AGENDA

1. Population of Interest, Sampling Frame and Sample Population
2. Sampling Basics and different sampling designs
3. Techniques in Random Sampling
4. Effect of the Sampling Design

Population of Interest, Sampling Frame and Sample Population

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Data

Population of Interest

- Population of interest, also called target population.
- Contains all the elements of which we want to measure certain characteristic (our *target variables*).
- Population data is commonly collected through censuses, i.e.
  - Housing and Population Census
  - Agricultural Census
  - Establishment Census
- or the administrative system, i.e.
  - Population registry
  - Business Register.

Sampling Frame

- From the population of interest we construct the sampling frame
- Sampling Frame acts as the basis for drawing the (survey) sample
- There are two main types of sampling frame:
  - List frames
  - Area frames
- Requirements for an ideal sampling frame are:
  - **Completeness** – each element is once and only once represented
  - **Timeliness** – the sampling frame needs to be up to data (i.e. full coverage of the population at the time of the survey)
  - **Informativeness** – provides information about the population elements (i.e. useful for stratification).

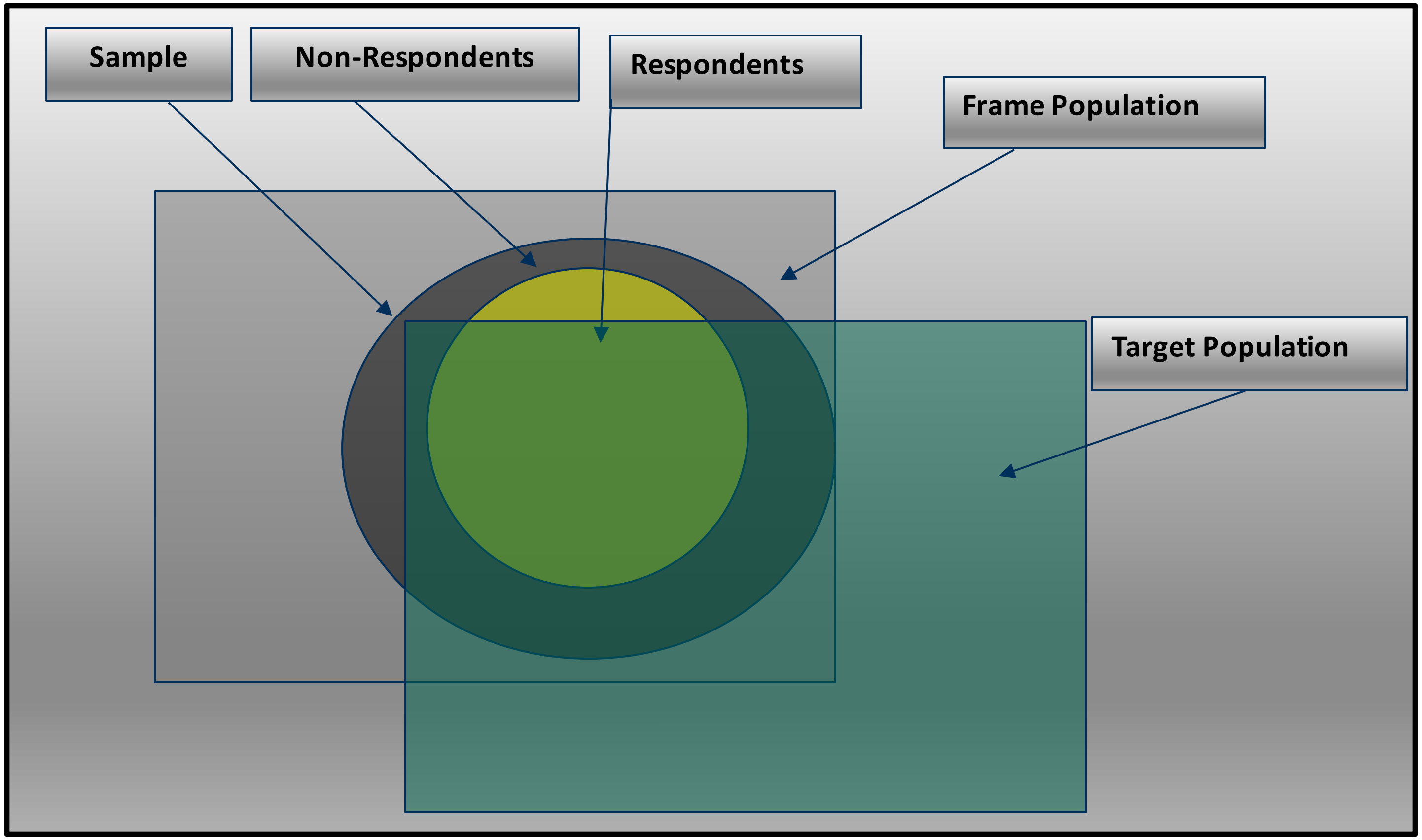
Sample Population

- Each element in the target population and subsequently in the sampling frame requires to have a non-zero probability of selection,
    - i.e. if a person is not registered in the census, then she/he may not be in the sampling frame, and therefore, has a zero probability of selection.

- A valid sample from this population must be selected at random, any other sample may lead to a biased sample.

- The final data will then be collected from the responding elements in the sample population, the remaining part are different types of non-response elements.

Reality Check



Sampling Basics and different sampling designs

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Data

Purpose of sampling

“To study a portion of the population – through observations at the level of the units selected, such as households, persons, institutions or physical objects – and make quantitative statements about the entire population”

Purpose of sampling

Why sampling?

- Saves **cost** compared to full enumeration
- Control the response **burden.**
- Easier to control **quality** of sample
- More **timely** results from sample data

Decreasing relevance in case of the use of new technologies for data collection.

Note on Notation

- A capital letter represents the true population and lower case letters represent the sample population, i.e.
  - Y is a value from the population,
  - y is a value from the sample,
  - N is the population size,
  - n is the sample size.
- A “hat” ($\hat{x}$) represents an estimate
- A “bar” ($\bar{x}$) represents a mean
- Variance is represented as either V(…) or v(…) or with S(…) or s(…)

Unit of analysis/(Final) Sampling Unit(s)

- Each unit provides us with a value y_i for the measurement of our variable of interest., with $i = 1,2,3 \ldots n$

- Most familiar units of analysis are persons, households, farms, and economic establishments

- Less familiar ones are land plots, segments of a coastline or lakes.

Parameters of interest

- Objective of sampling is to estimate the extend and/or the distribution of particular population parameters.
- The well familiar and most common parameters are :

  Population Averages (Means)

  Population Shares (Proportions)

- Less common parameters are:

  Regression coefficients

  Complex/Composite estimators

Unbiasedness of a Paramters

- When the estimate from a realized sample equals the true population value, then this estimate is called fully unbiased.

$$\Theta = \hat{\theta}(y_1, y_2, \dots, y_n)$$

- However this may only be true for a specific realization (or if n = N and correctly measured!),
- A less strict requirement for unbiasedness comes from convergence theory, which says, that the expcted value or the mean over ALL different realizations must converge to the true value.

$$\frac{\hat{\theta}_1 + \hat{\theta}_2 + \dots + \hat{\theta}_k}{k} \xrightarrow[k\rightarrow\infty]{} \Theta \text{ or } E(\hat{\theta}) = \Theta$$

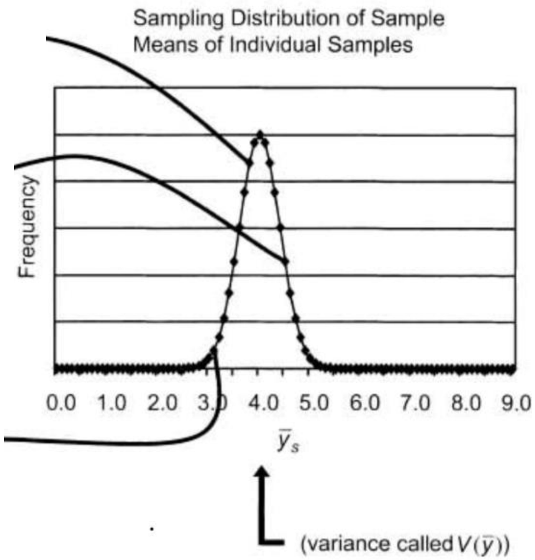
Sampling Distribution

- The Sampling Distribution is the distribution of the estimate(Mean, Total, …) across all possible samples.

- The survey is carried out only on a single, particular sample.

- The more centered the distribution is about its true population value, the less uncertainty is in our final estimate.

- The Mean Squared Error (MSE) measures the squared average deviation, and can be decomposed into:
  - MSE = VAR(x)+Bias(x)
  - $Bias(\bar{y}) = V(\bar{y})$
  - $Var(\bar{y}) = s^2$



Sampling Distribution

- Each sample realization in turn is therefore also subject to a particular selection strategy (= DESIGN) and its resulting selection probabilities.
- From a probabilistic point of view, we therefore have two probabilities:
  - One probability is the SELECTION within the sample, which is n/N.
    - » 1. order probability.
  - And another probability is the REALIZATION of a particular sample, which is 1/N(s)
    - » 2. order probability,
  - Where $N(s) = \binom{n}{N}$ possible samples.

Population Proportions

Mean

A proportion p is equal to the mean of a dummy variable.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad with\ y = \begin{cases} 1\ unit\ i\ has \\ 0\ unit\ has\ not \end{cases}$$

Variance

$$V(\hat{p}) \approx \frac{\hat{p}(1-\hat{p})}{n-1}$$

Population Mean

Mean

Similar to a proportion, just now $y \epsilon \mathbb{R}$ and

$$\hat{\boldsymbol{y}} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Variance

$$\mathrm{V}(\hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{Y})^2$$

Sample Size

- Depends on the distribution of our target variable(s) in the target population.
- The more homogenous this population is, the less sample size is required.
- The higher the sample size, the more precise the estimate will be.
- One important characteristic is, that probabilistic theory allows us to determine the degree of precision for our estimates, namely the standard error e.
  - If e is expressed relative to its underlying mean, then we talk about the Margin of Error (MoE), imagine you are asked to produce estimates with a precision of +/- 5%.
- The smaller e or MOE is, the more sample size is required.

Margin of Error (MoE)

The Margin of error is the absolute

$$P\left(\left|\hat{\theta} - \theta\right| \leq e\right)$$

or relative precision, required for our population estimates

$$P\left(\left|\frac{\hat{\theta} - \theta}{\theta}\right| \leq r\right)$$

And one of the basic parameters required to calculate our sample size

Sample Size

The sample size therefore depends mainly on two parameters:

- e (or MOE) as the desired level of precision
- and V the population variance.

All things equal, and leaving out a few steps, we can therefore write the sample size as the ratio

$$n = \frac{V(Y)}{e^2}$$

For a dichotomous variable this becomes: $n = \frac{p(1-p)}{e^2}$ and under the assumption of normal distribution: $\mathrm{n} = \frac{t_\alpha^2 \times p(1-p)}{e^2}$

And for a continuous one: $n = \frac{\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N}}{e^2}$ and with i.i.d. $n = \frac{t_\alpha^2 \times \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{N}}{e^2}$

Sample Size

Demonstration with Sampling App

Techniques in Random Sampling

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Data

Simple Random Sampling (SRS)

Also known as scientific sampling or probability sampling

Each unit has a **non-zero and known probability of selection**

Mathematical theory is available to predict the probability distribution of the sampling error

(*the error caused by observing a sample instead of the whole population*).

Simple Random Sampling (SRS)

Single stage, equal probability sampling

- Simple Random Sampling (SRS)
- Systematic sampling with equal probability

Stratified sampling

Cluster sampling

Multi-stages sampling

In real life those techniques are usually combined in various ways – most sampling designs are **complex**

Simple Random Sampling (SRS)

Single stage, equal probability sampling

Random selection of n “units” from a population of N units, so that each unit has an equal probability of selection

- N (population) → n (sample)
- Probability of selection (sampling fraction) = f = n/N
- Design Weight = 1/f =1/(n/N)

Is the most basic form of probability sampling and provides the theoretical basis for more complicated techniques

Simple Random Sampling (SRS)

Single stage, equal probability sampling (continued)

- **Simple Random Sampling**. The investigator mixes up the whole target population before grabbing “n” units.
  - » This is also our reference sample, when comparing other sampling designs.
- **Systematic Random Sampling.** The N units in the population are ranked 1 to N in some order (e.g., alphabetic). To select a sample of n units, calculate the step *k* (k= N/n) and take a unit at random, from the 1st k units and then take every kth unit.

Simple Random Sampling (SRS)

Single stage, equal probability sampling

Advantage

- self-weighting aka epsem (simplifies the calculation of estimates and variances)
- Gives a fully representative and unbiased estimate.

Disadvantages

- Required sampling frame may not be available
  - » List of all sampling units required
- May entail high transportation costs since in a national survey, final sampling units are spread out all over the country, therefore mostly used in phone- or web-survey.

Stratified Simple Random Sampling (STSRS)

Domain or Analytic Stratum

- Indicates two or more sub-groups in our population, for which we require our estimates to be at a particular level of precision.
- The population is divided into mutually exclusive subgroups called domains or strata, i.e.
  - Urban / Rural,
  - Administrative Areas,
  - Male / Female
- Then a random sample is selected from within each stratum/domain.
- Each domain estimate is calculated with its own weights, and independent of the other domain/stratum.
- Depending on the type of allocation overall estimates (i.e. national estimates) are a weighted sum of the individual estimates.

Stratified Simple Random Sampling (STSRS)

Design Stratum

- The purpose of a design stratum is to create homogenous sub-groups
  - Measured as the variance of the target variable in the stratum.
- If units within a stratum are homogenous, the sample size for the specific stratum can be lower
- A well stratified sample, can even require a lower sample size, than a SRS
  - See Neyman (1933)
- One problem in (design) stratification was the number of strata, however by simultaneously defining the optimal number of strata and the sample allocation, the stratification problem becomes nothing more than a simple optimization problem,
- either unconstrained: $Min\ e = \frac{\sum_1^k e_h}{k}$ or constrained $Min\ e = \frac{\sum_1^k e_h}{k}$ $\mathrm{s.t.:\ C} = \sum_1^k C_h$

Stratified Simple Random Sampling (STSRS)

The population is divided up into subgroups or “**strata**”.

A separate sample of units is then selected from each stratum.

There are two primary reasons for using a stratified sampling design:

- To potentially reduce sampling error by gaining greater control over the composition of the sample.
- To ensure that particular groups within a population are adequately represented in the sample.

These objectives are often contradictory in practice.

The sampling fraction generally varies across strata.

Sampling weights need to be used to analyze the data.

Stratified Simple Random Sampling (STSRS)

Establishment survey

- Stratification of establishments by economic activity and employment size

National household survey

- Geographic domains – regions, provinces
- Urban/rural
- Socio-economic groups

Agricultural survey

- Agro-ecological zones
- Land use
- Farm size

Stratified Simple Random Sampling (STSRS)

- Each stratum is treated as an independent population
- Estimate of stratified total is sum of stratum totals
- Estimate of stratified mean is weighted combination of stratum means, except if design is selfweighting.

Stratified Simple Random Sampling (STSRS)

Allocation

Three major types of sample allocation of sample units among the strata:

- Proportional allocation
  - Epsem, reduces overall variance
- Equal allocation
  - Allows for comparability
- “Optimum” allocation
  - Minimizes Stratum Variance

$$C = c_0 + \sum_{h=1}^{H} n_h c_h$$

$$n_h = \frac{(C - c_0)\, {}^{N_h s_h}\!\big/\!{}_{\sqrt{c_h}}}{\sum_{k=1}^{H} N_k s_k \sqrt{c_k}}$$

Stratified Simple Random Sampling (STSRS)

Estimation

Calculation of the Mean in a Stratified Sample:

$$\bar{x} = \sum_{h=1}^{L} \frac{N_h}{N} \quad \bar{x}_h$$

Calculation of the Variance in a Stratified Sample:

$$V(\bar{x}) = \sum_{h=1}^{L} (1 - n_h/N_h) \frac{N_h^2}{N^2} V(\bar{x}_h) \approx \sum_{h=1}^{L} \frac{N_h^2}{N^2} V(\bar{x}_h)$$

L = Number of strata

h = stratum number

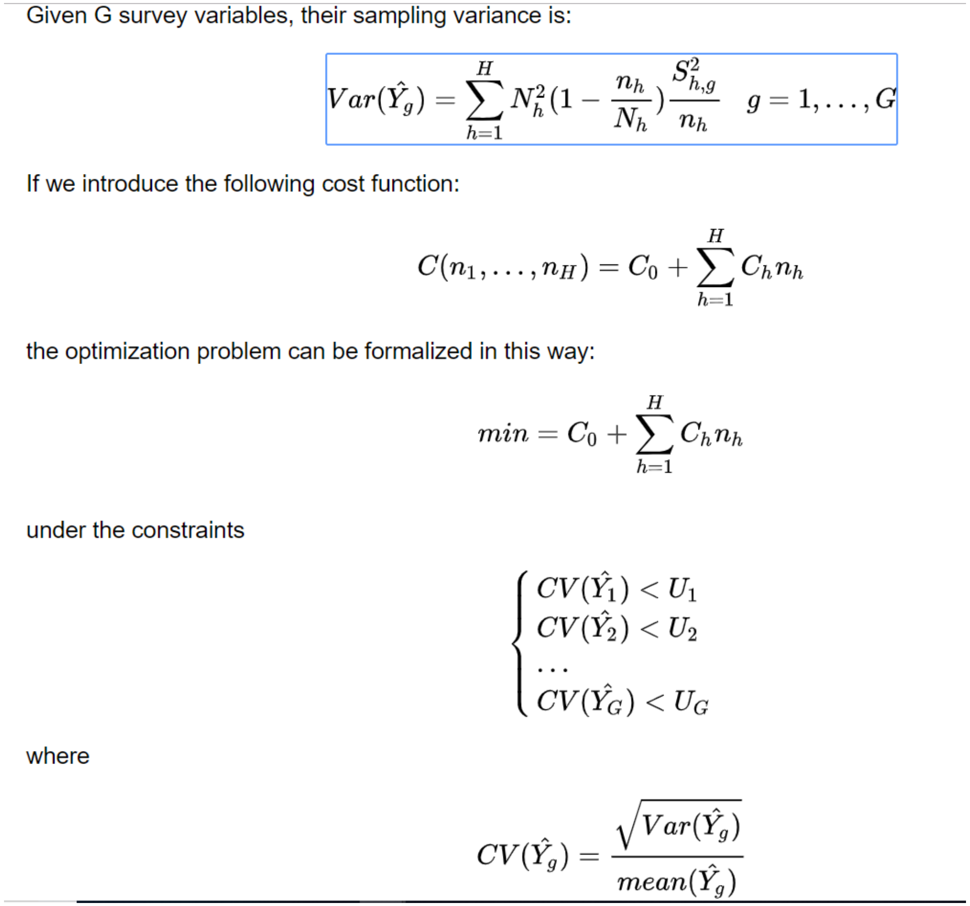
N_h =Population size in stratum h

n_h = sample size in stratum h

Stratification as an Optimization Problem

- When stratification is done purely for the purpose of sampling efficiency, then we can treat both, the number of strata as well as the corresponding sample size and its allocation as an optimization problem.
- The main reason behind this approach is based on the Neyman allocation, which attempts to minimize the overall variance through allocation.
- This algorithm used for this, also allows us to simultaneously optimize several variables and over several domains.
- The criteria according to which stratification is defined are crucial for the efficiency of the sample.
- With the same precision constraints, the overall size of the sample required to satisfy them may be significantly affected by the particular stratification chosen for the population of interest.

Stratification as an Optimization Problem



Given G survey variables, their sampling variance is:

$$\boxed{Var(\hat{Y}_g) = \sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_{h,g}^2}{n_h} \quad g = 1, \ldots, G}$$

If we introduce the following cost function:

$$C(n_1, \ldots, n_H) = C_0 + \sum_{h=1}^{H} C_h n_h$$

the optimization problem can be formalized in this way:

$$min = C_0 + \sum_{h=1}^{H} C_h n_h$$

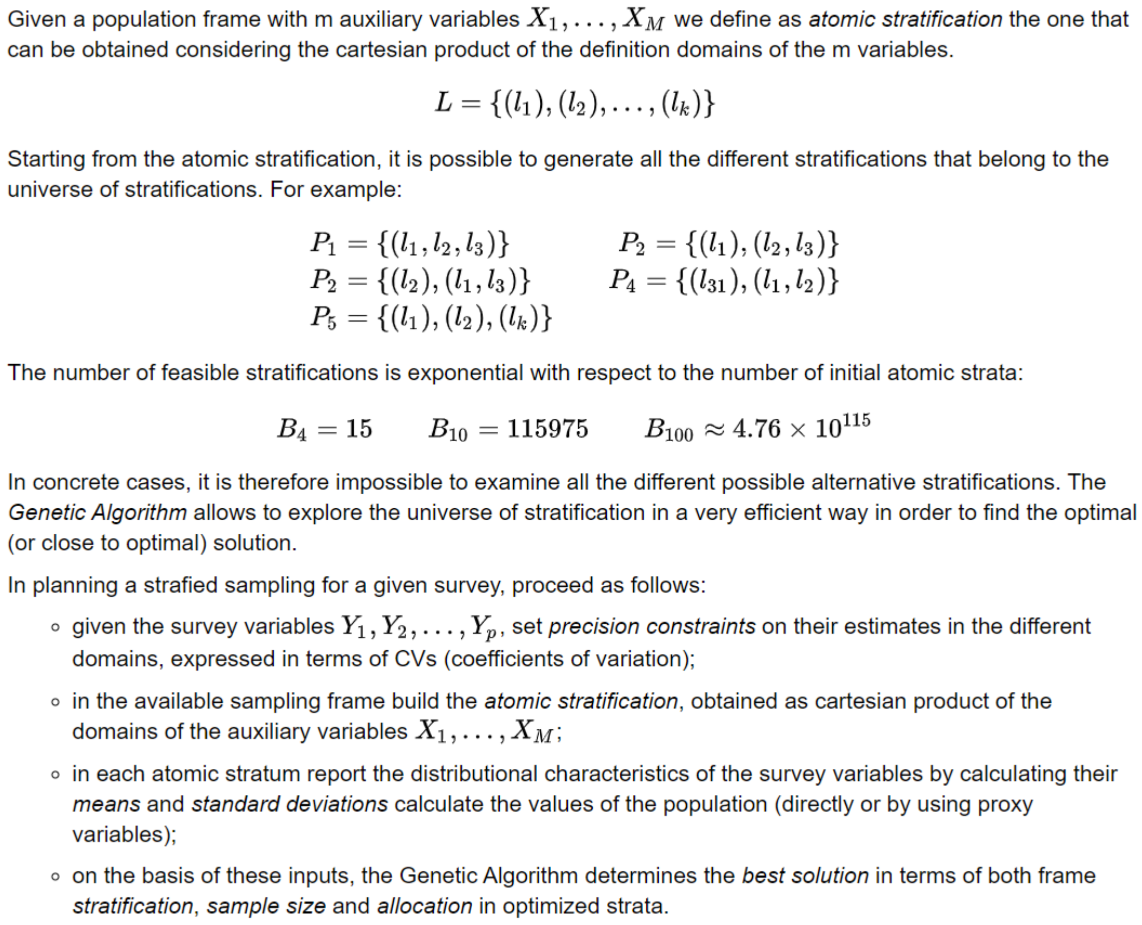
under the constraints

$$\begin{cases} CV(\hat{Y}_1) < U_1 \\ CV(\hat{Y}_2) < U_2 \\ \ldots \\ CV(\hat{Y}_G) < U_G \end{cases}$$

where

$$CV(\hat{Y}_g) = \frac{\sqrt{Var(\hat{Y}_g)}}{mean(\hat{Y}_g)}$$

Stratification as an Optimization Problem



Given a population frame with m auxiliary variables $X_1, \dots, X_M$ we define as *atomic stratification* the one that can be obtained considering the cartesian product of the definition domains of the m variables.

$$L = \{(l_1), (l_2), \dots, (l_k)\}$$

Starting from the atomic stratification, it is possible to generate all the different stratifications that belong to the universe of stratifications. For example:

$$\begin{array}{ll} P_1 = \{(l_1, l_2, l_3)\} & P_2 = \{(l_1), (l_2, l_3)\} \\ P_2 = \{(l_2), (l_1, l_3)\} & P_4 = \{(l_{31}), (l_1, l_2)\} \\ P_5 = \{(l_1), (l_2), (l_k)\} & \end{array}$$

The number of feasible stratifications is exponential with respect to the number of initial atomic strata:

$$B_4 = 15 \qquad B_{10} = 115975 \qquad B_{100} \approx 4.76 \times 10^{115}$$

In concrete cases, it is therefore impossible to examine all the different possible alternative stratifications. The *Genetic Algorithm* allows to explore the universe of stratification in a very efficient way in order to find the optimal (or close to optimal) solution.

In planning a strafied sampling for a given survey, proceed as follows:

- given the survey variables $Y_1, Y_2, \dots, Y_p$, set *precision constraints* on their estimates in the different domains, expressed in terms of CVs (coefficients of variation);
- in the available sampling frame build the *atomic stratification*, obtained as cartesian product of the domains of the auxiliary variables $X_1, \dots, X_M$;
- in each atomic stratum report the distributional characteristics of the survey variables by calculating their *means* and *standard deviations* calculate the values of the population (directly or by using proxy variables);
- on the basis of these inputs, the Genetic Algorithm determines the *best solution* in terms of both frame *stratification*, *sample size* and *allocation* in optimized strata.

Stratified Simple Random Sampling (STSRS)

Demonstration of efficiency gains through design stratification (*R Samplingstrata*)

Cluster Random Sample

- Units of analysis are divided into clusters of units.
- In population surveys, these clusters are commonly defined by the census (i.e. census districts).
- Clustering is done for the main purpose of saving costs in personal interview surveys.
- Cluster should be as heterogenous as possible, however very often this may not be the case.

Probability Proportional to Size Sampling

- Clusters are often selected with Probability Proportional to (estimated) Size (PP(e)S)
- This type of sample selection improves efficiency when the Measure of Size (MOS) is proportional to our variable of interest
- In multipurpose household surveys this may be true for some variables, but not for others, testing is required
- In some cases also a composite measure of size is possible
- Formally the selection probability is written as:

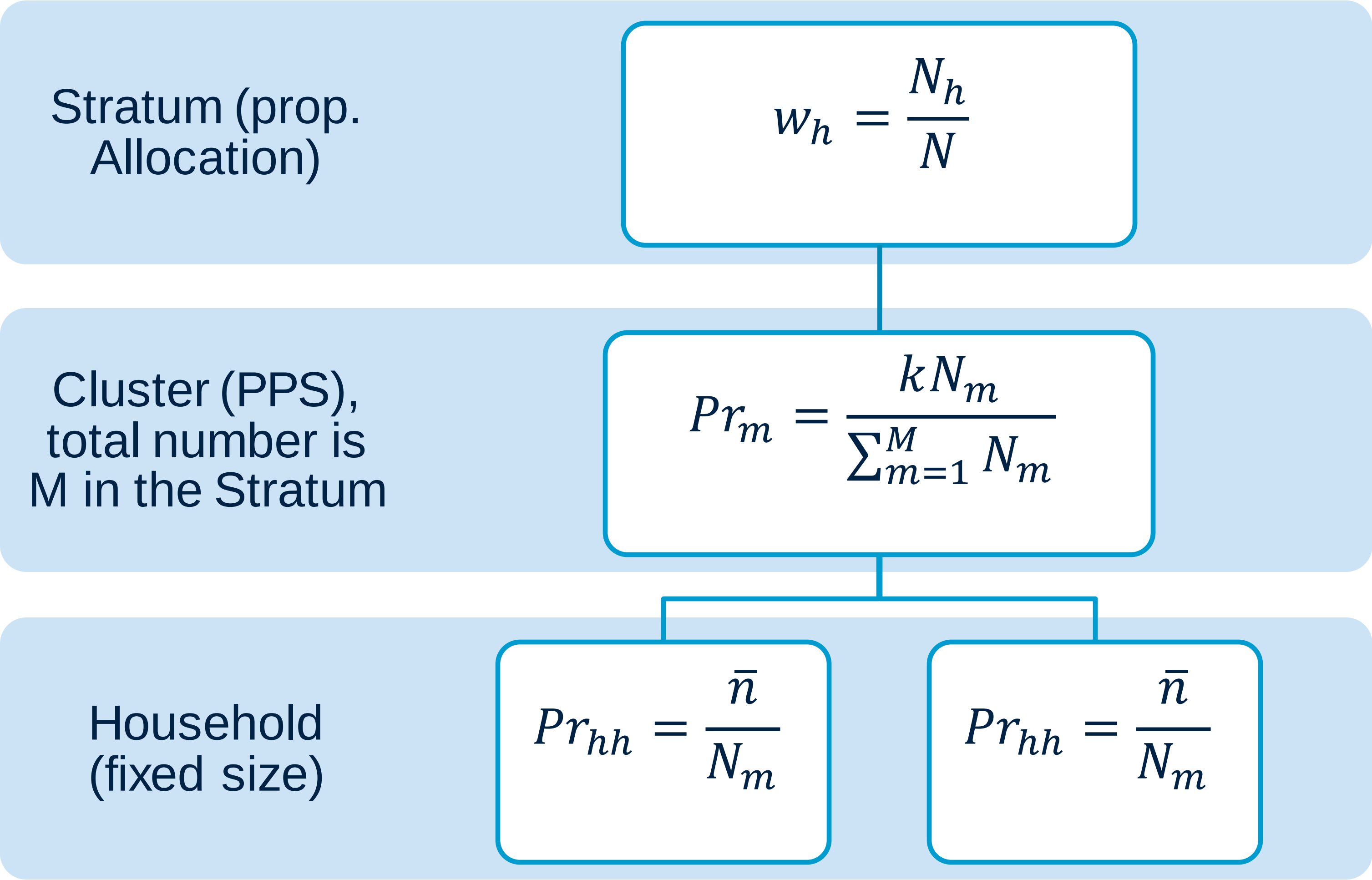
$$p = \frac{m * MOS_m}{\sum_{m=1}^{M} MOS_m}$$

Complex Survey Designs

- Is the result of putting all together, in other words, we use
  - ➢ Stratification to improve efficiency
  - ➢ Clustering to facilitate logistics and reduce costs

- When clustering is applied, then:
  1. A sample of clusters is taken independently within Domain/Stratum (Stage 1).
  2. Within each cluster we then take a sample of units of analysis (i.e. final sampling units). (Stage 2)

Complex Sampling Design Stages



Stratum (prop. Allocation)

$$w_h = \frac{N_h}{N}$$

Cluster (PPS), total number is M in the Stratum

$$Pr_m = \frac{kN_m}{\sum_{m=1}^{M} N_m}$$

Household (fixed size)

$$Pr_{hh} = \frac{\bar{n}}{N_m}$$

$$Pr_{hh} = \frac{\bar{n}}{N_m}$$

Effect of the Sampling Design

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Data

Design Effect - General

- Theoretical Sample Sizes are always based on SRS, it constitutes the “benchmark”-design.
- However any modification to SRS introduces a Design Effect (Deff), whereas the Deff for a stratified design may even be negative, in case of efficient design stratification, it most likely is positive for clustered designs.
- The total Deff can be decomposed into 3 components:
  - Design effect due to Weighting,
  - Design effect due to Stratification,
  - Design effect due to Clustering.

Design Effect - Clustering

- The theoretical definition of the design effect is the ratio between the variance of a specific design and a pure random sample
- The design effect is most relevant in clustering, and accounts for the average similarity (in the target variable) of the final sampling units within the cluster measured by its Intra Cluster Correlation.
- The more similar the values of the target variable are among the final sampling units, the higher the final sample size must be.

Design Effect - Importance

- Necessary to calculate sampling errors and design effects based on actual sample design

- If *deff* is ignored, the sampling errors will be underestimated, and the conclusions from any test of hypothesis or analysis will be biased

- Statistical software will always assume simple random sampling unless told otherwise

- In Stata standard errors and Deffs for complex designs can be calculated using the “svy” commands.
  - SPSS, R (ReGenesees package) all have specific packages

Design Effect – Calculation (Clustering)

where: $e^2_{complex} = e^2_{SRS} \times ceff = e^2_{SRS} \times [1 + \rho(m-1)]$

ρ = intraclass correlation coefficient – measure of homogeneity within a cluster

m = number of units per cluster

$$\hat{\rho} = \frac{\sum_{c=1}^{C}\sum_{j=1}^{m}\sum_{k\neq j}^{m}\left(x_{jc}-\bar{x}\right)\left(x_{kc}-\bar{x}\right)}{C\,m\left(m-1\right)\hat{s}^2}$$

C = number of clusters

Design Effects

The cluster effect measures the inefficiency of two-stage sampling relative to SRS

$$Deff = \frac{V_{complext}}{V_{SRS}}$$

In complex sample design (with many stages, stratification, etc.)

Type equation here.

$$deff = \frac{e^2_{complex}}{e^2_{SRS}} \rightarrow deff = \frac{n_{complex}}{n_{SRS}}$$

Thank you!

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Data