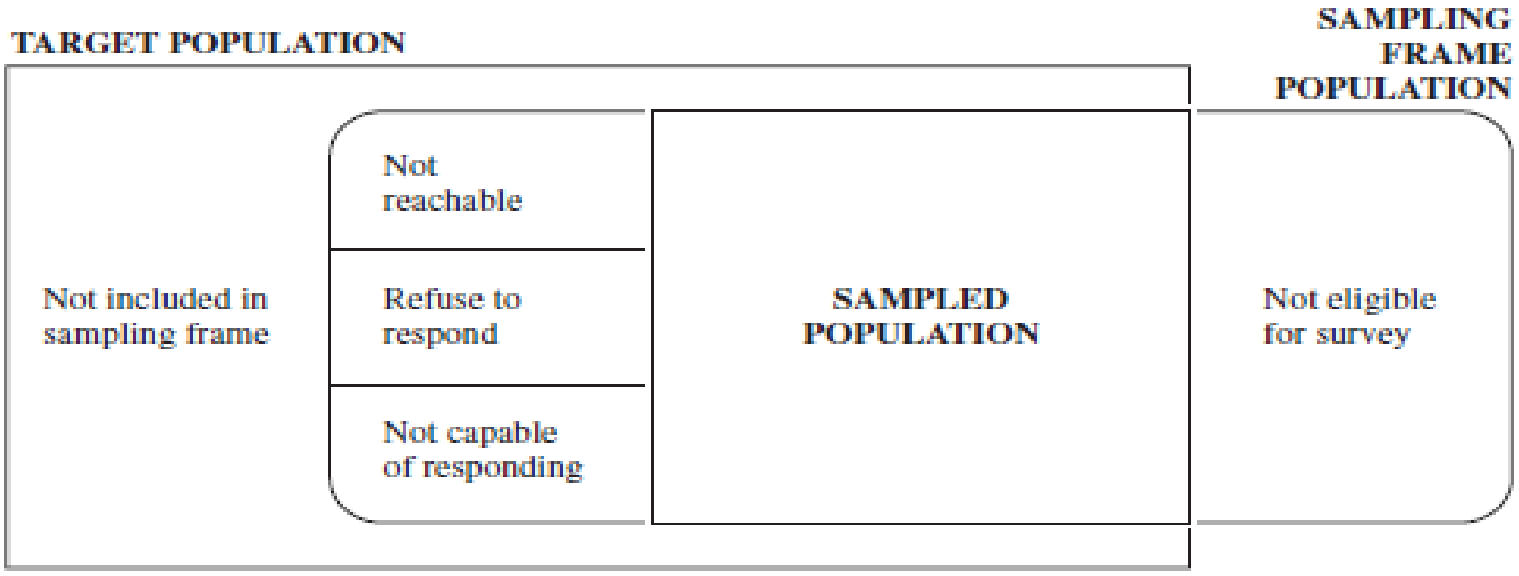
Sampling design: A definite plan for obtaining a sample from a given population.

Sampling unit: A unit that can be selected for a sample. We may want to study individuals, but do not have a list of all individuals in the target population. Instead, households serve as the sampling units, and the observation units are the individuals living in the households.

Basic concepts for sampling

Sample Statistics: Values calculated from the values in a sample in order to estimate population parameters

Sampling frame: A list, map, or other specification of sampling units in the population from which a sample may be selected.



Source: Lohr, S., Sampling design and analysis (2009)

Non-Probability and Probability Sampling

There are two types of sampling: non-probability and probability sampling.

Non-probability sampling,

- uses a subjective method of selecting units from a population
- provides a fast, easy and inexpensive way of selecting a sample
- is the sample representative for the population ? Response of this question is crucial to make inferences about the population
- not give chance to each unit of the population including in the sample
- descriptive statistics give only information about the sample
- not produce any estimation for sampling errors and confidence intervals

Non-Probability and Probability Sampling

Probability sampling,

- involves the selection of units from a population based on the principle of randomization
- is more complex, time consuming and usually more costly
- reliable estimates can be produced along with estimates of the sampling error
- give chance to each unit of the population including in the sample

There are several different ways in which a probability sample can be selected.

Non-Probability Sampling

Survey statisticians need to have some strong assumptions to make inferences about the population.

Non-Probability sampling can be applied,

- if there is no sampling frame and impossible to construct any
- to studies that are used as an idea generating tool
- to studies that are used as a preliminary or follow-up steps of probability survey
- for exploratory or diagnostic studies
- to surveys where volunteers are the only chance for the research.

Non-Probability Sampling

Advantages of Non-Probability sampling:

+ Quick and convenient
+ Relatively inexpensive
+ Not require a survey frame

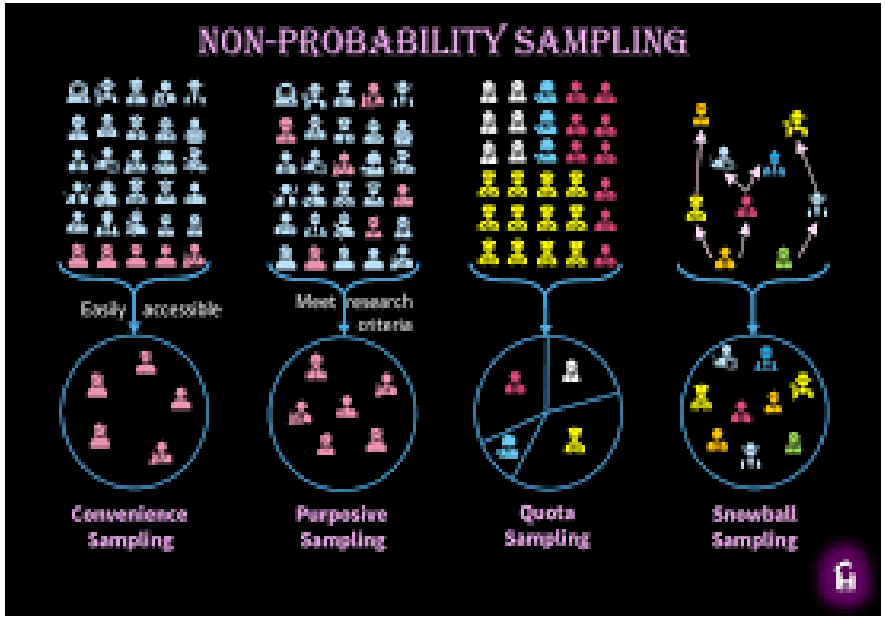
Disadvantages of Non-Probability sampling:

– Require strong assumptions about the representativeness of the sample
– Estimates are not reliable
– Sampling error of the estimates can not computed

Non-Probability Sampling

There are seven different types of Non-Probability sampling methods commonly used.

- Haphazard sampling
- Volunteer sampling
- Judgement sampling (Purposive sampling)
- Quota sampling
- Modified probability sampling
- Snowball sampling (Network sampling)
- Adaptive sampling



Probability Sampling

!!! The major strength is application of statistical theory

There are two main criteria for probability sampling:

1. The units be randomly selected

2. All units in the survey population have a non-zero inclusion probability in the sample and that these probabilities can be calculated.

There are many different types of probability sample designs.

- simple random sampling
- systematic sampling,
- probability-proportional-to size sampling,
- cluster sampling,
- stratified sampling,
- multi-stage sampling and multi-phase sampling

Probability Sampling

Advantages of Probability sampling:

+ Reliable estimates
+ Sampling error of the estimates can be produced
+ Relatively small samples may be enough for inferences

Disadvantages of Probability sampling:

– More difficult and longer to implement than Non-Probability sampling
– More expensive

How the probability sampling work ?

Steps for sample survey under probability sampling

Survey statisticians can prepare sample estimates (mean, total etc.) easily when **each unit has the same inclusion probability**.

- This kind of samples called as **self-weighting** samples
- To obtain the population totals inflate the sample results by a constant factor

Self-weighting designs have basic advantages like

- Easy to apply
- Because of non-complexity less analiysis required
- Variance of the estimates would not be high related to weights
- More comprehensible for non-statistical users and general public.

Steps for sample survey under probability sampling

However, there is various reasons for **not to use self-weighting** designs.

- To meet specific reporting requirements, different reporting domains or population subgroups may be sampled at different rates
- Due to defects in the frame, errors in selection, non-response etc. the resulting sample may not be self-weighting
- It may be possible to reduce bias and variance by weighting

Therefore, most of the time sample data have to be weighted to produce estimates for the population of interest. There are major advantages in following certain basic standards and a systematic approach in computing sample weights.

Steps for sample survey under probability sampling

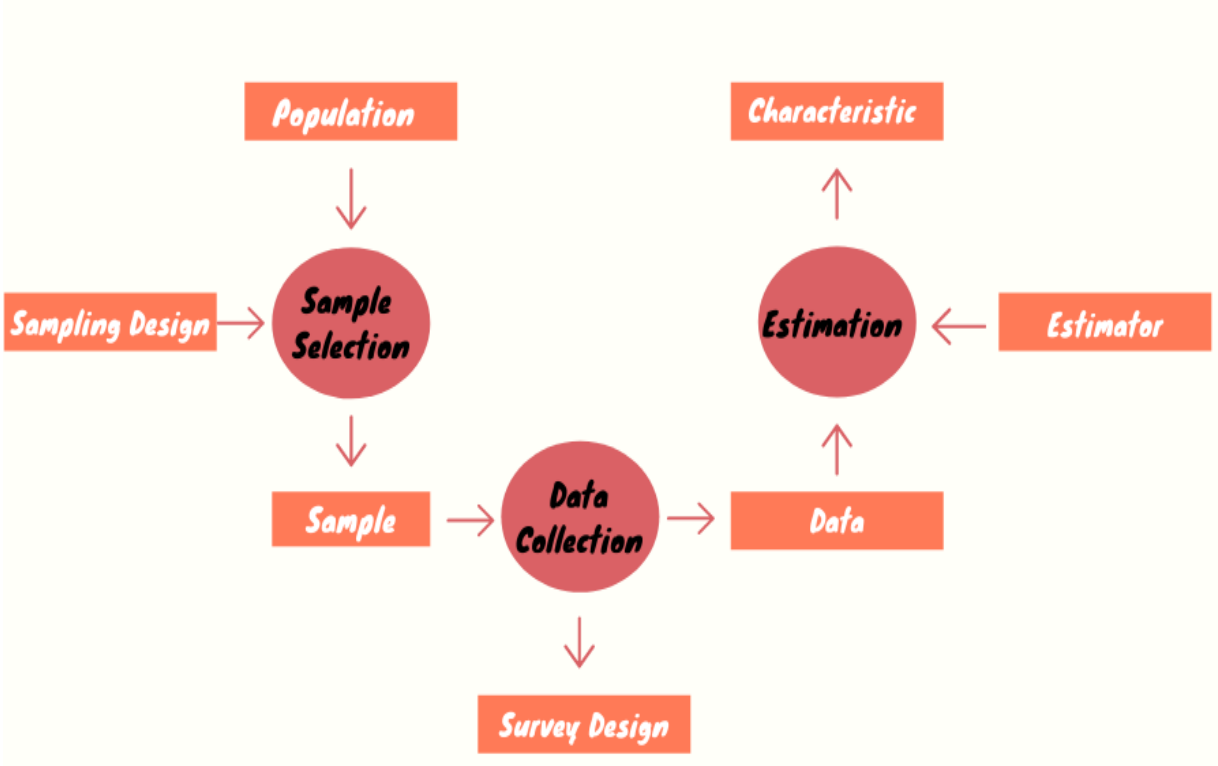
There are various information sources which can be used in a systematic manner to develop a step-by-step weighting procedure:

- sample design
- sample implementation,
- the sampling frame,
- significantly larger surveys with better coverage, higher response rates and more reliable information on certain characteristics of households and/or persons, compared to the survey under consideration;
- external sources - current registers, administrative records, population projections, etc. - providing information on population size and characteristics.

Steps for sample survey under probability sampling

Step-by-step procedure would be very useful to understand the different aspects of weighting. By the way, each step should be applied separately so that its contribution to the final weights can be identified.

- Design weights
- Non- response adjustment
- Calibration
- Trimming



References

Franklin, S., & Walker, C. (2010). Survey methods and practices. Statistics Canada. *Social Survey Methods Division, Ottawa*. (Originally published in October 2003)

Lohr, S. (2009). Sampling design and analysis (No. 519.52 L64).

Verma, V. (2014). Sampling: An Introduction, University of Siena, Siena, September 2014

Introduction to Sampling Design

- SRS, SYS, STR methods
- Cluster and Multistage Sampling methods
- How to construct a sampling frame
- Sample size determination and allocation

Cenker Burak METİN

Head of Survey and Questionnaire Design Group

Introduction to sampling techniques

A probability sample is selected from the sampling frame by using a sampling design consists of a combination of various sample selection techniques.

Alternative sampling techniques can be preferred with respect to statistical efficiency and/or other practical aspects like

- ✓ suitability to a given sampling task,
- ✓ requirements for application
- ✓ user friendliness.
- ✓ time and budget constraints
- ✓ the role of auxiliary information

In any sampling procedure, generally using auxiliary information on the population may be crucial. It can be useful in the construction of an efficient sampling design. Moreover, we may have efficiency gain at the estimation stage.

Introduction to sampling techniques

- **A simple random sample (SRS)** is the simplest form of probability sample.

  ✓ SRS of size n is taken when every possible subset of n units in the population has the same chance of being the sample.

  ✓ SRS provides a reference scheme when assessing the gain from the use of auxiliary information

- In **a systematic sample (SYS)**, a starting point is chosen from sampling frame using random no

  ✓ That unit, and every kth unit thereafter, is chosen to be in the sample.

  ✓ For standard application of SYS auxiliary information does not play a role

  ✓ SYS can be more efficient than SRS if there is a certain relationship between the ordering of elements in the sampling frame and the values of the study variable.

Introduction to sampling techniques

- **Stratified sampling (STR)** relies strongly on the use of auxiliary information.

  ✓ Sampling frame is first divided into non-overlapping subpopulations called strata,

  ✓ STR can be more efficient than SRS if the strata are internally homogeneous with respect to the study variable.

- **In cluster sampling (CLU),** the population is assumed to be readily divided into naturally formed subgroups called clusters.

  ✓ Firstly, clusters is drawn from the sampling frame, then all elements of the sampled clusters are taken (one-stage cluster sampling), or a sample of elements is drawn (two-stage cluster sampling).

  ✓ If the clusters are internally homogeneous, then CLU is less efficient than SRS.

!!! The availability of the auxiliary information

Introduction to sampling techniques

Parameters, estimators and quality measures

- ✓ Let our **parameter of interest** be population total $t = \sum_{k=1}^{N} y_k$ of study variable y.
- ✓ An **estimator** of the population total t is denoted by $\hat{t}$.
- ✓ The sample mean $\bar{y} = \sum_{k=1}^{n} y_k / n$ which is calculated using the n sample measurements.
- ✓ Using the sample mean, an estimate for the population total is calculated as $\hat{t} = N \times \bar{y}$
- ✓ For complex designs more complex evaluations needed

!!! In survey sampling, estimators are preferred that fulfil certain theoretical properties.

- **Unbiasedness**, $E(\hat{t}) = t$
- **Consistency** is a somewhat weaker property, $n\uparrow \ \hat{t} \to t$
- **Precision** of an estimator refers to its variability
- **Accuracy** of an estimator refers to $MSE(\hat{t}) = Var(\hat{t}) + Bias^2(\hat{t})$

Introduction to sampling techniques

Parameters, estimators and quality measures

!!! In survey sampling practice, estimators are used that are unbiased or at least consistent.

A challenge for survey statistician is for a given sampling **task to obtain efficient estimators whose design variances are as small as possible**. This is for high reliability of the results calculated by using the collected sample survey data.

Simple Random Sampling (SRS)

- ✓ SRS is often regarded as the basic form of probability sampling.
- ✓ If there is no previous info it may be applicable. (population assumed homogenous)
- ✓ Ensures that each element has an equal probability of selection. Thus, **SRS is an equal-probability sampling design**.
- ✓ SRS sets a baseline for comparing the relative efficiency of a sampling designs.
- ✓ For sampling of n elements, every element k in the sampling frame of N elements has exactly the same inclusion probability, $\pi_k = \pi = n/N$
- ✓ **SRSWOR** and SRSWR can be applied in practice. (why SRSWOR preferred ?)
- ✓ **Definition of SRSWOR** is as follows.

  Consider a population U of N elements SRSWOR is a method of selecting n elements out of U such that all possible subsets of U of size n have the same probability of being drawn as a sample. Note that there are $\binom{N}{n}$ possible subsets of U of size n.

Simple Random Sampling (SRS)

✓ Estimator of t can be written simply as $\hat{t} = N\sum_{k=1}^{n} y_k/n = N\bar{y}$, or

$$\hat{t} = \sum_{k=1}^{n} y_k/\pi = \sum_{k=1}^{n} y_k/(n/N) = \sum_{k=1}^{n} w_k y_k$$

$$\hat{v}(\hat{t}) = N^2(1 - \frac{n}{N})(\frac{\hat{s}^2}{n}) \qquad \hat{s}^2 = \sum_{k=1}^{n}(y_k - \bar{y})^2/(n-1)$$

✓ Note that if the sampling fraction (n/N) is small, fpc is minor, because fpc is close to one.

✓ If the sampling fraction (n/N) is small the fpc for SRS-WOR will be close to 1.

✓ If the sample size n approaches the population size N, the variance estimate will reduce.

✓ Thus, in a census the sampling variance is zero.

Simple Random Sampling (SRS)

SRS has a number of advantages:

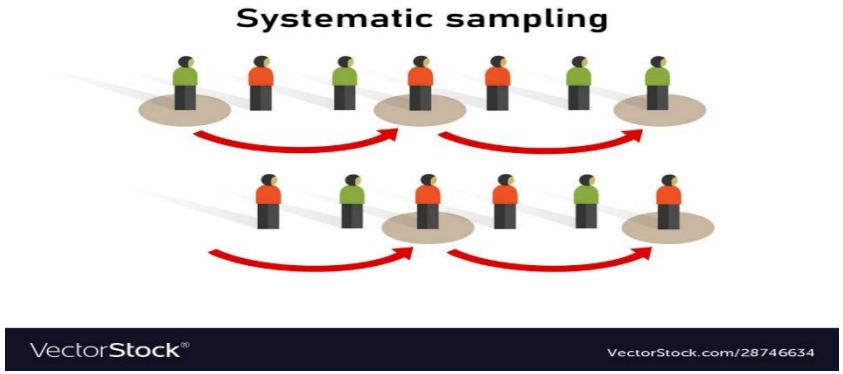
+ Simplest sampling technique
+ Requires no additional information
+ Needs no technical development

Disadvantages of SRS:

– Less statistically efficient estimates in order to no use of auxiliary information
– Expensive for face-to-face interviews
– Equal probability sampling may have bad effects in some situations

Systematic Sampling (SYS)

- ✓ SYS is a widely used where the sampling frame is an ordinary data base.
- ✓ SYS also is an equal probability sampling design
- ✓ Steps in the selection of SYS
  1. Define the sampling interval q = N/n, where an integer q is assumed.
  2. Select a random integer a with an equal probability of 1/q between 1 and q
  3. Select elements numbered a, a + q, a +2q, a +3q, ... , a + (n−1)q in the sample.
- ✓ In practice, there are several ways of selecting a systematic sample.
- ✓ For SYS, there is no known analytical variance estimator for the design variance,



Systematic Sampling (SYS)

✓ Estimation under SYS depends on the knowledge on the sorting order of the sampling frame:

1. **If the sorting order** of the sampling frame **can be assumed random** with respect to the study variables and all auxiliary variables, estimation with **SYS** will correspond to that of **SRSWOR.**

2. If the sampling frame is sorted by an auxiliary variable (or, several such variables), SYS sampling will produce a sample which tends to mirror correctly the structure of population with respect to the variables used in sorting. Sorting the frame before SYS sampling is called **implicit stratification.**

Systematic Sampling (SYS)

SYS has a number of advantages:

+ Proxy for SRS when there is no frame.
+ Does not require auxiliary frame information
+ Sample dispersion may be better than SRS
+ Estimates can be easily calculated
+ Simpler than SRS since only one random number is required.

Disadvantages of SYS:

– If the sampling interval matches some periodicity may result in a ‘bad’ sample
– Less statistically efficient estimates in order to no use of auxiliary information
– If a conceptual frame is used, the final sample size is not known beforehand.
– For variance estimation, SYS is often treated as if it were a SRS.

Stratified Sampling (STR)

- ✓ Target population is divided into non-overlapping subpopulations called strata by **STR**.
- ✓ SRS, SYS or PPS can be used for sample selection within the strata.
- ✓ There are several reasons to prefer stratified sampling:

  1. For administrative reasons,
  2. For both sampling and estimation, **stratification** allows for flexible stratum-wise use of auxiliary information
  3. Stratification may improve the efficiency if each stratum is homogeneous with respect to the variation of the study variables.
  4. Stratification can guarantee representation of small subpopulations or domains
  5. Different data collection techniques for different strata

- ✓ Stratification is particularly important in the case of skewed populations
- ✓ Regional, demographic and socioeconomic variables are typical stratifying variables.

Stratified Sampling (STR)

✓ The population is divided into H strata. An individual stratum is denoted with the index h=1,…, H and comprises N_h elements.

$$\sum_{h=1}^{H} N_h = N$$

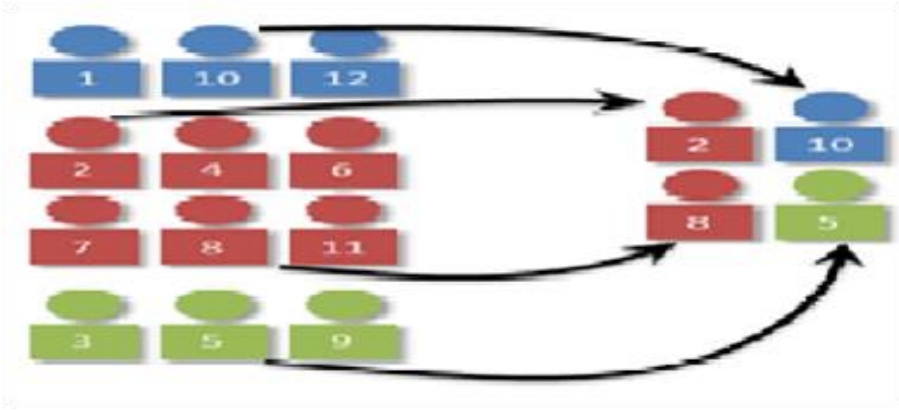
N_h of every stratum is known. The value of the target variable of element k in stratum h is denoted with Y_{hk}, for h=1,…, H, and k =1,…, N_h . The sample size of stratum h is expressed as n_h; by definition $\sum_{h=1}^{H} n_h = n$. The notation for the sample observations in stratum h is y_{hk}, with h=1,…, H, and k =1,…, n_h .

$$\hat{t} = \sum_{h=1}^{H} \hat{t}_h \text{ where } \hat{t}_h = \sum_{k=1}^{n_h} y_{hk}/\pi_{hk} = \sum_{k=1}^{n_h} w_{hk} y_{hk}$$

$$\hat{v}(\hat{t}) = \sum_{h=1}^{H} \hat{v}(\hat{t}_h) = N \sum_{h=1}^{H} \left(\frac{1-f_h}{f_h}\right) \left(\frac{N_h}{N}\right) \hat{s}_{yh}^2 \qquad f_h = n_h/N_h$$

Stratified Sampling (STR)

- ✓ Alternative strategies may be applied to determine stratum sample sizes.
- ✓ Generally, the overall sample size n is first fixed and then allocated to the strata.
- ✓ If each stratum is important to give estimate than ascertain large enough stratum sample sizes. Therefore, the stratum sample sizes n_h are first determined
- ✓ The most common allocation techniques for defining the stratum sample sizes are **proportional allocation, equal allocation, optimal or Neyman allocation.**
- ✓ Some special allocation techniques like power and comprimise allocations are also used by NSOs.



Stratified Sampling (STR)

STR has a number of advantages:

+ May increase the precision of overall population estimates
+ Provide statistically efficient domain estimators.
+ May be operationally or administratively convenient.
+ Protect against selecting a ‘bad’ sample.
+ Give a chance to use different sampling frames and procedures for different strata

Disadvantages of STR:

– Requires high quality auxiliary information for all units on the frame
– More costly and complex frame construction.
– If survey variables are not correlated to the stratification variables less efficient estimates
– Estimation is slightly more complex than SRS and SYS.

Cluster Sampling (CLU)

- ✓ **CLU** is the process of randomly selecting complete groups (clusters) from the sampling frame.
- ✓ Less statistically efficient sampling strategy than SRS but preferred for several reasons.
  1. Sampling clusters can greatly reduce the cost of collection.
  2. It is not always practical to sample individual units from the population.
  3. Allows the production of estimates for the clusters themselves
- ✓ Cluster sampling is a two-step process.
  1. Grouping the sampling frame into clusters
  2. Select a sample of clusters (PSU) and interview all units within the selected clusters or interview sample of elements (SSU) in the selected clusters
- ✓ Different sample designs can be used to select the clusters (SRS, SYS or **PPS)**

Cluster Sampling (CLU)

✓ For statistically efficient estimates, the units within a cluster should be as different as possible.

✓ The statistical efficiency of cluster sampling depends on

- how homogeneous the units within the clusters are
- how many population units are in each cluster
- the number of clusters sampled.

✓ If each cluster closely mirrors the population structure then efficient sampling possible

!!! When neighbouring units are similar, it is more statistically efficient to select many small clusters rather than a few, larger clusters.

For detailed description, notation and assumptions please look at Sampling theory (2012)

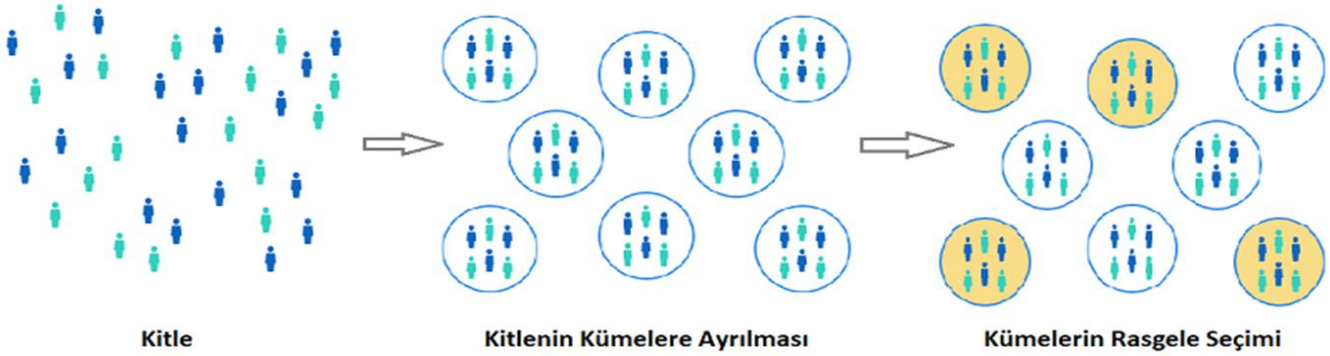
Cluster Sampling (CLU)

CLU has a number of advantages:

+ The cost efficiency may be high.
+ Sampling frame at the element level is not required

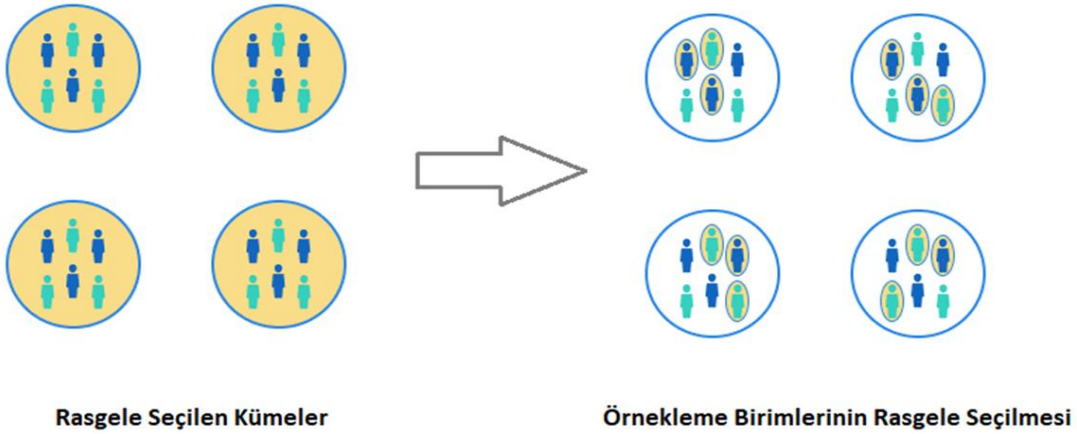
Disadvantages of CLU:

– Less statistically efficient than SRS if the units within the clusters are homogeneous with respect to the study variables.
– Final sample size is not usually known in advance
– Survey organisation can be more complex
– Variance estimation would be more complex



Multi-Stage Sampling

- ✓ Multi-stage sampling is the process of selecting a sample in two or more stages.
- ✓ The first stage are selected units are PSU's, and second stages are called SSU's etc.
- ✓ In two-stage sampling, the SSU's are often the individual units of the population.
- ✓ Commonly used with area frames to overcome the inefficiencies of one-stage cluster sampling,
- ✓ Each stage of a multi-stage sample can be selected by any sampling technique.
- ✓ It is flexible.



Multi-Stage Sampling

Advantages of Multi-Stage sampling :

+ May be more statistically efficient sampling strategy than a one-stage cluster design
+ May reduce the travel time and cost of personal interviews
+ Not necessary to have a list frame for the entire population.

Disadvantages of Multi-Stage sampling :

– Usually not as statistically efficient as SRS
– The final sample size is not always known
– Survey organisation is more complex than for one-stage cluster sampling.
– Calculation of the estimates and sampling variance may be complex.

Unequal probability sampling

- ✓ If **the target variable Y** in the population is proportional to a particular **auxiliary variable X** sampling with unequal inclusion probabilities is preferable.
- ✓ X must be known for all elements in the population.
- ✓ The great advantage of unequal inclusion probabilities is that the variances of the estimators can be reduced substantially,
- ✓ **Probability-proportional-to-size (PPS) sampling** is one technique that uses auxiliary data and yields unequal probabilities of inclusion.
- ✓ If population units vary in size and these sizes are known, such information can be used during sampling to increase the statistical efficiency.
- ✓ One advantage of PPS is that stratification in size classes is no longer necessary
- ✓ PPS can yield dramatic increases in precision if the size measures are accurate and the variables of interest are correlated with the size of the unit.

References

Franklin, S., & Walker, C. (2010). Survey methods and practices. Statistics Canada. *Social Survey Methods Division, Ottawa*. (Originally published in October 2003)

Lohr, S. (2009). Sampling design and analysis (No. 519.52 L64).

Banning, R., Camstra, A., & Knottnerus, P. (2012). Sampling theory. Sampling design and estimation methods. Statistics Netherlands. The Hague, 87.

Lehtonen, R., & Djerf, K. (2008). Survey sampling reference guidelines: introduction to sample design and estimation techniques.

Implementation of Sampling Techniques

- **Installation of R and R Studio**
- **Introduction to the R software with exercices**
- **Presentation of the R Packages “Sampling” and “Survey”**
- **Practical applications on data examples.**

Cenker Burak METİN

Head of Survey and Questionnaire Design Group

Installation of R and R Studio

Why R ?

- ✓ R is supported by academia
- ✓ R is an open source initiative,
- ✓ R is not just a statistics package, it's a statistical programming language
- ✓ R is designed to overcome the data scientist problems
- ✓ R is both flexible, powerful and endless
- ✓ no licence costs (freeware)
- ✓ allowed to copy and reu-use code (free software)
- ✓ source code is available and can be modified (open source)

Install R for UNIX platforms, Windows and MacOS from **https://www.r-project.org/**

Download Base R

Link for download: https://cran.r-project.org/bin/windows/base/

Installation of R and R Studio

Script Editors – Rstudio (https://www.rstudio.com/)

Advantages

- ✓ designed for R
- ✓ working with project philosophy
- ✓ script editor communicate with R
- ✓ objects in the workspace are listed
- ✓ version control systems (svn, git) are supported
- ✓ dynamical reports supported
- ✓ many add-ons (eg Rmarkdown, C++ code, ggplot2, . . .)

Download RStudio

Link for download: https://www.rstudio.com/products/rstudio/download/