INVITED FEATURE: STATISTICS PRIMER

Health related surveys: sampling methods

Ari Samaranayaka, Robin M Turner, Claire Cameron

Have you ever thought of running a survey? You could run an online survey! Word on the street is that they're easy to do. We wish that were true! Surveys can be very tricky and there can be a lot of hidden hooks for the unwary. These hooks could be hidden in how you ask questions, how you recruit people, how you manage and store the data, and how you deal with a lack of participation. Then there are the details around the statistics which can include how you decide to structure the sample of people you want to include and how you analyse the data you have collected.

The Cambridge Dictionary says that a survey is "an examination of opinions, behaviour, etc., made by asking people questions".[1] They are commonly used in health research to answer many important questions, such as "What is the prevalence of a particular disease?", "What are peoples' self-reported quality of life?", or "What side effects have people experienced from a vaccination?". Surveys, whilst appearing simple, have many aspects that need to be thought through to obtain a high-quality answer to any research question. These include, but are not limited to:

1. Defining the research question.
2. Deciding how the sample will be chosen (sampling method).
3. Designing, developing and validating the survey questions.
4. Working out how to analyse the data accounting for the design.

In this primer, we will focus on describing sampling methods (point 2) that are commonly used in health studies.

Often, we are interested in a specific population. This might be people in a particular country, those over a certain age, or those with a particular disease. If we want to answer a question about a specific group of people, we will need to approach the population of interest with relevant questions – we call this a population-based survey. An important step in designing a population-based survey is defining the 'population' of interest. Depending on the research question, the population is usually a defined group of people. As opposed to censuses, where information is collected from everyone in the population, surveys collect information only from a subset, known as a sample, of the population. A well-conducted population-based survey will have a target population and a sampling method (a.k.a. sampling design) that ensures the sample is representative of that population. As noted in a previous primer, defining the research question happens as part of the initial study design, and refinements to the definition of the population lead to improvements in the clarity of the research question and the study design.[2]

Surveys may collect information at one time point (cross-sectional) or at multiple time points on the same sample (longitudinal). Information can be collected in the present and into the future (prospectively) or about the past (retrospectively). Depending on the aspects of the population that are important for the study, the selection of a representative sample can be complex. In health surveys, sampling designs are often tailormade to account for these complexities (for example, the New Zealand Health Survey,[3] which we describe later when talking about complex surveys), and their sampling designs are usually published. In this primer, we will introduce three commonly used sampling designs which are the basis of many more complex designs.

How the sample is chosen is critical to applying the research findings back to the population. To add to the challenge, there may be very limited flexibility (if any) on how participants can be identified and recruited depending on the research setting, related ethical and pragmatic issues, and how rare the condition of interest is. How we sample people falls broadly into two categories based on whether we know the probability of an individual being in the sample ('probabilistic sampling') or not ('non-probabilistic sampling'). Probabilistic sampling uses random selection to sample people, whereas non-probabilistic sampling uses non-random selection methods.

We will first briefly introduce the non-probabilistic methods. Non-probabilistic methods include methods such as convenience sampling (sampling based on availability), quota sampling (aiming to fill quotas from each pre-determined category), snow-ball sampling (using already selected participants to find new participants), and judgmental or purposive sampling (using researchers' judgements to decide who are suitable to be selected). With these non-probabilistic methods, we cannot know what the probability is for people to be selected into the sample and we cannot ensure that everyone in the population has a chance of being selected. For example, in a survey of local pedestrians, if we were to interview the first 100 people who walk past us on a specific local street, there will be many local pedestrians who have zero chance of being included as they never walk down that street.

Generalising findings from non-probabilistic samples to the population can be misleading. This is because the findings are very likely to be subject to bias as a non-random selection process is very likely to produce a sample that does not properly represent the underlying population. Probabilistic samples are assumed to be representative of the population; therefore, findings from those samples can be generalised to the population. As an example, a convenience sample taken from a shopping mall in a particular area is likely to capture relatively healthy people who are mobile. If this was used to measure the health status of the general population in that area it would be clearly biased. A random sample of people living in the same area (not necessarily shopping) would probably include those who are unable to get to the shops and would provide a more reasonable estimate of health. Despite their drawbacks, non-probabilistic sampling methods are needed pragmatically in some situations where it is impossible to use a more representative sampling method. This is often the case when participants are difficult to find (rare conditions) and when the cost per-participant is highly variable therefore selecting participants that are cheaper to recruit can make the study feasible.

We will now focus purely on the probabilistic methods. All probabilistic sampling methods involve the random sampling of people at some stage of the selection process. Selecting participants randomly is a way of minimising bias in the sampling process. The three most commonly used basic sampling methods in health research are described below. More complex surveys often include combinations of these basic methods in multiple stages (multistage designs).

1. Simple random sampling (see Figure 1): The sample is chosen by randomly selecting individuals directly from the population, so that each individual in the population will have the same chance of being selected. For this, a complete list of individuals in the population is needed (called a 'sample frame') and from this list, participants are randomly chosen. Most basic statistical analysis methods are based on this 'equal probability of being selected' assumption. Due to the advantages of simplicity in sampling method and subsequent data analyses, this design is very common. Random sampling is usually a part of other probabilistic sampling methods, so, unfortunately, the terms 'random sampling' and 'simple random sampling' are used interchangeably in the literature. A drawback of simple random sampling is that it may not capture a sufficient number of people from minority groups in the population, which may undermine the ability to make group-specific assessments or comparisons between groups.
2. Stratified random sampling (see Figure 2): The population is divided into groups (strata) based on one or more relevant characteristic(s). People can only belong to one group (mutually exclusive) and everyone is included (exhaustive). Simple random sampling is then used to select people from within each stratum. The proportion of people chosen from each stratum can reflect the size of the stratum or can be altered to oversample certain groups. Stratified sampling can help increase the accuracy and precision of estimates because it reduces the variability within each stratum. It allows for the comparison and analysis of different subgroups in the sample and reduces survey cost by overcoming some practical challenges. However, incorporating the stratified sampling design into the analysis plan is complicated compared to a simple random sample.
3. Cluster sampling (see Figures 3A & B): Often the population is made up of groups or clusters (for example, aged care facilities where elderly residents live). Single-stage cluster sampling randomly selects clusters (such as aged care homes) where everyone in the selected clusters are included in the sample (see Figure 3A). Cluster sampling often extends to multistage sampling. This type of sampling involves randomly selecting smaller clusters within previously selected clusters, or randomly selecting individuals within those clusters (see Figure 3B), or both. For example, we may want to sample patients who underwent a specific treatment in New Zealand hospitals last year. We can start by sampling hospitals (first stage) and then sampling patients within these selected hospitals (second stage).

It is very important to understand the sampling design used to collect the data before starting any analysis of survey data. Even though everyone in the population will have the same probability of being selected using simple random sampling, this is not the case in the other two sampling methods. People in stratified, cluster and multistage sampling can have varying probabilities of being selected into the sample. Cluster sampling has additional complexity as it introduces an element of correlation (i.e. non-independence) among people within clusters. Information collected from people within a cluster are more likely to be similar to each other than to people from other clusters, which breaks the independent observations assumption in basic statistical methods. This must be accounted for in the statistical analysis with more complex analysis methods that allow for a non-equal probability of selection and/or non-independence of observations.

Complex surveys often combine these basic sampling methods: stratification, clustering and simple random sampling. All three could be used within the same design, for example, as the New Zealand Health Survey did.[3] Therefore, it is critical that data analysis methods pay special attention to the sampling design, especially when the survey data are not from a simple random sample.

Another natural question to ask as part of the survey design is, how many participants need to be recruited? For more complex designs, this may be extended to how many participants are needed in each stratum, or how many clusters are required? Survey designs become complex once they move away from simple random sampling and require specific expertise to estimate sample size beyond the simple methods discussed in our earlier primer.[2] We highly recommend collaborating with a biostatistician to work through the design of complex surveys, to ensure the design is appropriate for the research questions and the analysis plan is appropriate for the data structure.

Going back to our opening question, have you ever thought of running a survey? We hope this primer has highlighted some of the challenges of conducting a survey. In our experience, they are never as simple as they first appear. We have introduced three commonly used probabilistic sampling methods; however, there are many other issues beyond the scope of this statistics primer that can occur in surveys. These include other aspects such as questionnaire design, analysis methods that are used for complex designs, and sources of bias such as recruitment and participation challenges. Our advice is to spend time thinking through all aspects of the design and analysis before starting; that way, many of these issues can be avoided or minimised.

References

1. Cambridge Dictionary. Meaning of survey in English. Cambridge Dictionary; [cited 2023 Sept 13]. Available from: https://dictionary.cambridge.org/dictionary/english/survey.

2. Samaranayaka A, Cameron C, Turner R. Sample size in health research. NZ Med Stud J. 2021 Apr 22;32:52-4. doi: 10.57129/XPSC9819

3. Ministry of Health. New Zealand Health Survey Methodology Report. Wellington: Ministry of Health; 2012. Available from: https://www.moh.govt.nz/notebook/nbbooks.nsf/0/38A0B04E793F5F1ECC257DC60008B736/$file/new-zealand-health-survey-methodology-report-v2.pdf.

About the authors

› Dr Ari Samaranayaka, BSc, MPhil, PhD, is a Senior Research Fellow and Biostatistician, Biostatistics Centre, Division of Health Sciences, University of Otago
› Professor Robin M Turner, BSc(Hons), MBiostat, PhD, is the Director, Biostatistics Centre, Division of Health Sciences, University of Otago
› Associate Professor Claire Cameron, BSc(Hons), DipGrad, MSc, PhD, is a Research Associate Professor and Biostatistician, Biostatistics Centre, Division of Health Sciences, University of Otago

Correspondence

Dr. Ari Samaranayaka: ari.samaranayaka@otago.ac.nz

Figures

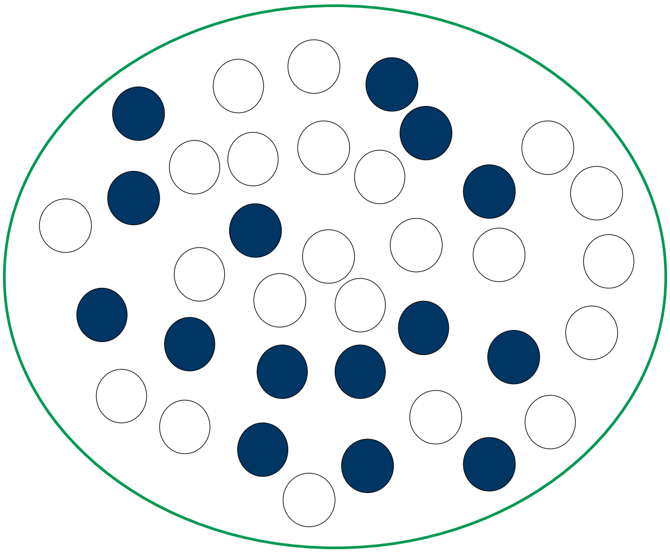

Figure 1: *Simple random sampling from a population (contained inside the green oval) of people (small circles). People have been randomly selected into the sample (filled-in blue circles) and all people in the population have an equal chance of being selected into the sample.*

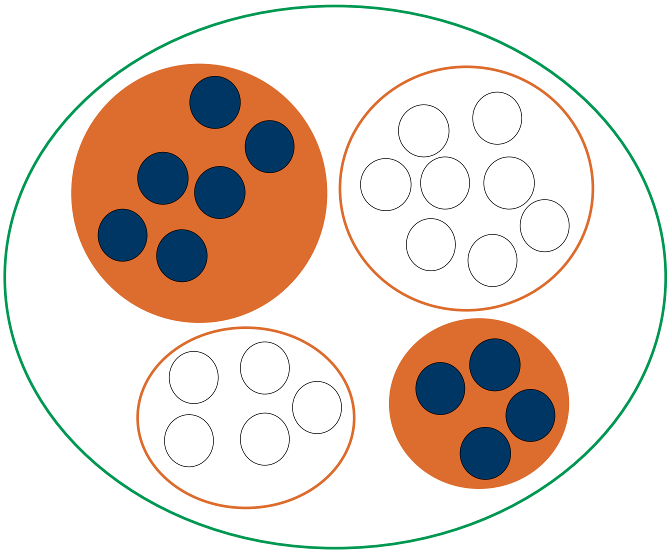

Figure 3A: *Single-stage cluster random sampling from a population (contained inside the green oval) separated into four clusters (large circles with orange outer lines). Two clusters are randomly selected (filled-in orange circles) with each cluster having an equal chance of being selected. Within randomly selected clusters, all people are included into the sample (filled-in blue circles).*

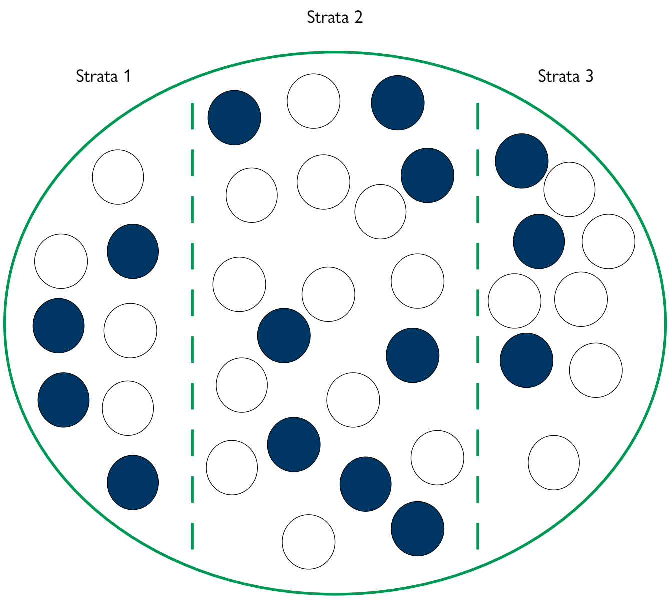

Figure 2: *Stratified random sampling from a population (contained inside the green oval) separated into three strata (shown by the dashed lines). Within each stratum, people (small circles) have been randomly selected into the sample (filled-in blue circles). All people within a single stratum have an equal chance of being selected into the sample.*

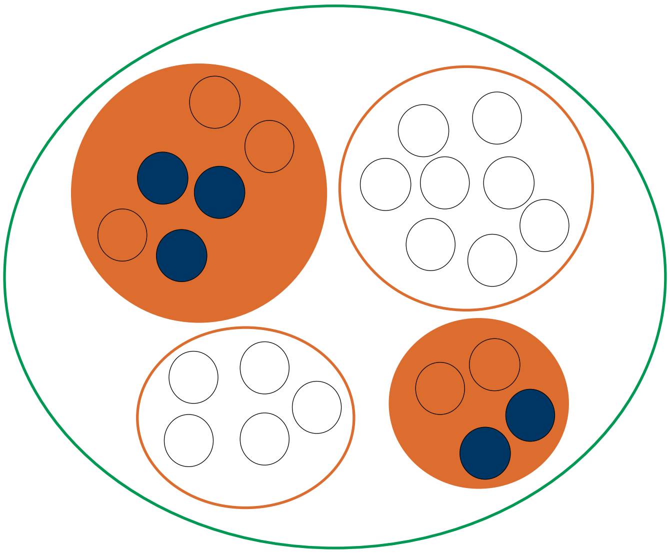

Figure 3B: *Multistage cluster random sampling from a population (contained inside the green oval) separated into four clusters (large circles with orange outer lines). Two clusters are randomly selected as in Fig 3A (filled-in orange circles). Within randomly selected clusters, people (small circles) are then randomly selected into the sample (filled-in blue circles). All people within a selected cluster have an equal chance of being selected into the sample.*