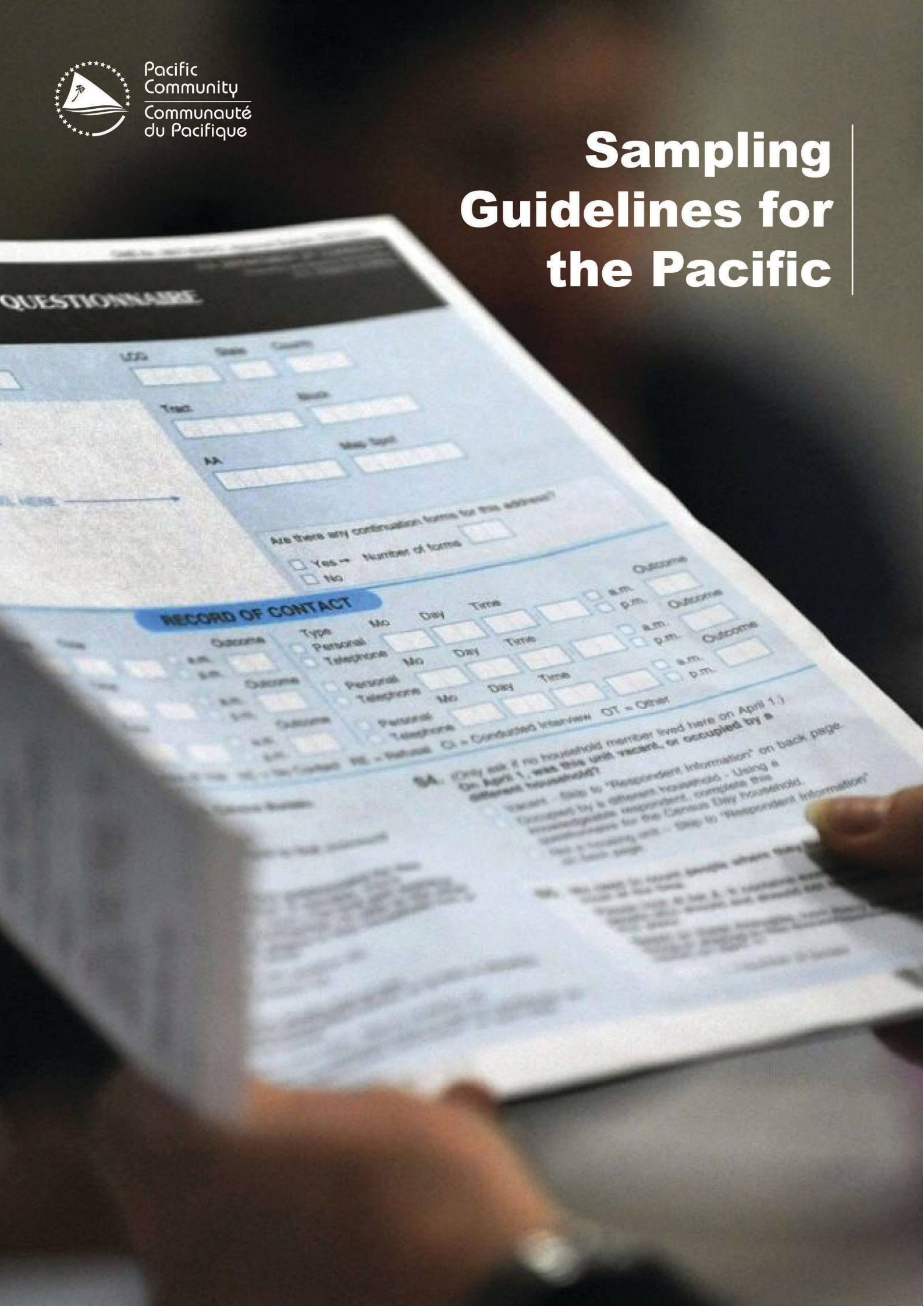
Pacific Community
Communauté du Pacifique

Sampling Guidelines for the Pacific



Sampling Guidelines for the Pacific

Noumea, New Caledonia
February 2024

Cover picture: Wikimedia Commons

Prepared for publication at SPC's Headquarters,

B.P. D5, 98848, Noumea Cedex, New Caledonia

Table of Contents

Tables

Figures

Acronyms

ABS	Australian Bureau of Statistics
CI	Confidence Interval
CPI	Consumer Price Index
deff	Design Effect
DHS	Demographic and Health Survey
EA	Enumeration Area
ESCAP	United Nations Economic and Social Commission for Asia and the Pacific
fpc	finite population correction
HIES	Household Income and Expenditure Survey
HH	Household
ICC	Intra-cluster Correlation Coefficient
LFS	Labour Force Surveys
ME	Maximum acceptable margin of Error
MICS	Multiple Indicator Cluster Surveys
MSE	Mean Square Error
NSO	National Statistics Office
PCE	Per Capita Expenditure
PICTs	Pacific Islands Countries and Territories
PPS	Probability Proportional to Size
PSMB	Pacific Statistics Method Board
PSU	Primary Sampling Unit
RSE	Relative Standard Error
SD	Standard Deviation
SDGs	Sustainable Development Goals
SE	Standard Error
SIDS	Small-Island Development States
SPC	Pacific Community
SRS	Simple Random Sampling
WB	World Bank
WOR	without replacement
WR	with replacement

Introduction

The Pacific Island countries and territories (PICTs) are generally characterized as having small populations that are geographically dispersed across multiple islands. While there are many cultural, social and economic differences among the countries and territories of the Pacific region, most are mostly are Small Island Developing States (SIDS) and the small size, scattered and sometimes hard to reach populations, coupled with underdeveloped administrative systems, pose significant common challenges to the development of statistics.

Despite these challenges being widely acknowledged, there is enormous and increasing demand for all PICTs to regularly produce high quality disaggregated statistics to report against economic, social, cultural and environmental development indicators. This is exemplified by the Pacific Region SIDS adopting the 2030 Agenda for Sustainable Development, which includes a commitment to report against the indicators designed to measure progress towards the 17 Sustainable Development Goals (SDGs). Among other national and regional demands for high quality statistics, the SGDs place enormous strain on the small, underdeveloped and under resourced statistical systems of the Pacific region.

All PICTs conduct household (HH) surveys. Sample surveys are and will remain the main source of information for many social and economic development indicators. Surveys are increasingly called upon as the sole data source for many national, regional and international development indicators, including those of the SDGs, and it is expected that HH surveys will continue to increasingly be called upon to meet emerging data needs across a range of sectors and themes.

Balancing the conflicting demand for more efficient data collection with the demand for more disaggregated development indicators requires efficient sampling strategies. These guidelines aim to provide Pacific survey practitioners with a resource to help achieve this balance while simultaneously strengthening capacity in sampling theory.

The Sampling Guidelines for the Pacific include the following chapters and topics:

- Chapter 1 provides an introduction to surveys and sampling. It covers basic concepts and definitions and provides an introduction to notation. It provides information on survey quality, sampling frames and sample design.
- Chapter 2 provides a theoretical overview of sampling theory, including different sampling strategies, such as Simple Random Sampling, Stratified Sampling and Multi-Stage Sampling. Further, this chapter provides an introduction to sample size calculations.
- Chapter 3 provides a practical example of sampling approaches for Household Income and Expenditure Surveys (HIES), which are an important data source for the measurement of poverty and food security, and for the rebase of the consumer price index (CPI). In this chapter, an overview of the different stages to Stratified two-stage cluster design is provided with additional information on replacement procedures and generation of sampling weights.

In addition to these chapters, five appendixes are provided, which present case studies for: i. computation of sample size for a Simple Random Sample; ii. Computation of sample size for a two-stage stratified sample; iii. Selection of a probability proportional to size sample; and iv. a glossary of key sampling terminology. The fifth appendix, finally, presents sampling issues when dealing with core and supplementary modules.

It is envisaged that additional chapters or appendices might be progressively added to these guidelines; they will provide practical examples for sampling approaches for other core national statistical collections that are frequently conducted in the Pacific region, such as: Multiple Indicator Cluster Surveys (MICS), Labour Force Surveys (LFS), Disability Surveys and Agricultural Survey. Chapters 1 and 2 of the Sampling Guidelines for the Pacific remain relevant; however, the sampling approach – and the unit – differs for those surveys, which is why these guidelines are considered to be a 'living' product, with a view for further development as the opportunity and need arise.

The preparation of the Sampling Guidelines for the Pacific was commissioned by the Pacific Statistics Methods Board (PSMB) and they were written by the Sampling Sub-Committee of the Board, which consisted of the following

members: Taggy Tangimetua, Cook Islands National Statistics Office (NSO); Bruce Fraser, Australian Bureau of Statistics (ABS); Bertrand Buffière and Michael Sharp, Pacific Community[1] (SPC), Tracey Savage, Stats NZ, Chris Ryan, United Nations Economic and Social Commission for Asia and the Pacific (ESCAP), and Kristen Himelein, World Bank (WB).

The Sampling Guidelines for the Pacific were introduced at the Regional Sampling and Planning Workshop held in Nadi, Fiji, in February 2020. The workshop was attended by 18 participants from 8 PICTs; the feedback received during the workshop was incorporated in the final preparation of the guidelines.

[1] The Pacific Community (SPC) gratefully acknowledges the funding provided from the Trust Fund for Statistical Capacity Building (TFSCB), through the World Bank, which financed SPC's contribution to the development of the Sampling Guidelines for the Pacific, as well as the funding provided by MFAT and DFAT to Stats NZ and ABS, respectively.

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 *or* ***units*** are the objects from which information is sought in a survey, and for which statistics are ultimately compiled. Examples of *elements or units* include: people, HHs, schools, hospitals, businesses, farms, and geographic areas (such as enumeration areas, or EAs).

The ***target population*** *or* ***population of interest*** for a survey is the population of elements or 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 HHs (i.e. excluding institutionalized populations such as those in dormitories, boarding schools, prisons, military barracks, etc.) in a country.

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

A ***sample survey*** refers to a survey that aims to collect or observe data from a subset or sample of the population of interest – i.e. from only some of the elements or units in the target population – but still make quantitative statements about the whole population.

The objective of a sample survey is to estimate ***parameters*** or ***indicators*** of the whole population, such as the mean, total, or proportion, from only a part of the population (i.e. from a sample). These are referred to as ***sample estimates*** or simply ***estimates*** because the true value is not known, as not all units in the target population were surveyed. Examples of sample estimates from HIES include poverty rates and average income.

An ***estimator*** is a rule or function that is used to calculate estimates from the data collected or observed in a sample survey. For example, an estimator for the population mean is the sample mean.

Estimates from a sample survey must be accompanied by associated ***measures of uncertainty***, such as the standard error or confidence interval. These measures describe the **expected precision** of the estimate obtained from the sample survey.

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 or 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 estimates from that sample.

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, HHs, 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 HHs – e.g. population statistics, labour force participation, household income and expenditure/consumption, poverty, education and health. Census, HIES, Demographic and

Health Survey (DHS), Labour Force Survey (LFS), Gender based Violence Survey and Multiple Indicator Cluster Survey (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.

1.3 Introduction to notation

This section introduces the basic notation used throughout the Sampling Guidelines for the Pacific. The aim is to keep the notation as simple and consistent as possible throughout. Unfortunately, statistical and sampling notation is not standardised. So, it's important to be aware that different authors, texts and presenters may use different notation to that used here. Where possible, we've noted the most common alternative notation(s) that you may see elsewhere in the footnotes.

The key principles underpinning the notation used in the Sampling Guidelines for the Pacific (see Table 1) are as follows – in general:

- upper case Roman letters are used to represent population characteristics or parameters[2]
- lower case Roman letters are used to represent sample characteristics of sample estimates of population parameters[3]
- the variable of interest[4] is Y (X is also commonly used elsewhere, and sometimes Z)

Table 1. Notations for Population and Sample characteristics

Characteristic	Notation for:	
	Population	Sample
Number of units/elements	N	n
Identifiers for each unit/element	$i = 1, \dots, N$	$i = 1, \dots, n$
Observed values/measurements for the variable of interest	$Y_1, Y_2, \dots, Y_i, \dots, Y_N$	$y_1, y_2, \dots, y_i, \dots, y_n$
Summation	$\sum_{i=1}^{N}$	$\sum_{i=1}^{n}$
	Population parameters	Sample estimates
Total	Y	y
Mean	$\bar{Y}$	$\bar{y}$
Proportion	P	p
Variance	S^2	s^2
Standard deviation	S	s

The **probability of selection** into a sample for a unit or element i is: π_i

The **sampling weight**[5] for unit or element i is the inverse of its probability of selection: $w_i = \frac{1}{\pi_i}$

---

[2] Greek letters and symbols are also commonly used to represent population parameters – e.g. μ for the population mean, σ^2 for the population variance, and σ for the population standard deviation.

[3] A caret or "hat" – ^ – symbol above a letter is also commonly used to represent sample estimates. The "hat" may appear above the Greek letter or upper-case Roman letter for the corresponding population parameter – e.g. $\hat{\mu}$ or $\hat{\bar{Y}}$ for an estimate of the population mean – or possibly above the lower case Roman letter for the same – e.g. $\hat{\bar{y}}$).

[4] The variable of interest may sometimes be referred to as the study variable.

[5] The sampling weight may also be referred to as the selection weight, sample weight or probability weight, or simply the weight.

Additional statistical and sampling terminology, notation and formulae will be introduced and explained in subsequent sections of the Sampling Guidelines for the Pacific, as they become relevant.

2. Survey quality

2.1 Sampling error versus non-sampling error

It is important that information on the 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, such as: frame error,[6] non-response error,[7] measurement error[8], and processing error[9]. These errors may be random or systematic. Both censuses and sample surveys are subject to non-sampling error. Given than non-sampling errors can come from multiple sources and are often unobservable (or costly to observe), it is usually not feasible to precisely quantify the total non-sampling error in a survey.

Together, sampling and non-sampling error are known as ***total survey error***. It is important to carefully control the total survey error when designing a sample survey, and to ensure both types of error are well-managed across the survey process (from survey design through to publication). A balance between sampling and non-sampling error needs to be found in sample surveys. 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.
- if money from a limited survey budget is put towards a large sample to reduce the sampling error, there is a risk that the remaining budget for interviewer training, field supervision etc. will be less than adequate, resulting in increased non-sampling error.

---

[6] A frame error consists of a difference between the information presented in the frame and the real world situation. The sources of information used to maintain and update the frame will generally contain irregularities of some sort: the frame may be subject to certain lags in the recording of real world events, or it may have gaps due to the lack of adequate sources for certain types of information. If these distortions compared to the real world situation are considered to be acceptable by the users of the Frame, they should not be considered to be errors. If they are not acceptable, procedures or sources need to be changed or improved (see: https://cros-legacy.ec.europa.eu/content/frame-error_en).

[7] Non-response error results from a failure to collect complete information on all units in the selected survey. It affects survey results in two ways. First, the decrease in sample size or in the amount of information collected in response to a particular question results in larger standard errors. Secondly, a bias is introduced to the extent that the distribution of some characteristics of non-respondents differs from the distribution of the respondents within a selected survey (see: https://cros-legacy.ec.europa.eu/content/non-response-error_en).

[8] Measurement errors are those errors in the survey observations that may be caused by interviewers, respondents, data processors, and other survey personnel. Often, the causes of measurement errors are poor questions or questionnaire design, inadequate personal training or supervision, and insufficient quality control. Measurement errors are often hidden in the data and are only revealed when the measurement process is repeated or responses are compared to a gold standard (i.e., error-free measurements). If repeated measurements are collected by the same measurement process, systematic errors may remain hidden (see: https://www.sciencedirect.com/science/article/abs/pii/S0169716108000126).

[9] Processing error includes processing-related errors in data capture, coding, editing and tabulation.

2.2 Variance versus bias

In survey sampling, preferred estimators are those fulfilling certain theoretical properties. One of these properties is ***unbiasedness***[10], meaning that the expected value of an estimator, across all possible samples, equals the true value of the population parameter being estimated – e.g. $E(\bar{y}) = \bar{Y}$

Bias is therefore the difference between the true value and the expected value of the estimator: $Bias(\bar{y}) = E(\bar{y}) - \bar{Y}$

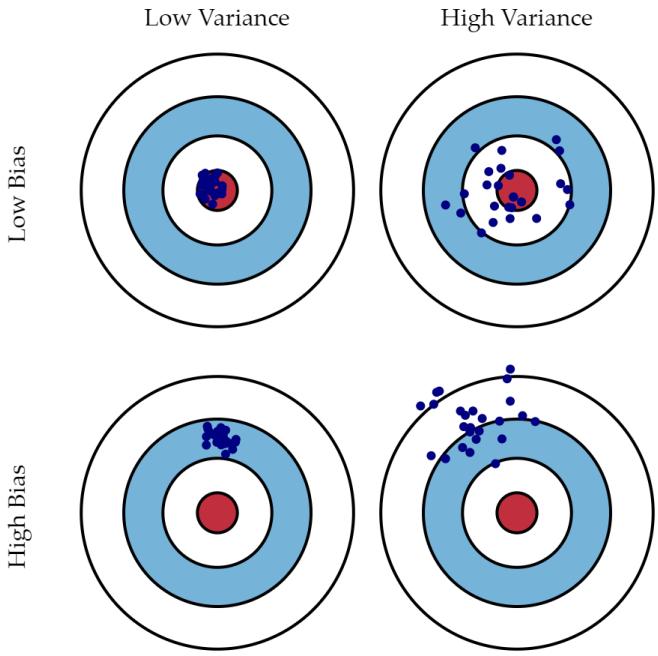
When nonzero, it represents systematic error.

The ***precision*** of an estimator reflects its variability across all possible samples (i.e. across the sampling distribution for the estimator) and is measured by the ***sampling variance***[11] $Var(\bar{y})$. The smaller the sampling variance, the better the precision of the estimator. A precise estimator is called ***efficient***, another desirable property.

As noted above, it is important to carefully control the total survey error when designing a sample survey. The total survey error of an estimator is measured by the ***mean square error***, which is the sum of the sampling variance and squared bias – e.g. $MSE(\bar{y}) = Var(\bar{y}) + Bias(\bar{y})^2$. Sample estimators with low total survey error or MSE are ***accurate***, which is another desirable property.

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

Figure 1. Variance and bias of an estimator



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 sampling variance, resulting in a low total survey error.

2.3 Quality measures for sample estimates

The ***standard error*** (SE) and ***the relative standard error*** (RSE) of an estimator are commonly used quality measures for sample estimates. The quality measures are derived from the theoretical properties discussed before.

---

[10] *Unbiased estimates may also be referred to as representative estimates.*

[11] *The sampling variance may also be referred to as the design variance.*

For an estimator $\bar{y}$ of a population mean $\bar{Y}$:

*The **standard error** is:* $SE(\bar{y}) = \sqrt{Var(\bar{y})}$, where $Var(\bar{y})$ is the sampling variance of the sample mean $\bar{y}$, calculated from the sample.

The ***relative standard error*** is: $RSE(\bar{y}) = \frac{SE(\bar{y})}{\bar{y}}$, i.e. the standard error divided by the estimate itself. The relative standard error is often expressed as a percentage $RSE(\bar{y}) \times 100\%$

When comparing the precision of different sample estimates (whether from the same survey or across different surveys), it is best to use relative standard errors, rather than standard errors, given RSEs are a standardised measure, i.e. calculated relative to the estimates themselves.

2.4 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, such as disability prevalence rates ▪ 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 or units that define a target population. 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 the units of interest 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 that are very small, and do not contribute much to the accuracy of the sample estimates

Care must be taken when doing this, though, as deliberate exclusion of a part of the target population can lead to bias in the sample estimates. 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 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 - Simple Random Sampling (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 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 that data will be collected for 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), unbiased estimates whose sampling variances (e.g. standard errors) are as small as possible, whilst also managing the costs of the survey, non-sampling errors, and the overall sample size.

4.3 Sample size determination

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 by survey 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 variables of interest to collect data on?
- is there any existing information or knowledge (informed guess) about the statistical distribution of the variables of interest?
- which are the most important population parameters or indicators to be estimated from the sample data?
- what level of precision is required for the sample estimates? how precise do the estimates of population 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 considered 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?
- what was the response rate from previous survey?

All of these questions – and many other factors – need to be considered before a sampling design (including the choice of the overall sample size) can be determined. Often 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 survey design process.

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 i in the population frame of N elements has exactly the same inclusion probability π_i that is:

$$\pi_i = \pi = \frac{n}{N}$$

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 after its selection; this means that a population element can only be sampled once. In a WR scheme, on the contrary, 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:

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

where: $n! = n \times (n-1) \times (n-2) \times \ldots \times 1$

For example, a small population of $\mathrm{N} = 10$ 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 $\mathrm{N} = 100$, and we select a sample of $n = 30$, then we find the number of different samples grows enormously, as such:

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

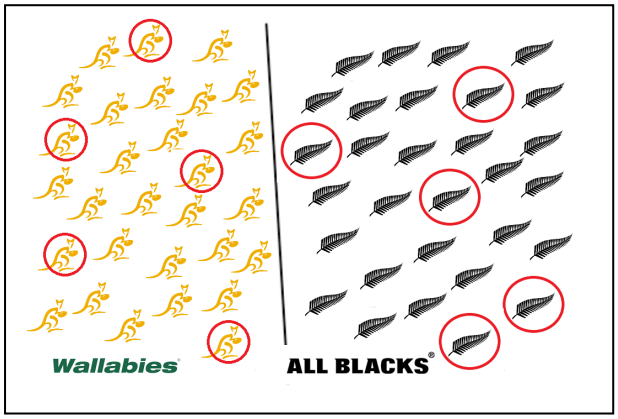
How representative is a SRS normally?

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. Figure 2 displays the case of 2 villages, each with a population of 30: one with all *Wallabies* supporters, and the other with 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

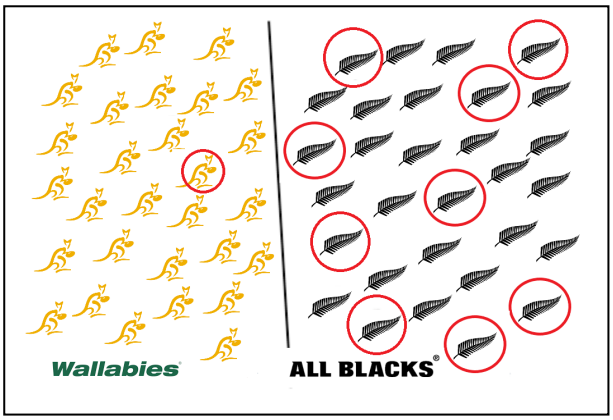
sample, Scenario 2 might happen from time to time (only 1 person selected from the village of Wallabies supporters), resulting in an outlier sample, and thus a poor estimate.

Figure 2. Representativeness of an SRS sample

Scenario 1: Representative sample

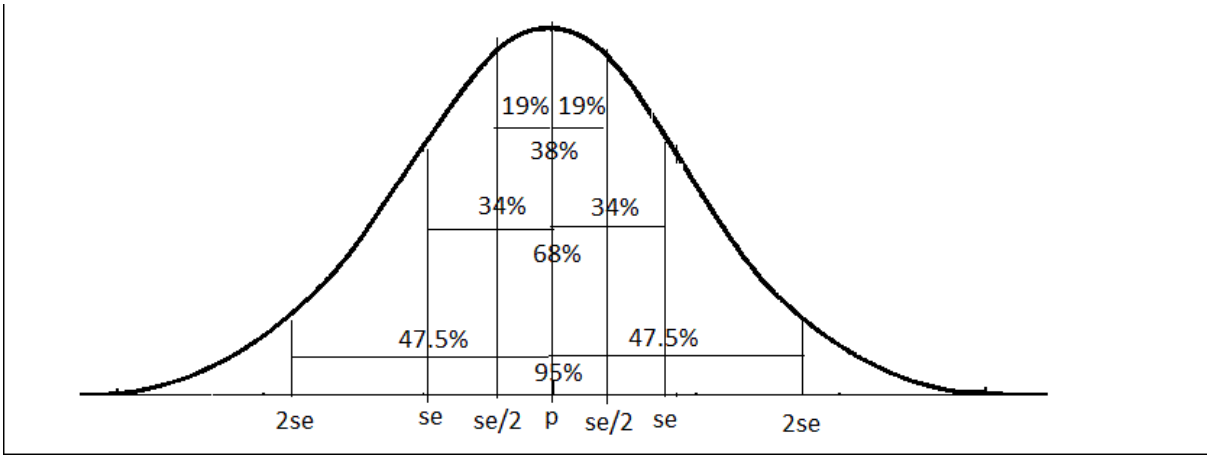


Scenario 2: Outlier sample



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% would be all *Wallabies* or all *All Blacks* supporters. Sampling theory is used to predict the likelihood of obtaining an outlier draw or sample. Figure 3 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.

Figure 3. Standard normal distribution



1.3 Estimation using SRS

Under SRS, estimators for the population total Y, mean $\bar{Y}$ and proportion P are as follows:

1.3.1 Estimate of Mean

The estimate of the mean is the sum of the variable of interest across all the elements in the sample, divided by the number of elements in the sample.

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

1.3.2 Estimate of Total

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

$$y = N\bar{y} = N\sum_{i=1}^{n}\frac{y_i}{n} = \sum_{i=1}^{n} w_i y_i$$

where $w_i = {}^{N}/_{n} = {}^{1}/_{\pi_i}$ is the selection weight for each sample unit, i

1.3.3 Estimate of Proportion

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

$$p = \frac{1}{n}\sum_{i=1}^{n} y_i$$

where: $yi = 1$ if sample unit i has the characteristic,

$yi = 0$ otherwise

1.3.4 Standard deviation and standard error of the proportion, and confidence interval

In the context of SRS, the standard deviation of the proportion is:

$$sd = \sqrt{p * (1 - p)}$$

And the standard error:

$$se = sd / \sqrt{n}$$

The confidence interval is given by:

$$C.I. = p +/- z * se$$

Where z is the z value associated to the confidence level.

1.4 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

1.5 Wallabies – All Blacks supporters

In terms of Wallabies – All Blacks example, the sampling theory indicates that the estimation of the proportion of the Wallabies supporters is 50%.

The standard deviation of the proportion is:

$$sd = \sqrt{0.5 * (1 - 0.5)} = 0.5$$

The standard error of the proportion is:

$$se = 0.5 / \sqrt{10} = 0.158$$

In terms of confidence interval, overall (see paragraph 1.3.4 and Figure 3):

- Around 38% of all the possible samples will estimate the proportion of Wallabies supporters between 42 and 58%, or within ½ of a standard error from the mean;
- About 68% of all possible samples will estimate the proportion of Wallabies supporters between 34.1 and 65.8%, corresponding to 1 standard error;
- and 83.4% of all possible samples will estimate the proportion between 19 and 81%, or 2 standard deviations[12].

As we will see in the sample size calculation section later, these precisions figures depend on the prevalence of the observed characteristic in the population and the size of the sample. A larger sample would generate more reliably precise estimates because the standard error would be smaller.

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[13] k, and ii) where on the list 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 skip interval k = N/n, where an integer k is assumed.
2. Select a random integer "a" with an equal probability of 1/k between 1 and k. In Excel, this can be achieved by rounding up the result of "=rand()*k".
3. Select elements numbered a, a + k, a +2k, a +3k, ..., a + (n−1)k 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 shown in Figure 4 we can see that N/n is no longer an integer, and produces the result k = 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 k = 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 k 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.

---

*[12] If p = 0.5, z = 1.96 and se = 0.158, then C.I. = 0.5 +/- 1.96 * 0.158 = 0.5 +/- 0.31.*

[13] May also be referred to as the skip interval, or simply the skip, or the sampling interval.

Figure 4. Example of systematic sampling when N/n is not an integer

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)

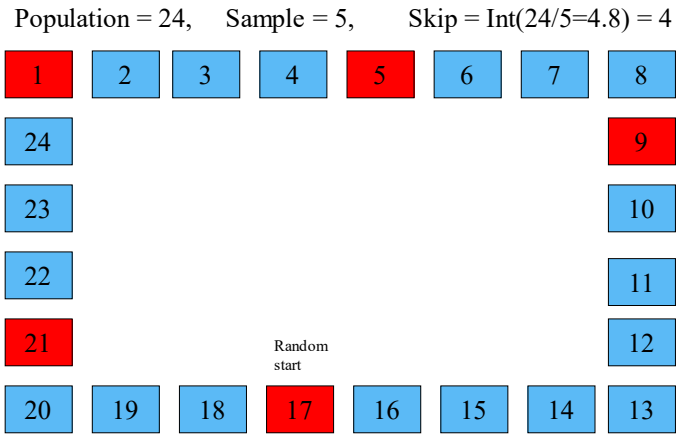
2.3.2 Apply Circular Sampling

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 skip interval k = N/n – rounding down to the integer nearest N/n. For example, if N = 24 and n = 5, then k is taken as 4 and not 5. In Excel, this is "=Int(24/5)".
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 Figure 5 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 k = 4.

Figure 5. Example of circular sampling



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

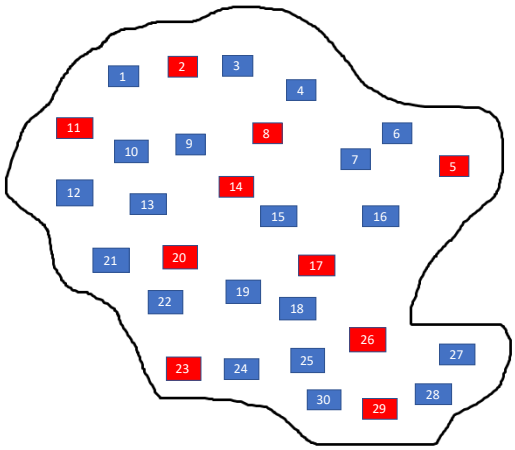
If the sorting order of the sampling frame can be assumed random with respect to the variables of interest and all auxiliary variables, the sample selected will correspond with that of SRS-WOR.

If the sampling frame – say, HH list – is sorted by an auxiliary variable (say, income) 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 achieve 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 HHs within previously selected small geographical areas. When this is the case, the HHs 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 not to select HHs next to each other for safety reasons. Systematic sampling can help reduce the likelihood of this occurring.

See Figure 6 for a graphical representation of how a HH listing may have been carried out in a village, and thus how a systematic sample helped ensure good representation of s throughout the village. In this diagram you can see how the HHs were numbered in the list, with a sample of 10 HHs chosen from the list of 30. The 10 HHs chosen in the village and highlighted in red, are nicely spread-out throughout the village, which is desirable.

Figure 6. Systematic sampling



2.5 Estimation using Systematic Sampling

Under systematic sampling, estimators for the population total Y, mean $\bar{Y}$ and proportion P are the same as SRS, and are thus as follows:

2.5.1 Estimate of Mean

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

2.5.2 Estimate of Total

$$y = N\bar{y} = N \sum_{i=1}^{n} \frac{y_i}{n} = \sum_{i=1}^{n} w_i y_i$$

where: $w_i = {}^{N}/_{n} = {}^{1}/_{\pi_i}$ is the **selection weight** for each sample unit, i

2.5.3 Estimate of Proportion

$$p = \frac{1}{n} \sum_{i=1}^{n} y_i$$

where: $yi = 1$ if sample unit i has the characteristic,

$yi = 0$ otherwise

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 representative 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 discussed in the next paragraph, 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

Stratified sampling involves dividing the target population into non-overlapping sub-populations 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:

$$N_1, N_2, \ldots, N_h, \ldots, N_H$$

such that their sum is equal to N, the total population size:

$$\sum_{h=1}^{H} N_h = N$$

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:

$$n_1, n_2, \ldots, n_i, \ldots, n_H$$

and their sum is equal to n, the overall sample size:

$$\sum_{h=1}^{H} n_h = n$$

Common examples of stratified sampling include:

- For establishment surveys – stratification by economic activity and by employee size.
- For household surveys – stratification by geographic areas (e.g. regions, provinces), by urban/rural areas and by socio-economic groups.
- For agricultural surveys – stratification by agri-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 sub-populations 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 variables of interest.
4. Allowing for flexible stratum-wise use of auxiliary information for sampling. For example, in a survey of school pupils, two-stage sampling can be used in the stratum of large schools, whilst single-stage, systematic sampling can be used in the stratum of small schools.

If dividing the sample into strata leads to differing probabilities of selection between strata, this results in an ***unequal probability sampling design***, and the analyst will need sampling weights to generate representative estimates.

3.2 Applications of stratified sampling

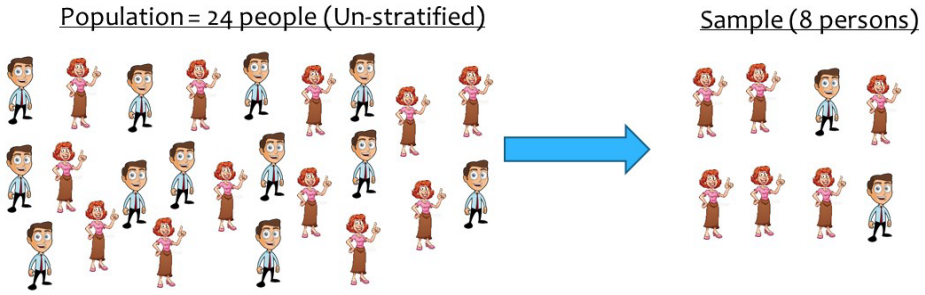
3.2.1 Example 1: Stratifying by sex

One of the reasons why stratified sampling is so popular is its ability to protect against outlier draws. In the example shown in Figure 7, we have a simple population of N = 24 people (12 male and 12 female), for which we wish to sample $n = 8$ 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 males in the sample, thus having too many females compared to the true value.

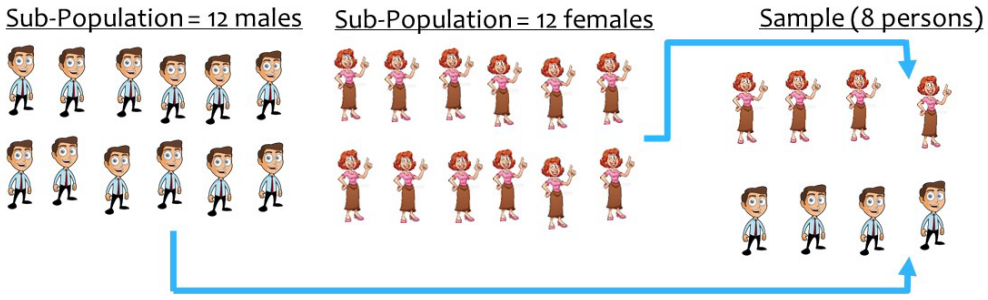
In scenario 2, we now stratify our population in to two sub-populations (N_1 = 12 males and N_2 = 12 females). Using stratification, we can now select the samples separately in both, and intentionally select exactly n_1 = 4 males and n_2 = 4 females from each population, thus guaranteeing a better representation of each sex.

Figure 7. Stratified vs. Non-stratified sampling

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% of the national population and a diverse

international economy, and rural areas with the remaining 25% 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% of the population would be in the urban sample and 25% 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 into 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 single right or wrong answer.

3.2.3 Example 3: Stratifying by administrative region

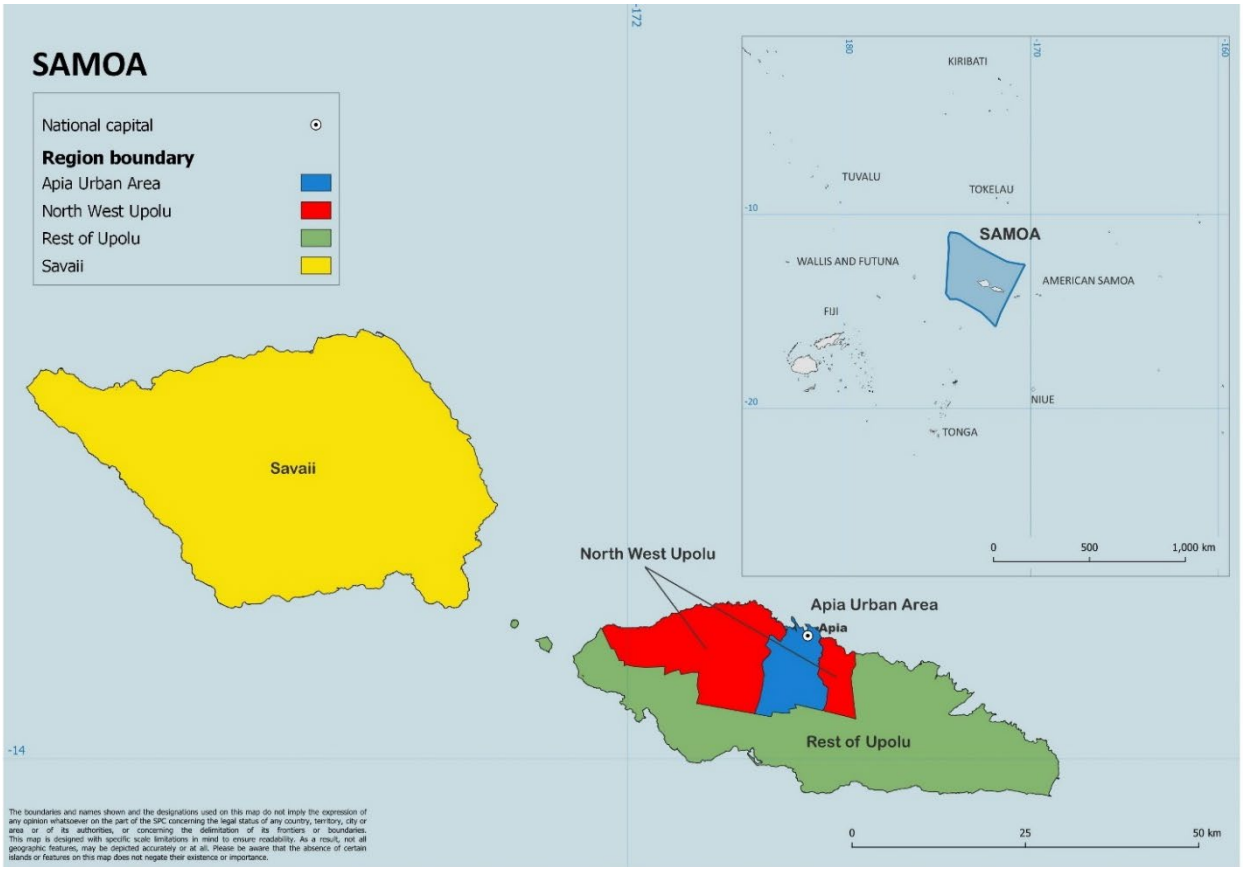
Samoa

Another common application of stratification in the Pacific is to stratify by geographical region. As illustrated in Figure 8, Samoa usually applies four levels of stratification to their HH 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 sample estimates 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.

Figure 8. Stratification by geographical region. The case of Samoa

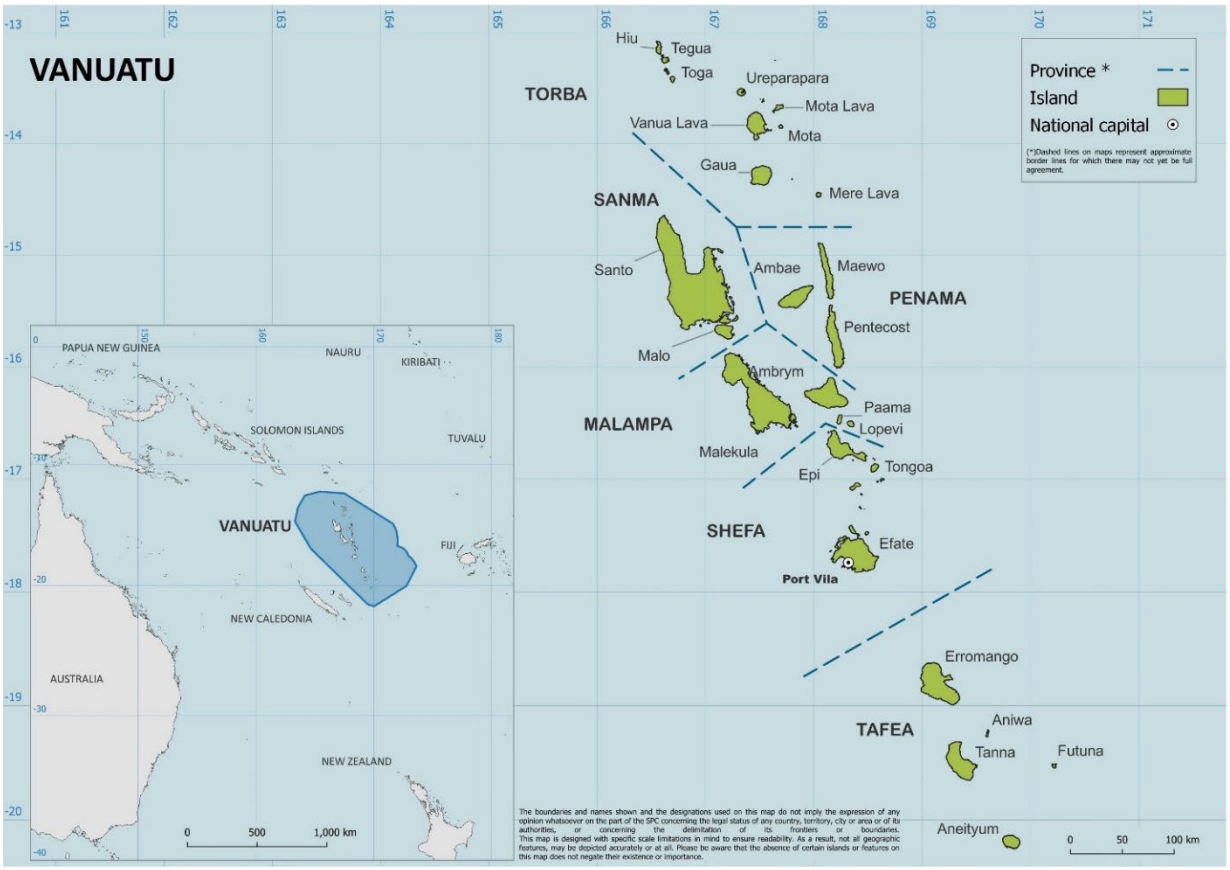


Vanuatu

Figure 9 shows the six main provinces which make up Vanuatu. For many HH surveys covered in Vanuatu, the National Statistics Office adopts eight strata, with the provinces of Sanma and Shefa being split into two (urban and rural), with the urban centers being Luganville and Port Vila, respectively.

Once again, as with Samoa, estimates 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.

Figure 9. Stratification by geographical region. The case of Vanuatu



Differing sample fractions

In both the Samoa and Vanuatu examples, as also discussed in paragraph 3.2.3 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 of roughly 1,960 HHs, was 320 HHs, which is approximately 16% of the total number of HHs. For Malampa, one of the larger provinces with a population of roughly 8,900 HHs, the suggested sample size was 480 HHs, which is approximately 5%. These differing sample fractions ensure estimates of acceptable precision 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 estimates of a population mean, total and proportion, 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 next formula assumes either SRS or systematic sampling is applied within each stratum.

3.3.1 Estimate of Mean

$$\bar{y} = \sum_{h=1}^{H} \frac{N_h \bar{y}_h}{N}$$

where:

$$\bar{y}_h = \sum_{i=1}^{n_h} \frac{y_{h,i}}{n_h}$$

3.3.2 Estimate of Total

$$y = \sum_{h=1}^{H} y_h$$

where:

$$y_h = N_h \bar{y}_h = N_h \sum_{i=1}^{n_h} \frac{y_{h,i}}{n_h} = \sum_{i=1}^{n_h} w_h y_{h,i}$$

and where, for each sample unit, i , in stratum h:

$y_{h,i}$ is the observed value of the variable of interest, and

$w_h = {N_h}/{n_h} = {1}/{\pi_h}$ is the selection weight

3.3.3 Estimate of Proportion

$$p = \sum_{h=1}^{H} \frac{N_h}{N} p_h$$

where:

$$p_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{h,i}$$

and where:

$y_{h,i} = 1$ if sample unit i in stratum h has the characteristic, and

$y_{h,i} = 0$ otherwise

3.4 Allocation of the sample across strata

There are countless ways to allocate sample elements 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 proportional to the number of units in the frame for the stratum. This method is the simplest form of sample allocation and does not require sampling 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 the subscript 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 and Neyman allocation

Optimal allocation divides the sample across strata in such a way as to deliver the most precision (i.e. minimize the standard error) at the national level, for the least cost. The formula for optimal 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 of the variable of interest 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:

$$n_h = n \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h},$$

a special case of optimal allocation known as Neyman allocation.

Neyman was a very common method of sample allocation historically; more recently, however, it is less used 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 Sustainable Development Goals (SDGs) indicators at national and sub-national level;
- a good allocation over urban/rural strata for the estimate of average HH income may differ from a good allocation for the estimate of poverty rate.

So, caution is needed when using Neyman allocation for surveys that have multiple objectives, or when the sample estimates of S_h are subject to large sampling errors themselves. Despite this, Neyman allocation still serves as a useful starting point for many designs and therefore is one of the most commonly used tools in sample design.

Similarly to equal allocation, the differing probabilities of selection across the strata under both Optimal and Neyman allocation necessitate the use of sampling weights.

3.4.2 Practical allocation

In current national HH 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 strata until all reach a minimum level of precision.

4 Multi-Stage Sampling

4.1 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 HH surveys it is common to adopt a 2-stage process of selecting HHs, where small geographical areas are selected at the first stage (Primary Sampling Units – PSUs[14]), and then a sample of HHs chosen from within each selected geographical area. Multi-stage sampling is used in nearly all official HH 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 HHs within stratum directly using SRS or systematic random sampling. As it will be discussed later, 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 (e.g. HHs). 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 only 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 into what are commonly referred to as Enumeration Areas (EAs)[15] that, generally speaking, correspond to the workload of one enumerator during the census data collection period. A list of EAs is produced and maintained between censuses, and generally used as the frame for selection of PSUs for HH surveys in a two-stage sample design for countries following a multi-stage methodology. A sample of HHs 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.

4.2 Surveys in the Pacific with a third stage of selection

Although many HH surveys in the Pacific only require a two-stage selection process, there are situations when just one member of each selected HH 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 HH members eligible to participate.

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

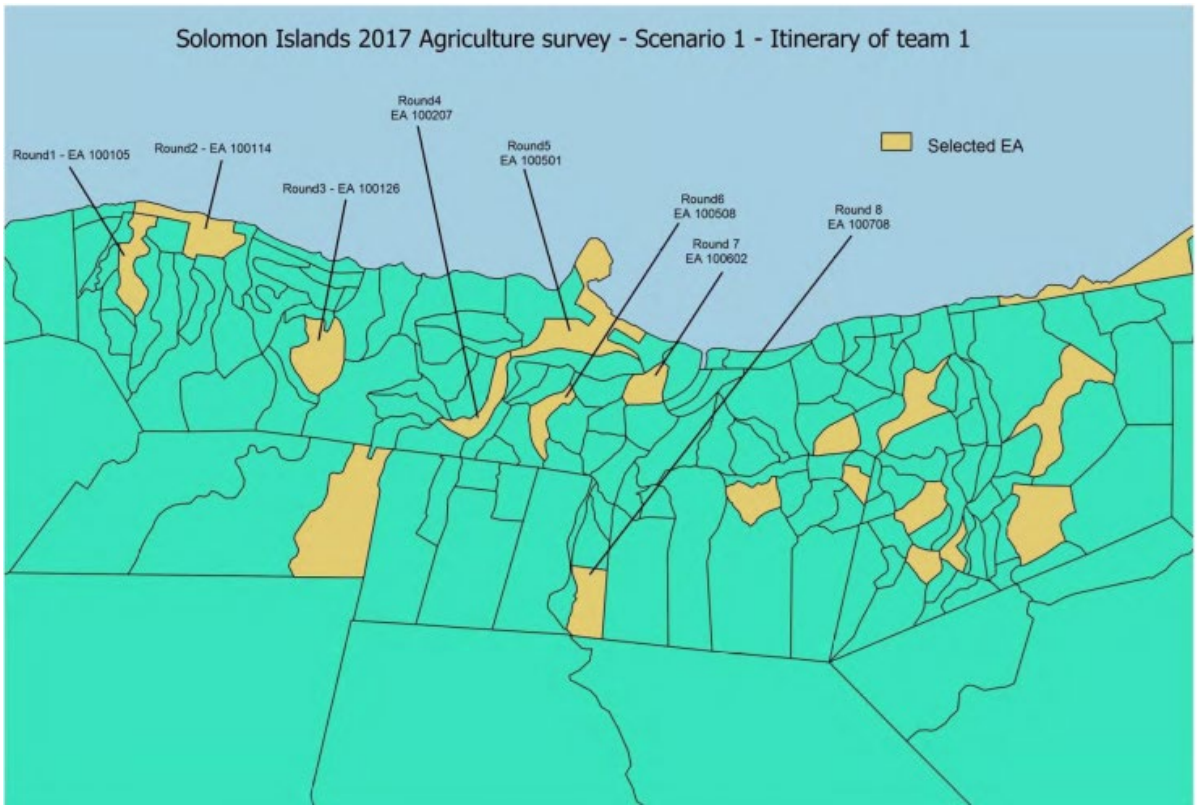
Figure 10 and Figure 11 show an example of the selection of first stage units for an Agriculture Survey in the Solomon Islands. The first map shows the selection of EAs in Honiara, which are represented by the yellow patches (selected EAs). Given the EAs were listed based on geographical position and selected using a form of systematic random sampling, the sample is dispersed across Honiara.

---

[14] PSUs may also be referred to as Primary Selection Units, or Primary Survey Units.

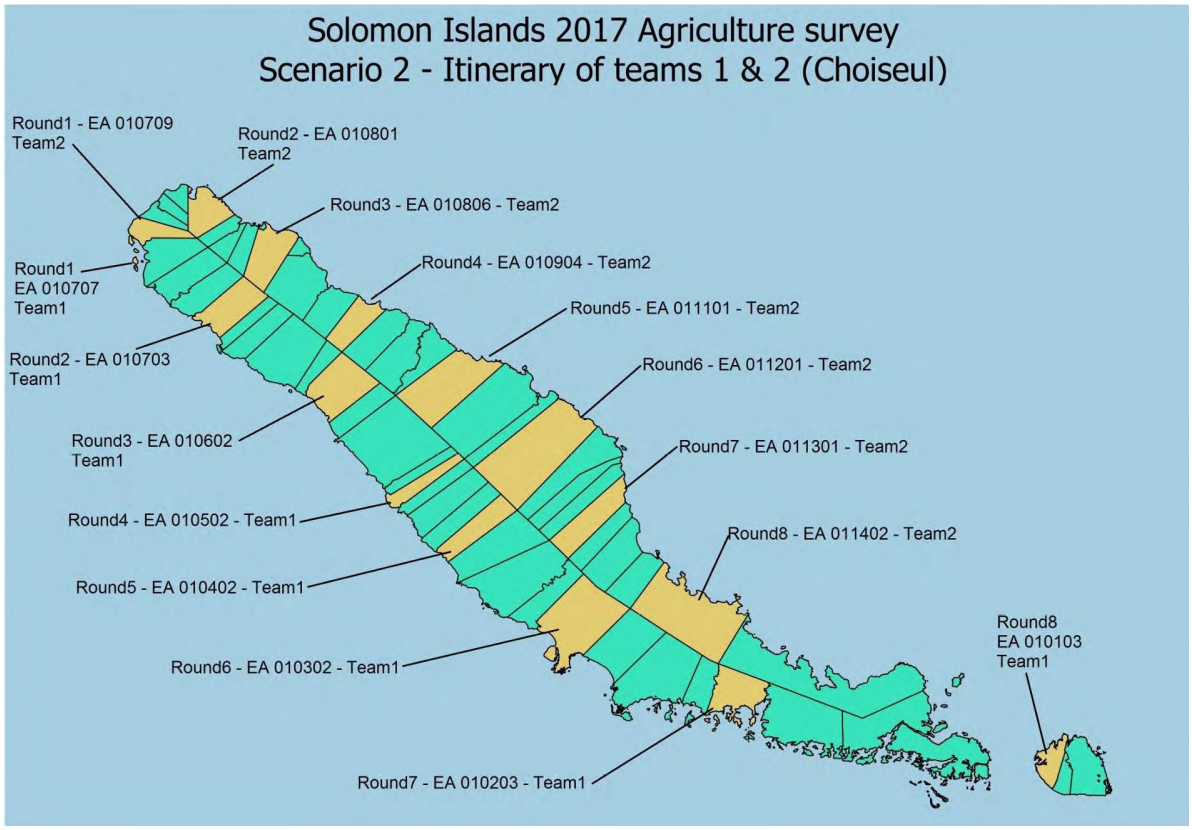
[15] In some Pacific countries – for example, Tonga - these are referred to as "census blocks".

Figure 10. Selection of Enumeration Areas (EAs) in Honiara, Solomon Islands, for an Agriculture Survey



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.

Figure 11. Selection of Enumeration Areas (EAs) in Choiseul, Solomon Islands, for an Agriculture Survey



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

4.4 Introduction to Probability Proportional to Size Sampling (PPS)

Probability proportional to size (PPS) sampling is a sampling method in which larger population elements (EAs, PSUs, businesses) 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_i of the auxiliary size variable Z is known for every population element i. 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 HH surveys conducted in the Pacific, PPS sampling often uses the number of HHs in a PSU (EA or Census Block) as the size measure. The number of HHs, rather than the total population, is more commonly used because HHs 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, with sample size n, the probability of selection for unit or element i of size Z_i is:

$$\pi_i = n \times \frac{Z_i}{Z}$$

where:

$$Z = \sum_{i=1}^{N} Z_i$$

where Z is the sum of the size measures Z_i over all N units in the population. In PPS, the probabilities of selection π_i vary between units and thus, PPS is an unequal probability sampling design. Sampling weights are therefore required for the analysis.

4.5 Household Survey Example: Probability Proportional to Size (PPS) Sampling

Many HH surveys conducted in the Pacific select EAs in the first stage using stratified PPS sampling and then select HHs in the second stage using systematic sampling from updated administrative lists or a HH listing operation within the selected EAs only. The practical process for selecting a PPS sample of EAs (say, in Excel) at the first stage of sample selection, within a single stratum, is illustrated in Figure 12 and described in the example below.

Step 1

Sort or order the *n* EAs (26 in this case) representing this stratum by geographical position. Such 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 Z_i (in this case the number of HHs) 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 k and random start for identifying the selections. In this example, it has been pre-determined that 6 EAs will be selected from the stratum, so the skip can be calculated as k = 1,365/6 = 227.5, where 1,365 is the number of HHs. The random start has then been generated between 0 and 227.5, with a value of 14.04546.

Step 4

Determine the six selection numbers required to select the six EAs in the sample (4th column in Figure 12). 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 1,155.

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 cumulative number of HHs ("Cum # HHs" in Figure 12) greater than the selection number, for which the previous EA has a "Cum # HHs" less than the selection number. For example, the second selection EA in the example below was EA 100105, as the selection number 243 is lower than 281 but greater than 237.

Figure 12. Practical steps for selecting a PPS sample

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

Seln = selection

Case Study 3 (in Appendix 3) illustrates an alternative practical process for selecting EAs using stratified PPS sampling in the first stage of a two-stage HIES sample design, using the same scenario as in the example above (a single stratum containing 26 EAs of varying sizes, from which 6 EAs are to be selected).

An advantage of the alternative approach shown in Case Study 3 is that the first-stage probabilities of selection for each EA (π_i) are calculated as part of the sample selection process. It is helpful to calculate and retain these probabilities upfront, as they will be a critical input when it comes time to calculate weights and produce estimates for the survey.

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

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 strata). Once the EAs have been selected using PPS sampling for this first stage, it is common practice at the second stage to select a fixed number of HHs from each selected EA, despite the size of the EA. Whilst the approach of selecting a fixed number of HHs from each selected EA helps with allocating even workloads across field staff, it also has other benefits in that it helps give HHs a similar chance of selection in the survey. This feature can be seen in Figure 13: a stratum has 15 EAs, of which we have selected 5 EAs during the first stage. Now let's assume we decide to select a fixed number of HHs per EA, say 15, at the second stage: we refer to this number as the cluster size. Adopting this approach gives each HH an equal probability of selection, if the size of the EA (with respect to the number of HHs) does not change.

Figure 13. Selecting a fixed cluster size of households at the second stage

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

Seln = selection

The probability of selection for HH i is calculated as follows:

> Prob (Household i selected in the survey) = Prob (EA containing Household i selected in Stage 1) x Prob (Household i selected in Stage 2)

Or, more generally for two-stage sampling:

> π_i = Prob (selecting the PSU in 1st stage) x Prob (selecting the Unit in 2nd stage)

In the example above, for the 5 selected EAs, the probabilities of a HH being selected from those EAs, with a skip of 127.4, is calculated as follows:

Table 2. Calculation of probability of selection for a generic household i

EA	Prob (EA containing HH i selected in Stage 1)	Prob (HH i selected in Stage 2)	Prob (HH i selected in the survey)
10403	47/127.4	15/47	= (47/127.4) x (15/47) = 15/127.4 = 0.1177
10405	56/127.4	15/56	= (56/127.4) x (15/56) = 15/127.4 = 0.1177
10408	42/127.4	15/42	= (42/127.4) x (15/42) = 15/127.4 = 0.1177
10411	37/127.4	15/37	= (37/127.4) x (15/37) = 15/127.4 = 0.1177
10415	40/127.4	15/40	= (40/127.4) x (15/40) = 15/127.4 = 0.1177

As shown in table 2 above, all HHs have the same chance of being selected in the survey (π_i= 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 sampling was adopted, but HHs within those EAs then have less chance of being selected, because a fixed number of HHs (cluster size) is selected.

In practice however, HH lists for each selected EA are generally updated after the first stage, and the fixed cluster of HHs selected during the second stage of selection are chosen from this updated list. When this occurs, the probabilities of selection for each HH will be the same within an EA, but different across EAs. For example, in the illustration above, if EA 10403 had its HH listing updated once it was selected and it was found that 52 HHs were now in the list (an increase over the original 42 HHs), then the probability of a HH from that EA being selected would now be:

Table 3. Adjusted probability of selection for a generic household i, after updating the household list

EA	Prob (EA containing HH i selected in Stage 1)	Prob (HH i selected in Stage 2)	Prob (HH i selected in the survey)
10403	47/127.4	15/**52**	= (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 HHs identified.

4.7 Estimation using two-stage sampling

4.7.1 Estimate of Mean

$$\bar{y} = \sum_{i=1}^{n} \frac{w_i y_i}{N} = \sum_{i=1}^{n} \frac{y_i}{N\pi_i}$$

where:

π_i = Prob (selecting the PSU in 1st stage) x Prob (selecting the Unit in 2nd stage)

4.7.2 Estimate of Total

$$y = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

where:

π_i = Prob (selecting the PSU in 1st stage) x Prob (selecting the Unit in 2nd stage)

4.7.3 Estimate of Proportion

$$p = \sum_{i=1}^{n} \frac{w_i y_i}{N} = \sum_{i=1}^{n} \frac{y_i}{N\pi_i}$$

where:

$y_i = 1$ if sample unit i has the characteristic, and

$y_i = 0$ otherwise

π_i = Prob (selecting the PSU in 1st stage) x Prob (selecting the Unit in 2nd stage)

4.8 Estimation using Probability Proportional to Size Sampling

The following are the formulae for estimating a population mean, total and proportion when PPS sampling is used in isolation, with sample size n. The formulae and probabilities of selection π_i need to be adapted further when stratified PPS is used – for example, when PPS is used as part of a two-stage stratified cluster design.

4.8.1 Estimate of Mean

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{N} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{N}$$

where:

$$\pi_i = n \times \frac{Z_i}{Z} \text{ and } Z = \sum_{i=1}^{N} Z_i$$

4.8.2 Estimate of Total

$$y = \sum_{k=1}^{n} w_i y_i = \sum_{k=1}^{n} \frac{y_i}{\pi_i}$$

where, as mentioned above:

$$\pi_i = n \times \frac{Z_i}{Z} \text{ and } Z = \sum_{i=1}^{N} Z_i$$

4.8.3 Estimate of Proportion

$$p = \frac{\sum_{i=1}^{n} w_i y_i}{N} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{N}$$

where:

$y_i = 1$ if sample unit i has the characteristic, and

$y_i = 0$ otherwise

$\pi_i = n \times \frac{Z_i}{Z}$ and $Z = \sum_{i=1}^{N} Z_i$

5 Introduction to sample size calculations

Sample size calculations determine the minimum number of sample elements required to achieve a minimum level of precision for the desired sample 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 (SRS)

Sample size calculations for a simple random sample (SRS) design require three pieces of information: the desired confidence level for the sample estimate, the variance of the variable of interest, and the maximum acceptable margin of error (ME) on the sample estimate. The formula for the required sample size differs slightly if the variable of interest is a continuous variable, Y (like HH consumption/ expenditure, income, or years of education) or a proportion, P (share of individuals with post-secondary education, share of HHs with electricity, or share of HHs with a mobile phone). The two formulas are shown in Table 4 below:

Table 4. Sample size calculation for a Simple Random Sample

Continuous variable	Proportion
$\boldsymbol{n_\infty = \frac{t_\alpha^2 \times S^2}{ME^2}}$	$n_\infty = \frac{t_\alpha^2 \times P(1-P)}{ME^2}$

In these formulas, the variance of the variable of interest is represented by either S^2 – 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 commonly used value for the confidence level α is 95%, which corresponds to a t-value of 1.96. There are other t-values for higher or lower levels of required precision. Finally, ME represents the maximum margin of error that is acceptable on the sample estimate.

An important point to note about the margin of error (ME) is that it must be in the same units as the variance. For example, if the variance is in dollars, then the ME should also be in dollars. A common error is to define ME as a percentage. Consider the example below.

Table 5. Calculation of the sample size: the right and wrong Margin of Error (ME)

Mean = $\boldsymbol{\bar{Y}}$ = 500 AUD; $\boldsymbol{S^2}$= 125 AUD^2; $\boldsymbol{\alpha}$ = 95%; $\boldsymbol{t_\alpha}$=1.96; $\boldsymbol{ME}$ = 5% of mean	
$\boldsymbol{n_\infty = \frac{t_\alpha^2 \times S^2}{ME^2} = \frac{1.96^2 \times 125^2}{25^2} \approx 96}$	$n_\infty = \frac{t_\alpha^2 \times S^2}{ME^2} = \frac{1.96^2 \times 125^2}{0.05^2} \approx 24{,}000{,}000$

In the equation on the left, the analyst correctly multiplied the mean $\bar{Y}$ = 500 by the percentage ME = 5% of mean, to obtain an ME value of 25 in AUD, and the calculations required a sample size of 96. In the equation on the right, the analyst erroneously used the percentage ME as-is, 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 required for an infinite population to the sample size required for

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 (finite) 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 HH 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*. However, in Pacific countries with smaller populations, the *fpc* may have quite an impact, and using it may lead to a significant reduction in the sample size required for a HH survey. See the three examples in Table 6 for n_∞ = 96.

Table 6. Examples of the impact of the finite population correction on final sample size

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

5.2 Sample size calculations for a stratified SRS sample design

Sample size formulae and calculations for a stratified SRS sample design are more complex than for a simple random sample design (SRS). 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 variable 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, ME_h, can be calculated for each stratum. Stratified sampling involves SRS within each stratum, so the formulae for ME_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:

Table 7. Calculation of the Margin of error in stratified sampling.

Continuous variable	Proportion
$\boldsymbol{ME_h = t_\alpha \times \sqrt{\frac{S_h^2}{n_h}} = t_\alpha \times \frac{S_h}{\sqrt{n_h}}}$	$ME_h = t_\alpha \times \sqrt{\frac{P_h(1-P_h)}{n_h}}$

The resulting margin of error at the overall or national level can then be calculated as follows, using a continuous variable as the example:

$$ME = t_\alpha \times \sqrt{\frac{S^2}{n}}$$

where, under stratified SRS sampling (again ignoring the fpc):

$$S^2 = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 S_h^2$$

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 margin 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.

Example: First iteration of a stratified sample design.

Initial sample design assumptions:

- two strata (Urban and Rural) with population size, mean and variance as shown in Table 8.
- a maximum overall sample size of n = 500, with Neyman allocation of the sample across strata
- α = 95%; t_α= 1.96

Table 8. First iteration of a stratified sample design

Stratum ($\boldsymbol{H}$ = 1,…,H)	$\boldsymbol{N_h}$	Mean = $\overline{\boldsymbol{Y}}$ (in AUD)	$\boldsymbol{S_h^2}$ (in AUD)
Urban	7,500	1,000	100^2
Rural	2,500	500	30^2
National	10,000	800	

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

Table 9. Final sample sizes and margins of error at the stratum level

Stratum ($\boldsymbol{H}$ = 1,…,h)	$\boldsymbol{N_h}$	Mean = $\overline{\boldsymbol{Y}}_{\boldsymbol{h}}$ (in AUD)	$\boldsymbol{S_h^2}$ (in AUD)	$\boldsymbol{n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}}$ where $\boldsymbol{S_h = \sqrt{Var_h\,(Y)}}$	$\boldsymbol{ME_h = t_\alpha \times \sqrt{\frac{S_h^2}{n_h}}}$
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:

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

$$S^2 = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 S_h^2 = \left(\frac{7{,}500}{10{,}000}\right)^2 100^2 + \left(\frac{2{,}500}{10{,}000}\right)^2 30^2 = \mathbf{5{,}681.25}$$

$$ME = t_\alpha \times \sqrt{\frac{S^2}{n}} = 1.96 \times \sqrt{\frac{5{,}681}{500}} = \mathbf{6.6}$$

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.

Chapter 3 - Household Income & Expenditure Surveys

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

National HIES are conducted every 5 years to measure a range of socioeconomic variables. The distinctive feature of the HIES is that it collects consumption/expenditure 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 weights, and to measure well-being.

In terms of the sample requirements for measuring HH well-being, the sample design is developed around being able to produce estimates with a given level of precision for the relevant indicator or parameter in the target population. The relevant indicator for a HIES is either mean HH per capita[16] consumption/expenditure or mean per capita income, $\bar{y}$ (with the associated variable of interest y being HH consumption/expenditure per capita or personal income, respectively). This 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 HH well-being is often correlated with other indicators, including those related to education, health, livelihoods, etc.

In terms of Sustainable Development Goals (SDGs), HIES is the preferred source of data to measure the following indicators[17]:

Goal 1: End poverty in all its forms everywhere.

- Indicator 1.1.1 - Proportion of population below the international poverty line, by sex, age, employment status and geographical location (urban/rural).
- Indicator 1.2.1 - Proportion of population living below the national poverty line, by sex and age.
- Indicator 1.2.2 - Proportion of men, women and children of all ages living in poverty in all its dimensions according to national definitions.

Goal 2: End hunger, achieve food security and improved nutrition and promote sustainable agriculture.

- Indicator 2.3.2 - Average income of small-scale food producers, by sex and indigenous status.

Goal 10: Reduce inequality within and among countries.

- Indicator 10.1.1 - Growth rate of HH expenditure or income per capita among the bottom 40% of the population and the total population.
- Indicator 10.2.1 - Proportion of people living below 50% of median income, by age, sex and persons with disabilities.

In addition, HIES can be used as a secondary source for several other SDG indicators, and may also be used to contribute to or measure some SDGs where it is neither the preferred nor the secondary source.

The HIES target population is people living in HHs (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.

2. Basic design: Stratified two-stage cluster design

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

- stratified SRS sample design, with simple or systematic random sample of HHs within each stratum;

---

[16] 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.

[17] Although secondary sources such as the Agriculture Census or Disability Survey can be used instead to measure some of these indicators.

- two-stage stratified cluster sample design, with PPS sampling of EAs at first stage, followed by the selection of a fixed number (cluster) of HHs within each selected EAs at the second stage.

Depending on the size of the population and the availability of updated administrative statistics, however, some countries may forego the clustering.

Appendix A contains a sample design Case Study for a HIES that uses a stratified SRS (without clustering). Appendix B contains a Case Study for a HIES using a two-stage stratified cluster sampling.

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 variable of interest as well as the necessary cluster and stratification variables. Because the HHs 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 HHs could have been allocated differently to produce a more precise estimate, given the same sample size.

The key ingredients from past HIES data needed for the new HIES sample design are: the sample size n, the sample estimate of mean per capita expenditure $\bar{y}$, the standard error of the mean $SE_{complex}(\bar{y})$, the cluster size m, and the design effect $deff$. The ***design effect (deff)*** measures the statistical efficiency of a sampling design with respect to simple random sampling (SRS), and is given by:

$$deff(\bar{y}) = \frac{Var(\bar{y})}{Var_{SRS}(\bar{y})}$$

where the numerator is the sampling variance of the mean $\bar{y}$ under the actual (possibly complex) sampling design, and the denominator represents the sampling variance under an assumption of simple random sampling for a sample of the same size n.

The formula for $deff$ gives rise to the following remarks:

(a) $deff < 1$ The actual sampling design is more efficient than SRS.

(b) $deff = 1$ The efficiency of the actual sampling design is the same as that of SRS.

(c) $deff > 1$ The actual sampling design is less efficient than SRS.

Each piece of information required for the sample design should be calculated at the level of the strata for the new HIES sample design, even if a different stratification was used in the analysis of the data from the previous sample design.

This information is then used to calculate the necessary components for the new HIES sample design. The sample size n, the standard error of the estimate from the previous survey $SE_{complex}(\bar{y})$, and the design effect $deff$ are used to estimate the standard deviation s of the variable of interest y (per capita HH expenditure for HIES). The standard deviation s, along with the Intra-cluster Correlation Coefficient[18] (ICC or ρ) – which itself is calculated from the design effect $deff$ and cluster size m - are used to calculate the required sample size n for the new survey. With this information, the sample designer can predict the precision of the sample design for the new survey.

2.2 Choosing the stratification

Typically, stratification attempts to group relatively homogeneous or similar population elements together. In the case of the HIES, a standard design defines the strata using administrative divisions split into urban and rural areas,

[18] *May also be referred to as the Intraclass Correlation Coefficient.*

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."[19] Analytical strata represent the lowest level at which the sample estimates 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 and its (say, 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 HIES sample designs, the precision is measured with the relative standard error (RSE), which is defined as the standard error divided by the mean:

$$RSE(\overline{y}) = \frac{SE(\bar{y})}{\bar{y}}$$

HIES sample designers usually target a RSE of around 5%, though RSEs are sometimes as high as 10% 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 of the population variation and increase the overall precision. For example, one region may be made up of 2 islands that are very different. One island may have a diverse economy and great variation in living standard between the HHs. 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 HHs. If the population of the 2 islands is the same, we would end up with roughly the same number of clusters and HHs 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 HHs. In contrast, because the first island has lots of variation, comparatively more information is gained by having more clusters and HHs in the first island. Therefore, the sample designer may want to explicitly decide to sample more HHs on the first island than on the island of fishermen, even though in the final tables, the estimates will only be presented at the level of the region. In that case, the sample design would include design strata at the island level within the analytical stratum of the region.

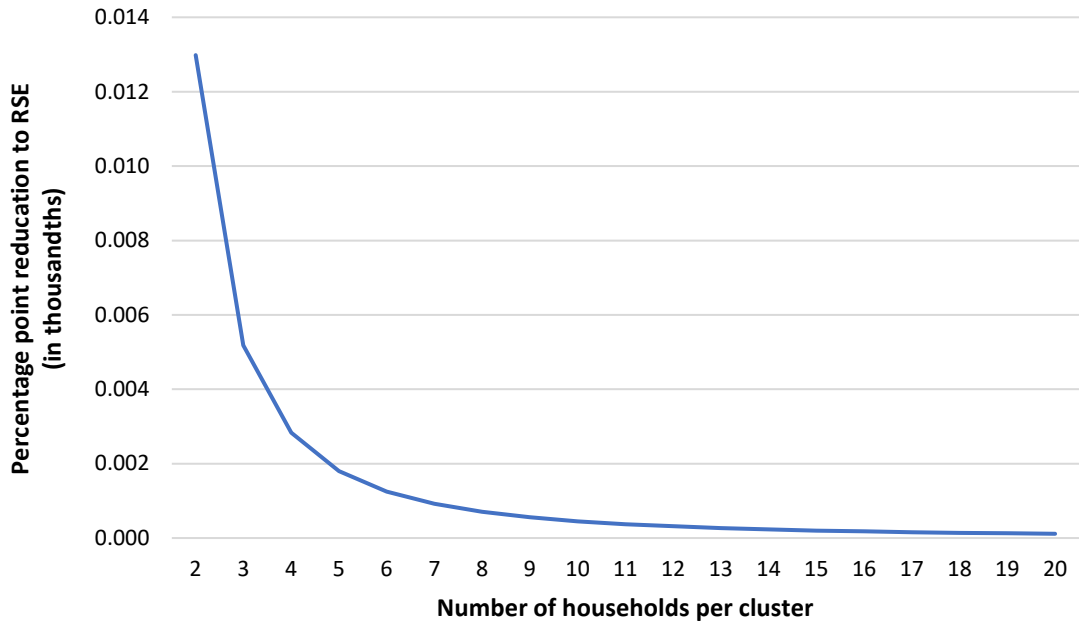
2.3 Defining cluster size (which can be different for different areas/strata)

The intra-cluster correlation coefficient (ICC) calculated from the previous survey is used to understand the relationship between the number of HHs per cluster and the relative standard error (RSE). Figure 14 shows the reduction of RSE at the national level for each additional HH per cluster in a simulated sample design with 500 clusters allocated with probability proportional to size (PPS).

---

[19] "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%, for example.

Figure 14. Percentage point reduction in national RSE per additional Household



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

The largest contribution to the precision of survey estimates comes from the first additional HH (moving from a cluster size of 2 to 3), reducing the RSE by 0.005 percentage points. Each additional HH applied to the cluster in the simulation reduces the RSE less. In this example, the marginal benefit of an additional HH becomes limited around 8 HHs per cluster, and basically non-existent after 12 HHs. In general, for a fixed sample size, sampling more smaller clusters over fewer larger clusters increases the level of precision of the resulting estimates.

3. First stage

3.1 Defining clusters

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

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 the size of each EA, in terms of number of HHs. Since HIES surveys select HHs instead of individuals, it is important that the list has the total number of HHs instead of total number of individuals. Also, if possible, the list should be updated with the latest population counts, though it is possible to proceed even if the list is dated. The sampling frame should be carefully checked for duplicates; EAs are generally include between 150 and 200 HHs, although in some Pacific countries the average size of an EA is 50 – 70 HHs. Any EA with outlier populations should be verified. It is important to start the process of verifying the sampling frame in the early stages, as checking outliers often requires contacting regional offices and can be a time-consuming process.

3.3 Selecting EAs

It is standard practice in HH surveys to select the enumeration areas within each stratum by using probability proportional to size (PPS) 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 of EAs by geographic location 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 HHs is a key component in generating unbiased estimates. Generally, there are two options: administrative statistics or conducting a HH listing operation. In either case, there are two objectives: to produce a list of HHs from which survey and replacement HHs can be drawn, and to have an accurate count of HHs as an input into the weight calculations.

4.1 Administrative statistics

In some cases, there will be administrative statistics available on which to base the selection of HHs. 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 HHs within the selected EA. It is important that there be a way to exclude HHs living outside of the selected EA if the administrative lists cover more than one EA. Missing information, such as HHs that have recently moved to the area and are not included on the administrative list, could lead to bias in the resulting estimates.

4.2 Household listing

If administrative lists are not available, a HH listing operation is required. To implement an accurate HH 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 HHs 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 HHs living in canvased areas, the missed HHs 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 be low as well. Similarly, if the counts are too high - if a team lists HHs 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 HHs. The division should be done using the EA map and using clearly defined landmarks (such as roads and rivers) as boundaries. Once segmented, one segment is randomly selected, and the listing process continues as previously discussed. 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 the fieldwork, it is inevitable that there will be certain selected EAs or HHs in which it is not possible to conduct interviews. In an ideal case, the sample design would have selected additional EAs and HHs to compensate for those that are not interviewed. However, this is rarely done in practice 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 HH.

5.1 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 cyclone." 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.

5.2 Household level replacements

Most replacements, however, take place at the HH 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 HH can be replaced with another randomly selected HH in the EA without introducing bias. If the HH, however, refuses to participate, even a randomly selected replacement will introduce bias into the sample. This is because those HHs 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 HH refusal is still unavoidable. In this case, due to high travel costs for most HH surveys implemented in the Pacific, replacement HHs are selected for HHs that refuse.

The standard procedure for replacing HHs at the EA level is to select more HHs than needed and to hold back some as replacements. If HHs are selected with systematic random sampling, the step should be calculated for the larger number of HHs and then the selected HHs randomly allocated into "selected" and "replacement." In an example where 10 HHs are needed for the survey and 5 as replacements, 15 total HHs 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 HHs remain unused. If a HH refuses, one HH 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 HH. In general, no more than 20% of HHs should be replaced in a given EA, as each replacement introduces some bias into the data. Detailed 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

6.1 Sampling weights

The weight calculations follow the steps of the selection. In the first stage, EAs are selected by stratified probability proportional to size (PPS). Within each stratum h, the first-stage probability of selection is then:

$$\pi_{hi1} = n_h \frac{Z_{hi}}{Z_h}$$

where n_h is the number of EAs selected in stratum h, Z_{hi} is the expected size or population of the EA (in HHs) based on the sampling frame, and $Z_{hi} = N_h$ is the population of strata h (in HHs).

If a selected EA i in stratum h requires segmentation, the probability of selection of one of the segments is:

$$\pi_{hi2} = \frac{1}{seg_{hi}}$$

where seg_{hi} is the number of segments that EA i in stratum h is divided into. For most of the EAs in which segmentation is not necessary, seg_{hi} = 1 and π_{hi2} = 1.

Within each selected EA or segment i, the second-stage probability of selection of each HH j is:

$$\pi_{ij} = \frac{m_h}{Z'_{hi}}$$

where:

m_h is the number of HHs to be selected from each EA in stratum h (i.e. the cluster size in stratum h)

Z'_{hi} is the number of HHs found in the HH listing exercise for EA or segment i in stratum h

The overall probability of selection of HH j, within EA *i* in stratum *h* is:

$$\pi_{hij} = \pi_{hi1} \times \pi_{hi2} \times \pi_{ij}$$

and the selection or sampling weight is:

$$w_{hij} = \frac{1}{\pi_{hij}} = w_{sel}$$

6.2 Non-response adjustment at the household level

It may be necessary to include a non-response correction for a limited number of EAs if a team exhausts all selected HH and replacements before completing their required number of HHs. In this case, it is necessary to assume that all non-responding HHs within an EA are statistically identical to those that chose to respond. Then:

$$w_{nr} = \frac{m_h}{m'_{hi}}$$

where:

m_h is the number of HHs to be selected from each EA in stratum h (i.e. the cluster size in stratum h)

m'_{hi} is the number of HHs that responded to the survey in EA i stratum h

The selection weight adjusted for non-response is then:

$$w_{sel,nr} = w_{sel} \times w_{nr}$$

6.3 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 that uses auxiliary or outside data (e.g. known population totals from a recent census).

With proper design and implementation, the population totals estimated from the sample survey (pop_{survey}) should be close to the known population figures from the auxiliary data source (pop_{known}) in each of the post-strata. Post-stratification should therefore be seen more as a fine-tuning adjustment rather than a major realignment.

To calculate the post-stratification adjustment, the formula is:

$$w_{ps} = \frac{pop_{known}}{pop_{survey}}$$

The final estimation weight would then be $w_{final} = w_{sel,nr,ps} = w_{sel\times}w_{nr} \times w_{ps}$

The level of disaggregation for the post-stratification adjustment should be at the lowest reliable level available from the 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.

However, both the precision of the population totals estimated from the survey and the reliability of the auxiliary data source underlying the post-stratification calculations need to be considered before deciding to use post-stratification. For example:

- if the weights are adjusted using poor quality auxiliary information, there is the possibility of reducing precision or introducing bias into the estimates. This is of particular concern when population projections based on censuses that are a number of years old are used.
- if the auxiliary data source is reliable at a low level, but the precision of the underlying survey estimates at this level is poor (subject to large sampling error), this can result in large post-stratification adjustments and increased variation in the sampling weights.

References

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

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.

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

UN Statistics Division. UNdata: Glossary http://data.un.org/Glossary.aspx

Appendices

A. Case Study 1 – Computation of sample size for a HIES with a stratified SRS sample design.

B. Case Study 2 – Computation of sample size for a HIES with a two-stage stratified cluster sample design.

C. Case Study 3 – Selecting a probability proportional to size (PPS) sample in the 1st stage of a HIES sample design.

D. Glossary of key terminology, notation, definitions, and formulae.

E. Sampling Methods for Core and Additional Modules.

Appendix A: Case Study 1 – Computation of sample size for a HIES with a stratified SRS sample design

In the case of a small island state with a small population (for example, Wallis and Futuna, Nauru, Tuvalu, Tokelau, Palau, Niue), often spread out across different islands, a stratified simple random survey is commonly implemented as the sampling strategy. The selection of HHs straight from the sampling frame is logistically possible in the context of small island countries and will provide the most accurate outputs. In larger countries with higher populations, often also spread out across many islands, stratified simple random sampling alone is not appropriate and will not be possible due to budget constraints.

Considering the type of survey that is to be conducted, it is important to firstly select the main parameter of interest and associated variable(s) of interest in the first place. For example:

- in the case of HIES, mean per capita expenditure would be one of the parameters of interest to be estimated, and the associated variable of interest (measured for each HH in the sample) would be HH expenditure per capita,
- in the case of the Labour Force Survey, the labour force participation rate would be one of the parameters of interest to be estimated, and the associated variable of interest (measured for each person in the sample) would be labour force participation, and
- in the case of MICS, stunting prevalence in children aged under 5 would be one of the parameters of interest to be estimated, and the associated variables of interest (measured for each child in the sample) would be height and weight.

The first case study deals with a stratified sample design and computation of sample size for a HIES conducted in a small island state (for example, Niue, Tuvalu, Tokelau). It shows various stratified sampling strategies and presents all the steps that the survey designer will have to take in order to generate the most efficient sampling strategy.

The second case study shows a complex sampling design (two-stage cluster sampling) for a larger country with higher population spread over a larger number of islands (for example: New Caledonia, Kiribati, Vanuatu, French Polynesia, Republic of Marshall the Islands, Cook Islands, Guam, FSM, Fiji, Tonga, Samoa, Solomon Islands).

Case study 1

1. Retrieve information from the previous HIES: the sample size, the estimate of the parameter of interest (e.g. mean per capita expenditure or mean PCE) and its standard error. In the case of a stratified sample design, this information will be required for each stratum (as well as the sample size and population size at the time the survey was selected). Then use this information to calculate the standard deviation of PCE, and the relative standard error of mean PCE.
2. Update of the sampling frame: it is recommended to update the sampling frame if the last count was conducted more than 2 years ago, or in the case of extreme events, such as a large population migration, or a natural disaster.
3. Use the information from the previous HIES and the update of the sampling frame to predict the best sample size to adopt for the next HIES.

Step 1: looking at the information from the previous HIES data

Information from the previous HIES might be available directly from the HIES report, and if not, it will be important to compute them using the dataset. In this case, the software Stata proposes a survey package that allows the computation of standard errors and design effects given a specific survey design.

The dataset must contain the following information[20]:

- Unique Household ID – *hhid*
- The sampling weight for each household – *hh_weight*
- Strata (associated to survey design) – *strata – e.g. province, island group, island*
- Finite population correction (total number of households within each stratum) – *fpc1*
- Per capita household expenditure (pce) – *pc_hh_exp*

The following Stata syntax shows how to compute the population size, sample size, mean and standard error for per capita household expenditure (PCE) from the previous HIES dataset:

```
*-------------------------------------------------------------------------------------------------------------------------*

use "$survey_dataset\previous_HIES.dta", clear

*1. Calculating population size:

total hh_weight, over(strata)                    //report to spreadsheet the total number of HHs in each strata

*2. calculating sample size:

tab strata                                       //sample size - report to spreadsheet (if no strata, leave blank)

*3. calculating mean, standard error:

svyset hhid [weight=hh_weight], strata(strata) fpc(fpc1)        //set up of the survey design

*national level: mean & standard error for per capita hh exp (variable of interest)

svy: mean pce_hh_exp                             //national level – report to spreadsheet

*strata level: mean & standard error for per capita hh exp (variable of interest)

svy: mean pce_hh_exp, over(strata)               //strata level – report to spreadsheet

*-------------------------------------------------------------------------------------------------------------------------*
```

The Stata command "survey set" (*svyset*) must be specified as it allows Stata to know about the sampling strategy used in this survey. Referring to the example, it says:

svyset hhid [weight=hh_weight], strata(strata) fpc(fpc1)

- we select HHs directly from the frame (simple random survey)
- it is a stratified sample and we know the finite population correction in both strata (the total number of HHs in both strata from the sampling frame)

From the Stata program shown above, the following spreadsheet can be filled in (Table 10):

[20] *In Italics the suggested name for each variable in the dataset.*

Table 10. Retrieving information from the previous HIES

Strata	Previous HIES					
	HH population N	Sample size n	Mean or average PCE ($\bar{y}$)	Standard error of $\bar{y}$ (SE_{SRS})	Standard deviation of PCE (s)	$RSE\%$
	1	2	3	4	5	6
Strata 1	2,111	588	1,581,059	75,416	1,828,741	4.8%
Strata 2	887	437	1,063,166	53,398	1,116,261	5.0%
National	2,998	1,025	1,427,840			

Column 1: the population size N (total number of HHs within the population) from the sampling frame in the previous survey (provided by the Stata program).

Column 2: the sample size n achieved in the previous survey (provided by the Stata program).

Column 3: the sample estimate of the parameter of interest ($\bar{y}$) – mean per capita HH expenditure (mean PCE) is commonly selected as the parameter of interest in the case of a HIES design (provided by the Stata program). This variable is expressed in local currency.

Column 4: the standard error of the mean PCE estimate (SE_{SRS}), assuming simple random sampling (SRS) or systematic sampling within each stratum (provided by the Stata program).

- ⇨ All information from columns 1 to 4 were sourced from the previous survey (provided by the Stata program). The remaining 2 columns in white are calculated in the spreadsheet using the following formulas.

Column 5: the standard deviation (s) of the variable of interest (for HIES, this is the standard deviation of PCE) which is calculated in the spreadsheet using the following formula:

$$s = \sqrt{n} \times SE_{SRS}$$

Column 6: the relative standard error (RSE) of the mean PCE estimate, which is calculated in the spreadsheet using the following ratio, and expressed as a percentage:

$$RSE = \frac{SE_{SRS}}{\bar{y}}$$

We can notice that the previous HIES provided good quality estimates for the mean per capita HH expenditure as both strata show a relative sampling error equal or less than 5%, which is commonly the goal.

Step 2: update the sampling frame

If, during the period between 2 surveys, a population census or HH listing occurred, it is important to take into account the new distribution of the population in the sampling strategy. The recommendation for using population count is if it happened within the previous 2 years, meaning that if the last count was conducted more than 2 years before the survey, it is recommended to update the sampling frame (by doing a new HH count, or update of the HH listing).

The second section of the spreadsheet presents the information from the most recently updated sampling frame (total number of HHs within both strata from the most recent listing – less than 2 years ago – and the % distribution).

Table 11. Updating the sampling frame

Strata	Updated sampling frame	
	HH population N	%
	7	8
Strata 1	2,214	72.2%
Strata 2	853	27.8%
National	3,067	100.0%

In this regard and for cost savings purposes, it is important to schedule the survey to occur within 2 years of a census or HH listing. Ideally, the survey should happen in the 2 years after census in order to avoid an update of the HH listing which would add cost and time to the survey budget and timeline.

Step 3: select the sample size for the next survey

The allocation of the sample across the strata depends on the total sample size, which is closely linked to (and often constrained by) the survey budget. The purpose of this exercise is to simulate different total sample size and different strata allocation and observe what the expected accuracy will be in terms of relative sampling error.

Table 12. Allocation of the sample across the strata

Strata	Next HIES				
	Original sample allocation n_1	Effective sample allocation (fpc adjusted) n_2	Final sample allocation (adjusted for response rate) n_3	Standard error of $\bar{y}$ SE_{SRS}	$RSE\%$
	9	10	11	12	13
Strata 1	500	408	453	81,784	5.2%
Strata 2	500	315	350	49,921	4.7%
National	1,000	723	803	60,649	4.2%

Column 9: the sample allocation across the strata (original sample distribution). In this case, the survey aims to sample 1,000 HHs equally distributed across the two strata (i.e. equal allocation).

Column 10: the adjusted sample size using the finite population correction (fpc)[21], which makes a big difference in small Pacific Island Countries.

$$n_2 = \frac{n_1}{1+\frac{n_1}{N}}$$

where: n_1 is the original sample allocation in the strata (referred to as n_∞ in the Glossary);

N is the household (HH) population size in the strata;

n_2 is the effective sample allocation in the strata, adjusted for the fpc (n_N in the Glossary).

Meaning that to achieve the same precision or RSE, 408 HHs will be required in strata 1 and 315 in strata 2 (instead of 500 in both) and a total sample of 723 HHs (instead of 1,000).

Column 11: the last sample size adjustment accounts for the response rate from the previous survey:

$$n_3 = \frac{n_2}{r}$$

where: n_3 is the final sample allocation in the strata

r is the response rate from the previous survey

In this example, the response rate from the previous survey was 90% (r = 0.9) in both strata.

Column 12: the expected standard error for the mean PCE estimates from the next HIES.

$$SE_{SRS} = \frac{s}{\sqrt{n_1}}$$

where: s is the standard deviation of PCE in the strata from the previous HIES

---

[21] The finite population correction factor (fpc) is an adjustment that applies at the final stage of the sample size computation and accounts for a finite (rather than infinite) population size. This adjustment reduces the sample size required to achieve the same precision and has a bigger impact in the context of the small Pacific Island countries. The factor is: $fpc = \frac{1}{1+\frac{n_1}{N}}$

n_1 is the original sample allocation in the strata for the next HIES

Note however, that SE_{SRS} at the National level is not simply a sum of SE_{SRS} across the two strata. Rather, the National level SE_{SRS} is calculated using the following formula:

$$SE_{SRS-National} = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times SE_{SRS-Stratum\ h}{}^2}$$

In Excel this can be achieved as follows: "=SQRT(SUMPRODUCT(I4:I5,I4:I5,M4:M5,M4:M5))". This assumes:

- column 8 in the spreadsheet above = $\frac{N_h}{N}$ = column I in Excel
- column 12 in the spreadsheet above = $SE_{SRS-Stratum\ h}$= column M in Excel
- Stratum 1 information is in row 4 in Excel
- Stratum 2 information is in row 5 in Excel

Column 13: the final expected RSE of the mean PCE estimates in the next HIES, computed using the same formula as column 6, but this time using the newly calculated standard error for the next HIES from column 12.

$$RSE = \frac{SE_{SRS}}{\bar{y}}$$

The following table presents some scenarios using different sample allocations and total sample sizes.

Table 13. Different scenarios obtained after using different sample allocations and total sample sizes

	n=1000				***n=700***			
	n_1	n_3	SE_{SRS}	$RSE\%$	n_1	n_3	SE_{SRS}	$RSE\%$
Equal allocation								
Strata 1	500	453	81,784	5.2%	350	336	97,750	6.2%
Strata 2	500	350	49,921	4.7%	350	276	59,667	5.6%
National	1,000	803	60,649	4.2%	700	612	72,489	5.1%
Square root allocation								
Strata 1	617	536	73,622	4.7%	432	402	87,985	5.6%
Strata 2	383	294	57,038	5.4%	268	227	68,186	6.4%
National	1,000	830	55,463	3.9%	700	629	66,285	4.6%
Proportional allocation								
Strata 1	722	605	68,059	4.3%	505	457	81,378	5.1%
Strata 2	278	274	66,949	6.3%	195	199	79,937	7.5%
National	1,000	879	52,540	3.7%	700	656	62,811	4.4%
Final practical allocation								
Strata 1	550	490	77,978	4.9%	400	376	91,437	5.8%
Strata 2	450	416	52,621	4.9%	300	294	64,447	6.1%
National	1,000	906	58,162	4.1%	700	670	68,397	4.8%

As expected, the 700 HHs sample provide less accurate estimates of mean PCE, especially at the strata level (when the National PCE estimates are still acceptable with less than 5% RSE, except for the equal allocation).

Given the purpose of this survey is to generate mean PCE estimates at the strata level, all scenarios with a total sample size of 700 should not be considered (as the RSE is greater than 5% by strata). The value of having a breakdown at the strata level resides in the cost of implementing a total sample size of 1,000 HHs.

Looking at the different sample allocations presented in Table 12 (other allocations could have been used like the Neyman, the optimal, the Kish, the Markwardt allocation…) it shows that:

- The equal allocation reports good RSE except in strata 1, which can be improved;

- Conversely, the square root allocation reports a higher RSE in strata 2;
- The proportional allocation provides better mean PCE estimates at the National level and for strata 1 but not for strata 2 (the smaller strata by population);
- Finally, a combination of those previous allocations led to the final practical allocation that provides acceptable quality of mean PCE estimates in both strata and at the national level (RSE within each strata, and at the national level, is less than 5%).

Appendix B: Case Study 2 – Computation of sample size for a HIES with a two-stage stratified cluster sample design

Case study 2

In the case of a larger, more populated country, spread out across many islands, the simple random survey cannot be implemented for obvious reasons – the cost of implementing a survey in a large country where HHs were selected directly from the sampling frame (i.e. simple random sampling or SRS) will be too high and the survey logistics would be unmanageable.

In those cases, a multi-stage survey is more appropriate; namely, a two-stage cluster survey. This sampling strategy consists of a random selection of two sampling units:

- Within each stratum, we first randomly select a sample of primary sampling units (PSUs), which are the most disaggregated census unit level (e.g. enumeration area, census block).
- Then, within each of the selected PSUs, we randomly select a cluster or group of the secondary sampling units, which are HHs.

Hence the name, a two-stage sample design. It is important to note that the more stages added to the design, the lower the accuracy of the survey outputs will be. By using a two-stage clustered approach to sampling, the survey budget will be lower and the logistics to implement the field work will be manageable (as you survey in defined areas, such as enumeration areas), but the price to pay is the higher design effect, which may result in decreased survey precision (compared with a simple random sampling design).

The method for computing sample size is very similar compared to case study 1 in a stratified simple random survey, but it is important to note that the design effect due to using two-stage cluster sampling will impact the formulas and the process.

The overall method is the same as the stratified SRS going through the following 3 steps:

1. Retrieve information from the previous HIES: the sample size, the cluster size, the estimate of the parameter of interest (e.g. mean per capita HH expenditure or mean PCE), its standard error and the associated design effect (in the case of a multi-stage survey). In the case of a stratified sample design, this information will be required for each stratum (as well as the sample size and population size at the time the survey was selected). Then, use this information to calculate the standard deviation and intra-cluster correlation coefficient of PCE, and relative standard error of mean PCE.
2. Update of the sampling frame: it is recommended to update the second-stage sampling frame (list of HHs within PSUs) if the last count or update was conducted more than 2 years before, or in the case extreme events (e.g. natural disaster) occurred in the meantime. This update or HH listing exercise only needs to be conducted in the PSUs selected from the first-stage sampling frame (list of PSUs within each stratum).
3. Use the information from the previous HIES and the update of the sampling frame to predict the best sample size to adopt for the next HIES.

Step 1: looking at the information from the previous HIES data

Information from the previous HIES might be available directly from the HIES report; if not, it will be important to compute them using the dataset. In this case, the software Stata proposes a survey package that allows the computation of standard errors and design effects given a specific survey design.

It is important at this stage to have a clear documentation of the previous survey design, particularly:

- How many stages of selection? (in this case, 2 stages)
- What were the corresponding sampling units at each selection stage?
- What were the levels of stratification?
- What was the cluster size? (number of HHs selected per PSU or EA)

In this example, the previous HIES was based on a 2-stage stratified cluster sampling design, where the primary sampling unit was the EA and the secondary sampling unit was HHs.

The dataset must contain the following information:

- Unique Household ID – *hhid*
- The sampling weight for each household – *hh_weight*
- Strata (associated to survey design) - *strata – e.g. province, island group, island*
- PSU / EA ID – *ea_id*
- Finite population correction 1 (total number of EAs within each stratum) – *fpc1*
- Finite population correction 2 (total number of households within each selected EA) – *fpc2*
- Per capita annual household expenditure (pce) – *pc_hh_exp*

This Stata syntax shows how to compute the population size, the sample size, the cluster size, and the mean, standard error and design effect for per capita HH expenditure (PCE) from the previous HIES dataset:

```
*-------------------------------------------------------------------------------------------------------------------------*

use "$survey_dataset\previous_HIES.dta", clear

*1. Calculating population size:

total hh_weight, over(strata)          //report to spreadsheet the total number of HHs in each strata

*2. Calculating sample size:

tab strata                             //sample size - report to spreadsheet

*3. Calculating the cluster size (EA size)

preserve
        gen n=1
        bys ea_id: egen clsz_nat = total(n)
        collapse (mean) clsz_nat, by(strata)
        list                           //cluster size strata - report to spreadsheet
restore
preserve
gen n=1
        bys ea_id: egen clsz_nat = total(n)
        collapse (mean) clsz_nat
        list                           //cluster size national– report to spreadsheet
restore

*4. calculating mean, standard error, and design effect

*set up the previous survey sampling design – 3 stages of selection
svyset ea_id [pweight=hh_weight], singleunit(centered) strata(strata) fpc(fpc1)     // 1. EA selection
                || hhid, fpc(fpc2)                                                  // 2. HH selection
*national level
```

```
svy: mean pc_hh_exp                    /* mean & standard error - report to worksheet */
estat effects                          /* design effect - report to worksheet */

*strata level
svy: mean pc_hh_exp, over(strata)      /* mean & standard error - report to spreadsheet*/
estat effects, srssubpop               /* design effect - report to worksheet */

*-----------------------------------------------------------------------------------------------------------------------*
```

The Stata command survey set (svyset) must be specified as it allows Stata to know about the sampling strategy used in this survey. Referring to the example, the program describes a 2-stage sample design where:

- stage 1 consists in the selection of EAs within the strata (fpc1 represents the total number of EAs within each strata):

  svyset ea_id [pweight=hh_weight], singleunit(centered) strata(strata) fpc(fpc1)

- stage 2 consists in the selection of HHs within each selected EA (fpc2 represents the total number of HHs within each selected EA):

  || hhno, fpc(fpc2)

Each level of selection is specified using the syntax || in Stata.

From this Stata program the following spreadsheet can be filled in:

Table 14. Calculating mean, standard error and design effect

Strata	Previous HIES									
	HH population N	Sample size n	Cluster size m	Mean or average pce $\bar{y}$	Standard error of $\bar{y}$ $SE_{complex}$	Standard deviation of pce s	Design effect $deff$	$deft$	Intracluster correlation coefficient ρ	$RSE\%$
	1	2	3	4	5	6	7	8	9	10
Strata 1	5,031	223	10.9	1,866	113	1,291	1.70	1.30	0.07	6.0%
Strata 2	3,022	243	11.5	1,338	78	1,153	1.10	1.05	0.01	5.8%
Strata 3	1,391	213	14.3	1,531	73	1,073	1.00	1.00	0	4.8%
Strata 4	2,930	243	11.5	1,136	212	1,020	10.50	3.24	0.90	18.7%
Strata 5	1,293	216	14.1	2,115	343	1,964	6.60	2.57	0.43	16.2%
National	13,667	1,138	12.4	1,582						

Column 1: the population size N (total number of HHs within the population) from the sampling frame in the previous survey (provided by the Stata program)

Column 2: the sample size n achieved in the previous survey (provided by the Stata program)

Column 3: the cluster size m achieved in the previous survey EA (provided by the Stata program). It represents the average number of HHs interviewed per EA.

Column 4: the sample estimate of the parameter of interest $(\bar{y})$ – mean per capita annual HH expenditure (mean PCE) is commonly selected as the parameter of interest in the case of a HIES design (provided by the Stata program). This variable is expressed in local currency.

Column 5: the standard error of the mean PCE estimate $(SE_{complex})$ given the previous HIES used a two-stage cluster sample design (provided by the Stata program).

Column 6: standard deviation (s) of the variable of interest (for HIES, this is the standard deviation of pce) which is calculated in the spreadsheet using the following formula:

$$s = \sqrt{n} * \frac{SE_{complex}}{deft}$$

Column 7: the design effect $deff$ (provided by the Stata program).

Column 8: $deft$ the square root of the design effect.

$$deft = \sqrt{deff}$$

Column 9: the intra-cluster correlation coefficient (ρ). It measures the relatedness or homogeneity of clusters with respect to the variable of interest, PCE.

$$\rho = \frac{deff - 1}{m - 1}$$

Column 10: the relative standard error RSE of the sample estimate (mean pce), which is calculated in the spreadsheet using the following ratio, and expressed as a percentage:

$$RSE = \frac{SE_{complex}}{\bar{y}}$$

⇨ Information from columns 1 to 5 and 7 were sourced from the previous survey (provided by the Stata program). All other columns in white are calculated in the spreadsheet using the formulas above.

We can see that the previous HIES design provided:

- a good quality estimate for the mean per capita HH expenditure in strata 3, as it shows a relative sampling error equal or less to 5% which is commonly the goal;
- estimates of medium quality in strata 1 and 2, both with relative sampling error between 5% and 10%, which ideally could be improved on in the next HIES sample design;
- poor quality estimates (with a high design effect and RSE greater than 15%) in strata 4 and 5, meaning that some improvements should be made to the sample design for those strata in the next HIES design.

Step 2: update the sampling frame

In a similar way to the stratified simple random sample design in Case study 1, the multi-stage sample design requires the use of an updated sampling frame (within the last 2 years). If this updated sampling frame is not available, there is a compromise that consists in updating only the EAs that have been selected in the first stage of the survey (which reduces the work). In many cases, this update of the HH listing happens during the fieldwork: just before the field team starts the interview within an EA, they proceed to update the HH listing within this EA. This method requires the team to proceed to the random selection of HHs within EAs as well.

In our example, an update of the HH listing occurred during the last Census, which was conducted after the previous HIES and within the last 2 years. The second section of the spreadsheet presents the information from this most recent sampling frame update (total number of HHs within each stratum and the % distribution).

Table 15. Updating the sampling frame

Strata	Updated sampling frame	
	HH population N	%
	11	12
Strata 1	8,412	47.2%
Strata 2	3,459	19.4%
Strata 3	1,520	8.5%
Strata 4	3,072	17.2%
Strata 5	1,370	7.7%
National	17,833	100.0%

Again, in this regard and for cost savings purposes, it is important to schedule the survey to occur within 2 years of a census or HH listing. Ideally, the survey should happen in the two years after census in order to avoid an update of the HH listing which added cost and time to the survey budget and timeline.

Step 3: select the sample size for the next survey

In a multi-stage survey, the sample allocation across strata depends on the total sample size and the cluster size (the number of HHs to select in each primary sampling unit or PSU). The cluster size in a HIES usually varies between 10 and 15 HHs to select in each PSU, and it is possible to have different cluster size in each stratum.

This scenario proposes an equal allocation of a total sample size of 1,500 HHs across the 5 strata, and a cluster size of 12 HHs in every stratum. The purpose of this exercise is to observe what the expected accuracy of this HIES sample design will be in terms of relative standard error.

Table 16. Sample size and sample allocation in a multi stage design

Strata	Next HIES								
	Original sample allocation n_1	Effective sample allocation (fpc adjusted) n_2	Final sample allocation (adjusted for response rate) n_3	Cluster size m	Standard error of $\bar{y}$ (assuming SRS in each stratum) SE_{SRS}	$deff$	$deft$	Standard error of $\bar{y}$ (accounting for multi-stage design) $SE_{complex}$	$RSE\%$
	13	14	15	16	17	18	19	20	21
Strata 1	300	290	322	12	74.55	1.8	1.3	99.4	5.3%
Strata 2	300	276	307	12	66.59	1.1	1.1	70.0	5.2%
Strata 3	300	251	279	12	61.92	1.0	1.0	61.9	4.0%
Strata 4	300	273	303	12	58.92	11.0	3.3	195.0	17.2%
Strata 5	300	246	273	12	113.38	5.7	2.4	270.7	12.8%
National	1,500	1,336	1,484					63.0	4.0%

Column 13: the sample allocation across the strata (original sample distribution). In this case, the survey aims to sample 1,500 HHs equally distributed across all strata (i.e. equal allocation).

Column 14: the adjusted sample size using the finite population (fpc) [22], which makes a big difference in small island countries. This adjusted sample size will provide the same precision or RSE as the original one.

$$n_2 = \frac{n_1}{1+\frac{n_1}{N}}$$

where: n_1 is the original sample allocation in the strata (referred to as n_∞ in the Glossary)

N is the household (HH) population size in the strata

n_2 is the effective sample allocation in the strata, adjusted for the fpc (n_N in the Glossary)

Column 15: the last sample size adjustment accounts for the response rate from the previous survey:

$$n_3 = \frac{n_2}{r}$$

where: n_3 is the final sample allocation in the strata

r is the response rate from the previous survey

[22] *The finite population correction factor (fpc) is an adjustment that applies at the final stage of the sample size computation and accounts for a finite (rather than infinite) population size. This adjustment reduces the sample size required to achieve the same precision and has a bigger impact in the context of the small island countries. The factor is:* $fpc = \frac{1}{1+\frac{n_1}{N}}$

In this example, the response rate from the previous survey was 90% (r = 0.9) in all strata.

Column 16: the cluster size m is the number of HHs to select in each of the selected PSUs or EAs.

Column 17: the expected standard error for the mean PCE estimates from the next HIES, assuming the next survey design is a stratified simple random sample (SRS):

$$SE_{SRS} = \frac{s}{\sqrt{n_1}}$$

where: s is the standard deviation of pce from the previous survey for the strata

n_1 is the original sample allocation in the strata

Column 18: the expected design effect or $deff$ for the next HIES based on the multi-stage design.

$$deff = 1 + \rho\,(m - 1)$$

where: ρ is the intracluster correlation computed from the previous HIES, as calculated in column 9

m is the cluster size proposed in the next HIES sample design

Column 19: $deft$ is the square root of the design effect, calculated using the same formula as column 8, but this time using the newly calculated $deff$ from column 18.

$$deft = \sqrt{deff}$$

Column 20: the expected standard error for the mean pce estimates in the next HIES, assuming a multi-stage design.

$$SE_{complex} = deft \times SE_{SRS}$$

Note however, that $SE_{complex}$ at the National level is not simply a sum of $SE_{complex}$ across the five strata. Rather, the National level $SE_{complex}$ is calculated using the following formula:

$$SE_{complex-National} = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times SE_{complex-Stratum\ h}{}^2}$$

In Excel this can be achieved as follows: "=SQRT(SUMPRODUCT(M4:M8,M4:M8,U4:U8,U4:U8))". This assumes:

- column 12 in the spreadsheet above = $\frac{N_h}{N}$ = column M in Excel
- column 20 in the spreadsheet above = $SE_{complex-Stratum\ h}$= column U in Excel
- Stratum 1 information is in row 4 in Excel
- Stratum 5 information is in row 8 in Excel

Column 21: the final expected RSE of the mean PCE estimates in the next HIES, assuming a multi-stage design. This is calculated using the same formula as column 10, but this time using the newly calculated standard error for the next HIES from column 20.

$$RSE = \frac{SE_{complex}}{\bar{y}}$$

This completes our first sample design scenario for a HIES with a two-stage stratified cluster design – a scenario with equal allocation across the strata of a total sample size of n = 1,500 HHs, with a cluster size of m = 12 HHs in all strata.

Now, as in Case study 1, you can try out different scenarios or sample designs – with different stratifications, allocation methods, cluster sizes, and total sample size – and look at how each different design impacts on the

precision (RSE) of your mean pce estimates. The aim is to find the best scenario or sample design, that meets your reporting needs and produces estimates with an acceptable level of precision.

Once you've investigated and considered various sample design options and selected the final design to be used in the next HIES, the next step is to select the sample. Case Study 3 illustrates how to select the first stage sample in a two-stage HIES sample design – that is, how to select a stratified probability proportional to size (PPS) sample of primary sampling units (PSUs) such as enumeration areas (EAs).

Appendix C: Case Study 3 – Selecting a probability proportional to size (PPS) sample in the 1st stage of a 2-stage HIES sample design

Case study 3

This case study illustrates how to select a probability proportional to size (PPS) sample, as part of a two-stage, stratified cluster HIES sample design (the focus of Case Study 2). This type of complex HIES sample design is most relevant in larger Pacific countries.

The case-study focuses specifically on the 1st stage of sample selection in this type of design – PPS selection of primary sampling units – such as enumeration areas (EAs) or census blocks - within strata.

The overall method for stratified PPS sample selection involves 3 steps:

1. Retrieve information from the previous Census or (if the Census was more than 2 years ago) a more recent HH listing exercise, and from the next HIES sample design, for each stratum. Information required from the previous Census or recent HH listing:
    - a list of all PSUs in the stratum (ideally sorted in a logical, geographical order), along with their EA identifier, and
    - the size (number of households) of each of these PSUs, Z_i

    and information required from the next HIES sample design:
    - the population (number of households), and
    - the number of PSUs to be selected in each stratum.

    This information is used to calculate the PPS probability of selection for each PSU in step 2.
2. Calculate the PPS probability of selection π_i for each PSU in the stratum, along with the cumulative PPS probabilities of selection for the stratum, calculated from a random start number between 0 and 1. This information is used to select the required number of PSUs into the sample in step 3.
3. Use the cumulative PPS probabilities of selection in the stratum to select the required number of PSUs into the sample. Each time the integer part of the cumulative PPS probability changes, the PSU is selected into the sample.

The remaining sections show this PPS sample selection process for a single stratum only – stratum 1 - which contains 26 enumeration areas (EAs) of varying sizes[23], with 6 EAs to be selected into the sample. However, the same PPS selection process can easily be replicated within each of the remaining strata to complete the first-stage selection of EAs for the next HIES.

Step 1: Retrieve information from the previous Census or HH listing and the next HIES sample design, for each stratum.

Information from the previous Census or HH listing should be readily available, as should information from the next HIES sample design provided it's been adequately documented.

---

[23] This matches the PPS sampling scenario illustrated in section 4.5 Household Survey Example: Probability Proportional to Size Sampling within Chapter 2 of the Sampling Guidelines for the Pacific.

For this scenario, assume the HIES sample design information for stratum 1 is as follows:

Table 17. Step 1: Retrieving information from next HIES

Strata	Next HIES sample design			
	HH population N	Effective sample allocation (fpc adjusted) n_2	Cluster size m	**Number of EAs to select**
Strata 1	**1,365**	72	12	**6**
Strata 2	…	…	…	…
…				

This shows that for stratum 1, the number of HHs in the population is 1,365, the sample size required for the next HIES is 72 HHs, and the required cluster size (number of HHs to select within each selected EA) is 12 HHs. The number of EAs that need to be selected in stratum 1 is therefore 72/12 = 6, as calculated in the right-hand column.

The HIES sample design information specifically required for PPS sample selection of EAs in stratum 1 is highlighted in blue bold. With this information, plus EA information (EA identifier and size) from the previous Census or HH listing, the following spreadsheet can be filled in:

Table 18. Next HIES sample design

Previous Census		Next HIES sample design	
EA id	EA size (HH) Z_i	Population size (HH) Z	# EAs to select n
1	2	3	4
100101	43	1,365	6
100102	81	1,365	6
100103	52	1,365	6
100104	61	1,365	6
100105	44	1,365	6
100106	38	1,365	6
100201	72	1,365	6
100202	49	1,365	6
100203	47	1,365	6
100204	33	1,365	6
100205	61	1,365	6
100206	63	1,365	6
100207	51	1,365	6
100301	48	1,365	6
100302	38	1,365	6
100303	71	1,365	6
100304	55	1,365	6
100305	51	1,365	6
100306	41	1,365	6
100307	49	1,365	6
100308	73	1,365	6
100309	48	1,365	6
100310	39	1,365	6
100311	32	1,365	6
100312	67	1,365	6
100313	58	1,365	6
Total	***1,365***		

Column 1: the list of all EAs i in stratum 1, with their EA identifier, sorted in a logical geographic order – by village, street etc. (from the previous Census).

Column 2: the size of each of these EAs, Z_i, measured in terms of the number of HHs in each EA (from the previous Census).

Column 3: the total number of households Z in stratum 1, where $= \sum Z_i$, i.e. where Z is the sum the EA sizes Z_i, across all EAs in stratum 1 (from the next HIES sample design).

Column 4: the number of EAs to select in stratum 1 (from the next HIES sample design, though originally sourced from the previous Census).

- ⇨ All information in columns 1 to 4 above is shaded in blue to show that it is retrieved from existing sources – either the previous Census or the next HIES sample design.

Step 2: Calculate the PPS probability of selection π_i for each EA in the stratum, along with the cumulative PPS probability of selection for the stratum, calculated from a random start number between 0 and 1.

The information retrieved in step 1 is used to calculate the PPS probability of selection π_i for each EA in the stratum step 2, as follows:

Table 19. Calculating the PPS probability of selection

Previous Census		Next HIES sample design		Next HIES sample selection	
EA id	EA size (HH) Z_i	Population size (HH) Z	# EAs to select n	PPS probability of selecting EA_i π_i	Cumulative PPS probability of selection (from random start)
1	2	3	4	5	6
				Random start	0.04257003
100101	43	1,365	6	0.1890	0.2316
100102	81	1,365	6	0.3560	0.5876
100103	52	1,365	6	0.2286	0.8162
100104	61	1,365	6	0.2681	1.0843
100105	44	1,365	6	0.1934	1.2777
100106	38	1,365	6	0.1670	1.4448
100201	72	1,365	6	0.3165	1.7613
100202	49	1,365	6	0.2154	1.9766
100203	47	1,365	6	0.2066	2.1832
100204	33	1,365	6	0.1451	2.3283
100205	61	1,365	6	0.2681	2.5964
100206	63	1,365	6	0.2769	2.8733
100207	51	1,365	6	0.2242	3.0975
100301	48	1,365	6	0.2110	3.3085
100302	38	1,365	6	0.1670	3.4755
100303	71	1,365	6	0.3121	3.7876
100304	55	1,365	6	0.2418	4.0294
100305	51	1,365	6	0.2242	4.2536
100306	41	1,365	6	0.1802	4.4338
100307	49	1,365	6	0.2154	4.6492
100308	73	1,365	6	0.3209	4.9700
100309	48	1,365	6	0.2110	5.1810
100310	39	1,365	6	0.1714	5.3525
100311	32	1,365	6	0.1407	5.4931
100312	67	1,365	6	0.2945	5.7876
100313	58	1,365	6	0.2549	6.0426
Total	***1,365***			***6***	

⇨ The information added in columns 5 and 6 of the spreadsheet is in white (rather than blue), to show that it is calculated in the spreadsheet using the following formulas.

Column 5: the PPS probability of selection π_i for each EA i in stratum 1:

$$\pi_i = n \times \frac{Z_i}{Z}$$

Column 6: the cumulative PPS probability of selection for each EA in stratum 1, calculated from a random start number between 0 and 1.

First, generate a random number between 0 and 1. In Excel this can be achieved as follows "=rand()". To ensure the random number doesn't subsequently change each time the spreadsheet is updated, copy the random number generated and then Paste -> Paste Values to fix at as a constant. In this example, the resulting random start is 0.04257003

Then, using the random number as the starting value, calculate the cumulative PPS probability of selection as follows:

For EA **1** (100101), the cumulative PPS probability of selection = ***Random start + π_1*** = 0.4257003 + 0.1890 = 0.2316

For EA **2** (100102), the cumulative probability of selection = ***cumulative PPS probability of selection for EA 1 + π_2*** = 0.2316 + 0.3560 = 0.5876

For EA **3** (100103), the cumulative probability of selection ***= cumulative PPS probability of selection for EA 2 + π_3*** = 0.5876 + 0.2286 = 0.8162, and so on.

Step 3: Use the cumulative PPS probabilities of selection to select the required number of EAs into the sample

The cumulative PPS probabilities of selection calculated in step 2 are used to select the required number of EAs n into the sample in step 3, as described in the table below.

⇨ Again, the information added in column 7 of the spreadsheet is in white (rather than blue), to show that it is calculated in the spreadsheet using the formula below.

Column 7: an indicator of whether each of the EAs in stratum 1 is selected into the sample or not:

0 = the EA is not selected into the sample

1 = the EA is selected into the sample

Effectively, when the integer part of the cumulative PPS probability increases from one EA to the next in the list, that next EA is selected into the sample. In the example above:

- for each of first three EAs (EA100101, EA100102, EA100103) the integer part of the cumulative PPS probability (column 6) is 0
- for the next EA in the list (EA100104), the integer part of the cumulative PPS probability is 1, an increase from the previous EA - so EA100104 is selected into the sample.

In Excel, this selection process can be automated in column 7 for the first EA (EA100101) as follows "= INT(F5) - INT(F4)", then filling the formula down for the remaining EAs (this assumes column 6 in Table 19 is column F in Excel, and the random start is in row 4).

Table 20. Selecting the required number of EAs into the sample

Previous Census		Next HIES sample design		Next HIES sample selection		
EA id	EA size (HH) Z_i	Population size (HH) Z	# EAs to select n	PPS probability of selecting EA_i π_i	Cumulative PPS probability of selection (from random start)	Selected EAs
1	2	3	4	5	6	7
				Random start	0.04257003	
100101	43	1,365	6	0.1890	0.2316	0
100102	81	1,365	6	0.3560	0.5876	0
100103	52	1,365	6	0.2286	0.8162	0
100104	**61**	**1,365**	**6**	**0.2681**	**1.0843**	**1**
100105	44	1,365	6	0.1934	1.2777	0
100106	38	1,365	6	0.1670	1.4448	0
100201	72	1,365	6	0.3165	1.7613	0
100202	49	1,365	6	0.2154	1.9766	0
100203	**47**	**1,365**	**6**	**0.2066**	**2.1832**	**1**
100204	33	1,365	6	0.1451	2.3283	0
100205	61	1,365	6	0.2681	2.5964	0
100206	63	1,365	6	0.2769	2.8733	0
100207	**51**	**1,365**	**6**	**0.2242**	**3.0975**	**1**
100301	48	1,365	6	0.2110	3.3085	0
100302	38	1,365	6	0.1670	3.4755	0
100303	71	1,365	6	0.3121	3.7876	0
100304	**55**	**1,365**	**6**	**0.2418**	**4.0294**	**1**
100305	51	1,365	6	0.2242	4.2536	0
100306	41	1,365	6	0.1802	4.4338	0
100307	49	1,365	6	0.2154	4.6492	0
100308	73	1,365	6	0.3209	4.9700	0
100309	**48**	**1,365**	**6**	**0.2110**	**5.1810**	**1**
100310	39	1,365	6	0.1714	5.3525	0
100311	32	1,365	6	0.1407	5.4931	0
100312	67	1,365	6	0.2945	5.7876	0
100313	**58**	**1,365**	**6**	**0.2549**	**6.0426**	**1**
Total	***1,365***			***6***		***6***

Then all EAs with a positive integer value (1, 2, …) in column 7 are selected into the sample, and those with a 0 in column 7 are not. The 6 EAs selected within stratum 1 in this example have then been (manually) highlighted in red bold.

This concludes the first-stage PPS selection of EAs within stratum 1. This selection process can be repeated within the same spreadsheet for the remaining HIES strata, using the same random start number for each of these. Then in the second stage, a sample of HHs is selected within each of the EAs selected at the first stage.

In most, if not all the selected EAs, a single cluster of HHs will be selected at the second stage. However, if the value in column 7 (Selected EAs) is greater than 1 – say, 2 or 3 – for a selected EA, this indicates the EA is particularly large and more than 1 cluster of HHs – 2 or 3 respectively – needs to be selected within it.

In the example above, the cluster size in stratum 1 is m = 12. So, if an EA in stratum 1 has a Selected EAs value of 2, this means 2 clusters, or 2 x 12 = 24 HHs need to be selected within this EA.

There are a couple of approaches for achieving this:

1. randomly select 24 HHs from the EA
2. segment or divide the EA into 2 parts of equal size (in terms of number of HHs) and select a one cluster of 12 HHs from each of the two new EAs.

Appendix D: Glossary of key sample design terminology, notation, definitions and formulae

Term	Notation/label	Definition	Formula
Cluster size	m	The number of units or elements selected into the sample in each cluster (e.g. in each Enumeration Area). For example: ▪ the number of HHs selected into the sample within each enumeration area (EA) ▪ the average number of people in the working age population selected into the sample within an enumeration area (EA)	
Confidence interval	CI (with upper and lower limits)	A confidence interval a useful way to illustrate the level of uncertainty in a sample estimate, providing an upper and lower limit on the estimate. The confidence interval is the sample estimate, plus or minus the margin of error of the sample estimate, ME In other words, the ½-width of the confidence interval is the margin of error ME. As for the margin of error, confidence intervals are usually given at the 95% confidence level.	For a sample mean $\bar{y}$, the confidence interval is: $CI = \bar{y} \pm ME = \bar{y} \pm t_\alpha \times SE$ So, at the 95% level of confidence, t_α = 1.96 and: $CI = \bar{y} \pm ME = \bar{y} \pm 1.96 \times SE$ *or* $CI = (\bar{y} - 1.96 \times SE, \bar{y} + 1.96 \times SE)$ For a sample mean $\bar{y}$, the confidence interval is: $CI = p \pm ME = p \pm t_\alpha \times SE$ So, at the 95% level of confidence, t_α = 1.96 and: $CI = p \pm ME = p \pm 1.96 \times SE$ *or* $CI = (p - 1.96 \times SE, p + 1.96 \times SE)$
Deft	$deft$	Another form or variant of the design effect that is sometimes used in sampling, it is the square root of the design effect.	$deft = \sqrt{deff} = \frac{SE_{complex}}{SE_{SRS}}$ Rearranging gives: $SE_{complex} = deft * SE_{SRS}$ *or* $SE_{SRS} = \frac{SE_{complex}}{deft}$ which are both useful formulae for sample designers. Substituting the formula for SE_{SRS} from above into this last formula gives:

			$s = \sqrt{n} \times \frac{SE_{complex}}{deft}$ another useful formula for sample designers working on complex sample designs (e.g. stratified two-stage cluster sampling).
Design effect	$deff$	The design effect for the sample estimate. The design effect is a measure of precision gained or lost on the sample estimate by use of a more complex sample design - *e.g. stratified sampling; two-stage cluster sampling* - instead of a simple random sample (SRS), for a fixed sample size n. In other words, it is the "bonus" or "penalty" that we get from stratification, or the "penalty" we pay for clustering.	$deff = \frac{SE^2_{complex}}{SE^2_{SRS}}$ Alternatively, $deff = 1 + \rho\,(m-1)$ where: ρ = intracluster correlation coefficient (see below for definition) m = cluster size (possibly an average)
Intra-cluster correlation coefficient[24]	ICC or ρ	The intra-cluster correlation coefficient or ρ (rho) for the variable of interest. It is a measure of the relatedness or homogeneity of clusters with respect to the variable of interest, that accounts for both: ▪ the variance of the variable of interest <u>within</u> clusters, and ▪ the variance of the variable of interest <u>between</u> clusters.	Rearranging the second formulae for the $deff$ gives: $\rho = \frac{deff - 1}{m - 1}$ another useful formula for sample designers
Margin of error	ME	The margin of error for the sample estimate, usually expressed at the 95% confidence level. Like the standard error, it is a measure of how precise the estimate from a sample is.	$ME = t_\alpha \times SE$ where t_α is the critical value of the Student's t-distribution for a confidence level of α. At a confidence level of 95%, α = 0.95 and t_α = 1.96, so that: $ME = 1.96 \times SE$ (There are other t-values for higher or lower levels of confidence, but α = 0.95 is the level of confidence most commonly used).

[24] The intracluster correlation coefficient may also be referred to as the intraclass correlation coefficient.

Population size	N	The total number of units or elements in the population of interest (target population) for a survey. For example: ▪ the total number of HHs ▪ the total number of people in the working age population	
Relative margin of error	RME	The relative margin of error for the sample estimate. It is the margin of error for the sample estimate, relative to (i.e. as a ratio of) the sample estimate itself and is usually expressed as a percentage. Like the margin of error, it is a measure of how precise the estimate from a sample is.	For a sample mean $\bar{y}$: $RME = \frac{ME}{\bar{y}}$ For a sample proportion p: $RME = \frac{ME}{p}$
Relative standard error	RSE	The relative standard error for the sample estimate. It is the standard error of the sample estimate, relative to (i.e. as a ratio of) the sample estimate itself and is usually expressed as a percentage. Like the standard error, it is a measure of how precise the estimate from a sample is.	For a sample mean $\bar{y}$: $RSE = \frac{SE}{\bar{y}}$ For a sample proportion p: $RSE = \frac{SE}{p}$
Sample estimate	$\bar{y}$ for a mean or p for a proportion	The estimate of a population parameter – such as the population mean $\bar{Y}$ or population proportion P – as calculated from a sample. For example: ▪ for HIES, the population parameter or indicator of interest to be estimated from the sample is mean per capita expenditure (PCE). ▪ for LFS, the parameter or indicator of interest to be estimated is the LF participation rate, or proportion of the working age population participating in the labour force. ▪ for MICS, the parameter or indicator of interest to be estimated is the prevalence (or proportion) of underweight children under 5 years of age.	For an SRS of size n , the sample estimate of the population mean $\bar{Y}$ is: $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ or the sample estimate of the population proportion P is: $p = \frac{1}{n}\sum_{i=1}^{n} y_i$ where: $y_i = 1$ if sample unit i has the characteristic, and $y_i = 0$ otherwise
Sample size	n	The number of units or elements selected into the sample, from the target population.	

		A subscript may be used to differentiate between: n_∞ – sample size assuming an infinite population n_N – sample size for a finite population, i.e. adjusted for the finite population correction (*fpc*) factor The sample size may also be adjusted (boosted) to account for the expected level of non-response to the survey.	Note that: $n_N = n_\infty \times fpc = \frac{n_\infty}{1 + \frac{n_\infty}{N}}$ where the finite population correction factor is: $fpc = \frac{1}{1 + \frac{n_\infty}{N}}$
(Estimated) Sampling variance	$Var(\bar{y})$ or $Var(p)$	▪ The variance of the sampling distribution of an estimator – such as a mean $\bar{y}$ or a proportion p – as estimated from a sample.	For a mean, assuming an SRS of size n: $Var(\bar{y}) = \frac{s^2}{n}$ For a proportion, assuming an SRS of size n: $Var(p) = \frac{p(1-p)}{n}$
(Estimated) Standard deviation	s	The standard deviation of the observed sample values for the variable of interest y_i. The standard deviation is a descriptive statistic. It is the degree to which individuals in the population differ from the mean of the population. It is the square root of s^2, the variance estimated from a sample.	$s = \sqrt{s^2}$
Standard error[25]	SE	The standard error of the sample estimate. The standard error (SE) is a measure of how precise the estimate from a sample is. It measures how far the sample estimate (e.g. mean or proportion) deviates from the actual population value. It is the square root of the variance of the sampling distribution of the estimator ($Var(\bar{y})$ or $Var(p)$)	For a mean, assuming an SRS of size n: $SE_{SRS} = \sqrt{Var(\bar{y})} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$ Rearranging gives: $s = \sqrt{n} \times SE_{SRS}$

[25] Note that in a stratified sample design, the standard error at the National level ($SE_{National}$) is <u>not</u> simply a sum of the standard errors across the strata ($SE_{Stratum\ h}$). Rather, the National level standard error is calculated using the following formula:

$$SE_{National} = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times SE_{Stratum\ h}{}^2}$$

		Note that a subscript may be used to indicate the sampling method underlying the standard error calculation – e.g. ▪ SE_{SRS} is the standard error estimated from a simple random sample (SRS) ▪ $SE_{complex}$ is the standard error estimated from a complex sample design, such as a two-stage cluster sample.	which is a useful formula for sample designers. For a proportion, assuming an SRS of size n: $$SE_{SRS} = \sqrt{Var(p)} = \sqrt{\frac{p(1-p)}{n}}$$
Stratum identifier	$h = 1, 2, \ldots, H$	Identifier for each stratum within a stratified sample design, where: h = the stratum identifier or number H = the total number of strata Examples of strata include: ▪ regions/provinces ▪ islands or Island groups ▪ urban vs rural ▪ type of economic activity ▪ number of employees	
(Estimated) Variance	s^2	The variance of the observed sample values for the variable of interest y_i. The variance is a descriptive statistic. It is the degree to which individuals in the population differ from the mean of the population.	Assuming an SRS of size n: $$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$ For a proportion where: $yi = 1$ *if sample unit* i *has the characteristic,* $yi = 0$ *otherwise* this formula simplifies to: $$s^2 = p(1-p)$$

Appendix E: Sampling Methods for Core and Additional Modules

One of the main methodological issues related to HH survey implementation in the Pacific is the excessive amount of time needed to complete the interview and the relative excessive burden for respondents. This issue was mentioned during the last two meetings of the Pacific Statistics Method Board [26] and is seen by the Pacific statistics stakeholder as one of the main methodological priorities to be addressed in the near future.

The average interview time is about 3-4 hours, depending on the country and the survey. In many cases, however, the interview time is several hours longer than that; the interviewers are often invited to eat at the interviewee's place, or have to remain until late in the evening or to come back in a following day to complete the interview. This is particularly true when the selected HH is big, with a number of members reaching 10-12 or even more: in such cases, which are not rare in the Pacific, the combination of a high number of questions and a high number of HH members determines a total interview time of up to 15 hours (P. Ellis 2023, forthcoming). This situation brings several negative consequences, including increased enumeration cost, more difficult planning and organisation of the field operations and an excessive burden on respondents – with consequent low quality of answers and reduced willingness to participate in following surveys.

While the most obvious solution to address this issue would seem to be reducing the number of survey modules and rationalise the number and structure of questions within them, the role of sampling in addressing it is not to be undervalued. What follows is a brief description of the contribution sampling can make to the goal of reducing the interview time while at the same time improving the quality of the survey results.

Most of the HH surveys in the Pacific produce results that are significant not only at the National level, but also at different types of sub-national geographical aggregations, such as regions, urban and rural areas, etc., depending on the country.[27] Sampling is performed in such a way that results are statistically significant at the level of each strata, which implies that all the survey modules are to be administered to all the selected HHs and HH members residing in the selected Enumeration Areas of all the strata.

The survey questionnaire, however, is composed of different modules: some of them, the "core modules", are to be included in all the survey questionnaires received by all respondents while other, "additional" modules, could be provided to only a sub-sample of respondents. The sub-sample is to be extracted in such a way that results are significant only at the national level, and not at the level of each selected strata. The clear consequence of this is the reduction of the total number of HHs that will receive the "additional" modules, and hence the reduction of the total time and human resources (and budget) needed for implementing the module, while at the same time keeping exactly the same number of survey modules and questions. This is how the climate change and natural disaster module is currently being implemented in Kiribati, with a sub-sample of HHs receiving it and a sampling strategy allowing for results significant at the National level.[28]

Because the additional modules will not be answered by all HHs – but only by those of the sub-sample – a second advantage is the possibility to reduce the number of additional modules each HH in the sub-sample has to respond. In fact, not all the additional modules would be included in the survey questionnaire received by the sub-sample of HHs: some HHs would receive the additional modules "A, B and C", while others would receive the modules "D, E and F". In the end, the number of selected HHs would be such that, for each additional module, the results will be statistically significant at the National level.

---

[26] See for example: https://sdd.spc.int/events/2023/05/11th-Pacific-statistics-methods-board-meeting-psmb

[27] For example, in Samoa there are 4 strata: Apia Urban Area, North-West Upolu, Rest of Upolu and Savai'i.

[28] Given the testing nature of this module, it was agreed to administer it only on half of the total sample; this will not allow for disaggregation of data by island group, but at best a urban/rural breakdown (or National only, depending on the response rate and the data quality). The sub-sample receiving the module has been selected in advance, based on the household id: only selected household with an odd id will be responding to the climate change section. This is made available by the CAPI survey methodology: the Survey Solution questionnaire triggers the climate change section based on the household id. The sampling weights for this section will have to be recomputed.

An interesting question would now be: how to determine which module is "core" and which one is "additional"? To answer this question, not only statistical considerations, but also political ones should be taken into account. The answer, therefore, should be provided after consultations with GS and other Pacific statistics stakeholders, and might not be the same across PICTs.

One possible method to distinguish between the "core" and the "additional" module is based on the use of the data derived from those modules to the aims of National Accounts. According to this method, all modules contributing to income and expenditure aggregates and poverty monitoring should be considered as "core", with results disaggregated at sub-national level, while other modules – such as climate change, disability, FIES, etc. – could be considered "additional" and provide data that are significant at the National level only. Should a country consider one of the "additional" topic a National priority (say: Climate Change in PICT "X"), then this module will be considered as core, and provided to all the HHs of the sample.

Another good reason for reducing the number of HHs receiving a given survey module is the contemporary presence of a similar module with slightly different questions collecting data on the same characteristics. Asking similar questions twice, in fact, would be negatively perceived by respondents who have already spent a big amount of their available time answering similar questions. This might occur, for example, when a new module is introduced in the survey questionnaire, say, for new monitoring purposes, but the previous version of the module is to be kept to avoid a break in the data series and to compare the results obtained with the two versions of the module. Including both modules in the survey questionnaire would cause a negative reaction in respondents; at the same time, it might not be not possible to simply drop the old module, especially if those data were requested by the country, nor to drop the new module, to allow the monitoring of internationally recognised indicators. The final solution would therefore be to provide the two survey modules to two different sub-sets of HHs, so that those responding to the module A would not receive the module B, and vice versa. The new probability of selecting HHs of the two sub-samples will then have to be reflected in the calculation of sampling weights.

A similar way to address the issue of excessive length interview would be the adoption of "rotating" modules. This approach implies that the "additional" modules – identified according to the methods described above – would be included in the HIES questionnaire every other survey.[29] In this way, the additional modules could be split between two consecutive surveys, and each HH would only receive half of the additional modules. A good example of implementation of rotating modules can be found in the FAO and World Bank Agricultural Integrated Survey (AGRIS) Programme,[30] with survey modules distinguished in "core" and "rotating". While the core modules are administered yearly to the selected HHs, the rotating modules are only provided to them at 3 to 5-year intervals between the 10-year agricultural census cycle. Survey schedule in the Pacific region, however – except a few cases – is often irregular, so it will be difficult to commit to a supplementary section in the next HIES due to the uncertainty of the next edition. Technical partners, experts and donors, in fact, usually prefer having more data points without having to wait too long.

The interview time can be further reduced – especially in big HHs – by extracting a sub-sample of HH members who have to respond to the survey questions asked at the individual level and that are not part of modules considered as "Core". Examples of such modules include those collecting information on Social Protection, Migrant Workers and Financial inclusion. In those cases, 1 HH member randomly selected in the HH can be interviewed, with an obvious reduction in the interview time, especially for big HHs. The drawback of this method is that, if the selected person is not at home at the time of the interview, then a new appointment should be made with the HH to interview that person. The overall cost/benefit balance in terms of time needed to complete the interview should therefore be evaluated after testing this method in the field.

Finally, another way to reduce not just the interview time but rather the respondents' fatigue is through coordination among the international development partners operating in the Pacific in the field of survey data collection (e.g.: SPC, ILO, UNICEF, etc.). It is not uncommon, in fact, that HHs selected for a given survey (say: MICS,

---

[29] *For example, every 10 – rather than 5 – years, under the hypothesis of a 5-year HIES implementation cycle in a given PICT.*

[30] *Available at: https://www.fao.org/in-action/agrisurvey/en/*

run by UNICEF) is then surveyed again in a subsequent survey undertaken by another organisation (say: HIES, run by SPC) after a short period of time. This causes of course a low willingness to be interviewed by the HH members and, as a consequence, lower response rates, lower data quality and lower acceptability of the survey by the local populations. Sharing information among agencies on the sampling strategies adopted would therefore be of great help to reduce the burden on respondents by avoiding their selection in multiple consecutive surveys. If this solution might appear conceptually easy, it is however not easy to implement, as it entails coordination among members of different organisations with different goals and programs, each of which has their own "relative advantage" in a given area compared to the other organisations. Finding a good compromise will therefore be more of a political, rather than technical, issue.