Dissertation Module: Research Skills Program
Topic 6: SAMPLING STRATEGY AND SAMPLE SIZE CALCULATION

The role of the sample is central in health research. We rarely have the opportunity of measuring the entire target population. Instead, we use a sample of this population. Sampling obviously offers a number of advantages, such as reduced cost and increased speed. However, incorrect sampling procedures can invalidate results.

There are two critical issues when considering the sample, firstly how the sample is chosen and secondly the size of the sample. The question of how to choose the sample is a question of validity. A well chosen sample will represent the target population and will therefore minimise the potential for selection bias. The question of sample size refers to reliability and addresses the question of whether the operational research hypothesis can be confirmed or rejected with statistical confidence.

Let us assume that we want to conduct a cross-sectional study about alcohol consumption and alcohol abuse in Australia with the target population being all Australians. If we recruited only students and staff from our university for our study then we will be unable to make any inferences about the target population as such a sample would *not* be representative of all Australians. In this case the sampling design was unsuitable and had an obvious selection bias.

In another attempt we carefully considered sampling a range of participants from teenagers to pensioners, and including men and women from different ethnic backgrounds across Australia in inner metropolitan cities as well as from rural and remote areas. Such a study is much more involved and resource intense, but is also more likely representative of “all Australians”. Now consider that in this second attempt we have (1) a sample size of 100 people or (2) a sample size of 10,000 people. For a start we will trust the larger sample much more. This “trust” will be mirrored by statistical “confidence”. The larger our sample size, the more precise our result will be, in statistical terms.

In this chapter we discuss some common forms of sampling and their application in epidemiology and we will also introduce the concept of sample size and show you how to calculate adequate sample sizes for specific (simple) situations.

Sampling Strategy

There are a number of sampling strategies which are commonly used in epidemiology with the choice of one specific strategy over another often depending on many factors, including the objectives of the study, the planned study design and the available resources, which is very important. The relative cost and benefit of the different methods should be carefully weighed.

Let's get started by defining two major branches of sampling strategies: probability and non-probability sampling.

Box 6.1: Definition of Probability Sampling

A sample is a **probability sample** if each individual in the target population has a known chance of being part of the sample.
Non-probability sampling means that either some groups of the target population have no chance of being sampled, or their chance of being sampled cannot be accurately determined.

Comments:

(1) **Random samples** are the best known examples of probability samples. If the target population has N individuals and the random sample will have size n, then the likelihood of each individual in the target population being selected into the sample is n/N. So let's say our target population is the 4.4 million New Zealanders and our random sample should have a size of 250, then each New Zealander has a chance of 0.006% (250 ÷ 4.4 million) of being selected into our random sample. This means that each New Zealander is equally likely to be part of the sample.
(2) Random sampling, stratified random sampling, systematic sampling, and cluster sampling are examples of probability sampling strategies. We will describe these sampling strategies in some detail later.
(3) Convenience sampling, snowball sampling, and purposive sampling will also be introduced; these are examples of non-probability sampling.
(4) Please note: Before selecting a sampling strategy, a clear definition of the target population is required! We can not determine a suitable sampling strategy if the target population is ill defined!
(5) The key feature we are aiming for is that the sample truly reflects all the (important) characteristics of the target population, in other words, we want the sample to be representative of the target population. For example, a true random sample will represent the target population.

In the following we will describe some common probability sampling methods.

Simple random sampling

Simple random sampling is theoretically the most straight forward way of obtaining a sample which represents the target population. Simple random sampling is the method of selecting n people from a target population of size N in the following way:
(1) People are independently selected, one at a time, until the desired sample size is achieved. That is simple random sampling is sampling without replacement.
(2) Each individual in the target population has an equal chance of being selected into the sample.

For example, we want to conduct a study of the consumption of ready-to-drink mixed drinks, also called alcopops, among Australian teenagers aged 13 to 18 years. We plan a cross-sectional study to estimate the prevalence of drinking alcopops. The target population is all Australians aged 13 to 18 years. In order to select a simple

random sample, a complete list of all 13 to 18 year-old Australians must be collated. This list is an example of a **sampling frame**. This list of the target population will be numbered from 1 to N and then n individuals will be randomly selected based on n different random numbers ranging between 1 and N obtained from a computer program.

This example highlights the fact that random sampling, although easy in theory, can be cumbersome or outright impossible in practice. Creating a complete list of all 13 to 18 year-old Australians is a challenge. But even if we succeeded in creating such a list, the collection of the information from the drawn random sample would require us to travel all over Australia.

Box 6.2: Notes about random sampling

(1) Please note random sampling is important because it is our best guarantee to represent the target population and hence minimize selection bias.

(2) For a simple random sample a complete list of the target population is required. Such a list is often difficult to obtain, or even impossible. If we want a random sample of all adult residents in New Zealand an almost complete list might be the electoral role or the telephone directory. Such lists allow us to identify individuals for our sample and constitute the simplest version of a **sampling frame**.

However, not everyone is in the telephone directory or on the electoral role! Hence sampling frames might be incomplete. We need to discuss the possible selection biases which can be introduced when we use these sampling frames. For example, nowadays with the ubiquitous use of mobile phones more and more young people no longer have a landline and are therefore not listed in the telephone directory.

However, incomplete sampling frames are not the worst possible situation. Assume, for example, we want to investigate language abilities and disabilities of deaf Australians. There is obviously no national register of deaf people – as there is no register for most disabilities or diseases – and hence we will not be able to obtain a sampling frame for random sampling in this group of people.

(3) Standard statistical analyses assume that the data were derived from a random sample or from a sample which can be treated as though it was a random sample. If the sample was obtained in any other way, results from statistical procedures might be invalid. For some other probability sampling strategies, such as cluster sampling, special statistical adjustments have to be made as the assumption of independent observations is violated by this sampling strategy. For non-probability samples, however, the information necessary for adequate adjustments is not usually available, probably leading to selection bias.

(4) Please do not confuse random sampling with randomisation!
Randomisation refers to the random allocation of participants to either an intervention or control group in an experimental study with the aim of achieving structural uniformity and avoiding confounding bias.
Random sampling can be used in both observational and experimental designs and refers to the way individuals from the target population are selected for the sample.

Random sampling aims to provide representativeness and is undertaken to avoid or minimize selection bias.

Stratified random sampling

In stratified random sampling the target population is first divided into sub-populations, called strata, according to the categories of an influencing variable, for example a strong known confounder. Simple random samples are then drawn from each stratum.

Suppose we want to conduct a study describing sun exposure and sun protective behaviours of outdoor workers in Queensland. Three large Queensland based companies who employ many outdoor workers have granted us access to their workers. We know from observation and information from the companies that about 95% of the outdoor workers are men. However, we also want to be able to describe sun exposure and sun protective behaviours of female outdoor workers, particularly because we suspect gender may influence our results. In this scenario a simple random sample of outdoor workers would not provide much information about female behaviour as a simple random sample of let's say 100 outdoor workers will only include about five females. Here stratified sampling could be the sampling strategy of choice. Stratified sampling in this instance would mean that two simple random samples are selected separately, one for male outdoor workers and one for the females.

The reason for using stratified random sampling in this example was the potential "under-representation" of female workers. Hence, in this example our aim would be to oversample female outdoor workers so that we could also confidently describe their sun exposure and sun protective behaviours.

Please note that in the example above, our sampling frame was based on employment with three companies in Queensland. Probability sampling within such a frame does not suggest the absence of selection bias. Depending on the companies, such a sample may or may not be representative of all Queensland outdoor workers.

Box 6.3: Notes about stratified random sampling

(1) Stratified random sampling permits us to make precise statements about the strata of the target population, even if these strata are small. Often the strata are defined by a known important confounder. In stratified random sampling we treat each stratum of the target population as a separate population in their own right.
In the example above, we could state different research hypotheses for the male and female outdoor workers and would consequently require different samples sizes to confirm or reject these hypotheses.

(2) Stratified random sampling ensures that each stratum is represented in the sample. It allows us to make inferences about sub-populations of the target population which would probably not be possible if we opted for one simple random sample.

(3) Stratified random sampling allows us to treat each stratum as a separate target population. Naturally we can then apply different sampling frames to different strata. This flexibility can sometimes provide a practical advantage.

(4) If the strata are more homogeneous with respect to the outcome measures being investigated, estimations within stratified random samples will be more precise than those obtained using overall simple random sampling. Weighted averages of stratum-specific estimations are used to obtain an overall result; however, the proportional composition of the target population with respect to the stratifying factor has to be known.

(5) Stratified random sampling might require a larger overall sample size than simple random sampling because often smaller strata are oversampled to achieve the statistical precision required. In our example of outdoor workers, we would deliberately oversample female outdoor workers so that we can describe their behaviour with statistical confidence.

Thus, there are specific situations in which stratified random sampling is preferable to simple random sampling – especially (1) if there are small but relevant strata in the target population that need to be included in sufficiently large numbers to ensure adequate outcome during statistical analysis; and (2) if a known, strong confounder exists. Since stratified random sampling equates to conducting several separate studies, more work is involved than with one simple random sample.

Often the choice between the two sampling strategies depends on the specific aim of the study. For instance, the example we used to explain simple random sampling was about a cross-sectional study to estimate the prevalence of drinking alcopops in a target population of all Australians aged 13 to 18 years. A simple random sample would be heavily weighted towards the metropolitan areas of Australia, as more than two-thirds of the Australian population lives in major cities (June 2008; Australian Bureau of Statistics, 2010a). From a market research perspective this sampling strategy might be suitable as the metropolitan market for alcopops is vast compared to rural or remote markets. However, from a public health perspective the drinking habits in rural and remote areas might be quite different from the metropolitan regions and thus might be of great interest. Hence, if our aim is identifying targets for public health campaigns, a stratified sampling approach may well be the preferred option.

Systematic sampling

Systematic sampling is another example of probability sampling. Quite a common example of systematic sampling is based on the telephone directory. For instance, the researcher could systematically select every person listed first in the left upper corner of each page of the phone book into the sample.

Systematic sampling is often much more practical than pure random sampling and in most instances also results in a sample that can still be considered random in the statistical sense. Just imagine using the phone directory as the sampling frame for a

true random sample. We would start with random numbers and would then have to identify the person by counting the respective entry - for each random number!
Problems with systematic sampling do occur, for instance, when there is some periodicity in the sampling frame which is related to the study aim. Imagine we are interested in health outcomes of cardiac surgery patients and systematically select the first cardiac surgery patient each week in a large hospital into our sample. These selected patients might not be representative of all patients as they may tend to be the more severe or the more acute cases, or they could all be operated on by one specific surgeon who usually has shift on Mondays.

The practical application of systematic sampling needs a more detailed description. So far we have only discussed the principle of systematic sampling, but we have not yet discussed sample size. The size of telephone directories varies with location. How do we make sure that we spread our sampling right across the entire target population and not just the first 100 pages, because we wanted a sample size of 100? The solution to this problem comes with the formal description of systematic sampling.

Systematic sampling can be formally described as:
Suppose that the N people (numbered from 1 to N) make up the target population and we want to select a sample of size n.
(1) We first calculate the proportion k that will be sampled:

$$\mathbf{k = round\ (\frac{N}{n})}$$

That is, we divide N by n and round the result to the next natural number.

(2) Then, a number r between 1 and k is chosen randomly.
(3) The sample with n observations is then obtained by selecting every k^{th} consecutive person from the target population starting with r: r, r + k, r + 2k, r + 3k, r + 4k,....., r + (n-1)k

For example, if we want to select a sample of 150 people from a target population of 50,000, we calculate k:

$$\mathbf{k = round\ (\frac{N}{n}) = round\ (\frac{50000}{150}) = round\ (333.33) = 333}$$

Find a random number r between 1 and 333; let's say r = 89.
And the actual systematic sample consists of the following 150 people from the target numbered target population:
89, 89 + 333, 89 + (2×333), 89 + (3×333), 89+ (4×333),....., 89 + (149 × 333) or
89, 422, 755, 1088, 1421, ..., 49706

With this systematic sampling approach, the people selected for the sample are evenly spread over the whole of the target population.

When we use the phone book for systematic sampling our sampling frame is the pages of the book. If the phone book has 500 white pages and we require a sample size of 1000, then we would systematically choose two addresses each page.

Box 6.4: Notes about systematic sampling

(1) In contrast to simple random sampling, systematic sampling requires only one random number, r, randomly chosen between 1 and k.

(2) Systematic sampling selects evenly over the entire target population.
If you were sampling houses from one street, systematic sampling makes sure that houses from the beginning, the middle and the end of the street are included. In simple random sampling, houses in the sample might – by chance – cluster at one end of the street.

(3) Systematic sampling may cause bias – especially if there is a type of periodicity (such as time or seasonal periodicity) in the original sampling frame! However, if the sampling frame of the target population is independent of the characteristics being studied, then a systematic sample can be treated like a simple random sample. In systematic sampling as in simple random sampling, everyone from the target population has the same chance of being selected – as long as the initial person is chosen randomly.

(4) Systematic sampling is often much more practical than simple random sampling. For instance, if you intend to conduct an audit and the files of your clients are all in a filing cabinet and are all of similar thickness, depending on the required sample size, you could draw a file every 10 cm. This approach might be speedier than simple random sampling where you would have to take out all the files, count them, number them etc.

But be aware that if file thickness varies, you could tend to draw out predominantly "difficult" cases which might not represent your entire clientele. However, you would end up with a nice example of bias in systematic sampling NOT related to any periodicity!

Cluster sampling

Cluster sampling is a special kind of probability sampling which involves sampling in naturally occurring clusters. Depending on the research question, these clusters could for example, be schools, day-care centres, suburbs, households, or communities. In cluster sampling a sample of clusters is chosen first and then people are sampled from these clusters.

Let's go back to the cross-sectional study aiming to estimate the prevalence of drinking alcopops among 13 to 18 year old Australians. For this example, there is no readily available complete list of Australians aged 13 to 18 years for simple random sampling. But most people in this age group attend school. Hence we might consider selecting schools (= clusters) from all over Australia and then sampling teenagers from within these schools. This approach is attractive as lists of schools exist and it would be relatively easy for us to access a large number of adolescents within

schools. However, we would need to be careful to select the schools appropriately to represent the whole of Australia.

The obvious downside of this cluster sampling frame is that not all 13 to 18 year olds are at school. Around three quarters of Australian students stay at school until they are 16 or 17 years old (Australian Bureau of Statistics, 2010b), so 18 year olds will be underrepresented and students who left school before they turned 17 will not be represented at all by this cluster sampling approach.

Let us consider a second example. Suppose an household survey is planned in an Australian city. A simple random sample of say 10,000 households covers a city more evenly than 50 city blocks containing an average of 200 households each. However, greater field costs will be incurred in locating 10,000 randomly chosen households than locating 50 blocks and visiting all the households in these blocks. Thus, in reality, the choice of using cluster sampling is often driven by external factors such as feasibility and available resources.

Please note that it is very important to select the clusters randomly from a list of all possible clusters in the target population. Otherwise the representativeness of the sample cannot be guaranteed.

Cluster sampling is sometimes called **multi-stage sampling** because it can be carried out in several stages potentially resulting in rather complex sampling strategies. In practice, a one-stage cluster sampling implies that we have a random selection of clusters and all people in these clusters are selected into our sample. Two-stage cluster sampling means that we have a random selection of clusters and then a random selection of people from these clusters. Alternatively, two-stage cluster sampling can also mean that we have a random selection of clusters and we then have a random selection of clusters within the initial clusters and from these second level clusters all people are selected into the sample. The initial clusters chosen are called **primary sampling units** (PSU).

A simple example of three-stage cluster sampling would be to choose a random selection of schools (stage 1) and a random selection of classes (stage 2) within the schools and then a random selection of students (stage 3) from the identified classes. In this example, the primary sampling units would be the schools recruited to the study. Because of random sampling at each stage we have ensured that there is representation at each stage.

Box 6.5: Notes about cluster sampling

(1) Cluster sampling is an effective sampling strategy if the individuals we intend to study occur in natural groups such as schools, suburbs, day-care centres, communities, or companies. Cluster sampling might be feasible when a complete list of the target population is unavailable and hence random sampling is not possible.

(2) Cluster sampling is often more convenient and economical than simple random sampling. For this reason, cluster sampling is often the method of choice.

(3) Cluster sampling is frequently used in cross-sectional studies. However, it is not

restricted to surveys or observational study designs; even randomised controlled trials can involve cluster sampling (see, for example, Harrison, S.L. et al., 2005)

(3) People from a cluster are often more similar to each other than people from different clusters. For example, people from one household tend to be more similar than people from different households; students attending one specific school might be more similar than students from another school; etc. The ratio of similarities between and within clusters is called the **design effect**. This design effect is often difficult to judge when planning a study – especially without in depth prior knowledge of the clusters.

The similarities within the clusters also imply that statistical estimates of the variances derived from cluster sampled data are smaller than from completely independent observations. Therefore, adjustments for cluster sampling are necessary (a) for sample size calculations as larger samples are required for clustered samples than for simple random samples; and (b) for data analyses.
The adjustment of the sample size requires knowledge of the design effect.

The analysis of clustered data from clustered samples requires specific statistical procedures which are often not routinely implemented in statistical software packages.

Please note that almost all studies use some form of cluster sampling; it is just rarely acknowledged! For example, conducting a study in one community is cluster sampling! The community being the single cluster considered. This might sound purist but is actually not! As soon as two communities are involved, the analysis should take the clustering effect into account.

Box 6.6: Warning

Irrespective of which sophisticated probability sampling technique is used, selection bias - in the form of volunteer bias – is still likely to occur simply because of the required informed consent and related ethical standards.

We insist that participants are asked to give their informed consent and that nobody can be coerced into participating in a study. As a consequence we face volunteer bias. Volunteer bias is almost always present and its effects can be highly distorting as the following historical example shows.

In the 1980s a nationwide survey of all US psychiatrists was conducted (Gatrell N., et al., 1986). A total of 5,574 psychiatrists practicing in the United States were approached by a postal survey and 1,442 (26%) returned the questionnaire. Of the 1057 male responders 7.1% admitted to having had sexual contact with one or more patients. Of the 257 female responders, 3.1% confessed to this practice. Eighty-eight percent of the sexual contacts occurred between male psychiatrists and female patients. All offenders who had been involved with more than one patient were male.

It is obvious that, although the initial sample was the entire target population of all US psychiatrists, the respondents were unlikely to be representative. Given that the

Hippocratic Oath outlaws sexual contact between medical professionals and their patients, it seems unlikely that there were claims of sexual contact that did not actually occur. It also seems highly likely that there were a number of psychiatrists who did in fact have sexual contact with patients but did not admit to them and/or did not return the survey. Thus in this example, it is most likely that information and selection bias occurred leading to underestimation of the true prevalence of sexual contact between psychiatrists and patients, despite the fact that initially all US psychiatrists were approached.

Non-probability sampling

From a theoretical perspective probability samples are always preferable as they provide the best chance of achieving a fair representation of the target population, assuming that the people selected will actually consent to participate. In contrast, non-probability sampling is likely to introduce selection bias and thus provide samples which are not representative of the target population.

However, many medical and epidemiological studies are actually based on non-probability samples. Why? Because achieving a probability sample is challenging, often resource intense and sometimes outright impossible. For instance if we do not have a sampling frame, then we cannot achieve a probability sample. For example, patients with a defined disease such as stroke or diabetes, or people with a disability such as blindness or having a foot amputated are generally not on a register. Consequently, if we wanted to sample such people for a study we would not be able to sample randomly as there is no complete list and hence no sampling frame for probability sampling! Of course we could try to create such a complete list, but as Martin Bland points out, this might be prohibitively complex as access to patient data from numerous health care providers across the country would be required (Bland, M., 2000).

In these situations it is much easier to opt for a **convenience sample** such as consecutive patients or clients from an hospital or other health care providers. Convenience samples are non-probability samples that are exactly that: conveniently available. “Man in the street” surveys are another example of convenience samples.

With convenience samples, we need to ask ourselves whether the patients who attend *our* clinic, practice or hospital are truly representative of all patients we wanted to talk about. Indeed magnitude and direction of the operating selection bias in convenience samples is often quite difficult to judge. However, as Bland suggested (Bland, M., 2000), consecutive patients for randomised controlled trials – while not optimal – are often still satisfactory as a randomised controlled trial aims to establish the difference between an intervention and a control group. And such a difference should remain similar whether the study was conducted with consecutive patients in Sydney or in Darwin; provided that the same eligibility criteria were used.

Another example of a non-probability sampling strategy is called **snowballing**. Here, selected participants are asked to recruit other people into the sample. A classic example of a snowball sample comes from one of the early American smoking studies by E. Cuyler Hammond and Daniel Horn (1954). Hammond and Horn

recruited large numbers of Americans in a very short time by asking all members of the American Cancer Society to recruit participants for the study from amongst their friends and families. However, in 1954, most members of the American Cancer Society were well educated white Americans as were most of their families and friends. Thus non-white Americans, immigrants, or people with lower levels of education were poorly represented in this otherwise impressively huge cohort study of 187,766 American men aged 50 to 69 years.

Although this example shows that snowball sampling has its advantages as it can be quick and effective in recruiting large numbers of people, it is generally frowned upon nowadays as the associated selection bias is usually substantial. However, snowball sampling is still sometimes used for sampling in "hidden" populations which are difficult to reach or engage otherwise, including sampling of illicit drug users, victims of violent crimes, or migrant minorities.

Lastly, the non-probability sampling method of **purposive sampling** should be mentioned where the researchers purposefully select into the study the people they think can provide valuable information. Purposive sampling is only sometimes used in quantitative epidemiological research when there are few people available to study but it is usually restricted to qualitative research where it is often the method of choice. Please note that sampling methods in quantitative and qualitative research differ markedly. Representativeness of the sample is an essential keystone in quantitative research while qualitative research which focuses on the richness of the data collected, often deliberately selects non-representative "key informants" as the sample.

Sample Size Calculation

In his book "Epidemiology" Leon Gordis (2008) begins his comments about sample size calculation in clinical trials with the following story:

> At a scientific meeting some years ago, an investigator presented the results of a study he had conducted to evaluate a new drug in sheep. "After taking the drug," he reported, "one third of the sheep were markedly improved, one third of the sheep showed no change, and the other one ran away." (Gordis, L., 2008)

Each quantitative epidemiological study requires an appropriate sample size to allow estimations and comparisons with some pre-defined statistical confidence. It is a waste of resources if the study is too small, that is under-powered, or unnecessarily large, that is over-powered. The term **power** here relates to its statistical meaning, that is, the probability of detecting an existing difference which is statistically significant.

Box 6.7: Developing a feel for statistical confidence

Assume we want to evaluate an improved oral rehydration formulation to reduce diarrhoea mortality in children from sub-Saharan Africa (based on Santosham, M. et al., 2010). We are planning a randomised controlled trial to compare childhood mortality caused by diarrhoeal diseases in a group of children who receive the new oral rehydration formulation with a group of children who receive standard care. We follow the children for one year to assess mortality.

(1) Let us assume that in a first study we randomise 5 children with diarrhoea to the intervention and 5 children with diarrhoea to the control group. After one year, one child in the intervention and two children in the control group have died.
This is, at face value, an impressive mortality improvement of 20%! However, based on 10 children only. Thus it seems intuitively likely that another such study which is based on another 10 children may have quite different results.
That is, in studies based on small sample sizes, it is likely that observed differences occur by chance alone (attributable to random error only).

(2) Now suppose we conducted our study with 1,000 children in each group and 100 children in the intervention group and 300 children in the control group died during the one year follow-up period.
This is again a mortality improvement by 20%, however, this time we are much more confident in the result because it is based on a large sample size. And this is exactly what is meant by **statistical confidence**.

Comments:

(1) Results from a large sample are much less likely to be due to random error alone than results from a small sample.
(2) While planning an epidemiological study, biostatistical methods are used to calculate the optimal sample size required to ensure that the operational research hypothesis can be assessed with statistical confidence. Please note that sample size calculations are important so that random error can be

controlled.

(3) Sample size calculations have to be conducted before the recruitment of participants begins.

We have regularly emphasized the importance of the operational research hypothesis and you now know that in quantitative epidemiology the operational research hypothesis states the expected outcome quantitatively. For example, we might wish to estimate the prevalence of hearing loss in children and hypothesize that it is about 2 per 1,000 Australian children in our target population. Alternatively, let us assume that we wish to improve the quality of life of younger and middle-aged people in residential care by expanding physiotherapy and speech pathology services. The intervention will be evaluated against a control group of usual care. The outcome measurement is a quality of life scale with scores ranging between 0 (= worse) and 100 (= best). We hypothesize that the intervention group will have an average quality of life score of 60 while the control group will have an average score of 40. These quantitative guesses of the expected hypothesized outcomes build the basis for estimating an appropriate sample size.

The **power of a study** in a statistical sense is the probability that the statistical procedure, which is later applied to the collected data to assess the outcome, is able to verify the stated hypothesis. A study is under-powered if there are not enough people in the sample to assess the data with adequate statistical confidence. An under-powered study has a sample size which is too small to verify the research hypothesis. A study is over-powered if there are too many people in the sample. In this case all statistical procedures will return a significant result, whether the results are clinically relevant or not. Resources are wasted in both under- and over-powered studies. As you can imagine, under-powered studies are much more frequent. Please note that studies which do not have an appropriate sample size calculation are unethical, as they are either over- or under-powered. Therefore, we should calculate a sample size based on our pre-stated research hypothesis. This calculated sample size is an optimal sample size which is minimally required to verify the outcome presented in the research hypothesis.

The appropriate formula for sample size calculations differs with the study design, the measure of association used and the operational research hypothesis. However, all formulae include the hypothesized outcome, a measure of the variability of this outcome, and the level of statistical confidence and power in some format. The latter are usually set to certain internationally accepted values, such as, 95% confidence (5% error) and 80% or 90% power.

There are countless formulae for sample size calculations depending on the study design and the hypothesis. In many situations conducting numerical calculations by hand can be a bit cumbersome, so nowadays readily available computer software is routinely used for sample size calculations. The Division of Biostatistics of the Department of Epidemiology and Biostatistics of the University of California, San Francisco, supports a website which provides a list with links to power and sample size programs (http://www.epibiostat.ucsf.edu/biostat/sampsize.html). Some software is freely available; other packages need to be purchased.

Sample size calculation for the prevalence estimated by a cross-sectional study

It is beyond the scope of this introductory text to delve too deeply into the subject of sample size calculations. Suffice here to introduce one simple formula for calculating an adequate sample size as an example; it is for a cross-sectional study, based on a single sample with prevalence as the outcome.

Let us consider the following example. Researchers in New Zealand conducted a cross-sectional study on *Helicobacter pylori* infection in female high school students (based on Fraser, A.G., et al., 2010). *Helicobacter pylori* is a bacterium that can inhabit the stomach. It was first discovered in 1982 by Dr. Barry Marshall and Dr. Robin Warren of Perth from Western Australia. At the time the conventional thinking was that no bacterium can live in the human stomach because of the acidic environment. Marshall and Warren rewrote the textbooks by providing new information about the causes of gastritis and gastric ulcers. In recognition of their discovery, they were awarded the 2005 Nobel Prize in Physiology or Medicine ("The Nobel Prize in Physiology or Medicine 2005", 2011).

Back to our example, let us assume the researchers hypothesized that the prevalence of infection with *Helicobacter pylori* varies with ethnicity and is highest in students from a Pacific Islander background. Suppose the quantification of this hypothesis was that the expected prevalence of *Helicobacter pylori* in female students from a Pacific Islander background is 50% and that they wanted to estimate this prevalence with 95% confidence and with a precision of 10%. The precision statement implies that the researchers aim to determine the true prevalence within a range of ± 10%. The correct formula for this situation is provided in Box 6.8.

Box 6.8: Sample size calculation for a cross-sectional study estimating prevalence

The formula: **Sample size = n =** $\dfrac{z^2 \times p(1 - p)}{d^2}$

gives the sample size to estimate an hypothesized prevalence p with a precision d and a confidence level defined by z.

In more detail:

(1) z describes the statistical confidence.
For example, z=1.96 translates to 95% confidence; z=1.68 translates to 90% confidence

(2) p is the expected, hypothesized prevalence

(3) d describes the intended precision when estimating the prevalence; d = 0.1 means that the true prevalence will fall within ± 10% of the estimated prevalence from the sample (with the intended confidence).

Please note that this formula is the formula for a 95% confidence interval of a

proportion. Please refer to a statistics textbook such as Bland's Introduction to Medical Statistics (2000) for a more in depth explanation of this formula and the theory behind it.

Let us now put the values from our New Zealand example into this formula. The hypothesized prevalence of *Helicobacter pylori* was 50% i.e. p = 0.50. The level of confidence was set to 95% which means that z = 1.96. The precision was chosen to be 10% which means that d = 0.10.
The optimal sample size for this situation is therefore:

$$\mathbf{n = \frac{z^2 \times p(1 - p)}{d^2} = \frac{1.96^2 \times 0.5(1 - 0.5)}{0.1^2} = 96.04}$$

Thus the New Zealand researchers would need 97 female high school students of Pacific Islander background to be able to estimate the prevalence of *Helicobacter pylori* infection in this group with 95% confidence and 10% precision. Ten percent precision means that the researchers are happy with an estimate of the true prevalence of *Helicobacter pylori* that lies somewhere between 40% and 60%.

Let us now have a closer look at this sample size formula.

$$\mathbf{n = \frac{z^2 \times p(1 - p)}{d^2}}$$

It is clear that as soon as we state the precision (d) and the confidence level (z), the sample size n is only dependent on p(1 - p). Table 1 lists the values for p(1 - p) for different prevalence values p. Please note that the values of p(1 - p) are symmetrical around p=0.5 (e.g. 0.16 for both p=0.2 and p=0.8) and therefore only one half needs to be listed.

It can be seen from Table 1 that p(1 – p) is largest for a prevalence of 0.5. In addition, there is little variation when the prevalence ranges between 0.5 and 0.3. Consequently, if you are unsure about the expected prevalence, assume a prevalence of 50%. This will give the largest sample size estimation, and your study will be on the safe side!

Table 1: Relationship between prevalence p and p(1 – p).

p	p(1 – p)
0.5	0.25
0.4	0.24
0.3	0.21
0.2	0.16
0.1	0.09

The sample size formula is also critically dependent on the intended precision d. Table 2 shows the required sample sizes for different precisions d, assuming a statistical confidence of 95% and a prevalence of 50%. Table 2 shows that the sample size is very sensitive to the precision intended when estimating the

prevalence. Thus, the achievable precision often depends on the resources available!

Table 2: Relationship between precision d and sample size, assuming p = 0.5 and z = 1.96.

Precision d	Sample size
0.2	25
0.1	97
0.05	385
0.025	1537

Box 6.9: Please note

(1) The above sample size formula applies only to **random samples** or samples which can be treated as random samples. If a more complex sampling strategy is used (such as cluster sampling) we need to adjust the formula for the design effect.

(2) The above sample size formula is correct if **one operational research hypothesis** is investigated. If we want to assess several operational research hypotheses within one study, then the sample size calculation needs adjustment for multiple testing.

The operational research hypothesis is at the centre of a study. A well constructed operational research hypothesis, based on previous published findings or on preliminary results of a pilot study, will state expected findings which are clinically relevant and close to reality. Such a carefully structured operational research hypothesis will enable the researchers to calculate an adequate sample size and conduct the research using an adequately powered study and hence in an economical and ethical way!

Summary

- A sample is a probability sample if each individual in the target population has a known chance of being part of the sample. Non-probability sampling means that either some groups of the target population have no chance of being sampled, or that the chance of being sampled cannot be accurately determined.
- Simple random sampling means selecting people out of a target population in such a way that each individual in the target population has an equal chance of being selected into the sample.
- Random sampling, systematic sampling, and cluster sampling are probability sampling approaches.
- A sample should represent the target population, that is, possess all the (important) characteristics of the target population. A sufficiently large random sample is likely to represent the target population if all the people selected actually agree to participate.
- Sample size calculations are important as they control random error. Results from small samples are more likely to be affected by random error than results based on large samples.

- A sample size calculation allows the researcher to compute the optimal size for the planned study and is important from a practical point of view as it provides information for assessment of resources required.
- A sample size calculation is equally important from a theoretical perspective. The optimal sample size allows us to reject (or otherwise) the operational research hypothesis with statistical confidence – and thereby answer our research question.
- An optimal sample size is also important from an ethical point of view.
- There are numerous different formulae available for sample size calculations, which depend on the study design, the operational research hypothesis, and the outcome measure. Special software programmes are available for dealing with sample size calculations.

References

Australian Bureau of Statistics (2010a) Australia's environment: issues and trends, January 2010; Population.
[http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/52DDB96A9F8B5A60CA2576C000193390?opendocument].

Australian Bureau of Statistics (2010b) Year Book Australia, 2009-2010. Primary and secondary education. [http://www.webcitation.org/5qvDapIMe] last accessed March 2010.

Bland, M. (2000) An introduction to medical statistics. Third edition; Oxford University Press, Oxford, UK.

Fraser, A.G., Scragg, R., Schaaf, D., Metcalf, P., Grant, C.C. (2010) Helicobacter pylori infection and iron deficiency in teenage females in New Zealand. N Z Med J; 123(1313): 38-45.

Gartrell, N., Herman, J., Olarte, S., Feldstein, M., Localio, R. (1986) Psychiatrist-patient sexual contact: results of a national survey. I: Prevalence. Am J Psychiatry; 143(9): 1126-31.

Gordis, L. (2008) Epidemiology. Fourth edition; Saunders, Elsevier Inc, Philadelphia, USA.

Hammond, E.C. and Horn, D. (1954) The relationship between human smoking habits and death rates: a follow-up study of 187,766 men. J Am Med Assoc; 155(15):1316-28.

Santosham, M., Chandran, A., Fitzwater, S., Fischer-Walker, C., Baqui, A.H., Black, R. (2010) Progress and barriers for the control of diarrhoeal disease. Lancet; 376(9734): 63-7.

"The Nobel Prize in Physiology or Medicine 2005" (2011) Nobelprize.org [http://nobelprize.org/nobel_prizes/medicine/laureates/2005/index.html] last accessed March 2011.