

Chapter 3

Sampling

Resources are always limited. It is usually not possible nor necessary for the researcher to study an entire target population of subjects. Most medical research usually involves taking a portion (sample) of an entire group (population). As we have seen generalization is then made from the sample to the population from which the sample was taken (the target population), and further beyond to a larger public, to other settings, populations and circumstances. The question then arises as to how accurately does the sample reflect the entire population. Secondly, how true are conclusions derived from a small number of subjects for numbers several times larger, bearing in mind chance variations. To put it differently, these considerations relate to methods of sampling, and sample size.

Sampling Methods

The aim of all sampling methods is to draw a representative sample from the population. With a representative sample one can confidently generalize the results to the rest of the population from whom the sample was drawn. This would not only save time and money but also dealing with problems of validity and generalization. If the sample is biased (not representative) one can generalize less validly from the sample to the population.

There are two basic methods of sampling viz. probability sampling and non-probability sampling.

Probability Sampling

Probability sampling assures that each subject in the population has a known chance of being included in the sample. In non-probability sampling, there is no way of assuring that each subject has a known probability of being included in the sample; and therefore the sample cannot be expected to reflect the entire population. The conclusions of the research in such instances become less reliable or generalizable. On the other hand there is the advantage of convenience and low cost which sometimes outweighs such risks. For example opinion polls which are common in some countries at the time of elections are examples of non-probability sampling. The most easily accessible members of the public are polled, and the results cannot claim to represent the trend in the general population. They are, in fact, examples of incidental sampling.

Simple Random Sampling is one in which all the members of a population have equal chance of selection. The method involves making a list of all members of the population (preparing a sampling frame) and using random tables, dice etc. to throw up numbers for inclusion. The advantage of random sampling is that it is possible to estimate exactly how representative the sample is. Because a random sample is usually more representative than non-random samples, the sample size needed for good

prediction of population characteristic is small. Say, for example, the research question relates to maternal characteristics, events in pregnancy and during delivery, low birth weight and how these relate to child survival. In a hospital with 5,000 deliveries per year 500 babies of low birth weight are expected to be born, and the researcher wishes to take a sample of half that number. Each of the low birth weight deliveries is assigned a unique number, and then by using the random number tables the required sample of 1 in 2 is taken.

In this example, the heavier weight groups are likely to be over represented. To avoid this happening, the researcher resorts to *quota sampling* i.e. selects say the first 20 in each subgroup of 2500g - 2000g; 1999g - 1500g; 1499g - 1000g, and so on. Quota sampling helps to make the sample truly representative of different subgroups in the population.

The disadvantages of random sampling are the following:

1. The researcher needs to list every member of the population. Often this is not possible.
2. Cost is higher. It is cheaper to use conveniently available groups. Random sampling involves considerable planning and expense.

Stratified Random Sampling is similar to quota sampling. The difference is that each quota is filled by randomly sampling from each subgroup. For example simple random samples are taken separately of men and women in a study of hypertension, or of different social classes. The main reason for choosing this type of sampling is to overcome the possibility that the subgroups may differ significantly on the variable of interest; for example, social class and mortality rates.

The advantages of this procedure are as follows:

- All the important groups are proportionately represented.
- The representativeness of the sample is known.

The disadvantages are the following:

- A list of all members of the population, and the proportion of important groups is needed.
- Cost is higher.
- Gain in accuracy is not that much more than simple random sampling.

Area Sampling. The sampling is done on the basis of location. Locations are randomly selected, and then the researcher interviews the residents of these locations.

Systematic sampling. Every tenth or twentieth case is selected. It is not a truly random technique, but will usually give a representative sample. This method of sampling may be employed for selecting cases for study from a queue of mothers waiting to be seen in an outpatients clinic.

Cluster sampling. In this method of sampling aggregates of subjects rather than individuals are selected. Cluster sampling has been advocated by the World Health Organization for evaluating major public health programmes like the Expanded Programme of Immunization and Oral Rehydration Therapy. The principle of cluster sampling in a given area is to select randomly 30 clusters of 7 individuals each for data collection.

The method comprises the following steps:

1. identification of the geographical area of interest.
2. identification of the age group of interest.
3. preparing a list of all cities, towns, villages and settlements, and the sizes of their populations.
4. calculation of the cumulative total population of the area.
5. determining the sampling interval by dividing the total population by 30.
($SI = \text{total population} \div 30$).
6. identifying the number of clusters for each residential area by dividing their population with the sampling interval.
7. in each residential area selecting at random the number of clusters so identified.
8. random selection of a starting point (household) at each site or cluster.
9. selection of 7 individuals within the desired age group at each site. Selection begins at the starting household and continues to the nearest household until a total of 7 individuals is obtained. All individuals of appropriate age living in the last household are included, even if this means a cluster of 8 to 10 individuals rather than the required 7.

Non-probability sampling

The examples of *Incidental sampling* as opinion polls and exit polls have been already mentioned. Another form is *convenience sampling*; as for example, a study of veno-occlusive episodes in patients of sickle cell anaemia attending the investigator's outpatients department.

The practical realities of clinical work are such that non-probability sampling will occur in most instances. For example, an investigation of the prevalence of hypernatraemia in children attending hospital with diarrhoea, or that of anaemia in pregnant women. Both these examples relate only to a community of patients attending a particular hospital, and in no way provide a basis for estimating prevalence rates at other hospitals and in other communities. It may be argued that probability sampling was unnecessary in both the examples, and that the research was mainly intended to gain insight in a given situation.

SELECTING SAMPLES

Since a major interest in research is concerning etiology, which often means the association between exposure and outcome, most studies will select their samples by either exposure or outcome. If either of these is rare, the selection of a sufficient number of subjects is necessary to obtain a meaningful result.

With regard to exposure, the investigator may compare samples of exposed and unexposed subjects, subjects with different levels of exposure, or subjects exposed to two or more agents (or treatments). This kind of sample selection is usually employed in cohort studies.

With regard to outcome, the investigator may take a sample of subjects with the outcome, and compare them with those without the outcome for previous exposure. This type of sample selection occurs in case - control studies. Cross - sectional studies can use either method of selection i.e. by exposure or by outcome, or without regard to either.

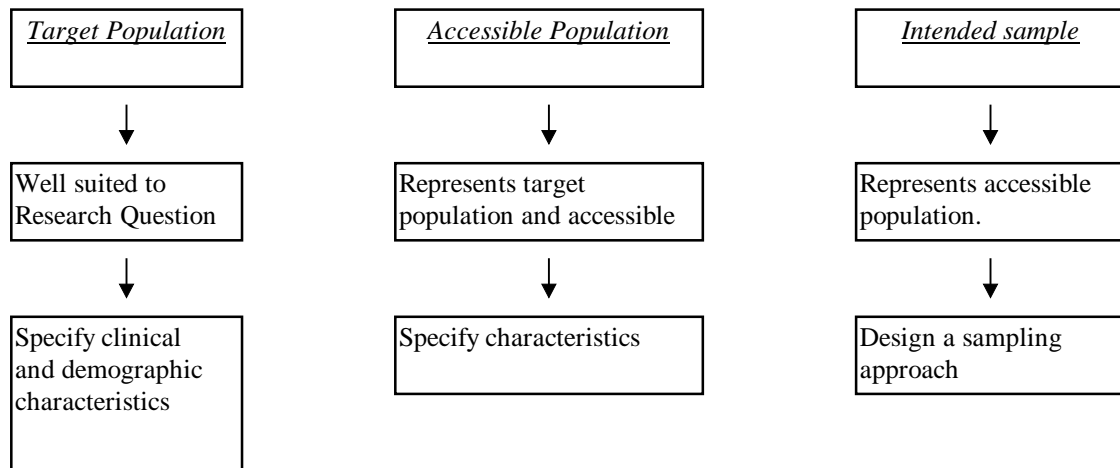


Figure 3.1: Characteristics of the sample and inferences from study results.

Internal validity

The crucial point in sampling is that regardless of what criteria are used to select study subjects, they should be representative of the target population. Because then only can the results of the study be safely extended to that population. This matter of internal validity has been already stressed. Since results which are not valid for the target population will certainly not be valid for the wider population, it is most important to achieve a high degree of internal validity (see figure 3.1).

The factors which can affect internal validity are the following:

Sample distortion

- The study sample became unrepresentative of the target population in one or more ways, resulting in erroneous conclusion.

Measurement bias.

- Errors in measurement of one or more interacting variables throw doubts on the results (e.g. growth studies in which accuracy of the measuring instruments have not been regularly checked or measurement techniques not standardized for different observers).

- Confounding factors. - One or more variables influence the interacting variables independently (e.g. education level of the mother influencing social class and health of the family independently).
- Reverse causality bias. - In studies where exposure and outcome are measured simultaneously there is no way of knowing which came first.

A study of the above list of factors would show that threats to internal validity can arise at any time during the course of a study. In the planning stage it occurs from faulty sampling. During the implementation stage sample distortion can occur because of drop-outs, migration, or death. Measurement bias can be another threat to internal validity during the implementation stage. Finally, during analysis and interpretation of findings internal validity may be at risk because of confounding and reverse causality.

Any observed outcome of a study depends on the particular sample drawn from the universe of subjects comprising the target population. Different samples from this population will necessarily provide different estimates of the outcome. This is due to the variability between subjects. A measure of the variability is the standard deviation. However, if a study is repeated on several samples, the difference in estimates would cancel out. In other words, the estimates would converge towards the true estimate.

Sampling error is the measure of the probability that any one sample is not completely representative of the population from which it is drawn. When data relating to a number of subjects have been assembled, the investigator may wish to apply the results to the actual as well as potential subjects. This collection of actual and potential subjects is referred to as population. The results of the analysis say something about this population. The subjects being reported on constitute a sample from the population. To make generalization possible they must be a representative sample.

With a very large sample it is almost always possible to obtain a significant result with statistical testing. Statistics are sensitive to sample size. A factor that produces a large difference which is statistically significant in a small sample is more worthy of attention than a factor that produces a small difference which is statistically significant in a larger sample.

SAMPLE SIZE

Since individuals in a population do not have the same value for any given variable (e.g. age, housing, social class, or income) but rather a range of values, a small sample might happen to have a mean or a rate that differs considerably from that of the entire population. As a rule small samples drawn from the same population are likely to have considerable sampling variation. On the other hand repeated large samples yield sample means or rates very close to the population values, and to each other. Thus sampling variation or *error* is inversely related to the sample size to a certain extent (see

figure 3.2) As a rule of thumb for a population survey a sample size of 250 is adequate in most instances.

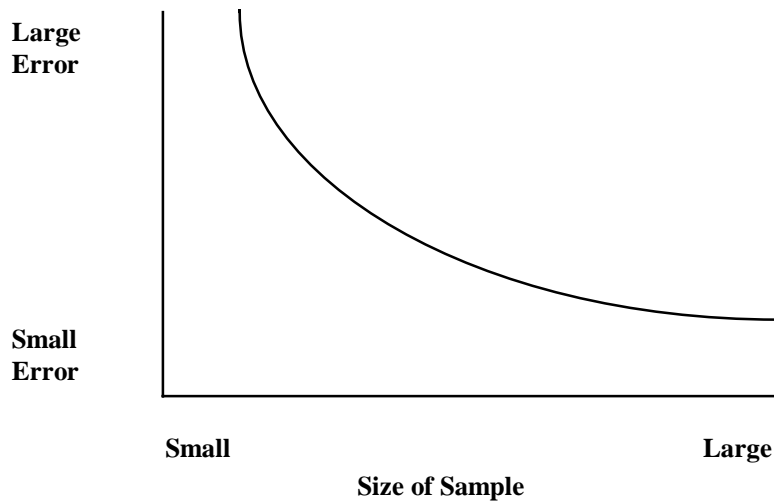


Figure 3.2: Sample size and sampling error.

When two groups are being compared the observed difference between them cannot be expected to represent the true differences exactly because of random variation in both the groups. In ordinary circumstances research is conducted on a sample, and there is always a possibility that the particular sample might not be truly representative even though selected in an unbiased way. But if the size of the sample was large enough the significant differences between study groups stand out better. On the other hand the researcher wants the sample size small enough to be practical. Thus deciding on the sample size is a balancing act.

There are four ways in which the results of a study relate to reality. Two of the four ways give correct conclusions and two incorrect as shown below:

	True Difference	True Difference
	Present	Absent
Conclusions of	Correct	Incorrect
Statistical test		(Type I error)
Different		
Conclusions of	Incorrect	Correct
statistical test	(Type II error)	
Not Different		

Type I error is similar to a "false positive" result i.e. saying there is a difference when there is not. Type II error is a "false negative" conclusion i.e. saying there is no difference when there is one. The larger the sample size the less likely it is that results would differ much from those for the population.

For the same difference between study groups the degree of statistical significance one intends to demonstrate has a bearing on sample size. Obviously a larger sample would be needed to demonstrate a significance of $p < 0.01$ than a significance level of $p < 0.05$. It has become customary to attach special significance to p values falling below 0.05. This is because it is generally agreed that one chance in twenty is a small risk of being wrong.

The likelihood of detecting an association between two or more variables depends on the actual magnitude of the association in the target population. If the association is large it would be easy to detect it in the sample, and difficult if it is small. But the investigator does not know the actual magnitude, and hence the reason for the study! Therefore, the investigator must *choose* the size of the association that he would like to detect, and for that purpose must make an informed guess. This is the trickiest bit in calculating sample size. The informed guess is made on the basis of findings from other studies, or from a pilot study that the researcher may have conducted. If there are no data at all, then the researcher must choose on the basis of the smallest effect size that would be considered meaningful.

Having decided about the association (or effect) size, the next step is to establish the maximum chance of avoiding Type I and Type II errors.

Type I (also called Alpha) error.

Sample size is also related to the risk of Type I error, i.e. a difference is shown in the study when there is not. The acceptable risk is a value judgment. If one is prepared to accept a large chance of falsely concluding that the observed difference is true one can take few subjects. On the other hand if one wishes to take only a small risk of concluding wrongly in this way, then a large number of subjects need to be recruited.

Type I error is related to statistical testing for significance. The convention is to take $P < 0.05$ as the cut-off point i.e. a 1 in 20 chance of the observed values arising simply by chance. A larger sample size is needed for $P < 0.025$, and an even larger for $P < 0.01$.

Type II (also called Beta) error.

The chosen risk of Type II error (i.e. the study concludes there is no difference whereas there is one) is another determinant of sample size. The choice is left to the investigator, but it is often set at 0.20, i.e. a 20% chance of missing true differences in a particular study. This is referred to as B, and $1 - B$ is called the *power* of the study. It is a measure of the probability that a study will find a statistically significant difference when such a difference really exists. A study is powerful if it has a high probability of detecting as different those observations that are really different. When the value of B is set at 0.2 the power of the study is said to be $1 - 0.2 = 0.8$ or 80%. Thus a power of 80 per cent is by convention considered a reasonable goal. As explained above it means that the probability of observing an association (or effect) of the specified level in the sample if it really exists is 80 per cent. If B is set at say 0.1, it would mean that the investigator is willing to accept a 10% chance of missing an association (or effect) of a given level. In other words, the power is equal to 0.9 or 90 per cent. That is the chance of finding an association (or effect) of the specified size.

In general, for any given number of subjects there is a trade-off between Type I and Type II errors. Everything being equal, the more one is prepared to accept one kind of error, the less it will be necessary to risk the other.

An example of a large number of subjects is the treatment of hypertension study in Britain. (*Br.Med.J.* 1985; 291:97 - 104) Because rate of outcome events was expected to be low a large number of subjects were recruited. It was estimated that 18 000 people will have to be followed for 5 years (90 000 person-years of observation) in order to detect a 40% reduction in the number of deaths due to stroke with an alpha error of 1% and a beta error of 5%.

In the event 17 354 people had been observed for 85 572 person-years by the end of the study. There was a 40% reduction in stroke rate and 18% reduction in all cardiovascular events in the treated group, but no difference in rate of coronary events or overall mortality.

An example of small sample size is Cimetidine study (*Lancet* 1977; 1: 4-7). 40 patients with active duodenal ulcer on endoscopy were randomly allocated to receive cimetidine or placebo. After 4

weeks, a second endoscopy was performed in this double blind trial. Ulcer healing was observed in 17 out of 20 patients on cimetidine, and in 5 out of 20 patients on placebo.

For most therapeutic questions a surprisingly large number of patients are required. The value of dramatic and powerful remedies can be established on small numbers. But such treatments come rarely. For many of the chronic diseases, a number of subjects and repeated trials are often needed. e.g. the United Kingdom Acute Lymphocytic Leukemia (UKALL) trials.

In practice Alpha and Beta cannot be zero, but they are made as small as is practical. Reducing Alpha and Beta increases the sample size. Hence planning the sample size aims at choosing a sufficient number of subjects which would keep Alpha and Beta at an acceptable level without making the study unnecessarily expensive or difficult. As stated in the preceding paragraphs, usually studies are set at a (Alpha) = 0.05, and B (Beta) at 0.20 (i.e. power of 80%). Researchers would keep a low alpha (i.e. $P < 0.05$) when the study requires to avoid Type I (False + ve) error, and low B (i.e. power $>80\%$) when it is particularly important to avoid Type II (false -ve) error.

To summarize, the steps to follow in calculating the sample size are:

- 1). Specify the probabilities of a and b errors. The convention for a error is $P < 0.05$ and for b a power of 80%.
- 2). For case-control studies make an educated guess of the proportion of the baseline population exposed to the factor of interest. Such a guess is often derived from similar studies reported in the literature.

In the case of cohort and intervention studies an educated guess is made about the proportion of the baseline population which has the disease of interest, and the magnitude of the expected effect which the investigator would consider meaningful. Methods of calculating the sample size are further described in the appendix at the end of this chapter.

Other points to consider in sampling and in sample size determination

VARIABILITY.

Variability also matters. Statistical tests of significance are designed to measure differences between groups. The greater the variability (spread) of the outcome variable in the subjects the more likely it is that the values in the groups will overlap. This would make it more difficult to demonstrate an overall difference between them.

SAMPLING UNIT.

The unit used for sampling dictates the unit used for analysis. For example, in a study of average stay in public and private hospitals, the sampling unit is hospital and not patients. In a study of immunization uptake between villages in a district the sampling unit is village and not children.

DROPOUTS.

Each sampling unit must be available for analysis. If dropouts are anticipated it is wise to increase the sample size by 10 to 20 per cent. Dropout rates in excess of 20% gives cause for concern.

CONFOUNDERS.

It is a good practice to think of possible confounders at the design stage of the study, and increase the sample size by roughly 10 to 20 per cent for each major confounder.

MULTIPLE HYPOTHESES.

A study may have several hypotheses and one may advance as many of them as makes sense. But it is prudent to specify one as the primary hypothesis. it would help to focus the study on its main objective besides providing a clear basis for the main sample size calculation.

During data analysis unexpected associations between variables may turn up. Many of them may even have significant P values. They are not to be used as established conclusions, but rather as a rich source for future research.

Appendix. 3.1 Calculating the desired sample size

The Principles.

Calculation of sample size depends on four factors:

- (i). sampling variation and variance.
- (ii). size of the effect of interest.
- (iii). level of significance.
- (iv). power of the test.

Sampling variation and variance.

Each sample has a slightly different value of the mean or proportion compared to the true value in the population. The measure of the variability is called variance and is equal to the square of the standard deviation. The smaller the sample the larger is the variance.

Size of the effect of interest.

This is a measure of the difference in the outcome that is being measured e.g. improvement by a new drug or improvement by a new form of treatment compared to the conventional one.

Level of significance.

This tells us how likely the observed result can arise by chance. For example $P < 0.05$ means that there is less than 1 in 20 likelihood of the observed difference arising by chance, and $P < 0.01$ means less than 1 in 100 likelihood of chance. (This value is also referred to as alpha).

Power of the test.

Power tells us how likely we are to detect an effect for a given sample size, effect size, and level of significance. (This value is referred to as 1 - Beta. By convention beta is set at 0.2 (i.e. 20%) and 1-Beta works out to be 0.8 or 80%.

For calculating the desired sample size, the first step is to decide what would be a clinically important difference that we would not want to miss. The next step is to set the power and the significance level.

Power as we have seen is the probability of obtaining a significant result given that such a difference does occur.

For a given significance level there are two ways of increasing the power. The first is to increase the size of the sample. The second is to increase the difference which is considered important enough not to miss.

There is a trade-off between power and significance level. Increasing the significance level will reduce the power. In general, a compromise has to be found between acceptable levels of significance and power. If for a given sample size the power is too low and we cannot increase the sample, there may be no point in proceeding with the study in its current form. The solution may be to rephrase the research question, including a broader range of patients.

Sample size may be calculated by using formulae or ready made tables which many textbooks of statistics provide. Whatever method is used it is important to remember that sample size calculations are based on our best guesses of a solution. The number arrived at simply provides an idea about the most suited number to be included in the study. It is in no way to be considered exact.

Below are examples of calculating sample sizes in most common situations. Moreover, appendices at the end of relevant chapters provide tables for calculating sample size. If more details are desired the reader should consult a textbook of statistics or be guided by a statistician.

Use of formulae for calculating minimal sample size

We first need to decide on appropriate values for the significance level and power. Using these two values obtain a value F from the table below:

Significance level required	Power required			
	80%	90%	95%	99%
0.100	6.18	8.56	10.82	15.77
0.050	7.85	10.51	12.99	18.37
0.025	9.51	12.41	15.10	20.86
0.010	11.68	14.88	17.81	24.03

For example, for a sample adequate to detect a difference significant at 5% level with a power of 90% we choose $F = 10.51$. This means that with F taken to be 10.51 the likelihood of detecting a difference if one actually exists is 90%.

Calculating sample size for comparing means

One sample

If d = acceptable difference between sample mean and the given mean (null hypothesis value), and SD = Standard deviation , then:

$$n > \frac{F(SD)^2}{d^2}$$

Two Samples

The calculated mean is the requirement for *each* sample.

Comparing two means

d = Difference between the means

SD = Standard deviation of each sample

then

:

$$n > \frac{F(2SD^2)}{d^2}$$

Comparison of two proportions

p₁, p₂ proportions

then

$$n > \frac{F\{p_1(1-p_1) + p_2(1-p_2)\}}{(p_2 - p_1)^2}$$