**INVITED FEATURE:** STATISTICS PRIMER

# Health related surveys: sampling weights

Ari Samaranayaka, Robin M Turner, Claire Cameron

## Introduction

In a previous article in this statistical primer series, we introduced commonly used sampling methods in health-related surveys.[1] They include simple random sampling, stratified random sampling, and cluster sampling. If you have not read that primer already, or feel you need a refresher on the topic, we suggest reading it before reading this article. A single complex survey can include a combination of the above sampling methods. All of these sampling methods aim to achieve a sample that is representative of the underlying study population. Simple random sampling achieves this representativeness by ensuring (as much as possible) that everyone in the population has the same chance of being selected. The assumption is made that observations from different participants are independent of each other, and each participant represents the same number of individuals in the population, therefore those observations contribute equally to the findings. These assumptions mean that basic statistical methods can be used to analyse the data from a simple random sample. As you can probably imagine, there are many situations where this sampling method is not possible or appropriate. Examples of circumstances where simple random samples are not possible include, but are not limited to: instances where we are not able to list the people in the population (non-availability of a sampling frame that includes everyone), which is needed to choose people at random; when participation of individuals who have been sampled are different between its subgroups; and when the number of people selected from minority groups is too small.

More complex situations require more complex sampling methods and more complex analyses. In stratified sampling, the population is divided into 'strata', which is the word used to describe subgroups within the structure of a sample. For example, a stratified sample may be grouped into strata by sex, ethnicity or occupation. Because the stratification is based on one (or more) characteristic(s) relevant to the research aim, the strata sizes are almost always different. Participants within a stratum are more likely to be similar to each other than those in different strata, thus participants' observations within a stratum tend to be more similar to each other. Similarly, in cluster sampling, irrespective of whether single-stage or multi-stage sampling is used, observations from participants within a cluster become correlated. This breaks the independence assumption required for many basic statistical methods. The independence assumption says that all observations are independent of each other. For example, if we select people into a sample from a household (a form of cluster) we know that people within households are much more similar, i.e., are correlated with each other, compared with people in other households. This means people across the sample are no longer independent of each other, they fall into clusters where they are correlated with each other and the basic assumption of independence is broken.

Complex sampling methods (i.e., sampling methods that are not simple random sampling) do not require everyone in the population to have the same chance of being selected. That means different participants in the sample represent different numbers of individuals in the population, making the 'contributions' from the observations unequal. This article will focus on this particular challenge, the non-equal contribution of the participants due to the sampling design, describing the use of sampling weights to address it in the analysis.

## What is a sampling weight and why it is important?

A sampling weight is a numerical measure that represents the 'relative contribution' of each person in the sample. For an individual in the sample, the sampling weight is (usually) one divided by the probability of the person being selected (also known as the inverse probability), and it can be understood as the number of individuals in the population that can be represented by the sampled person. For example, if the probability of being selected is 0.2 for a particular individual in the sample, the sampling weight is 5 – that is, 5 people in the population are represented by that one person. In simple random sampling, the equal probability of being selected makes the sampling weights the same for everyone (i.e. everyone in the sample represents the same number of people in the population), therefore sampling weights are not required when analysing such data. For more complex sampling methods where sampling weights are different between individuals, observations are weighted by this measure in the analysis to reflect their relative contributions.

Let us illustrate this using a simplified hypothetical example in the context of stratified sampling[1] compared to simple random sampling. We decide we want to estimate the prevalence of asthma in New Zealand using a survey. We need 1000 people in our sample to estimate that prevalence with a pre-determined level of precision.[2, 3] To keep this example simple, we will assume that the New Zealand population is 5 million. We will assume that 15% of the population (N=750,000) are living in substandard houses; we will refer to this group as Group A. The rest of the population (N=4,250,000) live in healthy homes; we call them Group B. First, we will look at what taking a simple random sample would look like and how it would be analysed. Using simple random sampling, rather than sampling from groups A and B separately, we will expect everyone in the population to have the same chance of being selected. If we took a sample of 1000 people, we would expect about 150 individuals to come from Group A (15% of 1000) and about 850 from Group B (the remaining 850, see Table 1). We use the word 'about' for those sample sizes because, in reality, when we take a sample of 1000 people, there are practical reasons why we may not get exactly 150 people from Group A and exactly 850 people from Group B. This is called sampling uncertainty. Suppose the prevalence of asthma is 40% among Group A and 20% among Group B (these are unknown in practice). This corresponds to 300 000 people from Group A (40% of 725 000) and

*Table 1: Estimating the prevalence of asthma using simple random sampling*

| Strata | Population | Sample Size | Prevalence in the Population (Unknown) | Number with Asthma in Sample | Prevalence Estimate from Sample |
|---|---|---|---|---|---|
| Group A | 750 000 | 150 | 40% | 60 | 40% |
| Group B | 4 250 000 | 850 | 20% | 170 | 20% |
| Total | **5 000 000** | **1 000** | **23%** | **230** | **23%** |

850 000 from Group B (20% of 4 250 000) having asthma, adding to 1 150 000 out of 5 000 000 people, giving a (unknown) population prevalence of 23%. In our sample, we can expect to have about 60 people with asthma from Group A (40% of 150), and about 170 people with asthma from Group B (20% of 850). In total, we expect to have about 230 people with asthma in our sample. This leads to an estimated prevalence of asthma in the sample of 23% (i.e., 230 out of 1000). This means that ignoring the stratification in this estimation did not alter our result because everyone in the population had an equal chance of being included in the sample under simple random sampling.

Now, assume we want to estimate the prevalence of the individual groups as well. Our sample from Group A is too small to provide a precise estimate. To get a more precise estimate of the prevalence of asthma in Group A, we would need to obtain a larger sample from that group. We call this 'oversampling'. We go away and do some sample size calculations[2] and decide that a minimum sample size of 400 will be required to estimate the prevalence with enough precision in each subgroup. We then set about recruiting a new sample from the population, but this time we will recruit 400 from Group A and 600 from Group B so that the total sample size remains the same (see Table 2). To plan this sampling, we need to use a stratified sampling design: stratify our population into Group A and Group B, then use simple random sampling within each stratum separately. We can now expect (ignoring sampling uncertainty) to have about 160 people with asthma from Group A (40% of 400), and about 120 people with asthma from Group B (20% of 600) in our sample. This gives a total of about 280 people with asthma in the sample. If we ignored the study design and calculated the prevalence we would get 280 out of 1000 or 28%. We can clearly see that this has overestimated the true population prevalence of 23%. Why does this happen? People in Group A who have a higher rate of asthma are much more likely to be included in the sample (because of the stratified design) than people in Group B, pushing up the prevalence estimate to be higher than it actually is. This shows that if we ignore these different probabilities of being selected into the sample, we will most likely get a wrong answer.

Another way to think about this is, if we look a bit deeper at the strata, we can see our sample of 400 people from Group A represents 750,000 people in the population. This equates to each person in the sample representing 1875 people in the stratified population (the population of Group A; see Table 2). In contrast, the 600 people in the sample from Group B represent 4,250,000 people in the population, or each Group B person in the sample represents 7083.3 people in the stratified population (the population of Group B). These 1875 and 7083.3 are the sampling weights. We can use these strata-specific sampling weights to estimate the true population prevalence correctly. We do this by 'weighting' each observation with the corresponding sampling weight, that is, multiplying the number of people with asthma in the stratified sample by the sampling weight, to estimate the number of people with asthma in the stratified population. In this way, we can estimate the number of people with asthma as 160 × 1875 = 300,000 among the Group A population, and as 120 × 7083.3 = 850,000 among the Group B population as shown in Table 2. This leads to a total of 1,150,000 people with asthma (out of 5,000,000), or an overall prevalence of 23%. This overall prevalence estimate agrees with the actual prevalence in the population.

### Adjustments of the sampling weights

Sampling weights, as described above, are purely based on the sampling design. Therefore, they are more precisely referred to as design weights or base weights to indicate that they were not adjusted for other factors. Sometimes they are called probability weights because they are based on the probability of being in the sample. In some studies, in addition to the features of the sampling design, other unavoidable factors influence the number of individuals in the population that are represented by a single person in the sample. For example, in the above example, suppose of the 400 people selected from Group A only 375 participated in the survey (the other 25 did not respond to the survey invitation, were incapable of participating or refused to participate). To account for that, we can multiply the design weight of 1875 in that group by the non-response weight of 400 / 375 = 1.0667, to arrive at the adjusted sampling weight of 2000. Such adjustments are context-specific and account for known discrepancies between the sample and the population to ensure representativeness.

### When should we use sampling weights?

The simple example above shows how the population level estimates (such as means and proportions) are not correct if we ignore the sampling weights when they are required. The weighting only applies to how we combine subgroups of the population (i.e., strata or clusters) to arrive at population level estimates. It does not make sense to apply weights to the stratum-specific estimates. We can clearly see in the tables above that the stratum-specific prevalence estimates are the same as the true prevalence in each stratum irrespective of using sampling weights. Weighting for oversampling of specific strata makes no difference to these estimates as stratum-specific estimates are direct estimates. To demonstrate this, we estimated the stratum-specific prevalence using sampling weights in Table 2 (last column), which shows estimates are the same as in Table 1 (last column) where sampling weights were not used. For comparison of stratum-specific estimates, oversampling a particular group does not alter the estimate of the difference (or ratio) between groups (however, the precision of that estimated difference can be influenced if sampling weights are not used. We ignore that issue in this simple article). Therefore, we can ignore the sampling weights when we compare strata. This can be generalised to other types of analyses, like hypothesis tests and regressions, where the aim is to compare strata rather than to measure across strata combined. Therefore, the decision on whether to use sampling weights depends on the aim of the analysis. This is an im-

*Table 2: Estimating the prevalence of asthma using stratified random sampling*

| Strata | Population | Sample Size | Prevalence in the Population | Number with Asthma in Sample | Prevalence estimate ignoring sampling weights | Sampling weight | Estimated number with asthma in population | Prevalence estimate using sampling weights |
|---|---|---|---|---|---|---|---|---|
| Group A | 750 000 | 150 | 40% | 160 | 40% | 1 875.0 | 300 000 | 40% |
| Group B | 4 250 000 | 850 | 20% | 120 | 20% | 7 083.3 | 850 000 | 20% |
| Total | **5 000 000** | **1 000** | **23%** | **280** | **28%** | | **1 150 000** | **23%** |

portant and often misunderstood point. If the purpose of an analysis is to estimate something at the population level, then weights should be used. If the purpose is to compare between strata or present stratum-specific estimates, then weights are not needed. It gets more complicated when an analysis requires only a subset of the data (e.g., only those in a certain age range that goes across strata). In this situation, the underlying population for that analysis is different from the underlying population on which the calculation of sampling weights was based on. Therefore, the original sampling weights, as described above, are unlikely to be correct for this subset of analyses. It is also worth noting that the use of sampling weights cannot fully correct for all sources of bias, such as non-coverage, non-response, measurement errors, processing errors, or population changes. Also, weighting may introduce new issues such as variance inflation (i.e., reduce the precision in estimates). Therefore, sampling weights should be used with caution and transparently.

Due to privacy issues, publicly available datasets from complex surveys (e.g., the NZ Health Survey or NHANES) do not usually include each individual's strata level and/or cluster level information required to calculate sampling weights as described above. Instead, they provide the sampling weights, sometimes a series of them, to suit the most common subsets of data.

In this article, we focus on sampling weights. In the literature, we can see several other types of weights such as frequency weights and analytic weights, and the analyses using them are called weighted analyses. Frequency weights are whole numbers (i.e., positive integers) used to tell the statistical software how many identical cases are in the data. It is a kind of shortcut for data entry. For example, if you have 10 rows of identical data you can use one single data row with a frequency weight of 10. Analytic weights are used to indicate the precision of the data rows when their precisions differ. For example, when the data comprises summary statistics with varying levels of precision (i.e., variance), we want the most precisely measured data (i.e., lowest variance) to have the greatest weight. This can be achieved by having analytic weights proportional to the inverse of the variance. You can see that when reading the literature, a weighted analysis may refer to a variety of situations. You need to look carefully at what kind of weights they are referring to. Sampling weights, frequency weights, and analytic weights are used for different purposes and have different implications in terms of an analysis.

You may have noted that the overall purpose of sampling weights is to make estimates from a complex sample reflect those same measures (i.e., parameters) from the population. This is done by appropriately weighting observations when they are known to be non-representative. Such samples are very common in health research. For example, it is easy to find multicentre studies in the literature that are used to estimate overall population characteristics. Statistically, a centre is a cluster, meaning the sample is not a simple random sample. Most often, the sample sizes from each of the centres are different and determined for non-statistical reasons, such as the availability of participants. Any analysis done on these data cannot ignore the clustering, and weights should be used if estimates are being made for the population. Three important skills for users of the literature on complex samples (e.g., medical practitioners) are (1) the ability to identify the source of any unequal contribution of participants to the outcome of a study; (2) the possible influence of those observations on study findings; and (3) the understanding that sampling weights can be used to enhance the validity of findings by making the participants representative of the population. In addition to those who use administrative data for health research, these skills are particularly useful for emerging clinicians when appraising clinical literature.

Sampling methods used in health research are often more complex than our simple example. Generally, the calculation of sampling weights and their use in subsequent statistical analyses will benefit from input from a biostatistician whenever the sampling method differs from simple random sampling.

**References:**

1. Samaranayaka A, Turner RM, Cameron C. Health related surveys: sampling methods. N Z Med Stud J. 2023;37:40-42. doi: 10.57129/001c.122998

2. Samaranayaka A, Cameron C, Turner RM. Sample size in health research. N Z Med Stud J. 2021;32:52-54. doi: 10.57129/XPSC9819

3. Cameron C, Turner R, Samaranayaka A. Understanding confidence intervals and why they are so important. N Z Med Stud J. 2021;33:42-43. doi: 10.57129/AGAG5939

**About the Authors**

› Associate Professor Ari Samaranayaka, BSc, MPhil, PhD, is a Biostatistician at the Biostatistics Centre, Division of Health Sciences, University of Otago

› Professor Robin M Turner, BSc(Hons), MBiostat, PhD, is the Head of the Department of Preventive and Social Medicine and a Biostatistician, Division of Health Sciences, University of Otago

› Associate Professor Claire Cameron, BSc(Hons), DipGrad, MSc, PhD, is the Director of the Biostatistics Centre and a Biostatistician, Division of Health Sciences, University of Otago

**Correspondence**

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