WORK IN PROGRESS

Sampling Guidelines for the Pacific
– Part 1 –

Table of Contents

Chapter 1 - Introduction to surveys and sampling

1. Surveys and sampling

1.1 Basic concepts and definitions

A ***survey*** refers to any form of data collection.

Elements (statistical units) are units from which information is sought and measurements are taken in a survey, and for which statistics are ultimately compiled. Examples of *elements (statistical units)* include: persons, households, schools, hospitals, businesses, farms, and geographic areas (such as Enumeration Areas, or EAs). There are different types of elements or statistical units, such as selection units, collection units, reporting units, and analysis units.

The ***target population*** for a survey is the population of elements (units) we are theoretically interested in surveying. The aim of a survey is to produce statistics that represent the whole of the *target population*, and often different sub-populations within it. The *target population* for a survey is assumed to be fixed and finite. For example:

- for HIES, the *target population* is people living in households (i.e. excluding institutionalized populations such as those in dormitories, boarding schools, prisons, military barracks, etc.).

A ***census*** refers to a data collection (or survey) that aims to collect data from the whole population of interest, i.e. from all elements in the target population.

A ***sample survey*** refers to a data collection (or survey) that aims to collect data from a subset or sample of the population of interest, i.e. from only some of the elements in the target population, but still make quantitative statements about the whole population. The objective of sampling is to estimate parameters of the whole population, such as the mean, total, or proportion, from only a part of the population. Parameters estimated by a sample survey must be accompanied by a measure of associated uncertainty, such as the standard error or confidence interval, that describes the expected accuracy of the sample survey compared to the true population parameter.

In general, there are two types of samples - ***probability*** and ***non-probability*** samples. Our focus in these guidelines is on ***probability samples.*** These are samples in which a known, non-zero probability of selection can be calculated for each element (statistical unit) in the sample.

Sampling method refers to the techniques used to select a sample (subset) of the target population to survey, and to produce statistical results (or estimates) from that sample. They are often referred to as "estimates" because the true value is not known, as not all units in the target population were surveyed. Examples of sample survey estimates from HIES include poverty rates and average income.
Practical considerations may dictate that some units in the target population are excluded from the survey (e.g., institutionalized individuals, the homeless, or those that are not be possible to access without incurring excessive cost). The ***survey population*** refers to the population of units that actually have a chance of inclusion in a survey.

Non-response (or unit non-response) arises when – during data collection - a certain number of respondents (businesses, farms, households, people) refuse to participate in the survey or are unable to be contacted. If the respondents that refuse to participate are systematically different in any way from those that choose to respond, non-response leads to bias. While there are statistical

techniques to minimise the impact of non-response, the only way to completely prevent non-response bias is to prevent non-response.

1.2 Social surveys versus Business surveys

Social surveys focus on topics about people and households - e.g. population statistics, labour force participation, household income and expenditure / consumption, poverty, education and health. Census, HIES, DHS, LFS, Gender based Violence Survey and MICS are all examples of social surveys which are common in the pacific.

Business / establishment surveys focus on topics about enterprises, establishments and other business units, including farms - e.g. business statistics / demographics, employment numbers, sales revenue, energy use, and agricultural production.

In general, ***social surveys*** and ***business surveys*** differ in terms of the topics they cover, and the types of elements (units) that make-up their target populations. They may also differ in other ways, such as use of different approaches to the survey, different sampling techniques, and different data collection modes.

2. Survey quality

2.1 Overall survey design

Still to be drafted

2.2 Sampling error versus non-sampling error

It is important that information on quality of a survey is published alongside survey results, and reported to survey users. There are different types of error that can occur during the design and operation of a survey, and impact on survey quality. Some of these errors may be a result of random effects, but some may result from systematic errors.

Broadly, there are two types of survey error:

- *Sampling error* refers to the error due to producing estimates for a target population based on a randomly selected sub-set or sample of population elements, rather than all population elements. It is a measure of the uncertainty resulting from using a sample survey instead of surveying all elements in the target population with a census. All sample surveys are subject to sampling error and there are mathematical formulas that can be used to accurately quantify the resulting uncertainty.

- *Non-sampling error* refers to the error in a survey from sources other than sampling error. There are various sources throughout the survey process that can introduce non-sampling error – for example: frame error, non-response error, measurement error, and processing error. Both censuses and sample surveys are subject to non-sampling error. Given than non-sample error can come from multiple sources and is often unobservable, it is not possible to precisely quantify non-sampling error.

Together sampling and non-sampling error are known as *total survey error*. It is important to minimize total survey error when designing a sample survey, and ensure both types of error are

controlled and reduced to an acceptable level. In many cases, there is a balance between sampling and non-sampling error. For example, a very complex sample design may minimize the expected sampling error, but if it is difficult to implement and leads to interviewer error in the field, the total survey error may be higher.

2.3 Variance versus bias

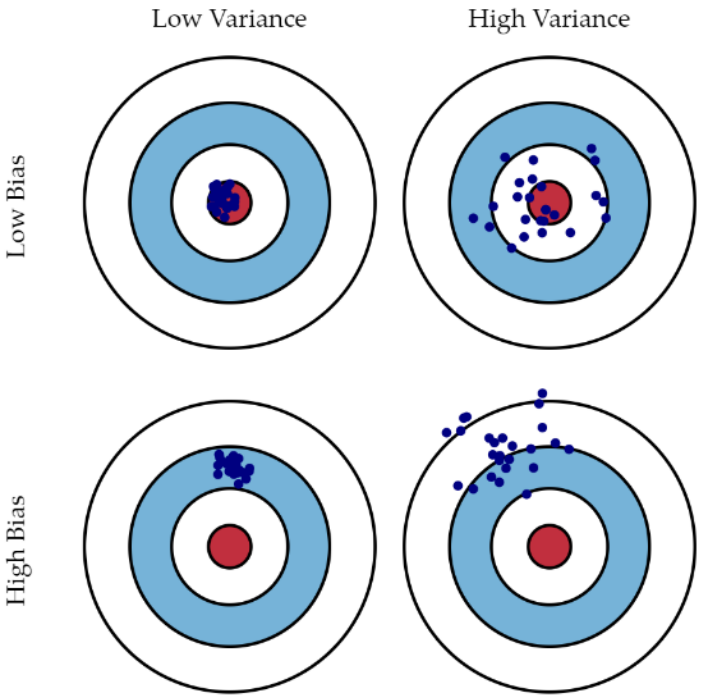
In survey sampling, estimators are preferred that fulfil certain theoretical properties.

One of these properties is ***unbiasedness***, meaning that the expected value of an estimator equals the true value of the population parameter being estimated, i.e. E(tˆ) ⍰⍰T. ***Bias*** is the difference between the true value and the expected value of the estimator, Bias(tˆ) = E(tˆ) – T, and when nonzero it represents systematic error. There are different types of bias, that can arise at different stages in the survey process, such as frame bias, selection bias, and non-response bias.

Precision of an estimator refers to its variability and is measured by the design or sampling variance Var(tˆ). The smaller the sampling variance, the better the precision of the estimator. A precise estimator is called ***efficient***, another desirable property.

The ***accuracy*** or ***total survey error*** of an estimator refers to the combined bias and precision properties of an estimator. It is measured by the mean square error (MSE), MSE(tˆ) =Var(tˆ) + Bias2 (tˆ) – i.e. the sum of the variance and squared bias.

The following diagram plots four different scenarios for an estimator, representing combinations of both low and high bias and variance:



Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

The top left hand scenario is the most desirable, given this estimator has both low bias and low variance, resulting in a low total survey error.

2.4 Quality measures for sample estimates

The standard error (s.e), coefficient of variation (c.v) and design effect (deff) of an estimator are commonly used quality measures of estimators. The quality measures are derived from the theoretical properties introduced above.

For an estimator $\hat{t}$ of population total, the measures are defined as follows.

> ***Estimated standard error*** - $s.e(\hat{t}) = \hat{v}(\hat{t})$, where $\hat{v}(\hat{t})$ is the estimated design variance or sampling variance of the total estimate $\hat{t}$.
>
> ***Estimated coefficient of variation or relative standard error*** - $c.v(\hat{t}) = s.e(\hat{t}) / \hat{t}$, i.e. the estimated standard error divided by the estimate itself. Coefficient of variation is often expressed in percentages, $100 \acute{} c.v\%$. Coefficient of variation is routinely reported in official statistics and used as a quality standard.
>
> ***Design effect (deff)*** - measures the statistical efficiency of a sampling design with respect to simple random sampling (SRS) and is given by $Var(\hat{t}) = \hat{v}(\hat{t}) / \hat{v}_{SRS}(\hat{t})$ where the numerator is the sampling variance of the total estimator under the actual (possibly complex) sampling design and the denominator represents the sampling variance under an assumption of simple random sampling of a sample of the same size.
> The formula for deff gives rise to the following remarks:

(a) deff < 1 The actual sampling design is more effective than SRS.
(b) deff = 1 The efficiency of the actual sampling design is similar to that of SRS.
(c) deff > 1 The actual sampling design is less effective than SRS.

2.5 Pros and cons of a census vs a sample

There are advantages and disadvantages to using a census or sample to produce statistics about a given target population:

Pros of a CENSUS	Cons of a CENSUS
• provides a true measure of the population (no sampling error) • benchmark data may be obtained for future studies • detailed information about small sub-groups within the population is more likely to be available	• may be difficult to enumerate all units of the population within the available time • higher costs, both in staff and monetary terms, than for a sample • generally takes longer to collect, process, and release data than from a sample
Pros of a SAMPLE	**Cons of a SAMPLE**
• costs would generally be lower than for a census • results may be available in less time • if good sampling techniques are used, the results can be very representative of the actual population • Can allow you to ask more detailed questions in the survey, if only a sample enumerated	• data may not be representative of the total population, particularly where the sample size is small • often not suitable for producing benchmark data • as data are collected from a subset of units and inferences made about the whole population, the data are subject to 'sampling' error • decreased number of units will reduce the detailed information available about sub-groups within a population – e.g. small geographical areas

Source: ABS website – *Statistical Language – Census and Sample* - http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+census+and+sample

The reality is that not all the data needs of a country can be met through census-taking; therefore, sample surveys provide a mechanism for meeting additional and emerging statistical needs on an on-going basis. In addition, since the logistics of a survey are generally less cumbersome than a census, the potential for non-sampling error is generally considered to be lower.

3. Sampling frames

3.1 Basic concepts and definitions

A *frame* is a list, map, or other specification of the elements (statistical units) which define a target population (whether via complete enumeration in a census, or via a sample survey).
Two commonly used types of frames in surveys are list frames and area frames.

For a sample survey, the term *sampling frame* may be used interchangeably with *frame.* In this case, the *sampling frame* is used to select population elements into the sample, and is also used as a basis for producing estimates for the population based on sample data.

Frame examples:

- List of establishments
- Census list of households
- Civil registrations

In a sample survey, there may be multiple stages of sample selection. In this scenario, multiple *sampling frames* need to be used or created, one for each stage of sample selection.

The *frame(s)* or *sampling frame(s)* used for a sample survey should be able to provide access to all the elements in the survey population, so that every element has a known and non-zero probability of selection into the sample.

3.2 Properties of a statistically sound frame

The *frame* for a survey should ideally:

- be exhaustive – i.e. list all of the units of analysis whose characteristics are to be measured in the survey - or at least provide very high coverage of the target population
- list each element of the target population once, and only once (i.e. no overlaps or duplicates)
- list only elements that are part of the target population (i.e. no irrelevant or erroneous elements)
- be up-to-date
- contain proper identification particulars for each element
- contain relevant and accurate information – referred to as *auxiliary information* - on each population element, such as size and other characteristics

For cost and other reasons, some elements in the target population may be removed from the *sampling frame*, so that they do not have a chance of selection into the survey. For example:

- people living on remote islands or inaccessible locations, which are difficult or costly to reach
- businesses which are very small, and do not contribute much to the variable of interest

Care must be taken when doing this, though, as deliberate exclusion of a part of the target population can lead to bias in the survey results. Any such exclusions from the sampling frame must be clearly documented and explained to users of the survey results.

4. Sampling design

In a sample survey, a probability sample is drawn from the frame population using a specified sampling design. Sampling designs can range from very simple through to quite complex. Typically in official statistics, sampling designs consist of a combination of various sampling techniques and sample selection methods, and may involve multiple stages of sampling.

4.1 Components of a sampling design

The key components of a sampling design are:

- the overall requirements and objectives of the sample survey – including a clear statement of the level of precision required for key estimates to be produced from the survey
- the overall sample size
- the stages of sampling (1 or more)
- within each stage:
    - the sampling frame
    - the sampling technique to be used - e.g. stratification, clustering
    - the sample size or fraction to be selected (allocation of the overall sample size)
    - the sample selection method to be used - e.g. SRS, sequential etc.

Chapter 2 introduces the basic sampling techniques and sample selection methods commonly used by Pacific NSOs - e.g. simple random sampling, systematic sampling and sampling with probability proportional to size (PPS).

4.2 Use of auxiliary information in sampling design

It is often useful to use auxiliary information on the population in designing a sample survey, and also in estimation procedures for such samples.

Auxiliary information is information obtained from pre-existing sources, and can be used either at the sampling stage (e.g. to create strata or clusters, sort frames, calculate measures of size etc.) or after data collection (e.g. to calculate weights, produce modelled estimates etc.).

To produce a more efficient sampling design, it is often desirable that auxiliary information should be:

- related to the variation of the variables of interest / data that will be collected in the survey
- available for every element in the frame population

Proper use of such auxiliary information can result in efficiency gains in the sampling design, as discussed in more detail in Chapter 2.

A challenge for survey statisticians is, for a given sample survey, to obtain efficient (precise) estimators whose design variances are as small as possible, whilst also managing the costs of the survey and the overall sample size.

4.3 Sample size determination

This section presents a general introduction to and overview of sample size determination. In practice, sampling designs and approaches for determining sample size vary depending on the type of survey (e.g. HIES vs. MICS, social surveys vs business surveys).

A key question often asked to surveys statisticians is: *what is the appropriate sample size for a sample survey?*

There is rarely a simple answer to this question. Rather, the survey statistician must consider the overall survey requirements and objectives, including:

- which are the most important study variables and parameters to be estimated from the sample?
- is there any existing information or knowledge (guess) about the statistical distribution of the study variables?

- what level of precision is required for the parameter estimates? how precise do the estimates of parameters from the survey need to be? - overall, and for different sub-groups of the population
- are there any specific factors that need to be taken into account in the survey? - e.g. special populations to be covered, certain analysis to be carried out etc.
- will the sample survey include complex sample design features such as stratification or clustering?
- what is the anticipated non-response rate for the survey?
- what are the financial (cost / budget) constraints for the survey?
- what are the time constraints for the survey?
- what is the size of the total target population?

All of these questions - and many other factors - need to be considered before a sampling design - including the overall sample size - can be determined. Oftentimes the above questions are not answered in sequence, rather the sample design is an iterative process that balances multiple objectives, desired precision, and available resources (human and financial) to arrive at the final design. Chapter 2 introduces the basic sampling techniques and sample selection methods commonly used by Pacific NSOs - e.g. simple random sampling, systematic sampling and sampling with probability proportional to size (PPS) – which are at the core of the design process. More detailed advice for specific surveys – including HIES, MICS, business surveys, agricultural surveys - are covered in the subsequent chapters of these guidelines.

5. Introduction to sampling notation used in these guidelines

Still to be drafted Chapter 2 - Overview of Sampling Theory

1. Simple Random Sampling (SRS)

1.1 Introduction to Simple Random Sampling (SRS)

Simple random sampling (SRS) is often regarded as the most basic form of probability sampling, and is applicable to situations where there is no previous information available on the population structure. Simple random sampling directly from the frame population ensures that each population element has an equal probability of selection, and thus, SRS is an *equal-probability sampling design*.

As a basic sampling technique, simple random sampling can be included as an inherent part of a sampling design – it provides the theoretical basis for more complicated techniques. In addition, simple random sampling sets a baseline for comparing the relative efficiency of a sampling design by using the design effect statistic which will be discussed later.

In simple random sampling of *n* elements, every element *k* in the population frame of *N* elements has exactly the same inclusion probability (π), that is:

$$\pi_k = \pi = {}^{n}/_{N}$$

In practice, SRS can be performed either without replacement (SRS-WOR) or with replacement (SRS-WR). WOR type sampling refers to the case where a sampled element is not replaced in the population; this also means that a population element can be sampled only once. In a WR scheme, a sampled element is replaced in the population. In both cases, the probability of selection $\pi = n / N$ remains.

1.2 How many different samples can you generate using SRS-WOR

For a population of N and a given sample of n, the total number of possible different samples is:

$$\text{Number of possible different samples} = \frac{N!}{(N-n)!\,(n)!}$$

Which for a small population of N = 10 and a sample of n= 4, will result in just 210 different sample possibilities, as such:

$$= \frac{10!}{(10-4)!\,(4!)} = 210$$

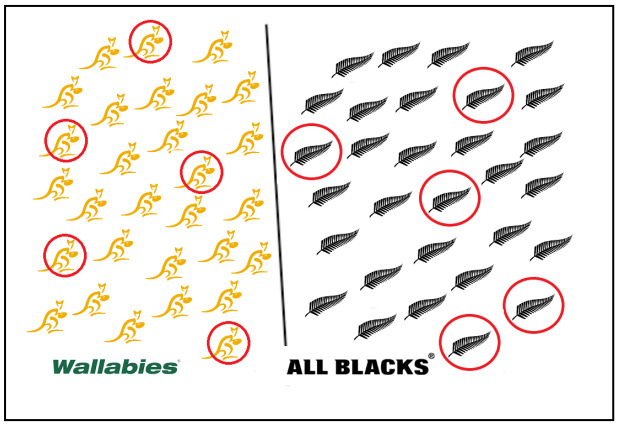
However, if we just increase the population a little, so now N = 100, and we select a sample of n = 30, then we find the number of different samples grows enormously, as such:

$$= \frac{100!}{(100-30)!\,(30!)} = 2.94\ x\ 10^{25}$$

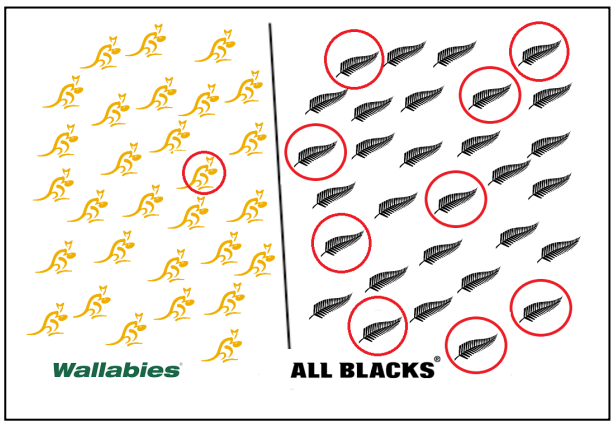
1.3 How representative is a SRS normally?

Given the random nature of the selection process, SRS normally generates samples close to the true population, especially for samples of significant size. However, because this cannot be controlled, this is not always the case. In the example below, we have two villages, each with a population of 30, one all Wallabies supporters, and the other all All Blacks supporters. Say you wished to select a sample of 10 persons using only SRS from the population of combined villages in order to estimate which national team gets more support. Whilst Scenario 1 would occur more frequently (5 persons from each village selected) and you thus generate a nice representative sample, Scenario 2 will happen from time to time (only 1 person selected from the village of Wallabies supporters), resulting in an outlier sample, and thus a poor estimate.

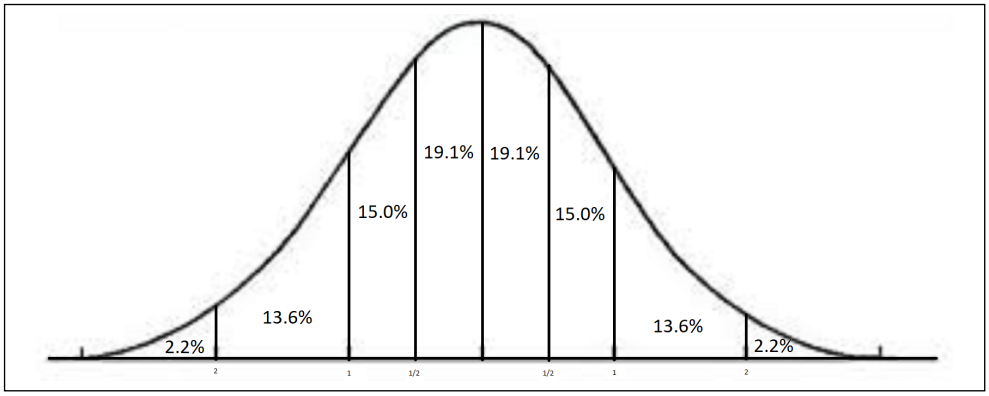
Scenario 1: Representative sample



Scenario 2: Outlier sample



Outlier samples, however, are rare. Using the formula above, we calculate that it is possible to draw 75,394,027,566 distinct samples of 10 rugby supporters from a population of 60. Of these, only 0.005 percent would be all Wallabies or all All Blacks supporters. Sampling theory is used to predict the likelihood of obtaining an outlier draw. The figure below shows the standard normal distribution of a sample. The highest point on the bell curve is the (true) 50% scenario and the most common outcome of the SRS procedure.



In terms of the Wallabies – All Blacks example from above, sampling theory indicates that 38.2 percent of all of the possible samples will estimate the proportion of Wallabies supporters between 42 and 58 percent, or within ½ of a standard deviation from the mean. 68.4 percent of all possible samples will estimate the proportion of Wallabies supporters 34.1 and 65.8, corresponding to one standard deviation, and 83.4 percent of all possible samples between 18.3 and 81.6 percent, or two standard deviations. As we will see in the sample size calculation section later, these precisions figures depend on the prevalence in the population and the size of the sample. A larger sample would generate more reliably precise estimates because the standard deviations would be smaller.

1.4 Estimation using SRS

Under SRS, an estimator of the target parameter for estimates of "total", "mean" and "proportion" are as follows:

1.4.1 Estimate of Mean

The estimate of the mean is the sum of the value of all the elements in the sample divided by the number of elements in the sample.

$$Est(\bar{Y}) = \sum_{i=1}^{n} \frac{y_i}{n}$$

1.4.2 Estimate of Total

The estimate of the total is the mean multiplied by the total population size. For example, if the average household size is 5 members, the total population would be 5 multiplied by the number of households.

$$Est(Y) = N \sum_{i=1}^{n} \frac{y_i}{n}$$

1.4.3 Estimate of Proportion

The estimate of a proportion is the sum of all of the elements in the sample with the characteristics divided by the total number of elements in the sample.

$$Est(Y_P) = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{where} \quad y_i = \begin{matrix} 1 & \text{if sample unit has characteristic} \\ 0 & \text{otherwise} \end{matrix}$$

1.5 Summary of advantages and disadvantages in using SRS

Advantages	***Disadvantages***

• It's simple - just need to generate a random number and use this for selection (provided you have a complete list of units)
• Generally produces low Sampling Errors
	• Can be costly if the sample is well spread out geographically
	• Can't control the representativeness of the sample
	  ◦ Can't control the sample for sub-populations
	  ◦ Sample may be highly skewed to one area

2. Systematic Sampling

2.1 Introduction to Systematic Sampling

Systematic sampling is a technique commonly used in the Pacific and – like SRS – is a type of *equal-probability sampling design*. The key element of systematic sampling is to skip through a list of elements with a constant interval each time, so two crucial bits of information are required; i) the constant interval, and ii) where to start.

The approach for applying systematic sampling in practice differs based on whether N/n is an integer, and as such, more than one approach will be addressed in this section.

2.2 Applying systematic sampling when N/n is an integer

The steps in the selection of a systematic sample of n elements from a population of N elements, when N/n is an integer, are as follows:

1. Define the sampling interval q = N/n, where an integer q is assumed.
2. Select a random integer "a" with an equal probability of 1/q between 1 and q [nb: in excel, this can be achieved by rounding up the result of "=rand()*q"]
3. Select elements numbered a, a + q, a +2q, a +3q,..., a + (n−1)q in the sample.

2.3 Applying systematic sampling when N/n is not an integer

There are two approaches to achieving your sample when N/n is not an integer; i) work with decimal places and round, or ii) apply circular sampling.

2.3.1 Work with decimal places and round

In the example below we can see that N/n is no longer an integer, and produces the result 32/7 = 4.571429. Systematic sampling can still be comfortably applied to this scenario if you are OK with using numbers with decimal places.

In this situation, your skip interval will still be N/n = 4.571429, and you will now be required to select a random start (no longer an integer) between 0 and 4.571429. In excel this can be achieved as follows "=rand()*4.571429". In the example below, the result was 2.636695.

The selection numbers are then achieved by adding the skip interval continuously to the random start until you have your required number of selections, n = 7. The units to be selected in the list are then identified by rounding up these selection numbers [in excel this is "=roundup(2.636695,0)"], resulting in 3, 8, 12, 17, 21, 26 and 31.

N	32	
n	7	
skip	4.571429	
R.Start	2.636695	
Sel1	2.636695	3
Sel2	7.208123	8
Sel3	11.77955	12
Sel4	16.35098	17
Sel5	20.92241	21
Sel6	25.49384	26
Sel7	30.06527	31

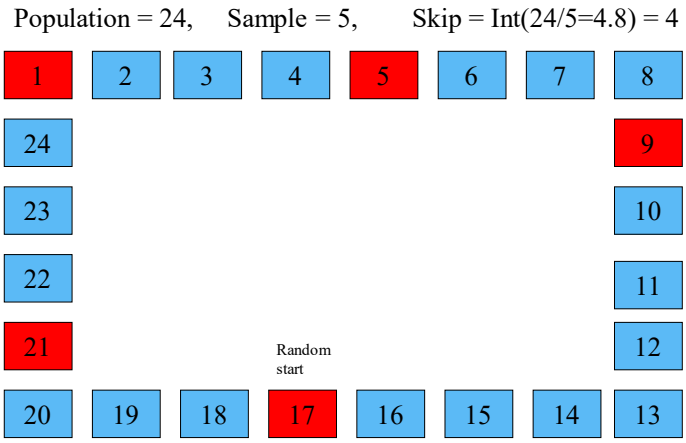
=32/7
=RAND()*4.571429
Same as the random start
=roundup(2.636695,0)

2.3.2 Apply Circular Sampling

Circular sampling is another way to select a systematic sample when N/n is not an integer.
The first thing you need to do is treat the list as a circle, so when you reach the end, you go back to the start. The approach is then applied in the following steps

1) Determine the interval ***k*** – rounding down to the integer nearest to ***N/n*** (If ***N*** = 24 and ***n*** = 5, then ***k*** is taken as 4 and not 5)
2) Take a random start between 1 and ***N***
3) Skip through the circle by ***k*** units each time to select the next unit until ***n*** units are selected
4) Thus there could be ***N*** possible distinct samples instead of ***k***

See the figure below for an example of a systematic sample chosen using circular sampling where N = 24 and n = 5. The random start was 17, and the skip used 4.



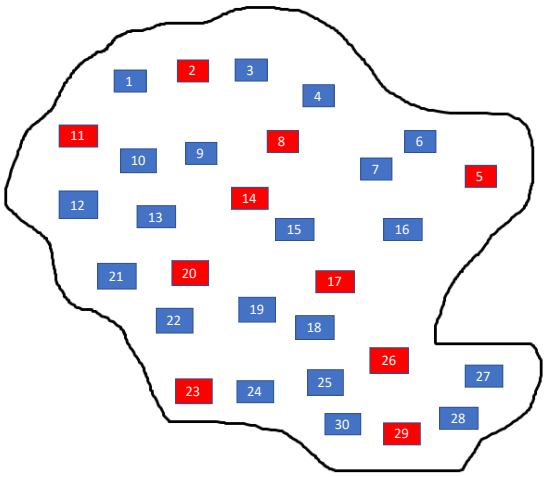
2.4 The value of sorting the list before selecting a systematic sample

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

If the sampling frame is sorted by an auxiliary variable (or, several such variables), systematic sampling will produce a sample which tends to mirror correctly the structure of population with respect to the variables used in sorting. This is desirable, as it will help in achieving a more representative sample. Sorting the frame before systematic sampling is called implicit stratification. For example, in some cases it is a good idea to sort the frame according to the regional population structure. Then a systematic sample will retain the appropriate population distribution across regions.

In the Pacific, systematic sampling is widely used when selecting households within previously selected small geographical areas. When this is the case, the households are often listed by geographical position, and as such, when a systematic sample is selected from that area, good geographical representation is often achieved. For surveys such as the Gender Based Violence surveys conducted in a number of Pacific countries, it is desirable to not select households next to each other for safety reasons. Systematic sampling can help reduce the likelihood of this occurring.

See below for a graphical representation of how a household listing may have been carried out in a village, and thus how a systematic sample helped ensure good representation of households throughout the village. In this diagram you can see how the households were numbered in the list, with a sample of 10 households chosen from the list of 30. The 10 households chosen in the village and highlighted in red, are nicely spread-out throughout the village, which is desirable.



2.5 Estimation using Systematic Sampling

Under systematic sampling, an estimator of the target parameter for estimates of “total”, “mean” and “proportion” are the same as SRS, and are thus as follows:

2.5.1 Estimate of Mean

$$Est(\bar{Y}) = \sum_{i=1}^{n} \frac{y_i}{n}$$

2.5.2 Estimate of Total

$$Est(Y) = N \sum_{i=1}^{n} \frac{y_i}{n}$$

2.5.3 Estimate of Proportion

$$Est(Y_P) = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{where} \quad y_i = \begin{matrix} 1 & \text{if sample unit has characteristic} \\ 0 & \text{otherwise} \end{matrix}$$

2.6 Summary of advantages and disadvantages in using Systematic Sampling

Advantages	***Disadvantages***
• Can sort the list prior to systematic sampling to achieve a more represetative sample • Generally produces low Sampling Errors Can be applied easily in the field	• Can be costly if the sample is well spread out geographically

3. Stratified Sampling

In actual sample design, simple random samples are very rare. Instead, most sample designs include elements of a *complex sample design*, or ways in which an SRS design is enhanced for greater accuracy or reduced cost. The most common element of complex sample design is stratification. As can be seen below, it has many benefits, which is why it is almost always adopted in official national surveys in all countries around the globe, including the Pacific.

3.1 Introduction to Stratified Sampling

S*tratified sampling* involves dividing the target population into non-overlapping subpopulations called *strata*. These strata are regarded as separate populations in which sampling of elements can be performed independently. Within the strata, some of the basic sampling techniques, SRS, systematic, etc, are used for drawing the sample of elements. Stratification allows flexibility because it enables the application of different sampling techniques for each stratum.

In stratified sampling, the population is divided into *H* non-overlapping subpopulations of size N_1, N_2, …., N_h, …..,N_H elements such that their sum is equal to *N*. For stratification, auxiliary information is required in the sampling frame. Regional, demographic, and socioeconomic variables are typical stratifying variables. A sample is selected independently from each stratum, where the stratum sample sizes are n_1, n_2, …., n_h, …, n_H elements, and their sum is equal to *n*, the overall sample size.

Common examples of stratified sampling include:

- For establishment surveys – stratification by economic activity and by employment size.
- For household surveys – stratification by geographic areas (e.g. regions, provinces), by urban / rural, and by socio-economic groups.
- For agricultural surveys – stratification by agro-ecological zones, by land use and by farm size.

In general, there are several reasons for the popularity of stratified sampling:

1. Preventing a weird or outlier draw by pure chance by more tightly controlling the selection.
2. Guaranteeing representation of small subpopulations or domains in the sample if desired.
3. Improving efficiency by dividing the population into homogeneous stratum (similar in nature) with respect to the variation of the study variables.
4. Allowing for flexible stratum-wise use of auxiliary information for sampling. For example, SRS sampling could be adopted in one stratum, and systematic sampling in another (although this isn't common).

If dividing the sample into strata leads to differing probabilities of selection between stratum, the analyst will need sampling weights to generate representative estimates.

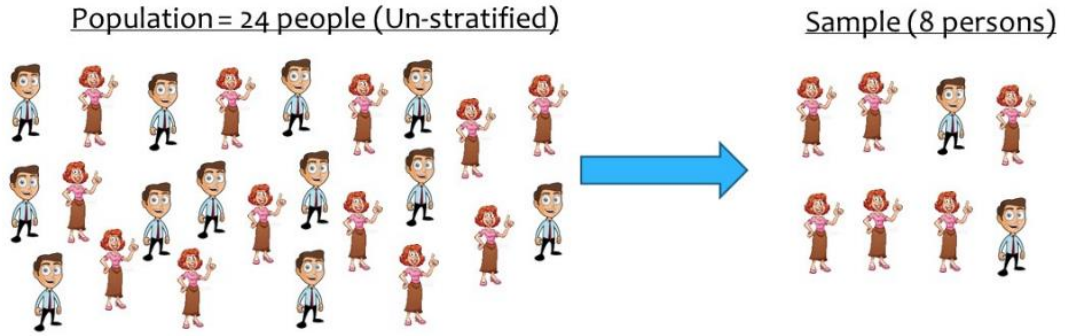
3.2 Applications of stratified sampling

3.2.1 Example 1: Stratifying by sex

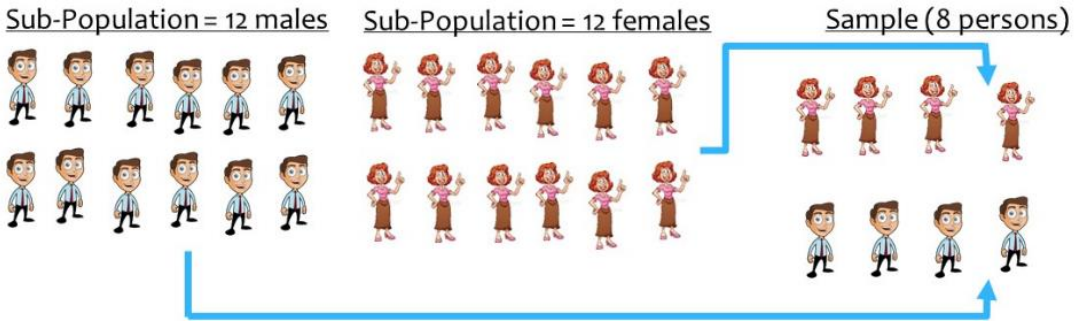
As discussed above in the reasons why stratified sampling is so popular, in point 1 it mentions the benefit of being able to protect against outlier draws. In the example below we have a simple population of 24 people (12 male and 12 female), for which we wish to sample 8 people in total. In scenario 1 we have the population not stratified, and thus cannot control the number of males and females in the sample with SRS. As a result, we may end up with only 2 males and 6 females in the resulting sample, thus having too many females compared to the true.

In scenario 2, we now stratify our population in to two sub-populations (12 male and 12 female). Using stratification, we can now select the samples separately in both, and intentionally select exactly 4 males and 4 females from each population, and thus guarantee better representation of each sex.

Scenario 1 – not stratified



Scenario 2 - stratified



3.2.2 Example 2: Stratifying by urban / rural location

In practice, reasons 2 - ensuring particular subgroups are adequately represented – and 3 – potentially reducing sampling error by controlling within stratum variance – are often contradictory or competing objectives. Take the example of a country that has two regions: an urban city with 75 percent of the national population and a diverse international economy, and rural areas with the remaining 25 percent of the population that are characterized by more traditional livelihoods and higher levels of hardship. If no stratification was used, it is expected that 75 percent of the population would be in the urban sample and 25 percent in the rural sample. If, however, there is little variation in the rural areas, it may be possible to achieve greater precision at the national level by shifting more of the sample in the diverse urban areas. Conversely, if we are interested in having a large enough sample size to closely study the rural population, we may want to shift more sample into rural areas to have greater precision there. Weighing these trade-offs is at the heart of sample design for complex surveys and there is no one right or wrong answer.

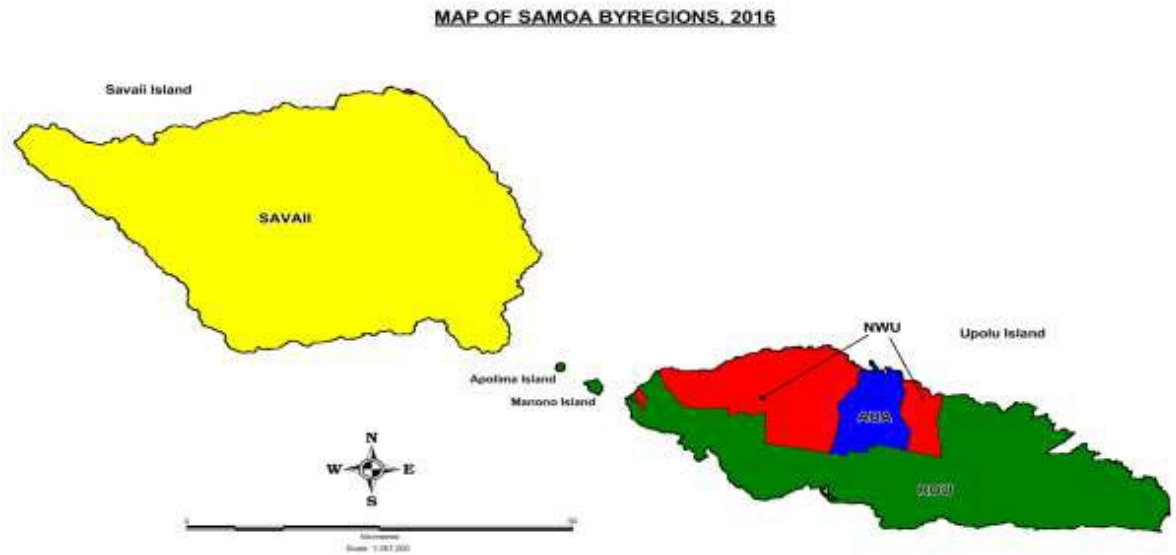
3.2.3 Example 3: Stratifying by administrative region

Samoa

Another common application of stratification in the Pacific is to stratify by geographical region. As illustrated in the diagram below, Samoa usually apply four levels of stratification to their household surveys which cover:

1. Apia Urban Area (Blue region)
2. North West Upolu (Red region)
3. Rest of Upolu (Green region)
4. Savaii (Yellow region)

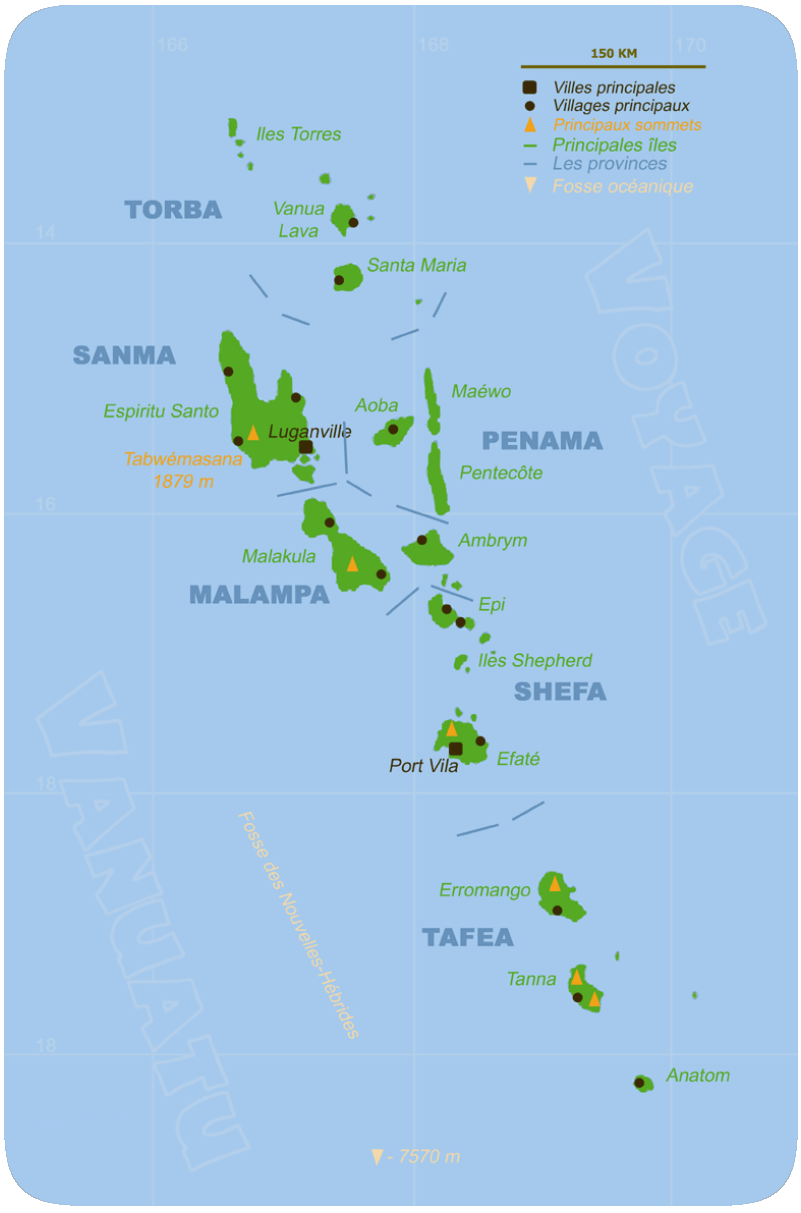
Sampling approaches are then applied independently within each of these 4 strata. The additional benefit of this stratification is that if results are required for urban/rural breakdown, this can easily be achieved with "Apia Urban Area" representing the urban population, and the remaining three strata representing the rural population.



Vanuatu

The map below shows the six main provinces which make up Vanuatu. For many household surveys covered in Vanuatu, the National Statistics Office adopt eight strata, with the provinces of Sanma and Shefa being split in to two (urban and rural), with the urban centers being Luganville and Port Vila respectively.

Once again, as with Samoa, results for urban and rural can be easily be created by combining Luganville and Port Vila together to form the urban population, and the remaining six strata (Torba, Sanma-rural, Penama, Malampa, Shefa-rural and Tafea) forming the rural population.



Differing sample fractions

In both the Samoa and Vanuatu examples, as also alluded to in 1.3.1 above, differing sample fractions may be adopted for different strata, to ensure enough sample is allocated to some of the smaller strata. For example, in recent documentation produced for the Vanuatu National Baseline Survey, the suggested sample size for the Torba province, with a population roughly 1,960

households, was 320 households, which is approximately 16 percent of the total number of households. For Malampa, one of the larger provinces with a population of roughly 8,900 households, the suggested sample size was 480 households, which is approximately 5 percent. These differing sample fractions ensure results of equal suitability can be achieved for all sub-populations of interest in the survey but require that sampling weights be used in the analysis.

3.3 Estimation using Stratified Sampling

In the following estimation formula, addressing both an estimate of total and mean, the symbol "h" will be used to notify stratum h, and "H" will be used to notify the total number of strata in the population. The formula below assume either SRS or systematic sampling is applied within each stratum.

3.3.1 Estimate of Total

$$Est(Y) = \sum_{h=1}^{H} Est(Y_h)$$

where

$$Est(Y_h) = N_h x \sum_{i=1}^{n_h} \frac{y_{h,i}}{n_h} = \sum_{i=1}^{n_h} w_h y_{h,i} \quad where \quad w_h = \frac{N_h}{n_h}$$

3.3.2 Estimate of Mean

$$Est(Y) = \frac{\sum_{h=1}^{H} N_h \overline{Y_h}}{N}$$

where

$$Est(\overline{Y_h}) = \sum_{i=1}^{n_h} \frac{y_{h,i}}{n_h}$$

3.4 Allocation of the sample across strata

There are a nearly infinite number of ways to allocate observations across the strata. Four of the more common examples are discussed below.

3.4.1 Proportional allocation

In a proportional allocation, the sample allocated to each stratum is proportionally to the number of units in the frame for the stratum. This method is the simplest form of sample allocation and does

not require sample weights for the analysis if the units within the stratum are chosen with SRS or systematic random sampling.

The formula for a proportional allocation is $n_h = n \times \frac{N_h}{N}$, where *n* is the sample size and *N* is the population size, with *h* denoting the stratum.

3.4.2 Equal allocation

In equal allocation each stratum is allocated an equal number of sample units, $n_h = \frac{n}{H}$, where the notation is the same as in a proportional allocation and *H* is the total number of strata. Equal allocation leads to differing probabilities of selection if the strata have different total populations, and therefore sampling weights are required for the analysis.

3.4.3 "Optimal" or Neyman allocation

A Neyman allocation divides the sample across strata in such a way as to minimize the national level standard error. Neyman was a very common method of sample allocation historically but is used less currently as surveys often have multiple objectives. For example, if in the past a survey was used to measure average national income, now additional objectives may include sub-national estimates and estimating SDG indicators nationally and for sub-groups. Still Neyman serves as a common starting point for many designs and therefore is one of the most common used tools in sample design.

The formula for a Neyman allocation is $n_h = n \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}}$, where S_h is the standard deviation and c_h is the cost in stratum *h.* As the cost is generally not known at the stratum level, this formula most commonly simplifies to $n_h = n \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$. Similar to equal allocation, the differing probabilities of selection across the strata necessitate the use of sampling weights.

3.4.4 Practical allocation

In current national household survey design, a practical allocation is the most common. A practical allocation does not follow a formula but rather divides the units of selection across the strata to meet multiple objectives. As long as the probabilities of selection are known, sampling weights can be calculated and representative estimates generated, even if there is not a single formula for the design. A common example of a practical allocation is to start with a Neyman allocation and increase the sample size in certain stratum until all reach a minimum level of precision.

4 Multi-Stage Sampling

4.4 Introduction to Multi-Stage Sampling

Multi-stage sampling as its name suggests, involves more than one stage of selection to the sampling approach adopted. For household surveys it is common to adopt a 2-stage process of selecting households, where small geographical areas are selected at the first stage (primary selection units – PSUs), and then a sample of households chosen from within each selected geographical area. Multi-stage sampling is used in nearly all official household surveys across the world. These PSUs are sometimes known as "clusters," so multi-stage sampling is also known as clustering. The Pacific is unique in that some of the smaller countries do not select PSUs, but instead sample households with stratum directly using SRS or systematic random sampling. As we will see below, SRS is a more efficient approach statistically, but it has a higher cost and more complicated logistics, leading it to be practical in only rare cases.

There are a couple of key reasons why multi-stage sampling is often used in practice:

- A list of all PSUs may be available for the population of interest, but not a list of all final units (eg, households). The first stage of the sampling process is to select a sample of PSUs, and then a list can be constructed of all final units in the selected PSUs.
- In face-to-face surveys, it is less expensive to concentrate the sample in selected areas, rather than spread the sample out over the entire country.

In the Pacific, all countries currently undertake a population and housing census, with frequency every 5-10 years. During this exercise, the population is divided in to what are commonly referred to as Enumeration Areas (EAs), which generally speaking, are the equivalent to a workload for one enumerator during the census data collection period. A list of these EAs is produced and maintained between censuses, and generally used as the PSU for household surveys in a two-stage sample design for countries following a multi-stage methodology. A sample of households is then selected within each EA, generally using systematic sampling to complete the second stage of selection. This two-stage process will normally take place once the survey strata have been identified and be performed within each stratum independently.

4.5 Surveys in the Pacific with a 3rd stage of selection

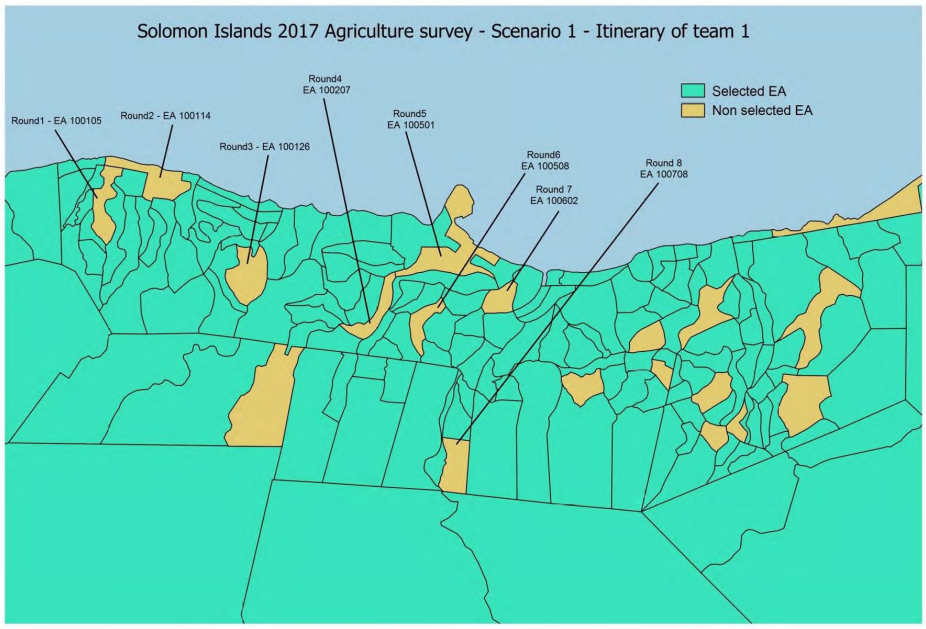
Although many household surveys in the Pacific require just two stages to the selection process, there are situations when just one member of the household may be required to be sampled, adding a third stage to the selection process. One such example of this is the Family Health and Safety Survey (sometimes referred to as the Gender based Violence Survey), where just one female in scope of the survey is required to be part of the survey, and thus needs to be selected randomly from the list of household members eligible to participate.

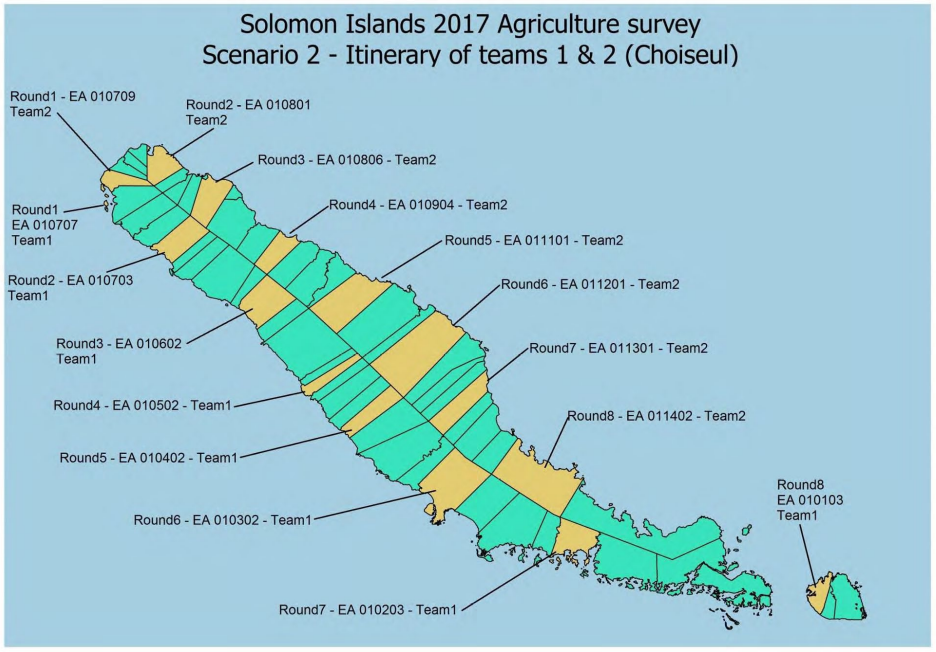
4.6 Solomon Islands: Example of 1st stage of selection (EAs)

Below is an example of the selection of first stage units for an Agriculture Survey in the Solomon Islands. The first map shows the selection of EAs in Honiara, which are represented by the light orange patches (selected EAs). Given the EAs were listed based on geographical position and selected using a form of systematic random sampling, the sample is dispersed across Honiara.

In the second map shows a similar exercise on one of the outer island provinces of Solomon Islands – Choiseul. The same approach was taken for the selection of PSUs, and the results are similar in terms of the dispersion of survey areas across the geography.

Within each selected EA, an updated list of households in produced and systematic sampling used to select the required sample size of households.





4.7 Introduction to Probability Proportional to Size Sampling (PPS)

Probability proportional to size (PPS) sampling is a special case of multi-stage sampling in which larger PSUs have a higher probability of being selected. Another way of stating this is that the probability of selection depends on the size of the population element. It is assumed that the value Z_k of the auxiliary size variable *z* is known for every population element *k*. Typical size measures are variables that physically measure the size of a population element. In business surveys, for example, the number of employees in a business firm can be used as a measure of size, and in a school survey the total number of pupils in a school is also a good size measure. In the case of household surveys conducted in the Pacific, PPS sampling often uses for the number of households as the size measure.

The number of households, rather than the total population, is more commonly used because households are the elements that are selected for the survey. In surveys where individuals are being selected, it would be more common to use the total population as the measure of size.

In PPS sampling, the probability of selection for a unit in a *n* unit sample is

$$\pi_k = np_k = n\frac{Z_k}{T_z}, \quad where \quad T_z = \sum_{k=1}^{N} Z_k$$

Where T_z is the sum of size measures over the N units in the population, and p_k is the single draw selection probability. In PPS, the probabilities of selection π_k vary between units and thus, PPS is an unequal probability sampling design and sampling weights are required for the analysis.

4.8 Examples of Probability Proportional to Size Sampling

4.8.1 PPS sampling in household surveys

Many household surveys conducted in the Pacific select EAs in the first stage using PPS sampling and then households in the second stage from updated administrative lists or a household listing operation. The process for undertaking this stage of selection in practice is described in the example below.

Step 1
Sort or order the 26 EAs representing this stratum by geographical position. This ordering can be east-to-west, north-to-south, or any other sorting that keeps nearby EAs close together in the list.

Step 2
Assign the size measure (in this case the number of households) to each listed EA, with a cumulative count alongside this value (the cumulative count will be used a little later to identify each selection).

Step 3
To the right you will see the calculations undertaken to determine the skip and random start for identifying the selections. In this example it has been pre-determined that six EAs will selected from the stratum, so the skip can be calculated as 1,365/6 = 227.5. The random start has then been generated between 0 and 227.5 and generated as 14.04546.

Step 4
Determine the six selection numbers required to select the six EAs in the sample. This is achieved by firstly assigning the random start as the first selection number (rounded up to the nearest integer), and then adding the skip five more times to produce the remaining five selection numbers (all rounded up). The resulting selection numbers were 15, 243, 471, 699, 927 and 1155.

Step 5
Assign the selection numbers to the EAs in the list (highlighted in red). This is done by choosing the EA which has a “Cum # HHs” greater than the selection number, for which the previous EA has a “Cum # HHs” less than the selection number. For example, the second selection in the

example below was EA 100105 – it was selected because the selection number 243, was below 281, but greater than 237.

EA	# HHs	Cum # HHs	Selection		EA	# HHs	Cum # HHs	Selection
100101	43	43	15		100301	48	743	699
100102	81	124			100302	38	781	
100103	52	176			100303	71	852	
100104	61	237			100304	55	907	
100105	44	281	243		100305	51	958	927
100106	38	319			100306	41	999	
100201	72	391			100307	49	1048	
100202	49	440			100308	73	1121	
100203	47	487	471		100309	48	1169	1155
100204	33	520			100310	39	1208	
100205	61	581			100311	32	1240	
100206	63	644			100312	67	1307	
100207	51	695			100313	58	1365	

Number of HHs	1365
Number of EAs	26
Number of EAs to select	6
Skip	227.5
Random Start	14.04546
Seln 1	15
Seln 2	243
Seln 3	471
Seln 4	699
Seln 5	927
Seln 6	1155

4.8.2 PPS sampling in ??? (include an example from a different type of survey – Business survey? Agricultural survey?)

4.9 Selecting a fixed cluster size of households at the second stage

As discussed above, the example for the Solomon Islands used PPS sampling to select the required number of EAs from each of the stratum, two of which were Honiara (urban population) and Choiseul (one of the rural stratum). Once the EAs have been selected using PPS sampling for this stage, it is common practice to select a fixed number of households from each selected EA, despite the size of the EA. Whilst the approach of selecting a fixed number of households from each selected EA helps with allocating even workloads across field staff, it also has other benefits in that it helps give households a similar chance of selection in the survey. This feature can be seen in the example below.

In the table below, a stratum has 15 EAs, of which we have selected 5 EAs during the first stage. Now presume we decide to select a fixed number of households per EA, say 15 – we refer to this number as the cluster size. Adopting this approach gives each household an equal probability of selection, if the size of the EA (with respect to number of households) does not change.

Computing the probability of selection of a household being selected in the survey:

Pr (HHold selected in the survey) = Pr (EA selected in Stage 1) x Pr (Hhold selected in Stage 2)

In the example below, for the 5 selected EAs, the probabilities of a household being selected from those EAs is calculated as follows:

- EA10403: Pr (EA sel in St1) x Pr (Hhold sel St2) = (47/127.4) x (15/47) = 15/127.4 = 0.1177
- EA10405: Pr (EA sel in St1) x Pr (Hhold sel St2) = (56/127.4) x (15/56) = 15/127.4 = 0.1177
- EA10408: Pr (EA sel in St1) x Pr (Hhold sel St2) = (42/127.4) x (15/42) = 15/127.4 = 0.1177
- EA10411: Pr (EA sel in St1) x Pr (Hhold sel St2) = (37/127.4) x (15/37) = 15/127.4 = 0.1177

- EA10415: Pr (EA sel in St1) x Pr (Hhold sel St2) = (40/127.4) x (15/40) = 15/127.4 = 0.1177

EA	# Hhs	Cum Hhs	Selection
10401	34	34	
10402	56	90	
10403	47	137	91
10404	29	166	
10405	56	222	219
10406	47	269	
10407	51	320	
10408	42	362	347
10409	51	413	
10410	32	445	
10411	37	482	475
10412	34	516	
10413	46	562	
10414	35	597	
10415	40	637	603

# Hholds	637
EA total	15
EA select	5
Skip	127.4
R.Start	90.79037
Seln 1	91
Seln 2	219
Seln 3	347
Seln 4	475
Seln 5	603

As can be seen from the calculations above, all households have the same chance of being selected in the survey (0.1177), regardless of the EA to which they belonged. This balancing is because larger EAs have more chance to be selected during the first stage of selection, as PPS was adopted, but households within those EAs then have less chance of being selected because a fixed number of households in selected.

In practice however, household lists for each selected EA are generally updated, and the fixed cluster of households selected during the second stage of selection are chosen from this updated list. When this occurs, the probabilities of selection for each household will be the same within an EA, but different across EAs. For example, in the illustration above, if EA 10403 had its household listing updated once it was selected and there was found to be 52 households now in the list, then the probability of a household from that EA being selected would now be:

- EA10403: Pr (EA sel in St1) x Pr (Hhold sel St2) = (47/127.4) x (15/**52**) = 0.1064

which is a little bit smaller than before because there is less chance of selection at the second stage with the extra households identified.

4.10 Estimation using two-stage sampling

4.10.2 Estimate of Total

$$Est(Y) = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

where

$$\pi_i = Prob(select\ the\ PSU\ in\ 1st\ stage)xProb(select\ the\ Unit\ in\ 2nd\ stage)$$

4.10.3 Estimate of Mean

$$Est(\bar{Y}) = \frac{\sum_{i=1}^{n} w_i y_i}{N} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{N}$$

4.11 Estimation using Probability Proportional to Size Sampling

4.11.2 Estimate of Total

$$Est(Y) = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

where, as mentioned above

$$\pi_k = np_k = n\frac{Z_k}{T_z}$$

4.11.3 Estimate of Mean

5 Introduction to sample size calculations

Sample size calculations determine the require minimum number of selected elements to achieve a minimum level of precision for a resulting estimate. These calculations start with precision requirements. Survey designs may also start with a total budget or total sample size, and then allocate this sample size across a given set of strata. Both methods are valid approaches, and both rely on balancing cost and precision to reach the final design.

5.1 Sample size calculations for a simple random sample

Sample size calculations for a simple random sample design require three decisions: the confidence level for our estimates, the variance of indicator of interest, and the maximum margin of error. The formula for the required sample size differs slightly if the indicator of interest is a continuous variable (like household consumption, income, or years of education) or proportion (share of individuals with post-secondary education, share of households with electricity, or share of households with a mobile phone). The two formulas are below:

Continuous variable	Proportion
$n_\infty = \frac{t_\alpha^2 \times Var(X)}{E^2}$	$n_\infty = \frac{t_\alpha^2 \times P(1-P)}{E^2}$

In these formulas, the variance is represented by either Var(X) for continuous variables or P(1-P) for proportions (where P is the proportion in the population). The confidence level is represented by α, which has a corresponding *t* value from Student's t-distribution table. The most common value of α is 95 percent, which corresponds to a t-value of 1.96. There are other t-values for higher or lower levels of required precision. Finally *E* represents the margin of error that we are willing to accept. An important point to note about E is that it must be in the same units as the variance. For example, if the variance is in dollars, then the E should also be in dollars. A common error is to define E as a percentage. Consider the example below.

Mean = $\bar{y}$ = 500 AUD; Var(X) = 125 AUD^2; α = 95%; t_α=1.96; E = 5% of mean	
$n_\infty = \frac{t_\alpha^2 \times Var(X)}{E^2} = \frac{1.96^2 \times 125^2}{25^2} \approx 96$	$n_\infty = \frac{t_\alpha^2 \times Var(X)}{E^2} = \frac{1.96^2 \times 125^2}{0.05^2} \approx 24{,}000{,}000$

In the equation on the left, the analyst correctly multiplied the mean by the percentage to obtain a value in AUD, and the calculations required a sample size of 96. In the equation on the right, the analyst erroneously used the percentage, which in this case was a maximum error of 0.05 AUD, and obtained a required sample size of more than 24 million.

The formulas above include the subscript ∞ , or calculate the sample size for an infinite population. In reality, all populations are finite. To convert the sample size for an infinite population to a finite population, the *finite population correction (fpc)* formula is required: $n_N = \frac{n_\infty}{1+\frac{n_\infty}{N}}$, where *N* is the size of the population. While the fpc is useful when designing facility and business surveys which may have limited population sizes, in many cases analysts skip this step for household surveys because the size of the sample relative to the size of the population is small, meaning there is very little impact from the fpc. See the three examples below.

N = 50,000	N = 10,000	N = 1,000
$n_N = \frac{n_\infty}{1+\frac{n_\infty}{N}} = \frac{96}{1+\frac{96}{50{,}000}} \approx 96$	$n_N = \frac{n_\infty}{1+\frac{n_\infty}{N}} = \frac{96}{1+\frac{96}{10{,}000}} \approx 95$	$n_N = \frac{n_\infty}{1+\frac{n_\infty}{N}} = \frac{96}{1+\frac{96}{1{,}000}} \approx 88$

5.2 Sample size calculations for a complex sample design: stratificationSample size

formulae and calculations for a stratified SRS sample design are more complex than for a simple random sample design. In addition to the confidence level for our estimates, the sample size calculations also require information about:

- the strata to be used
- the population size and the variance of the indicator of interest, both at the national level and within each stratum
- the method for allocating the sample across the strata – e.g. equal, proportional, Neyman, practical – along with all the information required for the allocation formulae
- the maximum margin of error that is acceptable on national level estimates, as well as the stratum-level estimates (where required).

In practice, sample designers tend to use existing tools (e.g. EXCEL templates, STATA code, R packages) that have been developed to assist with sample size calculations for stratified sampling.

Using these tools to decide on a final stratified sample design is often an iterative process, that requires some initial assumptions to be made – for example:

- assuming a fixed overall sample size, n, based on the previous sample design or the maximum budget
- then assuming a Neyman allocation of this overall sample size across the strata, to give n_h (as noted in section 3, Neyman allocation serves as a useful starting point for many sample designs, given it minimises the standard error at the national level)

Based on these assumptions, plus stratum level information on the population size and variance, the resulting margin of error, E_h, can be calculated for each stratum. Stratified sampling involves SRS within each stratum, so the formulae for E_h can be found be rearranging the sample size calculation formulae from section 5.1. Ignoring the finite population correction factor, the two formulas are:

Continuous variable	Proportion
$E_h = t_\alpha \times \sqrt{\frac{Var_h(X)}{n_h}}$	$E_h = t_\alpha \times \sqrt{\frac{P_h(1-P_h)}{n_h}}$

The resulting margin of error at the overall or national level can then be calculated as follows:

$$E = t_\alpha \times \sqrt{\frac{Var(X)}{n}}$$

where, under stratified sampling (again ignoring the fpc): $Var(X) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 Var_h(X)$

The resulting margins of error from this initial sample design can then be compared with the maximum acceptable margins of error, at both the overall or national level and the stratum level. If the margins of error resulting from the sample design are too high, then the sample design can be adjusted and the calculations re-done. Adjustments to the sample design might include:

- increasing the overall sample size (though this is generally not an efficient way of improving the margins of error in a stratified sample design)
- increasing the sample size to reduce the margin of error in specific strata (moving from a strict Neyman allocation to an adjusted Neyman or practical allocation)
- creating additional strata (e.g. by splitting one of the original strata into 2 or more sub-strata) .

The process is repeated, with the sample designer adjusting and refining the sample design along the way, until they find a sample design that strikes an acceptable balance between the desired precision of the estimates (maximum acceptable margin of error) and the overall sample size or budget.

<u>Example: First iteration of a stratified sample design</u>

Initial sample design assumptions:

- two strata – Urban and Rural, with population size, mean and variance as shown in the table below.
- a maximum overall sample size of 500, with Neyman allocation of the sample across strata
- α = 95%; t_α=1.96

Stratum (1…h)	N_h	Mean = $\bar{y}$ (in AUD)	Var(X) (in AUD)

Urban	7,500	1000	100^2
Rural	2,500	500	30^2
National	10,000	800	

Using this information along with formulas provided previously, the resulting sample sizes and margins of error at the stratum level are as follows:

Stratum (1…h)	N_h	Mean = $\bar{y}_h$ (in AUD)	$Var_h(X)$ (in AUD)	$n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$ where $S_h = \sqrt{Var_h(X)}$	$E_h = t_\alpha \times \sqrt{\frac{Var_h(X)}{n_h}}$
Urban	7,500	1000	100^2	**455**	**9.2**
Rural	2,500	500	30^2	**45**	**8.7**

From the stratum level information, we can then calculate the overall or national sample size (which should be 500, which was the original assumption), variance and margin of error as follows:

$$n = \sum_{h=1}^{H} n_h = 455 + 45 = \mathbf{500}$$

$$Var(X) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 Var_h(X)$$
$$= \left(\frac{7{,}500}{10{,}000}\right)^2 100^2 + \left(\frac{2{,}500}{10{,}000}\right)^2 30^2 = \mathbf{75.4}$$

$$E = t_\alpha \times \sqrt{\frac{Var(X)}{n}} = 1.96 \times \sqrt{\frac{75.4}{500}} = \mathbf{6.6}$$

If the maximum acceptable margin of error was 10 at both the overall level and the stratum level, then this stratified sample design meets the requirements. Provided the sample size of 500, and resulting survey budget are acceptable, then this design would be sufficient.

However, if the maximum acceptable margin of error is 10 at the stratum level, but 5 at the overall level, then this sample design does not meet the requirements. To achieve the requirements, the sample designer would need to make some adjustments to the sample design (e.g. creating additional strata to better control the overall variance), repeat the calculations and check the resulting margins of error again. They would continue iterating through this process until they find a design that strikes an acceptable balance between the desired precision of the estimates (maximum acceptable margin of error) and the overall sample size or budget, which may require some trade-offs.

5.3 Sample size calculations for a complex sample design: multi-stage selection

Chapter 3 - Household Income & Expenditure Surveys

1. Introduction to Household Income & Expenditure Surveys (HIES)

Household Income & Expenditure Surveys (HIES) are conducted every five years to measure a range of socioeconomic variables. The distinctive feature of the HIES is that it collects consumption data on food and non-food items. This information is then used to achieve the two main objectives of the HIES: to update the CPI basket and to measure well-being. [1-2 sentences on sample requirements for CPI baskets.] In terms of the sample requirements for measuring household well-being, the sample design is developed around being able to measure results with a given level of precision for the relevant indicator in the target population. The relevant indicator for a HIES is either total household consumption per capita[1] or income per capita. The well-being indicator is generally used as the basis for the sample size calculations even if the survey measures a range of socioeconomic indicators because household well-being is often correlated with other indicators, including those related to education, health, livelihoods, etc. The HIES target population is those living in households (i.e. excluding institutionalized populations such as those in dormitories, boarding schools, prisons, military barracks, etc.). If it is necessary to include these populations, separate sampling frames would be required.

In general, in the Pacific, there are 2 main types of HIES sampling design:

- stratified sample design, with simple or systematic random sample of households within each stratum
- stratified two-stage sample design, with PPS sampling of EAs at first stage, then simple or systematic random sample of households within selected EAs at second stage

2. Basic design (stratified two-stage cluster design)

Most HIES use a two-stage stratified cluster sample design in which census enumeration areas (EA) are randomly selected with probability proportional to size (PPS) in the first stage and households are randomly selected with systematic random sampling in the second stage. A version of this design will likely be the most common approach in the Pacific region as well, though depending on the size of the population and the availability of updated administrative statistics, some countries may forego the clustering.

2.1 Calculating design effects from previous survey

The starting point for all rigorous sample designs is existing information on the variable of interest from another data source. In the case of the HIES, most countries now have previously conducted a HIES, on which to guide future HIES designs. Even if the survey is several years old, it is often the best basis for designing the sample because it includes the indicator of interest as well as the necessary cluster and stratification variables. Because the households will be selected from an updated sampling frame in the second stage, basing the design on an outdated survey will not cause bias. It may, however, lead to a design that is inefficient, which means that, if more up-to-date information were available, the clusters and households could have been allocated different to produce a more precise estimate for the same sample size.

[1] The term “per capita” here is used to cover both traditional per capita measures, in which each household member is counted equally toward measuring the total household size, and “per adult equivalent” measures in which the age and gender of household members is considered.

The key ingredients from past HIES data for the sample design are the sample size, the mean, the standard error, the cluster size and the design effects. Each of these variables should be calculated at the level of the strata for the new survey, even if different stratification was used in the analysis of the original data. This information is then used to calculate the necessary components for the new design. The sample size and the standard error are used to estimate the standard deviation and variance of the variable of interest. The cluster size and design effects are used to calculate the intracluster correlation coefficient (icc). With this information, the sample designer can predict the precision on the sample design for the new survey.

2.2 Choosing the stratification

Typically, stratification attempts to group relatively homogeneous units of analysis together. In the case of the HIES, a standard design defines the strata using administrative divisions split into urban and rural areas, though not all countries in the Pacific use this approach. When defining the stratification, it is important to consider two types of strata: analytical strata and design strata.

2.2.1 Analytical strata

Analytical strata are what most people are referring to when they ask about the stratification design or when they ask at what level the survey is "representative."[2] Analytical strata represent the lowest level at which the results will be presented and can be thought of as the rows in the tables for the final report. For example, if an NSO wants to present well-being statistics separately for the capital city, as well as the three regions, then the survey would have four analytical strata. For analytical strata, there is an additional requirement for estimates at this level to have a certain level of precision. In many sample designs the precision is measured with the relative standard error (RSE), which is defined as the standard error divided by the mean. Sample designers usually target a RSE of around 5 percent, though RSEs are sometimes as high as 10 percent if there are budget limitations.

2.2.2 Design strata

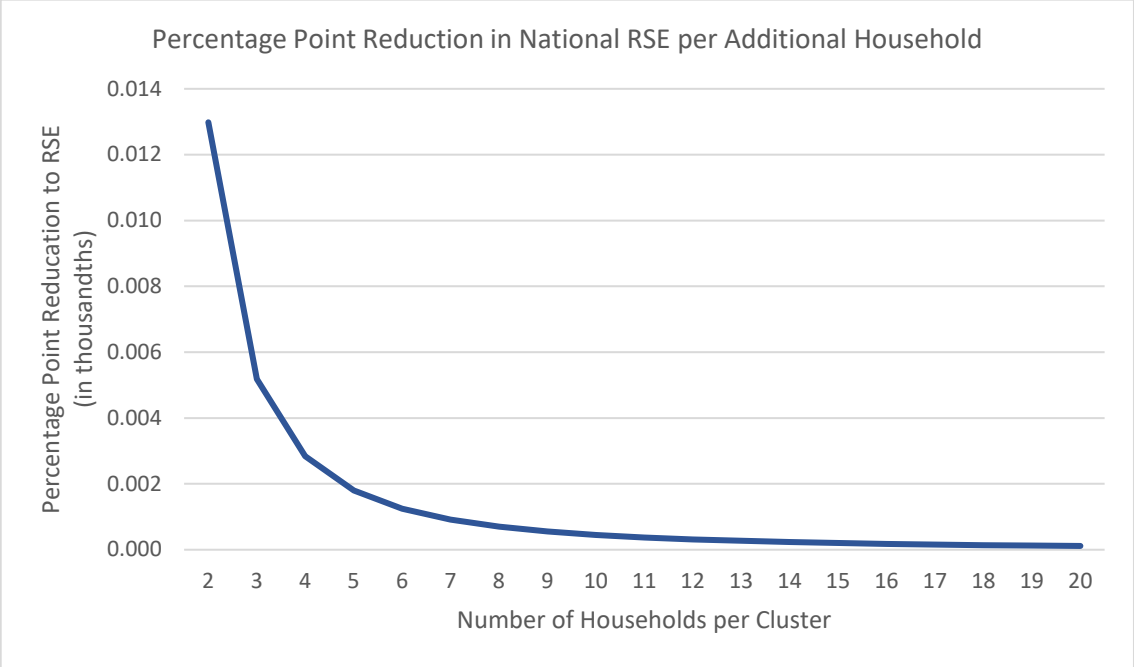
In addition to analytical strata, many surveys also use design strata to gain greater control over different elements in the variation and increase the overall precision. For example, one region may be made up of two islands that are very different. One island may have a diverse economy and great variation in living standard between the households. The other island could be a community of fishermen with few other sources of livelihood and little variation in the standard of living between the households. If the population of the two islands is the same, we would end up with roughly the same number of clusters and households being selected on both islands. However, for the island of fishermen, there is little variation so there is a limited amount of additional information that can be gained from additional clusters or households. In contrast, because the first island has lots of variation, comparatively more information is gained by having more clusters and households in the first island. Therefore, the sample design may want to explicitly decide to have more observations on the first island than on the island of fishermen, even though in the final tables, the results will

[2] "Representative" can be a confusing term because it has two different definitions depending on the audience. To a statistician, "representative" simply means an unbiased estimate of the population of interest. A sample of 10 people could therefore generate a representative estimate of a national population if those 10 people were randomly selected. To an economist and to most policymakers, there is an additional requirement of a minimum level of precision. For example, the random sample of 10 people would yield an estimate with a very wide confidence interval. To be representative to an economist, we would need a sample large enough such that the relative standard error is less than 5 percent, for example.

only be presented at the level of the region. In that case, the sample design would include design strata at the island level with in the analytical stratum of the region.

2.3 Defining cluster size (different for different areas?)

The icc calculated from the previous survey is used to understand the relationship between the number of households per cluster and the RSE. The figure below shows the reduction of RSE at the national level for each additional household in a simulated design with 500 clusters allocated proportional to size.



Source: Papua New Guinea National Household Survey 2009-2010, per capita household expenditure

The largest contribution to survey estimates comes from the first additional household (moving from a cluster size of 1 to 2), reducing the RSE by 0.005 percentage points. Each additional household applied to the cluster in the simulation reduces the RSE less. In this example, the marginal benefit of an additional household becomes limited around 8 households per cluster, and basically non-existent after 12 households. In general, more smaller clusters over fewer larger clusters increases the level of precision of the resulting estimates for a given sample size.

3. First stage

3.1 Defining clusters (EAs?)

Generally, surveys use census enumeration areas (EA) as the primary sampling unit because they are of a roughly uniform size (assuming a recent census) and maps are available clearly defining the boundaries.

3.2 Preparation of the sampling frame

In preparation for the sample selection, survey designers should first ensure that they have a clean sampling frame from which to select the sample. In general, the sampling frame is the list of all the enumeration areas (EA) from the most recent census along with their estimated number of

households. Since HIES surveys select households instead of individuals, it is important that the list have the total number of households instead of total number of individuals. Also, if possible, the list should be updated with the latest population estimates, though it is possible to proceed even if the list is dated. The sampling frame should be carefully checked for duplicates and, as EAs are generally by definition between 150 and 200 households, any EA with outlier populations should be verified. It is important to start the process of verifying the sampling frame early as checking outliers often requires contacting regional offices and can be a time-consuming process.

3.3 Selecting EAs

It is standard practice in household surveys to select the enumeration areas within each stratum by using "probability proportional to size" sampling. This increases the probability that larger, more populous EAs will be selected. This increases efficiency and decreases costs (as it is less likely that small remote clusters will be selected). In general, implicit stratification is also used within the strata to increase efficiency and prevent a chance bunching of selected EAs.

4. Second stage

A high-quality frame for the second stage selection of households is a key component in generating unbiased estimates. Generally, there are two options: administrative statistics or conducting a household listing operation. In either case, there are two objectives: to produce a list of households from which survey and replacement households can be drawn, and to have an accurate count of households as an input into the weight calculations.

4.3 Administrative statistics

In some cases, there will be administrative statistics available on which to base the selection of households. To qualify as the sampling frame for the second stage of selection, the administrative statistics should be up-to-date and provide a complete listing of all the household within the selected EA. It is important that there is a way to exclude households living outside of the selected EA if the administrative lists cover more than one EA. Missing information, such as households that have recently moved to the area and are not included on the administrative list, could lead to bias in the resulting estimates.

4.4 Household listing

If administrative lists are not available, a household listing operation is required. To implement an accurate household listing operation, it is important that listing teams know the boundaries of the selected EA, and carefully canvas each structure to identify the number of households residing there. (It is therefore also important that listing teams are familiar with the definition of "household" as it relates to the survey.) If the listing team does not cover all areas of the EA, or they miss households living in canvased areas, the missed households will have no probability of selection and the resulting data will be biased. In addition, if the EA-level counts are low, when the weight are calculated the total national population count will also be low. Similarly, if the counts are too high, if a team lists households outside the EA boundaries for example, the total population counts will be too high. Therefore, correct listing totals are important for both unbiased selection and accurate weight calculations.

In some rare cases, EAs will be too large to be effectively listed in a timely manner. In that case, the EA must be segmented. During segmentation, the EA is divided into 2-3 equally sized sub-clusters of

about 200 households. The division should be done using the EA map and with clearly defined landmarks (such as roads and rivers) as boundaries. One segmented, one of these segments is randomly selected and the listing process continues as described above. In the case of segmentation, it is extremely important that the field supervisor clearly notes that the EA was segmented and into how many sub-clusters it was divided. This information will be necessary for unbiased weight calculations.

5. Replacement procedures

During fieldwork, it is inevitable that there will be certain selected EAs or households in which it is not possible to conduct interviews. In an ideal case, the sample design would have selected additional EAs and households to compensate for those that are not interviewed. However, this is rarely done in practice in the developing world because refusal rates are difficult to estimate, and any underestimation of the refusal rate would lead to a smaller sample size than planned, while an overestimation would have cost implications.

Replacements can take place at the level of the EA or the household.

5.3 EA-level replacements

Replacements at the EA level are generally due to inaccessibility, either due to unexpected travel conditions or safety concerns. Since the EAs that cannot be accessed are fundamentally different from those that can be reached, being more remote or more dangerous for example, any replacement at the EA level leads to a loss of representativeness in the final dataset. As an example in a hypothetical country, if Long Island in the North region cannot be accessed, the survey would be "representative of the North region except for the Long Island." Similarly, if only part of Long Island were inaccessible, the survey would be "representative for the North region except for areas A, B, and C which were excluded due to inaccessibility related to the recent typhoon." It is important to note precisely in the survey documentation and all reports any areas which had to be excluded. As excluding and replacing EAs leads to a loss of representativeness, it should be avoided as much as possible. In areas of inaccessibility, it may be best to delay the survey in certain areas where there are short term issues rather than to replace them entirely. The only time that replacing EAs would not lead to a loss of representativeness is in the case that the entire EA has been destroyed, for example, if a village has been relocated due to a road or dam project. These EAs can be replaced from the original list with no loss of representativeness.

5.4 Household level replacements

Most replacements, however, take place at the household level. Households can be replaced for two reasons. First, a dwelling structure may have been occupied at the time of the listing but is now vacant. Unless there is a systematic reason that the dwelling is vacant (such as seasonal migration), the household can be replaced with another randomly selected household in the EA without introducing bias. If the household, however, refuses to participate, even a randomly selected replacement will introduce bias into the sample. This is because those households that refuse to participate are likely to differ in some ways from those that agree to participate.

For this reason, supervisors and interviewers should be trained in techniques to maximise response rates. In addition, HQ should monitor refusal rates by interviewer and team, to identify and remedy problems early. But in some cases, a household refusal is still unavoidable. In this case, due to high

travel costs for most household surveys implemented in the Pacific, replacement households are selected for households that refuse.

The standard procedure for replacing households at the EA level is to select more households than needed and to hold back some as replacements. If households are selected with systematic random sampling, the step should be calculated for the larger number of households and then the selected households randomly allocated into "selected" and "replacement." In an example where 10 households are needed for the survey and 5 as replacements, 15 total households should be selected using systematic random sampling from the listing. Once those 15 are selected, 10 should be selected using simple random sampling to be interviewed. If there are no refusals, then the other 5 households remain unused. If a household refuses, one household from the list of 5 replacements should be randomly picked to be interviewed as a replacement. This process should be done by a regional supervisor or in the headquarters, and field supervisors should be required to explain in detail why it was not possible to interview the original household. In general, no more than 20% of households should be replaced in a given EA, as each replacement introduces some bias into the data. Details notes should be kept for each team on the number of replacements and this information should be included in the basic information document.

6. Weights

6.3 Probability weights

The weight calculations follow the steps of the selection. In the first stage, EAs are selected with probability proportional to size. Within each stratum, the probability of selection is then $p_1 = \frac{kn'}{N}$ where *k* is the number of EAs selected in that strata, *n'* is the expected population of the EA based on the sampling frame (in households), and N is the population of the strata (in households). In the case in which an EA required segmentation, the probability of selection in the second stage is $p_2 = \frac{1}{s}$ where *s* is the number of segments. For most of the cases in which segmentation was not necessary, *s* = 1. Within each selected EA or segment, the probability of selection is $p_3 = \frac{m}{n}$ where *m* is the number of households to be selected and *n* is the number of households found in the listing form of the EA or segment. The probability of selection is $p = p_1 p_2 p_3 p_4$. The probability weight would then be $= \frac{1}{p}$.

6.4 Non-response adjustment

6.4.1 Household

It may be necessary to include a non-response correction for a limited number of EAs if a team exhausts all selected household and replacements before completing their required number of households. In this case, it is necessary to assume that all non-responding households within an EA are statistically identical to those that chose to respond. Then $w_{nr} = \frac{m}{m'}$ where *m'* is the number of EAs actually completed. The weight adjusted for non-response would then be $w_{p,nr} = w_p w_{nr}$.

6.4.2 Individual

6.5 Post-stratification

To reduce the overall standard errors and weight the population totals up to the known population figures, many weight calculations also include a post-stratification adjustment. The level of disaggregation for the post-stratification adjustment should be at the lowest reliable level available from an auxiliary data source, regardless of the level of representativeness for which the survey was designed. For example, if reliable census projections are available for the sub-regional level, these should be used even if the sample was designed only to be representative at the regional level. Note that close attention should be paid to the reliability of the outside data source underlying the post-stratification calculations. With proper design and implementation, the survey population estimates should be close to the actual population estimates. Post-stratification should therefore be seen more as a fine-tuning adjustment rather than a major realignment. If the weights are adjusted using poor quality auxiliary information, there is the possibility of reducing precision or introducing bias into the estimates. This issue is of particular concern when population projects are used from censuses that are a number of years old.

To calculate the post-stratification adjustment, the formula is $w_{ps} = \frac{pop_{known}}{pop_{survey}}$. The final weight would then be $w_{final} = w_{p,nr,ps} = w_p w_{nr} w_{ps}$.

References

Reference material and resources – Ch 1

Eurostat (2008). ***Survey sampling reference guidelines: Introduction to sample design and estimation techniques.***

- content for this chapter of the guidelines was based primarily on Ch 2, plus parts of Ch 3

Source: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-RA-08-003

UN Statistics Division (2008). ***Designing Household Survey Samples: Practical Guidelines.***

- content for this chapter of the guidelines was based primarily on Ch 1, Ch 8, Annex I

Source: https://unstats.un.org/unsd/demographic/sources/surveys/Series_F98en.pdf

UN Data Glossary http://data.un.org/Glossary.aspx

Reference material and resources – Ch 2

Reference material and resources – Ch 3

Appendices