Chapter 7

Experiments on People and Animals

7.1 Introduction

In this chapter we consider experiments on people and animals. Examples are clinical trials, testing educational strategies, psychological experiments, animal nutrition experiments, and trials on ambient conditions in the work-place. There are some issues peculiar to such experiments: these are discussed in Sections 7.2, 7.6, 7.7 and 7.8. On the other hand, the plot structure of such experiments is typically not very complicated, usually being one of those covered in Chapters 6, 4 or 2. These plot structures are specialized to experiments on people and animals in Sections 7.3, 7.4 and 7.5 respectively.

What is an observational unit in such an experiment? Usually it is a person or animal for a set time-period. A measurement may be taken before the experiment starts. This measurement is sometimes called a *baseline* measurement. It may be used to block the people or animals. It may be used as a covariate in the analysis of the data. Thirdly, the data to be analysed may be the differences between the post-treatment measurement and the baseline measurement.

Sometimes measurements are taken at several different times without changing the treatments. This can be enforced on animals or on hospital patients, but if non-hospitalized patients are asked to come back for measuremnt too often they may simply drop out of the trial.

What is the experimental unit? It may be a person or animal for the duration of the trial. It may be a group of people or animals for the duration of the trial if treatments can be administered only to whole groups. For example, teaching methods in schools can normally be changed only on a whole-class basis. It may

be a person or animal for a shorter time-period than the whole trial if treatments are changed during the trial.

What should the blocks be? This will be answered in Sections 7.3–7.5, after disposing of a specious argument for avoiding experimentation altogether.

7.2 Historical controls

When a new drug is introduced, there is often a good collection of data on the performance of the previous standard drug on a large number of people. It is, therefore, tempting to assess the efficacy of the new drug by administering it to several patients and comparing the results to those of the previous standard drug. In this situation, the patients who had received the previous drug are known as *historical controls*.

The use of historical controls is not satisfactory, for several reasons. In the first place, the new drug is given later in time than the previous one, and any observed difference between the results could conceivably be due to different conditions at different times. For example, a new drug to alleviate the symptoms of asthma might appear to be better just because it was tried out in a year when air pollution was less serious. Secondly, there is no doubt that people respond differently to treatments if they know that they are in a trial. If patients receiving the new drug know that they are taking part in a trial but the historical controls did not, then the patients on the new drug may well do better, even if there is no chemical difference between the new drug and the standard. Thirdly, if historical controls are used then the new drug is likely to be offered preferentially to those patients whom the doctors think can most benefit from having it rather than the standard drug. Again, the new drug may appear to be better without actually being so.

The only way to avoid these biases is to have a *randomized controlled* trial. Here *controlled* means that any control treatment, such as 'no treatment' or a standard drug, is included in the current trial. Patients are selected for inclusion in the trial in the knowledge that they may be allocated to any of the treatments. Randomization (and blocking) are used to ensure that differences between treatments are not confounded with time differences (such as the air pollution) or differences between patients.

Similar remarks apply to new diets, new educational methods, new training methods for athletes and new working conditions in the workplace.

7.3 Cross-over trials

A cross-over trial is suitable when each person or animal or group can be used more than once. Then we block by *person* and by *time-period*, to get a row-column design. A typical plot-structure (before allocation of treatments) is shown in Figure 7.1.

The number of periods should not be so large that people are very likely to drop out of the trial before it finishes, whether through boredom, adverse reactions,

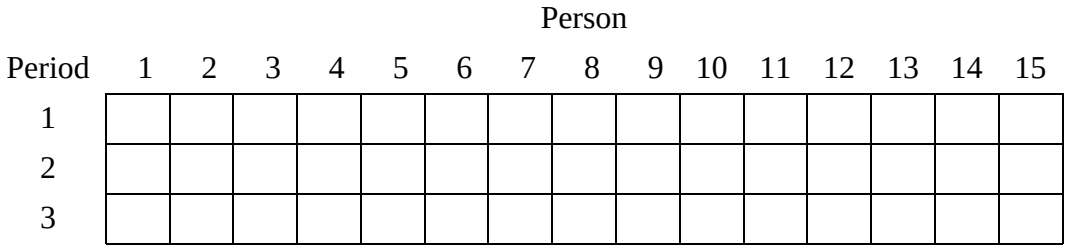

Figure 7.1: Typical plot structure in a cross-over trial

death or moving away from the area where the trial is conducted. Ideally, if the number t of treatments is small, use t periods and a multiple of t people. Then the construction methods of Chapter 6 can be used.

For a cross-over trial to be practicable and useful, several constraints must be satisfied.

(i) Sufficient people or animals or groups must be available at the same time.

(ii) Those people or animals or groups must be prepared to stay in the trial until the end. Since time-periods may be as short as hours or as long as years, it is the overall length of the trial that matters, not the number of periods. There is also a difference between involuntary subjects, such as animals or groups of school-children, and voluntary subjects, such as clinical patients, whether hospitalized or not.

(iii) No treatment should leave the subject in a very different state at the end of the period in which it is administered. In medicine, cross-over trials are best for chronic conditions, such as high blood pressure or asthma, and for testing drugs which may alleviate symptoms rather than curing the underlying illness. They are not good for drugs whose purpose is to cure the underlying illness; nor are they good for diseases where the patient has such a poor prognosis that he may not survive to the end of the trial.

Similarly, if the purpose of a particular teaching method is to improve pupils' understanding of Pythagoras' Theorem, then they cannot unlearn this and start afresh with another method: a cross-over trial is unsuitable. If the purpose of a particular diet is to confer long-term resistance to some disease, it should not be tested in a cross-over trial.

(iv) No treatment should have an effect which lasts into subsequent time-periods. Such effects are called *residual* effects or *carry-over* effects. Examples include drugs with delayed reactions and diets for cows which improve milk-yield over several subsequent weeks.

To some extent this problem can be circumvented by separating the time-periods in the trial by neutral periods (possibly of a different length from the trial periods) in which a common non-experimental treatment is given to

all subjects. Such periods are called *wash-out* periods. They must be long enough for residual effects to wear off. They increase the length of the trial, hence the cost and the probability that some subjects will drop out.

There are methods of designing and analysing cross-over trials when there are carry-over effects, but they are beyond the scope of this book.

7.4 Matched pairs, matched threes, and so on

The advantage of a cross-over trial is that people are blocks, so differences between people are eliminated from differences between treatments. If a cross-over trial is not suitable, then each person can be used only once and we must block in some other way. If there are t treatments, use everything relevant that you know about the people (age, weight, state of health, educational background, ...) to group people into blocks of size t. Then do a complete-block design. If $t = 2$, this is called a *matched pairs* design. Many clinical trials have only two treatments (standard drug and new drug, or 'no treatment' and new treatment), so matched pairs designs are quite common. However, the same principles apply equally well to matched threes, matched fours, etc.

One disadvantage of having a large number of small blocks is that many degrees of freedom are assigned to within-block differences, leaving fewer for the residual, so there may be a loss of power. If there are a large number of people in the trial and few relevant blocking factors then it may be better to have block size a multiple of t rather than t itself. Design and analysis is still as in Chapter 4. In clinical trials such large blocks are sometimes called *strata*, but this use of the word is different from our use of *strata* for the eigenspaces of the covariance matrix.

For such a design to be used, the experimenter must have prior information about the people or animals involved.

7.5 Completely randomized designs

If people or animals enter the trial at different times, with no prior information about them, then no blocking may be possible. This is a common situation in clinical trials, where it may be decided that the experimental units will be the next 120 patients who arrive at the surgery satisfying a list of criteria and who agree to participate in the trial. It can also happen in a nutrition trial on young animals if animals are to enter the trial soon after birth.

Since there is no prior information about the experimental units, no blocking is possible, so a completely randomized design is used.

Decide in advance on the number N of people or animals. Produce a randomized list of treatments for the N people as in Section 2.1. As each person enters the trial, allocate them to the next treatment on the list.

Sometimes a slight modification of this procedure allows a coarse form of blocking. If several hospitals are entering patients into the trial, it can be decided in

advance that each hospital will have k patients, where k is a multiple of t. Each hospital then has its own randomized list, like one of the blocks in Section 4.3. Anothe strategy is to block by time, so that the first k patients entering the trial form one block, the next k patients the second block, and so on.

7.6 Sequential allocation to an unknown number of patients

Some experimenters would like to start a trial without specifiying the number of patients in advance. If results on some patients become known before all patients have entered the trial, this presents the temptation to keep analysing intermediate results. Even if there is no effective difference between the new drug and the old, a test at the 5% significance level will show a difference one time in twenty, and the experimenter can then report the favourable result he is looking for. To avoid such spurious results, there is now an elaborate set of rules for intermediate analyses and for deciding when to stop the trial: these are beyond the scope of this book.

Even if no results are known while patients are still entering the trial, there are situations where it is unrealistic to specify the exact number of patients in advance. For example, should a trial on a rare disease have to be prolonged indefinitely while the experimenter waits for just two more patients? Some strategies have been suggested for sequential randomization of an unknown number of patients, but none is entirely satisfactory. For simplicity, I shall describe them for the case of two treatments.

The first is to simply toss a coin for each new patient. This has the disadvantage that the replications are most unlikely to be equal, so the variance of the estimator of the treatment difference will be larger than it needs be.

The second method is to toss a biased coin, so that the next patient is more likely to receive the treatment that has been under-represented so far. This has the disadvantage that a pair of successive patients are more likely to receive different treatments early on than they are later, which contradicts the randomization assumption made in Chapter 1.

The third method is to block by time. However, if blocks are large then there is still a danger of very unequal replication. If blocks are small then many degrees of freedom are lost to the between-blocks contrasts, which may lose power.

The fourth method is called *minimization*. This is a version of the biased-coin strategy that seeks to block on several factors unknown in advance (such as age, sex, smoking history). The coin is biased in such way that replication should be almost equal within each category of each blocking factor. Unfortunately, a side effect is that replication will be very unequal in categories of combinations of two blocking factors, so the design will be poor if there is, say, an interaction between age and sex.

7.7 Safeguards against biases

Early experiments on working conditions in factories showed that the productivity of the workers in the experiment improved no matter what conditions they had. Management was taking more notice of these workers than usual, and this produced a positive effect irrespective of the experimental conditions. Thus everyone who is taking part in an experiment should be equally aware that they are doing so.

Similarly, it is now well known that patients improve if they receive a treatment which they perceive to be beneficial, whether or not it is. Thus, in a clinicial trial, any null treatment should be given as a dummy treatment, which is called a *placebo*. For example, if the new drug is administered as a small blue tablet then the placebo should be a similar small blue tablet but without the active ingredient. Another reason for using placebos is to avoid confusing the effect of the drug with the effect of the regime of taking the drug.

Since people, or even animals, may respond for psychosomatic reasons, they should not know which treatment they are receiving. The trial is said to be *blind* is this condition is achieved. It is not always possible. In a comparison of methods of physiotherapy for treating lower back pain, patients will be well aware of the exercises they are doing. Similarly, in many psychological, social or educational experiments the subjects cannot avoid knowing what treatment they are getting.

Doctors, vets and other professionals involved should also not know the treatments allocated to each subject, because their own expectations or ethics may influence the result. *Assessment bias* occurs if a doctor recording a subjective rating does so more or less favourably according to whether he already thinks the treatment allocated to that patient is better or worse.

If neither patient nor doctor knows the treatment then the trial is said to be *double-blind*.

Example 7.1 (Educational psychology) An educational psychologist wanted to compare two different methods of presenting information. Her experimental units were thirty undergraduate volunteers from the Psychology department. They volunteered sequentially by arriving at the psychologist's office. In a private fifteen-minute session she presented the new information by either method A or method B and then gave the student a short test to find out how much of the information they had absorbed.

Her method of randomization was to toss a coin to decide between methods for the first volunteer, and to alternate between A and B thereafter. There are several reasons why this method is flawed. First, the students would talk among themselves and soon discover such an obvious pattern as $ABABA\ldots$ and so could deliberately present themselves in a special order; for example, deliberately alternating people who were good at method A with people who were good at method B. Secondly, the psychologist herself was aware of the simple pattern and would always know the next method to use, which might unconsciously affect her decision about whether to accept the next volunteer as suitable.

The third reason is more subtle, but has already been touched on in Section 7.6. If an experiment is to be analysed as 'unstructured' by the methods of Chapter 2, then the probability that the difference $y_\alpha - y_\beta$ contributes to residual or to an estimator of a treatment difference should be the same for all pairs of distinct experimental units α and β. In other words, there must be a fixed probability p such that

$$\Pr[T(\alpha) = T(\beta)] = p$$

whenever $\alpha \neq \beta$. (For an equireplicate design, p must be equal to $(r-1)/(N-1)$, because there are $r-1$ other experimental units which receive the same treatment as α does.) The psychologist's method of randomization does not achieve this, because

$$\Pr[T(\alpha) = T(\beta)] = 1$$

if α and β are volunteers whose positions in the sequence are both even numbers or both odd numbers, and

$$\Pr[T(\alpha) = T(\beta)] = 0$$

otherwise.

Example 7.2 (Lanarkshire milk experiment) An experiment in Lanarkshire in the early twentieth century demonstrated the conflict between the ethics of the professionals involved and the statistical needs of the experiment. Extra milk was given to a random selection of pupils at some schools to see if it affected their growth. At the end of the experiment it was discovered that the teachers had altered the random allocation to ensure that extra milk was given to the most under-nourished children. Their good intentions ruined the experiment, by confounding the effect of the extra milk with the initial state of health.

Example 7.3 (Doctor knows best) Similarly, a doctor's receptionist told me that her employer looked to see what was the next treatment in the randomization list before deciding whether or not to enter his next suitable patient in the trial. His argument was that, as the doctor, he knew better than any statistician what treatment was good for his patient. He did not see that his actions were biasing the whole trial and therefore possibly delaying the introduction of a more effective treatment. In fact, I hope that his boast was empty. The randomization list should be kept at a separate location, and the next treatment allocated only *after* the next patient has been entered in the trial.

Selection bias occurs if the experimenter or doctor consciously chooses which person to allocate to which treatment or consciously decides which person to include next in the trial on the basis of what treatment they will receive. This is what happened in Examples 7.2 and 7.3 respectively. Selection bias is not necessarily present with a randomization scheme like the one in Example 7.1, but the fact that the experimenter had the knowledge to enable her to bias her choice of subjects should make other scientists question the validity of the study. A different form of selection bias occurs if patients choose their own treatment.

Example 7.4 (AIDS tablets) An AIDS clinic in Bangkok offered its new tablets to 117 severely ill patients. Of these, 53 accepted the tablets. On average, these 53 lived for 5 weeks longer than the 64 patients who declined the tablets. The makers of the tablets claimed that this showed their efficacy. Other scientists pointed out that the patients who chose the tablets may have been healthier than the others, in that they still had the will to care about their future and seek cures.

In clinical trials it is rare for either the plot structure or the treatment structure to be complicated. The most important issues are usually replication and avoidance of bias.

7.8 Ethical issues

There are ethical issues in experiments on people and animals that simply do not arise in experiments on plants or on industrial processes.

An experimental treatment should not be applied to people if it is expected to cause harm. If a person under treatment appears to be suffering adverse side-effects they should probably be withdrawn from the trial. This is another reason why it is hard to achieve equal replication in clinical trials and why complicated plot structure should be avoided.

For a given illness, if there is already a standard drug which is known to be effective then it is not ethical to give no treatment in a clinical trial of a new drug for that illness. The control treatment must be the current standard drug.

Example 7.5 (Incomplete factorial) It is known that drugs A and B are both effective against a certain illness. Someone suggests a 2×2 factorial trial in which the treatments are all four combinations of (A or not) with (B or not). However, it is not ethical to give a patient no drug at all when an effective drug is available. Thus there can be only three treatments: A alone, B alone, and A and B together.

Because it is unethical to give patients harmful treatments, or to force people into occupations or recreations or diets, information sometimes has to be gathered non-experimentally. One possibility is historical controls: the long life-span of people born in the middle of the twentieth century compared to those born 100 years earlier may be associated with improved diet. Another possibility is the *observational study*, which is used for conditions which arise too rarely for any planned intervention.

A *controlled* study is better than either of these, if it is feasible. A controlled *retrospective study* (also called a *case-control* study) to investigate nystagmus might select one thousand men with the disease and one thousand without. The control group should be chosen to match the diseased men in overall age distribution and in urban/rural environment. If the aim of the study is to find out if nystagmus is associated with occupation then people should be selected for the groups without their occupation being known. If it is then found that the proportion of miners

among the diseased men is much higher than the proportion in the control group, then there is some evidence that miners are more likely to get nystagmus.

A controlled *prospective* study (sometimes called a *cohort study*) is better than a restrospective one, but more difficult to organize. Now two groups are selected which differ in terms of the possible *cause* of the disease: for example, smokers and non-smokers. The groups are matched on other factors as closely as possible. The people are monitored for a long time, perhaps for the rest of their life, and the proportions getting lung-cancer in the two groups noted. A prospective case-control study is likely to be more accurate and less biased than a retrospective one, because it does not rely on people's memories and there is no possibility of cheating in forming the groups. However, it is much more costly. Not only does it last for several years, but more people are required to allow for the fact that some of them will be *lost to follow-up* in the sense that they drop out, fail to keep in touch, or die from unrelated causes.

With animals, the ethical situation is less clear. Everyone agrees that animals used in experiments should not be subjected to unnecessary suffering, but there is less agreement on what suffering is necessary. Many people argue that animals may suffer, and even be killed, in the course of an experiment to find a cure for a serious human disease. Even so, the number of animals used should be kept to a minimum. Is it ethical to use animals in experiments on smoking or on cosmetics? Some people think that no animal suffering in experiments is permissible.

Example 7.6 (Frogs) Amphibian numbers declined world-wide in the 1980s and 1990s. In one experiment to find the cause, laboratory frogs were injected with one of three pesticides (DDT, malathion or dieldrin) or left in their normal state. After some days, the frogs' immune response was measured. The experiment showed that these three pesticides dramatically reduce the number of antibodies produced by the frogs. Such experiments can pave the way for banning of the harmful pesticides. Is it unethical to deliberately poison frogs in this way?

Replication is more of an ethical issue in experiments on people and animals than it is in general. Too few people (or animals) or time-periods will give insufficient power to detect genuine differences, so any suffering will be in vain. On the other hand, it is unethical to continue a clinical trial after one treatment is known to be better.

Humans in a trial should normally give their *informed consent* to participation. This means that they should be told enough details about the trial to be able to make an informed decision about whether they participate. In particular, they must be informed that there is a chance that they will be allocated an inferior treatment. In the past, such consent has not always been sought from prisoners or from people with low intelligence.

It is clear that it is unethical to conduct a clinical trial without informed consent. It is less clear whether consent is needed for educational experiments or experiments on working conditions. A particularly grey area is the clinical trial in which a whole group of people must receive the same treatment.

Example 7.7 (Educating general practitioners) A trial was conducted to test the effectiveness of implementing certain guidelines for the treatment of diabetes. Some general practioners were randomized to the 'intervention' treatment and asked to attend some educational sessions where the new guidelines were explained; the other general practioners in the experiment were not invited to such sessions. However, the observational units in the experiment were the diabetic patients of the two sets of general practioners.

Example 7.8 (Maternal dietary supplements in Gambia) In small rural communities an intervention at community level may be more effective than one at the individual level because members of the community encourage each other to take the new drug or the new diet. For this reason, a trial of maternal dietary supplements in rural Gambia used villages as experimental units. The mothers (or rather, their babies, whose birth-weight was recorded) were the observational units.

7.9 Analysis by intention to treat

For minor diseases, a doctor cannot force a patient to continue to take an unpleasant medicine or continue with an intrusive therapy, even if the patient has volunteered for the experiment. So long as the results are analysed according to which treatment was *allocated* to the patient, as opposed to which treatment he actually took, this ethical principle has the effect that the results from a clinical trial are more likely to generalize to the population at large.

Example 7.9 (Mouthwash) 1200 dental students take part in a trial to see if either of two mouthwashes (*A* and *B*) is effective in preventing gum disease. 300 students are randomized to each mouthwash, 300 to a placebo mouthwash that is actually tap water, and 300 are given no mouthwash at all (why are these last two 'treatments' included?) Suppose that *A* is actually more effective than *B* but either tastes more unpleasant or needs to be taken more often. Then it is likely that students allocated to *A* will be less diligent in using it than students allocated to *B*. Some of them will be honest about this lack of diligence but many will many will be ashamed to admit that they have not followed instructions.

Suppose that, on average, gum disease affects 6% of those of student age who use no mouthwash, 2% of those who use mouthwash *B* and 1% of those who use mouthwash *A*, but that one third of people who try mouthwash *A* give it up. Then the results of the trial on these two treatments might be as follows.

	gum disease	no gum disease	total
A	2	196	198
A but gave up	6	96	102
B	6	274	280
B but gave up	1	19	20

If the analysis uses only those students who claim to have persevered with their mouthwash, then the estimated disease rate for *B* is between 2.14% and 2.33%

while that for *A* is between 1.01% and 2.67%, depending on how many students admit that they have given up. Thus which mouthwash appears to be better depends on the willingness of the students to admit that they have not been diligent. On the other hand, if the analysis uses all the students then *A* appears worse than *B*, and this is probably the correct conclusion if a mouthwash is to be recommended to the general public.

Questions for Discussion

7.1 Describe the sort of design that would be most appropriate for each of the following situations:

(i) investigating a new drug which purports to reduce high blood pressure;

(ii) investigating a new 'miracle' cure for the common cold;

(iii) finding out what teaching method is best for brilliant young mathematicians who go to university at age 14 or less;

(iv) comparing four diets for piglets to see which gives the greatest weight gain in the first six weeks of life;

(v) comparing three winter feed supplements for sheep to see how they affect the quality of the wool shorn the following spring;

(vi) comparing two styles of conducting large amateur choirs;

(vii) assessing whether 'brisk' or 'motherly' policewomen obtain the most useful information from rape victims in the interview immediately after the rape;

(viii) seeing whether either of two proposed new drugs to help people give up smoking has any effect.

7.2 How would you investigate the following claims?

(a) Dentists are more likely to suffer mercury poisoning than other people.

(b) For men aged over 50, taking one aspirin a day helps to prevent stroke (heart attack?)?

(c) Vegetarians live longer than omnivores.

(d) Children whose fathers were working at nuclear reprocessing plants at or before the time of conception are more susceptible to leukaemia than the general population.

(e) Publishing 'league tables' of schools by examination results actually decreases the amount that schoolchildren learn.

(f) Cats are healthier if given skim milk rather than full milk.

(g) People who take more exercise in their teens will suffer less from serious diseases in their fifties.

7.3 An experiment was conducted to assess the effect of education about disease. A large selection of volunteers was split into two groups of equal size. The splitting was done carefully, in such a way the two groups were well matched in terms of proportion of males, age distribution and so on. A coin was tossed (in front of an independent witness!) to decide which group should receive the education. The chosen group was then split into subgroups of twenty people. Each subgroup had a one-hour session with a lecture about the disease followed by questions and discussion.

Subsequent incidence of the disease among the entire collection of volunteers was monitored.

How should this experiment have been designed and randomized? How should it have been analysed?