**NEC 2019**

08

**Fall**

**Bassirou Chitou**

[bdryan1210@gmail.com](mailto:bdryan1210@gmail.com)

Hurghada, Egypt

October 20 to 24, 2019

**Leaving No One Behind**

**National Evaluation Capacities (NEC) Conference 2019**

|  |
| --- |
|  |

**Introduction to**

**Sampling**

**for Development Evaluation**

Disclaimer 1

Fellow Evaluators,  
First, this workshop is **not** about “statistics” or “statistical methods”. Here, statistical methods will only serve as a “vehicle”. Second, this workshop is **not** about “STATA” or “XSTAT” or any software for that matter. Any software used will only serve as a “vehicle”.

Finally, this workshop is about “time tested principles” in “data collection”. These data in turn will help us discover evidences that makes our evaluation credible and usable.

Disclaimer 2

The contents, findings, and conclusions in this presentation are those of the author and do not necessarily represent the official position of the NEC nor that of UNDP.

Table of Contents

[1.1. Key Concepts 4](#_Toc20400888)

[1.1.1 Useful Definitions 4](#_Toc20400889)

[1.1.2. Random Sample Definition 6](#_Toc20400890)

[1.1.3. Sampling Without Replacement 6](#_Toc20400891)

[1.1.4. What a Random Sample is Not 6](#_Toc20400892)

[1.1.5. Advantages of a Random Sample 6](#_Toc20400893)

[1.2. Basic Types of Random Samples 7](#_Toc20400894)

[1.2.1. Simple Random Sampling (SRS) 7](#_Toc20400895)

[1.2.1.1. Definition 7](#_Toc20400896)

[1.2.1.2. How to Perform an SRS 7](#_Toc20400897)

[1.2.1.3. An Illustration 7](#_Toc20400898)

[1.2.2. Systematic Sampling 9](#_Toc20400899)

[1.2.2.1. Definition 9](#_Toc20400900)

[1.2.2.2. How to Perform a SYS 10](#_Toc20400901)

[1.2.2.3. Illustration 10](#_Toc20400902)

[1.2.3. Stratified Sampling 11](#_Toc20400903)

[1.2.3.1. Definition 11](#_Toc20400904)

[1.2.3.2. Two Types of Stratified Random Sampling 11](#_Toc20400905)

[1.2.3.3. STRS Illustration 11](#_Toc20400906)

[1.2.3.4. How to Perform an STRS 12](#_Toc20400907)

[1.2.3.5. Practical Application of an STRS 12](#_Toc20400908)

[2.1. What is a Multi-Stage Sampling 14](#_Toc20400909)

[2.1.1. Definition 14](#_Toc20400910)

[2.1.2. Reasons for Multi-Stage Sampling 14](#_Toc20400911)

[2.2. Two-Stage Cluster Sampling 14](#_Toc20400912)

[2.2.1. What is a Cluster 14](#_Toc20400913)

[2.2.2. How to Perform a two-stage cluster sampling: 15](#_Toc20400914)

[2.2.3. Advantages of Two-Stages Cluster Samples 17](#_Toc20400915)

[2.3. First-Stage Probability Proportional to Size Sampling 17](#_Toc20400916)

[2.3.1. What is PPS 17](#_Toc20400917)

[2.3.2. How to Perform a PPS sampling**:** 18](#_Toc20400918)

[2.3.3. Example of PPS Sampling 18](#_Toc20400919)

[2.3.4. Practice of PPS 20](#_Toc20400920)

[2.4. Second Stage Systematic Sampling 22](#_Toc20400921)

[2.4.1. Mapping and Listing of Second Stage Units 22](#_Toc20400922)

[2.4.2. Second-Stage Equal Systematic Sampling 24](#_Toc20400923)

[2.4.2.1. Advantage of the Second Stage SYS Scheme 24](#_Toc20400924)

[2.4.2.2. Use of an Excel template to Perform a Second-Stage SYS 24](#_Toc20400925)

[2.5. Possible Third-Stage Sampling 26](#_Toc20400926)

[2.5.1. Third-Stage Sampling a Subset of Already Sampled Households 26](#_Toc20400927)

[2.5.2. Third-Stage Sampling of One Respondent from Several Eligible 27](#_Toc20400928)

[2.6. The Importance of Sampling Weight 28](#_Toc20400929)

[2.6.1. What is a Sampling Weight? 28](#_Toc20400930)

[2.6.2. Why are weights needed? 28](#_Toc20400931)

[2.6.3. Illustrative Example 29](#_Toc20400932)

[2.6.4. Note and Caution 29](#_Toc20400933)

1 **Essentials of Probability Sampling in Evaluation**

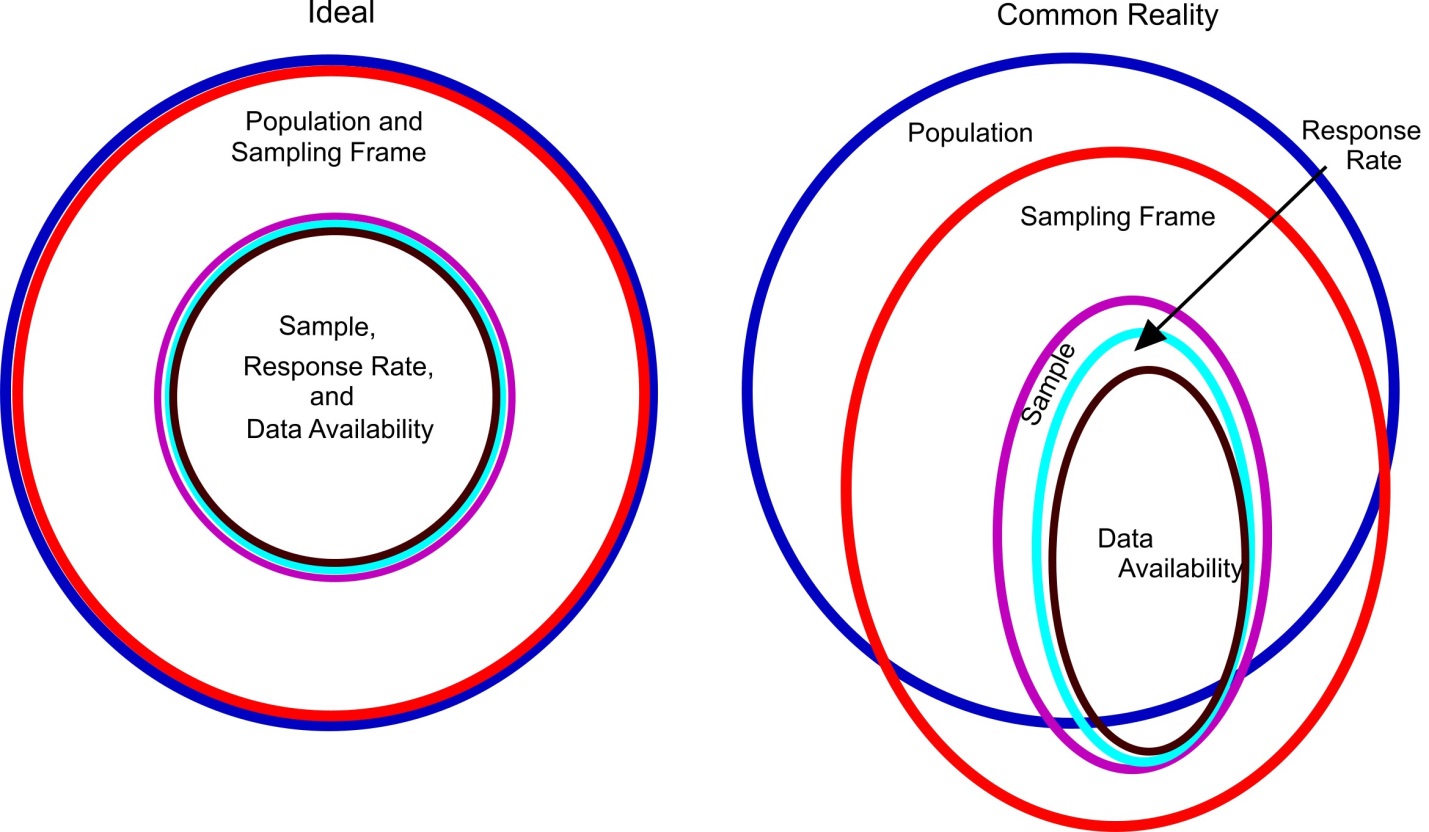
# Key Concepts

## Useful Definitions

Several terms are critical to the understanding of sampling theory and practice. The most common ones are:

* **Units of Analysis**: The type of entity on which data are collected and for which decision   
   are made (people, clinics, wells, reports, etc.).
* **Population**: The full set of units of analysis about which you desire to infer conclusions.
* **Sampling Frame**: A list of all units of analysis in a given population.
* **Census**: The collection of data from all units of analysis in a given population.
* **Sample:** A subset of units drawn from the sampling frame or sometimes directly from   
   the population using a known scheme.
* **Response Rate**: Percent of the intended units in the sample for which most of the   
   required data are actually collected.
* **Data Availability**: Percentage of “responding” units for which there are data on a given   
   variable.

**Figure 1: Relationship among Key Sampling Terms**



## 1.1.2. Random Sample Definition

A **random** sample is one in which the units of analysis have been selected with **known** and **non-zero probabilities** using a **random process**.

## 1.1.3. Sampling Without Replacement

In evaluation work, the random sampling is presumed to be done “**without replacement**,” meaning that **once** a unit has been selected to be in the sample, **it is not** placed back into the sampling frame for possible repeated selection. A unit is selected “**once and only once**”.

## 1.1.4. What a Random Sample is Not

It is important to understand that random sampling is Not the same as selecting cases by any of the following:

* Whoever walks by a given point in a market place;
* The first thirty patients who visit a health clinic on a given day;
* Cases that the evaluator “judges to be representative” of the population;
* Random sampling is quite different from a “**randomized trial**.” In many medical studies, people with a given medical condition are recruited with announcements and volunteer to participate (not a random sample). Then they are randomly assigned to either the treatment or control condition. It is possible for a randomized trial to use a random sample.

## 1.1.5. Advantages of a Random Sample

A probability sample has the following advantages. It:

* **precludes** any possibility of the evaluator consciously or unconsciously introducing **biases** in the sample;
* is **representative** of the population of interest;
* is also the **only type** of sampling **for which inferential statistics could be computed**? Tests of statistical significance and the computation of confidence intervals assume random sampling;
* Results are **generalizable** to the entire population of interest.

# Basic Types of Random Samples

There are several basic types of random samples. The most widely used ones are introduced below and then will be discussed in greater detail.

## Simple Random Sampling (SRS)

### Definition

|  |
| --- |
| **Simple random sample (SRS)**  In SRS, each unit of analysis has a unique identification number (ID number) or is assigned one. A random number of table or computerized random number generator indicates the ID numbers of the units to be selected for the sample. |

### How to Perform an SRS

To perform an SRS, use the following are key steps:

* **Define the population** carefully: indicate **what is within** and **outside** of it;
* Find or generate a **sampling frame** that lists all the units in the population or a very large proportion of them;
* Assign each a **unique ID number**. The numbers do not have to be consecutive;
* Decide on **the size of the sample** (called n);
* Use either a random number generator either:
* Online, such as [www.randomizer.org](http://www.randomizer.org/)
* MS Excel
* Statistical software such as STATA, R, SAS, or SPSS
* Old paper random table

### An Illustration

|  |
| --- |
| **ACTIVITY**  Assume you have a sampling frame numbered consecutively from 201 to 500. First, use the online tool: [www.randomizer.org](http://www.randomizer.org) and second, use MS Excel |
| **NOTES:** |

#### A. Using Randomizer.Org

* Go to [www.randomizer.org](http://www.randomizer.org) ;
* Scroll down to “GENERATE NUMBERS” ;
* Answer each of the 6 questions as follows:
  + Number of sets = “1”;
  + Number per set =” 12”;
  + Number range: From = “201” To = “500”;
  + Do you wish each number in a set to remain unique? = “Yes”
  + Do you wish to sort the numbers that are generated? = “Yes. Least to greatest”
  + How do you wish to view your random numbers? = “Place Markers Off”;
  + Click on “RANDOMIZE NOW”.

#### C. Using Excel

|  |
| --- |
| **SRS Using Excel**  Step 1: In cell A1, type  **= ROUND (RANDBETWEEN (201,500),0)**  Note: The “(201,500)” specifies that we want random numbers ranging from 201-500.The “,0” specifies that we want the random numbers rounded to zero decimal places.    Step 2: Put the cursor in the lower right corner of cell A1 (not A2) until you obtain the “**+**” sign (not the hollow plus sign),  Step 3: Press down on the left mouse button and drag the “**+**” downward as many cells as you want random numbers—in this example we want just 12. Release the left mouse button. The random numbers will appear in Column A. |

## Systematic Sampling

### Definition

|  |
| --- |
| **Systematic sample (SYS):**  A selection scheme where a **random starting point** is selected and then **every kth unit** is selected for the sample until the desired sample size is obtained as you run through the entire sampling frame. |

### How to Perform a SYS

An accurate SYS is obtained using the following steps:

|  |  |
| --- | --- |
| **Step 1**: | Determine the population size, say N; |
| **Step 2:** | Number each unit in the population from 1 to N: this is their “Pseudo ID (PSID) |
| **Step 2**: | Determine the sample size, say n; |
| **Step 4**: | Calculate the sampling interval (SI) by dividing the population size by the sample size. that is SI = N/n; |
| **Step 5**: | Select a random number (r) between 1 and (SI). The r may be a decimal number. If it is, it is a good practice to retain 2 digits after the decimal. |
| **Step 6:** | The first unit in the sample is the unit with the PSID equals to r |
| **Step 7**: | Select the other units in the sample by adding the sampling interval (SI) to r; that is r + SI, r + 2SI, …, r + (n-1)SI. |

### Illustration

|  |
| --- |
| **Systematic Sampling: An Illustration**  Let’s assume you are interested in taking a random sample of four health clinics from 20 clinics that are part of a project. The clinics are numbered from 1-20. Here are the steps to do it:  **Step 1**: Population size is N = 20;  **Step** **2**: Sample size is n = 4;  **Step 3**: Sampling Interval, SI = 20/4 = 5;  **Sep 4**: Using “Randomize Now”, we select “r = 3” as a random number between   1 and SI;  **Step 5**: Hence, the first clinic selected is “clinic 3”;  **Step 6**: Therefore, the next three clinics selected are:  Clinic 8 (8 = 3 + 5),  Clinic 13 (13 = 8 + 5),  Clinic 18 (18 = 13 + 5);  **In sum, the 4 clinics selected are: 3, 8, 13, and 18**. Voila! |
|  |

## Stratified Sampling

### Definition

|  |
| --- |
| **Stratified Random Sample (STRS):**  In an STRS, first, the sampling frame is divided into two or more sub-populations called strata. Second, in each stratum. and a random sample is drawn using either an SRS or a SYS. |

### Two Types of Stratified Random Sampling

**Proportional and Disproportional Stratified Random Sampling**

There are 2 types of STRS, namely, proportional and disproportional.

* Proportional STRS has three features:
  + It used when we want estimate at the full population level or at a national level;
  + The sample is an “exact mirror” of the population. That is, the ratio of each stratum size to the sample size is “exactly the same” as the ratio of each stratum to the population size;
  + this improves the “precision” of estimate at the population or national level.
* Disproportional STRS has three different traits:
  + It used when want to compare two or more groups or strata;
  + In the sample, all strata have the “same size”, regardless of their size in the population;
  + It provides the most statistical power and improve the are used to improve the precision of inferences to individual stratum; and comparisons among the strata.

### STRS Illustration

|  |
| --- |
| **EXAMPLE of Proportional STRS**  Zuglu land has a population of 8,000 Bete and 2,000 Guru households. An evaluation only has resources adequate for a sample of 100 households; and your priority objective is to generalize about all 10,000 households in the population.  The recommended sampling scheme is a “Proportional STRS”. The population is comprised of 80% of Bete and 20% of Guru. Therefore, a “Proportional Stratified Sample” of 100 households will have 80 Bete and 20 Guru households. |

|  |
| --- |
| **EXAMPLE of Disproportional STRS:**  Zizi village has a population of 6,000 Alfenas and 4,000 Beos households. There were allegations that the Beo minority was being excluded from the benefits of a WFP food relief project,  The primary purpose of your evaluation was to investigate the above allegations, but your budget will not allow for the total sample to exceed 100 households.  The best sampling scheme is a “Disproportional STRS” with 50 Alfenas and 50 Beos households. In this scheme, the percentage of Beos is 50% in the sample, although this ratio is only 40% in the population. This scheme will achieve two purposes, namely, it would:   * substantially increase the precision of inferences made about the Beos * improve the power any comparison between Beos and Alfenas. |

### How to Perform an STRS

The following steps describe how to select a sample using the SRTS scheme:

Step 1: Have a separate sampling frame for each stratum: the number of sampling frames must equal to the number of strata;

Step 2: Each sampling frame must be clean and update with a unique ID for each unit in the stratum;

Step 3: Determine how many units to be drawn from each stratum;

Step 4: In each stratum, draw the desired number of units determined at step 3 using either an   
SRS or a SYS;

Step 5: Repeat step 3 and step 4 for all strata.

### Practical Application of an STRS

|  |
| --- |
| **ACTIVITY**  Using the sampling frame in Appendix A, draw a disproportional stratified sample of three units from the 1991-94 stratum (which we will call the early stratum) and three units from the 1995-98 stratum (which we will call the later stratum). You may use Excel or “Randomize Now” on [www.randomizer.org](http://www.randomizer.org) . |
| **NOTES:** |
| ***SOLUTION*** |

**Note**: “Cluster Sampling” and “Multi-Stage Sampling” will be discussed in detail in the next chapter.

|  |  |
| --- | --- |
| 2 | **Multistage Sampling Essentials** What is a Multi-Stage Sampling?Definition The process of choosing a sample in **two or more successive stages** is called multi-stage sampling. Primary sampling units (PSU) are units selected at the first stage. Units selected at the second stage are called secondary sampling units (SSU) and so on. The most common multi-stage samples are two- or three-stage cluster samples. Reasons for Multi-Stage Sampling Multi-stage sampling is generally preferable under either of two conditions:   * the units to be sampled are scattered over a large geographical area, making it time-consuming and expensive to take a simple random or stratified random sample of the units; * or there is no up-to-date sampling frame for the units of analysis of interest but there is an up-to-date sampling frame for naturally occurring clusters of the units of interest—for instance, there may not be a good sampling frame of people living in rural areas but there may be a good sampling frame of the villages in which they live.   Although much of the discussion below is about samples of households, most of the procedures are applicable to samples of farms, small businesses, market places, schools, health clinics, or other units that meet one of the two above indicated conditions. Two-Stage Cluster SamplingWhat is a Cluster A cluster is a **naturally grouping of units *in a well-defined geographic* area** such as: census enumeration area (EA), village, village segment, city block, school, district hospital, health center, etc. The table below lists is some clusters that might be sampled in development evaluations. |

**Table 2.2.1: Some clusters that might be sampled**

|  |  |  |  |
| --- | --- | --- | --- |
| **Purpose** | **1st stage sampling unit:**  **Cluster** | **2nd stage sampling unit:** | **3rd stage sampling unit** |
| Condition of bridges rated at 5 tons or greater load | Province or state | Bridges rated at 5 tons or greater load | N/A |
| Condom use among Female Sex Workers | Brothel, massage parlor, bar, city block | Female Sex Workers | N/A |
| Mechanization by small manufacturing  facilities | City, town | Manufacturing facilities with less than ten employees | N/A |
| Youth Knowledge of HIV and Sexual Behaviors | School, household, locations where “street children” gather | Youth | N/A |
| Student’s aspirations for secondary education | Primary schools | Fifth grade classrooms | Students |
| Cell phone use by street vendors | Cities | Marketplaces | Vendors |

Source: Adapted from “Sampling of Populations. Methods and Applications, Third Edition”, Paul Levy and Stanley Lemeshow. John Wiley, 1999

## How to Perform a two-stage cluster sampling:

Two-stage cluster sampling is briefly described immediately below, and then subsequently will be described in more detail.

***Stage 1: Selection of clusters***

The first stage usually comprises two steps. In the first step, the clusters are selected using a Probability Proportional to Size (PPS) method. With PPS, the larger clusters have a greater probability of being selected for the sample. The PPS method is explained in 11.4.1 and 11.4.2.

In the second step, usually a listing and mapping procedure is performed, within each selected cluster, to list *all* secondary sampling units (SSUs). This is done if an accurate sampling frame is not available for each selected cluster. For household surveys, unless a census was recently completed, there is no way to know how many households there are in each selected cluster or where they are located, without performing a listing and mapping procedure.

***Stage 2: Selection of secondary sampling units (SSUs)***

In the second stage of sampling, a fixed (or proportional) number of SSUs is selected from each selected cluster, using *equal probability systematic sampling*.

**Examples of two-stage samples**

|  |
| --- |
| **EXAMPLE 3.1.3.A**  The Demographic and Health Surveys (DHS) are a component of the worldwide project *MEASURE & EVALUATION* hosted by USAID. The *MEASURE* DHS project has been implemented more than 200 surveys in about 100 countries. The reader is encouraged to refer to “*DHS Sampling Manual*” for more detailed discussions on all issues related to multi-stage household sampling. It is available at:  <http://www.measuredhs.com/pubs/pdf/AISM5/DHS_III_Sampling_Manual.pdf>  The MEASURE DHS Program has developed a convenient and practical sample selection procedure for households—*a two-stage cluster sampling* procedure. In the first stage of sampling, clusters are sampled rather than individual households. In most cases, a cluster is a census *Enumeration Area* (EA). The average EA size for a population census is usually between 100 and 300 households, which are adequate for a *sample take* of 20 to 30 households per EA. |

Clusters are usually stratified before the first stage of sampling, by region, urban/rural, or other characteristics that are thought to be importantly related to the variables of interest. Then a stratified sample of EAs is selected from each stratum with *probability proportional to size* (PPS).

In each selected EA, a listing and mapping procedure is performed to list *all* dwellings/households. This procedure is important for correcting errors existing in sampling frames that are several years old. The list and map become the sampling frame for the second stage of sampling.

In the second stage of sampling, a fixed (or varying) number of households is selected by *equal probability systematic sampling* from each of the EAs selected in the first stage. This is the same sampling procedure described much earlier (and is sometimes referred to as “fixed interval sampling with a random start.”

|  |
| --- |
| **EXAMPLE 3.1.3.B:**  The Canadian Labor Force Survey is an example of a two-stage cluster sampling. The whole country is divided into 1,100 strata. The first stage is the selection, in each stratum, of a sample of EAs using PPS. Then, the EAs are mapped and all the dwellings in them are listed. At the second stage, a systematic sample of dwellings is selected from each list. All persons within a selected dwelling are interviewed. |

## Advantages of Two-Stages Cluster Samples

The advantages of this two-stage cluster sampling procedure are:

1. Two stages sampling usually offers considerable efficiency with only moderate increases in sampling errors, but additional stages of selection often increase the sampling error considerably.
2. The use of residential households as the second-stage sampling unit guarantees the best coverage of the target population;
3. A household listing procedure, after the selection of the first stage and before the actual data collection, provides an up-to-date sampling frame for household selection in the second stage
4. It guarantees a representative sample of the target population when there is a list of all clusters to be sampled but not a list of all target individuals within each cluster.

It should be noted that two-stage cluster sampling is generally not advisable if there are only a small number of clusters in the population or only a small number of households within many of the clusters. Under these conditions, the estimated sampling error is likely to be high and thus the estimates are not likely to be precise

# First-Stage Probability Proportional to Size Sampling

## What is PPS

PPS sampling is a probability sampling scheme where the chance of a cluster being selected depends upon its size. That is, larger clusters have a higher likelihood of being selected into the sample. This is particularly important when the **measure of size** (mos) is thought to be **correlated** with the main indicator of interest. For instance, people in large villages may be more likely to visit a health clinic within their village than people from small villages who may have to travel to a larger village to reach a clinic.

The characteristic used as the mos depends on the main indicator of interest and the information available. The mos for census enumeration areas would usually be either the number of houses in the EA (for a survey of household head) or the total number of youths within the EA (for a survey of sexual behavior among youth).

## How to Perform a PPS sampling**:**

To implement PPS sampling, follow these seven steps:

|  |  |
| --- | --- |
| **Step 1**: | Prepare a list of clusters with corresponding mos for each. |
| **Step 2**: | Starting at the top of the list, calculate the **cumulative** mos and enter these figures in a column next to the measure of size for each unit. |
| **Step 3**: | Determine the range corresponding to each unit. The lower limit of the range is the previous row’s upper limit of the range plus one. The upper limit is the – cumulative mos. Record the range in the column to the right of mos. |
| **Step 4**: | Calculate the sampling interval (SI) by dividing the **total** cumulative mos,(Call this M) by the number of units to be selected (a, say), that is SI = M/a. |
| **Step 5**: | Select a random number (r) between 1 and (SI). The r may be a decimal number. If it is, it is a good practice to retain 2 digits after the decimal. |
| **Step 6**: | Compute subsequent random numbers are obtained by adding the sampling interval (SI) to r; that is r + SI, r + 2SI, …, r + (a-1)SI. |
| **Step 7**: | Select each cluster whose range contains the number obtained in step 6. |

Note: In selecting clusters, the SI decimal points MUST be retained and the following rule is applied. When the decimal part of the sample selection number is less than 5, the lower numbered cluster is chosen, and when the decimal part of the sample selection number is 5 or greater, the higher numbered cluster is chosen. L

n.

## Example of PPS Sampling

|  |
| --- |
| **EXAMPLE of a PPS Sampling**  The government of Zulu Land is ready to launch a vast program of economic empowerment and wealth creation called “One cow per family” program. The program consisted of giving each household in a village one cow and $12,000 in cash. The idea is that household who benefited from this program will “be better off or will have a better welfare” than the one who did not. As a pilot study coordinator, you were asked to select 4 villages in a province that has a total of 12 villages using the PPS method. This is how to proceed. |

**Table 4.3.3. Selection of four Villages (clusters) using PPS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster Num** | **mos:**  **# HH listed** | **Cumulative mos** | **Range** | **Sample Selection #** | **Cluster Selected** |
| 1 | 163 | 163 | 1 - 163 | 22 | \* |
| 2 | 250 | 413 | 164 - 413 |  |  |
| 3 | 110 | 523 | 414 - 523 |  |  |
| 4 | 210 | 733 | 524 - 733 | 638 | \* |
| 5 | 207 | 940 | 734 - 940 |  |  |
| 6 | 160 | 1100 | 941 - 1100 |  |  |
| 7 | 165 | 1265 | 1101 - 1265 | 1254 | \* |
| 8 | 180 | 1445 | 1266 - 1445 |  |  |
| 9 | 140 | 1585 | 1446 - 1585 |  |  |
| 10 | 309 | 1894 | 1586 - 1894 | 1870 | \* |
| 11 | 245 | 2139 | 1895 - 2139 |  |  |
| 12 | 325 | 2464 | 2140 - 2464 |  |  |
| Number of clusters to be selected = 4 ; SI = 2464 / 4 = 616; r = 22 | | | | | |
| \* = Selected clusters | | |  | |  |

## Practice of PPS

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ACTIVITY:**  Use PPS to select a sample of two villages from the following six villages, each with the indicated number of households. Assume that your initial random number is 53.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Village ID | Number of Households  in Each Village | Cumulative mos | Range | Sample Selection Number | | 1 | 50 |  |  |  | | 2 | 30 |  |  |  | | 3 | 10 |  |  |  | | 4 | 20 |  |  |  | | 5 | 40 |  |  |  | | 6 | 30 |  |  |  | |
| **NOTES:** |
| ***SOLUTION***:  The cumulative mos is 180, and we want a sample of 2 villages. Therefore, the SI=90 (180/2). We assumed the random number was **53**, which is within the cumulative mos of the village with ID number 2, and thus it is selected. Then **53+90=143,** which is within the cumulative mos of the village with ID number 5, and thus it is selected.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Village ID** | **Number of Households** | **Cumulative mos** | **Range** | | **Sample Selection Number** | | 1 | 50 | 50 | 1 | 50 |  | | 2 | 30 | 80 | 51 | 80 | **53\*** | | 3 | 10 | 90 | 81 | 90 |  | | 4 | 20 | 110 | 91 | 110 |  | | 5 | 40 | 150 | 111 | 150 | **143\*** | | 6 | 30 | 180 | 151 | 180 |  | | **Total** | 180 |  | | | | | **sample size** | 2 |  | | | | | **SI** | 90 |  | | | | | **\*** | village selected |  | | | | |

**Excel Template for First Stage PPS selection:** The MEASURE DHS staff have developed an Excel template that can greatly facilitate the drawing of large PPS samples, including those for which there is prior stratification of the clusters. It is “Program 1-First Stage PPS Sampling” and is being provided with the permission of Dr. Ruilen Ren, who created it.

# Second Stage Systematic Sampling

The second stage comprises two main activities, namely, mapping and listing, as well as the actual sampling procedure. Let’s take a look at each of these activities

## Mapping and Listing of Second Stage Units

In some cases, a good second-stage sampling frame might already exist, as will often be the case for fifth-grade classrooms in selected primary schools or government-built bridges designed for 5 ton or greater loads in selected states.

In other cases, there will not be an accurate sampling frame, as is often the case for household surveys, especially if they are done several years after the last census. In these cases, a complete listing and mapping of dwellings/households in each selected cluster should be compiled.

In each selected cluster (i.e. village, town, or EA), the mapping exercise consists of drawing a location map of the cluster as well as a sketch map of all the households in that cluster. In small villages, it can be done manually as shown in Appendix J. In large villages and urban areas, it is usually done with a help of a cartographer or GIS staff.

Google maps, aerial photography, and commercial satellite mapping services might be considered for the mapping process, depending on their availability, when they were produced, and the cost. An example of mapped area is given on the next page.

During the listing process, the data collectors go around and number every mapped household in each selected cluster. The actual number of households in each selected cluster is then recorded. This provides updated mos of each selected cluster. These mos will be used in computing the probability of selection of each household as well as its weight. The listing operation also usually involves recording a brief description of every household together with the name of the head of the household.

For instance, in the previous “One Cow per family” example (example 3.1.4), during the listing operation, we notice that the current number of households and each of the four selected clusters (villages) was somewhat different from initially estimated. It was actually: Village 1: 151; Village 4: 250; Village 7: 110; and Village 10: 322. Those updated mos will be used in the following examples.

The listing operation represents an appreciable field cost, but often there is no reliable method by which it can be avoided. The listing operation represents one of the most important bias prevention procedures in the survey because it allows second stage sampling of households by the project’s sampling experts prior to sending interviewers into the field and it allows easy field work quality control.

With a complete household listing and the households randomly pre-selected, the data collection procedure is completely replicable by the fieldwork supervisor; therefore, it does not leave any space for “deliberate errors” by the interviewer.

We need to list ALL of the households found in the cluster for two purposes: To prevent biased household selection (if only 50 or the 300 households are listed, there is no way to assure the 50 are representative of the 300), and to allow proper calculation of the sampling weights.

|  |
| --- |
| **ACTIVITY**  Our sampling team reaches the 10th town randomly selected in our first stage of cluster sampling. The town mayor has offered to help with our household survey by doing the following:  a) He will draw us a map of all the occupied dwellings in the town,  b) He will identify the ones that don’t like talking with strangers and will probably refuse to be interviewed, and  c) He will identify the households that do not provide their children with adequate nutrition, the subject of our survey.  How should we make use of his help? |
| **NOTES:** |
| ***SOLUTION***  Use your pre-planned procedures for mapping and listing households! His map may not be accurate, even if he knows the village perfectly, because he might have a personal stake in biasing the results. Good random sampling does NOT avoid households that are expected to refuse to participate, but instead it carefully plans inducements and follow-up efforts to try to complete interviews with every sampled unit. His identification of households providing children with inadequate nutrition might be inaccurate. On the other hand, you might want to check your independently created map against his to make sure you staff have not inadvertently overlooked some occupied households. You might also want to assign your most persuasive interviewer to the households most likely to refuse the interview. In addition, as a validity check on your interview data, you might want to check his list of households not providing adequate nutrition against your results, but if there are disparities, it is not clear whether his information or your data are in error. |

## Second-Stage Systematic Sampling

The second stage of sampling involves sampling of units from the already selected clusters. These are called secondary sampling units (SSU). A SYS scheme is generally used at this stage.

### Advantage of the Second Stage SYS Scheme

There are three advantages of using equal probability systematic sampling at this stage rather than simple random sampling:

1) Since the SSU listing procedure follows a pattern, systematic sampling assures that the selected households come from throughout the cluster, rather than being concentrated in one or two areas, which could happen by chance if a simple random sample is used;

2) If necessary or desirable, the actual sampling can be done easily in the field with a minimum of training of the field staff;

3) It allows for easy verification of fieldwork, walking a specified pattern and checking the selected SSUs.

It is possible to draw a second stage sample from each cluster selected in the first stage of sampling by using the procedures explained in Section 6.16, but that would be time consuming if you had had to do it for more than a few clusters. There are tools for doing this much more quickly.

2.4.2.2. Use of an Excel template to Perform a Second-Stage SYS

The MEASURE DHS program has developed an Excel template that can be used for second-stage equal probability systematic sampling of units from any type of cluster, although the template columns are labeled for the sampling of households. It also allows drawing a third-stage sample at the same time. This is Program 2 on this workshop CD, and is being provided by Dr. Ruilen Ren, who created it. Instructions for its use are provided immediately below and are illustrated. The user has to input the first three columns of data.” The template generates the rest of the data.

Let us illustrate this with our “One Cow per family” from example using a simplified representation of the Excel template shown in Figure

**Column 1**: Cluster Number: Input the cluster ID numbers. In our “One Cow per family” from example the four first-stage sampled clusters have ID numbers 1, 4, 7, and 10, and so we enter that information in the first column.

**Column 2**: Number HHs: Input the number of households obtained during the **listing and mapping** exercise. This is the mos. As mentioned in the fifth paragraph of section, the mapping and listing operation found the actual number of households to be 151, 250, 110, and 322, and thus we enter than data in the second column.

**Column 3**: Number Selected: Input the number of households (or other units) that you want sampled from each cluster in this second stage of sampling. The number can be the same number for all clusters, or it can be proportional to the mos (HHS Listed) of each cluster. In this example, it is 10 for every cluster.

After inputting all the required information in column 1 to column 3 in an empty Program 2 Excel template, click on any of the empty cells and all should fill with data.

* The Selection Interval (sampling interval) is computed by Excel (dividing Column 2 data by Column 3 data).
* The Random Number is generated by Excel.
* The 10 numbers in the right half of the spreadsheet are the cases numbers randomly selected by Excel from each cluster. Thus, for cluster # 1, the following ten households were selected in this order: Households # **1, 16, 31, 46, 61, 76, 91, 107, 122, and 137**.

**Figure 3.1.6: Example of household selection at a second stage using Systematic Sampling.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fill the first three cells with the required information | | | | | | | | | | | | | | |
|  | | | | Household Selected | | | | | | | | | |  |
| **Cluster Num** | **Num.**  **HHs** | **Num**  **Selected** | **Select Interval** | **Random**  **(0-1)** | **1st** | **2nd** | **3rd** | **4th** | **5th** | **6th** | **7th** | **8th** | **9th** | **10th** |
| 1 | 151 | 10 | 15.10 | 0.02422 | 1 | 16 | 31 | 46 | 61 | 76 | 91 | 107 | 122 | 137 |
| 4 | 250 | 10 | 25.00 | 0.89743 | 23 | 48 | 73 | 98 | 123 | 148 | 173 | 198 | 223 | 248 |
| 7 | 110 | 10 | 11.00 | 0.59064 | 7 | 18 | 29 | 40 | 51 | 62 | 73 | 84 | 95 | 106 |
| 10 | 322 | 10 | 32.20 | 0.72572 | 24 | 56 | 88 | 120 | 153 | 185 | 217 | 249 | 281 | 314 |

\*Note: The values shown for “Num HHs” differ from the values shown for those clusters in Table 11.4.3 because a mapping and listing operation was assumed to have been conducted in those clusters after they were selected for the sample, as is reported in the fifth paragraph of Section 11.5, and it found the numbers of households shown in this table.

It should be understood that, each time this Excel template is run, it is likely to select different households from each cluster, because it is selecting them randomly. So, if you try to replicate this example, please, enter the same data shown in the first three columns, and in column labelled “Random (0-1)”.

# Possible Third-Stage Sampling

There are two forms of third-stage sampling. One is sampling a subset of households from those already sampled in the second stage, and the other is sampling among eligible respondents within each household already been sampled in the second stage. We will illustrate both.

## Third-Stage Sampling a Subset of Already Sampled Households

Let us return to the “One cow per family” program of example 2.3.3. Recall that in section 3.1.6 you sampled 10 households in each of the selected villages (see fig. Now, you were asked by the government to select 5 households which will receive the program (beneficiary) and 5 of which will not (control). This way, any improvement in the welfare in the lives of the beneficiary households could rightly be attributed to the program. As a sampling specialist, how are you going to choose the 5 beneficiaries?

Answer: This calls for a three-stage cluster sampling. We could randomly select five of the ten cases selected from each cluster in Figure 3.1.6, but that would be time consuming if the number of clusters is large. Alternatively, we can have Program 2 on the IPDET Sampling Workshop CD, do this automatically. We will illustrate this in Figure 3.1.7.1, which now shows the full Excel spreadsheet, rather than the simplified version shown in Figure 3.1.7.1.

At the very top of the spreadsheet, specify a “**Run size**” of **1**. A run size of 1 means there is one “simple selection”. In household surveys, for convenience and efficiency, sometimes “runs” of five (or even more) houses will be sampled. This means that rather than sampling each house independently, each time a house is sampled, the next 4 (or more) houses will also be automatically selected.

Just below the very top, specify a “**Sub-sample take per cluster**” of **5**. This means that a third-stage sample of 5 households among the 10 selected at the second-stage is needed. A value of 0 means no third-stage is needed. The following Figure 3.1.7.1 displays the result.

**Figure 2.5.1: Combining Second and Third Stage Sampling of Households**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Fill the green cells with the required information** | | | | | | | | | | | | | | | | |
|  |  | |  | |  |  | **Household Selected** | | | | | | | | | |
|  |  | |  | | Run size | 1 |  |  |  |  |  |  |  |  |  |  |
|  | **Sub-sample take per cluster** | | | | | 5 | 1st |  | 2nd |  | 3rd | 4th |  |  | 5th |  |
| **Cluster Num** | **Num HHs** | **Num Selected** | | **Select Interval** | | **Random**  **(0-1)** | **1st** | **2nd** | **3rd** | **4th** | **5th** | **6th** | **7th** | **8th** | **9th** | **10th** |
| 1 | 151 | 10 | | 15.10 | | 0.02422 | 1 | 16 | 31 | 46 | 61 | 76 | 91 | 107 | 122 | 137 |
| 4 | 250 | 10 | | 25.00 | | 0.89743 | 23 | 48 | 73 | 98 | 123 | 148 | 173 | 198 | 223 | 248 |
| 7 | 110 | 10 | | 11.00 | | 0.59064 | 7 | 18 | 29 | 40 | 51 | 62 | 73 | 84 | 95 | 106 |
| 10 | 322 | 10 | | 32.20 | | 0.72572 | 24 | 56 | 88 | 120 | 153 | 185 | 217 | 249 | 281 | 314 |

These results are interpreted as follows: The five households that are to receive the “One cow per family” program intervention are in the columns where the top right-hand row shows 1st, 2nd, 3rd, 4th, and 5th. Thus, for Cluster 1, households 1, 31, 61 76, and 122 are to receive the program intervention and the other five are to serve as controls. For Cluster 4, households 23, 73, 123, 148, and 223 are to receive the program intervention and the other five are to serve as controls.

You might worry that this third-stage random sampling is not independent between clusters, because it selects the first, third, fifth, sixth, and ninth case from each of the four clusters ten selected households. But since a different random number selected the household from each cluster (see the fifth column of data in the table), the selection is independent across the clusters.

## Third-Stage Sampling of One Respondent from Several Eligible

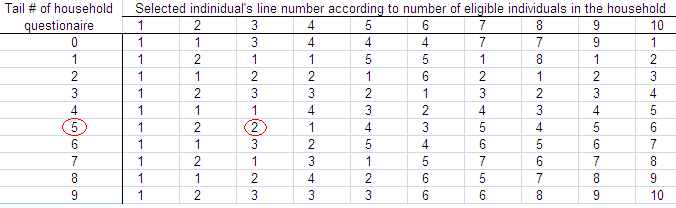
Sometimes, a survey will interview all eligible individuals found in a sampled household because it is easy to manage the fieldwork and is cost-efficient. In some circumstances, the survey is designed to interview **only one of two or more eligible individuals** per sampled household. In this case, it is better to select the individual to be interviewed through a random selection procedure. Otherwise, serious bias may be introduced, especially when the average number of eligible individuals per household is substantially larger than 1.

For instance, the Youth Behavioural Surveillance Survey (BSS) is designed to collect information on knowledge and sexual behaviour of youth aged 15 to 24. This survey is intended to interview only one eligible youth per household. However, in Zulu Land, the recent census finds an average of 3 eligible youths per household. In this instance, a third stage selection is needed using a random procedure.

A random selection of a single individual per household is easy to perform using a “Kish Grid” which is shown below in Figure 2.5.2. First, assign each eligible person in the household a unique consecutive number starting with 1. The questionnaire “tail numbers” listed in the left column of a Kish Grid are the last digit of the questionnaire ID number. For a household with questionnaire #135, the tail number is 5. If the household has three eligible individuals (whom have been assigned numbers 1, 2 and 3), read down the column labelled “3” until it intersects row 5. That cell has a “2” and thus the interviewer would select the individual assigned the number 2 for the interview.

Note: the Kish Grid only works when there are not more than 10 people within the second-stage sampled units that would be eligible for the third-stage sampling.

**Figure 2.5.2: Kish Grid for selecting one respondent from several eligible within a household**



With the selection of one individual per household, it is necessary for a statistician to calculate an individual sampling weight by taking into account the number of eligible individuals in the household for a specific interviewed individual. Otherwise, another source of bias might be introduced. It may worth to point out that each stage of sample selection is likely to bring additional sampling errors. Therefore, it is better to limit the sampling to two stages unless there is a compelling reason for a third stage

# The Importance of Sampling Weight

## What is a Sampling Weight?

In a simple term, a “Sampling weight” or “weight”, is the i**nverse of the probability of being in the sample**. This “multiplier” tell us “**how many people** with same characteristics in the population**” a given respondent in the sample represent**.

## Why are weights needed?

Weights are used to:

* ensure the sample **accurately represent** the actual target population
* **take into account the sampling design**, especially, if the sample design is  
   either a stratified sampling design, or a cluster sampling design, or a two-stage cluster sampling design, or any other design except simple random sampling;
* **take into account unexpected effects** such as non-response or  
   non-sampling error;
* **adjust** each respondent’s contribution to the overall results;
* produce **accurate results** and **make sound inferences**;
* accurately generalize our results to the target.

## Illustrative Example

Amakaru village in Central Africa, has a population made of two tribes, namely 8,000 Dene and 2,000 Dakota. For the past two years, there have been some serious allegations that a vast majority of the Dene people was being excluded from the benefits of a food relief project co-funded by FAO and WFP.

You were called in to conduct an evaluation of this project with the primary purpose of investigating the above allegation. However, your budget will not allow for the total sample to exceed 100 participants.

For this purpose, you decided to use a “**disproportional stratified random sample**” of 50 Dene and 50 Dakota. This would substantially increase the precision of inferences and improve the power of making comparisons between the Denes and the Dakotas. The weight for each stratum is as follows:

**Population = 10,000** **Sample = 100**

**Dene Dakota**

**8,000 people 2,000 people**

**Sample: 50 sample: 50**

**weight = 8,000/50 = 160. weight = 2,000/50 = 40.**

**Each Dene sampled Each Dakota sampled**

**represents 160 Denes in represents 40 Dakotas in**

**the population the population**

## Note and Caution

It is important to bear the following words of caution in mind:

* + - * the computation of weights can vary from **simple to very complex**. It is a good practice to seek the advice of an experienced survey expert;
      * The “default” analysis procedure in almost all software packages is the “unweighted” analysis. However, most packages provide a special procedure for “weighted” analysis. It is the responsibility of the researcher to provide the weight variable and to choose the appropriate procedure;
      * In case where all the participants in a sample has “**equal weight**”, the sample is said to be “**self-weighted or equal weighted**”. In this instance, the unweighted and weighted analysis are **equivalent**;

**APPENDIX A**

**Sampling Matrix**

**World Bank Funded Operations** (Simulated Data)

**Sustain-**

**Year**  **Cost: US$ Outcome ability**

**ID #** **Country**  **Region Approved Sector Thousands Rating Rating**

01. Algeria MENA 98 d 550 S M

02. Brazil A LAC 91 b 380 S H

03. Brazil B LAC 96 d 180 U L

04. Bulgaria ECA 97 b 120 S L

05. Burundi AFR 98 a 280 U L

06. China EAP 94 c 900 S L

07. Colombia A LAC 95 b 230 S M

08. Colombia BLAC 96 c 170 S L

09. Egypt A MENA 97 a 370 S M

10. Ethiopia AFR 96 c 300 S L

11. Ghana AFR 93 b 150 S M

12. Guatemala LAC 92 a 250 S L

13. Hungary ECA 97 a 220 S L

14. India A SAS 93 c 360 S L

15. India B SAS 96 d 430 U H

**APPENDIX B**

**Random Number Table**



Source: OFPP Pamphlet No. 4, Supplement No. 2 to OMB Circular No. A-76, October 1980