Chapter 8: Quantitative methods

Henry Lucas
Institute of Development Studies

1. Introduction

As discussed in Chapter 7, in order to ensure a degree of independence, implementation research will typically not be funded from the intervention budget and the level of funding will almost always be relatively limited. Implementation researchers are therefore not in a position to insist that the data they need to meet their specific objectives should be made available within the general monitoring and evaluation (M&E) system. However, where possible, they should aim to be involved in the design of that system and may be able to negotiate modifications that serve their purposes. This may be possible either because those modifications are also seen as valuable to those with overall responsibility for the implementation – for example supporting operations research activities – or because additional resources from the implementation research budget are made available to complement those allocated to M&E within the intervention.

The above implies that implementation researchers will rarely be able to embark on independently managed and funded large-scale primary data collection activities but will have to rely mainly on the intervention M&E system, special studies using the 'qualitative methods' described in Chapter 8 and secondary sources. The key responsibility in this case is to adopt a systematic approach to determine the quality of the data to be analysed. This will be an ongoing challenge, given the tendency for data quality to vary over time, possibly improving initially as innovative systems for data collection are introduced and the enthusiasm of those involved is stimulated by access to new equipment and training workshops, but often deteriorating as that enthusiasm declines and systems fail to work as intended.

An obvious starting point when assessing data quality is the existence and completeness of those data. Missing information on facilities, providers, patients, etc. not only limits the scope of the analysis that can be undertaken, it typically biases the findings of that analysis. As a rule it will be the less well-resourced, less well-managed, most remote locations that are most at risk of providing incomplete data. Failure to recognise this trend can lead to a seriously over-optimistic view of intervention progress. Given that data are available, the statistical agency of the European Commission (Eurostat 2007) defines data quality as having five additional desirable attributes:

1. Accuracy;
2. Timeliness;
3. Comparability;
4. Coherence;
5. Accessibility and clarity.

Accuracy – essentially whether data reflect the true value of a given quantity – is obviously very difficult to test. However, we can check for obvious outliers – values that are almost certainly too large or too small. This can be a very important check, as many statistical procedures are highly sensitive to outliers, which can seriously distort the findings of any analysis. If we have time series data, excessive changes between one period and the next may also indicate measurement or recording problems. In some cases, issues may become apparent if we calculate rate or ratio indicators. For example, is the number of patients seen per day per doctor plausible? Are the recorded financial data compatible with patient numbers? In addition, by examining the frequency distributions of selected data items, as discussed below, it may be possible to determine if initial assumptions about those items have

proved valid. For example, attempts to assess patient satisfaction using a scale often result in a distribution in which almost no patients chose the lowest points on that scale. This should raise questions as to whether the scale we are using is an accurate reflection of patient attitudes.

Timeliness reflects the delay between the occurrence of an event or phenomenon and the availability of the associated data items. It is relevant in terms of both routine data systems and the intervention M&E system, which should be providing time series data that allow us to track implementation progress and link intervention inputs to outputs and outcomes. For example, a training workshop at a given point in time may be intended to result in improved staff performance, which is expected to produce better health outcomes by some future date. Assessment of the extent to which this sequence of events has taken place may be complicated by excessive delays in the availability of data from some sources, given that, as with missing data, such delays will often be associated with facilities or agencies that are performing less well.

Comparability relates to differences in concepts and measurement tools/procedures between sources – e.g. facilities, geographical locations, etc. – or over time. This can be a particularly serious problem in research on health systems, where different providers often choose to specify their own diagnostic and treatment protocols. Some unqualified providers may record all patients with a fever as suffering from malaria, while others rely on a range of rapid diagnostic tests. Some hospitals may record a diagnosis of tuberculosis based only on a chest X-ray while others require identification of Mycobacterium tuberculosis from clinical specimens. Additional issues arise if definitions are modified over time, perhaps as a result of the intervention itself. For example, faced with restrictions on outpatient costs by health insurance schemes, providers may simply vary their accounting procedures such that costs are transferred to inpatient departments. Such possibilities need to be carefully considered to avoid misinterpretation of apparent trends over time.

Where analysis involves the combination of data items from different sources it will be necessary to assess the extent to which there is *coherence* between those items. For example, to determine the extent to which some conditions remain untreated in the population it may be necessary to combine aggregates calculated from facility routine data systems with estimates of the prevalence of those conditions based on data from existing surveys. Those responsible for compiling these two sources will typically have used very different concepts, methodologies and instruments because they had distinct objectives. To the extent possible, any analysis must evaluate and try to address the implications of these differences.

Accessibility and clarity are perhaps most usefully understood as denoting the extent to which a researcher has the required level of understanding about the true nature of the data they intend to analyse. At a minimum this implies a careful review of the necessary 'metadata' – documentation that describes how the data are intended to be collected, compiled and stored – assuming these are available. However, it will often be clear that this documentation has limited relevance in terms of how those responsible for these activities proceed in practice. For example, heath staff tasked with introducing new procedures for collecting and recording patient data will typically soon find ways to reduce the time and effort required for an activity that they will often regard as a pointless addition to their workload. They may make assumptions as to the personal characteristics of the patient rather than ask the appropriate questions, or decide to enter data into the computer at the end of the day rather than 'waste' time after each patient visit as intended. The potential for misinterpretation if such issues are not addressed is evident. It may be frustrating to discover that an intended analysis cannot be undertaken as planned because the required data are not as originally assumed, but, as emphasised in Chapter 7, implementation research demands the highest standards of integrity and that includes ensuring that data limitations are thoroughly examined and addressed.

2. Rapid surveys

One additional methodological tool that can prove very useful in implementation research is the rapid survey. This description is usually applied to relatively small-scale surveys, typically of around 200 subjects or less, which aim to collect a very limited number of items of quantitative data over a short time period, say 5–10 days, that can be analysed and interpreted within at most a few weeks. Rapid surveys can target a variety of populations including facility records, health staff, patients, households and individuals. They adopt the probability sampling approach described in Chapter 7, and can therefore be analysed using statistical inference procedures that provide unbiased estimates of population 'parameters' (quantitative information) and reliable estimates of error bounds on those parameters. They should not be used to address more complex questions, for example the detailed operation of new incentive schemes or implications of new mechanisms for reimbursement of user fees. It is often assumed that surveys can be used to 'find out about' a policy question. In fact, the successful planning and implementation of a rapid survey typically requires that a great deal is already known, both about the population to be surveyed, and about the subject matter under investigation.

Rapid surveys are usually cross-sectional but may also be used to track changes over time (Tipping and Segall 1996). The small sample size and limited number of data items greatly reduce the administrative and logistic burdens associated with large-scale, multi-topic sample surveys, particularly those associated with the recruitment and training of field staff. This does not imply that such surveys should be undertaken without due consideration of the implications in terms of resource allocation. Though they are sometimes described as 'lightweight', it is essential that they are designed and implemented with all the rigour that should be applied to any study that intends to claim the respect that is reserved by many policymakers for findings derived from traditional statistical surveys. The range of tasks to be undertaken is identical to that required for a large-scale survey, even if the content of each is much more limited:

- Questionnaire design;
- Sample design;
- Mapping/listing to create the sampling frame;
- Preparation of fieldwork manuals;
- Recruitment of field staff;
- Training of field staff;
- Field enumeration and supervision;
- Transport and communications;
- Data preparation and processing;
- Computer analysis.

A number of the above require human resource management skills that some researchers either lack or are reluctant to practise. Again as with large-scale surveys, a key point to bear in mind is that not all of those involved in a survey can be assumed to have a direct personal stake in achieving a successful outcome. Without effective management and supervision, supported by a system of incentives and penalties, many will not perform to the standard required. Apart from such practical issues, it is also necessary to give some thought to the legal and administrative context within which surveys are undertaken. Are there laws that prohibit data being taken from a patient record or doctors providing information on the health status of an individual? If these activities are legal, is it necessary to obtain permission from a relevant administrative agency before undertaking them? Even if we have such permission, perhaps only because that agency wishes to encourage the intervention with which we are associated, we should still consider if we are abiding by the ethical criteria described in Chapter 6. As this is intended as a practical guide, we would also advise consideration of any potential political implications of our survey. Are we addressing a sensitive issue? Might some

stakeholders be concerned that we are gathering information that might be used to their disadvantage? What are the possible implications in terms of our overall research activities?

Another set of potential constraints that may impact on the quality of the survey are those relating to the targeted respondents. Can we assume, given that we have the appropriate permissions, that they will be cooperative? If you have undertaken the detailed stakeholder analysis discussed in Chapter 5, the findings from that analysis should provide information that will help you make such a judgement. Does what you know of your intended respondents indicate that they may have reasons – guilt, embarrassment, suspicions about your motives – to be concerned about providing you with data? Might they be irritated by what they see as an interruption to their normal activities? For example, in many countries busy frontline health workers often tend to regard all record-keeping as a largely pointless chore that takes time away from patient care. On the other hand, might their desire to be helpful – perhaps because they regard you as a high-status individual, or simply out of a natural tendency to be polite – lead them to provide data that they know to be unreliable and/or incomplete, rather than risk disappointing you?

This raises another issue. Even if respondents are cooperative, do they have access to the data you require? Do those data relate to current knowledge that they almost certainly possess or memories that may have become less reliable over time? Will they need to consult records? If so, do such records exist and are they complete and reliable? Finally, it is important to remember that one major disadvantage of questionnaire surveys is that it is very difficult in practice to ensure that every question will be interpreted in precisely the same way by *all* respondents. As a first priority, we should, if possible, try to ensure that the enumerator and respondent share fluency in a common language and that the questionnaire has been translated into that language (standard practice is to translate the questionnaire and then translate back for comparison with the original). Obviously, every effort should be made to avoid ambiguity and complexity in language. One useful approach is to deliberately try to identify any remotely possible way in which questions might be open to misinterpretation. Health-related surveys raise particular issues, in that researchers sometimes casually use technical terms that seem commonplace to them but that may be interpreted quite differently by some sections of the surveyed population. For example, the term ‘inpatient care’ is usually taken to imply at least one night spent in hospital, but could be seen as applying to any individual who has received treatment in a hospital inpatient department, for example reclining on a bed to receive a saline drip. Similarly, to a researcher the word ‘doctor’ may signify a qualified, licensed professional. In a remote village the same word may be used for an unqualified traditional healer.

Sampling designs

The sampling designs used in rapid surveys, as in the great majority of large-scale surveys, are based on the combination of a relatively limited number of elements:

- Simple random sampling;
- Systematic (list) sampling;
- Stratified sampling;
- Sampling with probability proportional to size;
- Cluster sampling.

The differences between these procedures can best be understood by considering a simple example. Suppose we wish to estimate the proportion of hospital clinical staff who have understood the basic principles of an innovative procedure following a one-week training course. If we have a list of all the staff in the hospitals, we could take a random sample and calculate the proportion of our sample who can answer a few simple questions about the

procedure. That could be used as an unbiased estimate of the proportion of all staff that would have been able to answer those questions, which we could interpret as the proportion with a good knowledge of the procedure. There are two slightly different types of random sample. In simple random sampling (SRS), we would select members of staff sequentially from the full list, which allows for the possibility that we may select the same member more than once. In simple random sampling without replacement (SRSWOR), we again sample sequentially, but this time excluding any member previously selected.

SRS is used as a standard sample design to which others are compared. It allows calculation of simple estimates of the required sample size for a given level of *precision* (the size of the lower and upper error bounds for the estimate). Thus:

> A sample of size 100, selected using SRS allows estimation of a proportion to a precision of +/-10 per cent with 95 per cent confidence;
>
> A sample of size 400, selected using SRS allows estimation of a proportion to a precision of +/-5 per cent with 95 per cent confidence (note that improving precision by a factor of two requires increasing the sample size by a factor of four).

We can interpret '95 per cent confidence' as implying that only 5 per cent (1 in 20) of such samples would be so misleading as to result in an estimated proportion that was further away from the true population parameter. We simply assume that we have not been so unlucky as to have chosen one of those samples[i].

In theory, SRS can be used as the reference to calculate the *efficiency* of any given sample design:

Efficiency = precision of SRS/precision using alternative design and same sample size

However, typically we do not have sufficient information to estimate efficiency but may simply be aware that one design is almost certainly more efficient than another. This is important because even though we may not be able to calculate the cost of achieving a given level of precision, a more efficient design can deliver increased precision for the same cost – i.e. it will probably result in a better estimate.

As indicated in Chapter 7, taking a random sample requires repeated use of a set of random number tables, a computer program or more recently a mobile phone app. Systematic sampling is a simpler approach that involves selecting a starting point on the list at random and then sampling every kth entry, where k is equal to the total number of entries divided by the required sample size (k=N/n). When the end of the list is reached, the process continues from the first entry. Except in rare cases where the list happens to follow a pattern that increases the risk of a biased sample, it can be shown that this procedure produces unbiased estimates with sampling errors that are at least as small as those from a random sample. It can in theory be more efficient than SRS if the list is ordered by a variable that is related to staff performance, for example if the names are listed by hospital, because there is less risk of selecting an unrepresentative sample – that is, one that does not include staff from all hospitals. However, as indicated above, it will typically be impossible to estimate the extent of the gain in efficiency.

If we suspect that knowledge of the procedure among staff may differ substantially between hospitals, we can increase the precision of estimation – i.e. narrow the gap between the lower and upper bounds on our estimate – by ensuring that our sample must include staff from every hospital. For example, in each hospital we might take a constant proportion (k=n/N) of staff members. This would be a stratified sample, with each hospital being a separate stratum. This sampling design results in a reduced sampling error compared to random sampling because we have excluded the risk of selecting samples that were mainly from under- or over-

performing hospitals, samples which would have under- or over-estimated the proportion of knowledgeable staff in the total population. The unbiased estimator of this parameter is calculated as a weighted sum of the proportions in each hospital (p_i), where the weights are equal to the number of staff in each hospital (N_i) divided by the total number of staff (N), i.e.

$$P = \Sigma\, N_i/N \times p_i.$$

Stratification by a range of other variables, for example gender, age or grade of staff, etc., might similarly be used to reduce the sampling error, if it were suspected that they were also associated with differences in staff knowledge. A basic principle is that the more information we have about our target population, the easier it will be to develop an efficient sampling design. It is important to understand that a stratified sampling design is intended to provide better estimates of overall population parameters. In the above example, we would be calculating statistics for individual hospitals and it will be tempting to compare, for example, average performance levels between those hospitals. However, that was not the purpose of the sample design and we will usually find that we simply do not have sufficient observations in each hospital to make such comparisons reliably.

For the estimation of a range of key population parameters, including totals, averages and rates, large entities – villages or urban districts with large populations, hospitals with a large number of inpatient beds, diseases with a high prevalence rate – are obviously very important in terms of their contribution to those parameters. In the above example, failing to obtain data from a small district hospital would have little impact on our overall estimate of the proportion of staff with adequate knowledge. However, failing to include staff from the largest national hospital could easily make a substantial difference to our estimate. One way to address this issue would be to stratify staff by size of hospital. As an alternative, the probability P of including a staff member of a given hospital in our sample could be made proportional to the total number of staff in the hospital, for example for all staff members in hospital i, the probability of being selected is:

P(i) = number of staff in hospital(i) / number of staff in all hospitals

This approach, called sampling with ‘probability proportional to size’ (PPS), increases efficiency by increasing the probability of inclusion in the sample for staff from large hospitals, thus decreasing the risk of taking a sample that excludes staff from these hospitals.

The PPS design is most commonly used in *cluster sampling* (Bennett et al. 1991), which is also referred to as two-stage sampling. In our example the hospitals can be regarded as ‘clusters’ of staff. If we decided that it would be too expensive, for example in terms of travel and accommodation costs, to send enumerators to every hospital within our study area, we might decide to (1) select a sample of hospitals and then (2) select a sample of staff within each of those hospitals. A common sampling design would be to use PPS to sample hospitals and systematic sampling to sample staff within each hospital. Cluster sampling almost always involves a loss of efficiency for a given overall sample size. As not all hospitals will be surveyed, if there are differences between them this design introduces a risk of selecting a sample of hospitals that is not representative. This risk increases as the number of hospitals in the sample decreases. Cluster samples typically need to be two to ten times as large as an SRS to achieve the same precision.

Estimation of sampling errors

In each of the above designs, the sample selected is simply one of the many that might have been selected using the same design and with the same sample size. The sampling error of an estimate is essentially a measure of the variability between all possible values of that estimate that might have been obtained from different samples. One way to reduce that

variability is to increase the sample size but that will imply a higher cost. The other is to improve the sample design, adopting sampling procedures that attempt to maximise, for a given sample size, the proportion of possible samples that will provide estimates that are close to the population parameter. We can never ensure that the sample that we do obtain meets this requirement, but using probability sampling we can make a reasonable estimate of the sampling error, which determines the risk of a 'bad' sample, and use this to modify the sample design or increase the sample size to ensure that it is less than some designated level – typically 1 in 20 or 1 in 100.

For a simple random sample (SRS) the sampling error can be estimated using:

$$se_{srs} = \sqrt{[\sum(x_i - \bar{x})^2/n]}$$

where x_i (i=1..n) are the sample values, $\bar{x}$ is the arithmetic average or mean of those values and n is the sample size. If we are willing to take a risk of 1 in 20, a remarkable mathematical result called the *Central Limit Theorem*, which can be derived from the basic definitions of probability theory, allows us to construct a 95 per cent confidence interval for the mean of the sampling population:

$$\text{Population mean} = \bar{x} \pm 1.96\ se_{srs}$$

Or a 99 per cent confidence interval:

$$\text{Population mean} = \bar{x} \pm 2.58\ se_{srs}$$

Note that the more confident we wish to be, the wider must our interval be. A similar formula can also be applied to confidence limits for proportions, as in the example discussed above.

> Example: A study was designed to evaluate the effect of integrating ITN (insecticide treated bednet) distribution on measles vaccination campaign coverage in Madagascar (Goodson et al. 2012). A national cross-sectional survey was undertaken to estimate measles vaccination coverage, nationally, and in districts with and without ITN integration. To evaluate the effect of ITN integration, propensity score matching was used to create comparable samples in ITN and non-ITN districts. Relative risks (RR) and 95 per cent confidence intervals (CI) were estimated via log-binomial models. Equity ratios, defined as the coverage ratio between the lowest and highest household wealth quintile (Q), were used to assess equity in measles vaccination coverage.
>
> National measles vaccination coverage during the campaign was 66.9 per cent (95 per cent CI 63.0–70.7). Among the propensity score subset, vaccination campaign coverage was higher in ITN districts (70.8 per cent) than non-ITN districts (59.1 per cent) (RR = 1.3, 95 per cent CI 1.1–1.6). Among children in the poorest wealth quintile, vaccination coverage was higher in ITN than in non-ITN districts (Q1; RR = 2.4, 95 per cent CI 1.2–4.8) and equity for measles vaccination was greater in ITN districts (equity ratio = 1.0, 95 per cent CI 0.8–1.3) than in non-ITN districts (equity ratio = 0.4, 95 per cent CI 0.2–0.8).

It should be emphasised that the above formula is only appropriate where the sample is selected using simple random sampling. There is a tendency for researchers to ignore this requirement and use the formula whatever the sample design adopted. As indicated in earlier chapters, the argument in this volume is that implementation research findings are too important for such disregard of established analytical procedures to be considered acceptable. To illustrate the problem, consider that rapid surveys will almost always adopt some form of cluster sampling. This implies that the above has to be modified to include a 'design effect', which measures the ratio of the sampling error of the cluster sampling design to that which would have resulted if an SRS design had been used.

Population mean = $\bar{x}$ ± 1.96 $se_{cluster}$

= $\bar{x}$ ± 1.96 x design effect x se_{srs}

The design effect will vary depending on the extent to which the clusters differ from each other. If this is large compared to the variability between the individuals within each cluster, the risk of sampling clusters that are unrepresentative is large and the design effect is large. In the above example, if the staff in some hospitals had all been well trained in the new procedure while in others training had been minimal, taking a cluster sample of a small number of hospitals would run a substantial risk of over- or under-estimating the overall level of staff proficiency.

The implications of a large design effect on the appropriate confidence limit bounds can be substantial. Table 1 below compares standard errors using the SRS formula with the appropriate standard errors for a clustered design where the clusters were villages. Note that the design effect varies considerably, from 1.13 for primary completion rates (limited between village variation because pupils travel to school) to 4.08 for improved drinking water (high proportion of variation between villages because this relates to a village level facility).

Table 1: Comparison of SRS and cluster sampling errors

Item	m	SRS se	Design Effect	Cluster se	m-2se	m+2se
Availability of ITN*	0.05	0.01	1.54	0.01	0.03	0.07
Iodised salt consumption	0.82	0.01	1.77	0.02	0.78	0.87
Improved drinking water	0.75	0.01	4.07	0.06	0.64	0.87
Primary completion rate	0.86	0.03	1.13	0.03	0.80	0.93
Attends secondary school	0.81	0.01	1.48	0.02	0.77	0.85

*Insecticide treated bednets

One reason for the inappropriate use of the SRS formula by research was the difficulty of calculating the correct sampling error, which often requires a considerable familiarity with the methods of theoretical statistics. However, such calculations can now be undertaken using well-established software packages such as STATA and SPSS, which require only that the researcher provide a detailed description of the sample design adopted.

The WHO Expanded Programme on Immunisation (EPI) surveys

The origin of the 'rapid survey' concept is often dated to the WHO '30 by 7' cluster surveys that were introduced in 1978 to obtain rapid, inexpensive but reasonably reliable estimates of child immunisation coverage (Lemeshow and Robinson 1985). The target population is subdivided into a complete set of non-overlapping 'clusters', usually defined by geographic boundaries (typically villages or urban districts). A sample of 30 of these clusters is taken with probability proportion to size (PPS) and then a 'quasi-random' sample of seven households with children in the relevant age range is selected within each of these clusters. Following this procedure, coverage estimates can be obtained that can be confidently assumed to be within ±10 per cent of the true value. The basic immunisation coverage survey instrument (the seven children sampled per cluster are typically recorded on a one-page document) usually records simply the cluster location, the age and sex of the selected child and their immunisation status. A similar methodology has been applied in rapid nutrition surveys, which have often been applied in emergency situations (Prudhon and Spiegel 2007). In this case it is usually recommended that the second-stage sample size should be increased to 30 children (SMART 2005).

The approach has attracted some criticism. Turner et al. (1996) focus on the lack of formal probability sampling of households within clusters. For example, one popular technique involves selecting a random direction from a central location within a village or urban district (traditionally by spinning a bottle). The households from the central point to the edge of the community in the chosen direction are listed, one is selected at random and then that household and its nearest neighbours are visited until the required seven children have been enumerated. None of the commonly used methods meets the basic requirement of probability sampling, that every eligible member of the target population has a known, non-zero chance of being selected. Simulation exercises suggest that the risk of sampling bias is substantially higher than in a conventional cluster sample. The paper suggests that a relatively simple modification can retain the advantages of the '30 by 7' design while ensuring a true probability sample. This involves: the production of a simple sketch map of each selected cluster; dividing this into segments of roughly equal size; selecting one segment at random; and interviewing all eligible members of the target population(s). This approach also addresses the common situation where surveyors attempt to gather information on multiple indicators (e.g. vaccination and childhood illness incidence rates) from the same sample.

Myatt et al. (2005) argue that while the PPS approach used in the '30 by 7' surveys may result in improved estimates overall, the associated tendency to sample areas of high population density may lead to a judgement that reasonable coverage has been achieved even where more remote, low-density areas have been severely neglected. They argue that this is of special concern in the case of feeding programmes, where a priority objective may be to identify such areas before children become severely malnourished. They describe an alternative approach which was first trialled in 2002 in the Mchinji district of Malawi where a district-wide feeding programme had been implemented. A 10km by 10km grid was overlaid on a map of the district. All those squares (quadrats) with more than 50 per cent of their area within the district were sampled. Communities nearest the centre of each quadrat were then sampled, with the sample size determined as the number that could reasonably be surveyed in a single day, based on the size of each community and the distance between them. All children in a community were screened to identify those suffering from malnutrition using a standard anthropometric criterion. Coverage in each quadrat was calculated as the proportion of malnourished children included in the feeding programme and overall coverage estimated by treating the quadrats as a stratum in a stratified sample. The survey was reported as proving simple, inexpensive and rapid, providing results within just ten days.

3. Quantitative analysis

As discussed in Chapter 1, implementation research has two broad aims:

1. Understanding implementation processes, focusing on mechanisms that support or constrain those processes;
2. Communicating that understanding to the multiple stakeholders who may contribute to the integration of findings into current and/or future implementations.

Those stakeholders may include:

- The implementation team;
- Providers and other actors in the health sector;
- National and local policymakers/officials;
- NGOs and CBOs;
- Donor and other international agencies;
- Beneficiary communities;
- The general population.

A key issue is that very few of these stakeholders will have specialist knowledge of quantitative or qualitative methods. It is therefore of central importance that analysis and, most importantly, presentation of findings must be carefully considered to avoid potential misinterpretations that could lead to inappropriate responses. Emphasis needs to be placed on simplicity and interpretability – stakeholders need to both understand the information provided and interpret it correctly (Walker et al. 2007). In terms of quantitative analysis, this implies an emphasis on simple summary statistics such as:

- Counts, means, medians, ranges, percentiles;
- Rates, ratios and (for some stakeholders) risks;
- Frequency distributions, proportions and percentages.

This does not imply that complex analytical techniques are never appropriate; only that final communication of the analytical findings should meet the above criteria.

Designing analysis by purpose

A second important preliminary consideration is to clearly assess the primary objectives of any analysis – what specific issues are you trying to address? Implementation research is by nature intended not to simply describe specific implementations but to improve the process of implementation. For example, we might focus on:

Effectiveness: Research that aims to modify implementation procedures in order to improve the flow of benefits that result from a given level of resources. This is typically the primary aim of implementation research. It should also assess 'how effective' and 'for whom'?

Efficiency: Analysis that attempts to assess the implications of possible modifications to the implementation process in terms of the value of benefit flows relative to resource costs. The aim will be to improve the benefit/cost ratio.

Equity: Analysis of distributional issues, i.e. how are benefits and resource costs distributed, typically relating to population subgroups?

Sustainability: Focus on identification of essential inputs, potential constraints on their availability and other possible barriers to medium- and long-term sustainability.

The aim in this section is not to teach statistical methods but to consider, given the objectives described above, the most appropriate choice of methods in the context of implementation research. Five main areas are addressed:

1. Frequency distribution and summary statistics
2. Relationships and confounding variables
3. Subgroup analysis
4. Statistical models
5. Generalising from samples to populations.

A note on levels of measurement in quantitative studies

Variables are usually classified by their 'level of measurement':

1. Rational, e.g. weight of child, number of vaccinations;
2. Interval, e.g. temperature, some disability measures;
3. Ordinal, e.g. facility levels, quality of life indices;
4. Nominal, e.g. district names.

The level of measurement should determine the appropriate type of analysis – for example, using an ordinal dependent variable in a regression contravenes one of the assumptions of such models. Researchers often ignore such restrictions. However, as previously indicated, because the findings are explicitly intended to influence important implementation processes and to be interpreted and used by a wide variety of stakeholders, it is probably reasonable to set a higher standard in implementation research.

Distributions and summary measures

Implementation research data can be seen as distributions of the values of study variables over selected study populations. For example, we may consider the distribution of white blood cell counts across patients, the numbers of children under five across households or outpatient attendances on a given day across primary facilities. Analysis can be seen as the use of techniques intended to summarise those distributions and estimate the extent to which they are related. For example, in a sample of newborn children we might summarise the distribution of birth weights by calculating the frequency of low, normal and high weight births, classifying as 'normal' those in some standard range. If we also calculated the frequency of different education levels for the mothers of those children, we could estimate the strength of a possible relationship between these two variables.

This use of frequency distributions, which show the number of values of a given variable that fall in each of several non-overlapping (mutually exclusive) groups, for this purpose (table 2) has a number of advantages. They are useful for all types of variable, easy to explain and interpret for audiences without specialist knowledge and can be presented graphically (figure 1) and/or in different formats to aid interpretation.

Table 2: Provider education frequency distribution

Level of education of private providers	Frequency
Illiterate	106
Basic literacy	74
Primary school certificate	57
Secondary school certificate	11
Higher-level qualification	2
Total	250

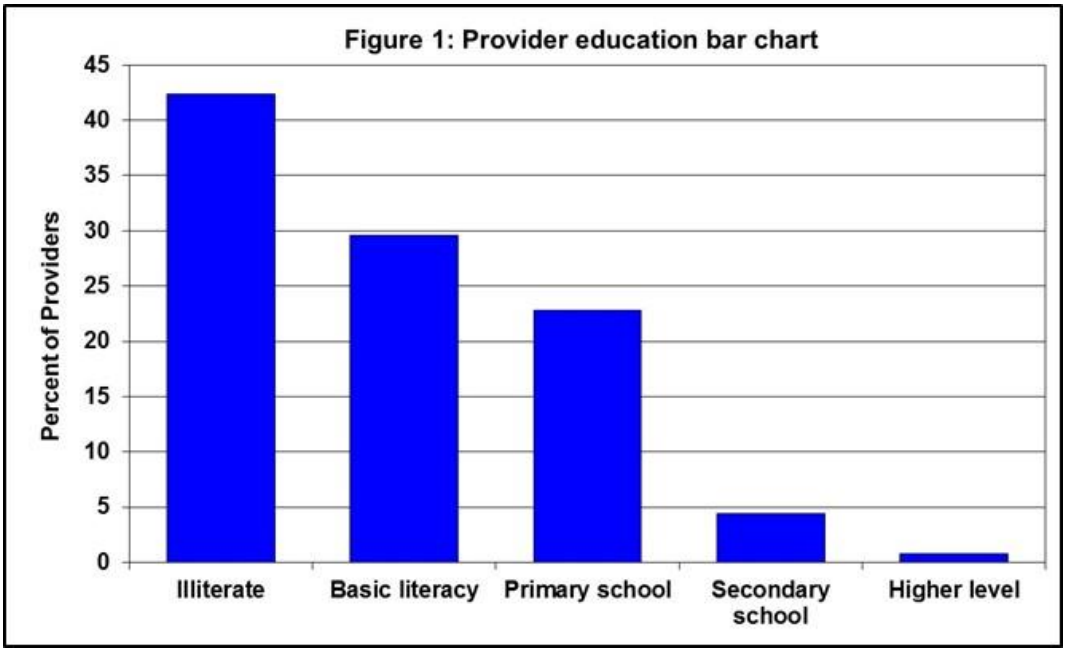


Figure 1: Provider education bar chart
Percent of Providers
45
40
35
30
25
20
15
10
5
0
Illiterate
Basic literacy
Primary school
Secondary school
Higher level

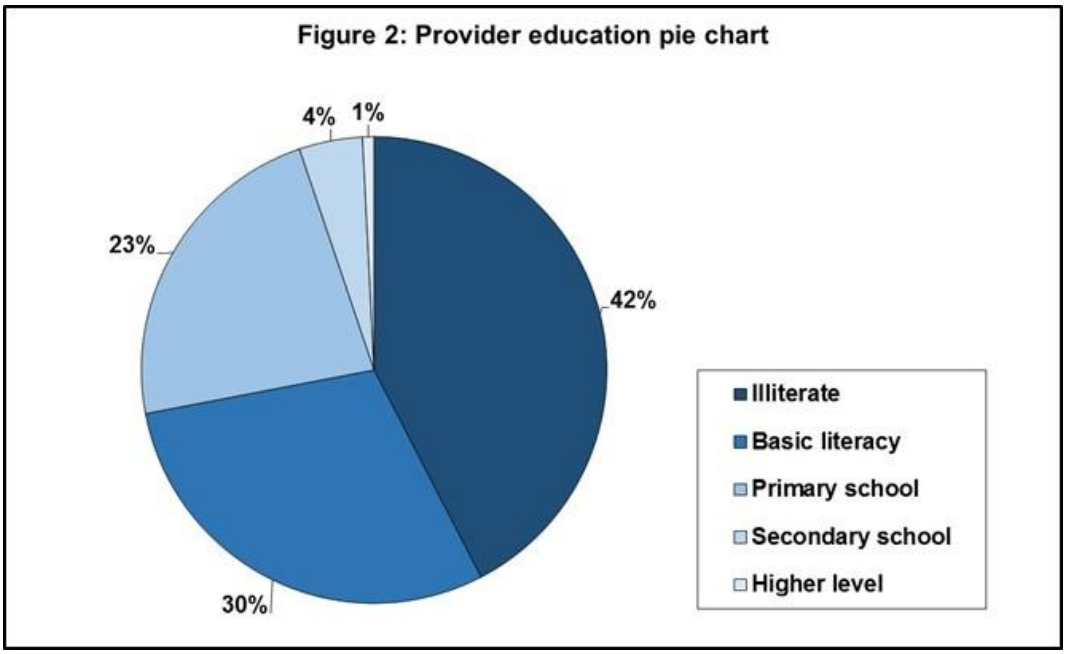
Frequency distributions provide an extremely useful approach to the presentation of large volumes of data. In the above example, information relating to 250 people has been used to construct one small table and, very importantly, no information has been lost in the process – that is, it would be possible to regenerate the original list of data values given the table. There are a number of interesting alternative ways of presenting the above data. We often, for example, calculate the 'relative frequency' (proportion, percentage) of data items that fall into a specific class. Again, to provide a slightly different perspective, we can 'cumulate' these percentages to show, for example, that 94.8 per cent of our population have at most a primary school certificate as in table 3.

Table 3: Alternative presentations of a frequency distribution

Level of education	Proportion	Percentage	Cumulative percentage
Illiterate	0.424	42.4	42.4
Basic literacy	0.296	29.6	72.0
Primary school certificate	0.228	22.8	94.8
Secondary school certificate	0.044	4.4	99.2
Higher-level qualification	0.008	0.8	100.0
Total	1.000	100.0	

Similarly we can experiment with different graphical displays. Figure 2 below shows the percentages as the segments of a pie chart. Note that the percentages are rounded to whole numbers. As a general rule, it makes sense to present data only to the degree of accuracy that (a) it warrants (estimates are almost always based on data that contain errors); and (b) makes the point we wish to make. Excessive precision (for example expressing numbers to more than one or two decimal places) confuses the eye of the reader and reduces impact.

Figure 2: Provider education pie chart



Defining groups for frequency distributions

A key decision in constructing a frequency distribution relates to the choice of groups. In the above examples, the educational attainment groups were predefined. However, we often have to

decide how to specify such groups in order to best summarise a given data set. For example, incomes will need to be grouped into ‘income bands’ and age data into ‘age bands’. The way in which this is done will depend on the aims of the analysis. Demographic analysis, for example, will often aggregate ages into fixed five- or ten-year age bands, such as 0–4, 5–9, 10–14, etc., with a final open-ended group such as ‘75 and over’ (note that these classes are defined such that there is no overlap – the second, for example, relates to ‘all children of five years or older but under ten years’). An educationalist, on the other hand might use groups such as 0–5, 6–12, 13–15, 16–18, 19+, where the groups are defined in line with the official age ranges for specific levels of education, for example pre-school, primary, lower secondary, etc.

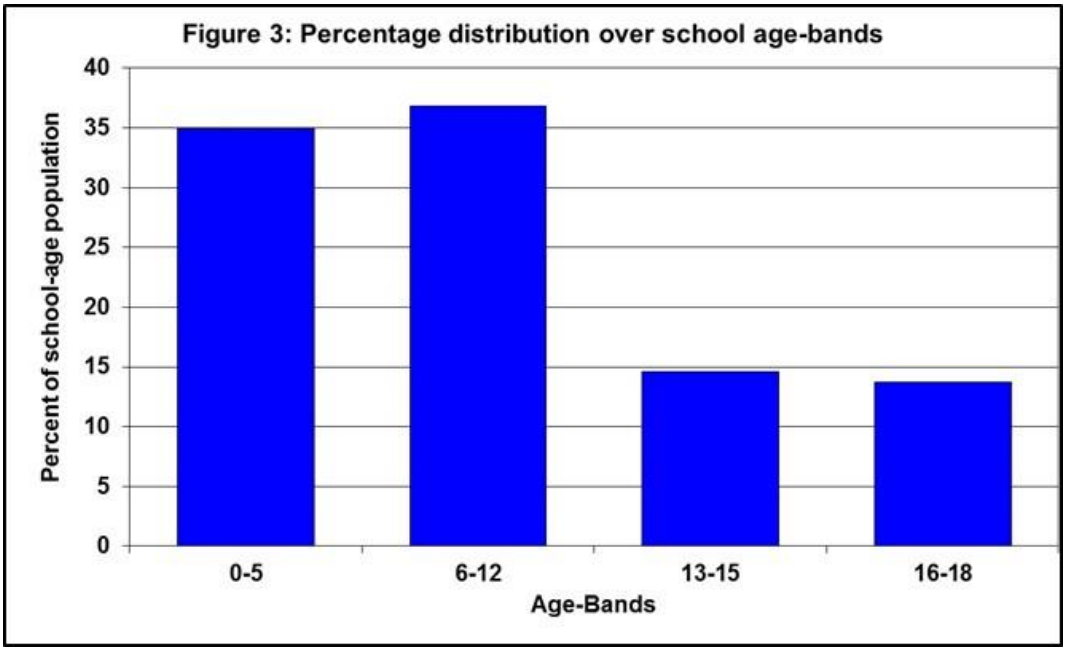
Just as above, we can construct a frequency distribution based on these groups, showing the number of people falling into each age band. However, here *the definition of groups does involve a loss of information*. Given the number of children in the 5–9 age band, we cannot deduce the ages of the individual children in this group. More frustratingly, we cannot, in the above example, derive the frequency distribution preferred by the educationalist if we are presented with that derived by the demographer. This can be a major problem because we often wish to combine distributions from more than one source. For example, we might know the number of children in primary school and wish to express this as a proportion of all children in the 6–12 age band. If we only know the numbers in the 5–9 and 10–14 age bands, we cannot directly calculate the number we require and have to resort to a weighting procedure based on more or less plausible assumptions.

How should groups be defined? In some cases, such as types of facility or staff salary ranges, official definitions may be most appropriate. If such obvious classifications do not exist or do not serve our purposes, we usually try to balance two conflicting objectives – limiting the loss of information (by using a relatively large number of groups) and providing a simple summary (by using a relatively small number of groups). In general, we would also prefer to make all the group intervals of equal width, because this simplifies comparisons between one group and another. In table 4, for example, a much higher percentage of the studied school-age population are in the second age band than are in the third. However, this is obviously at least partly because this group covers a greater range of ages – seven years as compared to three.

Table 4: Percentage distribution by school age-bands

Age band	Percentage
0–5	35
6–12	37
13–15	15
16–18	14
All	100

Note that the column chart below, which is derived from these data, does not reflect the variations in group ranges. The age bands are used simply as labels for the columns, which are all of equal width. It is the height of the column that shows the percentage falling into each age-band.



Joint frequency distributions

One of the simplest and yet most powerful techniques for analysing and presenting data involves comparing the frequency distributions of two groups within the study population. Table 5 takes the data used above and disaggregates by the gender of the respondent.

Table 5: Joint frequency distributions for two or more variables

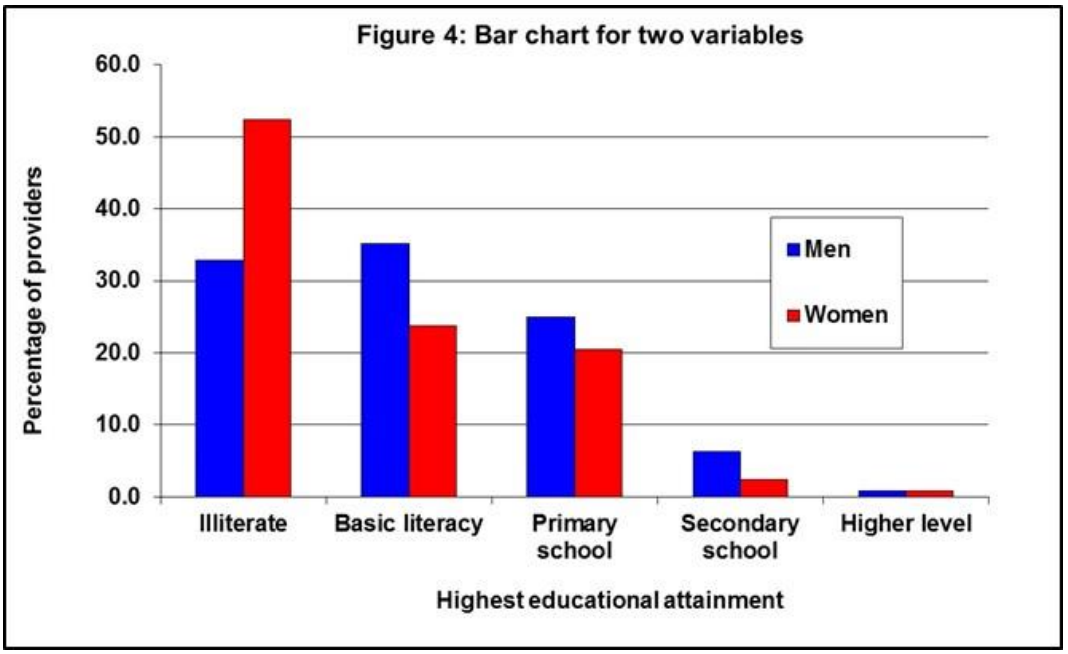
Highest level	Men	Women	All
Illiterate	42	64	106
Basic literacy	45	29	74
Primary school certificate	32	25	57
Secondary school certificate	8	3	11
Higher-level qualification	1	1	2
Total	128	122	250

Doing this reveals interesting new information. Although almost the same number of men and women were asked (128 and 122), it would appear from our sample that educational achievement is much higher for the former. We can make the comparison clearer by using the relative frequencies or percentages based on the total number of individuals in each group. Table 6 shows, for example, that 52.5 per cent of women are reported to be illiterate as compared to 32.8 per cent of men. Obviously the conversion to percentages would be even more useful if the numbers in the two groups differed more substantially.

Table 6: Joint distribution using column percentages

Highest level	Men	Women	All
Illiterate	32.8	52.5	42.4
Basic literacy	35.2	23.8	29.6
Primary school certificate	25.0	20.5	22.8
Secondary school certificate	6.3	2.5	4.4
Higher-level qualification	0.8	0.8	0.8
Total	100.0	100.0	100.0

The above table can again be presented graphically in a column chart, as in figure 4.



An alternative presentation, that can be useful if we wish to focus on the composition of each class, is obtained by calculating row percentages based on the number of individuals in each education group (table 7).

Table 7: Joint distribution using row percentages

Highest level	Men	Women	All
Illiterate	39.6	60.4	100.0
Basic literacy	60.8	39.2	100.0
Primary school certificate	56.1	43.9	100.0
Secondary school certificate	72.7	27.3	100.0
Higher-level qualification	50.0	50.0	100.0
Total	51.2	48.8	100.0

When interpreting percentage distributions it is always important to check on the absolute size of the denominator on which they are based. For example, the above table shows that 50 per cent of those with a higher-level qualification are men and 50 per cent women. Before getting too excited about this apparent example of gender equality, we should note that only one man and one woman are in this class!

Summary statistics and frequency distributions

Careful examination of the frequency distribution of a variable can be an extremely powerful and robust form of analysis. Unfortunately it is often bypassed. There is a tendency to move too quickly to the calculation of simpler 'summary statistics' that are intended – but often fail – to capture the essential features of the distribution. These usually focus on the derivation of measures:

- to indicate the overall 'location' of a distribution – how sick, poor, educated is a study population 'on average'?

- to indicate the extent of ‘variation’ within that population.

However, the reasons for selecting a particular summary statistic should obviously relate to the purpose for which it is intended. For example, if we ask the question ‘Has the recently implemented intervention reduced the problem of malnutrition among five-year-olds in this village?’, there is no doubt as to which of the following possible summary statistics would be more useful:

- Change in mean daily calorie intake of all five-year-olds in the village, or
- Change in proportion of five-year-olds in the village falling below a predetermined minimum calorie requirement.

Bearing in mind the above discussion about the need to present research findings in ways that are appropriate to the various stakeholder groups, appropriate criteria for the selection of summary statistics might be: (1) is the statistic clearly relevant to the specific concern we wish to address; (2) will stakeholders understand how it was derived; and (3) will stakeholders interpret it as intended – that is, are they taking what we would regard as the right message from the information we are providing? We can consider how to apply these criteria by considering some simple examples.

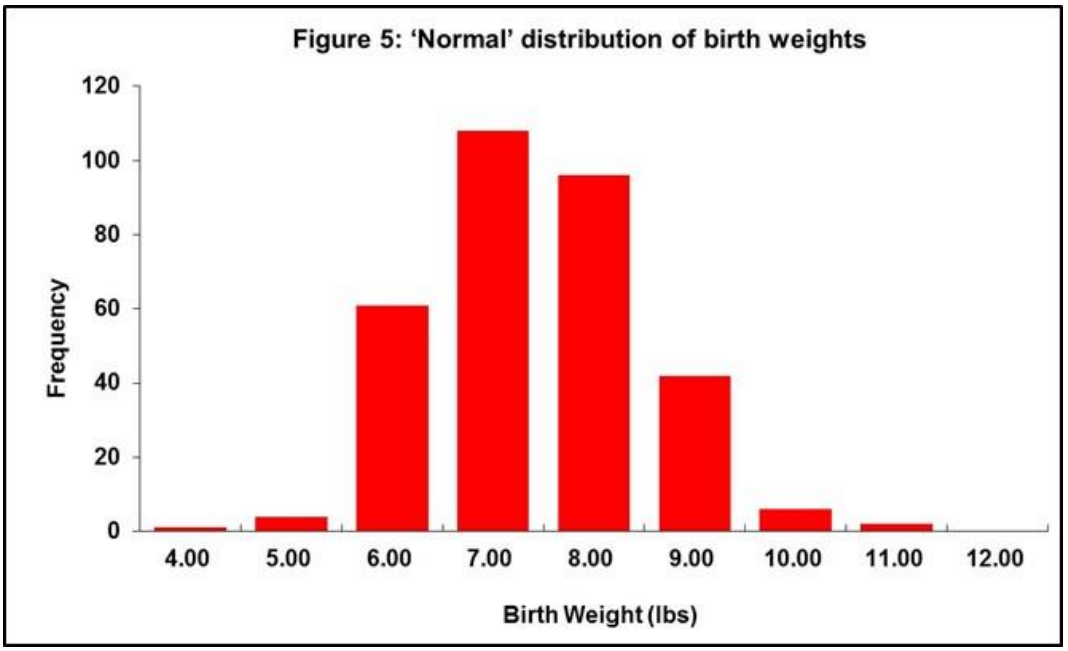
Mean or median?

There is a tendency for quantitative analysis of continuous variables to start by comparing mean values over time – for example by how much has the mean cost of treatment increased – or for different sections of the population, such as how does the mean length of stay in hospital vary between urban and rural populations? The mean is the most commonly used statistic, often seen as the ‘natural’ measure of central location and used without much thought. However, this is mainly because it is simple to calculate and manipulate. In the days before analysis was done by computer, it was relatively easy to calculate means either by hand or using a calculator. Moreover, given the means for two population groups (for example, two health districts) it was very easy to calculate the mean of the combination as:

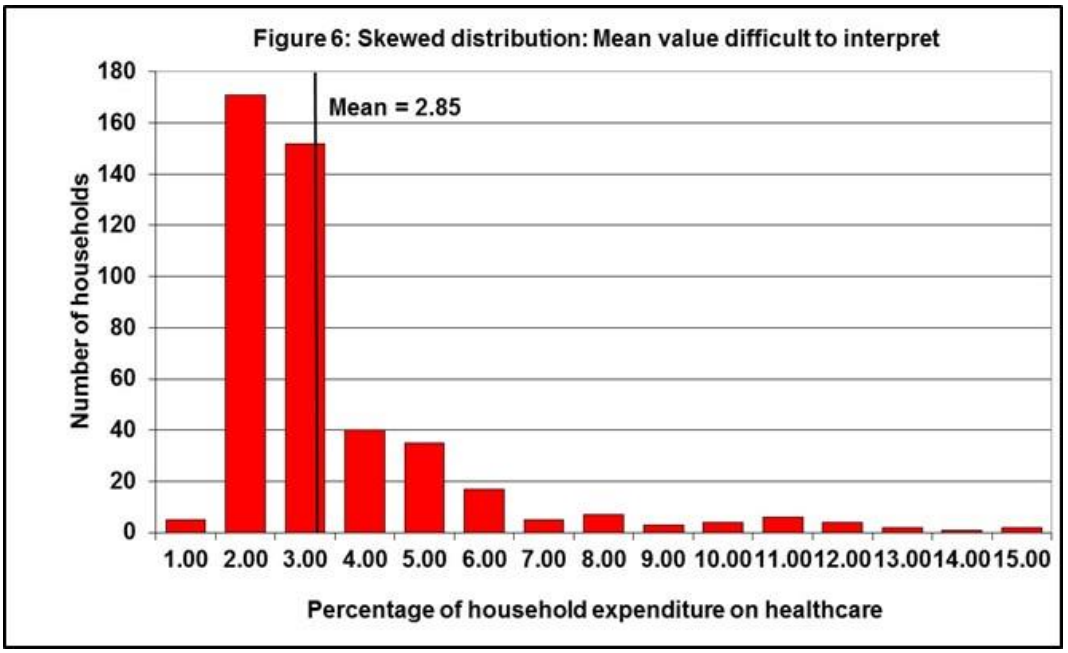
$$\text{combined population mean} = (n_1 \times \text{population mean}_1 + n_2 \times \text{population mean}_2) / (n_1+n_2)$$

Where n_1 and n_2 are the number of observations in the two populations.

On the other hand, we know that most people tend to misinterpret the mean. They assume that it can always be seen as representing the ‘typical’ value in a population, for example interpreting GDP/capita as the income of a typical person in a given country. In practice this is only a valid interpretation in the case where the underlying frequency distribution is symmetric, for example the so called ‘normal’ distributions that tend to occur for physical measures such as age-specific heights and weights. For example, in Figure 5 the mean birth weight is 7.5 lbs, which can be seen as a providing a reasonable idea of the typical birth weight of a baby in this population.



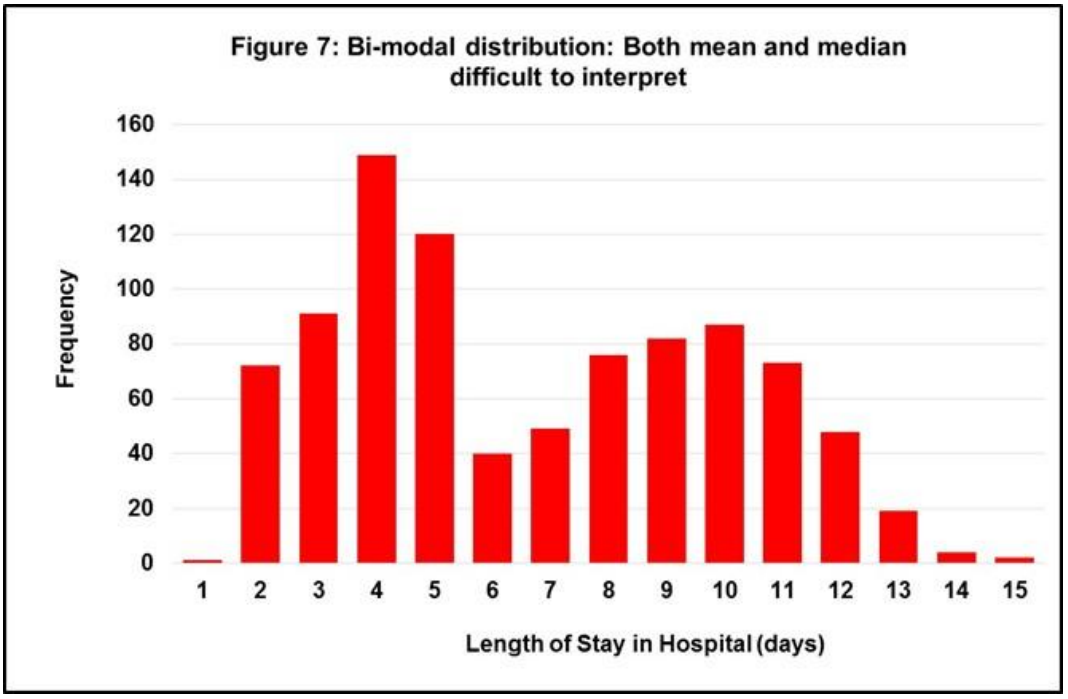
When the distribution is 'skewed', as in Figure 6, the mean can be seriously misleading as an indicator of the situation of most members of the population. It is pulled to the right by the limited number of individuals with high values. Such distributions are very common for variables such as expenditures, income, wealth, lengths of stay in hospital, etc.



Where the distribution is skewed in this way, the median value may be a better guide. It has the additional advantages of being easy to define and interpret – 'line up the population in order and identify the one in the middle' is relatively easy to explain to all stakeholders. The use of medians may be particularly important in analysis of data sets liable to errors that may include extreme outlier values (it is not unusual, for example, for an individual to accidentally add a zero to a number). Including these outliers in the calculation of the mean, which as indicated above is sensitive to large values, can give rise to biased results. The median is not

affected. An alternative approach sometimes used to deal with outliers is the 'trimmed mean'. For example, a 5 per cent trimmed mean removes the smallest and the largest 5 per cent of data values from the studied population and re-computes the mean using the reduced sample. This can be a useful approach but has the major disadvantage that it often appears somewhat arbitrary and increases the difficulty of explaining results to stakeholders.

Even the median is not much help in more complex distributions, such as the 'bi-modal' in figure 7. This type of distribution is often found where two subgroups are combined, for example patients in urban and rural hospitals. The most useful analysis in such cases involves identifying and separating the subgroups. This again emphasises the need to understand how variables are distributed in order to summarise them in ways that are helpful rather than misleading.



Measures of variation

In a population that has relatively limited *variability* in terms of the *variable* in which we are interested, a measures of location can be seen as reasonably 'representative' of the overall population and there is limited loss of information if we use this as a summary measure. If all those receiving treatment for malaria pay roughly the same amount, we lose little by describing the median or mean payment as 'the cost of malaria treatment'. On the other hand, if the amount paid for treatment of tuberculosis varies substantially across cases, use of the location measure alone would not be an appropriate summary of the data. We would be losing valuable information. Essentially, high variability implies that we have something to explain. Is the variability between urban and rural areas, between facilities within those areas, between patients who are insured and those who are not?

The variance is a measure of variability that considers all the data values relating to a study population. It asked the question 'how far away on average are the data values from the mean'? If we were considering length of stay, for example, and for most patients the stay in hospital was close to the mean, we would say that the distribution was relatively equal – with limited

variation ‘about the mean’. To calculate the variance we first determine the differences between each value and the mean, the ‘deviations from the mean’, square each of these differences, find their sum and divide by the number of values:

$$\text{variance} = \sum (x_i - m)^2 / n$$

Note that the size of the variance can often be determined by a limited number of deviations that are much larger than the rest. For example, if we have 100 inpatients and 48 stay in hospital for two days, 50 for three days and two for 20 days, the mean length of stay would be 2.86 days and the variance 6.24. Without the long-stay patients the variance would be 0.25. Simply using the mean and variance to summarise this data would lead to the incorrect interpretation that length of stay varied considerably, while in fact it would be much more useful to report that it appeared to be almost constant but for a few exceptional cases.

This effect results from the squaring of the deviations – squaring a large number produces a very large number. We saw above that the mean was influenced by outliers, but this effect is much more pronounced for the variance. The earliest use of the variance as an indicator of dispersion was in the field of scientific measurement and here it was considered an advantage that it was so influenced by outliers. These were either errors of measurement or extremely interesting data points – both of which required explanation. However, in social research it may often be an undesirable characteristic, first because the errors are typically of less interest (for instance caused simply by poor reporting), and second because it tends to focus analysis on attempts to explain the behaviour or experiences of a small number of individuals in what may be a fairly homogeneous population. Analysis of the differences in length of stay between small rural hospitals and the main teaching hospitals may be interesting, but from a policy perspective it would be probably be differences between the rural hospitals, if they were similar in most other respects, that would be more relevant.

The *standard deviation* is the square root of the variance. It has similar characteristics but also the advantage that it is expressed in the same unit of measurement as the original data. In the example above the standard deviation for all patients would be $\sqrt{6.24} = 2.5$ days and without the outliers it would be $\sqrt{0.25} = 0.5$ days. The standard deviation is a very important measure when we consider sampling from a population.

Another commonly used measure derived from the variance is the *coefficient of variation*. This is defined as:

$$\text{Coefficient of variation} = 100 \times \text{standard deviation} / \text{mean}$$

and provides a measure of *variation relative to the mean*. This is a useful statistic when comparing the variations of data sets that have substantially different means. For example, if we were to compare the variation in incomes for a population of hospital doctors as compared to a population of nurses, we would probably find, if we used any of the measures described above, that the former was considerably larger. However, this would be at least partly due to the generally higher incomes of doctors as compared to nurses. Essentially, the higher the incomes the more scope there is for variation. The coefficient of variation is not affected by this issue. Another advantage is that it is a pure ratio, which has no unit of measurement (because both the numerator and denominator have the same measurement unit). Thus, for example, we could directly compare the variation of incomes that are expressed in different currencies using this measure. It is also unaffected by inflation (as both numerator and denominator are equally affected), so we can consider if income variability has increased over time without worrying about the need to adjust using price indices.

Variances, standard deviations and coefficients of variation are widely used in statistical analysis. As with the mean, this is not because they are always the ‘best’ measures of variability (they can be easily interpreted for normally distributed variables but not for other

distributions) but mainly because they can be readily calculated and manipulated. For example, given the variances of two population subgroups it is easy to combine them to calculate the overall population variance. However, while they may have technical advantages, all these measures have serious limitations in terms of policy application, given that there is no way to provide a simple explanation of their derivation that would be understood by the great majority of stakeholders.

Alternative, more easily interpreted, measures of variation

Just as the median divides a data set into two halves, with 50 per cent above and 50 per cent below, *quartiles* can be used to divide it into four quarters with 25 per cent of the study population in each. There are three quartile values, usually denoted Q1, Q2 and Q3. If the data are listed in ascending order, Q2 is simply the median. Q1 is the median of the data points below the median and Q3 is the median of the points above the median. A useful and relatively easy to interpret and explain measure of variation is Q3–Q1, the *inter-quartile range*, which includes the 'middle 50 per cent' of a population.

When we have data on a reasonably large population (at least 100 members) we can extend the above to calculate *percentiles*. The p_{th} percentile divides the data into two parts with approximately p per cent having values less than the p_{th} percentile and (100 – p) per cent having values greater. Thus the 50_{th} percentile is the median, the 25_{th} percentile is the first quartile, etc. Other common percentiles, often used with incomes and expenditures, are the 20_{th} (which defines the first 'quintile group') and the 10_{th} (which defines the first 'decile group'). In describing inequality in income of doctors, for example, we might estimate the proportion of total incomes paid to the bottom and top decile groups.

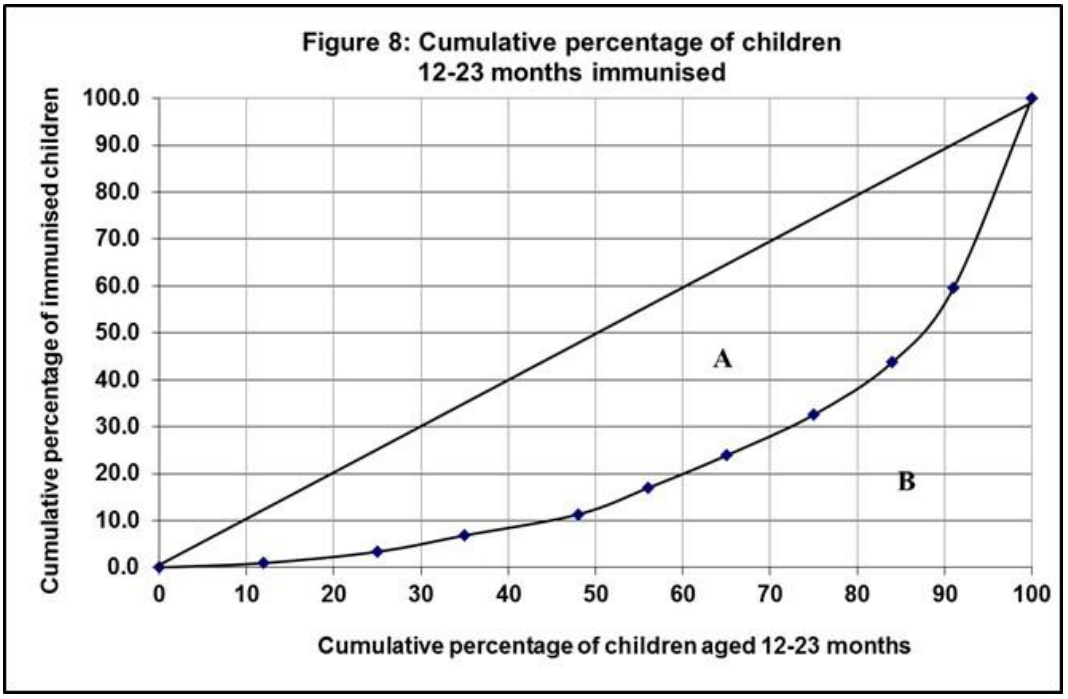
Precise formulae for calculating percentiles are available and used in computer statistical packages. However, because the number of data points is large, an approximation is usually perfectly adequate. For example, if there were 513 data points, it would be reasonable to calculate the quintiles as follows:

Q1 ≈ 513/5 ≈ 103
Q2 ≈ 2 x 513/5 ≈ 205
Q3 ≈ 3 x 513/5 ≈ 308
Q4 ≈ 4 x 513/5 ≈ 410

(rounding to the nearest integer) and use these to identify the four quintiles that divide the population into five quintile groups.

Lorenz curves and Gini coefficients

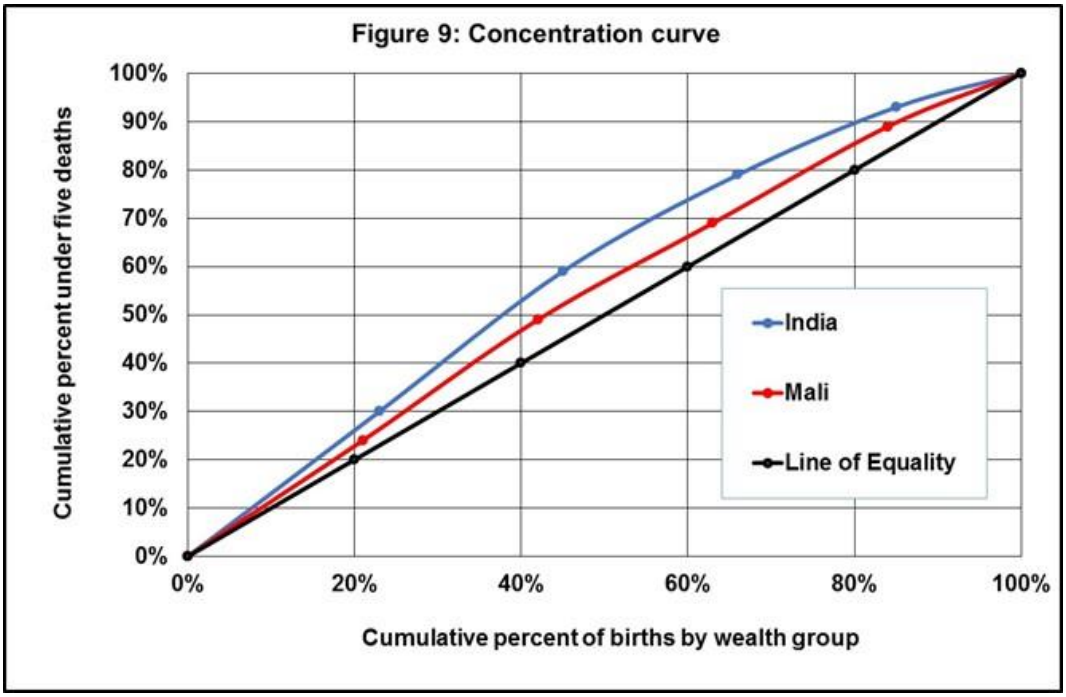
A *Lorenz curve* provides an alternative approach to measuring dispersion based on the cumulative distribution of a variable. The approach is often used for incomes or wealth distribution: for example, 'what share of the total income received by a given population goes to the 20 percent who receive the lowest incomes?', 'what share goes to the lowest 40 percent?', etc. By definition, the shares of each income group will increase as we move up through the income quintiles. However, the approach can also be used to analyse access to services. For example, we can ask 'what percentage of vaccinated children 12–23 months come from the sub-district with the lowest vaccination rate?', 'what percentage from the two sub-districts with the lowest rates?' etc. If we plot those percentages against the total percentage of children 12–23 months, cumulating over sub-districts, we obtain a Lorentz curve illustrating the variation in vaccination rates (figure 8).



The *Gini coefficient*, the ratio of area A to area (A+B), provides an alternative summary measure of variability that is often used when equity is a priority concern. If there is complete equality, the area A and the Gini coefficient equal 0. As inequality increases, area B becomes smaller and the Gini coefficient approaches 1. For any population the Gini coefficient will lie between 0 (complete equality) and 1 (complete inequality). However, there is no simple interpretation of other coefficient values. It is typically more useful, and certainly easier to communicate with stakeholders, if findings focus on the overall distribution illustrated by the Lorenz curve rather than exclusively on the Gini coefficient.

Concentration curves

Concentration curves (O'Donnell et al. 2008) can be seen as an extension of the Lorenz curve approach to include relationships between two variables. Typically, they show the cumulative percentage of a health status variable plotted against the cumulative percentage of a population ranked by socioeconomic status. For example, figure 9 shows the cumulative percentage of under five deaths plotted against the cumulative percentage of births, ranked by the wealth status of the households in which those births occurred (O'Donnell et al. 2008: Supplementary material). As might be expected the curves both lie above the line of equality because under-five mortality rates decrease with increases in wealth. The fact that the line for India is above that for Mali indicates that inequality in death rates was uniformly higher in India. The interpretation would have been more complicated if the lines for the two countries had crossed at some point. As with the Gini coefficient, it is possible to calculate a *concentration index* if a simple measure of inequality is desired.



Risk measures: Handle with care

Finally in this section, we can consider measures of 'risk'. These are widely used in health research but again are not well understood by the general population. For example, if the risk of contracting typhoid in an urban area over a one-year period is one in 10,000 and an intervention claimed to have reduced this to one in 20,000, this would probably be reported in local media as 'halving the risk of contracting typhoid'. There might then be a popular call for the intervention to be introduced at scale. However, this would disregard (a) the low risk prior to the intervention and (b) the likely estimation (sampling and non-sampling) errors when attempting to measure such rare events.

As another example, 'risk' and 'odds' are often confused. If we denote the risk of an event as P, then

Risk (P) of an event = number experiencing an event / population at risk.

Relative risk (P(A)/P(B)) = risk in group A / risk in group B.

Odds of an event = number experiencing / number not experiencing = P / (1-P)

Odds ratio = [PA/(1- PA)] / [PB/(1-PB)]

This distinction is particularly important when we consider reductions in risk, which are not equal to reduction in odds, for example:

Risk of malaria before intervention = P(B) = 0.5

Risk of malaria after intervention = P(A) = 0.1

Reduction in risk = 0.1/0.5 = 0.2

Reduction in odds = (0.1 / 0.9) / (0.5 / 0.5) = 0.11

The denominator problem

For the above calculation it is necessary to know the overall size of the population 'at risk'. Similarly, in clinical research one common summary statistic is the proportion or percentage of patients in the intervention and control groups whose condition improves. Calculating such proportions also requires data on the total membership and number improving in each group. In implementation studies, it is often very difficult to calculate or even reliably estimate these summary statistics because the *denominator* is not reliably known.

For example, we often have only a rough estimate of the number of children who should be immunised or could be sleeping under a net in a given district. Similarly, the catchment population of a facility or actual number of births over a period of time are often unknown. Because of this uncertainty, it is good practice to provide the estimates of both the numerator and denominator alongside any proportion, percentage or risk estimate and to indicate the sources used in the calculation.

4. Model building

As indicated above, we can regard analysis as essentially concerned with the explanation of variability. For example, why do the costs of care for a given condition vary between patients and/or between facilities? Can this be explained by variations in the severity of the condition or do other factors – patient gender or age, type of hospital, diverse treatment protocols, urban/rural location, etc. – play some role? In general terms, is variation in one variable associated with variation in another and does that association imply some causal relationship? As indicated above, this is an enormous topic to which we can only provide an introduction in this chapter. One excellent online course for those who wish to gain an in-depth knowledge of modelling techniques is that provided by the University of Bristol.

Subgroup analysis

During analysis we will often find that the outcomes of an intervention vary substantially between different subgroups of the target population. It would then seem natural to explore the possibility that the variables that define that subgroup may in some sense have caused the variation or alternatively been caused by it. However, subgroup analysis can be a contentious issue if the subgroups are not predefined. Large data sets containing multiple variables, whether from routine data systems or sample surveys, will often tend to exhibit patterns that may arise purely by chance. The term 'data mining' is often used to describe the process of exploring data sets to discover apparent relationships that may be of interest. It is generally regarded as useful when used to formulate new hypotheses but requiring great caution to avoid being misled – if you search through all possible combinations of variables in a data set containing perhaps 50 or more, there is a high probability that you will stumble across a number of apparent relationships purely by chance.

This is a particular issue in implementation research, where the emphasis is on detailed understanding of processes and an acceptance that relationships between inputs and outcomes may often be mediated by other variables. For example, suppose we find that the prevalence of chronic illness varies by age group and sex as in Table 4.1. If we obtained these findings using a rapid survey as defined above, we must first consider if the sample size was sufficient to provide reasonably reliable estimates of prevalence in each cell of the table.

Table 8: Prevalence of chronic Illness by sex and age group

	Percentage reporting at least one chronic illness	
Age group	Males	Females

15–24	0.55	0.80
25–44	1.79	4.01
45–64	4.91	12.28
65	12.86	20.00
All	1.77	4.25

One of these relationships, between chronic illness and age group, is long-established and well understood – as bodies age they tend to accumulate defects that are linked to various types of chronic illness. The other, the higher prevalence of chronic illness in women in all age groups, is less easily explained. It would not be correct to leap to the conclusion that women are naturally more prone to chronic illness than men. We might consider a range of possible hypotheses exploring, for example, the influence of childbirth, activities mainly undertaken by women and men, whether women were more likely to report illness than men or less likely to receive treatment for acute illnesses that then became chronic conditions. We might be able to examine some of these hypotheses by further analysis of this or other data sources, or by undertaking qualitative studies such as described in Chapters 9 and 10. The key requirement for researchers is to ensure that they have convincing evidence before advancing one or other theories to explain such observations.

Controlled and confounding variables

We sometimes describe such an analysis as one that assesses the relationship between inputs and outcomes *controlling* for other factors. Typically we know that in practice a very large number of other factors may influence this relationship, for example occupation, level of education, socioeconomic status, household size, type of dwelling, rural/urban location, etc. As indicated in Chapter 5, random allocation of subjects to subgroups would allow us to argue that the potentially 'confounding' effects of such variables average out. That will almost never be possible in the type of interventions we are considering and we therefore need to find some way to allow for these effects.

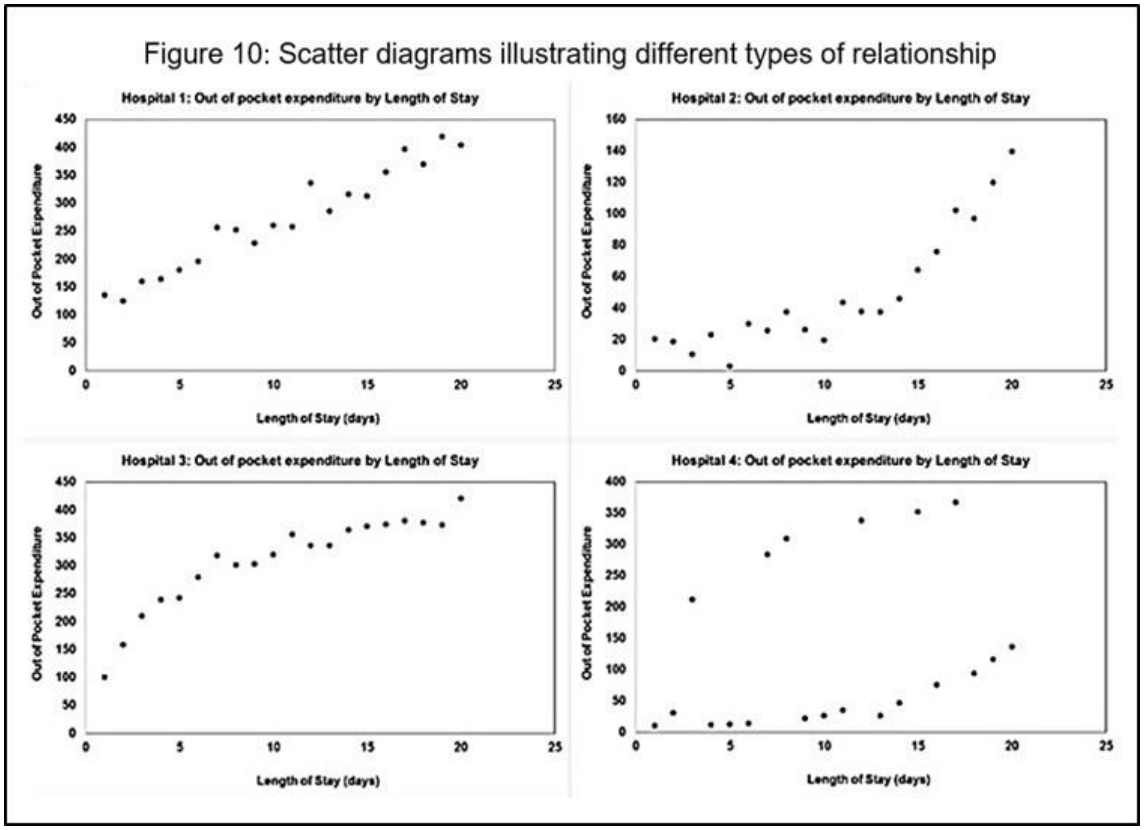
Cross-tabulating by all such factors, even with an apparently large data set, would almost always result in the numbers in most cells being too small to permit analysis. One alternative is to construct a model of the relationship between outcome and inputs that takes into account the effects of other confounding variables. This typically involves very strong assumptions both as to the nature of the multiple relationships between these variables and their individual distributions – assumptions that are often not adequately recognised or tested. As discussed above, it can be argued that the explicit intention to change implementation practice and influence a wide range of stakeholders requires implementation researchers to set higher standards than those conducting more exploratory research.

Models and presentation of findings

Models are typically very simplified versions of reality and we should be very cautious in their interpretation. In particular we should recognise that most stakeholders will typically have little understanding of the assumptions underlying those models. Modelling may be useful to explore our data but should be seen as an intermediary stage in the generation of findings that can be readily comprehended and interpreted. As with the step from distributions to summary measures, we should proceed cautiously and try to ensure that we understand the underlying form of the relations that we are trying to model.

Just as we can understand a great deal about individual variables by examining frequency distribution, much can be learned about two-way relationships from simple scatter diagrams.

Figure 10 illustrates that such relationships can take a great variety of forms. It shows possible scatter diagrams of a sample of out-of-pocket payments for inpatient care plotted against length of stay in four hypothetical hospitals, with very different policies on fees and with patients covered by different types of health insurance. Only the first might reasonably be assumed to reflect a linear relationship.



Our common, often unspoken, assumption of linear relationships between variables is frequently not only incorrect but may in many instances be simply irrational. The number of cases of tuberculosis identified cannot increase linearly with expenditure on case finding, because finding cases will become increasingly more difficult once the 'low-hanging fruit' have been identified. Hospital net revenues cannot increase linearly with the number of inpatients, because the marginal cost of an inpatient will decline as the number increases.

The *linear regression model*

By far the most common approach to model building is the use of some form of linear model and we can use this to illustrate modelling possibilities and limitations. The simple linear regression model is illustrated in figure 11. It is usually expressed by an equation of the form:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

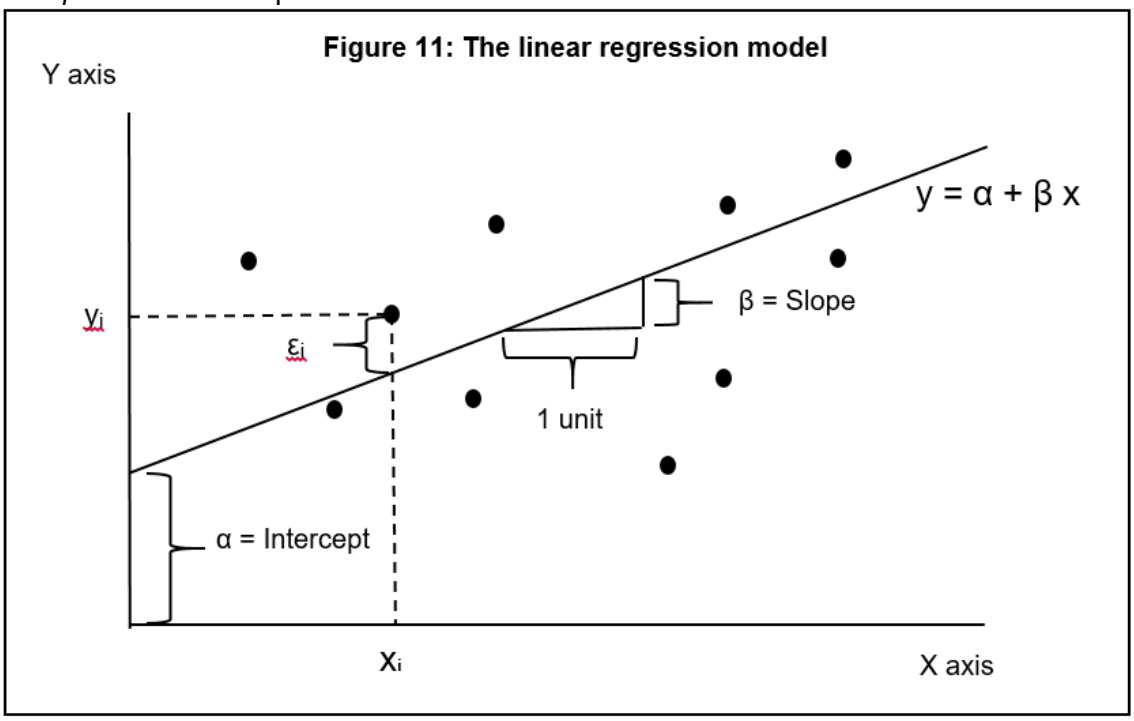
Where:

- y_i is the value of a response (outcome) variable for the i^{th} observation.
- x_i is the value of an explanatory (input) variable for the i^{th} observation.
- ε_i is the value of a random error term for the i^{th} observation.

This model is the equation of a straight line where:

α is the intercept and
β is the slope.

Figure 11: The linear regression model



Regression assumptions

The following strong assumptions, which many researchers choose to ignore, are required in order to argue that a regression model is appropriate:

- The relationship between X and Y is linear;
- The values of the independent variable X are assumed fixed (not random) – the only randomness in the values of Y comes from the error term ε;
- The errors ε_i are uncorrelated (independent) in successive observations;
- The errors ε_i are normally distributed with mean 0, variance σ^2 [$\varepsilon \sim N(0, \sigma^2)$].

We choose α and β such that the sum of squares of deviations from the regression line [Σ(observed value of y_i at x_i – predicted value of y_i at x_i)^2] is minimised. This is known as the error sum of squares (ESS) about the regression. ESS/(n-2), where n is the number of observations, provides an unbiased estimate of σ^2.

Variance components and the coefficient of determination

The error sum of squares can be compared to the sum of squared deviations about the mean (TSS) to see how much of this can be 'explained' by fitting the regression line. The division of the sum of squares about the mean (TSS) into two 'components', a regression sum of squares (RSS) and an error or residual sum of squares (ESS), is the simplest example of the 'variance components' approach to model building, which plays a central role in multilevel modelling. If we write $\hat{y}_i$ for the value of y_i predicted by the regression equation and $\overline{y}$ for the mean value of the observations, we can express the deviation of y_i from $\overline{y}$ as the sum of the deviation of y_i from $\hat{y}_i$ plus the deviation of $\hat{y}_i$ from $\overline{y}$:

$$(y_i - \overline{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

If we square both sides and sum over all values of y_i, we can derive the following result:

$$\Sigma(y - \overline{y})^2 = \Sigma (y - \hat{y})^2 + \Sigma(\hat{y} - \overline{y})^2$$

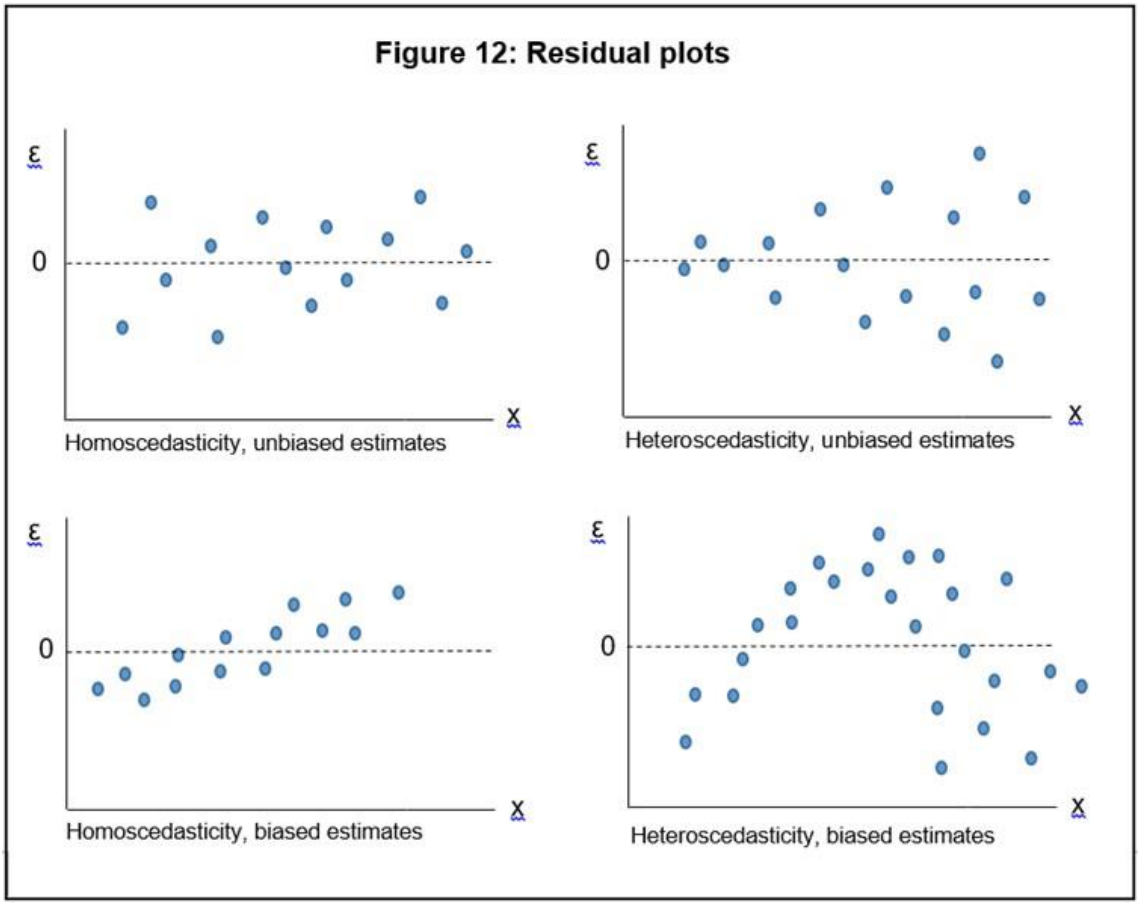
$$TSS = ESS + RSS$$

The ratio $R^2 = RSS/TSS = 1 – ESS/TSS$ is known as the *coefficient of determination* and is often loosely described as the proportion of variance 'explained' by the model.

Residuals

The use of the phrase 'explained by the regression line' should not be taken literally. It refers simply to the above ratio, which is interesting only if all the assumptions made in defining our model are correct – this is rarely the case. One way of exploring the value of our model is to look at the deviations of observations from the value 'predicted' by our regression line – the 'residuals'. We do this using a scatter plot with the explanatory variable (X) on the horizontal axis and the residuals on the vertical axis as in figure 12.

Figure 12: Residual plots



Specification errors

As indicated above, we use models to allow for the effects of a variety of potentially confounding variables. To do this we construct a multiple regression model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \delta_1 z_1 + \delta_2 z_2 + \ldots + \varepsilon$$

where:

y is the response variable

x_i are known explanatory variables

z_i are known confounding variables

However, one often intractable issue is that there are typically a range of factors which we have either ignored or cannot measure. The true model should be written:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \delta_1 z_1 + \delta_2 z_2 + \ldots + \gamma_1 c_1 + \gamma_2 c_2 \ldots + \varepsilon$$

where*:*

y is the response variable

x_i are known explanatory variables

z_i are known confounding variables

c_i are unknown confounding variables

ε is an error term.

This is described as a 'specification' error. In general, omitting such variables from the model has serious implications in terms of undermining the basic assumptions identified above.

Statistical inference in regression models

As discussed above, with the widespread availability of statistical software, it is expected that, where data have been collected using probability sampling, all estimates will be accompanied by estimated error margins. For example, in the simplest case of a random sample of size n we know that we can estimate 95 per cent confidence limits for a population mean as:

$$\text{sample mean} \pm 2\, s/\sqrt{n}$$

Where the term $s/\sqrt{n}$ is the 'standard error' of estimation (s.e.). It was also indicated above that for other probability sampling designs the formula for the standard error will vary, but the formula remains the same and can be extended to other statistics:

$$\text{estimated value} \pm 2 \text{ s.e.}$$

Multilevel modelling

Given that regression estimates will also require to be accompanied by error margins, we again have to address the issue that most surveys will use a sample design that involves cluster sampling at one or more levels. For example the DHS surveys typically involve:

- stratification by states/provinces and then by urban and rural areas;
- a PPS cluster sample of enumeration areas within each stratum;
- a systematic sample of 30 households per cluster.

As discussed above, failure to allow for the larger sampling errors associated with cluster sampling can result in the confidence limits for estimates that are too narrow, and incorrect assessment of tests of statistical significance. With a cluster sample the error sum of squares has two components:

$$\text{ESS (about mean)} = \text{ESS (between clusters)} + \text{ESS (within clusters)}$$

If the random sample formula for the error sum of squares is used, estimates of model parameters may be unbiased but estimated confidence limits will typically be far too narrow – that is, we will be substantially exaggerating the precision of our estimates. Multilevel modelling (Rashbash et al 2012, Diez Roux 2009, Goldstein 1999) explicitly builds the variation between clusters into the model and estimates the between-cluster variation.

Random intercept model

We can allow for between-cluster variation by formulating the model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + [u_j + \varepsilon_{ij}]$$

Where:

y_{ij} is the value of y for individual i in cluster j
x_{ij} is the value of x for individual I in cluster j
u_j is the deviation of the mean of cluster i from the global mean

We then have two random variables in the equation:

$u_j \sim N(0, \sigma_u)$
$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon)$

And can obtain correct estimates for: β_0, β_1, σ_u^2, σ_ε^2

Note that sometimes the variation between clusters may itself be of interest. For example, if we have clustered by health facility, we can estimate the proportion of total variability 'explained' by between-facility variation (Lopez-Cevallos and Chi 2009).

Sample survey software

Multilevel modelling can similarly be used in a wide range of other contexts where relationships exist between different 'levels' of a health system (district, facility, doctor, patient, etc.). Using modern survey analysis software it is relatively straightforward to describe even complex sampling design and obtain appropriate parameter estimates. These packages include the 'usual suspects': SAS, STATA, SPSS and some more specialist software such as MLwiN, etc. They can all readily address the most common health survey designs involving cluster sampling and unequal sampling probabilities.

Example: A multilevel analysis of self-reported tuberculosis disease in a nationally representative sample of South Africans was undertaken based on the 1998 DHS (Harling et al. 2008, Harling 2006)). Individual and household-level demographic, behavioural and socioeconomic risk factors were taken from the DHS; data on community-level socioeconomic status (including measures of absolute wealth and income inequality) were derived from the 1996 national census.

Of the 13,043 DHS respondents, 0.5 per cent reported having been diagnosed with tuberculosis disease in the past 12 months and 2.8 per cent reported having been diagnosed with tuberculosis disease in their lifetime. In a multivariate model adjusting for demographic and behavioural risk factors, tuberculosis diagnosis was associated with cigarette smoking, alcohol consumption and low body mass index, as well as a lower level of personal education, unemployment and lower household wealth. In a model including individual- and household-level risk factors, high levels of community income inequality were independently associated with increased prevalence of tuberculosis.

The multilevel analytic approach was seen as allowing for the differentiation between community- and individual-level mechanisms in the relationship between socioeconomic status and tuberculosis. Furthermore, these data allow strong inferences to be drawn regarding risk factors for tuberculosis disease across the country: a nationally representative cross-sectional survey provided evidence on individual and household

characteristics, while South African census data provided robust estimates of the true community-level socioeconomic characteristics across the nation

References

Bennett, S.; Woods, T.; Liyanage, W.M. and Smith D.B. (1991) 'A Simplified General Method For Cluster-Sample Surveys of Health in Developing Countries', *World Health Statistics Quarterly* 44: 98–106, http://apps.who.int/iris/bitstream/10665/47585/1/WHSQ_1991_44%283%29_98-106_eng.pdf?ua=1 (accessed 30 March 2015)

Diez Roux, A. V. (2009). Next steps in understanding the multilevel determinants of health. Journal of Epidemiology and Community Health 2008(62):957-959. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/64000/diez%20roux_next%20steps%20in%20understanding%20the%20multilevel%20determinants%20of%20health.pdf;sequence=1

Eurostat (2007) Handbook on Data Quality Assessment Methods and Tools. Wiesbaden: European Commission, http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK ON DATA QUALITY ASSESSMENT METHODS AND TOOLS_I.pdf (accessed 30 March 2015).

Goldstein, H. (1999) *Multi-level Statistical Models*. London: Arnold. http://www.ats.ucla.edu/stat/examples/msm_goldstein/goldstein.pdf

Goodson, J.L., Kulkarni, M.A., Eng, J.L.V., Wannemuehler, K.A., Cotte, A.H., Desrochers, R.E., Randriamanalina, B. and Luman, E.T. (2012). Improved Equity in Measles Vaccination from Integrating Insecticide-treated Bednets in a Vaccination Campaign, Madagascar. *Tropical Medicine and International Health* 17(4): 430–37

Harling, G., Ehrlich, R. and Myer, L. (2008). The Social Epidemiology of Tuberculosis in South Africa: A Multilevel Analysis. *Social Science and Medicine* 66: 492–505. https://www.researchgate.net/publication/5923792_The_social_epidemiology_of_tuberculosis_in_South_Africa_A_multilevel_analysis

Harling, G. The Social Epidemiology of Tuberculosis in South Africa: A Multilevel Analysis. Thesis. Master of Public Health. University of Cape Town. https://open.uct.ac.za/bitstream/handle/11427/9326/thesis_hsf_2006_harling_g.pdf?sequence=1

Lemeshow, S. and Robinson, D. (1985). Surveys to Measure Programme Coverage and Impact: A Review of the Methodology Used by the Expanded Programme on Immunization. *World Health Statistics Quarterly* 38.1: 65–75

Lin, Mingfeng, Henry C. Lucas Jr, Galit Shmueli (2013). Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem. Information Systems Research 24(4):906-917. http://eprints.exchange.isb.edu/280/1/isre%252E2013%252E0480.pdf

Lopez-Cevallos, D.F. and Chi, C. (2009). Health Care Utilization in Ecuador: A Multilevel Analysis of Socio-Economic Determinants and Inequality Issues. *Health Policy and Planning* 25(3):209–18. http://heapol.oxfordjournals.org/content/25/3/209.full.pdf+html

Myatt, M.; Feleke, T.; Sadler, K. and Collins, S.(2005) 'Field Trial of a Survey Method for Estimating the Coverage of Selective Feeding Programmes', *Bulletin of the World*

Health Organisation 83(1):20–6. www.who.int/bulletin/volumes/83/1/20arabic.pdf (accessed 30 March 2015)

O'Donnell, O.; van Doorslaer, E.; Wagstaff, A. and Lindelow, M. (2008). *Analyzing Health Equity Using Household Survey Data*. Washington, DC: The World Bank. http://siteresources.worldbank.org/INTPAH/Resources/Publications/459843-1195594469249/HealthEquityFINAL.pdf (accessed 30 March 2016)

Prudhon, C. and Spiegel, P.B. (2007). A Review of Methodology and Analysis of Nutrition and Mortality Surveys Conducted in Humanitarian Emergencies from October 1993 to April 2004. *Emerging Themes in Epidemiology*, www.ete-online.com/content/pdf/1742-7622-4-10.pdf (accessed 30 March 2015)

Rashbash, Jon, Fiona Steele, William J. Browne and Harvey Goldstein (2012). *A User's Guide to MLwiN*. Centre for Multilevel Modelling, University of Bristol. http://www.bris.ac.uk/media-library/sites/cmm/migrated/documents/manual-web.pdf

SMART (2005). *Measuring Mortality, Nutritional Status, and Food Security in Crisis Situations: SMART Methodology*. The SMART Initiative, www.smartindicators.org/SMART_Protocol_01-27-05.pdf (accessed 30 March 2015)

Tipping, Jill and Malcolm Segall. (1996) How to do (or not to do) . . . Using a longitudinal illness record to study household health care decision-making in rural communes of Viet Nam. *Health Policy and Planning* 11(2): 206-211. http://heapol.oxfordjournals.org/content/11/2/206.full.pdf

Turner, A.G.; Magnani, R.J. and Shuaib, M. (1996). A Not Quite as Quick but Much Cleaner Alternative to the Expanded Programme on Immunization (EPI) Cluster Survey Design. *International Journal of Epidemiology* 25.1: 198–203

Walker, N.; Bryce, J. and Black, R.E. (2007). Interpreting Health Statistics for Policymaking: the Story Behind the Headlines. *Lancet* 369: 956–63

[i] Note that confidence limits (and statistical significance tests) are most useful when sample sizes are relatively limited. With very large samples (~ 10,000) all estimates become extremely precise (and all test significant) (Lin et al. 2013).