CHAPTER 1

Some general concepts

1.1 Types of investigation

This book is about the design of experiments. The word *experiment* is used in a quite precise sense to mean an investigation where the system under study is under the control of the investigator. This means that the individuals or material investigated, the nature of the treatments or manipulations under study and the measurement procedures used are all settled, in their important features at least, by the investigator.

By contrast in an observational study some of these features, and in particular the allocation of individuals to treatment groups, are outside the investigator's control.

Illustration. In a randomized clinical trial patients meeting clearly defined eligibility criteria and giving informed consent are assigned by the investigator by an impersonal method to one of two or more treatment arms, in the simplest case either to a new treatment, T, or to a standard treatment or control, C, which might be the best current therapy or possibly a placebo treatment. The patients are followed for a specified period and one or more measures of response recorded.

In a comparable observational study, data on the response variables might be recorded on two groups of patients, some of whom had received T and some C; the data might, for example, be extracted from a database set up during the normal running of a hospital clinic. In such a study, however, it would be unknown why each particular patient had received the treatment he or she had.

The form of data might be similar or even almost identical in the two contexts; the distinction lies in the firmness of the interpretation that can be given to the apparent differences in response between the two groups of patients.

Illustration. In an agricultural field trial, an experimental field is divided into plots of a size and shape determined by the investiga-

tor, subject to technological constraints. To each plot is assigned one of a number of fertiliser treatments, often combinations of various amounts of the basic constituents, and yield of product is measured.

In a comparable observational study of fertiliser practice a survey of farms or fields would give data on quantities of fertiliser used and yield, but the issue of why each particular fertiliser combination had been used in each case would be unclear and certainly not under the investigator's control.

A common feature of these and other similar studies is that the objective is a comparison, of two medical treatments in the first example, and of various fertiliser combinations in the second. Many investigations in science and technology have this form. In very broad terms, in technological experiments the treatments under comparison have a very direct interest, whereas in scientific experiments the treatments serve to elucidate the nature of some phenomenon of interest or to test some research hypothesis. We do not, however, wish to emphasize distinctions between science and technology.

We translate the objective into that of comparing the responses among the different treatments. An experiment and an observational study may have identical objectives; the distinction between them lies in the confidence to be put in the interpretation.

Investigations done wholly or primarily in the laboratory are usually experimental. Studies of social science issues in the context in which they occur in the real world are usually inevitably observational, although sometimes elements of an experimental approach may be possible. Industrial studies at pilot plant level will typically be experimental whereas at a production level, while experimental approaches are of proved fruitfulness especially in the process industries, practical constraints may force some deviation from what is ideal for clarity of interpretation.

Illustration. In a survey of social attitudes a panel of individuals might be interviewed say every year. This would be an observational study designed to study and if possible explain changes of attitude over time. In such studies panel attrition, i.e. loss of respondents for one reason or another, is a major concern. One way of reducing attrition may be to offer small monetary payments a few days before the interview is due. An experiment on the effectiveness of this could take the form of randomizing individuals

between one of two treatments, a monetary payment or no monetary payment. The response would be the successful completion or not of an interview.

1.2 Observational studies

While in principle the distinction between experiments and observational studies is clear cut and we wish strongly to emphasize its importance, nevertheless in practice the distinction can sometimes become blurred. Therefore we comment briefly on various forms of observational study and on their closeness to experiments.

It is helpful to distinguish between a prospective longitudinal study (cohort study), a retrospective longitudinal study, a cross-sectional study, and the secondary analysis of data collected for some other, for example, administrative purpose.

In a *prospective study* observations are made on individuals at entry into the study, the individuals are followed forward in time, and possible response variables recorded for each individual. In a *retrospective study* the response is recorded at entry and an attempt is made to look backwards in time for possible explanatory features. In a *cross-sectional study* each individual is observed at just one time point. In all these studies the investigator may have substantial control not only over which individuals are included but also over the measuring processes used. In a *secondary analysis* the investigator has control only over the inclusion or exclusion of the individuals for analysis.

In a general way the four possibilities are in decreasing order of effectiveness, the prospective study being closest to an experiment; they are also in decreasing order of cost.

Thus retrospective studies are subject to biases of recall but may often yield results much more quickly than corresponding prospective studies. In principle at least, observations taken at just one time point are likely to be less enlightening than those taken over time. Finally secondary analysis, especially of some of the large databases now becoming so common, may appear attractive. The quality of such data may, however, be low and there may be major difficulties in disentangling effects of different explanatory features, so that often such analyses are best regarded primarily as ways of generating ideas for more detailed study later.

In epidemiological applications, a retrospective study is often designed as a *case-control* study, whereby groups of patients with

a disease or condition (cases), are compared to a hopefully similar group of disease-free patients on their exposure to one or more risk factors.

1.3 Some key terms

We shall return later to a more detailed description of the types of experiment to be considered but for the moment it is enough to consider three important elements to an experiment, namely the experimental units, the treatments and the response. A schematic version of an experiment is that there are a number of different treatments under study, the investigator assigns one treatment to each experimental unit and observes the resulting response.

Experimental units are essentially the patients, plots, animals, raw material, etc. of the investigation. More formally they correspond to the smallest subdivision of the experimental material such that any two different experimental units might receive different treatments.

Illustration. In some experiments in opthalmology it might be sensible to apply different treatments to the left and to the right eye of each patient. Then an experimental unit would be an eye, that is each patient would contribute two experimental units.

The treatments are clearly defined procedures one of which is to be applied to each experimental unit. In some cases the treatments are an unstructured set of two or more qualitatively different procedures. In others, including many investigations in the physical sciences, the treatments are defined by the levels of one or more quantitative variables, such as the amounts per square metre of the constituents nitrogen, potash and potassium, in the illustration in Section 1.1.

The response measurement specifies the criterion in terms of which the comparison of treatments is to be effected. In many applications there will be several such measures.

This simple formulation can be amplified in various ways. The same physical material can be used as an experimental unit more than once. If the treatment structure is complicated the experimental unit may be different for different components of treatment. The response measured may be supplemented by measurements on other properties, called *baseline variables*, made before allocation

to treatment, and on *intermediate variables* between the baseline variables and the ultimate response.

Illustrations. In clinical trials there will typically be available numerous baseline variables such as age at entry, gender, and specific properties relevant to the disease, such as blood pressure, etc., all to be measured before assignment to treatment. If the key response is time to death, or more generally time to some critical event in the progression of the disease, intermediate variables might be properties measured during the study which monitor or explain the progression to the final response.

In an agricultural field trial possible baseline variables are chemical analyses of the soil in each plot and the yield on the plot in the previous growing season, although, so far as we are aware, the effectiveness of such variables as an aid to experimentation is limited. Possible intermediate variables are plant density, the number of plants per square metre, and assessments of growth at various intermediate points in the growing season. These would be included to attempt explanation of the reasons for the effect of fertiliser on yield of final product.

1.4 Requirements in design

The objective in the type of experiment studied here is the comparison of the effect of treatments on response. This will typically be assessed by estimates and confidence limits for the magnitude of treatment differences. Requirements on such estimates are essentially as follows. First systematic errors, or biases, are to be avoided. Next the effect of random errors should so far as feasible be minimized. Further it should be possible to make reasonable assessment of the magnitude of random errors, typically via confidence limits for the comparisons of interest. The scale of the investigation should be such as to achieve a useful but not unnecessarily high level of precision. Finally advantage should be taken of any special structure in the treatments, for example when these are specified by combinations of factors.

The relative importance of these aspects is different in different fields of study. For example in large clinical trials to assess relatively small differences in treatment efficacy, avoidance of systematic error is a primary issue. In agricultural field trials, and probably more generally in studies that do not involve human sub-

jects, avoidance of bias, while still important, is not usually the aspect of main concern.

These objectives have to be secured subject to the practical constraints of the situation under study. The designs and considerations developed in this book have often to be adapted or modified to meet such constraints.

1.5 Interplay between design and analysis

There is a close connection between design and analysis in that an objective of design is to make both analysis and interpretation as simple and clear as possible. Equally, while some defects in design may be corrected by more elaborate analysis, there is nearly always some loss of security in the interpretation, i.e. in the underlying subject-matter meaning of the outcomes.

The choice of detailed model for analysis and interpretation will often involve subject-matter considerations that cannot readily be discussed in a general book such as this. Partly but not entirely for this reason we concentrate here on the analysis of continuously distributed responses via models that are usually linear, leading to analyses quite closely connected with the least-squares analysis of the normal theory linear model. One intention is to show that such default analyses follow from a single set of assumptions common to the majority of the designs we shall consider. In this rather special sense, the model for analysis is determined by the design employed. Of course we do not preclude the incorporation of special subject-matter knowledge and models where appropriate and indeed this may be essential for interpretation.

There is a wider issue involved especially when a number of different response variables are measured and underlying interpretation is the objective rather than the direct estimation of treatment differences. It is sensible to try to imagine the main patterns of response that are likely to arise and to consider whether the information will have been collected to allow the interpretation of these. This is a broader issue than that of reviewing the main scheme of analysis to be used. Such consideration must always be desirable; it is, however, considerably less than a prior commitment to a very detailed approach to analysis.

Two terms quite widely used in discussions of the design of experiments are *balance* and *orthogonality*. Their definition depends a bit on context but broadly balance refers to some strong symmetry

in the combinatorial structure of the design, whereas orthogonality refers to special simplifications of analysis and achievement of efficiency consequent on such balance.

For example, in Chapter 3 we deal with designs for a number of treatments in which the experimental units are arranged in blocks. The design is balanced if each treatment occurs in each block the same number of times, typically once. If a treatment occurs once in some blocks and twice or not at all in others the design is considered unbalanced. On the other hand, in the context of balanced incomplete block designs studied in Section 4.2 the word balance refers to an extended form of symmetry.

In analyses involving a linear model, and most of our discussion centres on these, two types of effect are orthogonal if the relevant columns of the matrix defining the linear model are orthogonal in the usual algebraic sense. One consequence is that the least squares estimates of one of the effects are unchanged if the other type of effect is omitted from the model. For orthogonality some kinds of balance are sufficient but not necessary. In general statistical theory there is an extended notion of orthogonality based on the Fisher information matrix and this is relevant when maximum likelihood analysis of more complicated models is considered.

1.6 Key steps in design

1.6.1 General remarks

Clearly the single most important aspect of design is a purely substantive, i.e. subject-matter, one. The issues addressed should be interesting and fruitful. Usually this means examining one or more well formulated questions or research hypotheses, for example a speculation about the process underlying some phenomenon, or the clarification and explanation of earlier findings. Some investigations may have a less focused objective. For example, the initial phases of a study of an industrial process under production conditions may have the objective of identifying which few of a large number of potential influences are most important. The methods of Section 5.6 are aimed at such situations, although they are probably atypical and in most cases the more specific the research question the better.

In principle therefore the general objectives lead to the following more specific issues. First the experimental units must be defined

and chosen. Then the treatments must be clearly defined. The variables to be measured on each unit must be specified and finally the size of the experiment, in particular the number of experimental units, has to be decided.

1.6.2 Experimental units

Issues concerning experimental units are to some extent very specific to each field of application. Some points that arise fairly generally and which influence the discussion in this book include the following.

Sometimes, especially in experiments with a technological focus, it is useful to consider the population of ultimate interest and the population of accessible individuals and to aim at conclusions that will bridge the inevitable gap between these. This is linked to the question of whether units should be chosen to be as uniform as possible or to span a range of circumstances. Where the latter is sensible it will be important to impose a clear structure on the experimental units; this is connected with the issue of the choice of baseline measurements.

Illustration. In agricultural experimentation with an immediate objective of making recommendations to farmers it will be important to experiment in a range of soil and weather conditions; a very precise conclusion in one set of conditions may be of limited value. Interpretation will be much simplified if the same basic design is used at each site. There are somewhat similar considerations in some clinical trials, pointing to the desirability of multi-centre trials even if a trial in one centre would in principle be possible.

By contrast in experiments aimed at elucidating the nature of certain processes or mechanisms it will usually be best to choose units likely to show the effect in question in as striking a form as possible and to aim for a high degree of uniformity across units.

In some contexts the same individual animal, person or material may be used several times as an experimental unit; for example in a psychological experiment it would be common to expose the same subject to various conditions (treatments) in one session.

It is important in much of the following discussion and in applications to distinguish between experimental units and observations. The key notion is that different experimental units must in principle be capable of receiving different treatments.

Illustration. In an industrial experiment on a batch process each separate batch of material might form an experimental unit to be processed in a uniform way, separate batches being processed possibly differently. On the product of each batch many samples may be taken to measure, say purity of the product. The number of observations of purity would then be many times the number of experimental units. Variation between repeat observations within a batch measures sampling variability and internal variability of the process. Precision of the comparison of treatments is, however, largely determined by, and must be estimated from, variation between batches receiving the same treatment. In our theoretical treatment that follows the number of batches is thus the relevant total "sample" size.

1.6.3 Treatments

The simplest form of experiment compares a new treatment or manipulation, T, with a control, C. Even here care is needed in applications. In principle T has to be specified with considerable precision, including many details of its mode of implementation. The choice of control, C, may also be critical. In some contexts several different control treatments may be desirable. Ideally the control should be such as to isolate precisely that aspect of T which it is the objective to examine.

Illustration. In a clinical trial to assess a new drug, the choice of control may depend heavily on the context. Possible choices of control might be no treatment, a placebo treatment, i.e. a substance superficially indistinguishable from T but known to be pharmacologically inactive, or the best currently available therapy. The choice between placebo and best available treatment may in some clinical trials involve difficult ethical decisions.

In more complex situations there may be a collection of qualitatively different treatments $T_1, \ldots, T_v$. More commonly the treatments may have factorial structure, i.e. be formed from combinations of levels of subtreatments, called *factors*. We defer detailed study of the different kinds of factor and the design of factorial experiments until Chapter 5, noting that sensible use of the principle of examining several factors together in one study is one of the most powerful ideas in this subject.

1.6.4 Measurements

The choice of appropriate variables for measurement is a key aspect of design in the broad sense. The nature of measurement processes and their associated potentiality for error, and the different kinds of variable that can be measured and their purposes are central issues. Nevertheless these issues fall outside the scope of the present book and we merely note three broad types of variable, namely baseline variables describing the experimental units before application of treatments, intermediate variables and response variables, in a medical context often called end-points.

Intermediate variables may serve different roles. Usually the more important is to provide some provisional explanation of the process that leads from treatment to response. Other roles are to check on the absence of untoward interventions and, sometimes, to serve as surrogate response variables when the primary response takes a long time to obtain.

Sometimes the response on an experimental unit is in effect a time trace, for example of the concentrations of one or more substances as transient functions of time after some intervention. For our purposes we suppose such responses replaced by one or more summary measures, such as the peak response or the area under the response-time curve.

Clear decisions about the variables to be measured, especially the response variables, are crucial.

1.6.5 Size of experiment

Some consideration virtually always has to be given to the number of experimental units to be used and, where subsampling of units is employed, to the number of repeat observations per unit. A balance has to be struck between the marginal cost per experimental unit and the increase in precision achieved per additional unit. Except in rare instances where these costs can both be quantified, a decision on the size of experiment is bound be largely a matter of judgement and some of the more formal approaches to determining the size of the experiment have spurious precision. It is, however, very desirable to make an advance approximate calculation of the precision likely to be achieved. This gives some protection against wasting resources on unnecessary precision or, more commonly, against undertaking investigations which will be

of such low precision that useful conclusions are very unlikely. The same calculations are advisable when, as is quite common in some fields, the maximum size of the experiment is set by constraints outside the control of the investigator. The issue is then most commonly to decide whether the resources are sufficient to yield enough precision to justify proceeding at all.

1.7 A simplified model

The formulation of experimental design that will largely be used in this book is as follows. There are given n experimental units, $U_1, \ldots, U_n$ and v treatments, $T_1, \ldots, T_v$; one treatment is applied to each unit as specified by the investigator, and one response Y measured on each unit. The objective is to specify procedures for allocating treatments to units and for the estimation of the differences between treatments in their effect on response.

This is a very limited version of the broader view of design sketched above. The justification for it is that many of the valuable specific designs are accommodated in this framework, whereas the wider considerations sketched above are often so subject-specific that it is difficult to give a general theoretical discussion.

It is, however, very important to recall throughout that the path between the choice of a unit and the measurement of final response may be a long one in time and in other respects and that random and systematic error may arise at many points. Controlling for random error and aiming to eliminate systematic error is thus not a single step matter as might appear in our idealized model.

1.8 A broader view

The discussion above and in the remainder of the book concentrates on the integrity of individual experiments. Yet investigations are rarely if ever conducted in isolation; one investigation almost inevitably suggests further issues for study and there is commonly the need to establish links with work related to the current problems, even if only rather distantly. These are important matters but again are difficult to incorporate into formal theoretical discussion.

If a given collection of investigations estimate formally the same contrasts, the statistical techniques for examining mutual consistency of the different estimates and, subject to such consistency, of combining the information are straightforward. Difficulties come

more from the choice of investigations for inclusion, issues of genuine comparability and of the resolution of apparent inconsistencies.

While we take the usual objective of the investigation to be the comparison of responses from different treatments, sometimes there is a more specific objective which has an impact on the design to be employed.

Illustrations. In some kinds of investigation in the chemical process industries, the treatments correspond to differing concentrations of various reactants and to variables such as pressure, temperature, etc. For some purposes it may be fruitful to regard the objective as the determination of conditions that will optimize some criterion such as yield of product or yield of product per unit cost. Such an explicitly formulated purpose, if adopted as the sole objective, will change the approach to design.

In selection programmes for, say, varieties of wheat, the investigation may start with a very large number of varieties, possibly several hundred, for comparison. A certain fraction of these are chosen for further study and in a third phase a small number of varieties are subject to intensive study. The initial stage has inevitably very low precision for individual comparisons and analysis of the design strategy to be followed best concentrates on such issues as the proportion of varieties to be chosen at each phase, the relative effort to be devoted to each phase and in general on the properties of the whole process and the properties of the varieties ultimately selected rather than on the estimation of individual differences.

In the pharmaceutical industry clinical trials are commonly defined as Phase I, II or III, each of which has quite well-defined objectives. Phase I trials aim to establish relevant dose levels and toxicities, Phase II trials focus on a narrowly selected group of patients expected to show the most dramatic response, and Phase III trials are a full investigation of the treatment effects on patients broadly representative of the clinical population.

In investigations with some technological relevance, even if there is not an immediate focus on a decision to be made, questions will arise as to the practical implications of the conclusions. Is a difference established big enough to be of public health relevance in an epidemiological context, of relevance to farmers in an agricultural context or of engineering relevance in an industrial context? Do the conditions of the investigation justify extrapolation to the work-

ing context? To some extent such questions can be anticipated by appropriate design.

In both scientific and technological studies estimation of effects is likely to lead on to the further crucial question: what is the underlying process explaining what has been observed? Sometimes this is expressed via a search for causality. So far as possible these questions should be anticipated in design, especially in the definition of treatments and observations, but it is relatively rare for such explanations to be other than tentative and indeed they typically raise fresh issues for investigation.

It is sometimes argued that quite firm conclusions about causality are justified from experiments in which treatment allocation is made by objective randomization but not otherwise, it being particularly hazardous to draw causal conclusions from observational studies.

These issues are somewhat outside the scope of the present book but will be touched on in Section 2.5 after the discussion of the role of randomization. In the meantime some of the potential implications for design can be seen from the following Illustration.

Illustration. In an agricultural field trial a number of treatments are randomly assigned to plots, the response variable being the yield of product. One treatment, S, say, produces a spectacular growth of product, much higher than that from other treatments. The growth attracts birds from many kilometres around, the birds eat most of the product and as a result the final yield for S is very low. Has S caused a depression in yield?

The point of this illustration, which can be paralleled from other areas of application, is that the yield on the plots receiving S is indeed lower than the yield would have been on those plots had they been allocated to other treatments. In that sense, which meets one of the standard definitions of causality, allocation to S has thus caused a lowered yield. Yet in terms of understanding, and indeed practical application, that conclusion on its own is quite misleading. To understand the process leading to the final responses it is essential to observe and take account of the unanticipated intervention, the birds, which was supplementary to and dependent on the primary treatments. Preferably also intermediate variables should be recorded, for example, number of plants per square metre and measures of growth at various time points in the growing cycle. These will enable at least a tentative account to be developed of

the process leading to the treatment differences in final yield which are the ultimate objective of study. In this way not only are treatment differences estimated but some partial understanding is built of the interpretation of such differences. This is a potentially causal explanation at a deeper level.

Such considerations may arise especially in situations in which a fairly long process intervenes between treatment allocation and the measurement of response.

These issues are quite pressing in some kinds of clinical trial, especially those in which patients are to be followed for an appreciable time. In the simplest case of randomization between two treatments, T and C, there is the possibility that some patients, called noncompliers, do not follow the regime to which they have been allocated. Even those who do comply may take supplementary medication and the tendency to do this may well be different in the two treatment groups. One approach to analysis, the so-called *intention-to-treat principle*, can be summarized in the slogan "ever randomized always analysed": one simply compares outcomes in the two treatment arms regardless of compliance or noncompliance. The argument, parallel to the argument in the agricultural example, is that if, say, patients receiving T do well, even if few of them comply with the treatment regimen, then the consequences of allocation to T are indeed beneficial, even if not necessarily because of the direct consequences of the treatment regimen.

Unless noncompliance is severe, the intention-to-treat analysis will be one important analysis but a further analysis taking account of any appreciable noncompliance seems very desirable. Such an analysis will, however, have some of the features of an observational study and the relatively clearcut conclusions of the analysis of a fully compliant study will be lost to some extent at least.

1.9 Bibliographic notes

While many of the ideas of experimental design have a long history, the first major systematic discussion was by R. A. Fisher (1926) in the context of agricultural field trials, subsequently developed into his magisterial book (Fisher, 1935 and subsequent editions). Yates in a series of major papers developed the subject much further; see especially Yates (1935, 1936, 1937). Applications were initially largely in agriculture and the biological sciences and then subsequently in industry. The paper by Box and Wilson (1951) was

particularly influential in an industrial context. Recent industrial applications have been particularly associated with the name of the Japanese engineer, G. Taguchi. General books on scientific research that include some discussion of experimental design include Wilson (1952) and Beveridge (1952).

Of books on the subject, Cox (1958) emphasizes general principles in a qualitative discussion, Box, Hunter and Hunter (1978) emphasize industrial experiments and Hinkelman and Kempthorne (1994), a development of Kempthorne (1952), is closer to the originating agricultural applications. Piantadosi (1997) gives a thorough account of the design and analysis of clinical trials.

Vajda (1967a, 1967b) and Street and Street (1987) emphasize the combinatorial problems of design construction. Many general books on statistical methods have some discussion of design but tend to put their main emphasis on analysis; see especially Montgomery (1997). For very careful and systematic expositions with some emphasis respectively on industrial and biometric applications, see Dean and Voss (1999) and Clarke and Kempson (1997).

An annotated bibliography of papers up to the late 1960's is given by Herzberg and Cox (1969).

The notion of causality has a very long history although traditionally from a nonprobabilistic viewpoint. For accounts with a statistical focus, see Rubin (1974), Holland (1986), Cox (1992) and Cox and Wermuth (1996; section 8.7). Rather different views of causality are given by Dawid (2000), Lauritzen (2000) and Pearl (2000). For a discussion of compliance in clinical trials, see the papers edited by Goetghebeur and van Houwelingen (1998).

New mathematical developments in the design of experiments may be found in the main theoretical journals. More applied papers may also contain ideas of broad interest. For work with a primarily industrial focus, see *Technometrics*, for general biometric material, see *Biometrics*, for agricultural issues see the *Journal of Agricultural Science* and for specialized discussion connected with clinical trials see *Controlled Clinical Trials, Biostatistics* and *Statistics in Medicine. Applied Statistics* contains papers with a wide range of applications.

1.10 Further results and exercises

1. A study to investigate the association between car telephone usage and accident rates was reported by Redelmeier and Tib-

shirani (1997a) and a further careful account discussed in detail the study design (Redelmeier and Tibshirani, 1997b). A randomized trial was infeasible on ethical grounds, and the investigators decided to conduct a case-control study. The cases were those individuals who had been in an automobile collision involving property damage (but not personal injury), who owned car phones, and who consented to having their car phone usage records reviewed.

(a) What considerations would be involved in finding a suitable control for each case?

(b) The investigators decided to use each case as his own control, in a specialized version of a case-control study called a case-crossover study. A "case driving period" was defined to be the ten minutes immediately preceding the collision. What considerations would be involved in determining the control period?

(c) An earlier study compared the accident rates of a group of drivers who owned cellular telephones to a group of drivers who did not, and found lower accident rates in the first group. What potential biases could affect this comparison?

2. A prospective case-crossover experiment to investigate the effect of alcohol on blood œstradiol levels was reported by Ginsberg et al. (1996). Two groups of twelve healthy postmenopausal women were investigated. One group was regularly taking œstrogen replacement therapy and the second was not. On the first day half the women in each group drank an alcoholic cocktail, and the remaining women had a similar juice drink without alcohol. On the second day the women who first had alcohol were given the plain juice drink and vice versa. In this manner it was intended that each woman serve as her own control.

(a) What precautions might well have been advisable in such a context to avoid bias?

(b) What features of an observational study does this study have?

(c) What features of an experiment does this study have?

3. Find out details of one or more medical studies the conclusions from which have been reported in the press recently. Were they experiments or observational studies? Is the design (or analysis) open to serious criticism?

4. In an experiment to compare a number of alternative ways of treating back pain, pain levels are to be assessed before and after a period of intensive treatment. Think of a number of ways in which pain levels might be measured and discuss their relative merits. What measurements other than pain levels might be advisable?
5. As part of a study of the accuracy and precision of laboratory chemical assays, laboratories are provided with a number of nominally identical specimens for analysis. They are asked to divide each specimen into two parts and to report the separate analyses. Would this provide an adequate measure of reproducibility? If not recommend a better procedure.
6. Some years ago there was intense interest in the possibility that cloud-seeding by aircraft depositing silver iodide crystals on suitable cloud would induce rain. Discuss some of the issues likely to arise in studying the effect of cloud-seeding.
7. Preece et al. (1999) simulated the effect of mobile phone signals on cognitive function as follows. Subjects wore a headset and were subject to (i) no signal, (ii) a 915 MHz sine wave analogue signal, (iii) a 915 MHz sine wave modulated by a 217 Hz square wave. There were 18 subjects, and each of the six possible orders of the three conditions were used three times. After two practice sessions the three experimental conditions were used for each subject with 48 hours between tests. During each session a variety of computerized tests of mental efficiency were administered. The main result was that a particular reaction time was shorter under the condition (iii) than under (i) and (ii) but that for 14 other types of measurement there were no clear differences. Discuss the appropriateness of the control treatments and the extent to which stability of treatment differences across sessions might be examined.
8. Consider the circumstances under which the use of two different control groups might be valuable. For discussion of this for observational studies, where the idea is more commonly used, see Rosenbaum (1987).