



NHS

*National Institute for
Health Research*

**The NIHR Research Design Service
for the East Midlands**

**The NIHR Research Design Service
for Yorkshire & the Humber**

Sampling and Sample Size Calculation

Authors

Nick Fox

Amanda Hunn

Nigel Mathers

This Resource Pack is one of a series produced by The NIHR RDS for the East Midlands / The NIHR RDS for Yorkshire and the Humber. This series has been funded by The NIHR RDS EM / YH.

This Resource Pack may be freely photocopied and distributed for the benefit of researchers. However it is the copyright of The NIHR RDS EM / YH and the authors and as such, no part of the content may be altered without the prior permission in writing, of the Copyright owner.

Reference as:

Fox N., Hunn A., and Mathers N. Sampling and sample size calculation
The NIHR RDS for the East Midlands / Yorkshire & the Humber 2007.

Nick Fox
School of Health and Related Research
(ScHARR)
University of Sheffield
Regent Court
30 Regent Street
Sheffield
S1 4DA

Amanda Hunn
Tribal Consulting, Tribal House,
7 Lakeside, Calder Island Way
Wakefield
WF2 7AW

Nigel Mathers
Academic Unit of Primary Medical
Care,
Community Sciences Centre,
University of Sheffield,
Northern General Hospital,
Herries Road,
Sheffield S5 7AU
United Kingdom

Last updated: May 2009

The NIHR RDS for the East Midlands www.rds-eastmidlands.nihr.ac.uk

Division of Primary Care,
14th Floor, Tower building
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: 0115 823 0500

Leicester: enquiries-LNR@rds-eastmidlands.org.uk

Nottingham: enquiries-NDL@rds-eastmidlands.org.uk

The NIHR RDS for Yorkshire & the Humber

www.rds-yh.nihr.ac.uk

ScHARR
The University of Sheffield
Regent Court
30 Regent Street
Sheffield
S1 4DA
Tel: 0114 222 0828

Sheffield: rds-yh@sheffield.ac.uk

Leeds: rds-yh@leeds.ac.uk

York: rds-yh@york.ac.uk

© Copyright of The NIHR RDS EM / YH
(2009)

Table of Contents

	Page
1. Introduction.....	4
2. The representative sample.....	5
3. Sample size and the power of research.....	12
4. Calculating sample size.....	17
5. Summary.....	30
6. Answers to exercises.....	31
7. Further reading and resources.....	35
8. Glossary.....	36

1. Introduction

Sampling and sample size are crucial issues in pieces of quantitative research, which seek to make statistically based generalisations from the study results to the wider world. To generalise in this way, it is essential that both the sampling method used and the sample size are appropriate, such that the results are representative, and that the statistics can discern associations or differences within the results of a study.

LEARNING OBJECTIVES

Having successfully completed this pack, you will be able to:

- distinguish between random and non-random methods of sample selection
- describe the advantages of random sample selection
- identify the different methods of random sample selection
- match the appropriate methods of sample selection to the research question and design
- realise the importance of estimating the optimal sample size, when designing a new study, and of seeking independent advice at this stage
- describe the factors influencing sample size
- make a preliminary estimate of the appropriate sample size.

Working through this pack

The study time involved in this pack is approximately 10 hours. In addition to the written text, the pack includes exercises for completion. We suggest that as you work through the pack, you establish for yourself a 'reflective log', linking the work in the pack to your own research interests and needs, and documenting your reflections on the ethnographic method. Include your responses to the exercises plus your own thoughts as you read and consider the material.

At all stages of your work, you may find the Glossary contained at the end of this resource pack to be of assistance.

2. The representative sample

It is an explicit or implicit objective of most studies in health care which 'count' something or other (quantitative studies), to offer conclusions that are *generalisable*. This means that the findings of a study apply to situations other than that of the cases in the study. To give a hypothetical example, Smith and Jones' (1997) study of consultation rates in primary care which was based on data from five practices in differing geographic settings (urban, suburban, rural) finds higher rates in the urban environment. When they wrote it up for publication, Smith and Jones used statistics to claim their findings could be generalised: the differences applied not just to these five practices, but to all practices in the country.

For such a claim to be legitimate (technically, for the study to possess 'external validity'), the authors must persuade us that their sample was not biased: that it was *representative*. Although other criteria must also be met (for instance, that the design was both appropriate and carried out correctly - the study's 'internal validity' and 'reliability'), it is the representativeness of a sample which allows the researcher to generalise the findings to the wider population. If a study has an unrepresentative or biased sample, then it may still have internal validity and reliability, but it will not be generalisable (will not possess *external validity*). Consequently the results of the study will be applicable only to the group under study.

It is essential to a study's design (assuming that study wants to generalise and is not simply descriptive of one setting) that sampling is taken seriously. The first part of this pack looks at how to gather a 'representative' sample which gives a study external validity and permits valid generalisation.

However, there is a second issue which must be addressed in relation to sampling, and this is predominantly a question of *sample size*. Generalisations from data to wider population depend upon a kind of statistic which tests *inferences* or *hypotheses*. For instance, the *t-test* can be used to test a hypothesis that there is a difference between two populations, based on a sample from each. To give an example, we select 100 males and 100 females and test their body mass index. We find a difference in our samples, and wish to argue that the difference found is not an accident (due to chance), but reflects an actual difference in the wider populations from which the samples were drawn. We use a *t-test* to see if we can make this claim legitimately.

Most people know that the larger a sample size, the more likely it is that a finding of a difference such as this is not due to chance, but really does mean there is a difference between men and women. Many quantitative studies undertaken and *published in medical journals* do not have a sufficient sample size to adequately test the hypothesis which the study was designed to explore. Such studies are, by themselves, of little use, and -- for example in the case of drug trials -- could be dangerous if their findings were generalised.

We will consider these issues of sample size, and how to calculate an adequate size for a study sample in the second half of this pack. Before that, let us think in greater detail about what a sample is.

2.1 Why do we need to select a sample anyway?

In some circumstances it is not necessary to select a sample. If the subjects of your study are very rare, for instance a disease occurring only once in 100 000 children, then you might decide to study every case you can find. More usually, however, you are likely to find yourself in a situation where the potential subjects of your study are much more common and you cannot practically include everybody. For example, a study of everybody in the UK who had been diagnosed as suffering from asthma would be impossible: it would take too long and cost too much money.

So it is necessary to find some way of reducing the number of subjects included in the study without biasing the findings in any way. Random sampling is one way of achieving this, and with appropriate statistics such a study can yield generalisable findings at far lower cost. Samples can also be taken using non-random techniques, but in this pack we will emphasise random sampling, which -- if conducted adequately -- will ensure external validity.

2.2 Random Sampling

To obtain a random (or probability) sample, the first step is to define the target population from which it is to be drawn. This population is known as the *sampling frame*, and can be thought of as a list of all the people / patients relevant to the study. For instance, you are interested in doing a study of children aged between two and ten years diagnosed within the last month as having *otitis media*. Or you want to study adults (aged 16-65 years) diagnosed as having asthma and receiving drug treatment for asthma in the last six months, and living in a defined geographical region. In each case, these limits define the sampling frame. If the research design is based on an experimental design, such as a randomised controlled trial (RCT), with two or more groups, then the population frame may often be very tightly defined with strict eligibility criteria.

Within an RCT, potential subjects are randomly allocated to either the intervention (treatment) group or the control group. By *randomly allocating subjects* to each of the groups, potential differences between the comparison groups should be much reduced. In this way *confounding variables* (i.e. variables you haven't thought of, or controlled for) *will be more equally distributed* between each of the groups and will be less likely to influence the outcome (or dependent variable) in either of the groups.

Randomisation within an experimental design is a way of ensuring control over confounding variables and as such it allows the researcher to have greater confidence in identifying real associations between an independent variable (a potential cause or predictor) and a dependent variable (the effect or outcome measure).

The term *random* may imply to you that it is possible to take some sort of haphazard or *ad hoc* approach, for example stopping the first 20 people you meet in the street for inclusion in your study. This is not random in the true sense of the word. To be a 'random' sample, every individual in the population must have an

equal probability of being selected. In order to carry out random sampling properly, strict procedures need to be adhered to.

Random sampling techniques can be split into *simple random sampling* and *systematic sampling*.

2.3 Simple Random Sampling

If selections are made purely by chance this is known as simple random sampling. So, for instance, if we had a population containing 5000 people, we could allocate every individual a different number. If we wanted to achieve a sample size of 200, we could achieve this by pulling 200 of the 5000 numbers out of a hat. This would be an example of simple random sampling - sometimes also called Independent Random Sampling because, as the probability of a person being selected is independent of the identity of the other people selected.

The usual method of obtaining random numbers is to use computer packages such as SPSS. Tables of random numbers may also be found in the appendices of most statistical textbooks.

Simple random sampling, although technically valid, is a very laborious way of carrying out sampling. A simpler and quicker way is to use systematic sampling.

2.4 Systematic Sampling

Systematic sampling is a more commonly employed method. After numbers are allocated to everybody in the population frame, the first individual is picked using a random number table and then subsequent subjects are selected using a fixed sampling interval, i.e. every *n*th person.

Assume, for example, that we wanted to carry out a survey of patients with asthma attending clinics in one city. There may be too many to interview everyone, so we want to select a representative sample. If there are 3,000 people attending the clinics in total and we only require a sample of 200, we would need to:

- calculate the sampling interval by dividing 3,000 by 200 to give a sampling fraction of 15
- select a random number between one and 15 using a set of random tables
- if this number were 13, we select the individual allocated number 13 and then go on to select every 15th person, i.e. numbers 28, then 43, then 58, and so on.

This will give us a total sample size of 200 as required.

Care needs to be taken when using a systematic sampling method in case there is some bias in the way that lists of individuals are compiled. For example, if all the husbands' names precede wives' names and the sampling interval is an even number, then we could end up selecting all women and no men.

2.5 Stratified Random Sampling

Stratified sampling is a way of ensuring that particular strata or categories of individuals are represented in the sampling process.

If, for example, we want to study consultation rates in a general practice, and we know that approximately four per cent of our population frame is made up of a particular ethnic minority group, there is a chance that with simple random sampling or systematic sampling we could end up with no ethnic minorities (or a much reduced proportion) in our sample. If we wanted to ensure that our sample was representative of the population frame, then we would employ a stratified sampling method.

- First we would split the population into the different strata, in this case, separating out those individuals with the relevant ethnic background.
- We would then apply random sampling techniques to each of the two ethnic groups separately, using the same sampling interval in each group.
- This would ensure that the final sampling frame was representative of the minority group we wanted to include, on a pro-rata basis with the actual population.

2.6 Disproportionate Sampling

If our objective were to compare the results of our minority group with the larger group, then we would have difficulty in doing so, using the proportionate stratified sampling just described. This is because the numbers achieved in the minority group, although pro-rata those of the population, may not be large enough to give a reasonable chance of demonstrating statistical differences (if such differences do in fact exist).

To compare the survey results of the minority individuals with those of the larger group, it is necessary to use a disproportion sampling method. With disproportionate sampling, the strata selected are not selected pro-rata to their size in the wider population. For instance, if we are interested in comparing the referral rates for particular minority groups with other larger groups, then it is necessary to over sample the smaller categories in order to achieve statistical power, that is, in order to be able to demonstrate statistically significant differences between groups if such differences exist.

(Note that, if subsequently we wish to refer to the total sample as a whole, representative of the wider population, then it will become necessary to re-weight the categories back into the proportions in which they are represented in reality. For example, if we wanted to compare the views and satisfaction levels of women who gave birth at home compared with the majority of women who have given birth in hospital, a systematic or random sample, although representative of all women giving birth would not produce a sufficient number of women giving birth at home to be able to compare the results, unless the total sample was so big that it would take many years to collate. We would also end up interviewing more women than we needed who have given birth in hospital. In this case it would be necessary to over-sample or over-represent those women giving birth at home to have enough individuals in each group in order to compare them. We would

therefore use disproportionate stratified random sampling to select the sample in this instance.)

The important thing to note here about disproportionate sampling is that sampling is still taking place *within* each stratum or category. So we would use systematic or simple random selection to select a sample from the 'majority' group and the same process to select samples from the minority groups.

2.7 Cluster (Multistage) Sampling

Cluster sampling is a method frequently employed in national surveys where it is uneconomic to carry out interviews with individuals scattered across the country. Cluster sampling allows individuals to be selected in geographic batches. For instance, before selecting individual people at random, the researcher may decide to focus on certain 'areas', e.g. towns, electoral wards or general practices - selecting these by a method of random sampling. Once this was done, they could either i) select all the individuals within these areas, or ii) use random sampling to select just a proportion of the individuals within these chosen areas.

Although cluster sampling is a very valuable technique and is widely used, it is worth noting that it does not produce strictly *independent* samples, since the knowledge that one person in a specific cluster has been selected will increase the probability that others in the same cluster will also be selected.

Obviously care must be taken to ensure that the cluster units selected are generally representative of the population and are not strongly biased in any way. If, for example, all the general practices selected for a study were single-handed, this would not be representative of all general practices.

2.8 Non-Random Sampling

Non-random (or non-probability) sampling is not used very often in quantitative medical social research surveys, but it is used increasingly in market research and commissioned studies. The technique most commonly used is known as quota sampling.

Quota Sampling

Quota sampling is a technique for sampling whereby the researcher decides in advance on certain key characteristics which s/he will use to stratify the sample. Interviewers are often set sample quotas in terms of age and sex. For example, consider a market research study where interviewers will stop people in the street to ask them a series of questions on consumer preferences. The interviewer might be asked to sample 200 people, of whom 100 should be male and 100 should be female - and, within each of these groups, there should be 25 people in each of the age-groups: under-20, 20-39, 40-59 and over-60. The difference with a stratified sample is that the *respondents in a quota sample are not randomly selected* within the strata. The respondents may be selected just because they are accessible to the interviewer. Because random sampling is not employed, it is not possible to apply inferential statistics and generalise the findings to a wider population.

EXERCISE 1

Read the descriptions below and decide what type of sample selection has taken place.

1. School children, some with asthma and some without, are identified from GP records.
Method: children are selected randomly within each of the two groups and the number of children in each group is representative of the total patient population for this age group.
2. Children with and without chronic asthma are identified from GP records.
Method: the children are selected so that in the sample exactly 50% have chronic asthma and 50% have no asthma. Within each of these groups, the children are randomly selected.
3. A survey of the attitudes of mothers with children under one year.
Method: interviewers stop likely looking women pushing prams in the street. The number of respondents who fall into different age bands and social classes is strictly controlled.
4. A survey of attitudes of drug users to rehabilitation services.
Method: drug users are recruited by advertising in the local newspaper for potential respondents.
5. A postal survey of the attitudes of males to use of male contraceptives.
Method: all male adults whose National Insurance numbers end in '5' are selected for a survey.
6. A study of the length of stay of patients at Anytown General Hospital.
Method: all patients admitted to wards 3, 5, and 10 in a hospital are selected for a study.

(Answers are at the end of the pack)

2.9 Sampling in Qualitative Research

Since the objective of qualitative research is to understand and give meaning to a social process, rather than quantify and generalise to a wider population, it is inappropriate to use random sampling or apply statistical tests. Sample sizes used in qualitative research are usually very small and the application of statistical tests would be neither appropriate nor feasible.

For details on this topic, please refer to the The NIHR RDS EM / YH resource pack "Qualitative Research" by Kate Windridge and Elizabeth Ockleford, 2007.

EXERCISE 2

This is an opportunity to review your learning on this first part of the pack. Read the extract from a journal article 'National asthma survey reveals continuing morbidity' given below.

National asthma survey reveals continuing morbidity (Prescriber, 19 March 1996 p.15)

A preliminary analysis of a survey of 44,177 people with asthma has revealed that for many the condition causes frequent symptoms and substantially interferes with daily life. There is also a trend for older people with asthma to experience more problems. More information about treatment was seen by many as the best way to improve care.

The Impact of Asthma Survey was conducted by Gallup on behalf of the National Asthma Campaign with funding from Allen & Hanburys. Questionnaires were given to people with asthma via surgeries, pharmacies, retail outlets, the media and direct mailing in the autumn of 1995; the respondents were therefore self-selected and may not be representative of the population with asthma.

Asthma symptoms were experienced on most days or daily by 41 per cent of survey respondents, ranging from 18 per cent of the under-11s to 55 per cent of pensioners. Waking every night with wheeze, cough or breathlessness was reported by 13 per cent and 43 per cent say they are woken by symptoms at least once a week.

About 20 per cent consider that asthma dominates their life, ranging from 17 per cent in children to 37 per cent in the over-60s; over 40 per cent of each age group say the condition has a moderate impact on their quality of life.'

Now answer the following questions:

1. How was the sample selected for this survey?
2. Did the researchers use random or non-random sampling methods?
3. What are the advantages of their approach?
4. What are the disadvantages of this approach?
5. The sample size was 44,177. Why was the sample size so large and was this necessary?

(Answers are at the end of the pack)

3. Sample size and the power of research

In the previous section, we looked at methods of sampling. Now we want to turn to another aspect of sampling: how big a sample needs to be in quantitative research to enable a study to have sufficient 'power' to do the job of testing a hypothesis. While this discussion will necessarily take us into the realm of statistics, we will keep the 'number-crunching' to a minimum: what is important is that you understand the concepts (and know a friendly statistician!).

At first glance, many pieces of research seem to choose a sample size merely on the basis of what 'looks' about right, or what similar studies have used in the past, or perhaps simply for reasons of convenience: ten seems a bit small, and one hundred would be difficult to obtain, so 40 is a happy compromise! Unfortunately a lot of published research uses precisely this kind of logic. In the following section, we want to show you why using such reasoning could make your research worthless. Choosing the correct size of sample is not a matter of preference, it is a crucial element of the research process without which you may well be spending months trying to investigate a problem with a tool which is either completely useless, or over expensive in terms of time and other resources.

3.1 The truth is out there: Hypotheses and samples

As we noted earlier, most (but not all) quantitative studies aim to test a *hypothesis*. A hypothesis is a kind of *truth claim* about some aspect of the world: for instance, the attitudes of patients or the prevalence of a disease in a population. Research sets out to try to prove this truth claim (or, more properly, to reject the null hypothesis - a truth claim phrased as a negative).

For example, let us think about the following hypothesis:

Levels of ill-health are affected by deprivation

and the related null hypothesis:

Levels of ill-health are not affected by deprivation

Let us imagine that we have this as our research hypothesis, and we are planning research to test it. We will undertake a trial, comparing groups of patients in a practice who are living in different socio-economic environments, to assess the extent of ill-health in these different groupings. Obviously the findings of a study -- while interesting in themselves -- only have value if they can be generalised, to discover something about the topic which can be applied in other practices. If we find an association, then we will want to do something to reduce ill-health (by reducing deprivation). So our study has to have *external validity*, that is, the capacity to be generalised beyond the subjects actually in the study.

The measurement of such generalisability of a study is done by statistical tests of inference. You may be familiar with some such tests: tests such as the *chi-squared test*,

the *t-test*, and tests of *correlation*. We will not look at these tests in any detail, but we need to understand that the purpose of these and other tests of *statistical inference* is to assess the extent to which the findings of a study can be accepted as valid for the population from which the study sample has been drawn. If the statistics we use suggest that the findings are 'true', then we can be happy to conclude (within certain limits of probability), that the study's findings can be generalised, and we can act on them (to improve nutrition among children under five years, for instance).

From common sense, we see that the larger the sample is, the easier it is to be satisfied that it is representative of the population from which it is drawn: but how large does it need to be? This is the question that we need to answer, and to do so, we need to think a little more about the possibilities that our findings may not reflect reality: that we have committed an error in our conclusions.

3.2 Type 1 and Type 2 errors

What any researcher wants is to be right! They want to discover that there is an association between two variables: say, asthma and traffic pollution, *but only if such an association really exists*. If there is no such association, then they want their study to support the null hypothesis that the two are not related. (While the former may be more exciting, both are important findings).

What no researcher wants is to be wrong! No-one wants to find an association which does not really exist, or - just as importantly - *not* find an association which *does* exist. Both such situations can arise in any piece of research. The first (finding an association which is not really there) is called a *Type I error*. It is the error of *falsely rejecting a true null hypothesis*.

(Think through this carefully. What we are talking about here could also be called a *false positive*. An example would be a study which rejects the null hypothesis that there is no association between ill-health and deprivation. The findings suggest such an association, but in reality, no such relationship exists.)

The second kind of error, called a *Type 2 error* (usually written as Type II), occurs when a study fails to find an association which really does exist. It is then a matter of *wrongly accepting a false null hypothesis*. (This is a *false negative*: using the ill-health and deprivation example again, we conduct a study and find no association, missing one which really does exist.)

Both types of error are serious. Both have consequences: imagine the money which might be spent on reducing traffic pollution, and all the time it does not really affect asthma (a Type I error). Or imagine allowing traffic pollution to continue, while it really is affecting children's health (a Type II error). Good research will minimise the chances of committing both Type I and Type II errors as far as possible, although they can never be ruled out absolutely.

3.3 Statistical Significance and Statistical Power

For any piece of research that tries to make inferences from a sample to a population there are four possible outcomes: two are desirable, two render the research worthless. Figure 1 shows these four possible outcomes diagrammatically.

		POPULATION	
		False	True
S T U D Y	False	Cell 1 Correct Result	Cell 2 Type I error (alpha)
	True	Cell 3 Type II Error (beta)	Cell 4 Correct Result

Figure 1: The Null Hypothesis (Ho), Statistical Significance and Statistical Power

Each cell in the figure represents a possible relationship between the findings of the study and the 'real-life' situation in the population under investigation. (Of course, we cannot actually know the latter unless we surveyed the whole population: that is the reason we conduct studies which can be generalised through statistical inference). Cells 1 and 4 represent desirable outcomes, while cells 2 and 3 represent potential outcomes of a study which are undesirable and need to be minimised. We shall now consider the relationship between these possible outcomes, and two concepts, that of *statistical significance* and of *statistical power*. The former is well-known by most researchers who use statistics, the latter is less well understood. Let us look at these four outcomes, in relation to the study of ill-health and deprivation given as an example above.

Cell 1. The null hypothesis has been rejected by the results of the study, and there is support for a hypothesis which suggests an association between ill-health and deprivation. In 'reality' such an association does exist in the population. In this outcome, the study *is* reflecting the world outside the limits of the study and it is a 'correct' result (that is, the result is both statistically significant *and* real).

Cell 4. The results from the study support the null hypothesis: there is no association between ill-health and deprivation, and this is also the situation which pertains in the population - so once again in such circumstances the study reflects 'reality'.

Cell 2. In this cell, as in cell 1, the study results reject the null hypothesis, indicating some kind of association between the variables of deprivation and health. However, these study results are false, because in the population from which we drew our sample the null hypothesis is actually true and there is no such association. This is the Type I error: the error of rejecting a true null hypothesis. The likelihood of committing a Type I error (finding an association which does not really exist) is known as the alpha (α) value or the statistical significance of a statistical test. Some of you may be familiar with α as p , the quoted level of significance of a test. The p value marks the probability of committing a Type I error; thus a p value of 0.05 (a widely used conventional level of significance) indicates a five per cent -- or

one in 20 -- chance of committing a Type I error. Cell 2 thus reflects an incorrect finding from a study, and the α value represents the probability of this occurring.

Cell 3. This cell similarly reflects an undesirable outcome of a study. Here, as in Cell 4, a study supports the null hypothesis, implying that there is no association between ill-health and deprivation in the population under investigation. But in reality, the null hypothesis is false and there is an association in the real world which the study does not find. This mistake is the *Type II error of accepting a false null hypothesis*. and is the result of having a sample size which is too small to allow detection of the association by statistical tests at an acceptable level of significance (say $p = 0.05$). The likelihood of committing a Type II error is the *beta* (β) value of a statistical test, and the value $(1 - \beta)$ is the *statistical power* of the test. Thus, the *statistical power of a test is the likelihood of avoiding a Type II error* i.e. the probability that the test will reject the null hypothesis when the null hypothesis is false. Conventionally, a value of 0.80 or 80% is the target value for statistical power, representing a likelihood that four times out of five a study will reject a false null hypothesis, although values greater than 80% e.g. 90% are also sometimes used. Outcomes of studies which fall into cell 3 are incorrect; β or its complement $(1-\beta)$ are the measures of the likelihood of such an outcome of a study.

All research should seek to avoid both Type I and Type II errors, which lead to incorrect inferences about the world beyond the study. In practice, there is a trade-off. Reducing the likelihood of committing a Type I error by increasing the level of significance at which one is willing to accept a positive finding reduces the statistical power of the test, thus increasing the possibility of a Type II error (missing an association which exists). Conversely, if a researcher makes it a priority to avoid committing a Type II error, it becomes more likely that a Type I error will occur (finding an association which does not exist). Now spend a few minutes doing this exercise to help you think about Type I and Type II errors in research.



EXERCISE 3

Risk, Type I and Type II Errors

If we knew everything about the world, we would not need to do research. But we don't know everything, and research projects try to find out something more. With limited resources, we have to accept that sometimes (despite all efforts to conduct good research) our findings will be wrong. Use your judgement to decide in each of the four following pieces of research which poses the greater risk: a Type I or a Type II error, and why.

Research Study 1 A randomised controlled trial of a proven but expensive drug and an unproven cheap drug to treat HIV infection, to see if there is a difference in efficacy in controlling the disease.

Research Study 2. A study to test whether arrhythmias are more likely in patients taking a new anti-histamine prescribed for hayfever, compared with those already in use.

Research Study 3. A study to investigate the effect of training ambulance staff in defibrillator use on reducing numbers of 'dead-on-arrivals' after road traffic accidents.

Research Study 4. A survey of causes of deaths among white and ethnic minorities in the USA.

(Answers are at the end of the pack)

4. Calculating sample size

In the rest of this pack, we will work through examples of the calculations needed to determine an appropriate sample size. First, we will look at descriptive studies (which do not test a hypothesis). Then we will consider issues of statistical significance and power in inferential, i.e. hypothesis-testing studies.

We will see that the formulae we need to use are relatively simple, and easy to calculate using a pocket calculator or computer software - but the choice of the numbers to put into the formulae is often not so straightforward, and the choices will often need to be justified, to the potential funding organisation and/or ethics committee who will be assessing the research proposal. We will also see that the consequences of getting these numbers wrong, especially by under-estimating the required sample size - can be very serious indeed - and have in the past often resulted in 'hopeless' studies being carried out, which had no realistic chance of detecting the treatment effects or risk factors for which they were designed!

For these reasons, we strongly advise getting independent advice on sample size when you are designing your study - and this will usually be from either a trained statistician or from a researcher in your field who has longstanding experience of study design. We hope that you will be able to use the material in this resource pack for carrying out a preliminary sample size calculation, and discussing these issues with the advisor.

4.1 Sample Size in Descriptive Studies

Not all quantitative studies involve hypothesis-testing, some studies merely seek to describe the phenomena under examination. Whereas hypothesis testing will involve comparing the characteristics of two or more groups, a descriptive survey may be concerned solely with describing the characteristics of a single group. The aim of this type of survey is often to obtain an accurate estimate of a particular figure, such as a mean or a proportion. For example, we may want to know how many times, in an average week, that a general practitioner sees patients newly presenting with asthma. In addition we may also want to know what proportion of these patients admit to smoking five or more cigarettes a day. In these circumstances, the aim is not to compare this figure with another group, but rather, to accurately reflect the real figure in the wider population.

To calculate the required sample size in this situation, there are certain things that we need to establish. We need to know:

1. The level of confidence we require concerning the true value of a mean or proportion. This is closely connected with the level of significance for statistical tests, such as a t-test. For example, we can be '95% confident' that the true mean value lies somewhere within a valid 95% confidence interval, and this corresponds to significance testing at the 5% level ($P < 0.05$) of significance. Likewise, we can be '99% confident' that the true mean value lies somewhere within a valid 99% confidence interval (which is a bit wider), and this corresponds to significance testing at the 1% level ($P < 0.01$) of significance.

2. The *degree of precision* which we can accept. This is often presented in the form of a *confidence interval*. For example, a survey of a sample of patients indicates that 35 per cent smoke. Are we willing to accept that the figure for the wider population lies between 25 and 45 per cent, (allowing a margin for random error (MRE) of 10% either way), or do we want to be more precise, such that the confidence interval is three per cent each way, and the true figure falls between 32 and 38 per cent? As we can see from the following table, the smaller the allowed margin for random error, the larger the sample must be.

Margin for random error	Sample size
+ or - 10%	88
+ or - 5%	350
+ or - 3%	971
+ or - 2%	2188
+ or - 1%	8750

Table 1: Precision (margin for random error) and necessary sample sizes for a population with 35 per cent smokers

In the following pages, we will look at how to calculate sample sizes for mean averages (for example, mean birth-weights) to supply different levels of precision. The confidence interval will depend upon the distribution of values in the sample: the more variability (as measured by the *standard deviation*) in the population, the greater the sample will need to be, in order to supply a given confidence interval indicating an acceptable degree of precision.

We also need to bear in mind the *likely response rate*. Allowance needs to be made for non-responses to a survey, so that this can be added on to the required sample size. For example, if our calculations indicate that we need a minimum sample size of 200, but we only expect a 70% response rate, then we will need to select an initial sample size of $286 = 200 / 0.7$ in order to allow for possible non-response. It is particularly important to make an allowance for non-response when planning a longitudinal survey, when the same individuals will be repeatedly contacted over a period of time, since cumulative non-response can result in the final wave of the survey being too small to analyse.

Worked Example 1: How large must a sample be to estimate the mean value of the population?

Suppose we wish to measure the number of times that the average patient with asthma consults her/his general practitioner for treatment?

a) First, the SE (standard error) is calculated by deciding upon the accuracy level which you require. If, for instance, you wish your survey to produce a very accurate answer with only a small confidence interval, then you might decide that you want to be 95% confident that the mean average figure produced by your survey is no more than *plus or minus two visits to the GP*.

For example, if you thought that your survey might produce a mean estimate of 12.5 visits per year, then your confidence interval in this case would be $12.5 \pm$ two visits. Your

confidence interval would then tell you that you could reasonably (more detail on what 'reasonably' means below!) expect the true average rate of visits in the population to be somewhere between 10.5 and 14.5 visits per year.

Now decide on your required significance level. If you decide on 95%, (meaning that 19 times out of 20 the true population mean falls within the confidence limit of 10.5 and 14.5 visits), the standard error is calculated by dividing the MRE by 1.96. So, in this case, the standard error is 2 divided by 1.96 = 1.02.

If you want a 95% confidence interval, then divide the maximum acceptable MRE (margin for random error) by 1.96 to calculate the SE.

If instead you want a 99% confidence interval, then divide the maximum acceptable MRE by 2.56 to calculate the SE.

b) The formula to calculate the sample size for a mean (or point) estimate is:

$$N = \left(\frac{SD}{SE} \right)^2$$

where N = the required sample size,
 SD = the standard deviation, and
 SE = the standard error of the mean

The standard deviation could be estimated either by looking at some previous study or by carrying out a pilot study. Suppose that previous data showed that the standard deviation of the number of visits made to a GP in a year was 10, then we would input this into the formula as follows:

$$N = \left(\frac{SD}{SE} \right)^2 = \left(\frac{10}{1.02} \right)^2 = 9.8^2 = 96.12 = 97 \text{ (rounded to nearest patient)}$$

If we are to be 95% confident that the answer achieved is correct \pm two visits, then the required sample is 97 - before making allowance for a proportion of the people leaving the study early and failing to provide outcome data.

Worked Example 2: How large must a sample be to estimate a proportion / percentage?

Suppose that we were interested in finding out what percentage of the local patient population were satisfied with the service they had received from their GP over the previous 12 months. We want to carry out a survey, but of how many people?

Once again we need to know:

- The confidence level (usually 95% or 99%)

- The Confidence Interval we are willing to accept, for example that our survey finding lies within plus or minus five per cent of the population figure.

Assume that we decide that the precision with which we decide the proportion of respondents who say that they are satisfied with the service must be plus or minus 5%. This then is our confidence interval. To calculate the standard error, we divide the confidence interval by 1.96. In this case the standard error is $5/1.96 = 2.55$.

We also need to estimate the proportion which we expect to find who are satisfied. In order to estimate P (the estimated percentage) we should consult previous surveys or conduct a pilot. Assume, for the time being, that a similar survey carried out three years ago indicated that 70% of the respondents said they were satisfied. We then use the following formula:

$$N = \frac{P (100\% - P)}{(SE)^2}$$

With P = 70% and SE=2.55, we have:

$$N = \frac{70\% (100\% - 70\%)}{(2.55\%)^2} = \frac{2100}{6.50} = 323.08 = 324 \text{ (rounded upwards)}$$

So, in order to be 95% confident that the true proportion of people saying they are satisfied lies within $\pm 5\%$ of the answer, we will require a sample size of 324. This assumes that the likely answer is around 70% with a range between 65% and 75%.

Of course, in real life, we often have absolutely no idea what the likely proportion is going to be. There may be no previous data and no time to carry out a pilot. In these circumstances, it is safer to assume the worst case scenario *and assume that the proportion is likely to be 50%*. Other things being equal, this will allow for the largest possible sample size - and in most circumstances it is preferable to have a slight overestimate of the number of people needed, rather than an underestimate.

(If we wished to use a 99% level of significance, so we might be 99% confident that our confidence parameters include the true figure, then we need to divide the confidence interval by 2.56. In this case, the standard error would be $5/2.56 = 1.94$. Using the formula above, we find that this would require a sample size of 558.)

EXERCISE 4

Calculating Sample Sizes for Descriptive Studies:

1. You want to conduct a survey of the average age of GPs in the UK.

You want to calculate the 95% confidence interval for the average age of the GPs

Your acceptable margin for random error is plus or minus 3 years

From previous work you estimate that the standard deviation of the GPs' ages is 13 years

a) Calculate the SE, using the formula

$$SE = \frac{MRE}{1.96} = \dots\dots\dots$$

b) Using the formula for the sample size for a mean estimate, calculate

$$N = \left(\frac{SD}{SE} \right)^2 = \dots\dots\dots$$

c) What would the sample size need to be if the response rate to the survey is 70 per cent?

$$N = \dots\dots\dots$$

2. You want to conduct a survey of the proportion of men over 65 who have cardiac symptoms

Your significance level is 95%

Your acceptable margin for random error is plus or minus 2 per cent

From previous work you estimate that the proportion is about 20 per cent

a) Calculate the SE =

b) Using the formula for the sample sizes for a proportion, calculate:

$$N = \frac{P(100\% - P)}{(SE)^2} = \dots\dots\dots$$

c) What would the sample size need to be if the response rate to the survey is 80 per cent?

$$N = \dots\dots\dots$$

(Answers can be found at the end of the pack)

4.2 Sample Size in Inferential Studies

As we saw earlier in this pack, studies which test hypotheses (seeking to generalise from a study to a population), need sufficient power to minimise the likelihood of Type I and Type II errors. Both statistical significance and statistical power are affected by sample size. The chances of gaining a statistically significant result will be increased by enlarging a study's

sample. Put another way, the statistical power of a study is enhanced as sample size increases. Let us look at each of these aspects of inferential research in turn. You may wish to refer back to Figure 1, on page 11.

The Statistical Significance of a Study

When a researcher uses a statistical test, what they are doing is testing their results against a gold standard. If the test gives a positive result (this is usually known as 'achieving statistical significance'), then they can be relatively satisfied that their results are 'true', and that the real world situation is that discovered in the study (Cell 1 in Fig 1). If the test does not give significant results (non-significant or NS), then they can be reasonably satisfied that the results reflect Cell 4, where they have found no association and no such association exists.

However, we can never be absolutely certain that we have a result which falls in Cells 1 or 4. Statistical significance represents the likelihood of committing a Type I error (Cell 2). Let us imagine that we have results suggesting an association between ill-health and deprivation, and a t-test (a test to compare the results of two different groups) gives a value which indicates that at the 5% or 0.05 level of statistical significance, there is more ill-health among a group of high scorers on the Jarman Index of deprivation than among a group of low scorers.

What this means is that 95 per cent of the time, we can be certain that this result reflects a true effect (Cell 1). Five per cent of the time, it is a chance result, resulting from random associations in the sample we chose. If the t-test value is higher, we might reach 1% or 0.01 significance. Now, the result will only be a chance association one per cent of the time .

Tests of statistical significance are designed to account for sample size, thus the larger a sample; the 'easier' it is for results to reach significance. A study which compares two groups of 10 patients will have to demonstrate a much greater difference between the groups than a study with 1000 patients in each group. This is fair: the larger study is much more likely to be 'representative' of a population than the smaller one. To summarise: statistical significance is a measure of the likelihood that positive results reflect a real effect, and that the findings can be used to make conclusions about differences which really exist.

The Statistical Power of a Study

Because of the way statistical tests are designed, as we have just seen, they build in a safety margin to avoid generalising false positive results which could have disastrous or expensive consequences. But researchers who use small samples also run the risk of not being able to demonstrate differences or associations which really do exist. Thus they are in danger of committing a *Type II error* (Cell 3 in Fig 1), of accepting a false null hypothesis. Such studies are 'under-powered', not possessing sufficient statistical power to detect the effects they set out to detect. Conventionally, *the target is a power of 80% or 0.8*, meaning that a study has an 80 per cent likelihood of detecting a difference or association which really exists.

Examination of research undertaken in various fields of study suggests that many studies do not meet this 0.8 conventional target for power (Fox and Mathers 1997). What this means is that many studies have a much reduced likelihood of being able to discern the effects which they set out to seek: a study, with a power of 0.66 for some

specified treatment effect, will only detect that effect (if true) two times out of three. *A non-significant finding of a study may thus simply reflect the inadequate power of the study to detect differences or associations at levels which are conventionally accepted as statistically significant.*

When a study has only small (say less than 50%) power to detect a useful result, one must ask the simple question of such research: **‘Why did you bother, when your study had little chance of finding what you set out to find?’**

Sample size calculations need to be undertaken prior to a study to avoid both the wasteful consequences of under-powering, (or of *overpowering* in which sample sizes are excessively large, with higher than necessary study costs and, perhaps, the needless involvement of too many patients, which has ethical implications.).

Statistical power calculations are also sometimes undertaken after a study has been completed, to assess the likelihood of a study having discovered effects.

Statistical power is a function of three variables: sample size, the chosen level of statistical significance (α) and effect size. While calculation of power entails recourse to tables of values for these variables, the calculation is relatively straightforward in most cases.

Effect Size and Sample Size

As was mentioned earlier, there is a trade-off between significance and power, because as one tries to reduce the chances of generating false negative results, the likelihood of a false positive result increases. Researchers need to decide which is more crucial, and set the significance level accordingly. In Exercise 3 you were asked to decide, in various situations, whether a Type I or Type II error was more serious - based on clinical and other criteria.

Fortunately both statistical significance and power are increased by increasing sample size, so increasing sample size will reduce likelihoods of both Type I and Type II errors. However, that does not mean that researchers necessarily need to vastly increase the size of their samples, at great expense of time and resources.

The other factor affecting the power of a study is the *effect size* (ES) which is under investigation in the study. This is a measure of ‘how wrong the null hypothesis is’. For example, we might compare the efficacy of two bronchodilators for treating an asthma attack. The ES is the difference in efficacy between the two drugs. An effect size may be a difference between groups or the strength of an association between variables such as ill-health and deprivation.

If an ES is small, then many studies with small sample sizes are likely to be under-powered. But if an ES is large, then a relatively small scale study could have sufficient power to identify the effect under investigation. It is sometimes possible to increase the effect size (for example, by making more extreme comparisons, or undertaking a longer or more powerful intervention), but usually this is the intractable element in the equation, and accurate estimation of the effect size is essential for calculating power before a study begins, and hence the necessary sample size.

An Effect Size can be estimated in two ways:

Preferably:

- One can make a *decision about the smallest size of effect which it is worth identifying*. To consider the example of two rival drugs, if we are willing to accept the two drugs as equivalent if there is no more than a ten per cent difference in their efficacy of treatment, then this effect size may be set, acknowledging that smaller effects will not be discernible.

Alternatively:

- From a *review of literature or meta-analysis*, which can suggest the size of ES which may be expected.

4.3 A case study of statistical power: primary care research

Power calculations may be used as part of the critical appraisal of research papers. Unfortunately it is rare to see values for statistical power quoted for tests in research reports, and indeed often the results reported are inadequate to calculate effect sizes. Appraisals of various scientific subjects including nursing, education, management and general practice research have been undertaken by various authors.

Now read the edited extract from an article by Nick Fox and Nigel Mathers (1997).

Empowering your research: statistical power in general practice research Family Practice 14 (4) 1997

To explore the power of general practice research, we analysed all the statistical tests reported in the British Journal of General Practice (BJGP) over a period of 18 months. Power was calculated for each test based on the reported sample size. This enabled calculation of the power of each quantitative study published during this period, to assess the adequacy of sample sizes to supply sufficient power.

Method

All original research papers published in the BJGP during the period January 1994 to June 1995 inclusive were analysed in terms of the power of statistical tests reported. Qualitative papers were excluded, as were meta-analyses and articles which, although reporting quantitative data, did not report any formal statistical analysis even though in some instances such tests could have been undertaken. A further six papers were excluded because they did not use standard statistical tests for which power tables were available. This left 85 papers, involving 1422 tests for which power could be calculated using power tables. Power was calculated for each test following conventions of similar research into statistical power. Where adequate data was available (for example details of group means and standard deviations, or chi-squared test results) precise effect sizes could be calculated. Where this was not possible (in particular for results simply reported as 'non-significant') the following assumptions were made, all of which considerably *over-estimate* the power of the test:

- a) For significant results, the effect size was assumed to be 'medium', which as noted earlier means an effect 'visible to the naked eye'. Non-significant results were assumed to have a 'small' ES.
- b) Alpha values were set at the lowest possible conventional level of 0.05, and where a directional test was used, a one-tailed alpha was used (equivalent to two-tailed *alpha* of 0.1).

From the calculations of power for individual tests, a mean power for each paper was derived. This strategy has been adopted in other research into statistical power: what is reported is *study* power, rather than test-by-test power, and offers an estimate of the quality of studies in terms of overall adequacy of statistical power.

Results

Eighty-five papers comprising 1422 tests were analysed. The median number of tests per paper was 12, with a minimum of one test and a maximum of 90. The median power of the 85 studies was 0.71, representing a slightly greater than two-thirds probability of rejecting null hypotheses. The proportions of tests in different power bands is summarised in Table A. Of the 85 studies, 37 (44%) had power of at least 0.8, while 48 (56%) fell below this conventional target. The lowest power rating was 0.24, while 10 studies (12%) reached power values of 0.99 or more.

Power Band	N	%
< 0.25	2	2
0.26 - 0.49	19	22
0.50 - 0.79	27	32
0.80 - 0.96	21	25
≥ 0.97	16	19

Table A. Power of Studies (N = 85)

Discussion

The results of this survey of general practice research published in the BJGP indicates somewhat higher power ratings than those reported for other disciplines including nursing, psychology, education, management and some medical journals. However over half of the studies fall below the conventional figure of 0.8, and 25% have a power of 0.5 or less, suggesting a chance of gaining significant results poorer than that obtained by tossing a coin.

Scrutiny of the distribution of powers indicated bimodality. Of the papers meeting or exceeding the 0.8 target, 16 out of 37 had powers of more than 97%. Such high powers were achieved by the use of very large samples. Given that it is necessary to double the sample size to increase power from 0.8 to 0.97, it is reasonable to argue that as such the studies were *overpowered*, using sample sizes which were excessively expensive in terms of researcher time for data collection and analysis. In some cases these studies used pre-existing data sets and so this criticism is less pertinent; elsewhere, researchers

may have devoted far greater efforts in terms of time and obtaining goodwill from subjects than may strictly have been necessary to achieve adequate power. The importance of pre-study calculations of necessary sample size to achieve statistical power of 0.8 or thereabouts is relevant both for those studies demonstrated to be under-powered and those for whom power is excessive.

Conclusions

More than half of the quantitative papers published in the BJGP between January 1994 to June 1995 were 'under-powered'. This means that during the statistical analysis, there was a substantial risk of missing significant results. Twenty five percent of papers surveyed had a chance of gaining significant results (when there was a false null hypothesis) poorer than that obtained from tossing a coin.

EXERCISE 5

What was the median power of research in the papers surveyed?

What proportion of papers had too high a power - of, say, at least 99% - and why is this an issue?

(Answers can be found at the end of the pack)

Worked Example: Calculating the sample size in inferential studies

We will now work through two examples of sample size calculations. These provide formulae for calculating power for chi-squared and t-tests, although when calculating sample sizes most people will refer to tables or use computer software (details of books, a *British Medical Journal* article with some tables included, and software are given at the end of this pack).

Worked Example: Sample Size for Tests of Contingency

Imagine a doctor wanted to set up a double-blind trial of a new drug, to compare mortality after a stroke among patients using the new drug or a placebo.

- Measure: death from any cause within one year of first treatment
- Analysis: comparison of proportion of deaths amongst new drug and placebo patients, using chi-squared at $\alpha = 5$ per cent significance
- Standard treatment: 90 per cent expected to survive at least one year on placebo
- Power required: if the new drug can halve the mortality (reduce deaths from 10 to 5 per cent), this should be detected 90 per cent of time (power = 0.9, $\beta = 0.1$)

In summary:

p_1 = proportion of successes on standard treatment = 90%

p_2 = proportion of successes on the new drug which indicate it as more effective = 95%

$$\alpha = 0.05$$

$$\beta = 0.1$$

K = constant which is a function of α and β (see Table 2)

The sample size for each of the two groups, N is given by

$$N = K \times \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2}$$

Look up the value of K for $\alpha = 0.05$ and $\beta = 0.1$ in Table 2

Table 2: Values of K, as used for sample size calculations

		Power:			
		50%	80%	90%	95%
		b = 0.5	b = 0.2	b = 0.1	b = 0.05
a:	0.10	2.7	6.2	8.6	10.8
	0.05	3.8	7.9	10.5	13.0
	0.02	5.4	10.0	13.0	15.8
	0.01	6.6	11.7	14.9	17.8

$$\text{Thus, } N = 10.5 \times \frac{(0.9 \times 0.1) + (0.95 \times 0.05)}{(0.95 - 0.90)^2} = 10.5 \times \frac{0.090 + 0.048}{(0.05)^2}$$

and we have N = 580 patients in each group (Total = 1160), before allowance is made for non-response.

Note that here the difference between p_1 (success rate of placebo) and p_2 (success rate of the new drug) was very small. In other words, it was a very small effect size. Also the power required here was high (90%). If the effect size was larger or the power required was lower, then the sample size would be substantially smaller.

Worked Example: Sample Size for Test of Differences (t-test)

A clinical trial tests the preventive effect upon neonatal hypocalcemia of giving Supplement A to pregnant women. Women are randomised and given either placebo or Supplement A.

- Measure: serum calcium level of baby one week postnatally
- Analysis: Comparisons of difference between two groups of babies using an independent-samples t-test at 5% significance ($\alpha = 0.05$)
- Serum calcium in babies of untreated women 9.0 mg/100 ml, standard deviation (σ) 1.8mg/100ml
- Study should detect clinically relevant increase in serum calcium of 0.5 mg/100ml, 80 per cent of the time ($\beta = 0.2$)

In summary:

$$\mu = \text{Mean serum calcium level} = 9.0 \text{ mg/100ml}$$

$$\sigma = \text{Standard Deviation} = 1.8 \text{ mg/100ml}$$

$$d = \text{difference in means } \mu_1 - \mu_2 = 0.5 \text{ mg/100ml}$$

$$\alpha = 0.05$$

$$\beta = 0.2$$

The number of patients required in each group is given by

$$N = 2 \times K \times \left(\frac{s}{m_1 - m_2} \right)^2$$

where K is taken from Table 2

$$\text{So, } N = 2 \times 7.9 \times \left(\frac{1.8}{0.5} \right)^2 = 205$$

EXERCISE 6

Calculating Sample Size for Inferential Statistics.

1. A randomised controlled trial is carried out to investigate whether aspirin can prevent pregnancy-induced hypertension and pre-eclamptic toxemia in women at high risk (Schiff et al 1989)
 - Measure: Trial and placebo group: develop or did not develop hypertension
 - Analysis: Chi-squared test at 5% significance
 - Current situation: 30 per cent of women develop hypertension
 - Power required: clinically useful reduction by one third to 20 per cent should be detected with 80% ($\beta = 0.2$) power.

Calculate the necessary sample size.

2. A double-blind placebo-controlled trial is designed to test the effect of adding salmeterol to current treatment with inhaled corticosteroids in asthma sufferers who control their dosage according to a management plan. (Wilding al 1997)
 - Measure: dosage of corticosteroids after 6 months of trial
 - Analysis: comparison of differences in dosage between test and placebo group using t-test at $\alpha = 5\%$ significance.

- Current level: mean dosage of corticosteroids 700 micrograms (standard deviation $\sigma = 200$ micrograms)
- Power required to detect clinically relevant fall in dosage of 100 micrograms is 80 per cent ($\beta = 0.2$)

Calculate the number of subjects required in each treatment.

(Answers can be found at the end of the pack)

5. Summary

Key points to remember when deciding on sample selection are:

- Try to use a random method where possible and remember that random does not mean 'haphazard' or 'arbitrary'.
- Random selection means that everybody in your sampling frame has an equal opportunity of being included in your study.
- If you need to be able to generalise about small or minority groups and to compare those with larger groups, consider using disproportionate stratified sampling, but remember to re-weight the results afterwards if you wish to generalise to the whole population.

Key points to remember when deciding on sample size are:

- We strongly recommend that researchers obtain independent advice on sample size when designing their study - and this will usually be from either a trained statistician or from a researcher in your field who has longstanding experience of study design.
This is because sample size estimation is such a crucial aspect of the design of a quantitative study, with important ethical as well as cost implications - and often very little can be done to 'salvage' the results from an insufficiently large sample.
- There is a trade off between committing a Type I error (false positive) and a Type II error (false negative), but historically science has placed the emphasis on avoiding Type I errors.
- Other things being equal, increasing the sample size increases the sensitivity of the study to detect a difference between the groups being compared - and enables both α and β to be set at lower levels, and so will help reduce both Type I and Type II errors, but remember that it is costly and unethical to have too large a sample size.
- To calculate statistical power, you need to estimate the effect size.
- To estimate the sample size for a descriptive study in order to estimate a mean or a proportion, it is necessary to specify the maximum acceptable margin for random error.

6. Answers to exercises

Exercise 1

1. Stratified random sample. The sample is stratified because the sample has been selected to ensure that two different groups are represented.
2. Disproportionate stratified random sample. This sample is stratified to ensure that equal numbers of children from the two groups are selected - even though, in the population, only a relatively small proportion of children suffer from chronic asthma. The total sample will therefore *not* be representative of the children in the population.
3. Quota. The sample is not randomly selected but the respondents are selected to meet certain criteria.
4. Convenience. The sample is not randomly selected and no quotas are applied.
5. Systematic random sample.
6. Cluster sample. The patients are selected only from certain wards - and whether this is a randomised or a convenience cluster sample depends on how these wards had been chosen.

Exercise 2

1. The researchers used a convenience sampling approach, i.e. they selected people on the basis that they were easy to access. Respondents were therefore self-selected.
2. The sampling method used was non-random.
3. The advantages of this approach were that they were able to obtain the views of a large number of people very quickly and easily with little expense.
4. Unfortunately the convenience sample approach means that the sample is not representative of the population of individuals with asthma. Because a large part of the survey is made up of people attending in surgery and pharmacies, the sample will tend to over-represent those individuals requiring the most treatment. It will also over-represent those individuals who are most interested in expressing their opinions.
5. The sample achieved was very large because it was self-selected, and therefore the researchers would have had little control over how many people participated.

The sample is unnecessarily large. In order to achieve a statistically representative view of the asthmatic population, it would not be necessary to select such a large sample. This study demonstrates the point that large samples alone do not necessarily mean that the study can achieve representativeness. The only true way of achieving a representative sample is to use random sampling methods. Reflect on the sample size in this study as you now go on to study the second part of this pack.

Exercise 3

1. Type II error of accepting a false null hypothesis. If the study shows no difference in efficacy, missing a difference which is present, an effective but expensive drug may be dropped because of its cost, making treatment of patients less effective.
2. Type I error of rejecting a true null hypothesis. If an effect (increased levels of arrhythmias) is found this may lead to a useful drug being abandoned.
3. Your answer will depend on your reasoning. You might suggest a Type I error is more serious: if the training actually makes no difference (a true null hypothesis) but a study shows it does, then the findings may lead to innovating a procedure which is expensive and has risks associated with it. Alternatively, you might say a Type II error is more serious: a study failed to discover a real reduction (a false null hypothesis), so a useful procedure is not implemented and lives are lost. Your value perspective will affect which you see as more risky or costly (in an economic or humanistic sense).
4. Neither. A survey does not test a hypothesis. However, if a direct comparison were being made, the answer would probably be that a Type II error was more serious. Race is a very sensitive issue in the US. Missing a difference that existed (and was later discovered) could be both unjust and socially and politically catastrophic.

Exercise 4

1. The sample size needs to be 72 GPs.

We first calculate the SE by dividing the confidence interval by 1.96.

$$SE = 3 / 1.96 = 1.53$$

We then calculate:

$$N = \left(\frac{SD}{SE} \right)^2$$

In this case SD is 13, so that

$$N = \left(\frac{13}{1.53} \right)^2 = (8.49)^2 = 72.2$$

So, $N = 73$ (rounding up, for safety)

If the expected response rate is just 70 per cent, then to have 73 subjects completing the study, you will need to recruit $73 / 0.7 = 104.3 = 105$ subjects (rounding up).

2. The sample size needs to be 1539.

First, the SE can be calculated by dividing the confidence interval by 1.96:

$$SE = \frac{2}{1.96} = 1.02$$

We then calculate:

$$N = \frac{P(100\% - P)}{(SE)^2}$$

With $P = 20\%$ and $SE = 1.02$,

$$\text{we have: } N = \frac{20 \cdot (100 - 20)}{(1.02)^2} = \frac{20 \times 80}{1.04} = 1539 \text{ (rounding up)}$$

If the response rate is 80 per cent, you will need to recruit $1539 / 0.8 = 1924$ subjects.

Exercise 5

1. The median power was 0.71 or 71 per cent.
2. Twelve per cent. It is important because these studies are thus more costly than necessary to achieve acceptable power.

Exercise 6

Question 1.

We have:

$p_1 =$ proportion developing hypertension on placebo = 0.3

$p_2 =$ proportion developing hypertension on placebo which we wish to detect = 0.2

$\alpha = 0.5$

$\beta = 0.2$

$$N = K \times \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2}$$

$$N = 7.9 \times \frac{(0.3 \times 0.7) + (0.2 \times 0.8)}{(0.3 - 0.2)^2}$$

N = 293 patients in each group, before making allowance for non-response.

Question 2.

We have:

Difference $\mu_1 - \mu_2 = 100$ micrograms

Standard Deviation $\sigma = 200$ micrograms

$$\alpha = 0.5$$

$$\beta = 0.8$$

$$N = 2 \times K \times \left(\frac{s}{m_1 - m_2} \right)^2$$

$$N = 2 \times 7.9 \times \left(\frac{200}{100} \right)^2 = 64$$

N = 64 subjects in each group, before making allowance for non-response.

7. Further reading and resources

For details of sampling techniques

V Barnett. (2002) *Sample survey - Principles and methods*. 3rd edition Publ: Hodder Arnold

Bland M. (2000) *An Introduction to Medical Statistics*. 3rd edition. Oxford: University Press

For details of power calculations

Campbell MJ, Julious SA and Altman DG (1995) 'Estimating sample sizes for binary, ordered categorical and continuous outcomes in two group comparisons' *BMJ* **311**: 1145-8.

Petrie A and Samib C. (2005) *Medical Statistics at a Glance*. 2nd edition. Oxford: Blackwell Publishing.

Machin D, Campbell MJ. (1997) *Sample Size Tables for Clinical Studies*. 2nd edition. Oxford: Blackwell Scientific.

The software package *nQuery Advisor* provides simple efficient means of calculating power and sample size. It may be obtained from *Statistical Solutions* (8 South Bank, Crosse's Green, Cork, Ireland. Tel: 00 353 21 319629; Fax: 00 353 21 319630, e-mail: sales@statsol.ie).

More details are available from http://www.statsol.ie/html/nquery/nquery_home.html

The software Epi Info also provides information on calculating sample sizes and can be obtained for little or no charge from either your local Public Health Department or from your local academic department of general practice or from your local audit advisory group. More details are available from <http://www.cdc.gov/EpiInfo/>

Studies of power in different scientific disciplines

Fox NJ, Mathers NJ. (1997) 'Empowering your research: statistical power in general practice research'. *Family Practice* **14** (4): 324-329.

Polit DF, Sherman RE. (1990) 'Statistical power in nursing research'. *Nursing Research* **39**: 365-369.

Reed JF, Slaichert W. (1981) 'Statistical proof in inconclusive "negative" trials'. *Archives of Internal Medicine* **141**: 1307-1310.

8. Glossary

ANOVA (Analysis of Variance)	See: Oneway ANOVA (Analysis of Variance)
Bias	is the systematic deviation of the results from the truth. This may often be caused by poor sampling or poor questionnaire design.
Chi-squared(•²) test	is a non-parametric test (q.v.) of statistical significance. It is most commonly used to compare the proportion (of some property or event) occurring between two groups of people/patients. More generally speaking, it can be used to test for association between two categorical variables.
Confidence interval, CI, (for a mean value)	<p>indicate the precision of an estimate, usually for the mean value of some measurement. The confidence interval is specified by plus or minus two standard errors either side of the mean value itself, and represents the range of values within which we are '95% confident' that the true value (see 'Population') lies.</p> <p>Other methods may be used to calculate the CI for other quantities, such as proportions, differences between proportions, and regression slopes.</p>
Control group (in a clinical trial)	is the group in a clinical trial which is not exposed to the trial intervention. The control group exists to provide a baseline comparison for the intervention group so as to measure the influence of the independent variable on the outcome variable.
Correlation	is the degree to which two variables change together. This relationship is often taken to be linear (Pearson's correlation), but need not be (Spearman's correlation).
Descriptive design	is one which seeks to describe the distribution of variables for a particular topic. Descriptive studies can be quantitative, for instance, a survey, but they do not involve the use of a deliberate intervention.
Descriptive statistics	are used to describe and summarise variables within a data set including describing relationships between variables - often presented as graphs and tables.
Effect Size	is the magnitude of difference between two groups. E.g. in a drug trial comparing a novel drug with an existing one, the difference (proportion or %) in efficacy of the two drugs may be known as the effect size.
Error	can be due to two sources: random error and systematic error. Random error is due to chance, whilst systematic error is due to an identifiable source such as sampling bias or response bias.
Experimental design	is one in which there is direct control by the study team over the use of an intervention. In the classic experimental design, the subjects are randomly divided into intervention and control groups, and the patient outcome is assessed both before and after 'treatment'. See 'RCT'.
External validity	relates to the extent to which the findings from a study can be generalised (from the sample) to a wider Population, q.v., (and be claimed to be representative). Roughly speaking, it also means 'truth' or 'accuracy'.

Frame	see 'Sampling frame'
Hypothesis	a statement about the relationship between the dependent (i.e. outcome) and the independent (i.e. predictor, or explanatory) variables to be studied. Traditionally the null hypothesis is assumed to be correct, until research demonstrates that the null hypothesis is incorrect. See 'null hypothesis'.
Incidence	can be defined as the number of new occurrences of a phenomenon e.g. illness, in a defined population within a specified period of time. An incidence rate would be the rate at which new cases of the phenomenon occur in a given population. Also see 'prevalence'.
Independent variable	also sometimes known as a 'predictor' or 'explanatory' variable. Our interest in a study is usually in the association between one or more independent variables, and the 'dependent' or outcome variable (q.v.). In an experiment such as a clinical trial, the independent variable takes the form of the intervention or treatment, and is manipulated to demonstrate change in the dependent variable.
Inferential statistics	is the whole family of statistical techniques used to make generalisations from a sample to a population.
Internal validity	relates to the validity of the study itself, including both the design and the instruments used. Roughly speaking, this means 'internal consistency'.
Intervention	is the independent variable in an experimental design. An intervention could take the form of treatment, such as drug or surgical treatment. Those subjects selected to receive the intervention in an experiment are placed in the 'intervention' group.
Mean	is the 'average' value of a measurement. It is calculated by summing all the individual values and dividing this figure by the total number of individual cases to produce a mean average. It is a descriptive statistic which should only be applied to data on an 'interval' scale.
Median	is another measure of central tendency. It is the mid-point or middle value where all the values are placed in order. It is less susceptible to distortion by extreme values than the mean, and is a suitable descriptive statistic for both ordinal and interval data. It is the basis of nonparametric statistics (q.v.).
Mode	is the most frequently occurring of a number of mutually exclusive categories into which the study patients/subjects are divided. N.B. the 'mode' can be misleading when applied to numerical/quantitative data.
Nominal	data, also known as categorical data, is a set of unordered categories. Each category may sometimes be represented using a different numerical code, but the codes or numbers are allocated on an arbitrary basis and have no numerical meaning. See also 'ordinal' and 'interval data'.
Non-parametric statistics,	unlike parametric statistics, do not make any assumptions about the underlying distribution of data. Non-parametric statistics are therefore suitable for skewed data and nominal and ordinal levels of measurement.
Null hypothesis	is a hypothesis (q.v.), and is usually chosen to 'say' that there is <u>no relationship</u> between the dependent and independent variables. The null hypothesis is assumed to be correct, until research demonstrates that it is

incorrect.

Oneway ANOVA (Analysis of Variance)	is a test of statistical significance for assessing the difference between two or more sample means. Also known as F test.
Ordinal data	is composed of a set of categories which can be placed in an order. Each category is represented by a numeric code which in turn represents the same order as the data. However, the numbers do not represent the distance between each category. For instance, a variable describing patient satisfaction may be coded as follows: Dissatisfied 1, Neither 2, Satisfied 3. The code 2 <i>cannot</i> be interpreted as having twice the value of code 1.
Parametric statistics	are generally based on the assumption that the data follows a Normal distribution, i.e. the data when plotted follows a bell-shaped curve. Examples of parametric statistics are t-tests and analysis of variance (ANOVA).
Panel study	is another term for a longitudinal, cohort, or 'repeated measures' study, where individuals are examined and assessed repeatedly over a period of time.
Parallel design	refers to the traditional experimental design where each study subject/patient is allocated to just a single study treatment. An alternative to a parallel design is a 'crossover' design.
Placebo	is usually an inert drug or 'sugar coated pill' used to simulate drug treatment in a control group in an experimental design. Placebo is Latin for 'I will please'.
Population	is a term used in research which refers to <u>all</u> the potential subjects or units of interest who share the same characteristics which would make them eligible for entry into a study. The population of potential subjects is also known as the sampling frame.
Power	of a study is the probability of showing a difference between two groups, when there is a real difference of a pre-specified size between the corresponding populations. Therefore increasing the power of a study will reduce the chance of committing a Type II error. Power calculations are used to find out how likely we are to detect an effect for a given sample size, effect size, and level of significance. Power is usually denoted as $1 - \alpha$. The minimum recommended power level is 80%.
Prevalence	Is the number of cases or subjects with a given condition or disease within a specified time period. The prevalence of a condition would include all those people with the condition even if the condition started prior to the start of the specified time period. Compare with Incidence.
Prospective study	is one that is planned from the beginning and takes a forward looking approach. Subjects are followed over time and interventions can be introduced as appropriate. Compare with 'Retrospective'.
Qualitative	research usually dealing with the human experience and which is based on analysis of words rather than numbers. Qualitative research methods seek to explore rich information usually collected from fairly small samples and include methods such as in-depth interviews, focus groups, action research and ethnographic studies.

Quantitative	research is essentially concerned with numerical measurement and numerical data. All experimental research is based on a quantitative approach. Quantitative research tends to be based on larger sample sizes in order to produce results which can be generalised to a wider population.
Quasi-experimental design	is one in which the researcher has no control over who receives the intervention and who does not. An alternative to randomisation as used in experimental research, is the process of matching.
Quota sample	is a form of non-random sampling and one that is commonly used in market research. The sample is designed to meet certain quotas, set usually to obtain certain numbers by age, sex and social class. The sample selected within each quota is selected by convenience, rather than random methods.
Randomisation	is the random assignment of subjects to intervention and control groups. Randomisation is a way of ensuring that chance dictates who receives which treatment. In this way all extraneous variables should be controlled for. Random allocation does not mean haphazard allocation.
Random error	is non-systematic bias which can negate the influence of the independent variable. Reliability is affected by random error.
Randomised control trial (RCT)	is seen as the 'gold standard' of experimental design. As the name implies subjects are randomly allocated to either the intervention or the control group.
Ratio level data	is similar to interval data in that there is an equal distance between each value except that ratio data does possess a true zero. An example of ratio data would be age.
Reliability	is concerned with the extent to which a measure gives consistent results. It is also a pre-condition for validity.
Representativeness	is the extent to which a sample of subjects is representative of the wider population. If a sample is not representative, then the findings may not be generalisable.
Response rate	is the proportion of people who have participated in a study, or who have completed a question / questionnaire. It is calculated by dividing the total number of people who have participated by those who were approached or asked to participate.
Retrospective study	is a study in which the 'outcome' (the event or characteristic recorded by the 'Independent variable') of study subjects/patients may have already happened before the study was planned - and is typically found from records such as patient notes. Compare with 'Prospective'.
Sample	is the group or subset of the study 'Population' that is recruited for the study. A sample can be selected by random or non-random methods. Findings from a representative sample can be generalised to the wider population. N.B. sometimes each treatment group in a study is referred to as a 'sample', and so the 'total sample' is the combination of these treatment groups of study subjects.
Sampling frame	is the pool of potential subjects who share similar criteria for entry in to a study. The sampling frame is also known as the 'population'. An example might be: all people who suffered a stroke during the calendar year 2005

while resident in the UK.

Significance level	usually means the highest value for 'P-value' in statistical test that will be considered to be significant. The significance level is commonly set at either 0.01 (giving a one in a hundred chance of giving a 'false positive' test result) or 0.05 (a corresponding chance of one in twenty chance).
Significance tests	may be parametric or non-parametric. A significance test is used to detect differences between groups or associations between variables that are too great to have reasonably occurred by chance. The result is recorded as a P-value - or as statistical significance, according to some pre-set Significance level (q.v.).
Snowballing	is a non-probability method of sampling commonly employed in qualitative research. Recruited subjects nominate other potential subjects for inclusion in the study.
Spurious correlation	is an apparent correlation between two variables when there is no causal link between them. Spurious relationships are often accounted for by a third confounding variable. Once this third variable is controlled, the correlation between the two variables disappears. Another term for this is 'artifact'.
SPSS	(Statistical Package for the Social Sciences) is a popular and easy-to-use menu-driven software package for data analysis - its main limitation is that in some common situations, it is difficult or impossible to obtain 95% confidence intervals for your study results.
Standard deviation	is a summary measure of 'dispersion', i.e. of the variability of a set of measurements. It is a summary of how closely clustered or dispersed the values are around the mean. For data that is normally distributed, 68% of all cases lay within one standard deviation either side of the mean and 95% of all cases are within two standard deviations either side of the mean.
Standard error	is the standard deviation of an estimate of some quantity.
Stratified sample	is one where the sample is divided up into a number of different strata based on certain criteria such as age or sex or ethnic group. The sample selection within each strata is however based on a random or systematic sampling method. A stratified sample is a way of ensuring that the sample is representative rather than leaving it to chance.
Survey	is a method of collecting large scale quantitative data, but it does not use an experimental design. With a survey there is no control over e.g. who receives a medical intervention or when - but the corresponding advantage is that the researcher can examine the 'real world' and describe existing relationships.
Theoretical sampling	see the RDSU Resource Pack on Qualitative Research Methods.
Type I error	is the error of falsely rejecting a true null hypothesis and thereby accepting that there is a statistical difference when one does not exist. When designing a study, the acceptable chance of committing a Type I error is known as alpha.
Type II error	is the error of failing to reject a false null hypothesis or wrongly accepting a false null hypothesis. The likelihood of committing a Type II error, for e.g. a

treatment of a pre-specified size, is known as beta (β). The conventional level of statistical power ($1 - \beta$) and is usually set at 80% or 0.8.

t-test	is a test of statistical significance for assessing the difference between two sample means. It can only be used if the data distribution for each group follows approximately a Normal distribution and if the two sets of data have similar standard deviations.
Validity	is the extent to which a study measures what it purports to measure. There are many different types of validity.
Variable	is a term indicating any characteristic, assessment or measurement - i.e. a recorded item of data. It will typically be represented as a column of either text or numerical data in the table of 'raw data' from a research study.
Weighting	is a correction factor which is applied to data in the analysis phase to make the sample representative. For instance, if a disproportionate stratified sampling technique has been used, then the total data may need to be re-weighted to make it representative of the total population. Weighting is also used to correct for non-response, when the respondents are known to be biased in a systematic way.