5

Experimental designs to assess the effect of a treatment

5.1 Introduction

As was pointed out in Section 1.6, the subject of experimental design is usually thought of beginning with the work of Sir Ronald Fisher at Rothamstead Experimental Station in Britain, in the early part of this century. Two of his classical books on the subject were *Statistical Methods for Research Workers*, first published in 1925, and *The Design of Experiments*, first published in 1935. The subject has come a long way in the years since then, and yet there are still aspects of it that are not covered well in most textbooks. One reason for this is that the theory of experimental design began with applications in agriculture, an area where the researcher has a great deal of freedom to manipulate the experimental materials and to apply different treatments. However, in many other areas there are very real constraints as to what can be done, either for ethical reasons, or because the experimental units are human subjects with minds of their own.

The example of assessing the effect on heart disease of a high or low salt diet that was discussed in Section 1.6 suggests that it is important, at least in some disciplines, to consider potential experimental designs that are less than perfect but at least can be used. It is important, then, to understand the weaknesses of these *quasi-experimental designs*. Much of the theoretical work in this area has been done by those working in the areas of education research and in the evaluation of social programmes. However, this work is relevant to experimenters in many other subject areas as well.

This chapter is concerned mainly with the simplest type of experiment where the *experimental units* (for example, human or animal subjects) are in two groups. One of the groups is a *control group*, which receives the usual conditions, while the other group is the *treatment or experimental group*, which receives some special treatment. The question of interest is whether some measurement

taken on the experimental units is changed by the special treatment. Situations where a treatment can take several different forms, so that several experimental groups are needed, are also considered to some extent in this chapter. For these, the question is whether some or all the treatments change a measurement on the experimental units.

More complicated types of experiment are discussed in the next two chapters. Chapter 6 is concerned with cases where the question of interest is whether a treatment has an impact on the values in a time series, while Chapter 7 is concerned with experiments in which two or more factors are manipulated at the same time. It is the material in Chapter 7 that is usually emphasized in books on the statistics of experimental designs.

In Chapter 1 a distinction has been made between an experimental study, where events are controlled by the investigator, and an observational study, where events are uncontrolled. However, in practice this distinction may be somewhat blurred because an observational study would have been an experimental study if it were not for the fact that the assignment of the treatment to experimental units was either an accident, or made by some agency for reasons unconnected with the assessment of the effects of the treatment.

For example, consider Case Study 4, which was concerned with the effects of a fire on the sales in a hardware store. Here the 'treatment' was the fire, which was not started deliberately, so that this is an observational study. However, the data would presumably have been exactly the same if the fire had been started deliberately to see what would happen. Therefore, this can be considered to be a type of unplanned experiment.

Apart from situations of this type, there are also, of course, factors that cannot be manipulated because of their essential nature, such as sex, ethnic origins, age, and so on. The point here is that to all intents and purposes many observational studies can be thought of as unplanned 'experiments' to which the discussions in this and later chapters are relevant.

5.2 External and internal validity

In the social science literature the terms *internal validity* and *external validity* are used in the assessment of the problems that may occur with particular types of experimental design. Internal

validity concerns whether the apparent effects or lack of effects shown by the experimental results are due to the factor being studied, rather than some alternative factor. External validity concerns the extent to which the results of an experiment can be generalized to some wider population of interest.

An experiment without internal validity is useless because what is being measured is wrong. In other words, the treatment effect is confounded with the effects of other factors. To take a specific example, suppose that an experimenter is interested in the effect of four different diets on the growth of rats. The available rats are divided at random into four groups, each of which is given one of the diets. The weight gain in one month is then measured for each rat, and significantly different mean weight gains for different diets are required as evidence of diet effects. This seems straightforward, but the internal validity of the experiment would be completely lost if all the rats on one diet were kept in the same cage, since the differences in diets would be completely confounded with the differences in cages. This means that one plausible explanation for the different weight gains with different diets is that the weight gains were dependent on the cages rather than the diets. In other words, all the rats in some cages would have had relatively high growth and all the rats in other cages would have had relatively low growth, irrespective of which diet they were given.

With care, the internal validity of an experiment like this can be ensured so that it is possible to say with confidence that there is either evidence for a diet effect or not, under the conditions used. However, it would still be questionable to use the experiment to draw conclusions about the effect of different diets on wild rats. Hence the external validity of the experiment may be low in this respect, although generalization to what will happen in other laboratories might be realistic.

The potential threats to the internal validity of an experiment are the *alternative explanations* for a treatment effect. To a large extent these threats depend on the specific situation, but the following list gives an idea of the possibilities (Campbell and Stanley, 1963):

(a) History: specific events occur during the course of the experiment in addition to what is required by the design
(b) Maturation: the experimental units change during the experiment simply as a result of the passing of time
(c) Testing: taking one measurement on a unit changes the values of later measurements
(d) Instrumentation: there are changes either in the method for measuring the responses of experimental units or in the agents making the measurements
(e) Regression to the mean: if the allocation of experimental units to the experimental and control groups is made on the basis of pretest scores, with those with low scores going to one group and those with high scores going to the other, then on retesting there is a tendency for the two groups to have closer means than on the original test
(f) Differential selection: because of the method for selecting groups, certain types of experimental unit are mainly in either the experimental or the control group
(g) Experimental mortality: there are losses of experimental units during the experiment, which are at different rates for the experimental and control groups

Factors which effect the extent to which the results of an experiment can be generalized (the external validity) are:

(a) Testing interaction: where a pretest changes the sensitivity of experimental units to the experimental treatments in such a way that the measured treatment effect is not what it would be in the population of interest
(b) Selection interaction: where the nature of the process for selecting experimental units means that treatment effects are likely to be higher or lower than for units in the population of interest
(c) Reactions to experimental arrangements: where the changed circumstances necessary for an experiment cause changes in the measurements on experimental units
(d) Multiple treatment effects: where each unit gets several treatments and some effects of earlier treatments are present when later treatments are applied

These threats to the internal and external validity of experiments are considered further throughout the remainder of this chapter.

5.3 Pseudoreplication

One of the pitfalls in designing an experiment is the inadvertent use of *pseudoreplication*, which has been defined by Hurlbert