1

Some general concepts

An ideal sequence is defined specifying the progression of an investigation from the conception of one or more research questions to the drawing of conclusions. The role of statistical analysis is outlined for design, measurement, analysis and interpretation.

1.1 Preliminaries

This short chapter gives a general account of the issues to be discussed in the book, namely those connected with situations in which appreciable unexplained and haphazard variation is present. We outline in idealized form the main phases of this kind of scientific investigation and the stages of statistical analysis likely to be needed.

It would be arid to attempt a precise definition of statistical analysis as contrasted with other forms of analysis. The need for statistical analysis typically arises from the presence of unexplained and haphazard variation. Such variability may be some combination of natural variability and measurement or other error. The former is potentially of intrinsic interest whereas the latter is in principle just a nuisance, although it may need careful consideration owing to its potential effect on the interpretation of results.

Illustration: Variability and error The fact that features of biological organisms vary between nominally similar individuals may, as in studies of inheritance, be a crucial part of the phenomenon being studied. That, say, repeated measurements of the height of the same individual vary erratically is not of intrinsic interest although it may under some circumstances need consideration. That measurements of the blood pressure of a subject, apparently with stable health, vary over a few minutes, hours or days typically arises from a combination of measurement error and natural variation; the latter part of the variation, but not the former, may be of direct interest for interpretation.

1.2 Components of investigation

It is often helpful to think of investigations as occurring in the following steps:

- formulation of research questions, or sometimes hypotheses;
- search for relevant data, often leading to
- design and implementation of investigations to obtain appropriate data;
- analysis of data; and
- interpretation of the results, that is, the translation of findings into a subject-matter context or into some appropriate decision.

This sequence is the basis on which the present book is organized.

It is, however, important to realize that the sequence is only a model of how investigations proceed, a model sometimes quite close to reality and sometimes highly idealized. For brevity we call it *the ideal sequence*.

The essence of our discussion will be on the achievement of individually secure investigations. These are studies which lead to unambiguous conclusions without, so far as is feasible, worrying caveats and external assumptions. Yet virtually all subject-matter issues are tackled sequentially, understanding often emerging from the synthesis of information of different types, and it is important in interpreting data to take account of all available information. While information from various studies that are all of a broadly similar form may be combined by careful statistical analysis, typically the important and challenging issue of synthesizing information of very different kinds, so crucial for understanding, has to be carried out informally.

Illustration: Synthesizing information (I) One situation in which such a synthesis can be studied quantitatively occurs when several surrogate variables are available, all providing information about an unobserved variable of interest. An example is the use of tree ring measurements, pollen counts and borehole temperatures to reconstruct Northern Hemisphere temperature time series (Li *et al.*, 2010; McShane and Wyner, 2011).

Illustration: Synthesizing information (II) The interpretation of a series of investigations of bovine tuberculosis hinged on a consistent synthesis of information from a randomized field trial, from ecological studies of wildlife behaviour and from genetic analysis of the pathogens isolated from cattle and wildlife sampled in the same areas.

Illustration: Synthesizing information (III) The evidence that the Human Immunodeficiency Virus (HIV) is the causal agent for Acquired Immunodeficiency Syndrome (AIDS) comes from epidemiological studies of various high-risk groups, in particular haemophiliacs who received blood transfusions with contaminated products, as well as from laboratory work.

It is a truism that asking the 'right' question or questions is a crucial step to success in virtually all fields of research work. Both in investigations with a specific target and also in more fundamental investigations, especially those with a solid knowledge base, the formulation of very focused research questions at the start of a study may indeed be possible and desirable. Then the ideal sequence becomes close to reality, especially if the specific investigation can be completed reasonably quickly.

In other cases, however, the research questions of primary concern may emerge only as the study develops. Consequent reformulation of the detailed statistical model used for analysis, and of the translation of the research question into statistical form, usually causes no conceptual problem. Indeed, in some fields the refinement and modification of the research questions as the analysis proceeds is an essential part of the whole investigation. Major changes of focus, for example to the study of effects totally unanticipated at the start of the work, ideally need confirmation in supplementary investigations, however.

An extreme case of departure from the ideal sequence arises if new data, for example a large body of administrative data, become available and there is a perception that it must contain interesting information about something, but about exactly what is not entirely clear. The term 'data mining' is often used in such contexts. How much effort should be spent on such issues beyond the simple tabulation of frequencies and pair-wise dependencies must depend in part on the quality of the data. Simple descriptions of dependencies may, however, be very valuable in suggesting issues for detailed study.

While various standard and not-so-standard methods may be deployed to uncover possible interrelationships of interest, any conclusions are in most cases likely to be tentative and in need of independent confirmation. When the data are very extensive, precision estimates calculated from simple standard statistical methods are likely to underestimate error substantially owing to the neglect of hidden correlations. A large amount of data is in no way synonymous with a large amount of information. In some settings at least, if a modest amount of poor quality data is likely to be modestly

misleading, an extremely large amount of poor quality data may be extremely misleading.

Illustration: Data mining as suggestive Carpenter *et al.* (1997) analysed approximately 10^6 observations from UK cancer registrations. The data formed a 39×212 contingency table corresponding to 39 body sites and occupational coding into 212 categories. Although there is a clear research objective of detecting occupations leading to high cancer rates at specific body sites, the nature of the data precluded firm conclusions being drawn, making the investigation more in the nature of data mining. The limitations were that many occupations were missing, by no means necessarily at random, and multiple occupations and cancers at multiple sites were excluded. Crucial information on the numbers at risk in the various occupational categories was not available. A largely informal graphical approach to the data was used. Some previously well-established relationships were recovered; otherwise the conclusions were very tentative.

Illustration: Data mining or a fresh start An industrial company that has specialized in making virtually the same product for many years may have very extensive records of routine tests made on the product as part of their quality control procedures. It may be reasonable to feel that important lessons could be learned from careful analysis of the data. But it is likely that the data have been recorded and the tests performed by different individuals using testing procedures that may have changed in important and possibly unrecorded ways and that many important changes affecting product quality are unrecorded. Extensive effort may be better directed at experimentation on the current system than at analyses of historical data.

Furthermore, while the sequence from question to answer set out above is in principle desirable, from the perspective of the individual investigator, and in particular the individual statistician, the actual sequence may be very different. For example, an individual research worker destined to analyse the data may enter an investigation only at the analysis phase. It will then be important to identify the key features of design and data collection actually employed, since these may have an important impact on the methods of analysis needed. It may be difficult to ascertain retrospectively aspects that were in fact critical but were not initially recognized

as such. For example, departures from the design protocol of the study may have occurred and it very desirable to detect these in order to avoid misinterpretation.

1.3 Aspects of study design

Given a research question, a first step will be to consider whether data that may give at least a partial answer are already available. If not, one or more studies may need to be set up that are specifically aimed to answer the question.

The first considerations will then be the choice of material, that is, the individuals or contexts to study, and what measurements are appropriate. For a generic terminology we will refer to the individuals or contexts as *units of analysis*. We will discuss the criteria for measurements in more detail in Chapter 4. For studies of a new phenomenon it will usually be best to examine situations in which the phenomenon is likely to appear in the most striking form, even if this is in some sense artificial or not representative. This is in line with the well-known precept in mathematical research: study the issue in the simplest possible context that is not entirely trivial, and later generalize.

More detailed statistical considerations of design tend to focus on the precise configuration of data to be collected and on the scale of effort appropriate. We discuss these aspects in Chapters 2 and 3.

1.4 Relationship between design and analysis

The design and data collection phases of a study are intimately linked with the analysis. Statistical analysis should, and in some cases must, take account of unusual features of the earlier phases. Interpretation is the ultimate objective and so one objective of analysis is to get as close to incisive and justified subject-matter interpretation as is feasible.

Moreover, it is essential to be clear at the design stage broadly how the data are to be analysed. The amount of detail to be specified depends on the context. There are two reasons for requiring the prior consideration of analysis. One is that conclusions are publicly more convincing if established by methods set out in advance. An aspect that is in some ways more important, especially in studies with a long time frame, is that prior specification reduces the possibility that the data when obtained cannot be satisfactorily analysed. If, for quasi-legal reasons, for example to satisfy a regulatory agency, it is necessary to pre-specify the analysis in detail, it will

be obligatory to follow and report that analysis but this should not preclude alternative analyses if these are clearly more appropriate in the light of the data actually obtained.

In some contexts it is reasonable not only to specify in advance the method of analysis in some detail but also to be hopeful that the proposed method will be satisfactory with at most minor modification. Past experience of similar investigations may well justify this.

In other situations, however, while the broad approach to analysis should be set out in advance, if only as an assurance that analysis and interpretation will be possible, it is unrealistic and indeed potentially dangerous to follow an initial plan unswervingly. Experience in collecting the data, and the data themselves, may suggest changes in specification such as a transformation of variables before detailed analysis. More importantly, it may be a crucial part of the analysis to clarify the research objectives; these should be guided, in part at least, by the initial phases of analysis. A distinction is sometimes drawn between pre-set analyses, called confirmatory, and exploratory analyses, but in many fields virtually all analyses have elements of both aspects.

Especially in major studies in which follow-up investigations are not feasible or will take a long time to complete, it will be wise to list the possible data configurations, likely or unlikely, that might arise and to check that data will be available for the interpretation of unanticipated effects.

Illustration: Explaining the unexpected In preliminary discussion of a study of hypertension, a group of cardiologists were unanimous that a certain intervention would lower blood pressure. When challenged as to their possible reaction if the data showed the opposite effect, their answer was that in five minutes a plausible explanation would be suggested and in ten minutes three different explanations. That is, even though there was an initial quite strong belief that a particular process would be operating, the possibility of alternative processes could and should not be totally excluded. This made it desirable, so far as was feasible, to collect data that might help clarify the situation should indeed blood pressure surprisingly increase following the intervention.

1.5 Experimental and observational studies

A crucial distinction is that we use the term *experiment* to mean a study in which all key elements are under the control of the investigator whereas a

study is *observational* if, although the choice of individuals for study and of measurements to be obtained may be made by the investigator, key elements have to be accepted as they already exist and cannot be manipulated by the investigator. It often, however, aids the interpretation of an observational study to consider the question: what would have been done in a comparable experiment?

Illustration: Conflicting observational and experimental evidence
A number of observational studies, reviewed by Grady *et al.* (1992), suggested that women using hormone replacement therapy (HRT) for long periods of time had a lower coronary heart disease rate than apparently comparable control groups. In these studies the investigators chose the women to enter the investigation and the variables to be recorded but had no influence over whether any specific woman did or did not use HRT. In a randomized experiment (Writing group for Women's Health Initiative Investigators, 2002), women giving informed consent were assigned at random either to HRT or to an inactive control, the decision and its implementation being in principle concealed from the women and their treating doctor. After a period, this trial was stopped because of evidence of a possible *adverse* effect of HRT on total cardiovascular events and because of strong evidence of the absence of any useful overall beneficial effect. That is, the observational and experimental evidence were inconsistent with one another.

In its simplest terms the interpretation is as follows. The two groups of women compared in the observational studies may have been systematically different not only with respect to HRT use but also on a range of health-related features such as socio-economic status, education and general lifestyle, including eating habits. While some checks of comparability are possible, it remains the case that the clearly statistically significant difference in outcome between the two groups may be quite unconnected with HRT use.

By contrast, in the randomized trial the two groups differed only by the play of chance and by the fact that one group was allocated to HRT and the other to control. A clearly significant difference could confidently be taken as a consequence of the treatment allocation.

In an experiment conducted in a research laboratory the investigator could ensure that in all important respects the *only* difference between the groups being compared lay in HRT use versus no HRT

use. The conclusion would thus be virtually unambiguous. In the more complicated environment of a clinical trial, however, especially one lasting an appreciable time, departures from the trial protocol might occur, such as a failure to take the allocated medication or the taking of supplementary medication, such departures being indirect consequences of the allocated treatment. Thus the primary comparison of outcomes in the clinical trial includes not only the direct effect of the medication but also its indirect effects. In this particular study, a modest but nonnegligible failure to comply with the study medication was reported but was judged not to modify the findings.

1.6 Principles of measurement

The primary requirements for measurements are:

- what is sometimes called *construct validity*, namely that the measurements do actually record the features of subject-matter concern;
- in particular that they record a number of different features sufficient to capture concisely the important aspects;
- that they are reliable, in the sense that a real or notional repetition of the measurement process gives reasonably reproducible results;
- that the cost of the measurements is commensurate with their importance; and
- that the measurement process does not appreciably distort the system under study.

We discuss measurements and most of the above points in more detail in Chapter 4. In particular, we note now that measurements can be classified by the structure of possible values (for example, binary or continuous) and, even more importantly, by their potential role in interpretation, for example as outcomes or as explanatory features.

The issue of dimensionality, especially that of the so-called outcome and response variables, depends strongly on the context.

Illustration: Summarizing multidimensional data That the economic activity of a nation, the quality of life of an individual or the status of a complex organization such as a university can be wholly encapsulated in a single number such as a gross domestic product (GDP), a quality-adjusted life year (QUALY) or a league table ranking of the world's universities is patently absurd. In general, the description of complex

multidimensional phenomena by a limited number of summary measures requires the careful specification of objectives. Pressure to produce one-dimensional summaries, to be resisted except for highly specific purposes, comes from the view that many situations can be explained in terms of the optimization of an appropriate one-dimensional criterion. This may be combined with the explicit or implicit assumption that utility can be measured in money terms.

1.7 Types and phases of analysis

A general principle, sounding superficial but difficult to implement, is that analyses should be as simple as possible, but no simpler. Some complication may be necessary to achieve precise formulation or to uncover the issue of interest from a confusing background or, somewhat less importantly, to obtain meaningful assessments of uncertainty.

Moreover, the method of analysis should so far as feasible be transparent. That is, it should be possible to follow the pathways from the data to the conclusions and in particular to see which aspects of the data have influenced key conclusions, especially any that are surprising. Black-box methods of analysis may to some extent be unavoidable in complex problems, but conclusions from them demand particularly careful checking and verification.

Four main phases of analysis are usually needed:

- data auditing and screening;
- preliminary analysis;
- formal analysis; and
- presentation of conclusions.

Data auditing and screening, which should take place as soon as possible after data collection, include inspection for anomalous values as well as for internal inconsistencies. Other relatively common sources of concern are sticking instruments, repeatedly returning the same value, for example zero rainfall, as well as the confusion of zero values and missing or irrelevant values. Sometimes, especially when extensive data are being collected in a novel context, formal auditing of the whole process of data collection and entry may be appropriate. Typically this will involve detailed study of all aspects of a sample of study individuals, and the application of ideas from sampling theory and industrial inspection may be valuable.

Methods of analysis are broadly either graphical or quantitative, the former being particularly helpful at the preliminary stages of analysis and in the presentation of conclusions. Typically, however, it will be desirable that in a final publication the key information is available also in numerical form, possibly as supplementary material. The reason is that reading data from graphs makes further analysis subject to (possibly appreciable) avoidable rounding error.

Graphical methods for handling large amounts of complex data, often studied under the name *visualization*, may require specialized software and will not be considered in detail here. For handling less complicated situations Section 5.4 suggests some simple rules, obvious but quite often ignored.

1.8 Formal analysis

Some methods of analysis may be described as *algorithmic*. That is to say, relationships within the data are recovered by a computer algorithm typically minimizing a plausible criterion. The choice of this criterion may not be based on any formal grounds and does not necessarily have any specific probabilistic properties or interpretation. Thus the method of least squares, probably the most widely used technique for fitting a parametric formula to empirical data, can be regarded purely algorithmically, as in effect a smoothing device; it gains some strength in that way. In most statistical settings, however, the method of least squares is formally justified by a probability model for the data.

We concentrate our discussions on analyses based on a formal probability model for the data, although certainly we do not exclude purely algorithmic methods, especially in the initial stages of the reduction of complex data.

1.9 Probability models

Most of our later discussion centres on analyses based on probability models for the data, leading, it is hoped, to greater subject-matter understanding. Some probability models are essentially descriptions of commonly occurring patterns of variability and lead to methods of analysis that are widely used across many fields of study. Their very generality suggests that in most cases they have no very specific subject-matter interpretation as a description of a detailed data-generating process. Other probability models are much more specific and are essentially probabilistic theories of