مـركـز الإحـصـاء
STATISTICS CENTRE

Statistical Sampling Guide

Methodology and Quality Guides - Guide No. (1)

Table of Contents

Introduction

The Statistical Sampling Guide related to the Statistics Centre – Abu Dhabi falls within the scope of works entrusted to the Statistical Research, Methodology and Quality Standards Department pertaining to the documentation of statistical operations evidence. This guide aims to provide the statistical technicians within the centre, along with data users outside the centre, with details related to the design and sample selection procedures conducted by the centre, regardless of the type of surveys (economic survey, household survey, etc.)

This guide consists of four main chapters, whereby Chapter one includes the statistical terms and concepts related to the statistical sampling techniques. Such concepts relate to the actual reference to the sampling theory as well as to the statistical evidence and dictionaries of statistical terms adopted by the international and regional organizations. Chapter two of this guide presents the various statistical sampling technique in addition to the restrictions and advantages related to the same. It also presents the standards of selection of the best technique to be adopted in statistical sampling. On another note, Chapter three is concerned with the sample size estimates in accordance with the used sample design techniques. Furthermore, it presents the main requirements to be met to determine the sample size.

Finally, Chapter four describes, in statistical terms, the statistical sampling frames adopted in the centre, whether such frames were related to economic, household, or any other type of surveys. This chapter also includes the statistical sampling designs used in the Statistics Centre – Abu Dhabi, as well as the mechanism used in determining the size of the samples and the random techniques used in the sample selection.

Chapter 1: Sampling Concepts

In this chapter, the main concepts and definitions related to the theoretical and applicable aspects of sampling and sample design will be presented, knowing that the same aligns with the international concepts and definitions adopted in this field:

Statistical Population

The term “Statistical Population” includes all statistical units on which the statistical survey is to be conducted. These units shall be clearly defined, as they might include one or more common characteristics. The majority of the statistical populations consist of statistical units, whereby such units change in terms of time (renewable societies), while other units constitute static populations on which the time factor does not have any effect.

Statistical Survey

The Statistical Survey refers to an organized statistical process based on scientific methods and the principle of inclusion of part of the statistical population. More often, the units are chosen by adopting probability sampling technique or by including all units of the population, for data collection.

Comprehensive Census

The comprehensive Census refers to the organized statistical process based on the inclusion all of the units related to the Statistical Population during the data collection process. Usually, the comprehensive census is carried out in the social, agricultural and economics populations. The comprehensive census shall also be carried out in the event the targeted population is usually small, thus the inefficiency of the sampling technique. In addition, in the event the statistician fails to acquire a clear background on the nature of the population, they shall be mistaken and carry out a comprehensive census instead of a sampling technique.

Sampling Techniques

These techniques shall be used to select a sample of units from the population to be subject to statistical methods, whereby the results reached based on the sample data represent the targeted population indicators.

Random Selection

This process is related to the selection of units from the statistical population in such a manner that avoids any personal control which interferes with the selection or exclusion of any of the population units while ensuring the provision of a chance for each unit to appear in the selected sample.

Sampling Frame

It is a list or record including all units of the statistical population. It usually includes names and addresses of the statistical units and some other relevant information. It might also refer to a map enabling access to the statistical units for data collection.

Sample Design

Sample Design refers to a specific plan aiming to select a sample from a specific population. It also refers to the technique that shall be adopted by the statistician in the sample unit's selection process.

Sample

The Sample is a subset of the Statistical Population. It shall be selected by one of the statistical sampling techniques. The sample shall be representative of the survey population. For this purpose, the sample must include the characteristics of the population in such a manner that enables its results to be generalized to estimate population parameters.

Types of Samples

Statistical samples shall be divided into two main sections:

1- Probability Samples:

Probability samples are selected in accordance with the laws of probability, whereby their units are selected successively and with a known probability. The Probability Samples include Simple Random Samples, Stratified Samples, Systematic Samples and Cluster, etc.

Probability samples are mainly characterized by their capability to generalize their results to all population units by calculating the sampling weights, whereby the amount of weight of the sample unit depends on the probability of selecting the relevant unit from the population. Probability samples also enable the analysis of the sample results as well as the calculation of standard errors and coefficient of variation in addition to the Design Effect. So, the Probability Sample enables to estimate the margin error and the confidence level in the resulting estimates.

Therefore, the official statistics indicators depend mainly on the Probability Samples designed in such a manner that they represent their results at the level of the population as a whole and estimate the sampling errors.

2- Non-Probability Samples:

Non-Probability Samples shall be selected based on a method without referring to the laws of probability. Non-Probability Samples include purposive sampling, quota sampling, convenient sampling, snowball sampling, etc. This type of sampling is often applied in poll surveys and studies conducted on limited phenomena within the population. Such samples also give results based on data representing the sample units rather than the population as a whole.

Sampling Proportional to the Size

This type of sample is characterized by the fact that the probability of selection of each sampling unit is proportional to the size of that same unit for the studied characteristic. For example, the size of an economic establishment shall be measured by the number of its employees. At the time of selection of the economic establishments sample through the sampling proportional to the sizing process, a higher probability or chance is given to the large establishments, with a larger number of employees.

Multi-Purpose Sample

The Multi-Purpose Sample refers to the sample by which data is collected for several topics included in a single statistical survey, such as income, expenditure, health and nutrition.

Successive Sample

It refers to a sampling technique, whereby the Successive Sample covers the population for several years, where the population is divided into several groups, each being covered for one year. This technique is usually used in small populations as an alternative to censuses.

Pilot Survey

The Pilot Survey is concerned with selecting sampling units to collect their related data for the purpose of questionnaire test and to tackle the challenges that might be faced by the researcher at the time of the survey conduction.

Matched Sample

Matched samples refer to sample units consisting of similar or matching pairs, whereby the studied characteristic related to the sample is measured twice under different conditions.

Self-Weighting Sample

The Self-Weighting Sample refers to the sampling design which includes equal sampling weights for all sample units. In other terms, the probabilities of selecting all sample units shall be equal.

Convenient Sample

It is a Non-Probability sampling technique by which a sample is taken from the population due to its availability and suitability to be part of the survey, provided that it is not taken into consideration the sample to represent the whole population.

Sampling Unit

The Sampling Unit refers to the unit selected in the sample, representing an element in the statistical population. In other terms, the sampling unit constitutes the unit for which statistical data is collected.

Primary Sampling Unit

It refers to the sampling unit selected during the first stage of selecting a multi-stage sample. The Primary Sampling Unit often represents a cluster, which is a set of Secondary Sampling Units.

Secondary Sampling Unit

It refers to the sampling unit selected during the second stage of selecting a multi-stage sample. Each Secondary sampling unit is a part of the Primary Sampling Unit.

Analyzing Unit

The Analyzing Unit is used during the analysis of the collected statistical data to achieve the statistical survey objectives. The Analyzing Unit could be the sampling unit itself.

Non – Coverage Errors

The Non-Coverage Errors are due to several factors. Such errors include under-coverage or over-coverage during the preparation of the frame, the inclusion of out of scope units, or the failure to describe the units correctly with the frame. The Non-Coverage Errors shall be divided into two types: 1- Under-coverage: the failure to include the required units; and 2- Over-coverage: the inclusion of non-required units.

Non-Response Errors

The Errors result from the refusal of the respondents in the sample to response to the survey, there is a unit nonresponse, it is a full non-response to all the required variables in the questionnaire, also there is an item nonresponse it is a partial non-response to some variables in the questionnaire.

Random Errors

Random Errors are the deviation of the data values from the population parameter, provided that such deviation is made by chance. The random errors cannot be concealed in the comprehensive census. More often, random errors are limited and thus, can be measured and identified. The size of random errors depends on two main factors: the extent of difference or contrast between the population units on one hand and the sample size in relation to the population from which it was selected. The higher variance in the population units, the higher the chance of occurrence of random errors. As for the sample size, the larger sample size leads to lower value of random errors.

Bias Error

(a) Bias error in relation to estimation:

This term refers to the deviation of the mean of all population parameter's possible estimates from their real value. Such error is not easily detectable and corrected unless through radical modifications to the study design or data collection technique or the modification of the results.

(b) Bias error in relation to sampling:

The relevant bias is intentional and arises following the provision of information by the respondent, whereby such information does not match the real facts, or due intention of the survey designers to design a survey that matches their orientation or for intended purposes. The bias may also be unintended and result from the misunderstanding of the respondent of the required data to be submitted or the fact that they did not have enough time to prepare accurate answers.

Standard Error

The Standard Error represents the square root of the estimated sample variance, whereby the sample variance refers to the mean of the squares of the differences between the values of the sample units and the arithmetic mean of the concerned units.

Relative Standard Error

It refers to the standard error of the sample data divided by the calculated sample estimate of such data. The relative standard error is equal to the coefficient of variation.

Accuracy of the Sample Survey

It refers to the difference between the sample estimate and related parameter of the studied variable in the population. It shall be noted that the accuracy levels are higher with lower difference values.

Random Start

It refers to a random number selected randomly when using the systematic sample, whereby the selected number is between 1 and K, where K represents the value of the systematic period.

Optimum allocation

It constitutes one of the stratified sampling unit's allocation techniques on the various strata, whereby the share of each stratum is directly proportional to the size of the stratum and the variance within the stratum, and inversely proportional to the cost related to the data collection of the sampling unit within the concerned stratum.

Nyman Allocation

It constitutes one of the stratified sampling allocation techniques on the various strata, whereby the share of each stratum is directly proportional to both, the size of the stratum and the variance within the same stratum.

Proportional Allocation

It constitutes one of the stratified sampling unit's allocation techniques for the various strata, whereby the share of each stratum is directly proportional to the size of the concerned stratum in terms of the number of population units in the stratum.

Confidence in Sample Estimate

This term measures the extent of reliance on the sample estimates. Confidence levels increase whenever the sample size is increased, and is closer to the population parameters.

Bound of Error

It refers to the standard error value multiplied by z-value at certain confidence level.

Design Effect

It is the ratio between the variance of the estimate of an identified sampling design and the variance that estimate in simple random sample design.

Weighting

Weighting constitutes the procedure for calculating the weights for the sample units (the weight of a sampling unit is equal to the inverse of the probability of selecting the unit from the population) to be used in sample estimation

Chapter 2: Statistical Sampling Techniques

The sampling theory includes several statistical sampling techniques, whereby they all focus on obtaining a statistical sample that produces sampling estimates with the lowest possible sampling error, while considering the estimated sample size. This chapter is concerned with the statistical sampling techniques.

1. Simple Random Sample

The Simple Random Sample Technique is considered one of the simplest and most common sampling techniques. This technique is characterized that it provides each of the sampling units in the population equal chance to select or appear in the sample.

Two sample selection methods are included in the Simple Random Sampling Technique, whereby the first is used to select the sample and return it to the population to allow it to re-appear in the sample, and it is called the simple random sample with replacement. The second technique is used to select the sample while excluding each other selected sample from the population without returning it to the latter, and it is called the simple random sample without replacement, given that it does not give the re-appearance opportunity of selection another once.

1.1 Using the Simple Random Sampling Technique

Considering the ease and simplicity of applying this technique of sampling, it is therefore highly common technique. However, it shall be noted that there are requirements to be met within the sampling units of the targeted population, whereby such units shall be homogeneous in terms of the studied variable. In other words, the variance between the units of the population shall be relatively minimum. Other determinants shall be available by data users, such as the dissemination levels of the sample estimates. For example, should it be required to get the results at the level of the region, the random sample shall not be selected at the level of the population as a whole. Other determinants include the survey costs as well as other such as the availability of trained field personnel and transportation costs within the population units, which limits the use of this technique. Furthermore, this technique may be affected by the survey cost unit more than another sampling technique. Therefore, caution must be exhibited when using the Simple Random Sampling Technique.

1.2 Simple Random Sampling Selection

When choosing the simple random sampling method, then a technique that guarantees the randomness in the selection of the sampling units shall be chosen, whereby each unit of the population is granted an equal probability of selection. On the practical level, different methods are used to select simple random sample, including marking all the units of the population with a serial number, followed by choosing a random number from the random numbers table. The next step includes the selection of the sampling unit which serial number matches the selected random number. In the event where the chosen random number is not matched with the serial number of any population unit, another number shall be selected, and so on.

Example (1):

Consider a population of 60 establishments where a simple random sample of 4 establishments shall be selected. In this case, the establishments of the population are marked with serial numbers, starting from 01, 02, 03…. to 60. A random number of two decimals is selected from the random numbers table or via the computer. In this example, let's consider that the number 45 has been selected. Accordingly, the establishment with the serial number 45 shall be selected from the sample. In the following selection process, if the selected random number is 73, it shall be neglected, and the selection process shall be repeated. Subsequently, 4 establishments representing a simple random sample are selected.

2. Systematic Random Sample

The Systematic Random Sampling Technique is characterized by its easy and simple application, as well as the wide spreading of the sample in the population due to the selection technique carried out as a sequential systematic procedure. The linear systematic sampling technique is the most common with the systematic samples and its application shall be summarized as follows:

Consider a population of N units, whereby the size of the required sample to be selected is n. if we divide the size of the population N by the size of the required sample n, we shall obtain the value k, where the equation is nk=N. Statistically, the value k is known as the systematic period. A random number between 1 and the value k shall be selected, whereby it shall be referred to as the Random Start Number and marked by (I). The serial number of the first sample shall be (I), followed by the second sample (K+I), the third sample (K+I2), and so on.

Example (2):

In order to estimate the number of employees in the establishments within a selected region consisting of 400 establishments, a Systematic Sample of 16 establishments has been selected from. The sample selection technique shall be as follows:

$$N = 400,\ n=16$$

$$k = N/n = 400/16 = 25$$

A random number between 1 and 25 is then selected. In this case, consider the number 14. Therefore, the serial numbers of the selected sampling units within the sample shall be as follows:

$$14, 14 + 25, 14 + 2 \times (25), 14 + 3 \times (25) \ \ldots$$

Which is equal to: 14, 39, 64, 89, …

2.1 Advantages and Determinants of Using the Systematic Random Sample

The best advantage characterizing the Systematic Sample is represented by the ease in the selection of the samples and the allocation of the same adequately to the population. The Systematic Sample is considered efficient in comparison with the Simple Random Sample for many populations, mainly in the event of the linear tendency of the studied variables.

The difficulty of establishing unbiased estimate for variance represents a determinant in this type of sample, as well as in the case of a cyclical variable in the population, which might lead to producing a bias in the selected samples and the estimation process. For example, the sample of one member has been selected from each household in a group of families. The start number shall be 1 and the Sample Period shall be 2. The members holding the number 3 were selected in the sample. It shall be noticed that all the members of the sample belong to the third rink in the household. In the event the arrangement of the members within the household was as follows: the father, the mother, the son, the daughter, it shall be noticed that all the units in the sample refer to sons, which might lead to a bias in the sample estimates.

3. Sampling Proportional to the Size

As already mentioned, the best advantage characterizing the Simple Random Sample is represented by the equal chance of selection for each unit of the population. For example, when selecting a sample of economic establishments using the simple random sampling technique, all establishments, of all sizes, whether small or large, have the same chance of selection in the sample. In some surveys, the nature of the studied variable might require giving a greater probability for the appearance of specific units from the population in the selected sample. For instance, the nature of the studied variables might require giving more chance for the selection of the large establishments (establishments with a greater number of employees). In this case, the Sampling Proportional to the Size is the best fit sampling technique.

In addition, the Sampling Proportional to the Size technique is known as the cumulative collection and is summarized by giving each population unit a number related to the importance or size of the unit in the population.

Example (3):

Within the frame of economic establishments, the establishment with 1,000 employees is considered to have a weight of 1,000, which means that it contains 1,000 hypothetical units. Accordingly, the establishment with 100 employees is considered to weight 100, and so on.

The following tables clarify the aforementioned example:

Establishment	Number of Employees	Cumulative Number of Employees	Accompanying Numbers
1	1000	1000	1 – 1000
2	700	1700	1001 – 1700
3	1200	2900	1701 – 2900
4	500	3400	2901 – 3400
5	300	3700	3401 – 3700
6	800	4500	3701 - 4500

In order to select three establishments, three random numbers between 1 and 4500 shall be selected using the random numbers table or via the computer. The random number shall then be located conveniently in the column entitled “Cumulative Number of Employees” as shown in the above table. The establishment with the equal or greater cumulative number of workers shall be selected. In this example,

if the random numbers 75, 2000, and 4000 were selected, the establishments within the sample shall bear the numbers 1, 3, and 6, respectively.

Whenever the population is relatively large, the aforementioned sampling technique might be very time-consuming. Therefore, another technique, Lahiri, is used to select the sample proportional to the size. This technique includes selecting pairs of random numbers, whereby the first number of each represents the number of the sampling unit. The selected random number shall be between 1 and N, where N represents the number of the total sampling units within the population. The second number of the pair represents the size of the sampling unit, and shall be between 1 and M, where M represents the size of the larger sampling unit within the population in terms of the variable according to which the sampling weights is calculated.

4. Stratified Random Sample

As clarified earlier, the optimum Simple Random Sampling Technique required high possible level of homogeneity between all the units of the population. Otherwise, the sample will lead to estimates with relatively high level of bias and inaccurate. In addition, the produced sample estimates will not be able to break down at low domains of the population, also it will increase the burden and costs incurred in the data collection process. Due to limitations in the homogeneity in several populations, the Stratified Random Sampling Technique shall be adopted, whereby the population is divided into several non-overlapping groups, each being homogeneous in terms of the studied variable and referred to as “stratum”. This will provide more accurate results. For instance, when studying the income mean or educational level of the head of household, the population may be divided into rural and urban groups.

For the efficient usage of the Stratified Random Sampling Technique, accuracy shall be maintained, mainly when conducting the following processes:

- Stratum formation;
- Number of desired strata;
- Size of the sample in each stratum;
- Estimation method.

Example (4):

Consider the following eight establishments with the number of employees:

Establishment Number	Number of Employees	Establishment Number	Number of Employees
1	3000	5	6000
2	1500	6	1200
3	7000	7	4500
4	2500	8	5000

In order to estimate the mean number of employees in the establishments, the Simple Sampling Technique might be adopted for the selection of three samples. In the event the selected sample consists of the establishments with the numbers 1, 2, and 6, the sample mean of employees shall be 1,900. On the other hand, the actual mean of the population is 3,838 employees, which exceeds the double value

of the mean estimated in the sample. Therefore, an error is clearly detected due to the usage of the simple random sampling technique for a non-homogenous population.

When applying the stratified sampling technique, the population shall be divided into strata in accordance with the categories of the number of Employees as follows:

Stratum I – Number of Employees (<3,000)		**Stratum II – Number of Employees (3,000 – 5,000)**		**Stratum III – Number of Employees (>5,000)**	
Establishment No.	Number of Employees	Establishment No.	Number of Employees	Establishment No.	Number of Employees
2	1500	1	3000	3	7000
4	2500	7	4500	5	6000
6	1200			8	5000

A sample consisting of one establishment has been randomly selected from each stratum, which are with the serial numbers 2, 7 and 5. The estimated sample mean of employees is given by:

$$\left(\frac{3}{8}\times 6000\right)+\left(\frac{2}{8}\times 4500\right)+\left(\frac{3}{8}\times 1500\right)=3938$$

It shall be noticed that the aforementioned result calculated by stratified random sample is very close to the actual mean of the population when compared with the result established by the simple random sampling technique. Therefore, it shall be concluded that using the stratified sampling technique is a necessity in such cases.

4.1 Advantages and Determinants of Using the Systematic Random Sample

The following shall be considered when dividing the population into strata:

- The total sum of units in all the strata shall be equal to the total units of the population, while ensuring then non-overlapping between strata.
- The population units inside each stratum should have highest possible homogeneity in terms of the studied variable.
- In order to obtain sample estimates at geographical or administrative divisions (district, region, etc.), the population shall be segmented into geographical strata in accordance with the required representation domain.
- Prior to stratification, the process of categorizing the population units into categories is considered a type of stratification, where each of the categories may be considered as a stratum itself.

Stratification Method:

This method aims to limit the variance between the units of the population in relation to the studied variable. For instance, in the event the study tackles the average income of the household, the population of the households may be divided into groups according to their income levels. On the other hand, if the study is concerned with the number of employees in the industries, the latter may be stratified according to the number of employees, and so on.

Many techniques shall be adopted to determine the suggested number of categories, among which are the following:

Number of categories = 1+3 Log(n).

Example (5):

In the event the number of establishments in a determined sector is 1,850 establishments, the number of categories into which the population may be divided shall be as follows:

$$3+1(\text{Log } 1850) = 11$$

In order to divide the population into strata, the Cochran technique may be used, whereby it considers both, the number of the units of the population and the weight of each unit, given that the number of required strata is priory determined. The following example clarifies the stated:

Example (6):

If the number of establishments in the population is 1,850, whereby they are divided into initial categories according to the number of employees of 12 categories, should it be required to stratify the concerned population into four strata, being the following:

Category	Number of Establishments	Number of Employees	Cumulative Number of Employees (C)	$\sqrt{C}$
1 - 5	200	720	720	26.8
6 - 10	250	2000	2720	52.2
11-15	300	3900	6620	81.4
16-20	300	5250	11870	108.9
21-25	250	5750	17620	132.7
26-30	200	5400	23020	151.7
31-35	100	3650	26670	163.3
36-40	80	3000	29670	172.2
41-45	80	3360	33030	181.7
46-50	50	2400	35430	188.2
51-55	30	1590	37020	192.4
56 and above	10	950	37970	194.9

The square root of the Cumulative Number of employees ($\sqrt{C}$) in the last column shall be divided by the number of suggested strata:

$$\frac{194.9}{4} = 48.7$$

The first stratum corresponds to the square root of the cumulative number of employees (48.7). approximately, the first stratum boundaries may be considered of 1-10. By using the same technique, we shall obtain the following strata boundaries:

- Stratum I: 1 x 48.7 1-10
- Stratum II: 2 x 48.7 11-20
- Stratum III: 3 x 48.7 21-30
- Stratum IV: 4 x 48.7 above 30

4.2 Advantages of Stratified Sampling

1. In the stratified sample, the population shall be homogenous in each stratum. The population shall be well represented, as the samples are selected from different strata, mainly those of special importance.
2. The usage of the stratified sample shall be more efficient in comparison with the other types of samples, mainly in the event of a non-homogenous population and in the event of extreme values of some population units.
3. The stratified sampling leads to the reduction of the incurred costs as it reduces the sample required to be covered at a certain accuracy level.
4. The stratified sample may be used to obtain results at specified geographical or administrative levels (districts, sectors, regions, etc.).
5. Controlling, supervising, and organizing the field work, as well as determining the area of work of each group is better achieved when dividing the population into strata according to the geographical and administrative areas.

4.3 Stratified Sample Allocation

There are different ways to distribute the total sample (n) to the different strata, whereby the size of the sample in the stratum h is n_h.

Among the most important ways, we note the following:

Equal Allocation:

This technique is usually used whenever the need to obtain results at the level of the administrative regions arises (in the event the stratum represents an administrative region) or in the event of equal allocation of works within all the strata (according to the availability of the fieldwork capabilities). In addition, the equal allocation technique is used when the population size is nearly equal in all the strata.

In this case, the size of the sample in a single stratum shall be calculated according to the following equation:

$$n_h = \frac{n}{L} \quad \text{..... (1)}$$

Where L represents the number of strata.

Example (7):

If the sample size determined by 2,000 sampling units, given that the number of strata is 8 strata, the size of the sample in e stratum h is equal to:

$$\frac{2000}{8} = 250$$

Proportional Allocation:

This is the most common allocation technique, given its easy application. In fact, when no information other than the number of the units of the population in each stratum is available, the following equation may be used to estimate the number of samples in the stratum (h):

$$n_h = n(\frac{N_h}{N}) \text{ ..... (2)}$$

Where N_h represents the number of population units in the stratum (h).

It may be concluded from the above that the sample percentage in a single stratum is equal for all the strata, which leads to a self-weighted sample. In this case, there is no need to calculate the sampling weights when establishing the sample estimates, which enables the ability to give quick and highly accurate estimations.

Example (8):

When estimating the number of employees in a specified population, a sample of establishments has been selected from each region through the stratified sampling technique and by using the Proportional Allocation methods for a total sample size of 35 establishments. The size of the establishments in each stratum according to the above equation (2) shall be as follows:

Stratum (Region)	Total number of establishments	Sample size in the stratum
1	200	20
2	100	10
3	50	5
Total	**350**	**35**

Nyman Allocation:

The most important advantage characterizing the Nyman Allocation technique is represented by its usage to reduce the variance and increase the accuracy and efficiency of the data, whereby it takes into consideration the variance of each stratum in addition to the size of the same when allocating the total sample to the strata. The size of the sample in the stratum is directly proportional to the standard deviation of the same in order to increase the efficiency of the sampling design in comparison with the proportional allocation technique. This technique is usually used when the standard deviation varies between the strata. Whenever the total sample size as well as the sample cost are fixed among the different strata, the sample size of the stratum (h) shall be estimated according to the following equation:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} n \text{ ..... (3)}$$

On the other hand, the standard deviation at the level of each stratum shall be obtained according to a previous census or may be estimated by reference to previous surveys or other similar surveys.

Example (9):

When estimating the household average income in a certain region, the population shall be divided into three strata according to the income category. A sample of each category shall be selected according to the Nyman Allocation technique, with a total sample size of 15 households. Considering the following data:

Stratum (Income Category)	Number of households in the stratum (N_h)	Standard Deviation to each stratum (S_h)	$N_h S_h$
Less than 1,000	30	20	600
Between 1,000 – 3,000	50	30	1500
Greater than 3,000	20	20	1000
Total	**100**	**-**	**3100**

The Sample size in the first stratum =

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} = 15 \times \frac{600}{3100} = 3$$

The sample size in the second stratum=

$$= 15 \times \frac{1500}{3100} = 7$$

The sample size in the third stratum=

$$= \frac{15 \times 1000}{3100} = 5$$

Optimum Allocation:

The Optimum Allocation technique aims to reduce the variance as little as possible with a determined cost or to reduce the cost as little as possible with a certain accuracy level, where the cost factor is included in the sample allocation to the strata. This technique is usually used in the case of a variability in the data collection cost between the strata. For instance, the data collection cost in certain areas is much higher than the data collection cost in others. Many equations are used for this purpose, of which we present the following:

$$n_h = \frac{N_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{C_h}} \; n \quad \text{..... (4)}$$

Where (C_h) represents the cost of a single sampling unit in stratum (h).

Example (10):

To estimate the field production, mean in several agricultural regions, a stratified sample has been selected, where each region constituted a single stratum. The total number of selected samples was 200 samples. According to the aforementioned equation (4), the size of the sample in a single stratum as established in the last column of the below table is as follows:

Stratum (Region)	Number of fields N_h	Standard deviation of the stratum S_h	$N_h S_h$	Unit Cost (C_h)	$\frac{N_h S_h}{\sqrt{C}}$	Sample Size (n)
1	200	10	2000	4	1000	58
2	100	20	2000	6	816	47
3	150	10	1500	6	612	36
4	100	15	1500	9	500	29
5	50	20	1000	12	287	17
6	50	20	1000	20	224	13
Total	**650**				**3439**	**200**

Proportional Allocation Technique with a specific characteristic:

The design and selection of some samples are required in accordance with the relative importance of the variables of the studied variable in the sample. The following example may be adopted to give a clearer idea thereon.

In the event the estimation of the revenues of the establishments is required in the various regions (strata), considering the following data:

Strata	Number of establishments	Production quantity
1	1000	20000
2	500	30000
3	1500	10000

The size of Stratum 3, according to the number of establishments therein, represents 50% of the population units, although it contains 0.01667 of the production quantities. Therefore, in the case the number of establishments is used to weigh the stratum, misleading results may be produced, whereby the share of Stratum 3 represents 50% of the size of the selected sample. However, the share of Stratum 2 represents 17% of the selected sample, although it contains 50% of the total production quantity related to revenues, which reduces the accuracy and efficiency of the sample and increases the related costs.

For this purpose, the best technique to be adopted consists of allocating the sample to the strata in accordance with the production quantity, being the variable to be studied. The sample shall be allocated between the strata based on the production quantity rather than the number of establishments, using the following equation to allocate the sample in accordance with the proportional technique with a specific variable:

$$n_h = \frac{X_h}{\sum_{n=1}^{L} X_h} n \quad \text{..... (5)}$$

Where X_h refers to the value of the variable contained in stratum (h), being the production quantity according to the above mentioned example.

If the required sample size consists of 100 establishments, the allocation of the sample to the strata in accordance with the proportional technique and the production quantity variable shall be as follows:

Stratum	Sample Size
1	33
2	50
3	17
Total	**50**

5. Cluster Sampling

The Cluster Sampling Technique is based on the principle of conveniently dividing the population into groups, whereby the latter are close in terms of size and homogenous in terms of the studied variables. Each of the groups is called “cluster”, and the group of clusters represents the complete population without omission or repetition.

The most important advantage characterizing the cluster sample is its efficiency in terms of the cost unit, this technique is usually used in the populations lacking sampling frames or where it is difficult to provide an updated frame to the population units. However, a cluster frame may be provided, saving time and effort. Another advantage exists in using this technique represented in saving the transportation costs during the data collection stage between the sampling units. It shall be kept in mind that the disadvantages of the cluster sample are represented by it being less effective than the simple random sample as it is less prevalent.

The following shall be taken into consideration when using the cluster sample:

- Matching the number and size of clusters, whereby the size of the cluster is relatively small, and their number is relatively large.
- When forming the clusters, units from the populations close spatially or within a certain geographic area shall be selected, being similar in terms of the studied variable.
- Consistency in terms of the sizes of clusters, whereby they shall be as close as possible in size.
- Each cluster shall be clearly defined in order to be distinguished from another.

6. Multi-Stage Sampling

The main challenge faced by several surveys is represented by the lack of an updated frame for the main sampling units, such as establishments, housing, etc., whereby it is difficult to develop an updated frame for the same. At the same time, a list or frame with a variable is provided at a collective rather than a detailed level, such as population communities or major regions. In the relevant cases, the multi-stage sampling technique may be used.

6.1 Multi-Stages Sampling Advantages

- Saving time and money, whereby setting a frame with the primary sampling units shall be sufficient.
- The multi-stage sample is characterized by flexibility, whereby the sampling technique may be used at each of the various stages.

Within the multi-stages sampling technique, it is preferable to divide the population into equal primary sampling units in the following cases:

- When the size of the primary sampling units is relatively large, it shall need more time in terms of preparing the secondary sampling units' frame.
- When the size of the primary sampling units is small, it shall need time in terms of navigating between samples.

6.2 primary sampling units' selection

- The simple random unit may be used in the event the primary sampling units are homogenous.
- The primary sampling units may be divided into strata, and a sample shall be selected from each stratum.
- In case of significant variance between primary sampling units, the proportional to the size sample may be used.
- The systematic sample may be used. However, it will not produce an unbiased estimate for the sampling error.

As for the secondary sampling units, they shall be selected through any of the following sampling techniques: simple random sampling, systematic sampling, cluster sampling, or proportional to the size sampling.

Chapter 3: Sample Size Estimation

This chapter is concerned with the sample size estimation when conducting a statistical survey in addition to the main requirements to be made available in order to access the accurate estimation of the sample size. Furthermore, it presents a sample size estimation mechanism in accordance with the various statistical sampling techniques.

1. Requirements for Sample Size Estimation

The sample size estimation process is based on formulations and math equations depending on a number of variables that shall be made available when conducting the convenient sample size calculation process. The relevant variables are as follows:

1. The confidence level in the estimates to be built based on the size of the sample, being 90%, 95%, 99%, etc., whereby it statistically represents the areas of the normal allocation under the standard normal curve where the values of Z are 1.64, 1.96, 2.58, etc., respectively. The confidence level pertaining to the estimation value is positively related to the sample size, meaning the higher the size, the higher the estimation confidence level.
2. The estimation margin error, which means the difference between the actual and estimated value of the parameter for which an estimation shall be found using the sample data. The sample size is directly related to the margin error, meaning the lower margin error in the estimation (reduction of the error) required the increase in the sample size.
3. The variance of the population. In case when the population variance is unknown, a convenient variance estimation shall be adapted. In the event the survey aims to estimate various indicators in the survey, a convenient indicator shall be selected, known as the Key Indicator, to estimate the sample size while depending on estimating all the required indicators with high accuracy.

The confidence level and margin error shall be determined as a requirement, while calculating the sample size. A larger sample size shall be needed whenever the margin error value (meaning the required bound of error when estimating (d)) is low and when the confidence level exceeding the margin of error value (d) is high. Sample size determination carried out by studying several values of each of the confidence levels (z) and the margin errors (d). On this basis, different scenarios shall be established for the sample size, which enables the survey management to establish a balance between the same in accordance with the costs and available requirements.

2. Pre-assessment of the population variance

The formulation of sample size equations is based on the availability of estimated value for the population variance value (σ^2). Therefore, it is necessary to conduct an estimation process for the variances. Statistically, there are several techniques by which the variance of a population can be estimated to calculate the sample size:

2.1. Two-stage sample segmentation

Following this technique, the sample shall be divided into two parts and executed in two stages. In the first stage, a simple random sample of size (n_1), based on which the variance value is estimated by s_1^2. This estimation is also used to calculate the final sample size (n).

The remaining sampling units of the complete sample size shall be selected in the second stage after estimating the variance value in accordance with the sample size depending on this estimation.

Example (1):
For the calculation of the sample required for estimating the annual average household expenditure, it shall be sufficient to randomly select 500 households from the population subject to the study. The variance shall be estimated accordingly. The sample size estimation shall be used, whereby it might show that the sufficient sample size (according to the aforementioned accuracy and confidence levels) is 2,300 households. The survey concerned with the remaining sample units shall be completed, which is the difference between 2,300 and 500, showing a result of 1,700 establishments.
This technique is mainly characterized by the provision of trusted estimations to the parameter S^2. However, it is limited by the fact that major efforts shall be exerted for long periods. In addition, the sample selection has been divided into two phases, each sample unit was selected with a different sampling fraction, meaning that the probability theory was not employed in selecting the sample in its optimal form.

2.2. Conducting a Pilot Survey

This technique is based on benefiting from the pilot survey data executed near the main surveys in order to achieve other objectives, such as testing the survey questionnaire and control auditing rules as well as the estimation for the preparation of the necessary number of survey team to carry out the survey. Pilot surveys may also aim towards benefitting from their data in the estimation of the population variance and calculation of the necessary standard deviation in order to estimate the main survey sample size.

2.3. Prior surveys results

Practically, this is the most common technique as, when estimating the convenient sample size, prior surveys shall be reviewed and conducted on the same population or a similar one in order to estimate the standard error value. Although the variance is more static in comparison with the changes occurring in the central tendency measures for the phenomenon under study, than the standard deviation, as changes within the phenomenon conduct might occur over time.

2.4. Ratio estimation Technique

The ratio indicators (the number of cases characterizing the studied variable to the total number of cases) are among the main indicators included in studies and surveys. The ratio (p) may be estimated. It shall also be well known that the closer the ratio to the actual parameter, the more accurate the variance

estimation. For instance, if the parameter value in the population is p=0.3, the variance value shall be $(S^2 = p(1-p) = 0.3 \times 0.7 = 0.21)$.

It shall be noticed that the variance value reaches its maximum when the ratio is equal to 0.5, as follows:

$$(S^2 = p(1-p) = 0.5 \times 0.5 = 0.25)$$

This assumption requires the selection of the largest sample size under the previously adopted accuracy and confidence levels.

If the variance value is unknown regarding a certain phenomenon, and if the convenient sample size estimation is to be carried out, the variance value shall be set at 0.25, providing the largest sample size possible as a precautionary measure.

3. Choosing the convenient variable for the estimation of the sample size

It has been practically well known that the objective of conducting a statistical survey is not limited to data collection in relation to a single indicator. However, several indicators shall be available to collect their respective data and analyze the studied phenomenon in its different aspects. On another hand, the majority of the national and international statistical institutions are focusing on conducting Multi-indicators Surveys, which have become common and widely spread with the data collection techniques development and their automatic processing. The results obtained by such surveys provide various indicators. Among these surveys, we shall note the Multi-Indicators cluster Survey which aims to collect detailed indicators regarding the status of the children and mothers as well as their surrounding conditions and factors.

The main challenge that may be faced when determining the sample size in this type of surveys is related to the mechanism to be adopted in order to choose the convenient variable that may produce a sufficient sample size to obtain an estimate of the value of the concerned indicator as well as estimations for other indicators included in the survey, whereby they shall all be characterized with accuracy and efficiency.

This challenge leads to the fact that the optimum technique for the determination of the sample size shall be according to the following:

- Determining all important indicators included in the survey, as well as selecting an important indicator requiring the largest sample size.
- Ensuring keeping the error rate at a specific value in terms of the various indicators regarding which data is collected.
- Determining the targeted sub populations for each indicator and select the one with the lower rate in the overall survey population while considering the importance of the same in the overall survey objectives.

Example (2):

Based on Household Income and Expenditure Survey, if the ratio of the households spending money on durable goods refers to 82%, and the ratio of households with children attending school refers to 60%,

respectively, the sample size selected from the first indicator shall be adopted, although the second indicator requires a sample larger in size. It shall also be evident from the following calculations, considering that Z=1.96 and d=0.05.

$$n_1 = \frac{(1.96)^2\,(0.82)(0.18)}{(0.05)^2} = 202$$

$$n_2 = \frac{(1.96)^2\,(0.6)(0.4)}{(0.05)^2} = 369$$

The first ration is extremely significant in comparison with the second one. Accordingly, increasing the sample size from 202 to 369 is not justified to reach an indicator that is not considered essential for conducting the Household Income and Expenditure Survey.

4. Sample size estimation to estimate the Ratio (p)

The estimation of the convenient sample size is related to the nature and type of the main indicator, whereby sample data shall be used in estimating the same. The indicators may be numerical, such as the mean, like height means (in centimetres), the weight mean of the package (in grams) or the student scores mean (in degrees). The indicators may also be concerned with the phenomenon occurrence rate (frequency), such as the percentage of adult smokers.

The sample size estimation method for estimating both of means or ratios, is based on the characteristic of normal distribution of the related population.

4.1 Sample size estimation for the population ratio with absolute margin error (d)

If $\hat{p}$ is the estimation of the percentage of the population p based on a previous survey, or based on pilot survey, in accordance with the simple random sampling technique, the sampling distribution for estimation shall be the approximate normal distribution, where

$E(\hat{p}) = p$ and the variance $var(\hat{}_P) = \frac{p(1-p)}{n}$.

The value (d) represents the margin error, being the difference between the actual population percentage and the estimated percentage of the sample. It shall be expressed according to the following:

$$d = Z_{\frac{1-\alpha}{2}} \sqrt{\frac{\hat{P}\hat{q}}{n}} \quad \text{..... (1)}$$

Where $\hat{q} = (1 - \hat{p})$

The value $Z_{\frac{1-\alpha}{2}}$ in the aforementioned relation (1) expresses the number of standard errors, for the difference value between the estimated percentage and the population parameter (P). The value of (d) refers to the accuracy level. In order to obtain a small value of (d), this usually requires selecting a larger sample size. In case the value of Z=1.96 was chosen, this means that 95% of the possible sample percentages will have a standard error range of 1.96.

The aforementioned leads to the conclusion stating that, as the sample size increase the margin error relatively decrease, which mean that there is strong relation between both of the data accuracy and the related sample. In order to obtain the sample size, the solution for equation (1) above with respect to (n) obtain the following:

$$n = \frac{\left(Z_{\frac{1-\alpha}{2}}\right)^2}{d^2} \hat{p}\hat{q} \quad \text{..... (2)}$$

The above equation leads to the conclusion that the size of the sample increases with the higher numerator value in equation (2) at a certain value of Z, resulting in the change of the value of n according to the change of the value $\hat{p}$.

The larger sample size possible is achieved when the percentage value $\hat{p}$ is equal to 0.5. As such, the value shall be as follows: $\hat{p}\hat{q}$ = 0.25.

4.2 Sample size Estimation for the population Ratio with the relative Standard Error ε

The sample size may be estimated depending on the error value as a percentage of the estimation value p (relative margin error) rather than the absolute error value, meaning in the case where the error value is d= εp. According to the aforementioned equation (1), the estimation of the sample size shall be calculated as follows:

$$n = \frac{\hat{q}\left(Z_{\frac{1-\alpha}{2}}\right)^2}{(\hat{p})(\epsilon)^2} \text{..... (3)}$$

For instance, of the value of $\hat{p}$ = 0.2 and the required estimation value is ε=0.10, the absolute error value is d = 0.2 x 0.10 = 0.02. In this example, the estimated sample size shall be:

$$n = \frac{0.8(1.96)^2}{(0.2)(0.10)^2} = 1536$$

5. Sample Size Estimation for the Population Mean (μ)

In order to estimate the population mean or population total, the statistical principle adopted to calculate the sample size is based on the given absolute or relative sample error is as follows:

5.1 Sample Size Estimation to the Population Mean with Absolute error in Points (d)

The required sample size to estimate the population mean (μ) for a certain variable shall be calculated according to the following equation:

$$n = \frac{\left(Z_{\frac{1-\alpha}{2}}\right)^2 \sigma^2}{d^2} \quad \text{..... (4)}$$

Where:

σ^2 represents the population variance and may be estimated using the s^2 value of a sample selected from a previous similar survey or based on the results of a previous pilot survey.

the value of d (in points) represents the margin error in estimation, being the absolute value of the difference between the actual population mean (μ) and the estimated mean concluded from the sample data ($\bar{x}$).

Example (3):

The available information shows that the variance in production from in several agriculture holdings is 140 kgs. When estimating the production means in a single agriculture holding, provided that the difference between the estimated mean from the sample data and the actual mean does not exceed 1 kg, with a confidence level of 95%, then the required sample size (number of holdings) to estimate the holding production mean shall be as follows:

The absolute margin of error is: d = 1. The value of Z, achieving a confidence level of 95%, is: Z = 1.96, and the population variance is: σ^2 = 140.

According to the aforementioned equation (4), the sample size estimation shall be as follows:

$$n= \frac{(140)(1.96)^2}{1^2} =538$$

The size of the sample, being (538), is relatively large, due to the selection of a small absolute error size. In the event a difference not exceeding 2 kgs (d = 2) is accepted, the sample size will significantly decrease, as follows:

$$n= \frac{(140)(1.96)^2}{2^2} =134$$

5.2 Sample Size Estimation to the Population Mean with relative Margin Error (ε)

In case when the relative value of margin error has been adopted instead of the difference in points, equation (4) shall be replaced with the following:

$$n= \frac{Z^2\sigma^2}{\varepsilon^2(\mu)^2} \quad .....(5)$$

Where:

ε represents the margin of error percentage; and

μ represents the population parameter (arithmetic mean).

The real values of σ^2 and μ may not be available. Therefore, they shall be replaced by the values $\bar{x}$ and s^2 respectively from previous surveys or pilot studies., and the size of the sample shall be expressed as follows:

$$n= \frac{Z^2 s^2}{\varepsilon^2(\bar{x})^2} \quad ..... (6)$$

Example (4):

In the previous example, if the margin of error is referred to when the sample mean represents 5% of the real values of the mean, which, according to a previous study, refers to a value of 80 kgs (μ = 80), and the value of the variance is 250, the size of the sample shall be:

$$n=\frac{(250)(1.96)^2}{(0.05)^2(80)^2}=60$$

If we consider the equation (5), it shall be noticed that the value of σ^2 divided by μ^2 represent the value of the square of the Coefficient of Variation, whereas:

$$C.V=\frac{\sigma}{\mu}$$

This coefficient is usually unknown and may be estimated from the sample data collected from a previous survey.

In the event it is required to adopt the value of the Coefficient of Variation to estimate the sample size of the current survey, it shall be expressed using the following equation:

$$n=\frac{\left(Z_{\frac{1-\alpha}{2}}\right)^2 (C.V)^2}{\varepsilon^2} \quad \text{..... (7)}$$

6. Sample Size Estimation in stratified sampling design

The previously methods related to the sample size estimation are mainly applicable in case of the simple random sampling technique. However, as concluded earlier, the stratified sampling design is based on dividing the population into independent strata, whereby the size of the sample and the sampling of each stratum shall be independently estimated in accordance with a number of variables, being the size of the stratum, the variance in each stratum, and the cost related to the sampling of a single unit.

The sample size equation is based on the relation between the margin of error (d) and the required standard error as follows:

$$d = Z.S(\bar{x})$$

If d refers to the value of the margin error and Z to the confidence level, and by adopting the following equation related to the standard error in estimating the stratified sample: $S^2(\bar{x}_{st})=\frac{d^2}{Z^2}=B^2$, the overall stratified sample size shall be as follows:

- In equal allocation:

$$n_{eq}=\frac{L\sum_{h=1}^{L}N_h^2S_h^2}{N^2B^2+\sum_{h=1}^{L}N_hS_h^2} \quad \text{..... (8)}$$

- In proportional allocation:

$$n_{prop}=\frac{N\sum_{h=1}^{L}N_h^2S_h^2}{N^2B^2+\sum_{h=1}^{L}N_hS_h^2} \quad \text{..... (9)}$$

- In Nyman allocation:

$$n_{Ney} = \frac{\left(\sum_{h=1}^{L} N_h S_h\right)^2}{N^2 B^2 + \sum_{h=1}^{L} N_h S_h^2} \quad \text{..... (10)}$$

- In optimum allocation:

$$n_{opt} = \frac{\sum(N_h S_h \sqrt{C_h}) \sum \frac{N_h S_h}{\sqrt{C_h}}}{N^2 B^2 + \sum_{h=1}^{L} N_h S_h^2} \quad \text{..... (11)}$$

6.1 Alignment of the stratified sample size in accordance with the survey costs

In case when the overall budget for the survey is the only determinant for the sample size, with total cost is given by amount C, with an equal sampling unit cost in all strata (C_h) the sample size shall be calculated according to the following equation:

$$n = \frac{C}{c_h} \quad \text{..... (12)}$$

Example (5):

Suppose that the total available costs to conduct a survey is given by AED 160,000, whereas the cost of data collection from each sampling unit is AED 100, the overall sample size shall be:

$$\text{n=} \frac{C}{c_h} = \frac{160000}{100} = 1600$$

- When the total cost is previously determined and.

Assuming that the total survey cost is already known and is equal to C and that the cost of sampling a single unit (c_h) differs from one stratum to the other, and the variances between the strata are fixed despite the difference in costs, the overall sample size shall be calculated as follows:

$$n = C. \frac{\sum \frac{N_h}{\sqrt{C_h}}}{\sum N_h \sqrt{c_h}} \quad \text{..... (13)}$$

Example (6):

Suppose that the total cost has been determined with a value of AED 160,000 with the following values attributed to the size and cost:

Stratum No.	N_h	C_h	$\sqrt{C_h}$	$N_h\sqrt{C_h}$	$N_h/\sqrt{C_h}$
1	4000	36	6	24000	667
2	2700	81	9	24300	300
3	1600	100	10	16000	160
Total	**8300**			**64300**	**1127**

The estimated sample size shall be as follows:

$$n = C. \frac{\sum \frac{N_h}{\sqrt{C_h}}}{\sum N_h \sqrt{C_h}} = \frac{(160000)(1127)}{64300} = 2805$$

- When the total cost of the survey has been previously determined, whereas a different value is attributed to each of the strata sizes and variances value among each, the sample size estimation shall be carried out according to the following equation:

$$n = C.\frac{\sum_{h=1}^{L} N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}} \quad \text{.....} \quad (14)$$

- When the costs and variance are previously determined in each stratum with a given accuracy level, the sample size shall be estimated as follows:

$$n = \frac{\sum(N_h S_h \sqrt{C_h})\left(\sum \frac{N_h S_h}{\sqrt{C_h}}\right)}{N^2 B^2 + \sum_{h=1}^{L} N_h S_h} \quad \text{.....} \quad (15)$$

Where B represents a certain level of accuracy expressed as follows: $B = \frac{d}{z}$, whereas the value of (d) represents the previously determined margin of error. And z value is related to the confidence level.

7. The Design Effect in Cluster Sample Size Estimation

The simple random sampling technique constitutes the cornerstone of sampling designs. As a result of various determinants, such as the non-homogeneity in the population, the high cost, and the lack of an accurate sampling frame on some occasions, more complex sampling techniques shall be adapted, knowing that all techniques designed based on the simple random sampling theory. In the case of the cluster sampling technique, for example, different probability values relate to each cluster, which further results in variables that may not apply with the independency and normal probability distribution assumptions. However, it shall lead to more complex allocations, and complications in the statistical analyses.

In the event, that it has been required to apply the cluster sampling technique on a certain survey, and to determine the required sample size with a known confidence level and margin error, it shall be referred to as the normal probability distribution assumption, estimated sample size shall be obtained in accordance with this purpose.

However, the determined sample size is for a simple random sample rather than a cluster sample. It is well known that the variance related to the cluster sampling is higher than that of the simple random sampling, which means that this challenge shall be faced by increasing the size of the sample in the cluster sampling to reduce the variance value to be as equal as possible to the simple random variance value.

Based on the above, a relative coefficient shall be used to show the expected effect related to the usage of a cluster sampling design rather than using the simple random sampling technique. The relevant coefficient is known as the “Design Effect”.

Design Effect: refers to the variance ratio of the sample cluster design to the simple random sample design. It constitutes a measure of relative efficiency and shall be mathematically expressed using the following equation:

$$deff(\hat{\theta}) = \frac{Var_{cluster\ Des.}(\hat{\theta})}{Var_{SRS}(\hat{\theta})} \quad \text{..... (16)}$$

The above has helped conclude that the reason behind variance inflation in the cluster sampling technique is related to the cluster design. Therefore, the Design Effect may be defined as the inflation value of the estimated variance due to the adoption of the cluster design rather than the simple random sample.

On another hand, the number of sampling units to be selected from a single enumeration area (cluster size) is directly related to the value of the design effect according to the following equation:

$$deff = 1 + \delta_x \times (\bar{n} - 1)$$

Where:

$deff$ refers to the design effect;

δ_x = the value of intraclass correlation between population units in the Primary sampling unit (PSU)

$\bar{n}$ = the mean of the number of the secondary sample units selected from the main enumeration area.

It may be noticed from the aforementioned equation that the design effect value is directly proportional to cluster size as well as to the intraclass correlation value between the population units in the primary sampling unit.

Example (7):

In order to study the income average of the household in a certain population, based on similar previous surveys conducted for the same population, where two samples were randomly selected at that time: a cluster sample and a simple random sample. The variance ratio between the two designs (Design Effect) is given by Deff=1.4. The required sample size selected from the current survey to obtain an estimation for the average income of the household according to a confidence level of 95%, and margin error of d = 1.5, according to the simple random sample, shall be as follows:

$$n_{SRS} = \frac{z^2 S_{SRS}}{d^2} = \frac{1.96^2 370}{1.5^2} = 632$$

This means that the required simple random sample size within a confidence level of 95% is 632 establishments. In case of using a cluster sample design, the variance amount shall be significantly higher than that attributed to the simple random sample. accordingly, the sample size shall be modified to cover the variance value and maintain the estimation value within the bounds of error. In this case, the sample size shall be modified using the design effect as follows:

$$n_{cluster} = deff \times n_{SRS}$$

Since the value of the design effect is equal to 1.4, the size of the cluster sample shall be 632 multiplied by the design effect value, being 1.4, resulting in 885 secondary sampling units.

Chapter 4: Sampling Technique in Statistics Centre – Abu Dhabi

Within its framework Statistics Centre – Abu Dhabi is concerned with conducting all kinds of statistical surveys, economic, agricultural, environmental, and household surveys, to provide the Emirate of Abu Dhabi with all required official statistics. Furthermore, the SCAD role in designing and selecting samples that may be required from the various institutions and departments of the Abu Dhabi government.
The Statistics Centre – Abu Dhabi has different types of sampling frames and follows up on a periodic and continuous basis on the related updating operation. These frames are used to design and select statistical samples able to represent the statistical population with utmost efficiency and accuracy. Based on the concerning frames, the required sample size estimation procedures shall be carried out and statistical samples of all types may be selected.

1. Sampling Frames in the Statistics Centre – Abu Dhabi

The sampling frame represents a list of all the units of the population or may be presented as geographical maps showing all the units of the population. The frame list also includes geographical variables related to the addresses through which population units are inferred, as well as other technical variables that help in studying the nature and characteristics of the population units with the frame, such as the number of employees in establishment's frame or the number of households in the enumeration area under in the households and housing units frame.

Both, the establishment's frame and the households and housing units frame, are highly relied on in the design and selection process of household and economic survey samples selection in the Emirate of Abu Dhabi.

1.1 Establishments Frame

Content: the establishment frame includes a list of all operating establishments practicing one or several economic activities in the Emirate of Abu Dhabi according to their legal entity. An establishment may be the headquarter, a single establishment, or a branch that keeps the accounts of an establishment whose headquarters is based in the Emirate of Abu Dhabi, or the branch of the establishment that does not keep its accounts and its headquarters is located outside the Emirate of Abu Dhabi.

Stratification: establishments within this frame are divided into strata according to the economic activity ISIC-4 and the size of the establishment according to the number of employees according to the classification of economic establishments, as follows:

Manufacturing Industries Sector:
Micro-establishments (1 to 9) employees, small establishments (10 to 100) employees, medium-sized establishments (101 to 250) employee and large establishments (250 employee and above).

Other Sectors (trade, services, etc.):

Micro-establishments (1 to 5) employees, small establishments (6 to 50) employee, medium-sized establishments (51 to 200) employee and large establishments (200 employee and above).

On other hand the variables included in the sampling frame categorized as follows:

Type of variables	
Address Variables	Region, area, Plot No., location of the establishment in the building, establishment address, etc.
Characteristics Variables	Name of the establishment, license No., name of the owner, name of the Director General, Telephone number, etc.
Analytical Variables	Establishment status characteristics, legal entity, economic activity, establishment description, paid-up capital, revenues, etc.

Data Sources: The sampling frame of economic establishments was constructed based on a statistical register for economic establishments in the Emirate of Abu Dhabi.

Updating the economic establishment's frame: To keep pace with the coverage and inclusion process of all economic establishments being established or closing or modifying their economic activity, ongoing updating processes shall be carried out based on the results of the annual economic surveys carried out by the Centre. An annual update is carried out about the status of the establishments to record whether the same have ceased their economic activity or modified the latter. It may be even carried out to record the change in the number of employees therein in addition to updating the definitional and geographical variables related thereto.

In addition to the previous updating processes, the Centre is implementing the updating project of the economic establishments based on the statistical registers made available by the Abu Dhabi government institutions to update the lists of the frame.

1.2 Housing Units and Households Frame

Content: This frame includes a list of all the occupied housing units existing in the Emirate of Abu Dhabi.

Data Sources: The frame is built based on the available records of buildings, housing, and households, whereby all housing units comprise households.

Geographical Structure of the Frame: The geographical structure of the frame is in line with the administrative divisions approved by the Abu Dhabi Government, in addition to detailed statistical divisions used for sampling purposes.

As for the approved administrative division levels, they include the Region, the District, and the community. Whereas the statistical division levels include Enumeration Areas.

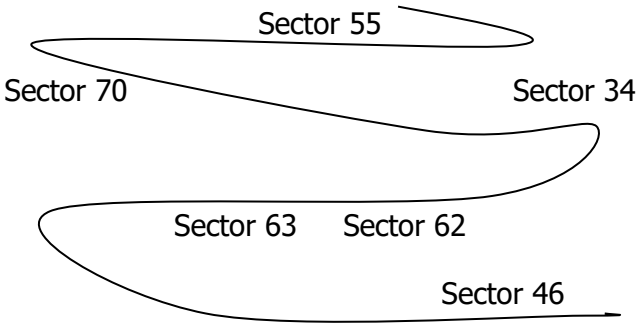
Definition: An Enumeration Area refers to a geographical area with natural or industrial borders. It includes building, housing units, and households, whereby the average number of households ranges between 100 and 200 households, with some exceptions for the areas extending over large spaces with a small population density.

Based on the above, the development and organization of the housing units and households frame is constructed in two stages, ensuring consistency with the statistical sampling designs that may be furtherly applied.

First Stage: Building Enumeration Areas based on geographical maps and data related to buildings, housing units, and households. The emirate of Abu Dhabi has been divided into independent and nonoverlapped Enumeration Areas.

Accordingly, a single enumeration area may form a single or part of a community. Sometimes, a group of sectors may be combined to form a single enumeration area, depending on the size and allocation of households and housing units in the area. It shall also be noted that these areas shall be adopted as Primary Sampling Units after the completion of the frame development process.

After the construction of the enumeration areas, the latter were arranged according to the aforementioned administrative divisions, within each region sectors were spirally arranged, from north to south, in order to ensure the geographical inclusion of the sample.



The arrangement of enumeration areas within each community was also spiral, as it includes the greatest spread within the sector, mainly in those including a large number of enumeration areas.

In addition to the list including all housing units and households within a single enumeration area, geographical maps exist to refer to the detailed locations of the buildings and housing units within a single enumeration area.

Second Stage: Preparing occupied housing units lists by households. The housing units and households framed within a single enumeration area include a list of detailed geographical names and addresses through which the housing units may be accessed as well as the households residing therein. In addition, the list includes the names of the heads of the families occupying the housing units as well as the type of household according to the nationality of the head of household (citizen household, non-citizen household, collective household). Accordingly, households are divided into two main types:

The Private Household: The private household consists of one or several members living in a housing unit and sharing their food. Among the members of the private households, only one is known as the head of the household, who shall be in charge of the living arrangements of the members. Usually, the household members' expenditure comes from the income of the head of the household. In the event the head of the household is a citizen, the household shall be referred to as a private citizen household. In a reverse condition where the head of the household is not a citizen, the household shall be referred to as a private non-citizen household.

The Collective Household: The collective family consists of a group of more than one member living together under a single housing unit, whereby no kinship relates any member to the other. The collective household does not have a head and its members do not share food or cooperatively spend their money. In this context, collective households shall be distinguished from labour camps, whereby the latter consists of large housing units managed by a specific institution or establishment providing job opportunities for the members residing in the camp. On the other hand, the members residing in the housing unit of the collective household are in charge of living arrangements as well as managing the housing they live in.

Updating the Housing Units and Households Framework

The process of ensuring the ongoing update of the sampling frame is the main priority upon which the accuracy of the frame is built, thus the efficiency and quality of the relevant surveys.

The updating frame process shall be carried out by two independent techniques that can be relied on according to the available capabilities, resources and timeframe.

The First Technique: it depends on the use of administrative records data, providing detailed information on the development of housing units and households residing therein. Water and electricity records, as well as other service records, provide data that may be used in the update of the new emerging enumeration areas after the establishment of the frame. New cities and housing units might have emerged after the census or in the event some housing units became occupied after being previously under construction. These new units may be added to the frame based on the administrative records data in order to increase the coverage levels.

The Second Technique: The second technique shall be carried out through a comprehensive updating of the enumeration areas in which the building, housing units, and the household residing therein information shall be updated. The update process is not mostly carried out according to the foregoing on all the enumeration areas, due to the time and costs that might be incurred as a result. Some countries tend to conduct partial update operations in the areas expected to have been subject to significant changes in terms of the establishment of new buildings, housing units and households.

On the other hand, another technique is used for the partial update of the enumeration areas located within the Master Sample, which is a relatively large sample selected to be used in several surveys. The Master Sample is divided into groups called "Replicates", where each replicate is selected independently

to represent the population as a whole. When conducting any survey, one or more replicates shall be chosen according to the selected survey sample size and the survey is carried out accordingly.
The Statistics Centre – Abu Dhabi updates the housing and buildings' current frame in accordance with the second technique by conducting partial annual field update processes for a specific percentage of enumeration areas in the Emirate. The update process also includes the addition of new emerging areas that were not existent at the time of preparation of the frame.

2. Designing Statistical Survey Samples

2.1 Annual Economic Surveys

Annual Economic Surveys aim to provide with economic indicators at the level of economic activities as well as representing the size of the establishment within a single activity (large, medium, small and micro establishments). The optimum sampling technique to be adopted in the design of the sample shall be the random stratified sampling technique, as the establishments are divided into independent strata:

- According to the economic activities, which may be at two or sometimes four ISIC digits.
- According to the size of the economic establishment which shall be comprised of 4 groups: large, medium, small, and micro establishments.

According to the foregoing, each stratum shall be considered an independent statistical population. Establishments belonging to a single stratum shall be determined according to the economic activity they carry out as well as the category of the number of employees (micro, small, medium or large establishments).

Sample Size Estimation

The sample size of the annual economic survey shall be estimated within each stratum according to strategies set out thereunder as follows:

- The sample size is determined at the level of a single stratum, being at the level of the economic activity and size of the establishments within the activity, while considering that all activities fall within the sizes of the various strata, mainly including 5 establishments or less, are completely included in the sample.
- for other activities and strata, the sample size is determined according to equations related to the estimation of the sample size based on the results of previously conducted statistical surveys, in such a manner that the margin error for the main sample variable, being the number of employees, ranges between 15% and 20%, except some strata characterizing some establishments with higher variance of the total employee’s variable where the margin of error may reach 25%. After determining the size of the sample within a single stratum, around 15% of total estimated sample size shall be added to substitute for the cases of non-response based on the results of the previously carried out economic surveys.
- The estimation of the sample according to the aforementioned criteria produces indicators related to the number of employees, employees’ compensation, total revenues, and total added value at the level of economic activity on two levels, with a margin error and at the level of the industry,

trade, services, and other sectors, whereby the margin of error does not exceed 15% for the main variables.

Sample Selection Technique

A sample of economic establishments is selected from each stratum by the proportional to the size sampling technique, where the size of a single establishment is expressed by the number of employees in the same.

The sample selection procedures within the stratum of large establishments slightly differs from other strata. It is well known that the concerned stratum comprises large economic establishments with an important economic weight, which is essential in economic calculations. At the same time, it cannot conduct a complete census due to its large number with a limited sample size. In this case, a single stratum shall be divided into two partial strata, whereby the first comprises the largest establishments in terms of economic activity and is completely selected in the sample. the rest of the establishments shall be comprised in the second partial stratum of which a random sample is selected in accordance with the proportional to the size sampling technique.

As for the mechanism adopted to divide the stratum related to large establishments into two partial strata, it relies on the allocation of the concentration of employees in the stratum, whereby the first partial stratum is composed of no less than 40% of the largest establishments in the stratum. This percentage differs in some activities according to the number and allocation of employees in the large stratum.

Sample Weights Calculation

It is well known that the sampling unit weight is the inverse of the probability of selecting the concerned unit from the sample. since the sampling design is stratified, the sampling weights for the sample establishments within each stratum will be calculated independently. The sample size of the economic establishments within a certain stratum is equal to the product of dividing the number of economic establishments in the frame by the number of responded economic establishments to the survey data.

As for the large establishments' stratum, mainly the first partial stratum, the weight of the economic establishment shall be equal to 1, meaning that it represents itself only, given that it has been selected rather than probably chosen. As for the rest of the establishments within the second partial stratum, the weight of each shall be calculated using the same previous technique.

2.2 Designing Household Surveys Samples

The frame of the household surveys has been developed and organized in line with the purposes of the concerned surveys as it ensures an efficient representation of the results. Being one of the basic procedures adopted when developing the frame, the household population in the Emirate of Abu Dhabi has been divided into four independent strata to ensure the minimum variance between the population units, which ensures reducing the sample size as much as possible while maintaining a high level of accuracy. The strata were divided as follows:

Stratum (1): it includes all the enumeration areas comprising citizen households representing less than 25% of the total households in each area.

Stratum (2): it includes all enumeration areas comprising citizen households representing 25% to 50% of the total households in each area.
Stratum (3): it includes all enumeration areas comprising citizen households representing 50% to 75% of the total households.
Stratum (4): it includes all enumeration areas comprising citizen households representing 75% to 100% of the total households in each area.
Based on the foregoing, the enumeration areas were arranged within the frame according to the following sequence: Region, District, community, Stratum, where it shall be classified within a single region (Abu Dhabi, Al Ain, Al Dhafra) into 4 strata according to the aforementioned classification. Accordingly, 12 implied strata were formed.
The sampling technique adopted in this case is the Stratified Two-Stage Cluster Sample Design. Within the first stage, a sample shall be selected from each stratum from the enumeration area, whereby these areas are known as the Primary Sampling Units. During the second stage, a sample of housing units occupied by households shall be selected from each Primary Sampling Unit already selected in the first stage.

Samples Selection in Household Survey

The selection procedure shall be carried out according to the Two-Stage Sample selection as follows:

- Enumeration areas are selected from the list of enumeration areas in the frame using the proportional to the size random sampling technique, which gives a greater chance of selection to the enumeration areas comprising the relatively large number of households to appear in the sample, which increases the levels of both, efficiency and quality.
- A household sample shall be selected from each enumeration area using the Systematic Random Sampling Technique, which ensures the allocation of the selected household sample over the geographical space of a single enumeration area. Accordingly, a relatively high variance shall be noticed between the households of the same enumeration area. At the same time, this technique provides a relatively low variance between the enumeration areas, which enables the role of this type of sampling design in providing high accuracy results.

References:

- Theory and Analysis of Sample Survey Designs, Caroga Singh, 1986.
- Sampling Techniques, William. Cochran, 1953.
- Glossary of Statistical Terms, Arab Institute for Training and Research in Statistics, 2005.
- Sampling Design methods in the application field, Dr. Mahdi Al Allaq, Adnan Shihab, 2001

مـركـز الإحـصـاء
STATISTICS CENTRE

الرؤية: ببياناتنا نمضي نحو غدٍ أفضل
Vision: Driven by data for a better tomorrow

www.scad.gov.ae
adstatistics