حكومة دبي
GOVERNMENT OF DUBAI

مركز دبي للإحصاء
Dubai Statistics Center

Sampling Techniques Manual

رؤيتنا | إحصاء داعم لصناعة السعادة والتنمية المستدامة.
Our Vision | Statistics that support creating happiness and sustainable development.

dsc.gov.ae

Contents

Introduction

Statistical Sampling is a statistical data collection method, which is limited to a part of the target population, which makes it different from the overall surveying method of all segments of the target population, which is known as "census." The sampling methods have developed until they reached the current status, and the development was associated with the Probability Theory. The evolution of Law of Large Numbers, Central Limit Theorem (CLT), Law of Normal Distribution, Student's t-distribution of 1908 for small samples, and the practical applications accompanied that, especially in the second half of the twentieth century, is of great importance in the development of statistical methods and sampling methods, particularly regards the methods of calculation of sampling errors and dissemination of the results of the sample to the population. Given the close interrelationship between the Sampling Theory and the Probability Theory, there are two different types of samples. **The first type "Probability Samples":** These are the randomly selected samples, and each unit has a probability other than zero. Focus is on this type in practical applications, since the results of these samples can be disseminated to the populations from which they were selected, as well as in determining the degree of confidence in the results and calculation of errors. **The second type "nonprobability samples"**: These are samples that don't use the principles of probability, since the units of the sample are selected according to purposive or personal method. Unlike the Probability Sampling, the accuracy of results and representation of the population cannot be defended. For this reason, and for the rare use of nonprobability samples, the following review will only be limited to the Probability Sampling methods. Prior to reviewing the special techniques of this method, a summary of some of the concepts for sampling is provided.

- **Target Population:** This means the units that should be covered by the survey. In the survey of a specific country or region, the target population is all the inhabitants of this region or country, and the target population in an economic survey is all economic and social facilities.

- **Survey Population:** In practice, there may be some limitations to cover certain target population units, because of the considerable cost for covering them or because of difficulties of accessing these units, such as population living abroad, or units that are located in disaster areas, and therefore this type of units should be deleted from the survey target population, and the survey units remaining after this procedure are then called “Survey Population”.
- **Sample frame:** This refers to the list of sampling units in each stage. It is also the cornerstone of the sample design and carrying out of all its subsequent stages. Furthermore, success of the sampling program depends heavily on the availability of modern and appropriate frame for the sample design. The sampling model frame is the latest frame encompassing all units. Examples of the most commonly-used frames for practical applications include the households' frame and the facilities frame. One of the most important characteristics that should be available in the frame is that it should be complete, i.e. including all the units in question. Each desired survey unit must be listed for once in the frame or the list that include these units, with no extraneous units. In this context, provision of the maps accompanying these frames is of great significance in locating the sample units, facilitating fieldwork, and cost-reduction, thus raising the efficiency of the survey.

The topics for this publication have been presented based on several sampling techniques, as in the works of many professionals, such as (Kish), (Cochran), (Nyman), (Hansen), and others. As well as the publications of many international institutions, especially the recent United Nations publications by the statistics department, such as (Household sample surveys in Developing and transition Countries 2005), Designing Household Survey Samples: (Practical Guidelines 2010 and Handbook of Household Surveys 1987). In addition to the most important global practices that have been applied in many countries around the world since 1980th, most importantly: the World Fertility Survey, conducted in many world countries since 1974 and Demographic and Health Survey (DHS), which has been applied in many world countries since 1984.

1. Types of Probability Samples

1.1. Simple Random Sampling

Simple random sample is defined as: a number of units randomly drawn from the sample frame, covering all units (individuals, households, establishments, land plots, etc.). Through the information collected on the sample units, all corresponding estimates for the survey population are made, in addition to the estimates of errors committed therein and the levels and areas of confidence in these estimates. The main factor in the probability sampling is the random sampling of the sample units. The basic condition for achieving randomization is that each unit of the sample frame should have an opportunity to be draws in the sample; which means that non-existence of any constraints for the absence of any unit in the sample. Such principle is expressed in the probability theory by saying that each unit has a different probability than zero for being drawn in the sample.The Simple Random Sampling is the basis for other statistical sampling methods, therefore a summary is provided for the theoretical basis for it, namely: Law of Large Numbers (LLN) and Central Limit Theorem (CLT).

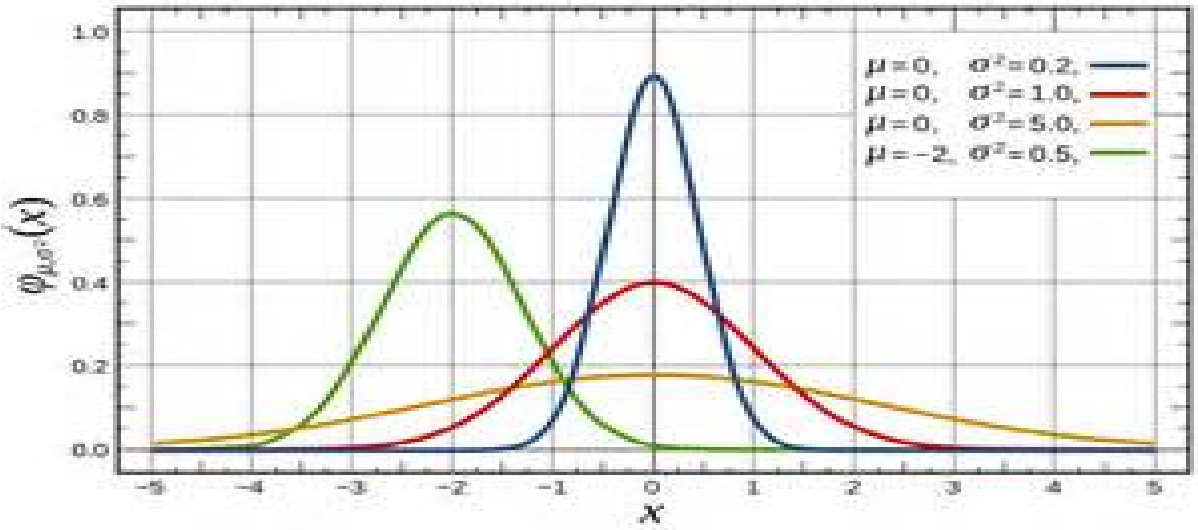
It has always been noted that through a limited number of observations of a certain phenomenon, some characteristics of this phenomenon can be identified, such as the mean or proportion. When this process is repeated on a limited number of other observations (greater than or equal to 30), these characteristics will differ slightly from those calculated from previous observation, and there will be minor differences from the real characteristics of population. Bernoulli, was the first to note that by sampling a limited number of balls from a box containing two types of balls, in the same circumstances and without any deliberate intervention (which was later called Random Selection) and he found that the proportion of a specific type of balls in the sample is close to the proportion of that particular type of balls in the box, this result was achieved when the units were selected with replacement (where the selected unit is returned to the population before the next selection), or without replacement (where the selected unit is not returned to the population). This was the first introduction to the mathematical formulation of the Law of Large Numbers. Similarly, the proportion of unemployed population in a certain area can be identified through sampling a large number of

the population, the household average income in a specific city can be identified through sampling a large number of households. The common factor in the previous examples, and other numerous examples of natural or social phenomena, is the large number of observations. This principle is called the "Law of Large Numbers" and it was the introduction to the CLT.

One of the key theories in statistical sampling is the widely-used CLT. Laplace and De Moivre were first to formulate this theory in the first half of the 18th century. The final version of the CLT was put forward in the end of the 19th century and beginning of the 20th century in the works of Aleksandr Lyapunov and Chebyshev. In the light of the results of practical experiments of the Law of Large Numbers, it became clear that if all random samples were formed, size of each is n of a population of the size N, while the sample size was greater than or equal to 30 (large samples), the means of these samples will be distributed according to the well-known Law of Normal Distribution, regardless of the law of distribution of population. As is known, the function of its distribution is:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

All these distributions share similar curves but differ in terms of mean $\overline{X}$ (μ) and variance σ^2, as in the next shape.



Source: http://www.bing.com/

Later in 1908, Student verified that principle when the sample size is smaller than 30 (small samples) and the population is distributed normally , the means of all samples will be distributed according to the law for small samples, known as (t-Student), which resembles the normal distribution, and will be matching when the sample size is greater than (30).

In order to facilitate their use, a distribution, called Standard Normal Distribution, was created based on the mean and variance of Normal Distribution, by replacing the random variable x in Normal Distribution by the random variable Z which is calculated by the following relation:

$$z = \frac{x - \overline{X}}{\sigma}$$

The new distribution function becomes $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

The resulting distribution is sometimes called Z Distribution. One of the key characteristics of the Standard Normal Distribution is that the mean of this distribution is equal to (0), and its variance is equal to (1), regardless of the mean and variance of the Normal Distribution which were varied, and for this reason the new distribution is called “Standard” because it will be the same regardless of different normal distributions, as the red curve in previous shape.

A special table was prepared for the values of that distribution which are matching the Z values, which can be found in many statistical references and software such as (Excel).

This result is of great significance in the theory of samples where the true values in population are estimated (such as mean or proportion) by the estimates corresponding to them in the sample.

If all the random samples were formed, the size of each is n of a population of the size N and the following samples means:

$$\overline{x}_1, \overline{x}_2, \ldots\ldots\ldots\ldots\ldots, \overline{x}_L$$

Where: (L) is the number of random samples that can be formed, and based on the CLT, the means of the random samples are distributed according to Normal Distribution. This distribution is determined by the mean equal to the mean of population $\overline{X}$, and its variance $\sigma^2_{(\overline{x})}$, and this result applies to all types of populations. The following, then, can be proved:

The mean of this distribution is equal to the mean of the population $\frac{\sum_{i=1}^{i=L} \overline{x}_i}{G} = \overline{X}$

Variance of this distribution, with replacement, equals $\sigma^2_{(\overline{x})} = \frac{\sigma^2}{n}$

Variance of the mean of this distribution without replacement equals

$$\sigma^2_{(\overline{x})} = \frac{\sigma^2}{n} * \frac{N-n}{N-1}$$

It is noted that the variance of the mean in the case of without replacement differs in the case of with replacement by the coefficient $\frac{N-n}{N-1}$, and when the size of the population is large, the difference between the value $N-1$ and the value N is too small, and by replacing it in the case of without replacement , it becomes:

$$\sigma^2_{(\overline{x})} = \frac{\sigma^2}{n} * \frac{N-n}{N} = \frac{\sigma^2}{n} * \left(1 - \frac{n}{N}\right) = \frac{\sigma^2}{n} * (1-f)$$

Where $f = \frac{n}{N}$ which is called Sampling Fraction. While $1-f$ is called Population Correction Factor (PCF).

Accordingly, the Standard Normal Distribution can be found for this population where $Z = \frac{\overline{x} - \overline{X}}{\sigma_{(\overline{x})}}$ while $\overline{X}$ represents the mean of any sample which was formed from this

population and $\sigma_{(\bar{x})}$ represents the standard deviation of the variance of this mean. It could be proved that the mean of this new variant is equal to zero, and its variance is equal to (1) and that it is distributed in accordance with the Standard Normal Distribution, and based on the properties of this distribution, it can be calculated that the probability that Z takes a value between any two corresponding values, if $f(Z)$ is the Standard Normal Distribution, this probability is calculated according to the following relationship:

$$\Pr\left(-Z \le \frac{\bar{x}-\bar{X}}{\sigma_{(\bar{x})}} \le +Z\right) = \int_{-Z}^{+Z} f(Z)^* dZ = 2\int_{0}^{+Z} f(Z)^* dZ$$

The latter relationship connects the mean of population and the mean of one sample of the size n. In practice, one sample formed from the population is used from which it could be calculated the difference between the sample mean and population mean. This difference may be positive or negative, and the absolute value of the difference represents the error in estimating the population mean, and is called: Margin of Error, referred to as e

$$e = \left|\bar{X} - \bar{x}\right|$$

Accordingly, the probability of the following relation can be calculated:

$$-Z \le \frac{\bar{X}-\bar{x}}{\sigma_{(\bar{x})}} \le +Z \Leftrightarrow -Z^*\sigma_{(\bar{x})} \le e \le +Z^*\sigma_{(\bar{x})}$$

If $Z = 1$ then the Margin of Error is between two corresponding values, as in the following relation:

$$-\sigma_{(\bar{x})} \le e \le +\sigma_{(\bar{x})}$$

In this case, it is called the Standard Error (SE), which is equal to one standard deviation of the sample mean. Depending on the Standard Normal Distribution tables, the probabilities corresponding to different Z values can be calculated. This probability is called the

Confidence Level, sometimes is called Confidence Interval. In most studies and statistical surveys a confidence level greater than 90%, and mostly 95%, is used which is corresponded by the value $Z = 1.96$ in the Standard Normal Distribution table. Generally, the following relation that links between the population mean, the mean of one of the samples, and the margin of error:

$$-Z * \sigma_{(\bar{x})} \leq \overline{X} - \bar{x} \leq +Z * \sigma_{(\bar{x})}$$

Accordingly, the population mean could be identified and confidence in this result, which is called the probability or confidence level. This can be expressed by the relation:

$$\bar{x} - e \leq \overline{X} \leq \bar{x} + e$$

The right side is called the maximum level of the population mean, and the left is the minimum, while the estimation of the population mean, accordingly, is "**estimtion by interval**", to differentiate it from the estimation that only depends on the sample mean without regard to the margin of error, which is called "**estimation by point**".

For example, a sample size of 100 units of a population of the size of 1000 was collected to estimate the mean of a given phenomenon, if the variance of this phenomenon in population is 2000, and the mean in the sample is equal to 80, what can be said about the margin of error in the case of with replacement and case of without replacement? What is the maximum and minimum population mean at 95% confidence level?

The margin of error with replacement

$$e_0{}^2 = Z^2 * \frac{\sigma^2}{n_0} = (1.96)^2 * \frac{2000}{100} = 76.83 \Rightarrow e_0 = 8.77$$

Estimation of the population mean:

$$\bar{x} - e \leq \bar{X} \leq \bar{x} + e$$

$$80 - 8.77 \leq \bar{X} \leq 80 + 8.77$$

$$71.23 \leq \bar{X} \leq 88.77$$

This result means that the population level will fall into this interval at 95% confidence level. The population mean can be outside this interval at a 5% confidence and in this case the population mean is greater than 88.77 or smaller than 71.23.This sample was selected from among all the samples that can be formed from the population. This important result should be taken into account when designing the sample so that it is a sample representative of the population, through testing the matching of a number of indicators of the population (average household size, age, sex, and others) with those in the sample sellected, where the sample is changed if its general characteristics are not matching the properties of the population.

The margin of error without replacement

$$e^2 = Z^2 * \frac{\sigma^2}{n} * \frac{N-n}{N-1} = 69.2 \Rightarrow e_0 = 8.32$$

which is smaller than in the case with replacement, noting that the margin of error in the case of replacement is not affected by the size of the population. In the case of large populations,

the margin of error is approximately the same in case of with replacement and without replacement. If the size of population is 1000000, the margin of error without replacement is equal to 8.77 too, which is the margin of error with replacement.

In a similar way to estimate the population mean, this can be generalized to the proportion of a phenomenon in the population, since the proportion in the population can be considered average, by assigning the value (1) for each value that match the phenomenon in the population, and value 0 for values that do not match. And then, the mean values that match the phenomenon, as is known, is calculated according to the following relation:

$$\overline{X} = \frac{X_1 + X_2 + \ldots\ldots\ldots\ldots + X_N}{N}$$

which is called the proportion of the phenomenon in the population and is referred to as P to distinguish it from the mean, and it is also the proportion that does not match the phenomenon (opposite), which is referred to as $Q = 1 - p$ and given the fact that any value in the population takes the value of (0) or (1), its variance according to the Bernoulli Law is equal to $\sigma^2 = P * Q$

And in a similar way to that of the mean, the square error of the proportion of the phenomenon in the population can be written as follows:

with replacement :

$$e_0^2 = Z^2 * \frac{PQ}{n_0}$$

without replacement:

$$e^2 = Z^2 * \frac{PQ}{n} * \frac{N-n}{N-1}$$

The following significant results can be concluded:

- The margin of error is directly proportional, (in the case of with replacement and without replacement) to the variance of the phenomenon and the desired level of confidence in the data.
- The margin of error is inversely proportional with the size of the sample (in the case of with replacement and without replacement).
- The margin of error is not related to the size of the population in the case of (with replacement), and the existence of this relation in the case of (without replacement).
- The margin of error without replacement is smaller than in the case of with replacement, for the same sample size in the two cases, because the proportion $\frac{N-n}{N-1}$ is always smaller than (1), considering that the sample size is always greater than (1).

1.2. Cluster Sampling

Sampling according to the Simple Random Sampling leads to scattering the sample units, which increases the cost of fieldwork in many statistical surveys. In order to reduce the cost of fieldwork, another method was developed, namely: Cluster Sampling. This method is based on the study of sampling units in limited areas instead of studying the units scattered in the population. Each of these areas is called a "cluster". Instead of studying the sample size of 1000 households, for example in a city, the sample units can be grouped in clusters, each is of a size equal to 50 units, and instead of collecting data about households scattered throughout the city, only 20 clusters covering all cluster units in the survey are used. Obviously, there is much cost that can be decreased accordingly. Although the Cluster Sampling Method reduces cost significantly compared with the Simple Random Sampling Method, but this method increases the sampling errors compared to the Simple Random Sampling errors, and this is one of the main disadvantages of this method. Clusters of equal sizes can be used as well as studying all the elements of the cluster. In addition, a number of cluster's elements could be studied and the number of these elements could be the same, or different, in each cluster.

1.3. Stratified Sampling

Often, it is desirable in many statistical studies to obtain independent data for some geographical areas for administrative or developmental purposes, such as: urban and rural areas, region, province, or citizens and non-citizens. In practice, there is strong justification for distinguishing between these regions, because rural population, for example, are characterized by different socio-economic and demographic characteristics than urban population. On the other hand, experience and practice proved that it is possible to reduce the random error by dividing the population into a set number of sections, in a way that the units of each of them are identical or homogeneous, and there are no significant differences between the units of each section, while there are such differences between the sections themselves. The term "stratum" is used to refer to each section, that is why this type of sampling is called: "Stratified Sampling.". This method is carried out by dividing the population into a number of strata, each is called "stratum". If the population is divided into (G) strata and the size of each are

$N_1, N_2 \ldots\ldots N_G$, respectively, the size of the population N is equal to: $N_1 + N_2 + \ldots\ldots + N_G = N$ If the size of the samples in the strata is $n_1, n_2, \ldots\ldots, n_G$ the sample size (n) is equal to: $n_1 + n_2 + \ldots\ldots + n_G = n$.

One of the most important characteristics of the Stratified Sampling is that it gives more accuracy under certain conditions, and it also allows the possibility of collecting information about each stratum. Each stratum of the target population could be considered an independent population.

1.4 Multistage Sampling

The use of the sampling methods discussed above mainly requires modern frame provided for the population in question so that the frame includes lists of all the units of the population. In practice, these frames may not be available sometimes. Usually, a modern frame is quite expensive. The study of the phenomenon of unemployment requires a modern frame for households, especially when the date of the last census is a while ago, and, definitely, a modern

frame for all households would also be very costly. In light of these consideration, a new method of sampling was used that help overcome the difficulties and problems related to the frame, by dividing the population into sub-groups, then dividing these groups into other sub-groups. And so on until the units to be studied are obtained and updated both in a short time and at an acceptable cost, provided that these sub-groups are randomly sampled. This method is called Multistage Sampling. The sampling units are selected in several stages. In the first stage, the population is divided into units that are relatively large (administrative or geographic regions), and some sub-groups are randomly selected. The units selected in the first stage are called the Primary Sampling Units (PSUs). In the second stage, the primary units selected in the first stage are then divided to appropriate units, such as the units that were used in previous census, for example, or any other suitable distribution. Then, a specific number of these sub-groups is randomly selected as well. The units selected at this stage are called Secondary Sampling Units (SSUs), and so on. In the third stage, the units selected in

the second stage are then divided to appropriate units, and then selecting a specific number in the third stage, and so on in the fourth stage till the final stage. The final sampling units are thus selected. The Cluster or Stratified sampling method can be used separately or together in every stage of the Multistage Sampling. In the surveys of households, for example, the population could be divided into strata such as urban, rural, or regions and then dividing each stratum into appropriate units, such as neighborhoods in urban cities, villages in rural areas, in the first stage, and in the second stage the neighborhoods and villages could be divided into smaller units, such as the enumeration areas (EAs) used in the last census.

The Simple or Systematic Random Sampling could be used for selecting the primary units in each stratum of the population in the first stage, or the secondary units in the second stage. In the third stage, all the households could be covered in the units selected in the second stage, or selecting a specific number randomly from each unit. The sample in this example is stratified three-stage cluster sample.

1.5 Master Sample

The housing and population census is suitable for providing modern and comprehensive frames for buildings, residential units, and households covered by the census. The frame provided by the census is a basis for designing various social and demographic surveys samples in the period following the census, however, because of the evolution of urban movement and continuous population movement and immigration, these frames become old after two or three years at most. It is even aggravated with time, i.e. it is not even possible to design a population-representative sample before updating the census frame. Definitely, this process requires effort and money, and is time-consuming. That is why the United Nations used the master sample approach in the 1980s, in the two documents issued by the United Nations Statistical Committee, the first was entitled the “National Household Survey Capability Program” issued in 1986, and the second one was entitled the “Handbook of Household Surveys”, issued in 1987.The basic principle of this approach is to design a sample of a suitable size depending on the latest household frame. This approach can be used in long periods after the last census in order to create appropriate frames for the master sample

design. The master sample frame will be substitute for the comprehensive frame provided by the census. This sample is used to design all the household surveys samples during a specific period, provided that such a frame should be updated and revised every two or three years at most to insert the changes which may have occurred after this period. This approach has been adopted and successfully implemented in many countries of the world. These publications and practices all agree on using the multistage stratified cluster method (two or more stages) for sample design. Initially, the target population is divided into two or more categories, each is called stratum, such as rural population, urban population, citizens and non-citizens. In the second stage, a random sample is selected from the units (with a certain size of population) formed in each stratum, each is called a cluster or (EAs). The EA's used in the latest census are widely-used in practical applications, usually ranging in size between about 100 and 200 households. Any other appropriate method could be used provided that the EA size is limited, so the sample units in the cluster are spread widely to increase the representation of the sample and reduce the cost of moving within the EAs. In the final stage, a sample of

households is covered in each EA of the sample units, and sometimes all households in the EA are covered.

The master sample is designed of a large EAs so as to form a basis for the formation of non-overlapping partial samples that meet the surveys need expected to be carried out during the period of two years at most. The number of EAs is determined according to the number of surveys to be carried out within a year or two. If it is expected to implement five surveys, and the sample size of each is 1000 households, for example, the master sample is formed in accordance with one of the following two approaches:

- **First Approach:** Forming a master sample of the size of 50 EAs, each is composed of 100 households, which is sufficient to sample five partial samples from each EA each is of the size of 20 households to ease the burden on households by not to be repeated across surveys more than once.
- **Second Approach:** Forming a master sample of the size of 150 EA, that constitute a basis for sampling partial non-overlapping three samples each is of the size of 50 EA,

and each is called (rotation). Then, it is possible to use the one rotation for sampling partial non-overlapping samples too for more than one survey, as in the first approach. In this case, five non-overlapping surveys for each rotated could be implemented.

1.6 Two Phases Sampling:

Focus in some surveys is on rare phenomena in the target population, such as the disabled, or high-income individuals and others. Although the size of the target population may be small, but reaching this number requires considerable cost to cover a large number of households. In order to reduce the cost, an exploratory survey is performed on a large sample of households containing a limited number of questions about the target group. This could be carried out by appending these points to other surveys being implemented, and classification of the units in which the target groups could be found. Later, in the second stage, the sample is selected from among these units, which is why this type of surveys is called: Two Phases Sampling, sometimes called Double Sample. If the decided sample size is 1000 disabled

individuals for a survey for the disabled, and it was also decided to carry out a survey of the labor force of the size of 15000 households, only one question about the presence of a disabled in the household could be included in the form, and this does not constitute a burden on the survey. In case this survey resulted in identifying 1500 disabled individuals, then a sample of the size of 1000 disabled individuals is to be selected to collect data in a detailed questionnaire for the disabled.

1.7 Rotation Sampling: The aim of the survey may be estimating change or studying the direction of some phenomena, through rotated or ongoing surveys, where the survey is rotated in successive annual or quarterly periods. According to the abovementioned publications and in light of practice, focus on this kind of surveys is on the expected margin of error to estimate the difference between the estimates of a rotation. This is related to the options of sample rotation that are limited into three methods:

- Rotation of the previous survey same sample.

- Partial rotation of the previous sample, 50% for example, and completing the sample with a new partial sample.
- Rotation of the survey with a new sample.

2. Sample Size

Sample size identification constitutes the first step in the preparation for conducting survey. Making decisions in many issues related to the expected survey results – with regard to sample accuracy and errors as well as the expected confidence in the results – is based on the sample size decision. The necessary financial and human resources as well as the period required for the survey will also be based on this basis. In practical applications, the focus is placed on calculating the sample size for average estimates or proportion of a target population, such as sample size calculation for the family's or individual's annual expenditure or income average estimate, economic sector's production estimate, unemployment estimate among populations, opinion poll proportion of satisfaction, etc. It is widely known that the simple random sample is the basis upon which the other methods have been based.

2.1 Sample Size with Replacement to Estimate the Average and proportion

- **Sample Size (with Replacement) for Average Estimate**

In order to distinguish between the sample size with replacement and without replacement, the sample size with replacement will be referred to as n_0 and the margin of error with replacement will be referred to as e_0. The relation between the sample size and the margin of error square in the simple random sample with replacement according to the Central Limit theory (CLT) is represented in:

$$e^2{}_0 = Z^2 * \frac{\sigma^2}{n_0}$$

Therefore, the sample size with replacement for the average estimate is calculated through the following formula:

$$n_0 = Z^2 * \frac{\sigma^2}{e_0{}^2} \qquad (1)$$

Where:

- n_0 is the size of the simple random sample with replacement
- Z is the value corresponding to the confidence level in the normal distribution table, which equals 1.96 at the confidence level of 95% and is adopted by most statistical surveys
- σ^2 variance is the examined phenomenon in the population (individuals, families, establishments, etc.) and is sometimes referred to as V or S^2. It is calculated through the following well-known equation: $V = S^2 = \sigma^2 = \frac{\sum_{i=1}^{i=N}(X_i - \overline{X})^2}{N}$, where N refers to the targeted population size and $\overline{X}$ is the average values of the examined phenomenon $X_1, X_2, \ldots\ldots\ldots, X_N$ in the targeted population. It is calculated through the following equation: $\overline{X} = \frac{\sum_{i=1}^{i=N} X_i}{N}$
- e_0 is the expected margin of error in the population average estimate. It refers to the total expected difference between the actual value in the targeted population and the same estimated value through the sample. Such difference might be positive or negative. The margin of error is determined through Z and the standard error (SE) at the following equation:

$$e_0 = \pm Z * SE$$

The standard error is defined as the standard deviation for the estimate variance (average or proportion) according to the Central Limit Theory. The margin of error in the practical application is determined so as the percentage between the margin of error and the estimate is not more than 15%, which is acceptable for result analysis and to meet the data users' needs. Sometimes, a maximum margin of error of 20% is adopted by some surveys. The percentage between the margin of error and the estimate (average or proportion) is called a margin of relative error and is referred to as $Mo\,\mathrm{Re}\,l$ and is calculated through the following

equations: $Mo\,\mathrm{Re}\,l_{(\bar{x})} = \frac{e}{\bar{x}}$ for the average margin of relative error and $Mo\,\mathrm{Re}\,l_{(P)} = \frac{e}{p}$ for the proportion margin of relative error. Similarly, the percentage between the standard error and the average or proportion could be calculated. It is called the relative standard error and is referred to as Rel. It is calculated through the following equations: $\mathrm{Re}\,l_{(\bar{x})} = \frac{SE}{\bar{x}}$ for the average and $\mathrm{Re}\,l_{(p)} = \frac{SE}{p}$ for the proportion. If the confidence level is 95%, the relative standard error should not be more than 7.5% (0.15/1.96); in rare cases, it reaches 10% (0.20/1.96).

It could be noticed that the sample size with replacement is not related to the population size. It is directly proportional to the variance and the confidence level; and it is inversely proportional to the margin of error square.

Example 1 : if the variance is $V = \sigma^2 = 4000$ in populatin of 1000 units, and the average according to the latest census or recent studies is $\overline{X} = 150$. And the expected margin of error is 5% of the avergae value, with a confidence level of 95%, i.e. the margin of error is $e_0 = .05 * \overline{X} = 7.5$. Then, the sample size through implementing equation No. (1) is 274.

$$n_0 = Z^2 * \frac{\sigma^2}{e_0{}^2} = 1.96\text{^}2 * \frac{4000}{7.5\text{^}2} = 273.18$$

It is worth mentioning that the calculated sample size is of the same type as the examined pnhenomenon (families, individuals, category of populations, establishments, etc.).

The sample size could be calculated through the standard error :

$$n_0 = Z^2 * \frac{\sigma^2}{Z^2 * SE^2} = \frac{\sigma^2}{SE^2} = \frac{4000}{\left(\frac{7.5}{1.96}\right)^2} = \frac{4000}{3.8265\text{^}2} = 273.18$$

Or through the relative margin of error, which equals $\mathrm{Re}\,l_{(\bar{x})} = \frac{e}{\bar{x}} = \frac{7.5}{150} = 0.05$

$$n_0 = Z^2 * \frac{\sigma^2}{\mathrm{Re}\,l^2_{(\bar{X})} * \bar{X}^2} = 1.96\text{^}2 * \frac{4000}{0.05\text{^}2 * 150\text{^}2} = 273.18$$ Or through the coefficient of variation (CV), where $\bar{X}$ is the average and the coefficient of variation $CV = \frac{\sigma}{\bar{X}}$ equals $CV = \frac{\sigma}{\bar{X}} = \frac{\sqrt{4000}}{150} = 0.42164$

$$n_0 = Z^2 * \frac{CV^2(\bar{X})}{\mathrm{Re}\,l^2_{(\bar{X})}} = 1.96\text{^}2 * \frac{0.42164\text{^}2}{0.05\text{^}2} = 273.18$$

The result is the same in all cases.

- **Sample Size (with Replacement) for Proportion Estimate**

The above items apply for the sample size calculation for proportion estimate. This is because the proportion is a special average case, where any value in the population could be either 1 when the phenomenon of interest is consistent or 0 in other cases. The proportion P is calculated through the previous average equation, where the numerator equals the number of cases in consistent with the phenomenon and the denumerator equals the number of cases in population (consistent and inconsistent with the phenomenon). The complementary proportion Q equals $Q = 1 - P$ and the proportion variance is according to Bernoulli equation: $V = \sigma^2 = P * Q$

The replacement in the previous margin of error square equation would lead to finding the following similar equations for sample size calculation to estimate the proportion with the margin of error:

$$n_0 = Z^2 * \frac{P * Q}{e_0^{\ 2}} \qquad (2)$$

Example 2 : if you need to calculate the sample size to estimate a proportion, which was – in previous studies – $P = 0.30$ (proportion of smokers out of total population at the age group of 18-40 year, for example) and the expected margin of error is 10% of the proportion with a confidence level of 95%, i.e. the margin of error is $e_0 = 0.10 * 0.30 = 0.03$. Then, the sample size with replacement and with the margin of error through applying equation (2) is 897 individuals at the targeted age group.

$$n_0 = Z^2 * \frac{P * Q}{e_0^{\ 2}} = 1.96\text{^}2 * \frac{0.30 * (1 - 0.30)}{0.03\text{^}2} = 896.37$$

2.2 Sample Size (without Replacement) for Average Estimate and proportion

- **Sample Size for Average Estimate**

The relation between the sample size and the margin of error square at the simple random sample without replacement according to the central limit theory is the following equation:

$$e^2 = Z^2 * \frac{\sigma^2}{n} * \frac{N - n}{N - 1}$$

For comparison and ease of calculation, the sample size with replacement and the sample size without replacement are ususally linked to each other, in practical application, according to same margin of error and cofidence level. Through replacing the margin of error square at the last equation with the margin of error with replacement, the equation would be:

$$Z^2 * \frac{\sigma^2}{n_0} = Z^2 * \frac{\sigma^2}{n} * \frac{N-n}{N-1} \Leftrightarrow \frac{1}{n_0} = \frac{1}{n} * \frac{N-n}{N-1} = \frac{1}{n} * \frac{N}{N-1} - \frac{1}{N-1}$$

Through multiplying the two sides of the last equation by $\frac{N-1}{N}$, it would be:

$$\frac{1}{n_0} * \frac{N-1}{N} = \frac{1}{n} - \frac{1}{N} \Leftrightarrow \frac{1}{n_0} * \frac{N-1}{N} + \frac{1}{N} = \frac{1}{n}$$

According to the last equation, the sample size without replacement is calculated through the following equation:

$$n = \frac{1}{\frac{1}{n_0} * \frac{N-1}{N} + \frac{1}{N}}$$

Through multiplying the numerator and denuminator in the second side by n_0, the equation would be:

$$n = \frac{n_0}{\frac{N-1}{N} + \frac{n_0}{N}} \qquad (3)$$

It is the general equation for the sample size without replacement and its relation with the sample size with replacement n_0. Note that the sample size is based on the target population size N, unlike the sample size with replacement.

The sample size without replacement for the average estimate, in example No. 1 through applying equation No. 3, equals 216, which is smaller than the sample size with replacement, i.e, 274, which is the general case for the same conditions.

$$n = \frac{n_0}{\frac{N-1}{N} + \frac{n_0}{N}} = \frac{274}{\frac{1000-1}{1000} + \frac{274}{1000}} = 215.24$$

When dealing with larger population, which is common in the practical applications, the percentage $\frac{N-1}{N}$ is close to one. In this case, the following equation for the sample size without replacement could be called 'short equation for large samples' and written as follown:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad (4)$$

Through applying the short equation on large population, the sample size would also be 216, as $\frac{N-1}{N} = 0.999$ nearly equals 1.

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{274}{1 + \frac{274}{1000}} = 215.07$$

When the percentage $\frac{n_0}{N}$ is less than 5%, which is the case in very large population, then the difference between the sample size with replacement and without replacement would be small. If the population size in this example is 100,000, the sample size without replacement is 274 $n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{274}{1 + \frac{274}{100000}} = 273.25$

which is the previously calculated sample size with replacememt, as the percentage $\frac{n_0}{N}$ is very small and less than 5%.

- **Sample Size (without Replacement) for proportion Estimate**

Following the previous steps for sample size calculation for average estimate, we could find a matching equation for the last two equations (general and brief ones) for sample size calculation for proportion estimate, taking into consideration that n_0 represents the sample size with replacement and is achieved through equation No. 2 above:

$$n_0 = Z^2 * \frac{pQ}{e_0^2}$$

The sample size without replacement for proportion estimate is given through the following general equation:

$$n = \frac{n_0}{\frac{N-1}{N} + \frac{n_0}{N}} \qquad (5)$$

Or through the following brief equation for larger population:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad (6)$$

The sample size without replacement through applying the short equation for larger population in the previous example is 473, which is less than the sample size with replacement 897, which is the general case. Therefore, the survey cost is less:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{897}{1 + \frac{897}{1000}} = 472.85$$

On the other hand, in the practical application, the sample size is not completely covered due to nonresponse. To avoid this, the sample size should be increased in all of the previous equations through multiplying them by a coefficient called 'nonresponse coefficient' k, which equals the inverted expected response rate $\frac{1}{RR}$, where RR is the response rate. The previous surveys are usually useful for estimating the expected response rate in the present survey. Therefore, the sample size calculated in previous equations should be multiplied by the nonresponse coefficient to get the final sample size.

If the expected nonresponse rate in example No. 1 equals 10%, then the response rate would be 90% and the nonresponse coefficient would be $k = \frac{1}{RR} = \frac{1}{0.90}$. So, the final sample size, taking into consideration nonresponse, equals the previously calculated sample size multiplied by nonresponse coefficient, i.e. the sample size should be 304 to get the required sample size, which is 274.

$$n = n_0 * k = Z^2 * \frac{\sigma^2 * k}{e_0^2} = 1.96^2 * \frac{4000}{7.5^2} * \frac{1}{0.90} = 303.5$$

2.3 Sample Size to Estimate a Number of Indicators

In the previous section, we calculated the sample size to estimate the average or proportion for a specific phenomenon. In rare practical applications, the objective of statistical survey is limited to estimating the average or proportion of a single phenomenon, but the survey objective is to estimate many indicators that could, usually, be in the form of proportions or averages. The sample size that corresponds to the estimate of one of such indicators is different than the sample size in the estimate of other indicators. The question here is what is the appropriate sample size for some, or all, of such indicators? Some aspects of this issue have been tackled in the survey techniques publications – such as the works of Kish, Cochran,

and others. – through the sample size calculation for a limited number of main indicators and adopting the larger sample size according to a certain confidence level.

In this context, we could benefit from the proportion variance or the average of appropriate sample size estimate for all of the survey expected indicators, not only for a limited number. This would be handled in case of replacement, due to the relation between the sample size with replacement and sample size without replacement to achieve the same confidence level and margin of error. Back to the sample size with replacement for proportion estimate

$$n_0 = Z^2 * \frac{pQ}{e_0^2}$$

We assume that $a^2 = \frac{Z^2}{e^2_0}$ and as $Q = 1 - P$, then if we reversed this, the sample size would be:

$$n_0 = a^2 * p * (1 - p) \qquad (7)$$

The last equation is a function for sample size – considering the proportion as an independent variable – and represents a Parabola equation with a maximum end when $p = 0.50$, taking into consideration that the proportion is $0 \le p \le 1$. Figure No. 1 below shows the graph line for this equation which is equivalent for the following both values: $Z = 1.96$ and $e_0 = 0.05$.

Back to the sample size for average estimate $n = Z^2 * \frac{\sigma^2}{e^2}$ in order to compare it with the sample size for proportion estimate, and assuming that the relative margin of error for the average estimate $\frac{e}{\overline{X}}$ equals the margin of error for proportion estimate e_0, which

could be represented as follows $\frac{e}{\overline{X}} = e_0 \Leftrightarrow e = e_0 * \overline{X}$, and through reverse, the sample size equation for average estimate would be $n = Z^2 * \frac{\sigma^2}{{e_0}^2 * \overline{X}^2}$.

It is known that the coefficient of variancion square is $CV^2 = \frac{\sigma^2}{\overline{X}^2}$ and through reverse, the sample size equation for average estimate would be $n = Z^2 * \frac{CV^2}{{e_0}^2}$.

Considering $a^2 = \frac{Z^2}{{e^2}_0}$ as before, the final equation would be:

$$n = a^2 CV^2 \qquad (8)$$

It is a function for the sample size for average estimate considering that the coefficient of variancion is the independent variable $CV \geq 0$. it also represents a Parabola equation with a small end when the coefficient of variancion is zero. For comparison, a part of it was graphically represented when $Z = 1.96$ and $e_0 = 0.05$, as in the case of proportion estimate. Refer to Figure No. 1. Many similar figures may be achieved through changing the margin of error and confidence level. We could note the following characteristics and results from Figure 1:

1. Sample size for proportion estimate calculated through equation No. 7 reaches its high end and equals 385 when the proportion is $P = 0.50$ and the proportion variance would be the maximum $V = P*(1-P) = 0.25$. If the proportion is 0.20, then the corresponding sample size would be 264; and if the proportion is 0.60, the corresponding sample size would be 369. Both of them are less than 385. This also applies to all of the other proportions. The maximum sample size would be appropriate for the estimate of all proportions less, or more, than 0.50 for the achievement of the same margin of error and confidence level. Note that the

corresponding sample size for 0.40 equals the corresponding sample size for 0.60; both of them are 369 as the proportion variance is the same in both cases $V = P*(1-P) = 0.40*0.60 = 0.24$

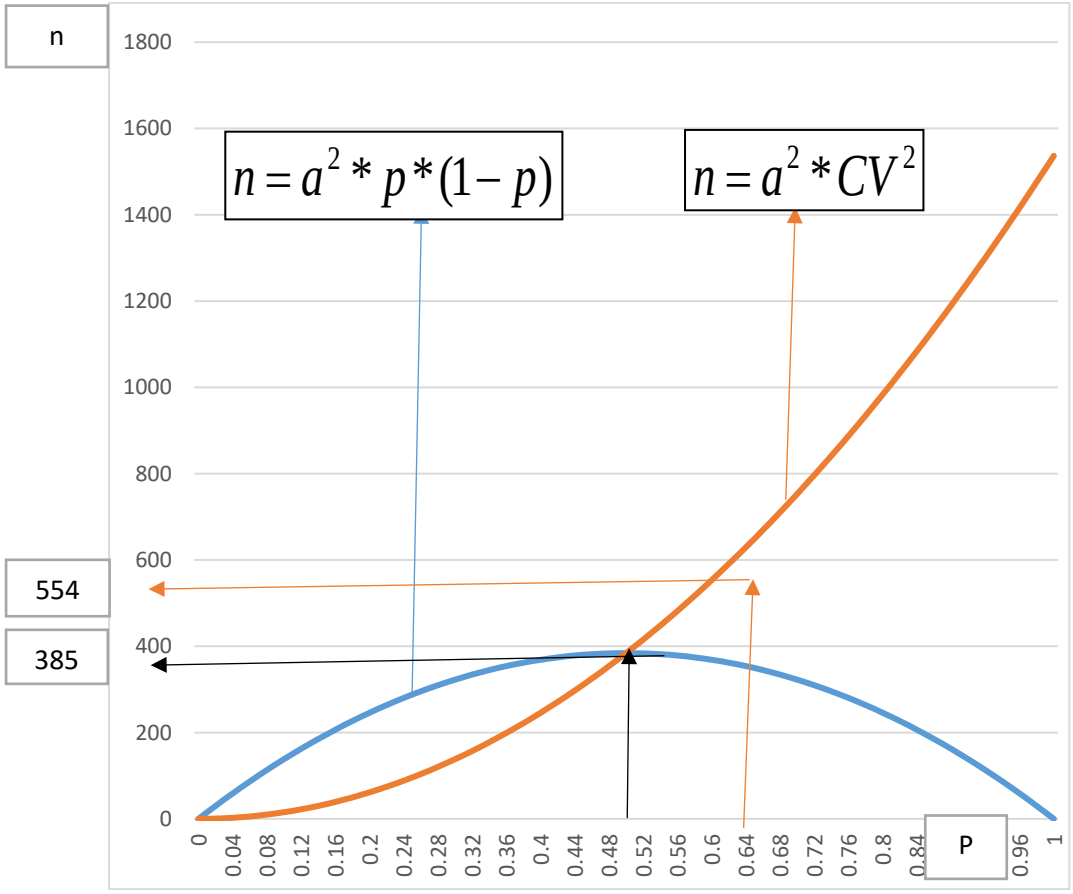


Figure No. (1)

This result is widely used in practical applications, especially for the sample size calculation for opinion polls, where the objective is often to estimate many relative indicators. Many international opinion poll institutions adopt this result to calculate the poll sample size.

2. If the coefficient of variancion also equals the same value $CV = 0.50$, the sample size calculated through equation No. 8 would be 385 and the sample size would be the same for both the proportion and the average. In all cases where the coefficient of variancion is less than 0.50, the proportion maximum sample size is more than the average sample size. The

average sample size for $CV = 0.40$ with a confidence level of 95% and a margin of error of 0.05 is 246, which is less than the proportion maximum sample size. This is the general case, provided that the coefficient of variancion must be less than 0.50. The proportion sample size would be appropriate for the estimate of all proportions, as shown above, and for the estimate of all averages with a coefficient of variancion less than 0.50. In this case, it would be sufficient to calculate the sample size to estimate the proportion through equation No. 7. This result has a great importance in practical applications. If the coefficient of variancion is less than or equals 50%, the maximum sample size calculated to estimate the proportion would also be appropriate for the average estimate. In this case, we could dispense with the (average or proportion) variance estimate in the targeted population, which is not available in most cases as the sample size corresponding to the proportion of 50% is a maximum one and is appropriate for the proportion and average estimate.

3. On the other hand, if the coefficient of variancion is more than 0.50, the sample size for average estimate would be larger than the maximum sample size for proportion estimate. And the sample size calculated for averge estimate through equation No. 8 would be appropriate for the estimate of all proportions. If the coefficient of variancion is 60%, the sample size for average estimate would be 554, which is larger than the sample size for proportion estimate, i.e.369.

4. Equation No. 3 is appropriate for the estimate of one average , but it would not be appropriate for the estimate of other averages of interest. It is clear that there is no maximum sample size for the estimate of all averages as the case in the maximum sample size for the estimate of all proportions. This is due to that the function in equation No. 8 always increases. Due to the impossibility of calculating the sample size for all of the averages to be estimated for the diffciulty of identifying the variance for each of them, the targetd population could be allocated to homogenuse strata, where the coefficient of variance for the averages of interest is less than or equals 0.50 in all strata. In this case, equation No. 1 could be applied for each strata through adopting the maximum sample size in each strata to estimate the proportion.

2.4. Sample Size for Household Surveys

In many social and demographic surveys, the targeted population is a certain category of population – individuals of the labor force in an unemployment survey, persons of a certain age, children less than five years, reproductive women (15-49), etc.). The simple random sample size with replacement – calculated to estimate the proportion or average through equations 1 and 2 – would refer to the sample size for the examined phenomenon, which represents the same targeted category members. If we need, for example, to calculate the unemployment rate, the sample size would be composed of members from the labor force; and if we need to estimate the proportion of a healthy habit among youth, the sample size would be composed of youths. Usually, no appropriate frame for the targeted category are available to get the required sample units, while there are available household frames from the last census or from other sources such as administrative records. The household could be adopted as a semi-final statistical unit to get the household sample corresponding to the individuals sample in the population targeted category. In this case, we should transfer the individuals sample size into household one. So, we should identify the phenomenon rate among population r and the average household size $\overline{n}$; the household sample size would, then, equal the individuals sample size divided by $r * \overline{n}$, which indicates the examined phenomenon size in each single household.

Through compensation for this in equation 1 and 2, with expected nonresponse coefficient k, they would be as follows:

Household sample size with replacement corresponding to the individuals sample for average estimate would be calculated through the following equation:

$$n = Z^2 * \frac{\sigma^2 * k}{e^2{}_0 * r * \overline{n}} \qquad \textbf{(9)}$$

The equation similar to the household sample size with replacement for proportion estimate would be:

$$n = Z^2 * \frac{p * Q * k}{e^2{}_0 * r * \overline{n}} \qquad \textbf{(10)}$$

In order to calculate the sample size without replacement, the equations from 3 to 6 – whether general or short for large population – should be applied.

Example 3 : if the sample size calculated in example No. 1 represents 274 individuals of a category with a fixed rate of 10% of population, i.e. $r = 0.10$; and if the average household size is $\overline{n} = 5$ and the nonresponse rate is 15%, i.e. the expected response rate is 85%, then the household sample size through applying equation No. 9 would be 643 households:

$$n = Z^2 * \frac{\sigma^2 * k}{e^2{}_0 * r * \overline{n}} = 1.96^2 * \frac{4000}{7.5^2 * 0.10 * 5 * 0.85} = 642.8$$

Example 4 : if the targeted age group in example 2 among population is 40%, the average household size is 5 members and the nonresponse rate is 5%, then the household sample size through applying equation No. 10 would be 1820 households:

$$n' = Z^2 * \frac{p * Q}{e^2{}_0 * r * \overline{n} * RR} = 1.96^2 * \frac{0.10 * 0.90}{0.01^2 * 0.40 * 5 * 0.95} = 1819.7$$

Sample size without replacement could be calculated through applying the equations from 3 to 6 – whether general or short for large population– afer identifying the population size N. If the targeted population size in this example is 10,000 households, then the simple random sample size without replacement in example No. 3 through applying equation No. 3 would br 605 households, which is somewhat less than the sample size with replacement due to that the targeted population is large. This is the general case; therefore, the sample size with replacement is adopted in most of the practical application in large population.

$$n = \frac{n_0}{\frac{N-1}{N} + \frac{n_0}{N}} = \frac{643}{\frac{10000-1}{10000} + \frac{643}{10000}} = 604.2$$

2.5. Sample Size for Other Sample Types

The sample size calculated in the previous equations represents the simple random sample size. In practice, a number of other sample types are adopted, such as stratified or cluster samples or adopting some, or all, of them at the same time, which is called the complex sampling method. The sampling theory provides the mathematical equations for sample size calculations for each type of these samples as the case for mathematical equations for the above mentioned simple random sample size calculation. However, the complexities in the application of such equations prevent their wide adoption in practical applications. This could be compensated in practice with the calculation of the design effect coefficient, referred to as $deff$. It represents the percentage between the estimate variance in other types sample and the corresponding variance in the simple random sample of the same sample size (KISH, 1965). Note that each of them represents the estimate random sampling error square. If the estimate variance in the simple random sample is V_{srs} and the corresponding variance is V_s in other types sample, then the design effect would be calculated through the following equation:

$$deff = \frac{V_s}{V_{srs}}$$

For example, if the cluster sample is used to estimate a phenomenon rate and its variance is $V_s = 0.0025$ and the variance for the same rate of the same size at the simple random sample is $V_{srs} = 0.0016$, then the design effect would be $deff = \frac{0.0025}{0.0016} = 1.56$.
It is clear that the cluster sample variance is larger than the simple random sample of the

same size with a value that equals the design effect value. Whereas the estimate variance is inversely proportional to the sample size, then the other types sample size n' is related to the simple random sample n through the following equation:

(11) $$n' = n * deff$$

This result is greatly important for other types sample size calculation with the significance of the simple random sample size, calculated through the previous equations, and of the design effect, calculated at previous surveys. Many publications recommend to consider the design effect value as 1.5 or 2, if not previously available. If the simple random sample size in the previous example is 350, the cluster sample size should be 546 (350*1.56) to get the same accuracy.

2.6 Sample Allocation to strata

In the previous section, we calculated the sample size for all types of samples in order to estimate the average or proportion at the level of the targeted population according to fixed confidence level and margin of error. In stratified sample, the targeted population is divided into strata, and the sample size should be defined at each stratum. This is applied in practical applications through one of the following methods:

First Method: calculating the sample size at the level of each stratum, considering each stratum as an independent population. Therefore, the sample size should be calculated through one of the above mentioned methods according to fixed confidence level and margin of error at the level of each stratum. The sample size could be multiplied compared to the sample size at the target population level and this may would not be consistent with the available human and financial resources.

Second Method: calculating the sample size at the target population level in accordance with fixed confidence level and margin of error and in consistent with the available resources. Then, the sample size should be allocated among the strata, whose number is h. If N is the target

population size and N_i is the stratum size, then $N = \sum_{i=1}^{h} N_i$. If σ_i^2 is the variance of stratum i , then the allocation of a fixed size n among the strata into which the target population was divided would be done through one of the following allocation methods that were recommended in the samples puplications:

- **Optimal or Neyman Allocation (1934):** it is the most commonly used method in practical application. Sample size is allocated to strata, according to this method, so as the variance σ^2_{st} at the level of the target population is as small as possible. This could be achieved if the sample size n_i in stratum i is in consistent with the product $N_i * \sigma_i$; the stratum sample size would be calculated throuh the following equation:

$$n_i = n * \frac{N_i * \sigma_i}{\sum_{i=1}^{h} N_i * \sigma_i}$$

- **Proportional Allocation:** the sample is allocated, according to this method, considering that the sampling fraction $f_i = \frac{n_i}{N_i}$ is the same in all strata and equals $\frac{n}{N}$. The sample size in each stratum would be consistent with its size for the targete population size. It should be calculated through the following equation. This allocation method is often used in practical application for its easeness.

$$n_i = n * \frac{N_i}{N}$$

- **Bankier Allocation or Power Allocation (1988):** sample allocation to strata through optimal or proportional allocation methods is appropriate for different estimates at the level of the target population according to fixed margin of error and confidence level. If the objective of the survey is achieving estimated of acceptable level of accuracy at the level of each stratum– as the case in most surveys – then the sample size in the small stratum, according to both of the above allocation methods, would be small and could not be appropriate for the desired level of accuracy for the estimates at the level of each stratum. Therefore, the sample should be allocated in proportion

  to the product $N_i^{\alpha} * CV_i$, where $0 \leq \alpha \leq 1$ and CV_i is the coefficient of variance in stratum i for a certain phenomenon. The sample size n_i in stratum i is calculated through the following equation:

$$n_i = n * \frac{N_i^{\alpha} * CV_i}{\sum_{i=1}^{i=h} N_i^{\alpha} * CV_i}$$

According to the above, an appropriate sample size is allocated for the small stratum and the sample size allocated for medium or large stratum would still be acceptable for the concerned estimates. The following examples shows the allocation of a 300-unit-sized sample to three strata of a population with a size of 850 units according to the different allocations, considering that $\alpha = 0.5$

stratum	Stratum Size N_i	Standard Deviation σ_i	Average $\bar{X}_i$	Coefficient of variance CV_i	Neyman Allocation	Proportional Allocation	Bankier Allocation
1	100	12	3	4	21	35	52
2	250	25	5	5	107	88	103
3	500	20	4	5	172	177	145
Total	850				300	300	300

3. Sample weights

3.1. Random selection and withdrawal probability

The first step in the design of the survey sample is to identify the target population, which is the units to be covered by the survey. In a survey on the residents of a specific country, the target population is all the residents in that country, while the target population in an economic survey is all Economic and social institutions. In practice, there may be some limitations and constraints for not covering certain units of the target population. For example, in a labor force survey, individuals under 15 years old are not covered. The lack of coverage may be due to the high cost or difficulties in accessing such units; e.g. national residents living abroad or in public housing, such as hospitals or hotels, as well as units located in disaster areas. This type of unit should then be deleted from the target population and the remaining units are then called the "Survey Population", where the units of this population are called the "Elementary Units"; the households in a family survey, the establishments in an economic survey, individuals at the age of 15 and above in labor force survey. Information is collected from a sample of the primary units, called Sampling Units, to be used in the estimates of the survey population.

The sample units may be identical to or different from the elementary units. Access to the elementary units may require the implementation of a number of phase. In the cluster sample, the survey population is divided into clusters, where each includes a specified number of elementary units. These clusters may be equal or different in size. A number of clusters are

randomly drawn and then all cluster units or a partial sample are also randomly selected in a similar manner to the multi-stage sample, where the elementary units are reached through a number of stages; e.g. reaching 15 years or older individuals in the labor force survey may

require drawing a sample of neighborhoods in cities and a number of villages in rural areas in the first stage. Sample units at this stage are called Primary Sampling Units (PSUs). In the second stage, a sample is withdrawn from the households from each initial unit, where the sample units at this stage are called the Secondary Sampling Units (SSUs). In the third stage, information is collected from qualified individuals at the age of 15 and older in the households withdrawn in the second stage. These units should be defined at each stage, hence the so-called Sampling Frame, which means the list of units from which the sample is drawn at each stage of sampling.

As sample units may be identical to or different from the elementary units of the survey population, the formula between the sample units and the elementary units should be determined at each stage so that each unit is ensured to have a specific probability of withdrawal that is different from zero at all stages of drawing.

The frame represents the cornerstone of the sample design and implementation of all subsequent stages. The success of the sampling program depends largely on the availability of appropriate frame for random sampling. The model sampling frame is the most recent one that covers all units of the survey population, with no units alien to the target population, ensuring that each unit occurs only once without repetition. The availability of maps

accompanying these frames is particularly important in locating sample units, facilitating fieldwork and reducing costs, thus increasing the survey efficiency.

Random sampling in expressed in the probability theory by each drawn unit having a probability different from zero. As is known, probability is a measure that reflects the degree of success in the occurrence of a specific event or measures the success in reaching a specific result in a given experiment. The probability value is calculated proportionally between the event-compatible cases and the number of possible cases; the probability of withdrawing one ball out of 10 balls is equal to 1/10 and the probability of getting one of the faces upon tossing a dice equals 1/6 (because the number of possible cases is 6). Also, the probability of obtaining a red card from the playing cards is equal to 26/52 (because the number of compatible cases is 26 and the number of possible cases is 52). Similarly, the probability of withdrawing one unit from a population of N units equals $P = \frac{1}{N}$ and the probability of withdrawing one unit in a sample of n size withdrawn from that population equals $P = \frac{n}{N}$ (because the number of compatible cases is n the number of possible cases is N), among other examples. Generally, if the number of compatible cases n for the occurrence of the event and the number of possible cases N , the probability of occurrence of this event, symbolized by P , can be is calculated by the following formula: $P = \frac{n}{N}$

As noted, the number of cases corresponding to the event is greater than or equal to zero, which is smaller or equal to the number of possible cases; hence the probability is greater than

or equal to zero and smaller or equal to 1 always, which can be expressed by the following formula: $0 \leq P \leq 1$

In practice, there is a wide range of units covering the term Unit. Unit in statistical surveys may be a member of the population, a family, a facility, an enumeration unit, a neighborhood or a village, among other examples. The above definition applies to the possibility of withdrawing any of these units in all these examples. The method of drawing these units, according to this principle, is called Simple Random Selection. This method is the direct application of LLN and CLT. The sample drawn in this way is called a Simple Random Sampling (SRS).

In practical applications, sample units are randomly drawn in a number of ways, including traditional ones; such as the known balloting method or random number table method. These two methods were widely used before using modern PC-based methods and several ready software; MINITAB, EXCEL, SASE and others. A distinction is made between the so-called Draw with Replacement and Draw Without Replacement, where the first includes returning the withdrawn unit to the population before the next drawing process, while the second method does not include returning the unit to the population after it is withdrawn. Drawing without replacement method is used in most practical applications, especially when dealing with units damageable after being tested, such as testing the duration of a sample lamp. The lamp cannot be returned after it has been tested and damaged. The second method is usually used to draw sample units in statistical surveys. The main methods of random drawing in practical applications and the probability of drawing sample units shall be discussed below.

1. Systematic Sampling Method

The method of systematic withdrawal is one of the most widely used methods in random sampling, by dividing the target population, size N , into a number of groups equal to the sample size n , selecting one unit from each group, according to the following steps:

Step 1: Numbering population units from (1) to (*N*).

Step 2: Target population shall be divided into n groups of equal size. To achieve this, the size of the population shall be divided by the sample size. The outcome is called Selection Interval, symbolized by (I); so $I = \frac{N}{n}$

The draw period may be an integer, such as 10, 7 or another, and it may be a number consisting of an integer and decimal fraction, such as 15,12. Sample units are withdrawn accordingly as follows:

- **Draw period is an integer**

A random number shall be selected between (1) and the number referring to the duration of the draw period, symbolized by (R) and called the Random Start. The unit which order in the population equals (R) is the first unit in the sample. The order of the second unit in the sample is obtained by adding the draw period to the random start, so that the second unit order in the sample shall be the corresponding order in the population. The same goes for the third unit until a number of units is withdrawn equal to the sample size. It can be noted that

the order of each unit is calculated by adding the draw period to the order of the unit preceding it, therefore the order of the sample units is calculated as follows:

$$R, R+I, R+2*I, ..., R+(n-1)*I$$

For example, drawing 3 out of 15 students systematically shall be conducted as follows:

- Numbering students from 1 to 15
- Draw period is calculated, which equals to $I = \frac{15}{3} = 5$
- Random number is chosen between (1) and (5) by drawing, using the random number tables, computers or the formula RANDBETWEEN (1,5) is 2, then it becomes the random start and the student (2) is thus the first student in the sample.
- To obtain the student's second order, add the draw period to the random start, then you get number 7, indicating the second student in the sample.

  Thus, the order of the third student is obtained by adding the draw period to the previous student's order, equaling 12.
- The sample is, then, students with numbers 2,7,12.

It is noted that the difference between the order of two consecutive units in the sample is equal to the draw period, which is the general rule in the systematic withdrawal method.

- **Draw period contains a decimal number**

If the number of students in the previous example is 16 students as in the following table:

The draw period is composed of an integer and a decimal number and a sample of 3 students shall be withdrawn, as follows:

- Draw period shall be calculated, equal to $I = \frac{16}{3} = 5.3$
- A random number between (1) and (0) shall be selected using the formula $= RAND(\ )$, let it be 0.12
- The random number is to be multiplied by the withdrawal duration. If the result is an integer, it shall be the order of the first unit in the sample. If the result is a number containing a decimal fraction, the order of the first unit shall be the integer that follows the number produced directly in the population. The result in this example is 0.636 (5.3*0.12), so that the first student order in the sample shall be the integer in the population that follows the resulting number, which is number (1), and corresponds to the student whose number is (1) in this class.

Student	Period	Random	Order of sample		Sample
	I	R	$I * (R + i)$		
1	5.3	0.12	0.639		1
2			5.936		6
3			11.236		12
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					

- To obtain the second student's order in the sample, add (1) to the random number and multiply the result with the draw period, so the result would be the order of the second student in the sample; hence the second student's order is 5.936 (5.3*1.12), corresponding to student number 6 in this class.
- Thus, number (2) is added to the random number and multiplied with the draw period to obtain the order of the third student, which is 11.236, and the order of third student shall be the 12th student in this class.

If the symbol for the random number is indicated by the code (r), the random order of the sample units shall be calculated as follows: $I*(r+i)$

Where $i = 0,1,2,\ldots\ldots,n-1$

It is noticeable that the difference between the random order of two consecutive units in the sample is equal to the draw period, as well and as if the draw period is an integer; that is to say, the order of the sample units is decided by adding the draw period to the order of the previous unit. A number of samples can be obtained by varying the draw duration, which are different and non-overlapping

samples, so that the units shall not be repeated in different samples. This result is of great importance in relieving the burden on the target elements in the various surveys.

This method can be applied to withdraw a number of clusters from the target population in the cluster sample, or to withdraw a number of primary sampling units (PSUs) in the multi-phase sample. Such units may be equal or unequal in size, which is the general case (Clusters in the cluster sample, count units in the census, villages in the countryside, urban neighborhoods, etc.). If the number of clusters or primary units A in the target population (equal or unequal in size) and the number of units to be withdrawn at this stage is a , drawing these units can be performed in a systematic way by either of the two previous methods as well, where the draw period equals $I = \frac{A}{a}$.

Based on the previous definition of the probability of withdrawing one unit of the sample, and based on the method of systematic withdrawal, the target population was divided into n group, where the size of each equals the draw period I . The probability of withdrawing one unit from each group, symbolized by the symbol (P) according to the above, is equal to the

number of withdrawn units (1) divided by the group size (the draw period) according to the following formula:

$$P = \frac{1}{I} = \frac{1}{\frac{N}{n}} = \frac{n}{N} = f$$

The sampling probability is constant and is called Sampling Fraction. In this case, the sample is called a self-weighted sample because of the probability of withdrawing the sample units is equal. The probability of a student's withdrawal

in the first example is 5 and the probability of the student's withdrawal in the second example is 5.3. When the initial units are equal or unequal in size, their withdrawal is similar to the withdrawal of the regular sample units. Instead of drawing one element, an initial unit consisting of a number of elements is withdrawn. The probability of drawing the initial unit is equal to the sum of (1) divided by the draw period. The probability of withdrawing the units is equal and the sample is also self-weighted. This probability is calculated in the following formula:

$$P = \frac{1}{I} = \frac{1}{\frac{A}{a}} = \frac{a}{A}$$

In contrast, the probability of withdrawing the units may be unequal, and then the sample shall not be self-weighted, as indicated in the following paragraph.

2. Probability Proportional to Size (PPS)

In practice, it is rare for initial units to be equal in size, then the initial units shall be drawn by PPS through the randomized systematic method; which is the method most commonly used in practical applications.

If M_i is the size of the initial unit i, then a shall be withdrawn from the primary units in this way according to the following steps:

- Preparation of the ascending assembly of the primary units' size.
- Determining the draw period which is equal to $I = \frac{\sum_{i=1}^{i=A} M_i}{a}$
- The steps in the previous paragraph 1.1 shall be applied if the draw period is an integer, and paragraph 1.2 if the draw period contains a decimal number, in order to determine the order of random units.
- A value is determined in the ascending assembly that is directly equal to or greater than the value obtained in one of the two preceding methods, so the corresponding unit shall be the required random unit.

For example, the following table shows the draw of 4 enumeration units out of 19 units of unequal size. This was done as follows:

- Prepare the ascending assembly of the initial units' size, as in the third column of the table.

- Determine the draw period, which is equal to $I = \frac{\sum_{i=1}^{i=A} M_i}{a} = \frac{2386}{4} = 596.5$

- A random number between (1) and (0) shall be selected using the formula $= RAND(\ )$; let it be 0.013
- Calculate the random order of the first unit in the sample using the previous formula in 1.2, which equals $I * (r + i) = 596.5 * (0.013 + 0) = 7.75$
- The first value in the ascending order is the one that is directly greater than this value, which is 104, corresponding to the first count unit, so that the count unit of number (1) is the first unit in the sample.
- The random order of the second unit equals $I * (r + i) = 596.5 * (0.013 + 1) = 604.25$, so the second unit in the sample is the count unit number (6).
- Thus, the order of the third and fourth units is determined. The sample units are the count units with numbers: 1,6,10,15. A number of different samples can be obtained by changing the random start each time.

Units	Mi	Comul	l=sumMi/4	r	l*(r+i)	Sampling units
1	104	104	596.5	0.013	7.75	1
2	130	234				
3	135	369				
4	134	503				
5	101	604				
6	100	704			604.25	6
7	114	818				
8	112	930				
9	150	1080				
10	146	1226				
11	109	1335				
12	146	1481			1200.8	10
13	131	1612				
14	147	1759				
15	103	1862			1797.3	15
16	135	1997				
17	132	2129				
18	139	2268				
19	118	2386				
Total	2386					

It is clear that the probability of withdrawing the initial unit i , which size is M_i equals its size divided by the draw period

$$P = \frac{M_i}{I} = \frac{M_i}{\frac{\sum_{i=1}^{i=a} M_i}{a}} = \frac{a}{\sum_{i=1}^{i=a} M_i} * M_i = \beta * M_i$$

Where $\beta = \frac{a}{\sum M_i}$ is a fixed number. It is obvious from this formula that the probability of withdrawing the units is different and the sample is not self-weighted. The probability of withdrawing the initial unit is commensurate with

its size with the coefficient β. This method is therefore called "Draw with probability proportional to size".

4. Sequential Simple Random Sampling

Sample units are withdrawn in this manner according to the following steps:

- Numbering frame units which number N is from 1 to N
- Each unit assigns a randomly assigned number between (1) and (0), using EXCEL program with formula $= RAND(\ )$

- These numbers are to be put in ascending order, and then a random number is to be selected between (1) and (N),using EXCEL, with formula $= RANDBETWEEN(1, N)$, so the corresponding number in the random order is random draw start or the corresponding number in the sequence of units.
- A number of serial numbers is to be determined after the random number, equal to the required sample size. If the end of the order is reached before the desired sample size is withdrawn, the sample size should be completed sequentially from the beginning of the order. Sometimes, for ease, the beginning of the ascending order is the beginning of the random draw, and a number of serial numbers are determined after this number, equals the sample size required.

For example, to randomly draw 3 students out of 15 students in a class, the following steps shall be adopted:

- Students are to be numbered from 1 to 15, as in the first column of the following table.
- A random number between (1) and (0) shall be assigned to each student in the previous formula, as in the second column.
- Random numbers shall be arranged in an ascending order, followed by the order of the students as in the third and fourth columns.
- If number (5) is the random number selected between (1) and 15 in the previous formula, it shall be assigned in the random order and shall be the number of the first student in the sample. The second student number in the sample is the directly next number in the random order. The same goes for the third student and the sample units are the students with numbers 5,6,12.

- If the start of the random draw is 15, the sample units are students with numbers 15,7,14.
- The selection process can be made considering that the randomly selected number is (5), which is the student number in the students' order sequence, then the corresponding student in the random order is the first student in the sample, that is the student with number (3). Then the second student is number (9) and the third student is number (5). Thus, the sample shall be the students with numbers 3,9,5. It is clear that many random samples can be withdrawn in this way. Therefore, sometimes for ease, the random start is considered the beginning of random order, and the sample units in this example are the students with numbers 14,8,11.
- Clearly, the probability of withdrawing one unit in this way equals $\frac{n}{N}$

Student N.	RAND	Random order	
1	0.669	14	0.191
2	0.829	8	0.2
3	0.367	11	0.309
4	0.752	10	0.352
5	0.556	3	0.367
6	0.577	9	0.404
7	0.891	5	0.556
8	0.2	6	0.577
9	0.404	12	0.658
10	0.352	1	0.669
11	0.309	13	0.713
12	0.658	4	0.752
13	0.713	2	0.829
14	0.191	15	0.854
15	0.854	7	0.891

4. (KISH) method to select a random individual among family members

In many household surveys, polls, marketing surveys, etc., a random individual is usually to be selected from among the family members qualified for the survey. The table prepared by Kish in 1965 is adopted in this process. It is called "Kish Table" and consists of eight tables. For example, a random individual is to be selected from among a group of qualified family members, of (18) years and older, as follows:

- Qualified male family members of (18) years and older are to be recorded, from the oldest to the youngest, followed by registering the eligible females from the oldest to the youngest as well, and the individuals shall be numbered sequentially.
- A random number between 1 and 8 is to be selected, so that the corresponding number in the table in the first column is the random table number on which the number of the eligible individual will be determined according to the number of family members.
- If Table 4 is selected and the number of eligible family members is 3, individual number (2) in the individual series shall be the individual required to be interviewed.
- Upon determining the random table number of the first family, the random table for the second family shall be the following table, which is the Table (5). In this example, qualified individuals should be registered in the second family and the individual qualified is to be selected as per the previous step.
- Thus, the random table for each family in the sample is determined by taking the table that follows the table corresponding to the preceding family, sequentially, until all sample households are covered. It is noted that the maximum number of the eligible

family members is (6). However, in practice, the number of the qualified family members can be more than six, and then another method of random withdrawal should be used.

The probability of individual withdrawal is (1) divided by the number of eligible family members.

KISH TABLE

Table No.		Number of Eligible Members					
		1	2	3	4	5	+6
		No. of Eligible Members					
A	1	1	1	1	1	1	1
B1	2	1	1	1	1	2	2
B2	3	1	1	1	2	2	2
C	4	1	1	2	2	3	3
D	5	1	2	2	3	4	4
E1	6	1	2	3	3	3	5
E2	7	1	2	3	4	5	5
F	8	1	2	3	4	5	6

5. Random selection of a number between two numbers

Population units are numbered from 1 to N and a number of random numbers equals to the size of the sample is obtained using the formula $= randbetween(1, N)$

If the size of the population is 50 households and the required sample size is 5 households, we can draw many random samples, each size is 5 households, using the above formula. The sample of households with numbers 49,3,10,13,24 is one of these samples. A distinction should be made between drawing with replacement and without replacement, where the drawn number cannot be repeated. If the number is repeated in the draw without replacement, a random number shall be taken without repeating the same number. The probability of household withdrawal in this example is 1/10 (5/50).

6. Random Digit Dialing Method

Mobile or landline phone subscribers' lists form a framework for random selection of respondents. Currently, Random Digit Dialing is widely used, especially in the landline or mobile phone surveys. This method can be summarized in the selection of random numbers randomly, making phone calls to the landline or mobile number corresponding to the randomly selected number. The probability of withdrawing one or several units shall be counted as in the above methods.

In practical applications, the process of withdrawing units may be carried out in several stages. One of the previous random draw methods can be adopted at any stage, and the probability

of withdrawal at each stage is conditional on the probability of withdrawal in the previous stage. For example, if a count unit is drawn in the first stage in a two-stage sample with a probability that equals P_1, the probability of a household being withdrawn from this unit in the second stage is conditional on the probability of withdrawing that unit in the first stage. The final probability to draw household P equals $P = p_1 * p_2$, where P_2 is the probability of drawing the household in the second conditional stage by drawing the count unit in the first stage.

3.2 Basic (Absolute) Weight

The primary purpose of using the sampling method is to obtain estimates of variables at the population level, based on the sample results. However, the estimates calculated from the sample results directly differ from those in the population, for several reasons:

1. Withdrawing sample units with unequal probabilities, which is almost the general case; as the probability of withdrawing equal sample units is rare.
2. Non-response of some sample units, which occurs in most cases upon failure in obtaining the required data from some sample units, because of rejection, the unit was not present at the time of survey or because of the existence of units alien to the population targeted in the study, etc.
3. Non-coverage of some population parts or units because of an old framework or for some force majeure due to natural disasters, or because some parts of population are excluded for reasons of cost or difficult access to such parts (desert areas, rugged

mountains, etc.) The final results shall be adjusted so that the basic distributions (total population, age, gender) would correspond to those of the population.

Weights are used to adjust the sample results to eliminate or reduce the impact of these factors on the differences between the sample estimates and the true values of such estimates in the target population, noting that sample estimates will be different from real values due to other types of errors, which are classified into two types of errors:

- Non-sampling Errors: known as bias errors; e.g. errors relating to frames, measurement, response, data manipulation, etc. Practical experience suggests that this type of error can be reduced rather than eliminated because of its sources and inability to control them fully.
- Sampling Errors: error resulting from limiting to one sample of the target population and not the entire population. Such errors include random errors, which are inherent to any random withdrawal and can be calculated for each type of random sampling, reducing this type of error. Both types of errors will be presented in guidelines dedicated to samples' error.

In the following paragraphs, the weights of the various survey samples and the modifications of sample results in these weights will be presented to compensate for the unequal probabilities, non-response, non-coverage, illustrative examples of multiple cases and the use of absolute weights and relative weights of different estimates.

The sample represents the target population from which the sample units were withdrawn, where each sample unit represents a number of units of the target population, known as the

unit weight. If the target population size is 20 units (individuals, households, facilities, areas, etc.) and a sample of 4 units was withdrawn, each unit in the sample represents 5 units (20/4 = 5) of the target population units. The number (5) represents the weight of each sample unit. By probability language, weight is expressed as being the reversed probability of unit withdrawal. The probability of drawing any unit from a 4-unit sample, withdrawn from a 20-unit population, as is known, equals to 1/5 (4/20 = 1/5). The reversed probability in this case equals (5), which represents the weight of each unit of the sample. To obtain the results at the level of the target population, the results of the sample are to be multiplied by the weight of each unit. This process is described as weighing the sample results or increasing sample results at the target population level. The sum of the weights after increasing sample results should be equal to the size of the target population. This is one of the rules that must always be achieved.

Generally, using mathematical formulas, if P is the probability of withdrawing the unit, the reversed probability is called the **basic weight**, symbolized by W_b and equals

$$W_b = \frac{1}{P}$$

The weights are increasing factors of unit values in the sample; for this reason, they are sometimes called Raising Factors. The probability of withdrawing the sample units may be equal and the unit weights shall then be equal and the sample becomes known as a self-

weighted sample. The characteristics of this type of samples include that the estimates (the proportions and averages) at the sample level are identical before and after weighting the results, because the weight will be multiplied by both the denominator and the numerator for both the ratio and the average and shall, thus, be eliminated. There is no need to weigh the results of the sample; that is why this type of sampling is preferred when the objective is to obtain population-specific estimates; such as the proportion or average. When it is desirable to obtain final totals, weights should be used.

Based on what has been presented in the previous paragraph, if a sample of n size is withdrawn from a population sized N, the probability of drawing any unit is equal to sampling fraction $f = \frac{n}{N}$ and the basic weight of the unit in the sample is equal to the reversed probability of drawing the unit or the reversed sampling fraction. The units of the sample may be withdrawn using the Systematic Selection method. The weight of the unit is equal to the withdrawal period

$I = \frac{N}{n}$. In general, the basic weight of the unit is calculated by one of the following forms of formula $W_b = \frac{1}{P} = \frac{1}{\frac{n}{N}} = \frac{N}{n} = \frac{1}{f} = I$

For example, in order to estimate the average monthly household income among a group of households of 20 units, a sample of 5 households was withdrawn by one random sampling method. The monthly household income in the sample was as shown in the following table.

Household	Household Income	Household Weight	Increasing (Weighting) the Sample	
			Income	Households Number
1	85000	4	340000	4
2	60000	4	240000	4
3	10000	4	40000	4
4	120000	4	480000	4
5	50000	4	200000	4
Total	325000		1300000	20
Average	65000		65000	

It is clear that the draw period equals $I = \frac{N}{n} = \frac{20}{5} = 4$ and the probability of household withdrawal equals $p = \frac{n}{N} = \frac{5}{20} = \frac{1}{4}$, which is also the sampling fraction. The basic weight of each household in the sample equals the draw period, the reversed probability of household withdrawal or the reversed sampling fraction. In all cases, the basic weight of each household is fixed and equal to 4.The sample is self-weighted. Sample results are being increased by multiplying the weight of each household by its income to obtain the total household income in the target population (20 households), the weight of each household is multiplied by (1) to

obtain the number of households in the population, as shown in the last two columns of the table. This table shows the following:

- Average household income in the target population after weighting the sample results is 65,000 monetary units, which is the same as the average household income in the sample before weighting. This is because the sample is self-weighted, as mentioned earlier.
- However, total income of the target population is estimated at 1,300,000 monetary units, which is different from the total sample population of 325,000 monetary units. For this reason, the sample results should be increased to obtain the final totals. Total income of the target population in this example can be obtained by multiplying the size of the population by the average of the sample (20*65000), since the weights of the sample households are equal and the sample is self-weighted. This will not work if the sample weights are different. If the weights of the sample households (by default) are as in the following table, the average monthly income of the household in the population is 67,000 monetary units after weighting the sample results, while it is equal to 65000 in the sample.

Household S//N	Household Income	Household Weight	Increasing (Weighting) the Sample	
			Income	Households Number
1	85000	2	170000	2
2	60000	4	240000	4
3	10000	3	30000	3
4	120000	5	600000	5
5	50000	6	300000	6
Total	325000		1340000	20
Average	65000		67000	

The final (target) unit may be withdrawn on several stages, as in the multi-stage sample, upon which each stage will then have its own weight. The weight of the final unit is the result of multiplying the weights by all stages. If W_1 is the weight of withdrawing the unit in the first stage, which may be a count unit, for example, and W_2 is the unit weight of household withdrawal in the second stage from the count unit households that was drawn in the first stage, and W_3 was the withdrawal weight of one household member (eligible for the survey) withdrawn in the second stage, them the basic weight of the final unit (the individual in this example) W_b shall be given by the following formula $W_b = W_1 * W_2 * W_3$

The sample may be self-weighted at any stage of the sample withdrawal and that self-weight may not apply to the next stage. The sample may be self-weighted when a household sample is withdrawn in the first stage. If a random individual is withdrawn from the household that was withdrawn in the first stage, the sample becomes not self-weighted for that random individual, where the weight of the individual will vary from one household to another because of the variance number of individuals among households. In practice, the sample is rarely self-weighted because some sample units are not responsive.

3.3 Compensation for the Non-Response

In practical applications, basic weights are rarely fixed because some sample units do not respond for a variety of reasons (rejection, absence, unqualified units, closed units, etc.). The core weights should be adjusted to compensate for the non-responsiveness, so that the unit weight shall be modified to match with the responsive units by calculating the response rate. To calculate the response rate, the sample units should be classified into the following types:

1. **Responsive units**: Eligible units that have given complete answers (in whole or in part); symbolized by H_r . Eligibility means that they belong to the target population and are not alien thereto.
2. **Unresponsive units**: Eligible units from which the required data was not obtained for various reasons; such as rejection, absence, illness, etc., symbolized by (H_{nr}).

3. **Ineligible units**: Fieldwork indicates that some of the units n' are alien and do not belong to the target population, hence they are not eligible. This category includes vacant dwellings and destroyed houses.

4. **Units unknown whether they are eligible or not (n_{un})**: As in the case of closed units or for unavailability at survey date. After completion of the fieldwork, some or all of these cases are confirmed. This relates to the modernity of the frames used in the withdrawal of the sample units. In practice, the following case may be addressed:

- Non-existence of ineligible or unknown units. Then the sample units shall then be categorized into respondent units and non-responding units. Response rate shall be calculated as follows: $R = \frac{H_r}{H_r + H_{nr}}$

If the sample size is 200 households and the number of non-responding households is 20, while the number of responding households was 180 households, the response rate then equals $R = \frac{H_r}{H_r + H_{nr}} = \frac{180}{180 + 20} = 0.90$

Since the basic weight of the unit $\frac{N}{n}$ is calculated based on the size of the basic sample and not all units are responsive, but only 90% thereof, the size of the basic sample should be adjusted to match the size of the responsive sample. Thus, the new sample size equals the size

of the basic sample multiplied by the response rate; i.e. the size of the responsive sample equals $n_r = 0.90 * n$

The actual weight per unit becomes $W = \frac{N}{n_r} = \frac{N}{0.90 * n} = \frac{N}{n} * \frac{1}{0.90}$

It is clear that the new weight per unit is equal to the basic weight multiplied by the reversed response rate, which is the general rule in case of non-responsiveness. If the reversed response rate is W_{rr}, then $W_{rr} = \frac{1}{R}$

The new weight becomes $W = W_b * W_{rr}$

If the unit base weight is 10 in this example, the new weight equals $W = 10 * \frac{1}{0.90} = 1.11$

- The existence of ineligible cases: the actual sample size equals $n_r = n - n'$ and the response rate $R = \frac{H_r}{n - n'}$

If the sample size is 250 and is classified after the fieldwork as in the table below, the response rate is $R = \frac{H_r}{n - n'} = \frac{200}{250 - 20} = 0.8696$

If the unit weight is 10, the new weight equals $W = 10 * \frac{1}{0.8696} = 1.15$

Responsive Households	200
Non-Responsive Households	30
Ineligible Households	20
Total	250

- Existence of cases of unknown eligibility, in which case the number of eligible cases should be estimated therefrom, by estimating the percentage of eligible units ε ,

  which is equal to: $$\varepsilon = \frac{H_r + H_{nr}}{H_r + H_{nr} + n'}$$

The number of cases expected to be eligible in this type of unit is equal to $H_{el} = \varepsilon * n_{un}$

After these cases are categorized, the actual sample size becomes

$$n_r = H_r + H_{nr} + H_{el}$$

The response rate is calculated by formula $$R = \frac{H_r}{H_r + H_{nr} + H_{el}}$$

To compensate for non-responsiveness, the baseline weight is multiplied by the reversed response rate $$W_{rr} = \frac{1}{R} = \frac{1}{\frac{H_r}{H_r + H_{nr} + H_{el}}} = \frac{H_r + H_{nr} + H_{el}}{H_r}$$

The new weight becomes $W = W_b * W_{rr}$

For example, a sample of 300 households was withdrawn and after the fieldwork was completed, the distribution of households visited was as follows:

Responsive Households	200
Ineligible Non-Responsive Households	30
Alien Households and vacant or destroyed houses	25
Households which eligibility is unknown	45
Total	300

The percentage of eligible households among households whose eligibility is unknown becomes: $\varepsilon = \frac{H_r + H_{nr}}{H_r + H_{nr} + n'} = \frac{200 + 30}{200 + 30 + 25} = 0.902$

The number of households expected to be eligible among households whose eligibility is unknown $H_{el} = \varepsilon * n_{un} = 0.902 * 45 = 41$

Response rate equals $R = \frac{H_r}{H_r + H_{nr} + H_{el}} = \frac{200}{200 + 30 + 41} = 0.738$

The basic weight should be adjusted by multiplying it by the reversed response ratio. If the base weight is 10, the adjusted weight equals

$$W = W_b * W_{rr} = 10 * \frac{1}{0.738} = 13.55$$

3.4 Compensation for Non-Coverage

When modern frameworks are used to represent the size of the target population or the size of its constituent classes, the size of the population after the weighting process will be commensurate with the size of the actual population. In practice, relatively old frameworks that do not reflect the actual reality of the changes that occurred in the size of the target population shall be used, resulting in unidentical final totals after the results are weighted with the actual totals at the date of survey. Therefore, the latest estimates of the total population allocations are used, so the results of the sample after the weighting process shall match those distributions. If the size of the basic population is N and the size of the population estimated at the survey date is (N'), the basic weight should be multiplied by the ratio ($\frac{N'}{N}$) and given the symbol W_c , so $W_c = \frac{N'}{N}$

The final weight W is $W = W_b * W_{rr} * W_c$

The results should be adjusted by multiplying the sample results by the final weight, so that the required unbiased estimates can be obtained (the differences between them and the real values are due to random errors only).

For example, a sample size of 100 was randomly drawn from a 1,500-size population. It is clear that the probability of sample units withdrawal is $P = \frac{n}{N} = \frac{100}{1500} = \frac{1}{15}$

The basic weight per unit is $W_b = \frac{1}{P} = \frac{1}{\frac{1}{15}} = 15$

This means that each sample unit represents 15 units of the population.

Assuming that there are 10 non-responsive units in the sample and assuming that there were no alien units $n' = 0$ and no unknown units $n_{un} = 0$, then the response rate would equal

$$R = \frac{H_r}{H_r + H_{nr}} = \frac{90}{90+10} = \frac{90}{100} = \frac{9}{10}$$

The weight of compensation for the non-responsiveness W_{rr}, which is equal to reversed response ratio: $W_{rr} = \frac{1}{R} = \frac{1}{\frac{9}{10}} = \frac{10}{9}$

The new weight (W') shall be calculated by multiplying the former two weights

$$W' = W_b * W_{rr} = 15 * \frac{10}{9} = 16.67$$

Each unit in the sample represents 16.67 in the population instead of 15. It should be noted that the size of the population remains fixed after the weighting process. In case of restricting to the base weight (in this case there is no non-responsiveness), the output of multiplying the sample by the basic weight equals $15 * 100 = 1500$, which is the size of the population.

If the frame of the targeted population is modern (which is rare) and non-responsiveness existed, the results of the sample are increased by weight W' and the size of the population

resulting from the new weight multiplied by the number of responding units (90) is matching to the size of population as well $15 * \frac{10}{9} * 90 = 1500$

If the sample frame is relatively old and assuming that the size of the population according to the latest estimates is 1700, the weight of compensation for the non-coverage W_c equals

$$W_c = \frac{N'}{N} = \frac{1700}{1500}$$

The final weight W is

$$W = W_b * W_{rr} * W_c = 15 * \frac{10}{9} * \frac{1700}{1500} = 18.89$$

The output of multiplying the final weight by the size of the responsive sample is 1700, which the size of the population according to the latest estimates.

When the population is divided into strata, as in the case of urban and rural areas for example, the different weights in each stratum should be calculated and the results of the sample in each stratum is to be multiplied by the weights of that stratum, so that the size of each layer after the weighting process shall be equal to its basic size before the weighting process according to the latest estimates of the size of each stratum, as above described, or the relative composition of the strata sizes after the weighting process should be identical to the relative composition of these strata in the target population. Otherwise, the estimates at the total level of the target proportion will be biased to any of the stratum and unacceptable.

For example, the following table shows a two-tier population. A sample was drawn from each stratum to estimate the proportion of males and females in each stratum and in the target population. The results of the sample were weighted in each stratum as above indicated and the final results for both males and females in each stratum, after weighing the results of both samples, were as shown in the table. Based on the latest population estimates of the size of each stratum, the size of the first stratum is 10400 individuals and the size of the second stratum is 2600 individuals. The size of each stratum after weighting the results is clearly different from its size in the target population and its total is different from the size of the targeted population. Also, the relative composition of both strata differs after the weighting process from their relative composition in the target population, because their relative composition is 70% and 30% respectively after weighting, which is different from their

relative composition in the target population; 80% and 20%, respectively. This is due to the following reasons:

- Non-coverage
- Different weights of sample units in the stratum or between strata; i.e. the sample is not self-weighted.

This is contrary to the basic principle of the outcome of the sample results' weighting process, where the strata sizes after the weighting should be equal to their sizes after the weighting process. Based on that principle, the weights should be adjusted so that the size of each strata shall be is equal to its size in the target population, or the relative composition of both strata

shall be compatible with the true proportional composition of the target population. This could be done by two methods:

Population			Weighted Sample					
Stratum	N_i	p	N_i'	p	Males	p	Females	P
1	10400	0.8	7700	0.7	4235	0.55	3465	0.45
2	2600	0.2	3300	0.3	1683	0.51	1617	0.49
Total	13000	1	11000	1	5918	0.538	5082	0.462

Method 1:

Adjusting the results so that the total of both strata after weighting are equal to their totals in the target population by compensating for the non-coverage as previously presented. Their relative composition will therefore be similar to their relative composition in the target population. This can be done by multiplying the weights in each stratum by the ratio between true stratum size and its size after weighing, as shown in the following table:

Population		**Weights Adjustment**							
Stratum	N_i	N_i'	N_i / N_i'	$N_i' * N_i / N_i'$	**p**	**Males**	**p**	**Females**	**p**
1	10400	7700	1.351	10400	0.8	5720	0.55	4680	0.45
2	2600	3300	0.788	2600	0.2	1326	0.51	1274	0.49
Total	13000	11000	1.182	13000	1	7046	0.542	5954	0.458

It is clear from the table that the size of both strata after adjusting the weights has become equal to the total of the target population and their size is equal to their size in the target population, as well as their relative composition. The ratio of males and females differed at the level of the target population. They are slight differences in this example. Other examples may be resented where such differences are significant.

It should be noted that weights adjustment has not changed the proportion of males and females at the level of each stratum. This is because the numerator and denominator in the

equation of the proportion will be multiplied by the same adjustment factor, leading to its deletion. This result can be expressed by the fact that the characteristics of each stratum (proportion and averages) do not change with these modifications, which is always a true principle. This should be verified upon conducting this type of adjustment.

Method 2:

In practical applications, the aim of the survey may be to estimate the characteristics (proportion and averages) only without total sums, as in many demographic surveys, Polls, etc. In practice, recent estimates may not be available for the size of each stratum. In both cases, another method may be adopted to modify the weights, so that the relative composition of the strata after weighting shall correspond to the relative composition of the strata in the target population, according to the relative estimates previously known to both strata in the target population (e.g. from the last census). This can be reached by multiplying the weights by a factor equal to the true proportion of the stratum and its a proportion after weighting,

as in the table below. Accordingly, the relative composition of both strata has become identical to their relative composition in the target population, with no change in their characteristics (male and female proportions). The proportion of males and females in the target population is equal to the proportion using Method 1.

Percentage between the stratum's percentage P	N'	$N' * P$	p	Males	p	Females	p
1.142857	7700	8800	0.8	4840	0.55	3960	0.45
0.666667	3300	2200	0.2	1122	0.51	1078	0.49
Total	11000	11000	1	5962	0.542	5038	0.458

The same results can be reached at the total level, depending on the known principle of calculating the average or the proportion at the sum of several strata. If the proportion P_i is in strata i and N_i is the size of strata i, proportion P at the total level is calculated by the following formula, depending on the size of each strata. $P = \frac{\sum N_i * P_i}{\sum N_i}$

The proportion can be calculated at the total level depending on the proportion of the size of each strata of the total stratum size $W_i = \frac{N_i}{\sum N_i}$, the above formula can be written as follows

$$P = \sum W_i * P_i$$

Two similar formulas can be written to calculate the total average.

Applying on this example, the proportion of males at the total level equals 0.542, which is the same ratio calculated by the above two methods, $P = 0.8 * 0.55 + 0.2 * 0.51 = 0.542$

Females s proportion hall be obtained by applying the same formula.

From the above, the following two important results are reached:

Result (1): Weights should be adjusted so that strata sizes or relative composition in the stratified sample shall be similar to the strata in the target population.

Result (2): If comprehensive and modern frames are adopted to withdraw sample units, sample results can be used to modify some of the characteristics of the population that relied on estimates based on the results of an old census and have undergone significant changes, so that the relative composition of the target groups in the estimated totals is adjusted to match the relative composition of these categories in the weighted sample results, in a manner opposite to the above Method 2.

3.5 compensation for incomplete responses (Imputation), and it means the compensation for some unfinished items due to refusal to answer by some sample units, hence these items can be compensated for with similar data for units similar to these units, in terms of geographical location, social situation or economic activity and other appropriate standards, or by using regression method to compensate for these items.

3.6 Trimming of Weights

After completing the adjustment weights, some units with excessive weights need to be scrutinized, as this will significantly increase the variance of the estimates thus inflating miscalculation, and can control the overall estimates. Typically these kinds of weights are reduced for a limited number not to exceed a certain limit, and so that the total weights do not change, so if the total weights before reduction was N and total weights after reduction was, N' hence the weights of all sample units including new reduced weights are to be multiplied by the rate $\frac{N}{N'}$. It should be noted that this procedure helps to reduce variation and thus reduce the estimates errors, but it increases the bias in some estimates, so be careful in lowering the weights and to resort to this procedure only in special and limited cases.

3.7 Relative Weights

In some cases, for ease, in calculating some indicators; such as the average or ratio, the relative weights are used instead of the absolute weights, so that the total weights of the sample equals (1) and the relative weight of one of the elements shall be equal to the result of multiplying the absolute weight of that element by the reversed average of all weights. If Wi is the absolute weight of element i and Wir is the relative weight thereto, the relative weight shall be calculated by the following formula

$$W_{ir} = W_i * \frac{1}{\frac{\sum W_i}{n}} = W_i * \frac{n}{\sum W_i}$$

The average relative weights should be equal to (1), as indicated in the following example, where the first column represents the sample households' numbers and the second column represents the basic weights of a sample size of 5 households. Relative weights are calculated using the following two steps.

- **Step 1**: Calculating the average basic (absolute) weights, which equals 80
- **Step 2**: Calculate the relative weight of each unit by multiplying the base weight by the reversed average basic weights, as indicated in the fourth column.

Household Number	Weight	Household Income	wi*1/∑wi/n	Wi*I	(wi*1/∑wi/n)*I
1	120	10	1.5	1200	15
2	75	20	0.9375	1500	18.75
3	15	35	0.1875	525	6.5625
4	90	50	1.125	4500	56.25
5	100	70	1.25	7000	87.5
Total	400		5	14725	184.0625
Average	80			36.8125	36.8125

The results of the sample are weighted by the new relative weights, only to calculate the averages and percentages, which will not differ if the results are weighed with absolute weights. This is because the value of the indicator in the sample remains the same before and

after the weighting process because both the numerator and denominator will be multiplied by the same weight. If the results are increased by the basic weights as in the fifth column and the results were increased by relative weights as in the sixth column, the average household income equals 36,8125 in both cases. When the aim is to obtain the total sums of the population, the basic weights should be used, so that the total income of the population in this example is 14725, using the basic weights or by multiplying the average household income by the total weights (400*36,8125). This differs from the total income using the relative weights, which equal 184,0625.

4. Sampling Errors

Sample data are subjected to two types of errors. Sampling errors and non-sampling errors. It is well known that non-sampling errors are common errors in the census, as well as when using the sampling method. These are many and varied errors, such as non-coverage errors, non-response errors, response errors and other known errors in surveys and censuses. sampling errors are errors concerning the samples, there are no such errors in the census. These errors result from studying part of the population and not the whole population. These errors are sometimes called, random errors or incidental errors. The importance of calculating errors in various estimates lies in being helpful in identifying areas where these estimates and desired confidence levels occur and making numerous tests on the sample results. This type of errors can be calculated for all estimates covering a large number of indicators. In practical applications, there are usually main estimates related errors such as overall totals, averages and proportions. The margin of error in the estimate is calculated, as explained earlier, according to The Central Limit Theorem, based on varied estimates and the desired levels of confidence. There are several publications concerning sampling techniques, methods of calculating errors in the samples based on the variance of the estimates (where the error in the estimate equals the square root of the variance), as in the work of Cochrane, Kish, Hansen, and others as well as the United Nations publications in this field. The following paragraphs will display some calculations relations the margin of error for the main estimates in each type of samples.

4.1 Errors of Simple Random Sampling (SRS)

In general, the Central Limit Theory (CLT) is used to calculate the standard error and the difference between the real value (average, proportion, total) and its estimated value is expressed by the sample, so it may be positive or negative, usually presented by the symbol (SE) and is sometimes called random error or error of chance. The simple random sample is a direct application of (CLT).

If a simple random sample is drawn from a population of size (N) and $X_1, X_2, \ldots\ldots\ldots, X_N$ are the values of the phenomenon studied in this population, then the average of this phenomenon is calculated by the following formula:

$$\overline{X} = \frac{\sum_{i=1}^{i=N} X_i}{N}$$

Its variance is calculated by the following known formula:

$$V = \sigma^2 = \frac{\sum_{i=1}^{i=N} \left(X_i - \overline{X}\right)^2}{N}$$

In order to calculate the various estimates, the difference is usually distinguished by both sample units sampling methods: Withdrawal with replacement, where the withdrawn unit is returned to the population before the next draw, and the withdrawal without replacement, where the drawn unit is not to be returned to the population in the next draw.According to the Central Limit Theory (CLT), the estimate calculated by the sample is distributed according

to Normal Distribution. The difference between the real value of the estimate in the target population and the estimate calculated by the sample, called Standard Error, equals the standard deviation of the estimation variance, which can be positive or negative. The variance of the estimate is indicated by the symbol ($V_{(\bar{x})}$) or the symbol ($\sigma_{(\bar{x})}^{2}$) for the average variance, and for proportion variance the symbols $V_{(p)}$ or $\sigma_{(p)}^{2}$, and $V_{(T)}$ or $\sigma_{(T)}^{2}$ for total variance. Therefore, the standard error in the estimation of the average is expressed in the following and similar formulas for the standard error of the proportion and the total:

$$SE = \pm\sigma_{(\bar{x})} = \pm\sqrt{V_{(\bar{x})}}$$

From that last formula it can be noticed that the standard error square is equal to the estimate variance. Since the different estimates are distributed according to Normal Distribution, the variance of the different estimates can be calculated according to the Central Limit Theory (CLT). as follows in the following paragraphs.

4.1.1 Random error square for sample average

- Average variance (with replacement)

$$V_{(\bar{x})} = \sigma^{2}{}_{(\bar{x})} = \frac{\sigma^{2}}{n}$$

- Average variance (without replacement)

$$V_{(\bar{x})} = \sigma^2{}_{(\bar{x})} = \frac{\sigma^2}{n} * \left(\frac{N-n}{N-1} \right)$$

When the population size is large, the difference between value ($N-1$) and value (N) becomes very small. By replacing that in the variance formula without replacement, it becomes: $\sigma^2{}_{(\bar{x})} = \frac{\sigma^2}{n} * \frac{N-n}{N} = \frac{\sigma^2}{n} * \left(1 - \frac{n}{N}\right) = \frac{\sigma^2}{n} * (1-f)$

Where ($f = \frac{n}{N}$) and is called Sampling Fraction; ($1-f$) is Population Correction Factor (PCF).

It is noticeable that the variance without replacement is smaller than the variance with replacement. That is why it is preferable to use the random sample without replacement in the practical applications because its errors are less than in random sample errors with replacement. When the sampling fraction is smaller than 5% and is achieved in larger populations, then ($1-f=1$) can be adopted and, hence, the variance of the estimate shall be the same in both replacement and without replacement cases.

4.1.2 Random error square of the proportion in the sample

The above applies to an obvious proportion because the proportion is a special case of the average, since any value in the population is either (1) when the phenomenon of interest is consistent with the pheomenon, or (zero) if otherwise. Proportion (P) is calculated by the

previous average formula, so that the numerator equals the number of units that corresponds to the phenomenon and the denominator equals the total number of the units of population (consistent and inconsistent with the phenomenon). The complementary proportion (Q) then is equal to ($Q = 1 - P$) and the variance of the proportion is according to (Bernoulli) Constitution

$$V = \sigma^2 = P * Q$$

The estimated proportion variance shall then be calculated similar to the estimated average variance, by one of the following two formulas:

- Proportion variance (with replacement)

$$V_{(p)} = \sigma^2{}_{(p)} = \frac{V}{n} = \frac{P * Q}{n}$$

- proportion variance (without replacement)

$$V_{(p)} = \sigma^2{}_{(p)} = \frac{P * Q}{n} * \left(\frac{N - n}{N - 1} \right)$$

When the population size is large, the difference between value ($N - 1$) and value (N) becomes very small. By replacing that in the variance formula without replacement, it becomes:

$$V_{(p)} = \sigma^2{}_{(p)} = \frac{P * Q}{n} * (1 - f)$$

4.1.3 Random error square for the total

The total population (T) is estimated based on the sample average ($\bar{x}$) in the following formula:

$$T = N * \bar{x}$$

Depending on the variance properties (if any of the values is multiplied by a fixed number, the variance of the new value is equal to the variance of the basic value multiplied by the square of this number). The following formula can be written to the estimated total variance in terms of the variance of the average calculated by previous formulas, which expresses the random error square of the total population

$$V(T) = V(N * \bar{x}) = N^2 * V(\bar{x})$$

The following similar formula can be written for a (T') total variation of one of the phenomena of interest, which proportion (p) as estimated in the sample

$$V(T') = V(N * p) = N^2 * V(p)$$

Example: In order to estimate the average daily wage of a worker in a company with 10,000 workers, a simple random sample of 50 workers was withdrawn and their daily wages were as follows:

104	129	146	121	123
122	131	147	150	129
121	145	127	149	147
102	122	138	138	142
150	145	131	111	132
132	123	103	140	126
135	114	130	129	139
122	106	146	102	102
145	115	108	124	119
134	115	113	107	101

In applying the previous relations, the average wage of worker in this company will be estimated and the error made in estimating the average by the sample, noting that the variance worker wage in that company, as per previous studies, equals 450.

Average worker wage in the sample equals $\overline{x} = \frac{\sum_{i=1}^{i=50} x_i}{n} = 126.64$

To calculate the error in estimating the average, the above formulas shall be applied to calculate the average with and without replacement, as follows:

Average variance with replacement by applying formula (1.1):

$$V_{(\bar{x})} = \sigma^2{}_{(\bar{x})} = \frac{\sigma^2}{n} = \frac{450}{50} = 9$$

Then the standard error committed by estimating the average (with replacement) is equal to:

$$SE = \pm\sigma_{(\bar{x})} = \sqrt{9} = 3$$

The average variance without replacement with applying formula (1.2):

$$V_{(\bar{x})} = \sigma^2{}_{(\bar{x})} = \frac{\sigma^2}{n} * \left(\frac{N-n}{N-1}\right) = \frac{450}{50} * \left(\frac{10000-50}{10000-1}\right) = 8.96$$

The standard error committed by average estimate (without replacement) is 2.99, which is smaller than the standard error with replacement, as above indicated, yet the difference between them is very small due to the large size of the population and the small sampling fraction ($f = \frac{n}{N} = \frac{50}{10000} = 0.005$) (as above indicated).

$$SE = \pm\sigma_{(\bar{x})} = \pm\sqrt{8.96} = 2.99$$

4.2 Stratified Sampling Errors

Different estimation errors in each stratum shall be calculated in the same manner as the simple random sampling errors presented in the previous paragraph, given that each stratum is a stand-alone population. The errors of the different estimates are calculated at the level of the target population, using variance characteristics for both the average and proportion at the total level, in terms of variance at the level of the strata as shown in the following paragraph:

4.2.1 Error in estimating the average

It is known that the average sample in the population, symbolized by ($\overline{X}_{(st)}$), is a weighted average of the samples' averages drawn from the strata, calculated with the following formula:

$$\overline{x}_{(st)} = \frac{\sum_{1}^{l} N_i * \overline{x}_i}{\sum_{1}^{l} N_i} = \frac{\sum_{1}^{l} N_i * \overline{x}_i}{N} = \sum_{1}^{l} W_i * \overline{x}_i$$

Where $i = 1,2,.....,l$ and $W_i = \frac{N_i}{N}$

$\overline{x}_i$ is the estimated average in stratum i

The estimated population average variance is then calculated by the following formula:

$$\sigma^2_{(\overline{x}_{st})} = \sigma^2\left(\sum W_i * \overline{x}_i\right)$$

Depending on the variance properties (if a number of values are multiplied by a fixed number, the variance of the new values is equal to the variance of the basic values multiplied by the square of the fixed number). The formula can be written as follows:

$$\sigma^2(\bar{x}_{st}) = \sum W_i^2 * \sigma^2(\bar{x}_i)$$

Based on the preceding paragraph, $V_{(\bar{x}_i)}$ is the variance of the estimated average in the stratum and is calculated when without replacement by the above formula

$$\sigma^2(\bar{x}_i) = \frac{\sigma_i^2}{n_i} * (1 - f_i)$$

Based on the above, the average variance is calculated in terms of the average variance in the strata by the following formula

$$\sigma^2(\bar{x}_{st}) = \sum W_i^2 * \sigma^2(\bar{x}_i) = \sum W_i^2 * \frac{\sigma_i^2}{n_i} * (1 - f_i)$$

This formula represents the random error square of the estimated average population, which is commensurate with the variance in the strata; the greater the strata variance is, the greater the random error square becomes, and the smaller the strata variance is, the smaller the random error square becomes. This leads to the following important result which is the basis for the use of the stratified sampling method, which can be expressed as follows: In order to reduce the random error in using the stratified sampling method, the population should be divided into homogeneous strata, so as to reduce the variation in the strata.

Example: the following table shows sample size and variance in two strata

Stratum	N_i	n_i	σ_i^2	$\bar{x}_i$	P_i
1	70	10	200	15	0.20
2	30	4	100	22	0.35

Random error square is 10.35 according to the previous formula

$$\sigma_{(\bar{x}_{st})}^2 = \left(\frac{70}{100}\right)^2 * \frac{200}{10}\left(1-\frac{10}{70}\right) + \left(\frac{30}{100}\right)^2 * \frac{100}{4}\left(1-\frac{4}{30}\right) = 10.35$$

Random (standard) error equals $e = \sqrt{\sigma^2_{(\bar{x}_{st})}} = \sqrt{10.35} = 3.22$

Calculating the estimated average of the previous formula that equals

$$\bar{x}_{(st)} = \frac{\sum_1^l N_i * \bar{x}_i}{\sum_1^l N_i} = \frac{70*15+30*22}{100} = 17.1$$

Average population according to the confidence level at 95% falls in the scope

$$17.1-1.96*3.22 \le \bar{X} \le 17.1+1.96*3.22$$

$$10.8 \le \bar{X} \le 23.4$$

The same steps used to estimate the average can be used to estimate the proportion in population. Estimating the proportion in population gives the following formula:

$$P_{(st)} = \frac{\sum_{i=1}^{L} N_i P_i}{\sum_{i=1}^{L} N_i} = \sum W_i * P_i$$

Where P_i is the estimated percentage of the proportion in stratum (i) and, similarly, it can be proved that the variation of the estimated proportion is given in the following formula

(2.2) $$\sigma^2_{(P_{st})} = \sum W_i^2 * (1 - f_i) * \frac{P_i * Q_i}{n_i}$$

Where $f_i = \frac{n_i}{N_i}$ and $Q_i = 1 - P_i$

Reference to the previous example and applying the latter formula, the random error square of the estimated proportion of the population can be calculated as follows:

$$\sigma^2_{(P_{st})} = (\frac{70}{100})^2 * (1 - \frac{10}{70}) * \frac{0.20 * 0.80}{10} + (\frac{30}{100})^2 * (1 - \frac{4}{30}) * \frac{0.35 * 0.65}{4} = 0.0067 + 0.0044 = 0.0111$$

The random (standard) error then equals $e = \sqrt{\sigma^2_{(P_{st})}} = \sqrt{0.0111} = 0.105$

The estimated population proportion is calculated with the previous formula

$$P = \frac{70}{100} * 0.20 + \frac{30}{100} * 0.35 = 0.14 + 0.105 = 0.245$$

The upper and lower limits for estimating the population proportion can be determined according to confidence level at 95% in the following formula

$$0.245-1.96*e \le P \le 0.245+1.96*e$$

4.3 Cluster sample errors

In practical applications, clusters of equal/unequal size may be used. The study may include all cluster units or on a specified number of cluster units. In this paragraph, the random errors of the various estimates (average, proportion and total) shall be calculated for equal size clusters and the random errors of the different clusters shall be calculated in the subsequent paragraph.

4.3.1 Clusters of equal size

Assuming that a population consists of A cluster and the number of elements in each cluster equals B, where clusters a of this population has been selected, randomly and without replacement. Then it shall be $n = a*B$ and $N = A*B$. This sample can be considered as a simple random sample without replacement and its units are the clusters, which number is a. Since the clusters are equal in size, the average sample equals the average values in all clusters or equals the averages in all clusters, as in the following formula:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{a=1}^{a} \overline{x}_a}{a}$$

This is similar to the simple random sample, so that cluster averages can be considered as simple random sample units, which number is a. Applying the previous formula (1.2), the average variance (random error square) equals

$$\sigma^2_{(\bar{x})} = (1-f) * \frac{S_{(a)}^2}{a}$$

Where $f = \frac{a}{A}$ $s^2_{(a)} = \frac{\sum_{a=1}^{a} (\bar{x}_a - \bar{x})^2}{a-1`}$

By switching, the random error square is calculated with the following formula

$$\sigma^2_{(\bar{x})} = (1-f) * \frac{\sum_{1}^{a} (\bar{x}_a - \bar{x})^2}{a*(a-1)}$$

Example. The following table shows a sample of four clusters, and assuming that $f = \frac{a}{A}$ is small that it can be neglected.

	C1	C2	C3	C4	
	17	17	19	12	
	29	34	15	10	
	40	34	41	41	
	26	29	39	18	
	40	18	35	27	
	13	35	22	28	
	39	31	34	33	
	12	48	17	17	
	20	16	21	20	
	43	41	11	17	
Total	279	303	254	223	1059
$\bar{x}$	27.9	30.3	25.4	22.3	26.475
$(\bar{X}-\bar{x})^2$	2.030625	14.63063	1.155625	17.43063	35.2475

Assuming that $f=\frac{a}{A}$ is too small to be disregarded and applying the above formula, the variance of average shall be

$$\sigma^2_{(\bar{x})}=(1-f)*\frac{\sum_1^a(\bar{x}_a-\bar{x})^2}{a*(a-1)}=\frac{35.2475}{4*(4-1)}=2.937$$

The random error in estimating the average equals 1,7138, which is the square root of its variance. Random error is calculated to estimate the proportion with the following similar formula

$$\sigma^2_{(\bar{x})}=(1-f)*\frac{\sum_1^a(p_a-p)^2}{a*(a-1)}$$

If the values colored in (red) in the above example represent females, then female proportion is 45% (18/40)

	C1	C2	C3	C4	
	17	17	19	12	
	29	34	15	10	
	40	34	41	41	
	26	29	47	18	
	40	18	35	27	
	13	35	22	28	
	39	31	34	33	
	12	48	17	17	
	20	16	21	20	
	43	41	11	17	
pi	0.3	0.5	0.4	0.6	0.45
(pi-p)^2	0.0225	0.0025	0.0025	0.0225	0.050

The random error square for females is 0.00417 and the error is 0.065

$$\sigma^2_{(\bar{x})} = (1-f) * \frac{\sum_1^a (p_a - p)^2}{a*(a-1)} = \frac{0.05}{4*(4-1)} = 0.00417$$

4.4 Multistage (complex) Sampling errors

In practice, it is rare for a standard, simple, stratified or cluster sampling method to be used independently, as previously indicated. Rather, more than one method is being used, in the so-called complex samples, as in the multistage sample, where some or all of these methods are used simultaneously. Calculation of sampling errors in this case becomes more complex. Because of the complexity of the variance calculation in such cases, some methodologies and

programs for calculating this type of error have been developed by some international organizations and scientific research centers, where some are available free of charge, including**: CENVAR, CSPRO, SUDAAN** and others. One of the most widely used methods for estimating variance in complex samples is the Ultimate Cluster Method, where the focus is placed on the calculation of the variance in the final clusters created from the Primary Sampling Units in the multistage sample, regardless of the number of selection stages that were performed beforehand. The United Nations publication Designing Household Survey Samples: Practical Guidelines, 2008. F No.98 will be used to calculate the random error of different estimates in this method.

4.4.1 Standard Total Error

The total sum variance estimate is calculated at the level of each stratum to which the target population is divided. Assume that C is the number of clusters selected in a certain stratum and that X_c is the weighted total in cluster c and the various clusters' totals in this stratum is as follows:

$$X_1, X_2 \ldots\ldots\ldots, X_c$$

Its total at stratum level equals:

$$X = \sum_{c=1}^{C} X_c$$

Average clusters $\overline{X}$ equals its total divided by the number of clusters:

$$\overline{X} = \frac{\sum_{c=1}^{C} X_c}{C}$$

The variance in the previous totals, as is known, is calculated as follows:

$$V(X_c) = \frac{\sum_{c=1}^{C} (X_c - \overline{X})^2}{C-1}$$

The estimated total population variance in this stratum of all clusters C equals:

$$V(X) = C * \frac{\sum_{c=1}^{C} (X_c - \overline{X})^2}{C-1} * \left(1 - \frac{C}{T}\right)$$

Where T is the number of clusters in the stratum and $\left(\frac{C}{T}\right)$ is sometimes very small (less than 5%) so that it can be disregarded from the total variance formula.

The previous variance formula can be simplified to facilitate the calculation, as follows:

$$V(X) = C * \frac{\left(\sum X_c{}^2 - \frac{(\sum X_c)^2}{C}\right)}{C-1} * \left(1 - \frac{C}{T}\right)$$

4.4.2 Ratio Standard Error Calculation

Averages and proportions can be calculated by ratio $R = \frac{Y}{X}$ between two values Y and X In case of a multistage sample, the averages and proportions are special cases for this ratio. In case of the average, the numerator is equal to the estimated values being estimated and the denominator is equal to 1 for each element. Thus, the denominator equals the sum of weights for all elements. In the case of the proportion, the denominator is also equal to 1 for all elements and each element in the numerator is equal to 1 when the element is matching with the characteristic being estimated and equals zero if otherwise. The variance of ratio calculates by following

formula:

$$V(R) = \frac{1}{X^2}\{V(Y) + R^2V(X) - 2 * R * cov(Y,X)\}$$

Where:

$$cov(Y,X) = \frac{c}{c-1}\left\{\sum\left(Y_i - \frac{Y}{c}\right)\left(X_i - \frac{X}{c}\right)\right\} * \left(1 - \frac{c}{T}\right)$$

Or

$$ov(Y,X) = \frac{c}{c-1}\left\{\sum Y_iX_i - \frac{(\sum Y_i)(\sum X_i)}{c}\right\} * \left(1 - \frac{c}{T}\right)$$

References

1 - Cochran, W.G, Sampling Techniques, 1977.

2- Dalenius Elements of Survey Sampling, Swedish Agency for Research Cooperation with Developing Countries, 1985 .

3 - Hansen, M.H, Hurwitz, W.N. and Madow. W. g. Sample Survey Methods and Theory, Vol. I. Methods and Applications, Vol, II. Theory, Willey, New York, 1953.

4 - Kish, L., Survey Sampling, Willey New York, 1965.

5 - International Statistical Institute , World Fertility Statistical Institute , World Fertility Survey , guidelines for Country Report No. 1 , Basic Documentation , The Hague , 1977 .

6 - Yates, F., Sampling Methods for Census and Surveys, 1981.

7 - Sampling frames and sample designs for integrated household survey programmes1986

Household Survey Capability Program, New York United Nations. Department of Technical

8 - League of Arab States (1990), Sampling Manual, Arab Maternal and Child Health Survey

9 - Macro International Inc. (1996). Sampling Manual. DHS-III Basic

10 - Household Sample Surveys in Developing and transition Countries UN 2005

11-Designing Household survey Samples: Practical Guidelines. United Nation, New York, 2008.