PART F

STUDY DESIGN, ANALYSIS AND INTERPRETATION

Our aim in this final part of the book is to facilitate the overall planning and conduct of an analysis, and to cover general issues in the interpretation of study results. We start in Chapter 34 by explaining how to link the analysis to study design. We include guides to aid the selection of appropriate statistical methods for each of the main types of study, and draw attention to design features that influence the approach to analysis.

In the next three chapters, we address three different issues related to interpretation of statistical analyses. Chapter 35 tackles the calculation of sample size, and explains its fundamental importance in the interpretation of a study's results. Chapter 36 covers the assessment and implications of measurement error and misclassification in study outcomes and exposures. Chapter 37 outlines the different measures that are used to assess the impact of an exposure or of a treatment on the amount of disease in a population.

Finally, Chapter 38 recommends general strategies for statistical analysis.

This page intentionally left blank

CHAPTER 34

Linking analysis to study design: summary of methods

34.1 Introduction

34.2 Randomized controlled trials
- Analysis plan
- Participant flow
- Analysis of baseline variables
- Intention to treat analysis
- Adjustment for baseline variables
- Subgroup analyses
- Crossover trials
- Cluster randomized trials
- Choosing the statistical method to use

34.3 Other designs to evaluate interventions

34.4 Longitudinal and cross-sectional studies
- Choosing the statistical method to use
- Types of sampling scheme and their implications

34.5 Case–control studies
- Analysis of unmatched case–control studies
- Analysis of matched case–control studies
- Interpretation of the odds ratio estimated in a case–control study

34.1 INTRODUCTION

The main focus of this book is on the statistical methods needed to analyse the effect of an exposure (or treatment) on an outcome. In previous parts, we have categorized these methods according to the types of outcome and exposure (or treatment) variables under consideration. These are summarized in the inside covers of the book. In this chapter, we now look more generally at how to link the analysis to the study design. In particular, we:

- summarize the range of methods available for each of the following:
  - randomized controlled trials;
  - other designs to evaluate the impact of an intervention;
  - cross-sectional and longitudinal studies;
  - case–control studies;
- highlight the key elements of each design that determine the choice of statistical method(s);
- discuss any specific issues that need to be considered in the interpretation of the results;
- draw attention to design-specific considerations that need to be built into the analysis plan, in addition to the general strategies for analysis outlined in Chapter 38.

Detailed discussions of the design of different types of study are outside the scope of this book, but are available in the following textbooks:

Clinical trials: Friedman *et al*. (1998) and Pocock (1983)
Interventions in developing countries: Smith & Morrow (1996)
Cluster randomized trials: Donner & Klar (2000) and Ukoumunne *et al*. (1999)
Case–control studies: Breslow & Day (1980) and Schlesselman & Stolley (1982)
General epidemiology: Gordis (2000), Rothman (2002), Rothman & Greenland (1998) and Szklo & Nieto (2000)

34.2 RANDOMIZED CONTROLLED TRIALS

Randomized controlled trials (RCTs) provide the best evidence on the effectiveness of treatments and health care interventions. Their key elements are:

- The comparison of a group receiving the treatment (or intervention) under evaluation, with a control group receiving either best practice, or an inactive intervention.
- Use of a **randomization** scheme to ensure that no systematic differences, in either known or unknown prognostic factors, arise during allocation between the groups. This should ensure that estimated treatment effects are not biased by confounding factors (see Chapter 18).
- **Allocation concealment**: successful implementation of a randomization scheme depends on making sure that those responsible for recruiting and allocating participants to the trial have no prior knowledge about which intervention they will receive. This is called allocation concealment.
- Where possible, a **double blind design**, in which neither participants nor study personnel know what treatment has been received until the 'code is broken' after the end of the trial. This is achieved by using a **placebo**, a preparation indistinguishable in all respects to that given to the treatment group, except for lacking the active component. If a double-blind design is not possible then outcome assessment should be done by an investigator blind to the treatment received.
- An **intention to treat analysis** in which the treatment and control groups are analysed with respect to their random allocation, regardless of what happened subsequently (see below).

It is crucial that RCTs are not only well designed but also well conducted and analysed if the possibility of systematic errors is to be excluded. It is also essential that they are reported in sufficient detail to enable readers to be able to assess the quality of their conduct and the validity of their results. Unfortunately, essential details are often lacking. Over the last decade concerted attempts to improve the quality of reporting of randomized controlled trials resulted in the 1996 **CONSORT statement** (Begg *et al*., 1996), with a revised version in 2001 (Moher *et al*., 2001). CONSORT stands for **CON**solidated **S**tandards **O**f **R**eporting **T**rials. The statement consists of a prototype flow diagram for summarizing the different phases of the trial, with the numbers involved in each (Figure 34.1), and a checklist

of items that it is essential for investigators to report (Table 34.1). Details of its rationale and background together with a full description of each component can be found on the website http://www.consort-statement.org/.

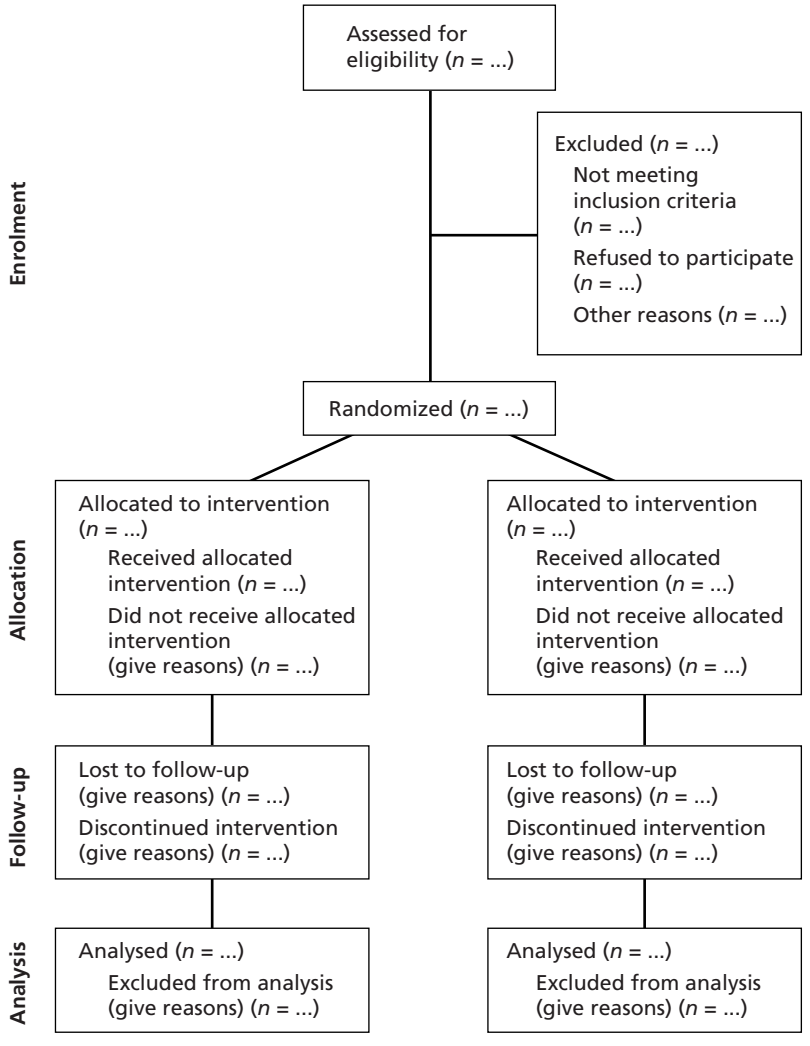


Fig. 34.1 Revised template of the CONSORT diagram showing the flow of participants through each stage of a randomized trial, reprinted with permission of the CONSORT group.

Analysis plan

In this section we will focus in particular on the features of the CONSORT statement pertinent to the **analysis plan**, key stages of which are outlined in

Table 34.1 The revised CONSORT statement for reporting randomized trials: checklist of items to include when reporting a randomized trial, reprinted with permission of the CONSORT group.

Paper section and topic	Item no.	Descriptor
TITLE AND ABSTRACT	1	How participants were allocated to interventions (e.g. 'random allocation', 'randomized', or 'randomly assigned')
INTRODUCTION		
Background	2	Scientific background and explanation of rationale
METHODS		
Participants	3	Eligibility criteria for participants and the settings and locations where the data were collected
Interventions	4	Precise details of the interventions intended for each group and how and when they were actually administered
Objectives	5	Specific objectives and hypotheses
Outcomes	6	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (e.g. multiple observations, training of assessors, etc.)
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules
Randomization:		
Sequence generation	8	Method used to generate the random allocation sequence, including details of any restriction (e.g. blocking, stratification)
Allocation concealment	9	Method used to implement the random allocation sequence (e.g. numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned
Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. When relevant, how the success of blinding was evaluated
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analyses and adjusted analyses
RESULTS		
Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analysed for the primary outcome. Describe protocol deviations from study as planned, together with reasons
Recruitment	14	Dates defining the periods of recruitment and follow-up
Baseline data	15	Baseline demographic and clinical characteristics of each group
Numbers analysed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was by 'intention-to-treat'. State the results in absolute numbers when feasible (e.g. 10/20, not 50%)
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (e.g. 95% confidence interval)

(*continued*)

Table 34.1 (*continued*)

Paper section and topic	Item no.	Descriptor
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory
Adverse events	19	All important adverse events or side effects in each intervention group
DISCUSSION		
Interpretation	20	Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes
Generalizability	21	Generalizability (external validity) of the trial findings
Overall evidence	22	General interpretation of the results in the context of current evidence

Table 34.2. Although CONSORT has been designed primarily for two-group parallel designs, most of it is also relevant to a wider class of trial designs, such as equivalence, factorial, cluster and crossover trials. Modifications to the CONSORT checklist for reporting trials with these and other designs are in preparation.

Table 34.2 Outline of analysis plan for a randomized controlled trial.

1. Complete flow diagram showing number of participants involved at each phase of the trial
2. Summarize baseline characteristics of trial population
3. Compare treatment groups with respect to baseline variables – focus on subset of variables thought to be associated with main outcome(s). Avoid formal tests of the null hypothesis of no between-group differences, since the null hypothesis *must* be true if the randomization was done properly
4. Conduct *simple* analysis of main outcome(s) by *intention to treat*
   (a) Present the estimated effect of treatment together with a CI and test of the null hypothesis of no treatment effect
   (b) Consider *sensitivity analyses* examining the possible effect of losses to follow-up, if these might affect the treatment effect estimate
5. Repeat analysis including adjustment for baseline variables if appropriate
6. Carry out any subgroup analyses if there is an *a priori* justification
7. Analyse side effects and adverse outcomes
8. Analyse secondary outcomes

Participant flow

An important first stage of the analysis is to work out the flow of the number of participants through the four main phases of the trial: enrolment, allocation to intervention groups, follow-up and analysis, as shown in Figure 34.1. In particular, it is important to note the number excluded at any stage and the *reasons* for their exclusion. This information is crucial for the following reasons:

- Substantial proportions lost at any stage have important implications for the **external validity** of the study, since the resulting participants may no longer be representative of those eligible for the intervention.

- Any imbalance in losses between treatment groups has implications for the **internal validity** of the study, since they may lead to non-random differences between the treatment groups which could influence the outcome.
- Knowing the difference between the number allocated to receive an intervention, and number who actually received it (and/or adequately adhered to it), is important for the interpretation of the estimated effect, as explained below under 'intention to treat analysis'.

Analysis of baseline variables

'Baseline' information collected at enrolment is used in the analysis of a trial in the following ways:

1 To describe the characteristics of the trial participants, which is essential for assessing the generalizibility of the results.
2 To demonstrate that the randomization procedure has successfully led to comparability between trial groups.
3 To adjust treatment effects for variables strongly related to the outcome (see below).
4 To carry out subgroup analysis (see below).

In their review, 'Subgroup analysis and other (mis)uses of baseline data in clinical trials', Assmann *et al*. (2001) found that the first two objectives are often confused, and that the approach to the second is often methodologically flawed. They recommend that:

- A general and detailed description is given of the trial participants, but that the analysis of comparability between groups should be restricted to a few variables known to be strong predictors of the primary outcome(s).
- Significance tests for baseline differences are inappropriate, since any differences are either due to chance or to flawed randomization. In addition, a non-significant imbalance of a strong predictor will have more effect on the results than a significant imbalance on a factor unrelated to the outcome.

Intention to treat analysis

In an '**intention to treat**' analysis, participants are analysed according to their original group assignment, whether or not this is the intervention they actually received, and whether or not they accepted and/or adhered to the intervention. Alternatively, analysis can be based on actual intervention received, with criteria for exclusion if inadequate adherence to the intervention was achieved. This is sometimes known as a '**per protocol**' analysis. The primary analysis of a RCT should always be an intention to treat analysis, since it avoids the possibility of any bias associated with loss, mis-allocation or non-adherence of participants. For example, consider a placebo-controlled trial of a new drug with unpleasant side-effects. If the sickest patients are unable to take the new drug, they may withdraw from the assigned treatment. Such problems will not affect the

placebo group, and therefore a per-protocol analysis would give a biased result by comparing the less sick patients in the drug group with all patients in the placebo group.

If there is a substantial difference between those allocated to receive an intervention and those who actually receive it (and adequately adhere to it), then we recommend that in addition analyses are carried out adjusting for actual treatment received, and that the results are compared with the intention to treat analysis. A valid method to correct for non-adherence to treatment in randomized trials was developed by Robins and Tsiatis (1991), but has not been widely used in practice, partly because it is conceptually difficult. However, software implementing the method is now available (White *et al*. 2002). It is important to report the numbers involved, and the reasons for the losses in order to assess to what extent the intention to treat analysis may lead to an underestimate of the efficacy of the intervention under ideal circumstances, and to what extent the per protocol analysis may be biased.

Adjustment for baseline variables

The analysis of the main outcome(s) should always start with simple unadjusted comparisons between treatment groups. For most randomized controlled trials, this is all that should be done. We recommend adjustment for covariates measured at baseline *only* in the following circumstances:

- Where there is clear a priori evidence about which baseline factors are likely to be strongly related to the outcome. Even where strong predictors exist, adjustment for them in the analysis is only necessary if the outcome is numerical.
- In particular, where the outcome is numerical and where a baseline measurement of it has been taken. An example would be a trial of an anti-hypertensive drug, where blood pressure is measured at baseline and following treatment. In this case the baseline measurement is likely to be strongly correlated with the outcome, and including it as a covariate in the analysis improves the precision of the treatment effect (see Section 29.8). Note that this is a better approach than taking differences from the baseline as the outcome variable, since the latter tends to overcorrect (see Snedecor & Cochran, 1989).
- Where the trial is sufficiently small that an imbalance sufficiently large to bias the treatment effect is possible. (Such a situation may occur in *cluster-randomized* trials; see below.)

Note that:

- The decision concerning covariates should *not* be made on the basis of statistically significant differences between the treatment groups at baseline, although this is often the practice (see above discussion on analysis of baseline variables).
- It is not necessary to adjust for centre in multi-centre studies, unless it is a strong predictor of outcome and the proportion of patients in the treatment group differs between centres.

Subgroup analyses

In their review, Assmann *et al*. (2001) found that the use of subgroup analyses is widespread in clinical trials, and often flawed. The choice of subgroups used is often not justified, their analysis is often inadequate and their results are given undue emphasis. They note that of all the problems that have been identified in the conduct, analysis and reporting of clinical trials, subgroup analysis remains the most over-used and over-interpreted.

- Subgroup analyses should only be conducted if there is a clear *a priori* reason to expect the treatment effect to differ between different groups of patients, such as between males and females, or between different age groups. Only a few predefined subgroups should be considered and analysis restricted to the main outcomes.
- They should include formal tests for interaction, as described in Section 29.5, and should not be based on inspection of subgroup *P*-values. A particularly common error is to assume that a small *P*-value in one subgroup, but not in another, provides evidence that the treatment effect differs between the subgroups. If the subgroups are of different sizes then this situation may arise even if the subgroup treatment effects are identical!
- In addition, in multi-centre trials it may be useful to present the results by centre as well as overall, as a means of data quality and consistency checking between centres. The results of such analyses may be presented in a forest plot (see Chapter 32). However, this should not lead to undue emphasis being placed on any apparent differences seen, unless these are supported by strong evidence supporting their plausibility.

Crossover trials

Crossover trials are trials in which both treatments (or the active treatment and the placebo control) are given to each patient, with the order of allocation decided at random for each patient. They are suitable in situations such as trials of analgesics for pain relief or therapies for asthma, where outcomes can be measured at the end of successive time periods, and where there is unlikely to be a carry-over effect of the first treatment into the period when the second treatment is being given. To address this issue, such trials may incorporate a 'washout' period between the periods when treatments under investigation are administered.

The main advantage of crossover trials is that by accounting for between-patient variability in the outcome they may be more efficient than a corresponding trial in which treatments are randomly allocated to different individuals (**parallel group trial**). The analysis of such trials should take account of the design by using methods for paired data. For numerical outcomes, the *mean difference* between each patient's outcomes on the first and second treatment should be analysed (see Section 7.6), and the standard deviation of the mean differences should always be reported, to facilitate meta-analyses of such trials, or of trials using both crossover

and parallel group designs. For binary outcomes, methods for matched pairs should be used (see Chapter 21).

Cluster randomized trials

The development, and the major use, of RCTs is in the evaluation of treatments or medical interventions (such as vaccines) applied at the individual level. In recent years, however, the use of RCTs has extended to the evaluation of health service and public health interventions. This has led to the development of **cluster randomized trials**, in which randomization is applied to clusters of people rather than individuals, either because of the nature of the intervention, or for logistical reasons. Some examples are:

- Evaluation of screening of hypertension among the elderly in the UK in which the unit of randomization was the GP practice.
- Evaluation of the impact on HIV transmission in Tanzania of syndromic management of sexually transmitted diseases, where the unit of randomization was STD clinics and their catchment populations.
- Evaluation in Glasgow of the impact on adolescent sexual behaviour of a sex education programme delivered through school, in which the schools were the unit of randomization.
- Evaluation in Ghana of the impact of weekly vitamin A supplementation on maternal mortality, where the unit of randomization is a cluster of about 120 women, the number that a fieldworker can visit in a week.

Three essential points to note are that:

1 Any clustering in the design must be taken into account in the analysis, as described in Chapter 31.

2 Because the number of clusters is often relatively small, a cluster randomized design may not exclude the possibility of imbalance in baseline characteristics between the treatment and control groups and careful consideration should be given to measurement of known prognostic factors at baseline and whether it is necessary to adjust for their effects in the analysis.

3 A cluster randomized trial needs to include more individuals than the corresponding individually randomized trial. Sample size calculations for cluster randomized trials are described in Chapter 35.

Choosing the statistical method to use

Table 34.3 provides a guide to selecting the appropriate statistical method to use. It shows how this depends on:

- the type of outcome;
- whether adjustment for baseline variables is needed;
- whether subgroup analyses are being conducted;
- and, in the case of survival outcomes, whether the proportional hazards assumption is satisfied.

Table 34.3 Analysis of clinical trials/intervention studies: summary of methods.

	Type of outcome			
	Numerical	Binary	Rate	Survival time
Data displays	Mean outcome in each group, with standard error	2 × 2 table, or k × 2 table for a trial with k treatment groups	Number of events, person-years and rate (with confidence interval) in each group	Number of events and person-years in each group Kaplan–Meier survival curves
Measure of the effect of treatment	Difference between means t-test	Risk difference/risk ratio/odds ratio (OR): z-test/χ^2 test Number needed to treat (see Chapter 37)	Rate ratio z-test	Mantel–Cox hazard ratio Log rank test
Adjustment for baseline variables	Multiple linear regression	Mantel–Haenszel methods Logistic regression	Mantel–Haenszel methods Poisson regression	Cox regression
Analysing for different treatment effects in different subgroups	Include interaction terms in regression model			Also check for non proportional hazards (i.e. whether effect of treatment changes with time)
Special cases	Cluster randomized trial or other clustering of outcome data (see Chapter 31 for methods) Crossover trials (use methods for *matched* data)			

In addition, it highlights two special cases that need to be considered:

- whether the data are **clustered**, either in group allocation (cluster randomized trials), or in outcome measurement (repeated measures in longitudinal studies/multiple measures per subject), and
- **crossover** trials, where for each patient, treatment and control outcomes are matched.

Details of the methods can be found in the relevant sections of Parts B–E.

34.3 OTHER DESIGNS TO EVALUATE INTERVENTIONS

As discussed in Section 32.8, while the large-scale, randomized, controlled trial is the 'gold standard' for the evaluation of interventions, practical (and ethical) considerations may preclude its use. In this section, we summarize the alternative evaluation designs available, and the analysis choices involved (*see* Kirkwood *et al.*, 1997). Essentially, we have one or more of three basic comparisons at our disposal in order to evaluate the impact of interventions. These are:

1 The **pre-post comparison** involves comparing rates of the outcome of interest in several communities before the intervention is introduced (pre-intervention), with rates in the same communities after they have received the intervention (post-intervention). Such a comparison clearly requires the collection of baseline data. The plausibility of any statement attributing an impact to the intervention will be strengthened if it is demonstrated that both the prevalence of the risk factor under intervention and the rate of adverse outcome have diminished following the intervention. However, pre-post comparisons alone, without adequate concurrent controls, rarely provide compelling evidence that an intervention has successfully impacted on health, since changes in both the prevalence of risk factors and outcome are frequently observed to occur over time in the absence of any intervention. It is therefore difficult to conclude that an observed change is due to the intervention and not due to an independent secular trend. An exception to this occurs when assessing mediating factors in programmes which seek to introduce into a community a new treatment or promote a product or behaviour that did not previously exist. It will, however, still be difficult to attribute any change in health status to the programme since the improvement may still be part of a secular trend, rather than a direct consequence of the intervention.

2 **The intervention–control comparison** following the introduction of the intervention is of course at the heart of a randomized controlled trial, but this comparison may be applied in a wider context. Thus the intervention versus control comparison may be randomized or non-randomized, matched or unmatched, double-blind or open. When the comparison is double-blind and randomized, with a large number of units, as is the case with an ideally designed randomized controlled trial, the plausibility of attributing any difference in outcome observed to the intervention is high. In the absence of double-blindness or

randomization on a reasonably large scale, inference concerning the impact of the intervention becomes more problematic and it becomes essential to control for potential confounding factors.

3 **Adopters versus non-adopters comparison**: this is carried out at the individual level even if the intervention is delivered at the community level. Individuals who adopt the intervention are compared with those who do not adopt the intervention. Such a comparison is essentially a 'risk factor' study rather than an 'impact' study in that it measures the benefit to an individual of adopting the intervention rather than the public health impact of the intervention in the setting in which it was implemented. This would be the case, for example, in comparing STD incidence rates among condom users versus non-condom users following an advertising campaign. Great care needs to be taken to control potential confounding factors, since adopters and non-adopters of the intervention may differ in many important respects, including their exposure to infection. The magnitude of this problem may be assessed by a comparison of the non-adopters in the intervention area(s) with persons in control areas.

Each of these three comparisons has its merits. In the absence of a randomized controlled design, we recommend that an evaluation study include as many as possible, since they give complementary information. From Table 34.4 it can be seen that both a longitudinal design and a cross-sectional design with repeated surveys in principle allow measurement of all three of the basic types of comparison. A single cross-sectional survey can make intervention–control comparisons and adopter versus non-adopter comparisons but not pre-intervention post-intervention comparisons. The longitudinal approach can more accurately establish outcome and exposure status and the time sequence between them, but is considerably more expensive and logistically complex than the cross-sectional approach. Randomized controlled trials usually measure outcomes using a longitudinal or repeated cross-sectional design in order to maximize follow-up. However, they are not restricted to do so and, where appropriate, outcome can be measured using a single cross-sectional survey. For example, in a cluster randomized trial of the impact of a hygiene behaviour intervention, both hygiene practices and prevalence of diarrhoea could be ascertained through a single cross-sectional survey carried

Table 34.4 Matrix showing the relationship between the 'classical' study designs and the three comparisons of interest in evaluating an intervention.

	Comparisons		
Data collection	Pre-post	Intervention–control	Adopters *vs* non-adopters
Longitudinal	Yes	Yes	Yes
Cross-sectional (repeated)	Yes	Yes	Yes
Cross-sectional (single round)	No	Yes	Yes
Case-control	No	No	Yes

out, say, six months after the introduction of the intervention. A case–control evaluation can only yield an adopter versus non-adopter comparison.

The choice of analysis methods for longitudinal and cross-sectional observational studies and for case control studies are summarized in the next two sections.

34.4 LONGITUDINAL AND CROSS-SECTIONAL STUDIES

We now turn to the analysis of observational studies to investigate the association of an exposure with an outcome. In this section we cover methods relevant to cross-sectional surveys and longitudinal studies, and in the next section those relevant to case–control studies.

A **cross-sectional** study is carried out at just one point in time or over a short period of time. Since cross-sectional studies provide estimates of the features of a community at just one point in time, they are suitable for measuring prevalence but not incidence of disease (see Chapter 15 for the definition of prevalence and Chapter 22 for the definition of incidence), and associations found may be difficult to interpret. For example, a survey on onchocerciasis showed that blind persons were of lower nutritional status than non-blind. There are two possible explanations for this association. The first is that those of poor nutritional status have lower resistance and are therefore more likely to become blind from onchocerciasis. The second is that poor nutritional status is a consequence rather than a cause of the blindness, since blind persons are not as able to provide for themselves. Longitudinal data are necessary to decide which is the better explanation.

As described in Chapter 22, in a **longitudinal** study individuals are followed over time, which makes it possible to measure the incidence of disease and easier to study the natural history of disease. In some situations it is possible to obtain follow-up data on births, deaths, and episodes of disease by **continuous monitoring**, for example by monitoring registry records in populations where registration of deaths is complete. Occasionally the acquisition of data may be **retrospective**, being carried out from past records. More commonly it is **prospective** and, for this reason, longitudinal studies have often been alternatively termed **prospective studies**.

Many longitudinal studies are carried out by conducting **repeated cross-sectional surveys** at fixed intervals to enquire about, or measure, changes that have taken place between surveys, such as births, deaths, migrations, changes in weight or antibody levels, or the occurrence of new episodes of disease. The interval chosen will depend on the factors being studied. For example, to measure the incidence of diarrhoea, which is characterized by repeated short episodes, data may need to be collected weekly to ensure reliable recall. To monitor child growth, on the other hand, would require only monthly or 3-monthly measurements.

Choosing the statistical method to use

Table 34.5 provides a guide to the statistical methods available for the analysis of cross-sectional and longitudinal studies and Table 34.6 summarizes the possible

Table 34.5 Analysis of observational studies: summary of methods.

	Type of outcome			
Type of exposure	Numerical	Binary	Rate	Survival time
Binary	Difference between means *t*-test	Risk ratio/odds ratio (OR) χ^2 test	Rate ratio *z*-test	Mantel–Cox hazard ratio Log rank test
Categorical	Group means Analysis of variance Multiple linear regression	ORs against baseline Logistic regression	Rate ratios against baseline Poisson regression	Hazard ratios against baseline Cox regression
Ordered categorical (dose–response effect)	Increase in mean/group Linear regression	Increase in log odds/group Logistic regression/χ^2 test for trend	Increase in log rate/group Poisson regression	Increase in log hazard/group Cox regression
Numerical	Regression coefficient (increase in mean/unit) Linear regression	Regression coefficient (log odds ratio/unit) Logistic regression	Regression coefficient (log rate ratio/unit) Poisson regression	Regression coefficient (log hazard ratio/unit) Cox regression
Adjustment for confounders	Multiple linear regression	Mantel–Haenszel methods Logistic regression	Mantel–Haenszel methods Poisson regression	Cox regression
Special cases	Clustered data (see Chapter 31 for methods) (Repeated measures in longitudinal studies/Multiple measures per subject/ Family studies/Cluster sampling)			Non-proportional hazards

Table 34.6 Observational studies: guide to the appropriateness of types of outcome, for each study design.

	Type of outcome			
Study design	Numerical	Binary	Rate	Survival time
Longitudinal (complete follow-up)	Yes	Yes	Yes	Yes
Longitudinal (incomplete follow-up)	Yes*	Yes*	Yes	Yes
Longitudinal (repeated cross-sectional surveys)	Yes**	Yes**	Yes	Yes
Cross-sectional	Yes	Yes	No	No
Case–control	No	Yes	No	No

* Methods beyond the scope of this book

** Analyse taking into account repeated measures of outcome, using methods for clustered data (see Chapter 31).

types of outcome according to the study design. The choice of which method to use is determined by:

- the sampling scheme used to recruit participants into the study;
- whether measures are made at a single point in time, continuously over time, or at repeated points in time;
- the types of the outcome and exposure variables.

The bottom line of the guide highlights two special cases that need to be considered:

- whether the data are clustered, either because of the sampling scheme (cluster sampling or family studies), or in outcome measurement (repeated measures in longitudinal studies/multiple measures per subject); and
- in the case of survival outcomes, whether the proportional hazards assumption is satisfied.

Details of the methods can be found in the relevant sections of Parts B–E.

Types of sampling scheme and their implications

Occasionally a study includes the whole population of a confined area or institution(s), but more often only a **sample** is investigated. Whenever possible any selection should be made at random. Possible schemes include:

1 **Simple random sampling:** the required number of individuals are selected at random from the **sampling frame**, a list or a database of all individuals in the population.
2 **Systematic sampling:** for convenience, selection from the sampling frame is sometimes carried out systematically rather than randomly, by taking individuals at regular intervals down the list, the starting point being chosen at random. For example, to select a 5%, or 1 in 20, sample of the population the starting point is chosen randomly from numbers 1 to 20, and then every 20th person on the list is taken. Suppose 13 is the random number selected, then the sample would comprise individuals 13, 33, 53, 73, 93, etc.

3 **Stratified sampling:** a simple random sample is taken from a number of distinct subgroups, or **strata**, of the population in order to ensure that they are all adequately represented. If different sampling fractions are used in the different strata, simple summary statistics will not be representative of the whole population. Appropriate methods for the analysis of such studies use weights that are inversely proportional to the probability that each individual was sampled, and robust standard errors (see Chapter 30) to correct standard errors.
4 **Multi-stage or cluster sampling:** this is carried out in stages using the hierarchical structure of a population. For example, a **two-stage sample** might consist of first taking a random sample of schools and then taking a random sample of children from each selected school. The **clustering** of data must be taken into account in the analysis.
5 **Sampling on the basis of time:** for example, the 1970 British Cohort Study (BCS70) is an ongoing follow-up study of all individuals born between 5th and 11th April, 1970 and still living in Britain.

34.5 CASE–CONTROL STUDIES

In a case–control study the sampling is carried out according to *disease* rather than *exposure* status. A group of individuals identified as having the disease, the **cases**, is compared with a group of individuals not having the disease, the **controls**, with respect to their prior exposure to the factor of interest. The overriding principle is that *the controls should represent the population at risk of the disease*. More specifically, they should be individuals who, if they had experienced the disease outcome, would have been included as cases in our study. The outcome is the case–control status, and is therefore by definition a binary variable. The methods to use are therefore those outlined in Part C. These are summarized in Table 34.7. The main feature that influences the methods for analysis is whether controls were selected at random or using a matched design.

Analysis of unmatched case–control studies

For **unmatched case–control studies**, standard methods for the analysis of binary outcomes using odds ratios as the measure of association are used. Analysis of the effect of a binary exposure starts with simple 2×2 tables, and proceeds to the use of Mantel–Haenszel methods and logistic regression to control for the effect of confounding variables. These methods were described in detail in Chapters 16 to 20.

Analysis of matched case–control studies

In a **matched case–control study**, each case is matched with one or more controls, who are deliberately chosen to have the same values as the case for any potential confounding variables. There are two main reasons for matching in case–control studies:

Table 34.7 Analysis of case–control studies: summary of methods.

Sampling scheme for controls	Single exposure	Adjustment for confounding variables
Random (unmatched case–control study)	2×2 table showing exposure $\times$ case/control OR = cross-product ratio Standard χ^2 test	Logistic regression or Mantel–Haenszel methods
Stratum matching (frequency matched case–control study)	Stratified analysis: 2×2 table for each stratum Mantel–Haenszel OR and χ^2 test	Logistic regression or stratified analysis, controlling for *both* the matching factor(s) and the confounding variables
Individual matching (one control per case)	2×2 table showing agreement between case–control pairs with respect to risk factor OR = ratio of discordant pairs McNemar's χ^2 test	Conditional logistic regression
Individual matching (multiple controls per case)	Mantel–Haenszel OR and χ^2 test, stratifying on matched sets	Conditional logistic regression

1 Matching is often used to ensure that the cases and controls are similar with respect to one or more confounding variables. For example, in a study of pancreatic cancer occurring in subjects aged between 30 and 80 years it is likely that the cases will come from the older extreme of the age range. Controls might then be selected because they are of similar age to a case. This would ensure that the age distribution of the controls is similar to that of the cases, and may increase the efficiency of the study, for example by decreasing the width of confidence intervals compared to an unmatched study. Note that unless the matching factor is strongly associated with both the outcome *and* the exposure the increase in efficiency may not be large, and therefore may not justify the increased logistical difficulties and extra analytic complexity.

2 In some case–control studies it is difficult to define the population that gave rise to the cases. For example, a large hospital specializing in the treatment of cardiovascular disease may attract cases not just from the surrounding area but also referrals from further afield. In developing countries, there may be no register of the population in a given area, or who attend a particular health facility. An alternative way of selecting controls representative of the population that gave rise to the cases is to select them from the neighbourhood of each case. For example, controls might be selected from among subjects living in the third-closest house to that of each case.

It is essential to note that *if matching was used in the design, then the analysis must always take this into account*, as described in Chapter 21. In summary:

1 In the simple case of individually matched case–control studies with one control per case and no confounders, the methods for paired data described in Sections 21.3 and 21.4 can be used.

2 When there are several controls per case, Mantel–Haenszel methods may be used to estimate exposure odds ratios by stratifying on the case–control sets. However, they are severely limited because they do not allow for further control of the effects of confounding variables that were not also matching variables. This is because each stratum is a single case and its matched controls, so that further stratification is not possible. For example, if cases were individually matched with neighbourhood controls then it would not be possible to stratify additionally on age group. Stratification can be used to control for additional confounders only by restricting attention to those case–control sets that are homogeneous with respect to the confounders of interest.
3 The main approach is to use **conditional logistic regression** (see Section 21.5), which is a variant of logistic regression in which cases are only compared to controls in the same matched set. This allows analysis adjusting for several confounders at the same time. There is also no restriction on the numbers of cases and controls in each matched set.
4 However, if cases and controls are only **frequency matched** (e.g. if we simply ensure that the age distribution is roughly the same in the cases and controls), then the matching can be broken in the analysis, and standard logistic regression used, *providing the matching variable(s) are included in the model*. Mantel–Haenszel methods are also valid, with the analysis stratified on all matching variables.

Interpretation of the odds ratio estimated in a case–control study

For a *rare* disease, we saw in Chapters 16 and 23 that the odds ratio, risk ratio and rate ratio are numerically equal. For a *common* disease the meaning of the odds ratio estimated in a case–control study depends on the sampling scheme used to select the controls, as described by Rodrigues and Kirkwood (1990). Briefly, there are three possibilities:

1 The most usual choice is to select controls from those still disease-free at the end of the study (the denominator group in the odds measure of incidence); any controls selected during the course of the study who subsequently develop disease are treated as cases and not as controls. In this case the odds ratio estimated in the case–control study estimates the odds ratio in the population.
2 An alternative, in a case–control study conducted in a defined population, is to select controls from the disease-free population at each time at which a case occurs (**concurrent controls**). In this case the odds ratio estimated in the case–control study estimates the rate ratio in the population.
3 More rarely, the controls can be randomly selected from the initially disease-free population (if this can be defined). In this case the odds ratio estimated in the case–control study estimates the risk ratio in the population.

CHAPTER 35

Calculation of required sample size

35.1 Introduction
35.2 Principles of sample size calculations
35.3 Formulae for sample size calculations
35.4 Adjustments for loss to follow-up, confounding and interaction
35.5 Adjustment for clustered designs
35.6 Types of error in significance tests
35.7 Implications of study power for the interpretation of significance tests

35.1 INTRODUCTION

An essential part of planning any investigation is to decide how many people need to be studied. A formal **sample size calculation**, justifying the proposed study size and demonstrating that the study is capable of answering the questions posed, is now a component of a research proposal required by most funding agencies. Too often, medical research studies have been too small, because the sample size was decided on purely logistic grounds, or by guesswork. This is not only bad practice: it is considered by many to be unethical because of the waste of time and potential risk to patients participating in a study that cannot answer its stated research question. On the other hand, studying many more persons than necessary is also a waste of time and resources. In a clinical trial, conducting a study that is too large may also be unethical, because this could mean that more persons than necessary were given the placebo, and that the introduction of a beneficial therapy was delayed. In this chapter we will:

1 Illustrate the principles involved in sample size calculations by considering a simple example in detail.
2 Present the different formulae required for the most common sample size calculations and illustrate their application.
3 Discuss the implications of loss to follow-up, control of confounding and examination of subgroup effects.
4 Describe the principles of sample size calculation for clustered designs.
5 Define the two types of error that can occur in significance tests.
6 Illustrate the implications of study power for the interpretation of statistical significance.

35.2 PRINCIPLES OF SAMPLE SIZE CALCULATIONS

Calculating the required sample size requires that we *quantify the objectives of our study*. For example, it would not be sufficient to state simply that the objective is

to demonstrate whether or not formula-fed infants are at greater risk of death than breast-fed ones. We would also need to state:

1 The *size* of the increased risk that it was desired to demonstrate since, for example, a smaller study would be needed to detect a fourfold relative risk than to detect a twofold one.
2 The *significance level (or P-value)*, that is the strength of the evidence, that we require in order to reject the null hypothesis of no difference in risk between formula- and breast-fed infants. The greater the strength of evidence required, that is the smaller the *P*-value, the larger will be the sample size needed.
3 The probability that we would like to have of achieving this level of significance. This is required since, because of sampling variation (see Section 4.5), we cannot rule out the possibility that the size of the effect observed in the study will be much smaller than the '*true*' effect. This means that we can never guarantee that a study will be able to detect an effect however large we make it, but we can increase the probability that we do so by increasing the sample size. This probability is called the **power** of the study.

For example, we might decide that a study comparing the risk of death among formula-fed and breast-fed infants would be worthwhile if there was a 90% probability of demonstrating a difference, at 1% significance, if the true risk ratio was as high as 2. We would then calculate the number of children required. Alternatively, if we knew that a maximum of 500 children were available in our study, we might calculate the power of the study given that we wanted to detect a true risk ratio of 3 at 5% significance.

The principles involved in sample size calculations will now be illustrated by considering a simple example in detail.

Example 35.1

Consider a hypothetical clinical trial to compare two analgesics, a new drug (A) and the current standard drug (B), in which migraine sufferers will be given drug A on one occasion and drug B on another, the order in which the drugs are given being chosen at random for each patient. For illustrative purposes, we will consider a simplified analysis based on the drug stated by each patient to have provided greatest pain relief. How many patients would we need in order to be able to conclude that drug A is superior?

First, we must be specific about what we mean by superiority. We will state this as an overall preference rate of 70% or more for drug A, and we will decide that we would like a 90% power of achieving a significant result at the 5% level.

Under the null hypothesis of no difference between the efficacies of the two drugs, the proportion of patients stating a preference for drug A will be 0.5 (50%). We can test the evidence that the observed preference proportion, p, differs from 0.5 using a z-test, as described in Section 15.6:

$$z = \frac{p - 0.5}{\text{s.e.}(p)} = \frac{p - 0.5}{\sqrt{(0.5 \times (1 - 0.5)/n)}} = \frac{p - 0.5}{\sqrt{(0.25/n)}}$$

This result will be significant at the 5% level ($P < 0.05$) if $z \geq 1.96$, or in other words if p is 1.96 standard errors or more away from the null hypothesis value of 0.5.

We will illustrate the principles behind sample size calculations by considering different possible sample sizes and assessing their adequacy as regards the power of our study.

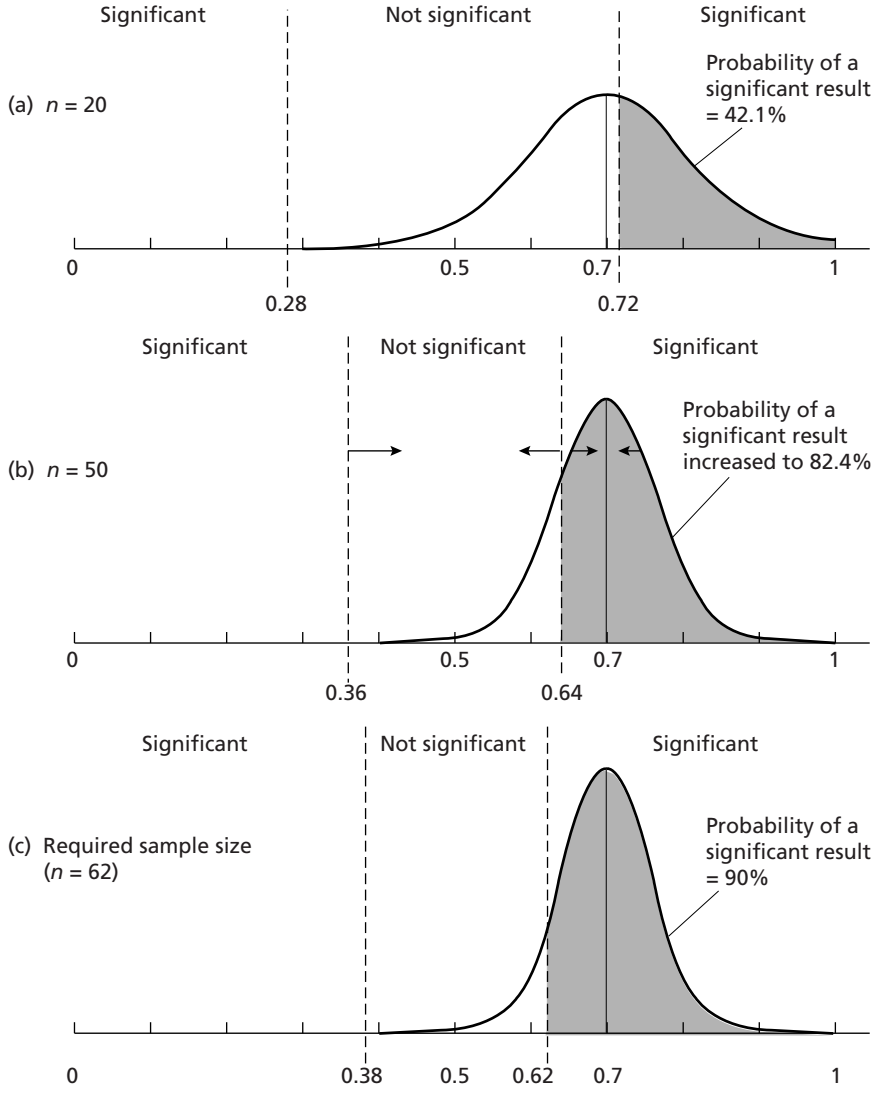


Fig. 35.1 Probability of obtaining a significant result (at the 5 % level) with various sample sizes (n) when testing the proportion of preferences for drug A rather than drug B against the null hypothesis value of 0.5, if the true value is 0.7.

(a) We will start with a sample size of $n = 20$, as depicted in Figure 35.1(a). Here:

$$\text{s.e.} = \sqrt{(0.25/20)} = 0.1118$$
$$0.5 + 1.96 \times \text{s.e.} = 0.5 + 1.96 \times 0.1118 = 0.72$$
and $$0.5 - 1.96 \times \text{s.e.} = 0.5 - 1.96 \times 0.1118 = 0.28$$

Thus observed proportions of 0.72 and above, or 0.28 and below, would lead to a result that is significant at the 5% level.

If the *true* proportion is 0.7, what is the likelihood of observing 0.72 or above, and thus getting a result that is significant at the 5% level? This is illustrated by the shaded area in Figure 35.1(a). The curve represents the sampling distribution, which is a normal distribution centred on 0.7 with a standard error of $\sqrt{(0.7 \times 0.3/20)} = 0.1025$. The z-value corresponding to 0.72 is:

$$\frac{0.72 - 0.7}{0.1025} = 0.20$$

The proportion of the standard normal distribution above 0.20 is found from Table A1 (in the Appendix) to equal 0.421, or 42.1%. In summary, this means that with a sample size of 20 we have only a 42.1% chance of demonstrating that drug A is better, if the true preference rate is 0.7.

(b) Consider next what happens if we increase the sample size to 50, as shown in Figure 35.1(b). The ranges of values that would now be significant have widened to 0.64 and above, or 0.36 and below. The sampling distribution has narrowed, and there is a greater overlap with the significant ranges. Consequently, the probability of a significant result has increased. It is now found to be 82.4%, but this is still less than our required 90%.

(c) Thus we certainly need to study more than 50 patients in order to have 90% power. But exactly how many do we need? We need to increase the sample size, n, to the point where the overlap between the sampling distribution and the significant ranges reaches 90%, as shown in Figure 35.1(c). We will now describe how to calculate directly the sample size needed to do this. A significant result will be achieved if we observe a value above

$$0.5 + 1.96 \times \text{s.e.} = 0.5 + 1.96 \times \sqrt{(0.5 \times 0.5/n)}$$

(or below $0.5 - 1.96 \times$ s.e.). We want to select a large enough n so that 90% of the sampling distribution is above this point. The z-value of the sampling distribution corresponding to 90% is -1.28 (see Table A2), which means an observed value of

$$0.7 - 1.28 \times \text{s.e.} = 0.7 - 1.28 \times \sqrt{(0.7 \times 0.3/n)}$$

Therefore, n should be chosen large enough so that

$$0.7 - 1.28 \times \sqrt{(0.7 \times 0.3/n)} > 0.5 + 1.96 \times \sqrt{(0.5 \times 0.5/n)}$$

Rearranging this gives

$$0.7 - 0.5 > \frac{1.96 \times \sqrt{(0.5 \times 0.5)} + 1.28 \times \sqrt{(0.7 \times 0.3)}}{\sqrt{n}}$$

Squaring both sides, and further rearrangement gives

$$n > \frac{[1.96 \times \sqrt{(0.5 \times 0.5)} + 1.28 \times \sqrt{(0.7 \times 0.3)}]^2}{0.2^2}$$
$$= \frac{1.5666^2}{0.2^2} = 61.4$$

We therefore require at least 62 patients to satisfy our requirements of having a 90% power of demonstrating a difference between drugs A and B that is significant at the 5% level, if the true preference rate for drug A is as high as 0.7.

35.3 FORMULAE FOR SAMPLE SIZE CALCULATIONS

The above discussion related to sample size determination for a test that a single proportion (the proportion of participants preferring drug A to drug B) differs from a specified null value. In practice it is not necessary to go through such detailed reasoning every time. Instead the sample size can be calculated directly from a general formula, which in this case is:

$$n > \frac{[u\sqrt{\pi(1-\pi)} + v\sqrt{\pi_{\text{null}}(1-\pi_{\text{null}})}]^2}{(\pi - \pi_{\text{null}})^2}$$

where:

n = required minimum sample size
π = proportion of interest
π_{null} = null hypothesis proportion
u = one-sided percentage point of the normal distribution corresponding to 100% − the power, e.g. if power = 90%, (100% − power) = 10% and u = 1.28
v = percentage of the normal distribution corresponding to the required (two-sided) significance level, e.g. if significance level = 5%, $v = 1.96$.

For example, in applying this formula to the above example we have:

$$\pi = 0.7,\ \pi_{\text{null}} = 0.5,\ u = 1.28 \text{ and } v = 1.96$$

giving

$$n > \frac{[1.28 \times \sqrt{(0.7 \times 0.3)} + 1.96 \times \sqrt{(0.5 \times 0.5)}]^2}{(0.7 - 0.5)^2} = \frac{1.5666^2}{0.2^2} = 61.4$$

which is exactly the same as obtained above.

The same principles can also be applied in other cases. Detailed reasoning is not given here but the appropriate formulae for use in the most common situations are listed in Table 35.1. The list consists of two parts. Table 35.1(a) covers cases where the aim of the study is to demonstrate a specified difference. Table 35.1(b) covers situations where the aim is to estimate a quantity of interest with a specified precision.

Note that for the cases with two means, proportions, or rates, the formulae give the sample sizes required for *each* of the two groups. The total size of the study is therefore *twice* this.

Table 35.2 gives adjustment factors for **study designs with unequal size groups** (see Example 35.4). Note also that the formulae applying to rates give the required sample size in the same unit as the rates (see Example 35.3).

The use of Table 35.1 will be illustrated by several examples. It is important to realize that sample size calculations are based on our best guesses of a situation. The number arrived at is not magical. It simply gives an idea of the sort of numbers to be studied. In other words, it is useful for distinguishing between 50 and 100, but not between 51 and 52. *It is essential to carry out sample size calculations for several different scenarios, not just one*. This gives a clearer picture of the possible scope of the study and is helpful in weighing up the balance between what is desirable and what is logistically feasible.

Example 35.2

A study is to be carried out in a rural area of East Africa to ascertain whether giving food supplementation during pregnancy increases birth weight. Women attending the antenatal clinic are to be randomly assigned to either receive or not receive supplementation. Formula 4 in Table 35.1 will help us to decide how many women should be enrolled in each group. We need to supply the following information:

1. The size of the difference between mean birth weights that we would like to be able to detect. After much consideration it was decided that an increase of 0.25 kg was an appreciable effect that we would not like to miss. We therefore need to apply the formula with $\mu_1 - \mu_0 = 0.25$ kg.
2. The standard deviations of the distributions of birth weight in each group. It was decided to assume that the standard deviation of birth weight would be the same in the two groups. Past data suggested that it would be about 0.4 kg. In other words we decided to assume that $\sigma_1 = 0.4$ kg and $\sigma_0 = 0.4$ kg.
3. The power required. 95 % was agreed on. We therefore need $u = 1.64$.
4. The significance level required. It was decided that if possible we would like to achieve a result significant at the 1 % level. We therefore need $v = 2.58$.

Applying formula 4 with these values gives:

$$n > \frac{(1.64 + 2.58)^2 \times (0.4^2 + 0.4^2)}{0.25^2} = \frac{17.8084 \times 0.32}{0.0625} = 91.2$$

Therefore, in order to satisfy our requirements, we would need to enrol about 90 women in each group.

Example 35.3

Before embarking on a major water supply, sanitation, and hygiene intervention in southern Bangladesh, we would first like to know the average number of episodes of diarrhoea per year experienced by under-5-year-olds. We guess that this incidence is probably about 3, but would like to estimate it within ±0.2. This means that if, for example, we observed 2.6 episodes/child/year, we would like to be able to conclude that the true rate was probably between 2.4 and 2.8 episodes/child/year. Expressing this in more statistical terms, we would like our 95% confidence interval to be no wider than ±0.2. As the width of this confidence interval is approximately ±2 s.e.'s, this means that we would like to study enough children to give a standard error as small as 0.1 episodes/child/year. Applying formula 9 in Table 35.1 gives:

$$n > \frac{3}{0.1^2} = 300$$

Note that the formulae applying to rates (numbers 2, 5, 9, 12) give the required sample size in the same unit as the rates. We specified the rates as per child per year. We therefore need to study 300 child-years to yield the desired precision. This could be achieved by observing 300 children for one year each or, for example, by observing four times as many (1200) for 3 months each. It is important not to overlook, however, the possibility of other factors such as seasonal effects when deciding on the time interval for a study involving the measurement of incidence rates.

Example 35.4

A case–control study is planned to investigate whether bottle-fed infants are at increased risk of death from acute respiratory infections compared to breast-fed infants. The mothers of a group of cases (infant deaths, with an underlying respiratory cause named on the death certificate) will be interviewed about the breast-feeding status of the child prior to the illness leading to death. The results will be compared with those obtained from mothers of a group of healthy controls regarding the current breast-feeding status of their infants. It is expected that about 40% of controls ($\pi_0 = 0.4$) will be bottle-fed, and we would like to detect a difference if bottle-feeding was associated with a twofold increase of death (OR = 2).

Table 35.1 Formulae for sample size determination. (a) For studies where the aim is to demonstrate a significant difference. (b) For studies where the aim is to estimate a quantity of interest with a specified precision.

		Information needed	Formula for minimum sample size
(a) Significant result			
1 Single mean	$\mu - \mu_0$	Difference between mean, μ, and null hypothesis value, μ_0	$\frac{(u+v)^2\sigma^2}{(\mu-\mu_0)^2}$
	σ	Standard deviation	
	u, v	As below	
2 Single rate*	μ	Rate	$\frac{(u+v)^2\mu}{(\mu-\mu_0)^2}$
	μ_0	Null hypothesis value	
	u, v	As below	
3 Single proportion	π	Proportion	$\frac{\{u\sqrt{[\pi(1-\pi)]}+v\sqrt{[\pi_0(1-\pi_0)]}\}^2}{(\pi-\pi_0)^2}$
	π_0	Null hypothesis value	
	u, v	As below	
4 Comparison of two means (sample size of each group)	$\mu_1 - \mu_0$	Difference between the means	$\frac{(u+v)^2(\sigma_1^2+\sigma_0^2)}{(\mu_1-\mu_0)^2}$
	$\sigma_1,\ \sigma_0$	Standard deviations	
	u, v	As below	
5 Comparison of two rates* (sample size of each group)	$\mu_1,\ \mu_0$	Rates	$\frac{(u+v)^2(\mu_1+\mu_0)}{(\mu_1-\mu_0)^2}$
	u, v	As below	
6 Comparison of two proportions (sample size of each group)	$\pi_1,\ \pi_0$	Proportions	$\frac{\{u\sqrt{[\pi_1(1-\pi_1)+\pi_0(1-\pi_0)]}+v\sqrt{[2\bar{\pi}(1-\bar{\pi})]}\}^2}{(\pi_0-\pi_1)^2}$ where $\bar{\pi}=\frac{\pi_1+\pi_0}{2}$
	u, v	As below	
7 Case–control study (sample size of each group)	π_0	Proportion of controls exposed	$\frac{\{u\sqrt{[\pi_0(1-\pi_0)+\pi_1(1-\pi_1)]}+v\sqrt{[2\bar{\pi}(1-\bar{\pi})]}\}^2}{(\pi_1-\pi_0)^2}$ where $\bar{\pi}=\frac{\pi_0+\pi_1}{2}$
	OR	Odds ratio	
	π_1	Proportion of cases exposed, calculated from $\pi_1=\frac{\pi_0\,\mathrm{OR}}{1+\pi_0(\mathrm{OR}-1)}$	
	u, v	As below	

All cases	u	One-sided percentage point of the normal distribution corresponding to 100 % – the power e.g. if power = 90 %, $u = 1.28$	
	v	Percentage point of the normal distribution corresponding to the (two-sided) significance level e.g. if significance level = 5%, $v = 1.96$	
(b) Precision			
8 Single mean	σ	Standard deviation	$\frac{\sigma^2}{e^2}$
	e	Required size of standard error	
9 Single rate*	μ	Rate	$\frac{\mu}{e^2}$
	e	Required size of standard error	
10 Single proportion	π	Proportion	$\frac{\pi(1-\pi)}{e^2}$
	e	Required size of standard error	
11 Difference between two means (sample size of each group)	σ_1, σ_0	Standard deviations	$\frac{\sigma_1^2 + \sigma_0^2}{e^2}$
	e	Required size of standard error	
12 Difference between two rates* (sample size of each group)	μ_1, μ_0	Rates	$\frac{\mu_1 + \mu_0}{e^2}$
	e	Required size of standard error	
13 Difference between two proportions (sample size of each group)	π_1, π_0	Proportions	$\frac{\pi_1(1-\pi_1) + \pi_0(1-\pi_0)}{e^2}$
	e	Required size of standard error	

*In these cases the sample size refers to the same units as used for the denominator of the rate(s). For example, if the rate is expressed per person-year, the formula gives the number of person-years of observation required (*see* Example 35.3)

How many cases and controls need to be studied to give a 90% power ($u = 1.28$) of achieving 5% significance ($v = 1.96$)? The calculation consists of several steps as detailed in formula 7 of Table 35.1.

1 Calculate π_1, the proportion of cases bottle-fed:

$$\pi_1 = \frac{\pi_0 \mathrm{OR}}{1 + \pi_0(\mathrm{OR} - 1)} = \frac{0.4 \times 2}{1 + 0.4 \times (2 - 1)} = \frac{0.8}{1.4} = 0.57$$

2 Calculate $\bar{\pi}$the average of π_0 and π_1:

$$\bar{\pi} = \frac{0.4 + 0.57}{2} = 0.485$$

3 Calculate the minimum sample size:

$$n > \frac{[1.28\sqrt{(0.4 \times 0.6 + 0.57 \times 0.43)} + 1.96\sqrt{(2 \times 0.485 \times 0.515)}]^2}{(0.57 - 0.4)^2}$$
$$= \frac{[1.28\sqrt{0.4851} + 1.96\sqrt{0.4996}]^2}{0.17^2} = \frac{2.2769^2}{0.17^2} = 179.4$$

We would therefore need to recruit about 180 cases and 180 controls, giving a total sample size of 360.

What difference would it make if, rather than recruiting equal numbers of cases and controls, we decided to recruit three times as many controls as cases? Table 35.2 gives appropriate adjustment factors for the number of cases according to differing number of controls per case. For $c = 3$ the adjustment factor is 2/3. This means we would need $180 \times 2/3$, that is 120 cases, and three times as many,

Table 35.2 Adjustment factor for use in study designs to compare unequal sized groups, such as in a case–control study selecting multiple controls per case. This factor (f) applies to the smaller group and equals $(c + 1)/(2c)$, where the size of the larger group is to be c times that of the smaller group. The sample size of the smaller group is therefore fn, where n would be the number required for equal-sized groups, and that of the larger group is cfn (see Example 35.4).

Ratio of larger to smaller group (c)	Adjustment to sample size of smaller group (f)
1	1
2	3/4
3	2/3
4	5/8
5	3/5
6	7/12
7	4/7
8	9/16
9	5/9
10	11/20

namely 360, controls. Thus although the requirement for the number of cases has considerably decreased, the total sample size has increased from 360 to 540.

35.4 ADJUSTMENTS FOR LOSS TO FOLLOW-UP, CONFOUNDING AND INTERACTION

The calculated sample size should be increased to allow for possible non-response or loss to follow-up. Further adjustments should be made if the final analysis will be adjusted for the effect of confounding variables or if the examination of subgroup effects is planned.

1 It is nearly always the case that a proportion of the people originally recruited to the study will not provide data for inclusion in the final analysis: for example because they withdraw from the study or are lost to follow-up, or because information on key variables is missing. The required sample size should be adjusted to take account of these possibilities. If we estimate that $x\%$ of patients will not contribute to the final analysis then the sample size should be multiplied by $100/(100 - x)$. For example if $x = 20\%$, the multiplying factor equals $100/(100 - 20) = 1.25$.

$$\text{Adjustment factor for } x\% \text{ loss} = 100/(100 - x)$$

2 Smith and Day (1984) considered the effect of controlling for confounding variables, in the context of the design of case–control studies. They concluded that, for a single confounding variable, an increase in the sample size of more than 10% is unlikely to be needed. Breslow and Day (1987) suggested that for several confounding variables that are jointly independent, as a rough guide one could add the extra sample size requirements for each variable separately.

3 In some circumstances we wish to design a study to detect differences between associations in different subgroups, in other words to detect *interaction* between the treatment or exposure effect and the characteristic that defines the subgroup. The required sample size will be *at least* four times as large as when the aim is to detect the overall association, and may be considerably larger. For more details see Smith and Day (1984) or Breslow and Day (1987).

35.5 ADJUSTMENT FOR CLUSTERED DESIGNS

The analysis of studies that employ a clustered design was described in Chapter 31. These include cluster randomized trials, in which randomization is applied to clusters of people rather than individuals (see also Section 34.2), family studies and studies which employ a cluster sampling scheme (see also Section 34.4). Because individuals within a cluster may be more similar to each other than to individuals in other clusters, a cluster randomized trial needs to include more

individuals than the corresponding individually randomized trial. The same is true of studies that employ a cluster rather than individual sampling scheme.

The amount by which the sample size needs to be multiplied is known as the **design effect** (Deff), and depends on **the intraclass correlation coefficient** (ICC). The ICC was defined in Section 31.4 as the ratio of the between-cluster variance to the total variance.

$$\text{Design effect (Deff)} = 1 + (n' - 1) \times \text{ICC}$$

ICC = intraclass correlation coefficient

n' = average cluster size

It can be seen that two factors influence the size of the design effect:

1 the greater the ICC, the greater will be the design effect; and
2 the greater the number of individuals per cluster, the greater will be the design effect.

The number of clusters required is given by:

$$\text{No. of clusters} = \frac{n}{n'}[1 + (n' - 1) \times \text{ICC}]$$

n = uncorrected total sample size

n' = average cluster size

Estimation of the ICC, at the time that a study is designed, is often difficult because published papers have not tended to report ICCs. Although attempts have been made to publish typical ICCs, for different situations (for example see Gulliford *et al.*, 1999), it will usually be sensible to calculate the number of clusters required under a range of assumptions about the ICC, as well as using a range of values for the cluster size. In particular, it may be useful to present the results graphically, with lines showing the number of clusters required against number of individuals per cluster, for various values of ICC.

For more details about sample size calculations for cluster randomized trials, see Donner and Klar (2000) or Ukoumunne *et al.* (1999). Alternatively, Hayes and Bennett (1999) suggested a method based on the **coefficient of variation** (standard deviation/mean) of cluster rates, proportions or means. They give guidance on how to estimate this value with or without the use of prior data on between-cluster variation, and provide formulae for both unmatched and pair-matched trials.

35.6 TYPES OF ERROR IN SIGNIFICANCE TESTS

A significance test can never prove that a null hypothesis is either true or false. It can only give an indication of the strength of the evidence against it. In using significance tests to make decisions about whether to reject a null hypothesis, we can make two types of error: we can reject a null hypothesis when it is in fact true, or fail to reject it when it is false. These are called **type I** and **type II errors** respectively (Table 35.3).

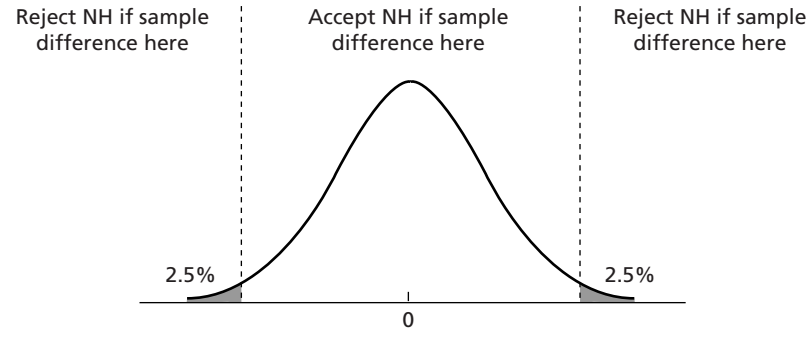
As explained in Chapter 8, the *P*-value (significance level) equals the probability of occurrence of a result as extreme as, or more extreme than, that observed if the null hypothesis were true. For example, there is a 5% probability that sampling variation alone will lead to a $P < 0.05$ (a result significant at the 5% level), and so if we judge such a result as sufficient evidence to reject the null hypothesis, there is a 5% probability that we are making an error in doing so, if the null hypothesis is true (see Figure 35.2a).

The second type of error is that the null hypothesis is not rejected when it is false. This occurs because of overlap between the real sampling distribution of the sample difference about the population difference, $d (\neq 0)$ and the acceptance region for the null hypothesis based on the hypothesized sampling distribution about the incorrect difference, 0. This is illustrated in Figure 35.2(b). The shaded area shows the proportion (b%) of the real sampling distribution that would fall within the acceptance region for the null hypothesis, i.e. that would appear consistent with the null hypothesis at the 5% level. The probability that we *do not* make a type II error ($100 - b$%) equals the **power** of the test.

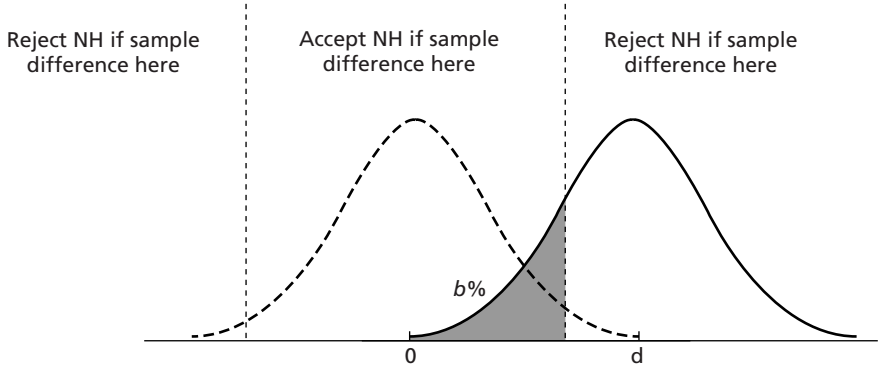
If a lower significance level were used, making the probability of a type I error smaller, the size of the shaded area would be increased, so that there would be a larger probability of a type II error. The converse is also true. For a given significance level, the probability of a type II error can be reduced by increasing the power, by increasing either the sample size or the precision of the measurements (see Chapter 36). Each of the curves in Figure 35.2 would be taller and narrower, and overlap less; the size of the shaded area would therefore be reduced.

Table 35.3 Types of error in hypothesis tests.

	Reality	
Conclusion of significance test	Null hypothesis is true	Null hypothesis is false
Reject null hypothesis	*Type I error* (probability = significance level)	Correct conclusion (probability = power)
Do not reject null hypothesis	Correct conclusion (probability = 1 − significance level)	*Type II error* (probability = 1 − power)



(a) Type I error. Null hypothesis (NH) is *true*. Population difference = 0. The curve shows the sampling distribution of the sample difference. The shaded areas (total 5%) give the probability that the null hypothesis is wrongly rejected.



(b) Type II error. Null hypothesis is *false*. Population difference = d≠0. The continuous curve shows the real sampling distribution of the sample difference, while the dashed curve shows the sampling distribution under the null hypothesis. The shaded area is the probability (*b*%) that the null hypothesis fails to be rejected.

Fig. 35.2 Probabilities of occurrence of the two types of error of hypothesis testing, illustrated for a test at the 5% level.

35.7 IMPLICATIONS OF STUDY POWER FOR THE INTERPRETATION OF SIGNIFICANCE TESTS

Unfortunately, significance tests are often misused, with investigators using a 5% threshold for statistical significance and concluding that any non-significant result means that the null hypothesis is true. Another common misinterpretation is that *the P-value is the probability that the null hypothesis is true*.

Table 35.4(a) demonstrates why such thinking is incorrect. It is based on considering what would happen if 1000 different null hypotheses were tested and significance at the 5% level ($P < 0.05$) used as a threshold for rejection, under the following plausible assumptions:

Table 35.4 Implications of study power for the interpretation of significance tests.

(a) Conclusions of significance tests of 1000 hypotheses, of which 10% are false, using $P = 0.05$ as threshold significance level, and conducted with 50% power (adapted from Oakes, 1986).

	Reality		
Conclusion of significance test	Null hypothesis true	Null hypothesis false	Total
Reject null hypothesis ($P < 0.05$)	45 (*Type I errors*)	50	95
Do not reject null hypothesis ($P \geq 0.05$)	855	50 (*Type II errors*)	905
Total	900	100	1000

(b) Proportion of false-positive significant results, according to the *P*-value used for significance, the power of the study and the proportion of studies in which the null hypothesis is truly false (adapted from Sterne and Davey Smith 2001). The result corresponding to Table 35.4(a) is in bold.

Proportion of studies in which the null hypothesis is false	Power of study	Percentage of significant results that are false-positives		
		$P = 0.05$	$P = 0.01$	$P = 0.001$
80%	20%	5.9	1.2	0.1
	50%	2.4	0.5	0.0
	80%	1.5	0.3	0.0
50%	20%	20.0	4.8	0.5
	50%	9.1	2.0	0.2
	80%	5.9	1.2	0.1
10%	20%	69.2	31.0	4.3
	50%	**47.4**	15.3	1.8
	80%	36.0	10.1	1.1
1%	20%	96.1	83.2	33.1
	50%	90.8	66.4	16.5
	80%	86.1	55.3	11.0

1. 10% of the null hypotheses tested are in fact false (i.e. the effect being investigated is real), and 90% are true (i.e. the hypothesis tested is incorrect). This is conceivable given the large numbers of factors searched for in the epidemiological literature. For example by 1985 nearly 300 risk factors for coronary heart disease had been identified; it is unlikely that more than a fraction of these factors actually increase the risk of the disease.
2. The power of the test is 50%. This is consistent with published surveys of the size of clinical trials (see, for example, Moher *et al.*, 1994); a large proportion having been conducted with an inadequate sample size to address the research question.

Assumption (1) determines the column totals in the table; the null hypothesis is true in 900 of the tests and false in 100 of them. The type I error rate will be 5%, the significance level being used. This means that we will incorrectly reject 45 of

the 900 true null hypotheses. Assumption (2) means that that the type II error rate equals 50% (100% – power). We will therefore fail to reject 50 of the 100 null hypotheses that are false. It can be seen from the table that of the 95 tests that result in a statistically significant result, only 50 are correct; 45 (47.4%) are type I errors (false positive results).

Table 35.4(b) extends Table 35.4(a) by showing the percentage of false positive results for different *P*-value thresholds under different assumptions about both the power of studies and the proportion of true null hypotheses. For any choice of significance level, the proportion of 'significant' results that are false-positives is greatly reduced as power increases. The table suggests that unless the proportion of meaningful hypotheses is very small, it is reasonable to regard *P*-values less than 0.001 as providing strong evidence against the null hypothesis.

CHAPTER 36

Measurement error: assessment and implications

36.1 Introduction

36.2 The evaluation of diagnostic tests
- Sensitivity and specificity
- Predictive values
- Choosing cut-offs

36.3 Assessing reproducibility of measurements
- Kappa statistic for categorical variables
- Numerical variables: reliability and the intraclass correlation coefficient
- Links between weighted kappa and the intraclass correlation coefficient

36.4 Numerical variables: method comparison studies

36.5 Implications for interpretation
- Regression dilution bias
- The effects of measurement error and misclassification in multivariable models
- Regression to the mean

36.1 INTRODUCTION

In this chapter we consider how to examine for errors made in measuring outcome or exposure variables, and the implications of such errors for the results of statistical analyses. Such errors may occur in a variety of ways, including:

1 **Instrumental errors**, arising from an inaccurate diagnostic test, an imprecise instrument or questionnaire limitations.
2 **Underlying variability**, leading to differences between replicate measurements taken at different points in time.
3 **Respondent errors**, arising through misunderstanding, faulty recall, giving the perceived 'correct' answer, or through lack of interest. In some instances the respondent may deliberately give the wrong answer because, for example, of embarrassment in questions connected with sexually transmitted diseases or because of suspicion that answers could be passed to income tax authorities.
4 **Observer errors**, including imprecision, misuse/misunderstanding of procedures, and mistakes.
5 **Data processing errors**, such as coding, copying, data entry, programming and calculating mistakes.

Our focus is on the detection, measurement and implications of random error, in the sense that we will assume that any errors in measuring a variable are independent of the value of other variables in the dataset. Detailed discussion of **differential bias** arising from the design or conduct of the study, such as **selection bias**, is outside the scope of this book. Readers are referred to textbooks on epidemiology and study design: recommended books are listed at the beginning of Chapter 34. We cover:

1 How to evaluate a diagnostic test or compare a measurement technique against a **gold standard**, that gives a (more) precise measurement of the true value. Often, the gold-standard method is expensive, and we wish to examine the performance of a cheaper or quicker alternative.
2 How to choose the 'best' cut-off value when using a numerical variable to give a binary classification.
3 How to assess the *reproducibility* of a measurement, including:
   - agreement between different observers using the same measurement technique,
   - the agreement between replicate measurements taken at different points in time.
4 The implications of inaccuracies in measurement for the interpretation of results.

36.2 THE EVALUATION OF DIAGNOSTIC TESTS

The analysis of binary outcome variables was considered in Part C, while methods for examining the effect of binary exposure variables are presented throughout this book. In this section we consider how to assess the ability of a procedure to correctly classify individuals between the two categories of a binary variable. For example, individuals may be classified as diseased or non-diseased, exposed or non-exposed, positive or negative, or at high risk or not.

Sensitivity and specificity

The ability of a **diagnostic test** (or procedure) to correctly classify individuals into two categories (positive and negative) is assessed by two parameters, **sensitivity** and **specificity**:

$$\begin{aligned}\text{Sensitivity} &= \text{proportion of true positives correctly identified as such}\\ &= 1 - \text{false negative rate}\end{aligned}$$

$$\begin{aligned}\text{Specificity} &= \text{proportion of true negatives correctly identified as such}\\ &= 1 - \text{false positive rate}\end{aligned}$$

To estimate sensitivity and specificity, each individual needs to be classified definitively (using a 'gold-standard' assessment) as true positive or true negative and, in addition, to be classified according to the test being assessed.

Example 36.1

Table 36.1 shows the results of a pilot study to assess parents' ability to recall the correct BCG immunization status of their children, as compared to health authority records. Of the 60 children who had in fact received BCG immunization, almost all, 55, were correctly identified as such by their parents, giving a sensitivity of 55/60 or 91.7%. In contrast, 15 of the 40 children with no record of BCG

Table 36.1 Comparison of parents' recall of the BCG immunization status of their children with that recorded in the health authority records.

BCG immunization according to health authority records ('gold standard' test)	BCG immunization according to parents (procedure being assessed)			
	Yes	No	Total	
Yes	55	5	60	*Sensitivity* = 55/60 = 91.7%
No	15	25	40	*Specificity* = 25/40 = 62.5%
Total	70	30	100	
	PPV = 55/70 = 78.6%	NPV = 25/30 = 83.3%		

immunization were claimed by their parents to have been immunized, giving a specificity of 25/40 or 62.5%.

Sensitivity and specificity are characteristics of the test. Their values do *not* depend on the prevalence of the disease in the population. They are particularly important in assessing **screening tests**. Note that there is an inverse relationship between the two measures, tightening (or relaxing) criteria to improve one will have the effect of decreasing the magnitude of the other. Where to draw the line between them will depend on the nature of the study. For example, in designing a study to test a new leprosy vaccine, it would be important initially to exclude any lepromatous patients. One would therefore want a test with a high success rate of detecting positives, or in other words a highly sensitive test. One would be less concerned about specificity, since it would not matter if a true negative was incorrectly identified as positive and so excluded. In contrast, for the detection of cases during the post-vaccine (or placebo) follow-up period, one would want a test with high specificity, since it would then be more important to be confident that any positives detected were real, and less important if some were missed.

Predictive values

A clinician who wishes to interpret the results of a diagnostic test will want to know the probability that a patient is truly positive if the test is positive and similarly the probability that the patient is truly negative if the test is negative. These are called the **positive** and **negative predictive values** of the test:

$$\text{Positive predictive value (PPV)} = \text{proportion of test positives that are truly positive}$$

$$\text{Negative predictive value (NPV)} = \text{proportion of test negatives that are truly negative}$$

In Example 36.1, BCG immunization was confirmed from health authority records for 55 of the 70 children reported by their parents as having received immunization, giving a PPV of 55/70 or 78.6%. The NPV was 25/30 or 83.3%.

The values of the positive and negative predictive values *depend on the prevalence of the disease in the population*, as well as on the sensitivity and specificity of the procedure used. The lower the prevalence of true positives, the lower will be the proportion of true positives among test positives and the lower, therefore, will be the positive predictive value. Similarly, increasing prevalence will lead to decreasing negative predictive value.

Choosing cut-offs

Where binary classifications are derived from a numerical variable, using a cut-off value, the performance of different cut-off values can be assessed using a **Receiver Operating Characteristic** curve, often known as a **ROC curve**. This is a plot of sensitivity against 1 − specificity, for different choices of cut-off. The name of the curve derives from its original use in studies of radar signal detection.

Example 36.2

Data from a study of lung function among 636 children aged 7 to 10 years living in a deprived suburb of Lima, Peru were introduced in Chapter 11. For each child the FEV_1 (the volume of air the child could breathe out in 1 second) was measured before and after she or he exercised on an electric treadmill for 4 minutes, and the percentage reduction in FEV_1 after exercise was calculated. This ranged from −17.9% (i.e. an increase post-exercise) to a 71.4% reduction.

A total of 60 (9.4%) of the parents (or carers) reported that their child had experienced chest tightness suggestive of asthma in the previous 12 months. There was strong evidence of an association between % reduction in FEV_1 and reported chest tightness in the child (odds ratio per unit increase in % reduction 1.052, 95% CI 1.031 to 1.075). To examine the utility of % reduction in FEV_1 as a means of diagnosing asthma, a ROC curve was plotted, as displayed in Figure 36.1, showing sensitivity (vertical axis) against 1 − specificity (horizontal axis) for different choices of cut-off values for FEV_1. In this example, we can see that if we required 75% sensitivity from our cut-off then specificity would be around 50%, while a lower cut-off value that gave around 60% sensitivity would yield a specificity of about 75%.

The overall ability of the continuous measure (in this case FEV_1) to discriminate between individuals with and without disease may be measured by the **area under the ROC curve**. If perfect discrimination were possible (the existence of a cut-off with 100% sensitivity and 100% specificity), the ROC curve would go across the top of the grid area, and yield an area of 1. This is because decreasing the specificity by lowering the cut-off would maintain sensitivity at 100%, since a lower cut-off can only capture an equal or higher percentage of cases. In contrast, if the continuous measure is not able to discriminate at all, then 100% sensitivity

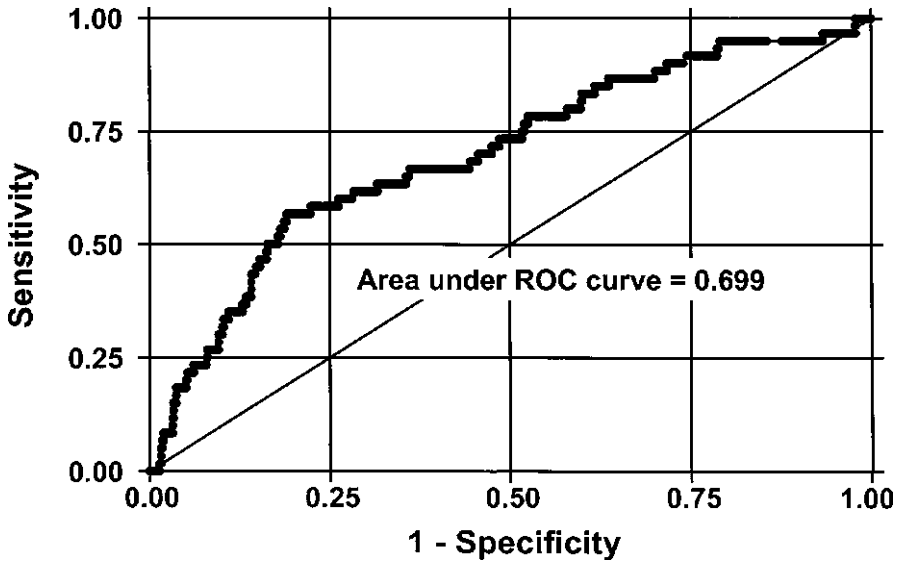


Fig. 36.1 ROC curve showing the sensitivity and specificity corresponding to different choices of cut-off for % reduction in FEV_1 as a test for chest tightness suggestive of asthma in children in Peru.

can only be achieved with 0 % specificity and vice versa. The ROC curve will be the straight line in Figure 36.1 showing sensitivity = 1 − specificity, and the area under the curve will be 0.5. In this example the area under the ROC curve is 0.699. The area under the ROC curve may also be used to quantify how well a predictor based on a number of variables (for example based on the linear predictor from a logistic regression model) discriminates between individuals with and without disease.

36.3 ASSESSING REPRODUCIBILITY OF MEASUREMENTS

In this section we describe methods to assess the extent of **reproducibility** of a measurement (also known as **reliability**), including:

- agreement between different observers using the same measurement technique;
- agreement between replicate measurements taken at different points in time.

This is particularly important for any variable that is subjectively assessed, such as in Example 36.3, or for which there may be underlying natural variation, such as the composition of a person's daily nutritional intake (see Example 36.5), which will show some day-to-day variations, as well as possible marked seasonal differences.

Kappa statistic for categorical variables

For categorical variables, the extent of reproducibility is usually assessed using a **kappa** statistic. This is based on comparing the *observed* proportion of agreement

(A_{obs}) between two readings made by two different observers, or on two different occasions, with the proportion of agreements (A_{exp}) that would be *expected* simply by chance. It is denoted by the Greek letter kappa, κ, and is defined as:

$$\kappa = \frac{A_{obs} - A_{exp}}{1 - A_{exp}}$$

If there is complete agreement then $A_{obs} = 1$ and so $\kappa = 1$. If there is no more agreement than would be expected by chance alone then $\kappa = 0$, and if there is *less* agreement than would be expected by chance alone then κ will be negative. Based on criteria originally proposed by Landis and Koch:

- kappa values greater than about 0.75 are often taken as representing excellent agreement;
- those between 0.4 and 0.75 as fair to good agreement; and
- those less than 0.4 as moderate or poor agreement.

Standard errors for kappa have been derived, and are presented in computer output by many statistical packages. These may be used to derive a *P*-value corresponding to the null hypothesis of no association between the ratings on the two occasions, or by the two raters. In general, *such P-values are not of interest*, because the null hypothesis of no association is not a reasonable one.

We will illustrate the calculation of kappa statistics using data from a study of the way in which people tend to explain problems with their health. We will do this first using a binary classification, and then a fuller 4-category classification.

Example 36.3: Binary classification

Table 36.2 summarizes data from a study in which 179 men and women filled in a Symptom Interpretation Questionnaire on two occasions three years apart. On the basis of this questionnaire they were classified according to whether or not they tended to provide a *normalizing* explanation of symptoms. This means discounting symptoms, externalizing them and explaining them away as part of normal experience. It can be seen that while 76 participants were consistently classified as normalizers, and 47 as non-normalizers, the classification changed for a total of 56 participants. More participants were classified as normalizers on the second than the first occasion.

The *observed* proportion of agreement between the assessment on the two occasions, denoted by A_{obs} is therefore given by:

$$A_{obs} = (76 + 47)/179 = 123/179 = 0.687\ (68.7\,\%)$$

Part (b) of Table 36.2 shows the number of agreements and disagreements that would be expected between the two classifications on the basis of chance alone. These expected numbers are calculated in a similar way to that described for the

Table 36.2 Classification of 179 men and women as 'symptom normalizers' or not, on two measurement occasions three years apart. Data kindly provided by Dr David Kessler.

(a) Observed numbers

	Second classification		
First classification	Normalizer	Non-normalizer	Total
Normalizer	76	17	93
Non-normalizer	39	47	86
Total	115	64	179

(b) Expected numbers

	Second classification		
First classification	Normalizer	Non-normalizer	Total
Normalizer	59.7	33.3	93
Non-normalizer	55.3	30.7	86
Total	115	64	179

chi-squared test in Chapter 17. The overall proportion classified as normalizers on the second occasion was 115/179. If this classification was unrelated to that on the first, then one would expect this same proportion of second occasion normalizers in each first occasion group, that is $115/179 \times 93 = 59.7$ classified as normalizers on both occasions, and $115/179 \times 86 = 55.3$ of those classified as non-normalizers on the first occasion classified as normalizers on the second. Similarly $64/179 \times 93 = 33.3$ of those classified as normalizers on the first occasion would be classified as non-normalizers on the second, while $64/179 \times 86 = 30.7$ would be classified as non-normalizers on both occasions. The *expected* proportion of chance agreement is therefore:

$$A_{\text{exp}} = (59.7 + 30.7)/179 = 0.505\ (50.5\%)$$

Giving a kappa statistic of:

$$\kappa = (0.687 - 0.505)/(1 - 0.505) = 0.37$$

This would usually be interpreted as representing at most moderate agreement between the two classifications made over the three-year follow-up period.

Example 36.4: Categorical classification

Table 36.3(a) shows a more complete version of the data presented in Table 36.2, with each participant now assessed as belonging to one of four groups according to the way in which they tended to explain symptoms. Those classed as non-normalizers (see earlier explanation) have been divided into *somatizers*, those who

Table 36.3 Classification of the dominant style for explaining symptoms of 179 men and women as normalizers, somatizers, psychologizers or no dominant style, on two measurement occasions three years apart. Data kindly provided by Dr David Kessler.

(a) Observed numbers

Dominant style at first classification	Dominant style at second classification				Total
	Normalizer	Somatizer	Psychologizer	None	
Normalizer	76	0	7	10	93
Somatizer	2	0	3	1	6
Psychologizer	17	1	15	8	41
None	20	3	5	11	39
Total	115	4	30	30	179

(b) Expected numbers of agreements

Dominant style at first classification	Dominant style at second classification				Total
	Normalizer	Somatizer	Psychologizer	None	
Normalizer	59.7				93
Somatizer		0.1			6
Psychologizer			6.9		41
None				0.2	39
Total	115	4	30	30	179

tend to explain their symptoms as indicating a potentially more serious physical illness, *psychologizers*, those who tend to give psychological explanations for their symptoms, and those with no dominant style. The *observed* proportion of agreement between the two occasions using the four category classification is:

$$A_{\text{obs}} = (76 + 0 + 15 + 11)/179 = 102/179 = 0.570\ (57.0\,\%)$$

The expected numbers for the various combinations of first and second occasion classification can be calculated in exactly the same way as argued in the two-category example. For the kappa statistic, we need these only for the numbers of agreements; these are shown in Table 36.3(b).

$$A_{\text{exp}} = (59.7 + 0.1 + 6.9 + 0.2)/179 = 72.9/179 = 0.407\ (40.7\,\%)$$

giving

$$\kappa = \frac{A_{\text{obs}} - A_{\text{exp}}}{1 - A_{\text{exp}}} = (0.570 - 0.407)/(1 - 0.407) = 0.27$$

representing poor to moderate agreement.

As the *number* of categories increases, the value of kappa will tend to decrease, because there are more opportunities for misclassification. Further, for ordinal

measures we may wish to count classification into adjacent categories as *partial* agreement. For instance, classification into adjacent categories might count as 50% agreement, such as normalizers classified as somatizers and vice versa in Table 36.3. This is done using a **weighted kappa** statistic, in which the observed and expected proportions of agreement are modified to include partial agreements, by assigning a weight between 0 (complete disagreement) and 1 (complete agreement) to each category. Kappa statistics can also be derived when there are more than two raters: for more details see Fleiss (1981) or Dunn (1989).

Numerical variables: reliability and the intraclass correlation coefficient

We now describe how to quantify the amount of measurement error in a numerical variable. As with the kappa statistic, this may be done using replicate measurements of the variable: for example measurement of blood pressure made on the same patient by two observers at the same time, or using the same automated measuring device on two occasions one week apart.

The **reliability** of a measurement is formally defined as the ratio of the variance of the 'true' (underlying) values between individuals to the variance of the observed values, which is a combination of the variation between individuals (σ_u^2) and measurement error (σ_e^2). It can be measured using the **intraclass correlation coefficient** (ICC), defined in Section 31.4 in the context of random-effects models:

$$\text{Intraclass correlation coefficient (ICC)} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

$$\sigma_u^2 = \text{variance between true measurements}$$

$$\sigma_e^2 = \text{measurement error variance}$$

Here the 'clusters' are the individuals on whom measurements are made, and the observations within clusters are the repeated measurements on the individuals. ICC can range from 0 to 1, with the maximum of 1 corresponding to complete reliability, which is when there is no measurement error, $\sigma_e^2 = 0$. The smaller the amount of measurement error, the smaller will be the increase in the variability of the observed measurements compared to the true measurements and the closer will be the reliability (and ICC) to 1. If all individuals have the same 'true' value, then $\sigma_u^2 = 0$ and ICC $= 0$; all observed variation is due to measurement error.

The intraclass correlation coefficient may be estimated using a one-way analysis of variance (see Chapter 11), or by using a simple random-effects model (see Chapter 31). When there are paired measurements, the ICC can also be derived by calculating the Pearson (product moment) correlation with each pair entered twice, once in reverse order.

Example 36.5

As part of a case–control study investigating the association between asthma and intake of dietary antioxidants (measured using food frequency questionnaires), replicate measurements of selenium intake were made 3 months after the original measurements, for 94 adults aged between 15 and 50 years. Figure 36.2 is a scatter plot of the pairs of measurements; note that because estimated selenium intake was positively skewed the measurements are plotted on a log scale (see Chapter 13). While there is clearly an association between the measurements on the first and second occasions, there is also substantial between-occasion variability.

The mean and standard deviation of log selenium intake (measured in log (base e) μg/week) in the 94 subjects with repeat measurements were 3.826 (s.d. = 0.401) on the first occasion and 3.768 (s.d. = 0.372) on the second occasion. There was some evidence that measured intake declined between the two measurements (mean reduction 0.058, 95% CI −0.008 to 0.125, $P = 0.083$). The estimated components of variance were:

$$\text{Within-subject (measurement error) variance, } \sigma_e^2 = 0.0535$$
$$\text{Between-subject variance, } \sigma_u^2 = 0.0955$$
$$\text{Total variance} = \sigma_u^2 + \sigma_e^2 = 0.1491$$

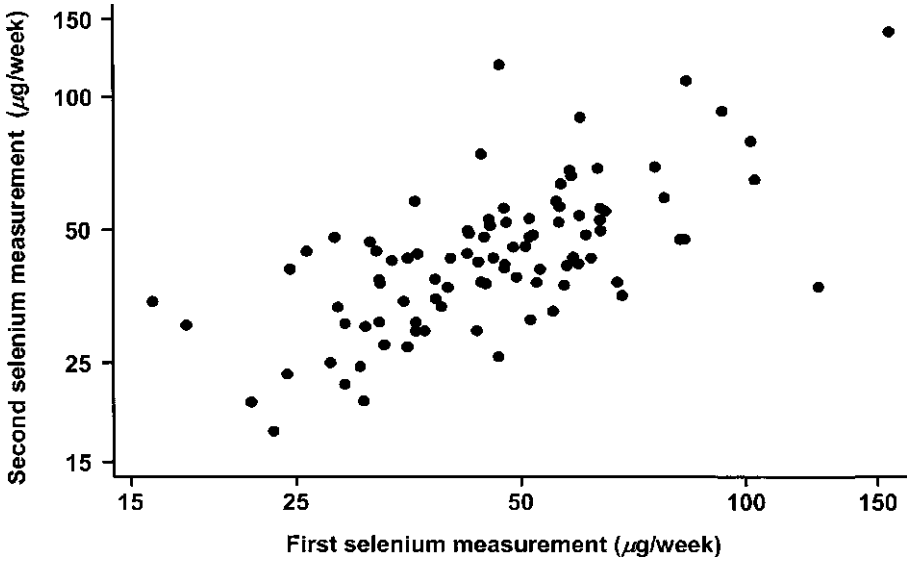


Fig. 36.2 Scatter plot of weekly selenium intake (μg/week) on a log scale among 94 participants in a study of asthma and intake of antioxidant vitamins, measured using a food frequency questionnaire on two occasions three months apart. Data displays and analyses from the FLAG study (Shaheen SO, Sterne JAC, Thompson RL, Songhurst CE, Margetts BM, Burney PGJ (2001) *American Journal of Respiratory and Critical Care Medicine* **164**: 1823–1828).

Therefore,

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = 0.0955/0.1491 = 0.6410$$

Thus in this example, 64.1 % of the total variability was between-subject variability, indicating fairly good reliability of assessing selenium intake using a single application of a food frequency questionnaire.

Links between weighted kappa and the intraclass correlation coefficient

For ordered categorical variables, there is a close link between the weighted kappa statistic (defined above) and the intraclass correlation coefficient. If the variable has k categories, and the weight, w_{ij}, for a subject in category i at the first measurement and j at the second measurement is chosen to be:

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}$$

then the value of the weighted kappa will be very close to the ICC. For example, for an ordered categorical variable with four categories the weights would be

$$\begin{aligned}
w_{11} &= w_{22} = w_{33} = w_{44} = 1 - \frac{0}{3^2} = 1 \\
w_{12} &= w_{21} = w_{23} = w_{32} = w_{34} = w_{43} = 1 - \frac{1^2}{3^2} = 0.889 \\
w_{13} &= w_{31} = w_{24} = w_{42} = 1 - \frac{2^2}{3^2} = 0.556 \\
w_{14} &= w_{41} = 1 - \frac{3^2}{3^2} = 0
\end{aligned}$$

36.4 NUMERICAL VARIABLES: METHOD COMPARISON STUDIES

We will now consider analyses appropriate to **method comparison studies**, in which two different methods of measuring the same underlying (true) value are compared. For example, lung function might be measured using a spirometer, which is expensive but relatively accurate, or with a peak flow meter, which is cheap (and can therefore be used by asthma patients at home) but relatively inaccurate. The appropriate analysis of such studies was described, in an influential paper, by Bland and Altman (1986).

Example 36.6

We will illustrate appropriate methods for the analysis of method comparison studies using data on 1236 women who participated in the British Women's

Regional Heart Study. The women were asked to report their weight as part of a general questionnaire, and their weight was subsequently measured using accurate scales. Figure 36.3 is a scatter plot of self-reported versus measured weight.

The two measures are clearly strongly associated: the Pearson correlation between them is 0.982. It is important to note, however, that the correlation measures the strength of association between the measures and *not* the agreement between them. For example, if the measurements made with the new method were exactly twice as large as those made with the standard method then the correlation would be 1, even though the new method was badly in error. Further, the correlation depends on the range of the true quantity in the sample. The correlation will be greater if this range is wide than if it is narrow.

The diagonal line in Figure 36.3 is the **line of equality**: the two measures are in perfect agreement only if all measurements lie along this line. It can be seen that more of the points lie below the line than above it, suggesting that self-reported weight tends to be lower than measured weight.

Bland and Altman suggested that the extent of agreement could be examined by plotting the *differences* between the pairs of measurements on the vertical axis, against the *mean* of each pair on the horizontal axis. Such a plot (often known as a **Bland–Altman plot**) is shown in Figure 36.4. If (as here) one method is known to be accurate, then the mean difference will tell us whether there is a systematic **bias** (a tendency to be higher or lower than the true value) in the other measurement. In

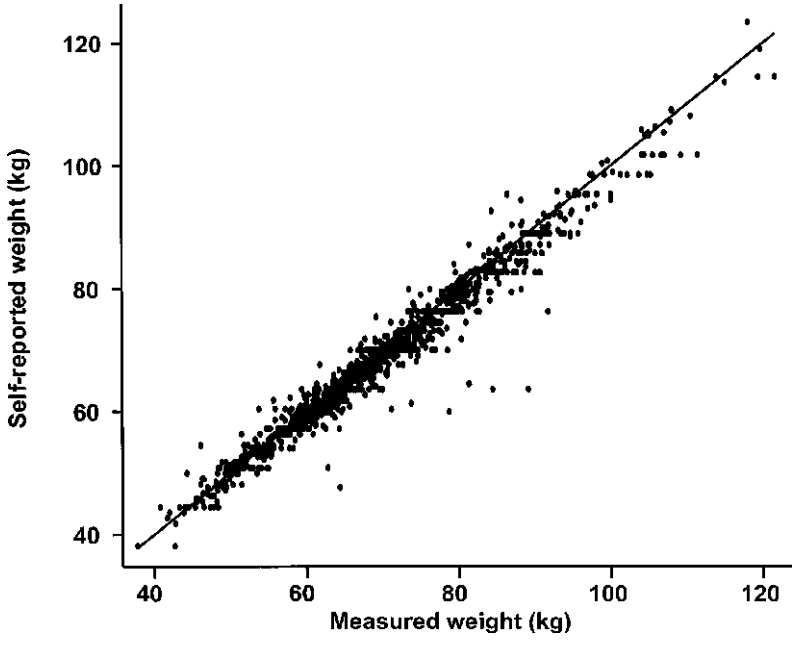


Fig. 36.3 Scatter plot of self-reported versus measured weight (kg) in 1236 women who participated in the British Regional Women's Heart Study. The solid line is the *line of equality*. Data displays and analyses by kind permission of Dr Debbie Lawlor and Professor Shah Ebrahim.

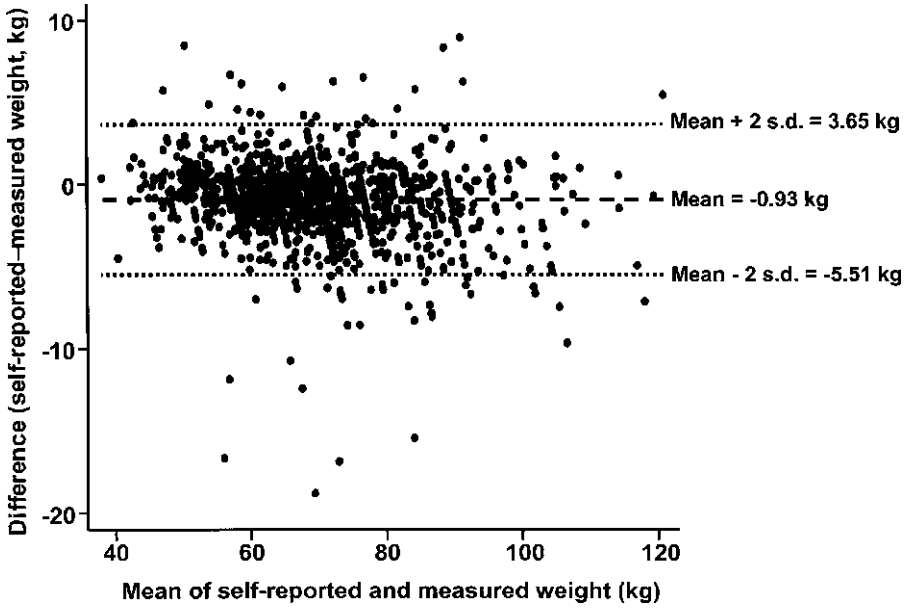


Fig. 36.4 Scatter plot (Bland–Altman plot) of self-reported minus measured weight (vertical axis) against mean of self-reported and measured weight (horizontal axis) in 1236 women who participated in the British Regional Women's Heart Study. The dashed horizontal line corresponds to the mean difference (−0.93 kg) while the dotted horizontal lines correspond to the 95% limits of agreement.

this example, mean self-reported weight was 68.88 kg, while the mean measured weight was 69.85 kg. The mean difference between self-reported and measured weight was −0.93 kg (95% CI −1.07 to −0.80 kg). There was thus a clear tendency for the women to under-report their weight, by an average of 0.93 kg. This is shown by the dashed horizontal line in Figure 36.4.

The dotted horizontal lines in Figure 36.4 correspond to the **95% limits of agreement**, given by the mean difference plus or minus twice the *standard deviation* of the differences. If the differences are normally distributed then approximately 95% of differences will lie within this range. In this example the 95% limits of agreement are from −5.51 kg to 3.65 kg. Inspection of Figure 36.4 also shows that the differences were negatively skewed; there were more large negative differences than large positive ones. Further, there was a tendency for greater (negative) differences with greater mean weight.

Note that *the difference should not be plotted against either of the individual measurements*, because of the problem of '**regression to the mean**' described in Section 36.5.

Having calculated the mean difference and the 95% limits of agreement, it is for the investigator to decide whether the methods are sufficiently in agreement for one (perhaps the cheaper method) to be used in place of the other. In this example, the systematic underreporting of weight in questionnaires, and the reduced

accuracy, would have to be considered against the increased cost of inviting women to a visit at which their weight could be measured accurately.

36.5 IMPLICATIONS FOR INTERPRETATION

The problems that may result from errors that occur when measuring outcome or exposure variables are summarized in Table 36.4. Each type of problem will be addressed in the sub-sections below. Note that the focus here is on random errors, in the sense that we are assuming that any errors in measuring a variable are independent of the values of other variables in the dataset.

Table 36.4 Summary of implications of random misclassification and measurement error.

	Type of error	
Type of variable	Misclassification (binary/categorical variable)	Measurement error (numerical variable)
Outcome	Regression dilution bias	Regression to the mean
Exposure	Regression dilution bias Potential problems if adjusting for confounders	

Regression dilution bias

Regression dilution bias means that the estimated regression coefficient of the exposure-effect estimate has been biased towards the null value of no exposure effect, so that the magnitude of the association between the exposure and outcome will tend to be underestimated:

1 For a numerical *exposure* variable, the degree of bias depends on the intraclass correlation coefficient (ICC). For linear regression the relationship is:

$$\text{Estimated coefficient} = \text{correct coefficient} \times \text{ICC}$$

For other regression models, such as logistic regression and Cox regression, the same relationship holds approximately, providing that the correct coefficient is not too large, and that the measurement error variance is not too large compared to the variance between true measurements. Frost and Thompson (2000) compare a number of methods to correct for regression dilution bias.

2 The estimated effect of a categorical (or binary) *exposure* variable can be corrected using replicate measurements on some or all individuals. However, methods to do this are more complex than those for numerical exposure variables, because *the errors will be correlated with the true values*. For example, if the true value of a binary variable is 0 then the size of the error is either 0 or 1, while if the true value is 1 then the size is 0 or -1. For this reason, applying

methods appropriate for numerical exposure variables will *overcorrect* the regression dilution in the effect of a binary exposure variable. Appropriate methods for this situation are reviewed by White *et al*. (2001).

3 For a binary *outcome* variable, if the sensitivity and specificity with which it was measured are known then estimated odds ratios from logistic regression may be corrected, as described by Magder and Hughes (1997).

4 Measurement error in a numerical *outcome* variable does *not* lead to regression dilution bias, although the greater the measurement error the lower the precision with which exposure-outcome associations are estimated.

As mentioned above, correcting for regression dilution bias requires that we make *replicate measurements* on some or all subjects. If each subject-evaluation costs the same amount, then we must trade off the benefits of increasing the number of *subjects* in our study with the benefits of increasing the number of *measurements per subject*. Phillips and Davey Smith (1993) showed that it will sometimes be better to recruit a smaller number of subjects with each evaluated on more than one occasion, because this leads to more precise estimates of subjects' exposure levels and hence to reduced bias in exposure effect estimates. They suggested that attempts to anticipate and control bias due to exposure measurement error should be given at least as high a priority as that given to sample size assessment in the design of epidemiological studies.

Before applying any method to correct regression coefficients for measurement error, it is important to be aware of the potential problems associated with measurement error in a number of exposure variables included in multivariable models, as described in the next sub-section.

The effects of measurement error and misclassification in multivariable models

When there are measurement errors in a *number* of exposure variables, and we wish to control for the possible confounding effects of each on the other, the effects are less straightforward to predict than is the case when we are considering the association between an outcome and a *single* exposure variable. For example, consider the situation in which:

1 the correct (underlying) value of exposure A is associated with the disease outcome, but is measured with substantial error;

2 the correct (underlying) value of exposure B is not associated with the disease outcome after controlling for exposure A; and

3 the amount of measurement error in exposure B is much less than the measurement error in exposure A.

In this situation, including A and B in a multivariable model may give the misleading impression that B is associated with the outcome, and that A is not associated with the outcome after controlling for B: the opposite of the true situation if there were no measurement error.

Such possible problems are frequently ignored. Note that the bias caused by differing amounts of measurement error in the two exposure variables may act in either direction, depending on:

1 the direction of the association between the two variables;
2 the relative amounts of error in measuring them; and
3 whether the measurement errors are correlated.

Regression to the mean

Regression to the mean refers to a phenomenon first observed by Galton when he noticed that the heights of sons tended to be closer to the overall mean than the heights of their fathers. Thus, tall fathers tended to have sons shorter than themselves, while the opposite was true for short fathers.

The same phenomenon occurs whenever two repeat measurements are made, and where they are subject to measurement error. Larger values of the first measurement will, on average, have positive measurement errors while smaller values of the first measurement will, on average, have negative measurement errors. This means that the repeat measurement will tend to be smaller if the first measurement was larger, and larger if the first measurement was smaller. It follows that the size of the first measurement will be negatively associated with the difference between the two measurements.

The implications of this will be explained in more detail by considering the repeated measurement of blood pressure and the assessment of anti-hypertensive drugs in reducing blood pressure. For a more detailed discussion of regression to the mean, and methods to correct for it, see Hayes (1988).

Example 36.7

Figure 36.5 shows the relationship between two diastolic blood pressure readings taken 6 months apart on 50 volunteers, while Figure 36.6 is a scatter plot of the difference between the two readings (vertical axis) against the initial reading (horizontal axis). This gives the impression that there is a downward gradient, so that those with a high initial level have a reduced blood pressure 6 months later, while the opposite is true for those with an initial low level. However, for the reasons explained above, this downward gradient may be the result of measurement error. If there is *no* association between the *true* reduction and the *true* initial value, the regression coefficient β_{obs} for the *observed* association between the difference and the initial value is given by:

$$\beta_{obs} = \text{ICC} - 1$$

in absence of 'true' association

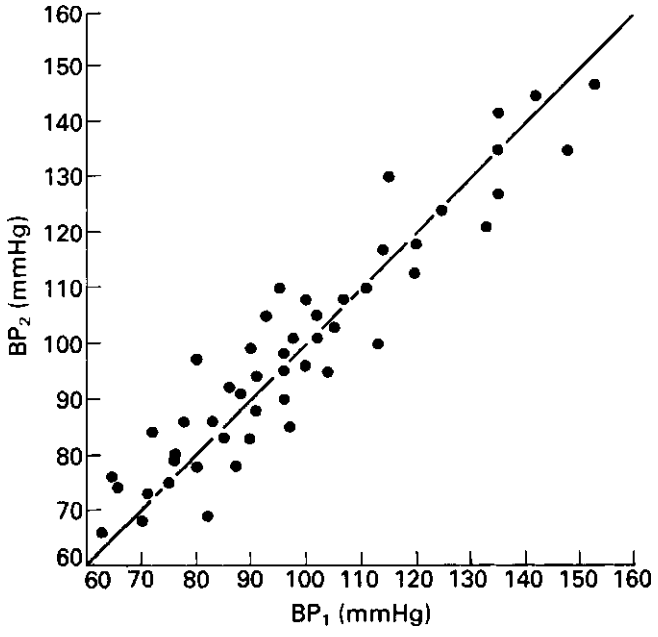


Fig. 36.5 The relationship between two diastolic blood pressure readings taken six months apart on 50 volunteers, showing little change on average. The straight line is the relationship that would be seen if the readings on the two occasions were the same.

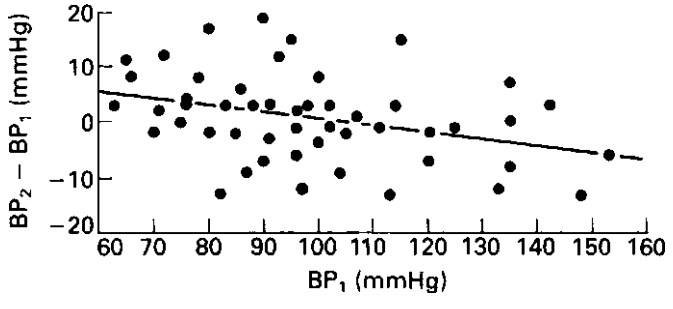


Fig. 36.6 Change in diastolic blood pressure plotted against initial value. An artificial negative correlation ($r = -0.35$, d.f. $= 48$, $P = 0.013$) is observed. The straight line is the regression line corresponding to this association.

Thus the greater the measurement error variance, the smaller is the ICC and so the greater is the slope of this apparent negative association.

Thus measurement error has important implications when the focus of interest is *change* in outcome measurement, for example in a clinical trial to evaluate the ability of an anti-hypertensive drug to reduce blood pressure:

1 If, as is often the case, the trial is confined to people with high initial diastolic blood pressure, say 120 mmHg or above, then it can be seen from Figure 36.6 that their repeated blood pressure measurements would show an average

reduction, even in the absence of any treatment. It is therefore essential to have a control group, and to compare any apparent reduction in the treatment group with that in the control group.

2 Analyses investigating whether the size of any change in blood pressure is related to the initial value must correct for *regression to the mean*. Blomqvist (1977) suggested that the true regression coefficient can be estimated from the observed regression coefficient using:

$$\beta_{\text{true}} = \frac{\beta_{\text{obs}} + (1 - \text{ICC})}{\text{ICC}}$$

To apply this method in practice requires an external estimate of the within-person (measurement error) variance.

3 Oldham (1962) suggested plotting the difference, $BP_2 - BP_1$, against the *average* of the initial and final blood pressure readings, $\frac{1}{2}(BP_1 + BP_2)$, rather than against the initial reading as shown in Figure 36.7, to correct for regression to the mean. (Note the similarity with Bland–Altman plots, described in Section 36.4.) The correlation is attenuated to −0.19, suggesting that much or all of the apparent association between blood pressure reduction and initial blood pressure was caused by regression to the mean. However, there are at least two circumstances when this can give misleading results. The Oldham plot will show a positive association when the true change is unrelated to the initial level, if:
   - the true change differs between individuals; or
   - individuals have been selected on the basis of high initial values.

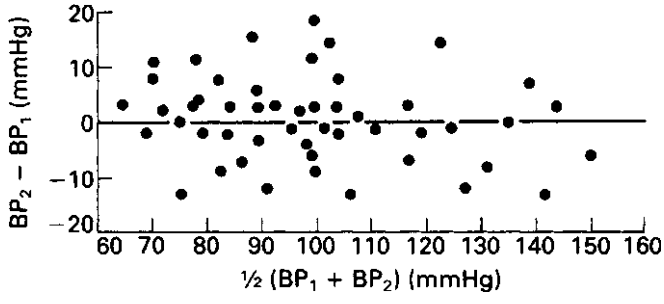


Fig. 36.7 Change in diastolic blood pressure plotted against the average of the initial and final readings, The correlation is attenuated to −0.19, suggesting little or no relationship between $BP_2 - BP_1$ and blood pressure.

CHAPTER 37

Measures of association and impact

37.1 Introduction

37.2 Measures of association

Risk ratios

Odds ratios

Rate ratios

Comparison of the rate ratio, risk ratio and odds ratio

37.3 Measures of the impact of an exposure

Attributable risk

Comparing attributable and relative measures

Population attributable risk

Potential impact of reducing prevalence of exposure

37.4 Measures of the impact of a treatment or intervention

Efficacy

Number needed to treat

Number needed to harm

37.5 Estimates of association and impact from multivariable analyses

37.1 INTRODUCTION

In this chapter we focus on the different measures that are used to assess the **impact** of an exposure or of a treatment on the amount of disease in a **population**. We start by summarizing the three different ratio measures of the association between an exposure (or treatment) and outcome, used throughout the book, and show how these relate to measures of impact.

37.2 MEASURES OF ASSOCIATION

Table 37.1 summarizes the three ratio measures that we use to assess the strength of the association between an exposure (or treatment) and an outcome. These are the risk ratio, the rate ratio and the odds ratio.

Risk ratios

A risk ratio > 1 implies that the risk of disease is higher in the exposed group than in the unexposed group, while a risk ratio < 1 occurs when the risk is lower in the exposed group, suggesting that exposure may be protective. A risk ratio of 1 occurs when the risks are the same in the two groups and is equivalent to no association between the exposure and the disease. The further the risk ratio is from 1, the stronger the association. See Chapter 16 for methods to derive confidence intervals for the RR.

Table 37.1 Summary of ratio measures of the association between exposure and disease, and the different study designs in which they can be estimated.

Definitions of different ratio measures	Study design(s) in which they can be estimated			
	Longitudinal (complete follow-up)	Longitudinal (incomplete follow-up)	Cross-sectional	Case–control
Risk ratio $= \frac{\text{risk in exposed group}}{\text{risk in unexposed group}}$	Yes	No	Yes	No
Rate ratio $= \frac{\text{rate in exposed group}}{\text{rate in unexposed group}}$	Yes	Yes	No	No
Odds ratio $= \frac{\text{odds in exposed group}}{\text{odds in unexposed group}}$	Yes	No	Yes	Yes

Odds ratios

Interpretation of odds ratios is the same as that for risk ratios (see above), but the odds ratio is always further away from 1 than the corresponding risk ratio. Thus:

- if $RR > 1$ then $OR > RR$;
- if $RR < 1$ then $OR < RR$.

For a rare outcome (one in which the probability of the event not occurring is close to 1) the odds ratio is approximately equal to the risk ratio (since the odds is approximately equal to the risk).

Rate ratios

While the calculation of the risk is based on the population at risk at the start of the study, the rate is based on the total person-years at risk during the study and reflects the changing population at risk. This was illustrated for a cohort study in Figure 22.2. When the outcome is not rare, the risk ratio will change over time, so that the rate ratio (providing that it is constant over time) may be a more appropriate measure of the association between exposure and disease. In particular, if all subjects experience the disease outcome by the end of the study, then the risk ratio will be 1 even if the time to event was much greater in the exposed than the unexposed group (or vice versa).

Comparison of the rate ratio, risk ratio and odds ratio

It was shown in Chapters 16 and 23 that for a rare outcome

$$\text{Risk} \approx \text{Odds} \approx \text{Rate} \times \text{Time}$$

so that

$$\text{Risk ratio} \approx \text{Odds ratio} \approx \text{Rate ratio}$$

For a **common disease**, however, the *three measures are different*, and will lead to three different measures of association between exposure and disease. The preferred choice in longitudinal studies is to use rate ratios (or hazard ratios when data on times to event occurrences are available and disease rates change over time: see Chapter 26). The rate ratio is the only choice when follow-up is incomplete, or individuals are followed for differing lengths of time. The use of risk ratios is more appropriate, however, when assessing the protective effect of an exposure or intervention, such as a vaccine, which it is believed offers full protection to some individuals but none to others, rather than partial protection to all (Smith *et al*., 1984).

The risk ratio and odds ratio can both be estimated from longitudinal studies with complete follow-up and from cross-sectional studies. Although the risk ratio would generally be regarded as more easily interpretable than the odds ratio, the odds ratio is often used because the statistical properties of procedures based on the odds ratio are generally better. In case–control studies the odds ratio is always used as the measure of effect.

37.3 MEASURES OF THE IMPACT OF AN EXPOSURE

We now show how ratio measures (of the strength of the association between exposure and disease) relate to measures of the impact of exposure. The formulae we present apply identically whether risks or rates are used.

Attributable risk

The risk ratio assesses how much more likely, for example, a smoker is to develop lung cancer than a non-smoker, but it gives no indication of the magnitude of the excess risk in absolute terms. This is measured by the **attributable risk**:

$$\begin{aligned}\text{Attributable risk (AR)} &= \text{risk among exposed} - \text{risk among unexposed}\\ &= \text{the } \textbf{risk difference} \text{ (see Section 16.3)}\end{aligned}$$

Example 37.1

Table 37.2 shows hypothetical data from a cohort study to investigate the association between smoking and lung cancer. Thirty-thousand smokers and 60 000 non-smokers were followed for a year, during which time 39 of the smokers and six of the non-smokers developed lung cancer. Thus the risk ratio was:

Table 37.2 Hypothetical data from a one year cohort study to investigate the association between smoking and lung cancer. The calculations of relative and attributable risk are illustrated.

	Lung cancer	No lung cancer	Total	One year risk
Smokers	39	29 961	30 000	1.30/1000
Non-smokers	6	59 994	60 000	0.10/1000
Total	45	89 955	90 000	
RR $= \frac{1.30}{0.10} = 13.0$	AR $= 1.30 - 0.10 = 1.20$/1000		Prop AR $= \frac{1.20}{1.30} = 0.923$ or 92.3 %	

$$\text{RR} = \frac{39/30000}{6/60000} = \frac{1.30}{0.10} = 13.0$$

so that there was a very strong association between smoking and lung cancer. The attributable risk of lung cancer due to smoking, given by the difference between the risks among smokers and non-smokers, was:

$$\text{AR} = 1.30 - 0.10 = 1.20 \text{ cases per 1000 per year}$$

Attributable risk is sometimes expressed as a proportion (or percentage) of the total incidence rate among the exposed, and is then called the **proportional attributable risk**, the attributable proportion (exposed), the attributable fraction (exposed) or the **aetiologic fraction (exposed)**.

$$\text{Proportional AR} = \frac{\text{risk among exposed} - \text{risk among unexposed}}{\text{risk among exposed}} = \frac{(\text{RR} - 1)}{\text{RR}}$$

In the example, the proportional attributable risk was $1.20/1.30 = 0.923$, suggesting that smoking accounted for 92.3 % of all the cases of lung cancer among the smokers.

Comparing attributable and relative measures

Example 37.2

Table 37.3 shows the relative and attributable rates of death from selected causes associated with heavy cigarette smoking. The association has been most clearly demonstrated for lung cancer and chronic bronchitis, with rate ratios of 32.4 and 21.2 respectively. If, however, the association with cardiovascular disease,

Table 37.3 Relative and attributable rates of death from selected causes, 1951–1961, associated with heavy cigarette smoking by British male physicians. Data from Doll & Hill (1964) *British Medical Journal* **1**, 1399–1410, as presented by MacMahon & Pugh (1970) *Epidemiology – Principles and Methods*. Little, Brown & Co., Boston (with permission).

	Age-standardized death rate (per 1000 person-years)			
Cause of death	Non-smokers	Heavy smokers	RR	AR
Lung cancer	0.07	2.27	32.4	2.20
Other cancers	1.91	2.59	1.4	0.68
Chronic bronchitis	0.05	1.06	21.2	1.01
Cardiovascular disease	7.32	9.93	1.4	2.61
All causes	12.06	19.67	1.6	7.61

although not so strong, is also accepted as being causal, elimination of smoking would save even more deaths due to cardiovascular disease than due to lung cancer: 2.61 compared to 2.20 for every 1000 smoker-years at risk. Note that the death rates were age standardized to take account of the differing age distributions of smokers and non-smokers, and of the increase in death rates with age (see Chapter 25).

In summary, the risk (or rate) ratio measures the strength of an association between an exposure and a disease outcome. The attributable risk (or rate), on the other hand, gives a better idea of the excess risk of disease experienced by an individual as the result of being exposed.

Population attributable risk

It is important to realize that the overall impact of an exposure on disease in the population also depends on how prevalent the exposure is. In population terms a rare exposure with a high associated risk ratio may be less serious in the total number (or proportion) of deaths that it will cause than a very common exposure with a lower associated risk ratio. The impact at the population level is assessed by the excess overall risk (or rate) in the population as compared with the risk (or rate) among the unexposed. The resulting measure is the **population attributable risk**:

$$\text{Population AR} = \text{overall risk} - \text{risk among unexposed}$$

This may also be expressed as a proportion (or percentage) of the overall risk. The resulting measure is the **population proportional attributable risk**, alternatively named the **aetiologic fraction (population)** or the attributable fraction (population).

$$\text{Population proportional AR} = \frac{\text{overall risk} - \text{risk among unexposed}}{\text{overall risk}} = \frac{\text{prevalence}_{\text{exposure}}(\text{RR} - 1)}{1 + \text{prevalence}_{\text{exposure}}(\text{RR} - 1)}$$

Figure 37.1 shows how the value of the population proportional attributable risk increases independently with the prevalence of the exposure and with the size of the risk ratio. If all the population are exposed (prevalence = 100%), then the value of the population proportional attributable risk is the same as the proportional AR (exposed) defined above.

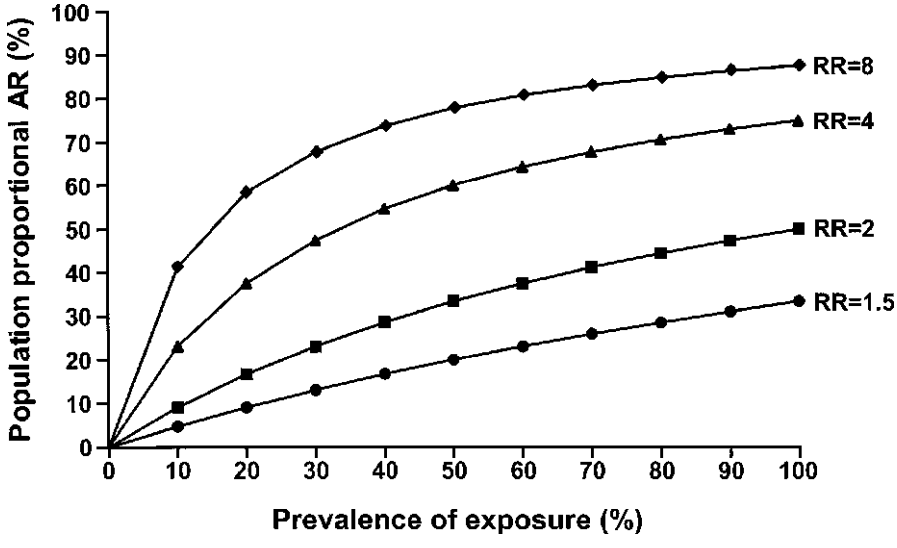


Fig. 37.1 Relationship between population proportional attributable risk and prevalence of exposure for various values of the risk ratio.

Potential impact of reducing prevalence of exposure

The population attributable and proportional attributable risks give a measure of the burden of disease in the population associated with a particular exposure. They also give a measure of the impact that would be achieved by a totally successful intervention which managed to eliminate the exposure. This is a theoretical maximum impact that is unlikely to be realized in practice. For example, it is unlikely that any approach to control smoking would result in all smokers giving up. If the intervention *reduces* the prevalence of exposure by r%, then the actual impact will be as follows:

$$\text{Percentage impact} = r\% \times \text{Population proportional AR}$$

Example 37.3

Figure 37.2 illustrates the difference between potential impact and population proportional attributable risk in a hypothetical population of 1000 children, followed for one year without loss to follow-up. There are 400 children exposed to a risk factor that is associated with a three-fold risk of death, and 600 children who are not exposed. The 600 children in the unexposed group experience a mortality rate of 50/1000/year which means that $600 \times 50/1000 = 30$ of them will die during the year. If the 400 children in the exposed group were at the same risk as the unexposed children, then $400 \times 50/1000 = 20$ of them would die. However, they are at 3 times this risk. Their mortality rate is therefore 150/child/year, which translates into $400 \times 150/1000 = 60$ deaths during the year, an excess of 40 deaths associated with exposure. Thus if it were possible to eliminate exposure to the risk factor, the total number of deaths per year would be reduced by 40, giving a total of 50 rather than 90 deaths a year. The population proportional attributable risk, which is the percentage of deaths attributable to exposure, equals 40/90, or 44%.

Suppose now that an intervention took place which successfully reduced the prevalence of exposure by one half, that is from 40% to 20%. The right hand panel in Figure 37.2 shows that there would then be 70 deaths a year. As the size of the exposed group would be halved, the number of excess deaths

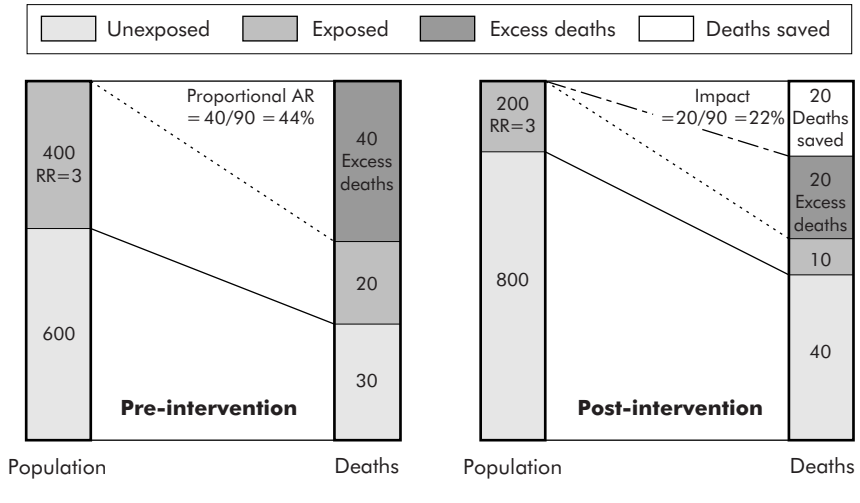


Fig. 37.2 Example showing potential impact of an intervention, assuming (i) 40% of population exposed pre-intervention, (ii) RR associated with exposure equals 3, (iii) mortality rate among unexposed equals 50/1000/year, and (iv) the intervention reduces the prevalence of exposure by 50%.

would also be halved, and would now be 20 rather than 40. Such an intervention would therefore prevent 20 of the pre-intervention total of 90 deaths. That is, its impact would be 20/90, or 22%.

37.4 MEASURES OF THE IMPACT OF A TREATMENT OR INTERVENTION

Efficacy

The **efficacy** of a treatment or intervention is measured by the proportion of cases that it prevents. Efficacy is directly calculated from the risk ratio (or rate ratio) comparing disease outcome in the treated versus control group. For a successful treatment (or intervention) this ratio will be less than 1.

$$\text{Efficacy} = 1 - \text{RR}$$

Example 37.4

Table 37.4 shows the *hypothetical* results from a randomized controlled trial of a new influenza vaccine. A total of 80 cases of influenza occurred in the placebo group. If this group had instead received vaccination one would have expected only 8.3% (the rate experienced by the vaccinated group) of them to have developed influenza, that is $220 \times 0.083 = 18.3$ cases. The saving would therefore have been $80 - 18.3 = 61.7$ cases, giving an efficacy of $61.7/80 = 77.2\%$.

The efficacy can be calculated directly from the risk ratio, which gives the risk in the vaccinated group as a proportion of the risk in the control group. If the vaccination had no effect, the risks would be the same and the risk ratio would equal 1. In this case, the risk is considerably lower in the vaccine group. The risk ratio equals 0.228, considerably less than 1. In other words the risk of influenza in the vaccine group is only 0.228 or 22.8% of that in the placebo group. The vaccine has therefore prevented 77.2% of influenza cases.

Table 37.4 Results from an influenza vaccine trial, previously presented in Table 16.2.

	Influenza		
	Yes	No	Total
Vaccine	20 (8.3%)	220 (91.7%)	240
Placebo	80 (36.4%)	140 (63.6%)	220
Total	100 (21.7%)	360 (78.3%)	460

$$\text{RR} = \frac{20/240}{80/220} = \frac{0.083}{0.364} = 0.228; \quad \text{Efficacy} = 1 - 0.228 = 0.772, \text{ or } 77.2\%$$

The **confidence interval for efficacy** is calculated from the confidence interval for risk ratio, as follows. Recall from Section 16.5 that:

$$\begin{gathered}95\%\ \text{CI (RR)} = \text{RR/EF to RR} \times \text{EF},\\ \text{where EF} = \exp[1.96 \times \text{s.e.}(\log \text{RR})]\\ \text{and s.e.}(\log \text{RR}) = \sqrt{[(1/d_1 - 1/n_1) + (1/d_0 - 1/n_0)]}\end{gathered}$$

Since efficacy equals one minus RR, its 95% confidence interval is obtained by subtracting each of the RR confidence limits from one.

$$95\%\ \text{CI (Efficacy)} = 1 - \text{RR} \times \text{EF to } 1 - \text{RR/EF}$$

Note that the lower efficacy limit is obtained from the upper RR limit, and the upper efficacy limit from the lower RR limit. In this example:

$$\begin{aligned}\text{s.e.}(\log \text{RR}) &= \sqrt{[(1/20 - 1/240) + (1/80 - 1/220)]} = 0.2319\\ \text{EF} &= \exp(1.96 \times 0.2319) = \exp(0.4546) = 1.5755\\ 95\%\ \text{CI (RR)} &= \text{RR/EF to RR} \times \text{EF} = 0.228/1.5755 \text{ to } 0.228 \times 1.5755\\ &= 0.145 \text{ to } 0.359\end{aligned}$$

$$\begin{aligned}95\%\ \text{CI (Efficacy)} &= 1 - \text{RR} \times \text{EF to } 1 - \text{RR/EF} = 1 - 0.359 \text{ to } 1 - 0.145\\ &= 0.641 \text{ to } 0.855\end{aligned}$$

Thus the 95% confidence interval for the efficacy of this influenza vaccine is from 64.1% to 85.5%.

Number needed to treat

An additional way of measuring the impact of treatment, which has become popular in recent years, is the **number needed to treat** (NNT). This is the number of patients who we must treat in order to prevent one adverse event. It is defined as:

$$\text{Number needed to treat (NNT)} = \frac{1}{|\text{risk difference}|}$$

The vertical bars in the formula mean the *absolute* value of the risk difference, that is the size of the risk difference ignoring its sign. NNT is best used to illustrate the

likely impact of treatment given a range of possible risks of the outcome event in the treated population.

Example 37.5

Consider the effect of a new treatment that reduces the risk of death following myocardial infarction by 25% (risk ratio = 0.75). The impact of using such a treatment will depend on the frequency of death following myocardial infarction. This is illustrated in Table 37.5, which shows that if the risk of death is 0.5 then 125 lives will be saved by treating 1000 patients with the new treatment, while if this risk of death is 0.02 then only five lives will be saved. The reduction in the number of deaths is simply the risk difference multiplied by the number of patients (risk difference = risk of event in treated patients minus risk of event in control patients). Therefore the risk difference measures the impact of treatment in *reducing* the risk of an adverse event in the same way that the attributable risk measures the impact of exposure in *increasing* the risk of an adverse event.

The values of the NNT are also shown in the table. When the risk of death in the absence of treatment is 0.5, the NNT equals $1/0.125 = 8$. Thus we will prevent one death for every eight patients treated. If, on the other hand, the risk of death in the absence of treatment is only 0.02, the NNT equals $1/0.005 = 200$, meaning that we will prevent one death for every 200 patients treated.

Table 37.5 Number of deaths in 1000 patients suffering a myocardial infarction according to whether a new treatment is used, assuming different risks of death in the absence of the new treatment and a treatment risk ratio of 0.75.

Risk of death			Number of deaths			
Current treatment (a)	New treatment (b) = 0.75 × (a)	Risk difference (c) = (b) − (a)	Current treatment (d) = 1000 × (a)	New treatment (e) = 1000 × (b)	Reduction in number of deaths (f) = (d) − (e)	NNT (g) = 1/|(c)|
0.5	0.375	−0.125	500	375	125	8
0.1	0.075	−0.025	100	75	25	40
0.02	0.015	−0.005	20	15	5	200

Number needed to harm

It is important to distinguish between beneficial effects of a treatment (risk ratio <1, risk difference <0) and harmful effects (risk ratio >1, risk difference >0). If the treatment is harmful then the NNT is referred to as the **number needed to harm** (NNH). This can be useful to assess the adverse impact of a treatment which has known side effects. For example, if our treatment for myocardial infarction was known to increase the risk of stroke, we might compare the number of patients treated to cause one stroke (NNH) with the number of patients treated to prevent one death (NNT).

Note that if the treatment has no effect (risk ratio $= 1$, risk difference $= 0$) then the NNT is $1/0 = \infty$ (infinity). This has a sensible interpretation: if the treatment is ineffective then we will not prevent any outcome events however many patients we treat. However problems can arise when deriving confidence intervals for the NNT, if one limit of the CI is close to the point of no treatment effect.

37.5 ESTIMATES OF ASSOCIATION AND IMPACT FROM MULTIVARIABLE ANALYSES

In most circumstances, multivariable analyses are based on ratio measures of the effect of exposure or treatment. This is because, both on theoretical grounds and on the basis of experience, the assumption of no interaction between the exposure and confounding variables is more likely to hold (at least approximately) for ratio measures. In the context of randomized trials, there is good empirical evidence that meta-analyses based on risk differences tend to be more heterogeneous than meta-analyses based on risk ratios or odds ratios (see Engels *et al.*, 2000; or Egger *et al.*, 2001, pages 313–335).

It is therefore usually sensible to derive a *ratio* estimate of the strength of association in a multivariable analysis of an observational study or meta-analysis of randomized trials, whatever measure of impact is required. Estimates of NNT or NNH are then derived by considering a range of levels of risk in the unexposed group, and/or prevalence of exposure.

CHAPTER 38

Strategies for analysis

38.1 Introduction
38.2 Analysis plan
38.3 Data checking
38.4 Initial analyses
Descriptive analysis
Specifying variables for analysis
Data reduction
Univariable analyses
38.5 Allowing for confounding
The need for external knowledge in assessment of confounding
Choosing confounders
38.6 Analysing for interactions
38.7 Making analyses reproducible
38.8 Common pitfalls in analysis and interpretation
38.9 Conclusions

38.1 INTRODUCTION

It is essential to plan and conduct statistical analyses in a way that maximizes the quality and interpretability of the findings. In a typical study, data are collected on a large number of variables and it can be difficult to decide which methods to use and in what order. In this final chapter we present general guidelines on strategies for data analysis.

38.2 ANALYSIS PLAN

The formulation of a written plan for analysis is recommended. The extent to which it is possible to plan analyses in detail will depend on the type of study being analysed:

- For a randomized controlled trial (RCT), which by its nature addresses a set of clearly defined questions, the analysis plan is usually specified in detail. It will include the precise definition of primary and secondary outcomes, the statistical method to be used, guidelines on whether to adjust for baseline variables and, possibly, a small number of planned subgroup analyses. See Section 34.2 for a description of the analysis of RCTs.
- For an observational study, which is exploratory in nature, it is often not possible to completely specify a plan for the analysis. However it is helpful to write down, in advance, the main hypothesis or hypotheses to be addressed. This will include the definitions of the outcome and exposure variables that will be needed to answer these question(s), the variables thought a priori to be possible confounders of the exposure–outcome association(s) and a small number of possible effect modifiers.

Well-written analysis plans both serve as a guide for the person conducting the analysis and, equally importantly, aid the interpretation and reporting of

results. For example, if we find evidence of a subgroup effect (interaction) we should report whether this was specified *a priori* or whether it is an unexpected finding.

38.3 DATA CHECKING

Careful checking and editing of the data set are essential before statistical analysis commences. The first step is to examine the distribution of each of the variables to check for possible errors. For categorical variables, this means checking that all observations relate to allowed categories, and that the frequencies in each category make sense. For numerical variables, **range checks** should be performed to search for values falling outside the expected range. Histograms can also be used to look for '**outliers**' that look extreme relative to the rest of the data.

The next step is to conduct **consistency checks**, to search for cases where two or more variables are inconsistent. For example, if sex and parity are recorded, a cross-classification of the two can be used to check that no males were recorded with a parity of one or more. Scatter plots can be useful for checking the consistency of numerical variables, for example of weight against age, or weight against height. Further outliers can be detected in this way.

Possible errors should be checked against the original records. In some cases it may be possible to correct the data. In other cases, it may be necessary to insert a missing value code if it is certain that the data were in error (for example an impossible birth weight). In borderline cases, where an observation is an outlier but not considered impossible, it is generally better to leave the data unchanged. Strictly speaking, the analysis should then be checked to ensure that the conclusions are not affected unduly by the extreme values (either using sensitivity analyses in which the extreme values are excluded, or by examining influence statistics; see Section 12.3). Note that when numerical values are grouped into categories before analysis, a small number of outliers are unlikely to have a marked influence on the results.

For studies in which individuals are classified as with and without disease, checks should generally be made separately in the two groups, as the distributions may be quite different.

38.4 INITIAL ANALYSES

Descriptive analysis

Once the data have been cleaned as thoroughly as possible, the distributions of each of the variables should be re-examined (see Chapter 3), both (i) as a final check that required corrections have been made, and (ii) to gain an understanding of the characteristics of the study population. Individuals with and without disease should again be examined separately.

Specifying variables for analysis

In addressing a particular question we will need to specify both the *outcome variable* and the *exposure variable* or variables (see Section 2.4). In observational studies, the control of confounding (see Chapter 18) is a key issue in the analysis, and so we should identify:

1 variables believed in advance to confound the exposure–outcome association (*a priori* confounders); and

2 other variables to be investigated as possible confounders, since a plausible argument can be made concerning their relationship with the exposure and outcome variables, but for which there is little or no existing evidence.

We should also specify any variables considered to be possible *effect-modifiers*: in that they modify the size or even the direction of the exposure–outcome association. As described in Sections 18.4 and 29.5, effect modification is examined using tests for interaction.

In practice, variables may play more than one role in an analysis. For example, a variable may confound the effect of one of the main exposures of interest, but its effect may also be of interest in its own right. A variable may be a confounder for one exposure variable and an effect-modifier for another. Many studies have an exploratory element, in that data are collected on some variables which may turn out to be important exposures, but if they do not they may still need to be considered as potential confounders or effect-modifiers.

Data reduction

Before commencing formal statistical analyses, it may be necessary to derive new variables by *grouping* the values of some of the original variables, as explained in Section 2.3. Note that *the original variables should always be retained in the dataset*; they should never be overwritten.

Grouping of *categorical* exposure variables is necessary when there are large numbers of categories (for example, if occupation is recorded in detail). If there is an *unexposed* category, then this should generally be treated as a separate group (e.g. non-smokers). The *exposed* categories should be divided into several groups; four or five is usually sufficient to give a reasonable picture of the risk relationship.

Grouping of *numerical* exposure variables may be necessary in order to:

1 use methods based on stratification (see Chapters 18 and 23), as recommended for the initial examination of confounding (see below);

2 use graphical methods to examine how the level of a non-numerical outcome changes with exposure level (see Section 29.6); and

3 to examine whether there is a linear association between a numerical *exposure* variable and a non-numerical outcome (see Section 29.6).

Note that grouping entails *loss of information*: after checking linearity assumptions or performing initial analyses using the grouped variable it may be appropriate

to use the original variable, or a transformation of the original variable (see Chapter 13), in the final analysis.

One strategy for numerical exposures is to divide the range of the variable using, say, quintiles, to give five groups with equal numbers of subjects in each group. This helps to ensure that estimates of effect for each category are reasonably precise, but can sometimes obscure an important effect if a few subjects with very high levels are grouped with others with more moderate levels. Alternatively, cut-off points may be chosen on the basis of data from previous studies, the aim being to define categories within which there is thought to be relatively little variation in risk. Using standard cut-off points has the advantage of making comparisons between studies easier. For example, Table 38.1 shows the different possibilities for including body mass index (BMI), defined as weight/(height^2), in an analysis to examine its association with a disease outcome.

For variables included in the analysis as *confounders*, three or four categories may be sufficient to remove most of the confounding. However, more categories will be needed if the confounding is strong, as would often be the case with age, for example. It is often necessary to examine the strength of the association between the potential confounder and the outcome variable before deciding on the number of categories to be used in analysis. The weaker the association, the more one may combine groups. However it would be unwise to combine groups with very different risks or rates of disease.

A further consideration is that for analyses of binary or time-to-event outcomes, groups in which there are no, or very few, outcome events must be combined with others before inclusion in analysis.

Table 38.1 Possible ways of deriving variables based on measured body mass index (BMI).

Choice
(i) Original variable
(ii) A transformation of the original variable (for example log BMI)
(iii) Quintiles of BMI, coded 1–5
(iv) Quintiles of BMI, coded as the median BMI in each quintile
(v) Standard cut-offs for BMI focusing on high levels of BMI as risky (<25 = normal; 25–30 = overweight; ≥ 30 = obese)
(vi) Standard cut-offs including an underweight group (<20 = underweight; 20–25 = normal; 25–30 = overweight; ≥ 30 = obese)

Univariable analyses

It is usually helpful to begin with a univariable analysis, in which we examine the association of the outcome with each exposure of interest, ignoring all other variables. This is often called the **crude association** between the exposure and the outcome. Although later analyses, controlling for the effects of other variables, will supersede this one, it is still a useful stage of the analysis because:

1 Examination of simple tables or graphs, as well as the estimated association, can give useful information about the data set. For example, it can show that there were very few observations, or very few outcome events, in particular exposure categories.
2 These analyses will give an initial idea of those variables that are strongly related to the disease outcome.
3 The degree to which the crude estimate of effect is altered when we control for the confounding effects of other variables is a useful indication of the amount of confounding present (or at least, the amount that has been measured and successfully removed).

For exposures with more than two levels, one of the levels has to be chosen as the baseline (see Section 19.2). Often this will be the *unexposed group* or, if everyone is exposed to some extent, the group with the lowest level of exposure. If there are very few persons in this group, however, this will produce exposure effect estimates with large standard errors. It is then preferable to choose a larger group to be the baseline group.

38.5 ALLOWING FOR CONFOUNDING

This section should be read in conjunction with Section 29.8, which describes general issues in the choice of exposure variables for inclusion in a regression model.

In any observational study, the control of confounding effects will be a major focus of the analysis. We have two tools available for this task: classical (Mantel–Haenszel) methods based on stratification, and regression modelling. We have emphasized the similarities between the two approaches (see Chapters 20 and 24), so they should not be seen as in conflict. Regression methods *controlling* for the effect of a categorical variable involve exactly the same assumptions, and hence give essentially the same results, as Mantel–Haenszel methods *stratifying* on the categorical variable.

A major reason for using classical methods in the initial phase of the analysis is that the output encourages us to examine the exposure–outcome association in each stratum, together with the evidence for interaction (effect modification). In contrast, it is easy to use regression models without checking the assumption that there is no interaction between the effects of the different variables in the model.

However, regression models are generally the best approach when we wish to control for the effects of a number of confounding variables, because stratifying on the cross-classification of all the confounders is likely to produce a large number of strata. As explained in Section 29.5, by assuming in regression models that there is no interaction between the effects of confounding variables, we can greatly reduce the number of strata (the number of parameters used to model the effect of the confounders). In addition, dose–response effects can be examined more flexibly in regression models (see Section 29.6).

The need for external knowledge in assessment of confounding

As explained in Chapter 18, a confounding variable, C, is one that is associated with both the exposure variable (E) and the outcome variable (D), and is not on the part of the causal chain leading from E to D. It is important to realize that *external knowledge is more important than statistical strategies* in choosing appropriate confounders to be controlled for in examining a particular exposure–outcome association. *This is because statistical associations in the data cannot, on their own, determine whether it is appropriate to control for the effects of a particular variable*.

Example 38.1

In their article on the appropriate control of confounding in studies of the epidemiology of birth defects, Hernán *et al*. (2002) considered the following example. Should we control for C, a binary variable which records the event that the pregnancy ended in stillbirth or therapeutic abortion, when examining the association between folic acid supplementation in early pregnancy (the exposure variable, E) and the risk of neural tube defects (the outcome, D) using data from a case–control study? They pointed out that controlling for C would *not* be the correct analysis, although:

1 controlling for the effect of C leads to a substantial change in the estimated association between E and D; and

2 C is strongly associated with both E and D, and is not on the causal pathway between them.

The reason is that C is affected by both E and D, rather than having any influence on either of them. Therefore C, in this instance, cannot confound the E–D association. Yet it is not uncommon to find epidemiological analyses controlling for C in situations such as this. Note that restricting the analysis to live births (i.e. considering only one of the strata defined by C) will also produce a biased estimate of the E–D association in this situation.

This example shows that careful consideration of the likely *direction of associations* between E, D and C is required in order to decide whether it is appropriate to control for C in estimating the E–D association. Figure 38.1 gives examples of circumstances in which C will and will not confound the E–D association.

Example 38.2

Because of the frequent introduction of new antiretroviral drugs for treatment of HIV-infected persons, and the large number of different possible combinations of these, many relevant questions about the effect of different drugs or drug combinations have not been addressed in randomized trials with ‘hard’ outcomes such as development of AIDS or death. There is therefore great interest in using longitudinal studies of HIV-infected individuals to address these questions.

Consider a comparison of drug regimens A and B. Because antiretroviral therapy may involve taking a large number of pills per day, and may have serious

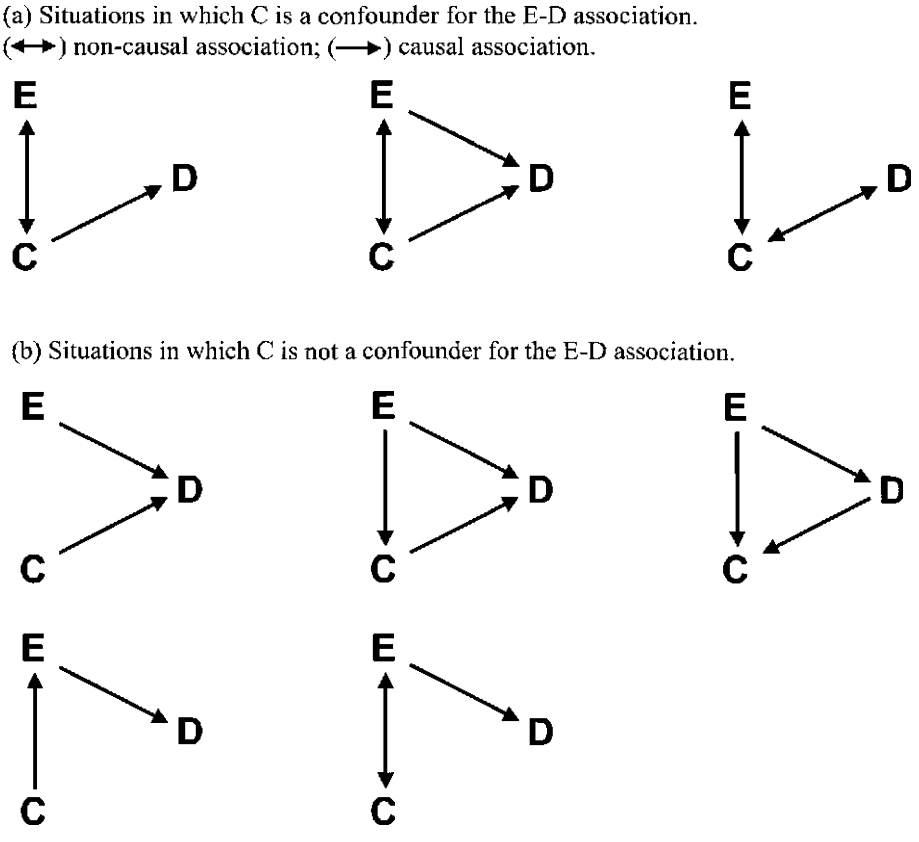


Fig. 38.1 Circumstances in which C will and will not confound an exposure–disease (E–D) association. (Adapted from *Case Control Studies MEB2* by James J. Schlesselman, copyright 1982 by Oxford University Press, Inc., with permission.)

side-effects, adherence to the prescribed regime is likely to be associated both with the probability of progressing to AIDS (D) and with the drug regimen (E). However, in this example *the drug regimen used is likely to influence adherence to therapy*. It would not, therefore, be appropriate to control for adherence in estimating the E–D association, as it will be on the pathway between them.

Example 38.3

The 'fetal origins' hypothesis suggests that there are associations between prenatal growth, reflected in measures such as birthweight, and adult heart disease. Huxley *et al*. (2002) reviewed 55 studies that had reported associations between birthweight (exposure) and later systolic blood pressure (outcome). Almost all of the reported regression coefficients were adjusted for adult weight. However, these need to be interpreted with caution since adult weight is on the causal pathway between birthweight and blood pressure. Removing the adjustment for adult weight, in 12 studies, halved the size of the estimated association.

Choosing confounders

Taking into account the need to combine external knowledge with statistical associations, we recommend the following strategy for choosing confounders:

1 Formulating a conceptual, hierarchical framework for the relationships between the different variables and the disease outcome is strongly recommended, as described by Victora *et al*. (1997) in the context of determinants of childhood diarrhoea mortality. This is particularly useful both as a way of summarizing existing knowledge and for clarifying the direction of any associations.
2 As a general rule, variables that are known *a priori* to be important confounders, based on previous work should be controlled for in the analysis.
3 In addition, other possible confounders may be selected as a result of exploratory analysis. This should be:
   - restricted to variables that are associated with *both* the outcome and exposure, and are not on the causal pathway between them;
   - based on both the data being analysed and external knowledge, and after careful consideration of the direction of associations.
4 Note, however, that for **multiple linear regression**, *all* exposure variables that are clearly associated with the outcome should be included when estimating the effect of a particular exposure, *whether or not they are confounders* (with the exception that variables on the causal pathway between the exposure of interest and the outcome should *not* be included; see Section 29.8).
5 Note also that automated 'stepwise' regression procedures are unlikely to be appropriate in analyses whose aim is to estimate the effect of particular exposures (see Section 29.8).

38.6 ANALYSING FOR INTERACTIONS

Three sorts of interaction may be distinguished:

1 *Interaction between confounders*. The main difference between regression models and classical methods is that classical methods always allow for all interactions between confounders. This is in fact usually unnecessary.
2 *Interaction between a confounder and an exposure of interest*. Strictly speaking, the calculation of exposure effect estimates controlled for confounding variables is appropriate only if the exposure effect is the same for all levels of the confounder. In practice, of course, the effect will vary to at least some extent between strata; in other words there is likely to be some interaction between the exposure and the confounders controlled for in the analysis. In the presence of substantial interaction, the stratum-specific effects of the exposure should be reported.
3 *Interaction between exposures of interest*. If present, this may be of importance both for the scientific interpretation of an analysis and for its implications for preventive intervention.

An exhaustive search for interactions with all possible variables, however, is unlikely to be useful. Formal tests for interaction lack power, and statistically significant interactions identified by a systematic sweep of all variables may well be chance effects, while real interactions may go undetected. Sample sizes are typically inadequate to have high power of detecting any but the strongest interactions (see Section 35.4). Combining groups in the interaction parameter may increase the power of tests for interaction (see Section 29.5).

The purpose of a statistical analysis is to provide a simplified but useful picture of reality. If weak interactions are present, this is probably of little intrinsic interest, and the calculation of an overall pooled estimate of effect for an individual exposure is a reasonable approximation to the truth.

For these reasons, we suggest delaying analysis for interactions to the final analysis. Exposure–exposure and exposure–confounder interactions should then be examined, paying particular attention to those thought *a priori* to be worth investigation. These should be examined one at a time, to avoid a model with too many additional parameters. In assessing the evidence for interactions, as much attention should be paid to the presence of meaningful trends in effect estimates over the strata, as to the results of formal tests for interaction.

38.7 MAKING ANALYSES REPRODUCIBLE

In the early stages of a statistical analysis it is useful to work *interactively* with the computer, by trying a command, looking at the output, then correcting or refining the command before proceeding to the next command. However, we recommend that all analyses should eventually be done using files (programs) containing lists of commands.

It is usually the case that, after analyses are first thought to be complete, changes are found to be necessary. For example, more data may arrive, or corrections may be made, or it may be discovered that an important confounder has been omitted. This often means that the whole analysis must be performed again. If analyses were performed interactively, this can be a daunting task. The solution is to ensure that the whole analysis can be performed by running a series of programs.

A typical series of programs is illustrated in Table 38.2. We strongly recommend that you add frequent comment statements to your programs, which explain what is being done in each section; especially in complicated or long programs. This is useful for other members of the project team, and also invaluable when returning to your own program some time later to rerun it or to modify it for a new analysis. It is also important to *document* the analysis by recording the function of each program file, and the order in which they should be run.

Following this strategy has two important consequences. Firstly, it will now be straightforward to reproduce the entire analysis after corrections are made to the raw data. Secondly, you will always be able to check exactly how a derived variable was coded, which confounders were included in a particular analysis,

Table 38.2 Typical sequence of programs to perform the analyses needed to analyse a particular exposure–outcome association.

Program 1:	Read the raw data file into the statistical package, label variables so that it is easy to identify them, check that they have the correct value ranges, check consistency between variables, create derived variables by recoding and combining variables, save the resulting dataset
Program 2:	Use the new dataset to examine associations between the outcome variable and the exposures and confounders of interest, by producing appropriate graphs and tables and performing univariable analyses
Program 3:	Use Mantel–Haenszel and regression analyses to estimate exposure effects controlled for potential confounders
Program 4:	Examine interactions between exposures and between exposures and confounders
Program 5:	Produce final tables for the research report

and so on. Remember that reviewers' comments on a draft manuscript that was submitted for publication tend to be received many months after the paper was submitted (and even longer after the analysis was done). Minor modifications to the analysis will be straightforward if the analysis is reproducible, but can waste huge amounts of time if it is not.

38.8 COMMON PITFALLS IN ANALYSIS AND INTERPRETATION

Even when the analyses of primary interest are specified at the start of the study, a typical analysis will involve choices of variable groupings and modelling strategies that can make important differences to the conclusions. Further, it is common to investigate possible associations that were not specified in advance, for example if they were only recently reported. Three important reasons for caution in interpreting the results of analyses are:

1 **Multiple comparisons**. Even if there is no association between the exposure and outcome variables, we would expect one in twenty comparisons to be statistically significant at the 5% level. Thus the interpretation of associations in a study in which the effect of many exposures was measured should be much more cautious than that for a study in which a specific a priori hypothesis was specified. Searching for all possible associations with an outcome variable is known as 'data-dredging' and may lead to dramatic but spurious findings.
2 **Subgroup analyses**. We should be particularly cautious about the interpretation of apparent associations in subgroups of the data, particularly where there is no convincing evidence of an overall association (see Section 34.2). It is extremely tempting to emphasize an 'interesting' finding in an otherwise negative study.
3 **Data-driven comparisons**. A related problem is that we should not group an exposure variable in order to produce the biggest possible association with the outcome, and then interpret the *P*-value as if this had always been the intended comparison. For example, when rearranging ten age groups into two larger

groups, we could compare 1 with 2–10 or 1 and 2 with 3–10 and so on. If we choose a particular grouping out of these nine possible ones because it shows the largest difference between 'younger' and 'older' individuals, then we have chosen the smallest *P*-value from nine possible ones. It is sensible to decide how variables will be grouped as far as possible before seeing how different groupings affect the conclusions of your study.

These problems *do not* mean that all studies must have hypotheses and methods of analysis that are specified at the outset. However, the interpretation of a finding will be affected by its context. If a reported association is one of fifty which were examined, this should be clearly stated when the research is reported. We would probably view such an association (even with a small *P*-value) as generating a hypothesis that might be tested in future studies, rather than as a definitive result.

38.9 CONCLUSIONS

In all but the simplest studies, there is no single 'correct' analysis or answer. Fast computers and excellent statistical software mean that it is easy to produce statistical analyses. The challenge to medical statisticians is to produce analyses that answer the research question as clearly and honestly as possible.

APPENDIX

STATISTICAL TABLES

A1 Areas in tail of the standard normal distribution 470

A2 Percentage points of the standard normal distribution 472

A3 Percentage points of the *t* distribution 473

A4 Two-sided *P*-values for the *t* distribution, according to the value of the test statistic 474

A5 Percentage points of the χ^2 distribution 476

A6 Probits 477

A7 Critical values for the Wilcoxon matched pairs signed rank test 479

A8 Critical ranges for the Wilcoxon rank sum test 480

A9 Random numbers 482