Sampling Errors, Bias, and Objectivity

10

David M. Flinton

10.1 Introduction

This chapter covers much of what should be considered before you undertake your research: what the population is; how to get a sample; and why sampling is important, and probability. The different types of bias that can exist in study design are covered. Finally, there is a brief section covering power and its importance in research.

10.2 Sampling

One of the key issues of research is how to choose a sample to be studied. Sampling is inherently different in quantitative studies compared to qualitative studies. A general difference between quantitative and qualitative research is that qualitative research utilises an inductive approach, taking specific information and making a broader generalisation. It is important in qualitative research that there is detailed, rich, complex data that are set in context. Qualitative data usually pertain to 'meanings' and as such are mainly in the form of words, themes, or patterns.

Quantitative research is more commonly associated with deductive approaches where a hypothesis is tested using statistical methods. We go from the general statement, the hypothesis to the specific: the observations. Because of this fundamental difference qualitative research tends to be based on purposive sampling where a small sample is selected because of certain characteristics. Quantitative research, which this chapter focuses on, relies on a sample, a sub-set of a population, being large enough to be representative of a population; otherwise, the results will be biased and not represent the population parameter.

D. M. Flinton (✉)
School of Health Sciences, City, University of London, London, UK
e-mail: d.m.flinton@city.ac.uk

A. Ramlaul (ed.), *Medical Imaging and Radiotherapy Research: Skills and Strategies*, https://doi.org/10.1007/978-3-030-37944-5_10

A parameter is a figure that is derived from the whole population. Samples are collected because it is usually not possible to collect data from the whole population, even when it is small. The term 'population' in this instance does not refer to the population as a whole, but rather is the set of individuals or items from which a sample is taken. For example, if a researcher was interested in quality of life after a heart attack in the United Kingdom (UK), the population would be all UK 'heart attack' patients; it is from this population that a sample would be taken. If we provide and report figures from a sample, they are referred to as statistics.

Parameters are very rarely known for a number of reasons. The main issue mentioned above is the inability to include all of a population due to the huge task of tracking down and collecting data on every single person in it. Another concern is the ethical issue of using a population when we know a sample would give an efficient estimate. There are also issues relating to the identification of subjects that could affect the ability to collect a whole population. Let us consider men with prostate cancer. How can we include those who die from undiagnosed prostate cancer? Finally, the transient nature of the population of interest can give problems. If we look at breast cancer, in England and Wales there are roughly 110 new cases every day and 30 deaths, so the breast cancer population changes on a day-to-day basis. Because of all these issues relating to collecting data from a population we take a smaller number of cases and assume that they are representative of the population: we sample.

The first stage of effective sampling is to define a population precisely and then construct a sampling frame. The ideal sampling frame is a list including all the items/people that you are trying to sample. So, for a study investigating radiography practitioners working in the UK, a comprehensive list of all UK based practitioners would be ideal. In practice the ideal hardly ever exists. In this case the closest we could probably get is the UK Health and Care Professions Council (HCPC) register. The list, if we could get permission to access it, would not be perfect as it would include practitioners taking a sabbatical, those who have recently retired or are maintaining their registration but working abroad.

The relationship between population, sampling frame, and sample can be seen in Fig. 10.1. Note in the diagram that the sampling frame is incomplete and does not cover the population so reflecting the discussion above. The figure on the left shows no sampling frame error as it takes evenly from the population, whereas we have a sampling frame bias on the right as we are not evenly covering the population. A sample is chosen to reflect a population from a sampling frame; from the sample we will get responses which we can analyse.

10.2.1 Sample Bias

If the data collected are not representative of a population then the estimates will not be accurate, and we say that the sample is biased. A selected sample is in some way systematically different to its population. The first and most obvious reason sampling bias may exist is if a sample size is too small and will therefore not produce a reliable

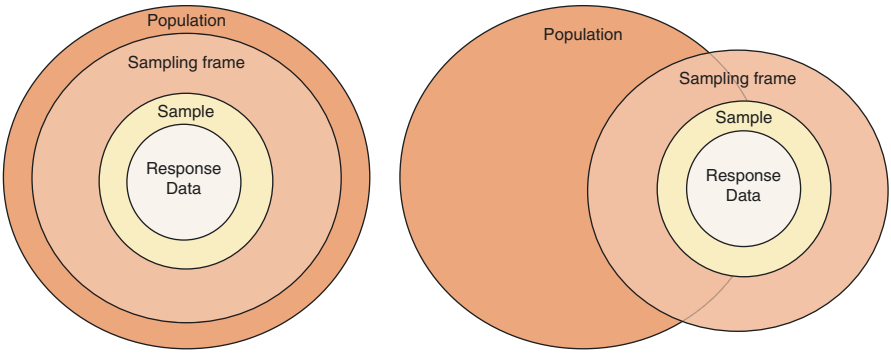

Fig. 10.1 Relationship between data and population

estimate of the population. Taking this to its extreme, my wife and I are both radiographers. If I used this sample, I might conclude that the gender ratio of radiographers is 50% male and 50% female, and 50% are therapeutic radiographers and 50% diagnostic radiographers. This obviously is not representative of the true population. This issue is explained further in the section on power calculations in Sect. 10.7.

Another factor to consider is the sampling method. There are various methods of sampling, some of which are better than others at obtaining a representative sample of a population and reducing sampling bias. In order to reduce bias a method is needed where every subject has an equal chance of inclusion/exclusion in a study. A sample is biased if certain members are under-represented or over-represented compared with others in the population; this bias can occur in the selection of both a study and control group. This may arise for several reasons some of which are detailed below.

10.2.2 Types of Sampling Bias

- Self-selection bias

  Imagine that a researcher was interested in what radiographers thought about the use of technology in the clinical teaching of students. A poster could be sent to all NHS radiography departments asking for volunteers for the study. Radiographers interested in technology or clinical teaching would probably be more likely to respond to this request, so they are effectively self-selecting themselves for the study and they are likely to differ in important ways such as age and gender from the population the experimenter wishes to draw conclusions about. This type of bias can lead to a polarisation of responses with extreme perspectives; those strongly supporting and those strongly against the issue being researched being more likely to respond than those who are more neutral.
- Selection from a specific real area.

  The above example also includes this type of bias as it only included radiographers from the NHS. Radiographers working in private hospitals might have

slightly different issues and considerations and therefore opinions compared with NHS departments. Another example of this type of bias might be if questionnaires were handed out at the front entrance of a hospital to those entering the hospital. This might give an overrepresentation of healthier individuals as the more infirm might use transport and thus might have a separate entrance, as might some separate clinics/wards. In this scenario these populations would be under-represented.

- Healthy user bias

  This form of bias occurs when a study population is likely healthier than the general population. The healthy user effect has been cited as a likely source of bias in observational studies looking at the use of hormone replacement therapy (HRT) [1]. It was postulated that women who took HRT were systematically healthier than those who did not use HRT; this implied that the benefits observed in the studies might not be due wholly to HRT.
- Berkson's bias

  This form of bias was first recognised by Dr. Berkson when looking at case–control studies where both the cases and controls were sampled from a hospital rather than from the population at large. A classic example of this was first published by Roberts et al. in 1978 [2] who, when looking at hospitalised cases, found a large positive association between the presence of both respiratory disease and locomotor disease. The association between respiratory disease and locomotor disease came about because the hospitalisation rate of patients with both these conditions was a lot higher than that for people who had only one of the conditions and even lower for patients without either disease. The observed association was therefore false; the finding would have been very different if the sample had been taken from non-hospitalised individuals.

10.3 Sampling Methods

Various sampling methods are at our disposal, some are better than others at removing sampling bias.

10.3.1 Random Sampling

The idea behind random sampling is to remove sampling bias. There are a number of ways of performing a randomisation process, some practicable, some not so practicable. Some types of sampling are detailed below, but this list is not exhaustive.

10.3.1.1 Simple Random Sampling

If we were interested in investigating patients' perception of the care received during their visit to casualty, a list of all patients who had attended casualty could be obtained from the picture archiving and communication (PAC) system during the timeframe in question. A random sample could be obtained by each name being written on a piece of paper, placed in a drum, and then randomly drawn.

58	06	96	03	51	50	09	96	67	74	08	97	06	71	22
89	08	40	54	15	03	69	94	98	91	94	21	91	29	06
91	88	08	83	54	54	13	04	94	67	70	01	31	25	18
38	66	48	14	30	31	03	96	65	30	53	43	55	20	97
24	35	69	21	18	55	71	78	54	94	80	58	47	46	48
45	60	39	34	12	91	57	51	73	08	01	18	58	92	87
74	04	28	68	68	60	67	37	34	48	22	86	73	51	53
06	57	05	72	96	97	27	78	55	27	57	77	50	08	68
24	74	76	86	46	82	64	38	07	30	42	09	48	15	05
88	81	89	45	85	68	79	50	38	10	80	74	93	23	39
55	96	15	31	08	60	04	04	98	24	21	81	45	12	83
50	42	28	55	02	16	49	48	46	14	72	41	83	08	56
50	72	79	30	45	88	47	51	44	73	31	99	76	80	18
58	84	67	26	03	86	96	77	42	59	04	01	58	99	86
48	07	34	94	44	45	14	79	40	72	48	14	01	05	92
48	58	32	58	97	87	76	42	29	20	11	83	94	89	92
48	94	21	60	13	93	48	44	82	39	74	85	68	11	13
14	78	45	22	08	11	77	20	35	75	41	43	25	31	44
35	67	95	35	86	02	03	29	35	42	87	53	10	18	46
95	44	89	14	56	52	25	47	96	79	52	04	59	73	04

Fig. 10.2 Table of random numbers

A better approach would be to use a table of random numbers. Each patient would be given a number and would be selected by matching numbers generated from the table. The table is random and so it does not matter where you start or in which direction you move. Assuming there were 820 patients available to a study, we could start at the top left corner of Fig. 10.2 and reading down move from right to left, so the first number generated would be 589, the next 891, then 008, 688 … and so on. Patients 589, 008, and 688 would then be approached to join the study. There was no patient 891 as there were only 820 patients in total so any number above 820 would be ignored. This process, although still time-consuming, is better than the first option proposed, but is still only really suitable for studies of relatively small size.

An important issue to consider is when we want to allocate subjects to groups: for example, if we want to produce two subject groups, one to receive the intervention and the other the placebo or alternative intervention. Again, we would like to remove selection bias and have a system of allocating patients to the groups which is random, and again using the random numbers table allows us to do this.

Again, let us make an assumption, this time that we want 24 subjects randomly allocated into two groups. We arbitrarily pick a starting point on the random numbers table and we have decided to move down and then along each block from left to right. The data selected are shown surrounded by a box in Fig. 10.2.

The numbers selected are then used to code which group they will belong to; odds go in the intervention group (I) and evens into the placebo group (P). Randomisation using the data would mean the allocation of subjects as shown in Fig. 10.3.

60	04	04	96	24	16	49	48	46	14	88	47	51
P	P	X	P	P	P	I	P	P	P	P	I	I

44	73	86	96	77	42	59	45	14	79	40	72
P	I	P	P	I	P	I	I	P	I	P	P

Fig. 10.3 Random numbers and sample sequences

1 I, I, P, P	2 I, P, I, P	3 I, P, P, I
4 P, P, I, I	5 P, I, P, I	6 P, I, I, P

Fig. 10.4 Block randomisation

5 P, I, P, I	6 P, I, I, P	6 P, I, I, P	3 I, P, P, I	5 P, I, P, I	1 I, I, P, P

Fig. 10.5 Sample sequence

The method is simple and random. It can give rise to uneven sizes of the groups, particularly in small trials, which can be a problem as if you calculated the sample size for the study it would have assumed equal group size. In the example above twice as many subjects were allocated to the placebo group compared with the intervention group.

10.3.1.2 Block Randomisation Sampling

Block randomisation overcomes the problem of the different number of subjects in different arms of a trial by keeping the subjects balanced throughout the study. The blocks can be of any size; they are usually a multiple of the number of treatments. If we use blocks with a size of four, we get six possible ways of assigning the two possible (placebo or intervention) treatments keeping the balance between treatments equal as shown in Fig. 10.4.

The allocation sequence is then decided by using the random numbers table to decide the sequence of the blocks. If we read horizontally on the table starting at the top left position, we get the figures 5, 8, 0, 6, 9, 6, 0, 3, 5, 1 so the selection of the 24 patients is as shown in Fig. 10.5. Note how we now have 12 subjects in each arm of the trial. This method can be further refined by varying the block length.

10.3.1.3 Stratified Sampling

Stratified sampling is a further development of block randomisation. It is used when it is important to achieve a balance between important characteristics in the subjects. A separate block randomisation is carried out for the important characteristic. If we were comparing alternative treatments for reducing stress in radiography

practitioners, for example, it might be important to stratify by gender. Each gender would have its own block randomisation, so each gender would be equally distributed between the two different treatments.

10.3.2 Purposeful Sampling

Snowball sampling is a technique for developing a research sample where existing study subjects suggest further recruits to take part in a study from among their acquaintances. It is of particular use when a researcher is studying a hidden population, a population with no sampling frame.

Judgement sampling is where a researcher actively selects subjects who are believed will be the most productive sample to answer a research question. This can be done via a framework looking at the possible variables that might influence a subject's contribution to a study.

Convenience sampling is where you sample subjects easiest to reach. It is generally considered as being the poorest way of getting a study sample, having the lowest credibility since the subjects are selected arbitrarily.

10.4 Bias and Error

One type of bias associated with sampling was discussed above. There are two other main forms: response and information bias. As we saw with selection bias a number of different variants exist and the same exists with these other two forms of bias.

Bias can be defined as an error in sampling or testing that will systematically affect the outcome of a study. If present, it infers that the findings of a study are less meaningful.

10.4.1 Response Bias

This is a type of bias that can affect the results of a study. It arises if there is a tendency for participants to respond inaccurately or falsely to questions.

10.4.2 Acquiescence Bias

Acquiescence bias occurs when respondents have a tendency to agree with all the questions in a measure. This may be made worse due to bad question design that encourages respondents to reply in a way they think the questioner wants them to answer, rather than what they actually think. An example of a badly phrased question is shown below. It leads respondents to an affirmative answer even if they disagree. Put differently the way the question is phrased makes you want to agree.

Please indicate the extent to which you agree or disagree with the following statement.

I was extremely satisfied with my radiotherapy treatment.

Agree	*Somewhat agree*	*Undecided*	*Somewhat disagree*	*Disagree*

Two ways acquiescence bias may be reduced is to 'reverse' some of the items on a questionnaire and to carefully consider question design. See the question below which asks the same question as above.

How satisfied/dissatisfied were you with your radiotherapy treatment?

Very satisfied	*Somewhat satisfied*	*Undecided*	*Somewhat dissatisfied*	*Very dissatisfied*

10.4.3 Demand Characteristics

Demand characteristics refer to a type of response bias where participants alter their response or behaviour because they become or think they become aware of what a researcher is investigating. They are trying to please the researcher (get it right) by conforming to what they see is the purpose of the experiment. Subjects might pick up on subtle cues such as body language or phrasing of the questions, which might be enough for the subjects to work out what the expectancy of a researcher is.

10.4.4 Social Desirability Bias

This form of bias occurs when a respondent provides an answer that they consider is more sociably acceptable, trying to conform to the social norm rather than revealing their own true opinions. The reason this is thought to happen is that respondents feel uncomfortable revealing their true answers. Latkin et al. 2016 [3] indicate that there are two dimensions to this: firstly altering the response to influence how the respondent is perceived by others; secondly self-deception, undertaking an action to enhance their own self-perception. Social desirability bias is more common when looking at sensitive or controversial subject matters. For example, consider what a potential respondent might think when confronted with a question that states: "indicate your level of racism on the scale below".

A number of ways of reducing this bias exist. Not revealing the purpose of a study and allowing anonymous responses help. Indirect rather than direct questions are also thought to reduce this form of bias. The use of a social desirability scale as part of the questionnaire can determine if respondents have responded with a high social desirability bias and therefore can be excluded from the study.

10.4.5 Extreme Responding

This occurs due to a respondent's tendency to pick the most extreme options from a ratings scale. It is most commonly observed in self-reporting questionnaires.

It is suggested that culture affects the amount of extreme responding and can be more pronounced in certain cultures. Culpepper and Zimmerman [4] found that Hispanic Americans evidence a lot of extreme responding. Van Herk et al. [5] noted higher levels in Spanish and Italian respondents compared to British, German, and French samples.

10.4.6 Question Order Bias

This occurs when a respondent may answer differently to questions based on the order in which questions appear. This is important when considering the design of a questionnaire or interview. This can occur as the respondents are unconsciously trying to apply meaning based on the order, i.e., a list of possibly responses might be interpreted as best on top, worse on the bottom. Another example might be if you draw attention to certain aspects in a preceding question as shown in the example below.

1. Which of the following features of your mobile X-ray unit would you most like to see improved? Battery life, physical size, limited mAs output.
2. What do you find most frustrating about your mobile unit?

10.5 Information Bias

10.5.1 Recall Bias

A subject's recall is thought to be dependent on their disease status, the exposure, even if irrelevant, being remembered better by cases than controls thus leading to exposure being under-reported by the controls. This can be exacerbated by certain issues such as a patient's preconceptions about the link between exposure and disease. These in turn can sometimes be influenced by the media, which may emphasise links between certain exposures to certain factors and the related health outcome. This problem was identified with case–control studies in the epidemiology section in Chap. 9 (see Sect. 9.4.2). It relies on subjects remembering an event in their past.

10.5.2 Observer Bias

The bias here is with the researcher(s). This occurs when researchers know the aim(s) of a study and allow this knowledge to influence their observations. In research the effect of observer bias can be removed by carrying out a specific type of study: a double-blind study. This is a study where neither the researchers nor the participants know which arm of a trial a subject is in. Note that there is also something called observer effect which is different to observer bias. The observer effect occurs when subjects change their behaviour because they know they are being watched.

This list is not exhaustive and other types of bias occur. Another area where bias exists is in scientific publications that can be used to support or refute work. Examples of bias that occur here are publication bias and reporting bias.

- Publication bias: the predisposition of journals to accept for publication studies that have a positive finding. Again, this can be exacerbated by authors, who have a tendency to only submit articles with a positive outcome.
- Reporting bias: the tendency of authors in studies that have multiple outcomes to only report the outcomes that are significant and ignore the non-significant outcomes.

10.6 Error

Error is generally considered to refer to a difference between an observed value and true value. There are two different types: random error and systematic error.

Random errors tend to mainly affect the variability around the mean; if a sample size is small, it may impact on the mean. Figure 10.6 represents two data sets: the blue line represents accurate data and the green line a repeat of the data collection with more random errors present. As can be seen in Fig. 10.6 random errors affect the precision of results as they are more spread out around the mean, i.e., have a larger standard deviation, but the mean is unaffected.

Systematic errors affect the mean value being reported, moving it higher or lower and their effect tends to be independent of the sample size. Systematic errors are often referred to as bias as they affect the accuracy of the results which are then no longer a true reflection of the population. Figure 10.7 shows the original data with a blue line. The green line represents the data recollected with systematic errors introduced.

Consider the example of an ionisation chamber being used to measure radiation dose. If the chamber was incorrectly positioned for one reading, we would have a random error and as the sample size is increased this one value would have less effect on the mean reading. If, however, the wrong chamber factor was given to a

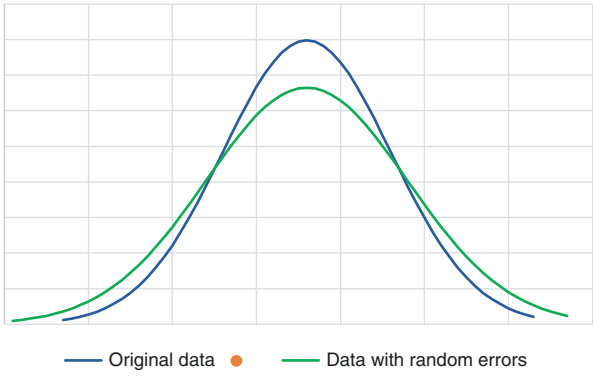

Fig. 10.6 Distribution change with random errors

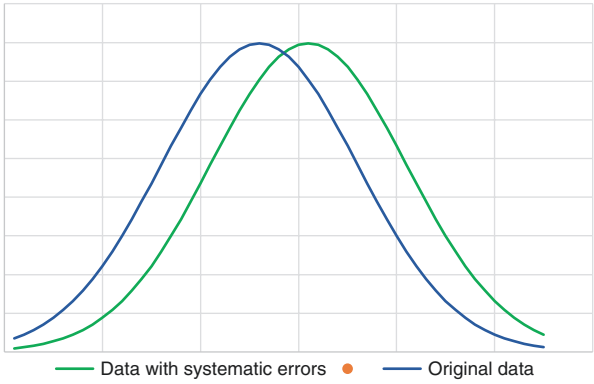

Fig. 10.7 Distribution change with systematic errors

researcher, this would affect all the readings, moving them in one direction (higher or lower), and the number of readings taken would not affect the results. With a systematic error all readings are affected to the same degree; the bias will not be apparent when looking at the data, but we might spot the random error as it might be very different to all the other readings. This is why it is important to look at the data before you start your analysis. You can simply look at the figures to see if one stands out as being different or you could do some simple plots to check the data to see if there are any outliers. This check should always be done as it highlights errors that occur when you input the data. Another way to reduce error coming in at this stage of the process is to carry out double entry of the data.

There are various types of bias that can be introduced into a study, some of which have already been mentioned. The types of bias a study is open to will depend on the type of study and how well it was performed.

10.7 Power Calculations

A frequently encountered problem in quantitative research is deciding how big a sample needs to be to find a 'reliable' result. The smaller a sample is, the less likely it is that if you repeated the sampling you would find you had the same result and the less reliable it is. On the other hand you do not want to waste time and resources collecting unnecessary data when a smaller sample would give a sufficiently reliable result. In reality this only applies to instances when it is feasible to repeat the exercise (such as handing out a questionnaire), but it is possible to extend the idea to non-repeatable samples as well in an imaginary way. So how do you measure reliability? Information on reliability and validity is given in Sect. 10.8 below.

When a decision is made to reject a null hypothesis it can be made on the basis that the *p*-value falls in a particular range of values (fixed level testing). One way to measure reliability could be to look at how often you would make the correct decision for a given significance level. This is what power calculations try to do.

There are four possible outcomes to a fixed level hypothesis test.

1. You accept the null hypothesis incorrectly.
2. You accept the null hypothesis correctly.
3. You reject the null hypothesis incorrectly.
4. You reject the null hypothesis correctly.

Power calculations work out the chance of the last possibility occurring. Notice that this does not consider all of the times that you might be correct in accepting the null hypothesis (case no. 2). Power calculation does not consider all of the ways that you could make the correct choice, only one of them (see also Chap. 15).

There are ways in which power calculations are useful.

- They can be done before a study starts in order to predict how big a sample needs to be to give a result of a required power (sampling requires time so this is a way of figuring out how little you have to do in order to get a 'good' result). This is called an a priori calculation.
- They can be done after a sample has been performed to see the power the study had (maybe a much smaller number of questionnaires were returned than you wanted and you want to find out how worthwhile it is to use the limited number that you have). This is called a post hoc calculation.

Trying to collect new data when an original sample has proved insufficient can be problematic; it involves more time and there are potential problems with dependency (i.e., one answer affecting another, for example, if subjects in the group you are sampling have talked about the research before you sample the second time).

The best way to use power calculations is as an a priori tool to try to predict how effective certain sample sizes will be. When collecting data using questions with human subjects there will always be problems about compliance. When conducting studies over a period of time, subjects may also drop out (failing to complete a study). This can to some extent be corrected by trying to predict the rates at which data might be lost and combining this with information from power calculations to find out how big a sample should be to leave data that give a reliable result. For example, a patient satisfaction survey might generate one return for every two patients in which case the sample needs to be twice the size of the sample calculated by a power calculation in order to achieve the required reliability.

It is important to realise that if you have obtained ethical approval for a study, then there are ethical issues surrounding the failure to collect enough data as this may involve waste of resources and misuse of subjects (especially if a study is patient based).

The 'power' of a test is often given as a percentage. The higher the percentage means the better the chance that your findings will be reliable. For example, the power of a test may be calculated as 80%. This means that if the sampling was repeated many times, then for 80% of those occasions the null hypothesis would be correctly rejected. By now you will hopefully have spotted that this is a probability of making the correct choice (for 80%, $p = 0.8$—it is just two different ways of

writing the same thing). A very important point here is that because it is a probability it cannot tell you about the exact occasions when you make the right or wrong choice, only how likely it is.

In addition to a decision about what level of reliability you will accept you also need to have an idea of what constitutes a 'significant' change or difference for your study. For example, if you are trying to find out whether recovery rates following treatment are improving, at what point does a change of a significant size occur? Finally, you will need to have an idea of the amount of spread in your data. If it is an a priori calculation you will need some estimate (maybe from a pilot study). This needs to be expressed in terms of standard deviation in order to perform the calculation. While it is possible to manually perform power calculations, the theory is heavily reliant on mathematics. In practice these calculations are performed using statistics software packages. You can do the calculations yourself or you can ask a statistician to do them for you.

10.8 Reliability and Validity

Two other terms often used in statistics are reliability and validity. Reliability seeks to describe the consistency or repeatability of a measurement. Validity refers to the strength of the conclusions drawn. In order for a study to be 'good' it must be both valid and reliable: neither by themselves is enough.

To illustrate this many authors refer to a metaphor of darts, as described and shown in Fig. 10.8. Four people have each thrown three darts at the board. Each used

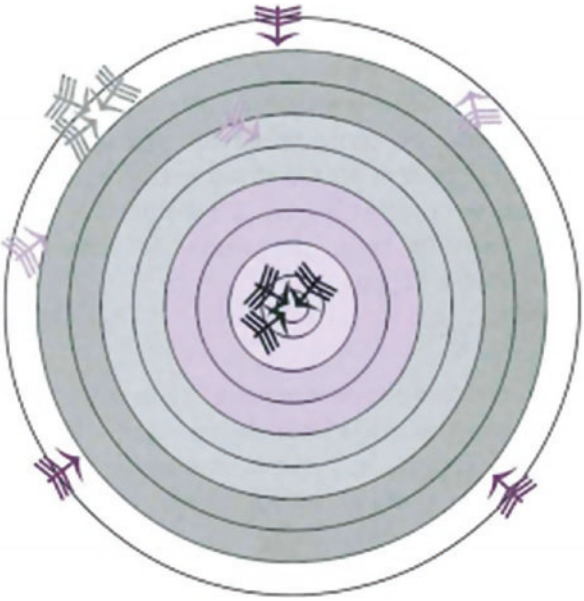

Fig. 10.8 Reliability and validity metaphor

a different colour dart. We are interested in how far on average they are from the bull. Two people (grey and black darts) managed to group the darts closely together, i.e., they were both consistent and reliable. Two people could also be described as being accurate (black and purple darts); the average of both colours is being the bull which equates to validity. Only one person (black darts) was consistent, accurate, reliable, and valid. This is what we should be aiming for in our research. The remaining person (mauve darts) was neither accurate nor consistent, so lacks both reliability and validity. But how do we know if we have a reliable and valid study?

10.8.1 Reliability

There are four common types of reliability: the inter-observer reliability, test–retest reliability, split halves reliability, and parallel forms reliability. Each is briefly described below. At the end of each description there is a test quoted. It is a common report of the type of reliability described.

- Inter-observer reliability

  In this instance a questionnaire is tested by a number of people or judges before being given to the test population. For example, if we wanted to know if a form was good at distinguishing between good and bad radiographs we could ask three experienced practitioners to each use the tool (questionnaire) to examine a number of different radiographs we have given them. As they will all be reporting on the same set of radiographs they should get the same or similar scores for each radiograph. If the raters do not agree with each other, all is not lost; we could retest the tool after training the testers to see if we get a better score. If we do, we may still have a reliable tool, but we would have to train the users to ensure this. Test: Cohen's kappa.
- Test–retest

  For this the test or questionnaire is administered to a sample and then re-administered after a time gap. The time gap is very important. Too short a time and the subjects remember what they said and try to emulate their first response, too long and there may have been a construct change, (things may have happened which means their response to the same question would be different). Test: Bland–Altman plot/test.
- Split halves test

  The test is only given once and then the test is divided into equivalent halves; a Pearson's correlation is then calculated between the scores from each half of the test. The closer the scores are between the two halves then the better the internal consistency. Test: Correlation.
- Parallel forms reliability

  This is used to assess the consistency of the results of two tests that were constructed in the same way. A large set of questions, which measure the same

construct, first needs to be produced. A major problem with this approach is that a lot of items that reflect the same construct have to be generated. The questions are then randomly divided into two sets and both instruments administered concurrently to the same sample. Test: Correlation.

10.8.2 Validity

As with reliability there are different types of validity: internal validity and external validity.

Internal validity is concerned with causality in the sample group studied. Internal validity asks the basic question, 'Did the experiment make a difference?' Another way of saying this: 'Was the experiment carried out in such a way that we are confident that the independent variable altered the dependent variable?' i.e., how confident are we about the cause and effect? Establishing internal validity can be threatened by a number of issues, such as confounding variables, outside influences/events, regression towards the mean and attrition from the study.

External validity is concerned with how results can be generalised to a population. The main threats to external validity are: the sample itself; sampling method; and time. For example, if a study was performed looking at radiography practitioners' perception of continuing professional development and was conducted predominantly using newly qualified radiographers, or the data were collected during a year when the HCPC audit to monitor radiographer registrants' compliance, these conditions would both affect external validity. In the first instance ensuring a random sample from all radiographers would help reduce the threat; in the latter instance a replication of the study would help eliminate the threat and demonstrate the generalisability of the results.

10.8.2.1 Construct Validity

Construct validity relates to a survey instruments, questionnaires, or tests and gauges how well we might expect the selected tool to perform at measuring what we think it is measuring. Do not get confused here. It does not refer to how well a questionnaire is constructed. A construct is the attribute, proficiency, ability, or skill that is being measured. Three variants of construct validity are briefly described below.

- Convergent validity

  This relates to the degree to which the test is similar to (converges on) other tests that it theoretically should be similar to. For instance, to show the convergent validity of a questionnaire that purports to measure fatigue in radiotherapy patients, we could compare the scores to a second fatigue test; high correlations would be evidence of convergent validity.

- Discriminant validity

  This is almost the opposite of the above. It is validity obtained when we measure two constructs that are thought to be dissimilar and the measures can discriminate between them. For instance, to show the discriminant validity of a spatial ability test, we might compare the scores with a test that looks at intelligence. Low correlations would be evidence of discriminant validity.

10.9 Conclusion

In this chapter the need for sampling was considered, as were the various forms of bias, random and systematic errors, and the concepts of reliability and validity. Statistics are created when describing/investigating samples. Sampling relies on being able to define a population precisely and to use an appropriate technique to avoid sampling error and to obtain an unbiased sample. There are many types of sampling, some are best in certain circumstances, but the best overall group of methods use probability sampling, which utilises some form of random selection. A random method of sampling gives each person an equal chance of being included in a study. Bias, a systematic error, and errors may be introduced into a study if it is designed incorrectly. Different types of bias occur depending on the study type. Deciding on the sample size of the study is very important: too small and it may not be representative of the population; too large and it is wasteful. It is also possible to calculate the power of a study after it has been conducted. In order for a study to be deemed 'good', the results must be both valid and reliable.

References

1. Gleason CE, Dowling NM, Friedman E, Wharton W, Asthana S. Using predictors of hormone therapy use to model the healthy user bias: how does healthy user status influence cognitive effects of hormone therapy? Menopause. 2012;19(5):524–33.
2. Roberts RS, Spitzer WO, Delmore T, Sackett DL. An empirical demonstration of Berkson's bias. J Chronic Dis. 1978;31:119–28.
3. Latkin CA, Mai VT, Ha TV, Sripaipan T, Zelaya C, Minh NL, Morales G, Go VF. Socially desirability response bias and other factors that may influence self-reports of substance use and HIV risk behaviors: a qualitative study of drug users in Vietnam. AIDS Educ Prev. 2016;28(5):417–25.
4. Culpepper RA, Zimmerman RA. Culture-based extreme response bias in surveys employing variable response items: an investigation of response tendency among Hispanic-Americans. J Int Bus Res Arden. 2006;5(2):75–83.
5. Van Herk H, Poortinga Y, Verhallen TMM. Response styles in rating scales. Evidence of method bias in data from six EU countries. J Cross-Cult Psychol. 2004;35(3):251–62.

Further Readings

Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
Bland M. An introduction to medical statistics. 4th ed. Oxford: Oxford University Press; 2015.
Petrie A, Sabin S. Medical statistics at a glance. 3rd ed. Oxford: Wiley Blackwell; 2009.

Web Resources

There are a number of web pages that deal with sampling and probability, including some with Java applications. Other useful resource on the web are random number generators.