
Chapter 1

Forward look

1.1 Stages in a statistically designed experiment

There are several stages in designing an experiment and carrying it out.

1.1.1 Consultation

The scientist, or other investigator, comes to the statistician to ask advice on the design of the experiment. Sometimes an appointment is made; sometimes the approach is by telephone or email with the expectation of an instant answer. A fortunate statistician will already have a good working relationship with the scientist. In some cases the scientist and statistician will both view their joint work as a collaboration.

Ideally the consultation happens in plenty of time before the experiment. The statistician will have to ask questions to find out about the experiment, and the answers may not be immediately available. Then the statistician needs time to think, and to compare different possible designs. In complicated cases the statistician may need to consult other statisticians more specialized in some aspect of design.

Unfortunately, the statistician is sometimes consulted only the day before the experiment starts. What should you do then? If it is obvious that the scientist has contacted you just so that he can write 'Yes' on a form in response to the question 'Have you consulted a statistician?' then he is not worth spending time on. More commonly the scientist genuinely has no idea that statistical design takes time. In that case, ask enough questions to find out the main features of the experiment, and give a simple design that seems to answer the purpose. Impress on the scientist that this design may not be the best possible, and that you can do better if given more notice. Try to find out more about this sort of experiment so that you are better prepared the next time that this person, or one of her colleagues, comes to you.

Usually the scientist does not come with statistically precise requirements. You have to elucidate this information by careful questioning. About 90% of the statistician's input at this stage is asking questions. These have to be phrased in terms that a non-statistician can understand. Equally, you must not be shy about asking the scientist to explain technical terms from his field if they seem relevant.

If the scientist does have a preconceived idea of a 'design', it may be chosen from an artificially short list, based on lack of knowledge of what is available. Too many books and

courses give a list of three or four designs and manage to suggest that there are no others. Your job may be to persuade the scientist that a better design is available, even if it did not figure in the textbook from which she learnt statistics.

Example 1.1 (Ladybirds) A famous company (which I shall not name) had designed an experiment to compare a new pesticide which they had developed, a standard pesticide, and ‘no treatment’. They wanted to convince the regulatory authority (the Ministry of Agriculture, Fisheries and Foods) that their new pesticide was effective but did not harm ladybirds. I investigated the data from the experiment, and noticed that they had divided a field into three areas, applied one pesticide (or nothing) to each area, and made measurements on three samples from each area. I asked the people who had designed it what the design was. They said that it was completely randomized (see Chapter 2). I said that I could see that it was not completely randomized, because all the samples for each pesticide came from the same area of the field. They replied that it must be completely randomized because there were no blocks (see Chapter 4) and it was not a Latin square (see Chapter 6). In defence of their argument they quoted a respectable textbook which gives only these three designs.

1.1.2 Statistical design

Most of this book is about statistical design. The only purpose in mentioning it here is to show how it fits into the process of experimentation.

1.1.3 Data collection

In collaboration with the scientist, design a form for collecting the data. This should either be on squared paper, with squares large enough to write on conveniently, or use the modern electronic equivalent, a spreadsheet or a hand-held data-logger. There should be a row for each observational unit (see Section 1.4) and a column for each variable that is to be recorded. It is better if these variables are decided before the experiment is started, but always leave space to include extra information whose relevance is not known until later.

Emphasize to the scientist that all relevant data should be recorded as soon as possible. They should never be copied into a ‘neater’ format; human beings almost always make errors when copying data. Nor should they be invented later.

Example 1.2 (Calf feeding) In a calf-feeding trial each calf was weighed several times, once at birth and thereafter on the nearest Tuesday to certain anniversaries, such as the nearest Tuesday to its eight-week birthday. The data included all these dates, which proved to be mutually inconsistent: some were not Tuesdays and some were the wrong length of time apart. When I queried this I was told that only the birthdate was reliable: all the other dates had been written down at the end of the experiment by a temporary worker who was doing her best to follow the ‘nearest Tuesday’ rule after the event. This labour was utterly pointless. If the dates had been recorded when the calves were weighed they would have provided evidence of how closely the ‘nearest Tuesday’ rule had been followed; deducing the dates after the event could more accurately and simply have been done by the computer as part of the data analysis.

Sometimes a scientist wants to take the data from his field notebooks and reorganize them into a more logical order for the statistician’s benefit. Discourage this practice. Not only does

1.1. Stages in a statistically designed experiment

3

Plot 8	6	Plot 23	0
	0		0
	7		0
	3		0
	6		0
	0		0
	4		0
	5		28
	6		0
	4		0
Average	<u>4.1</u>	Average	<u>28</u>

Fig. 1.1. Data sheets with intermediate calculations in Example 1.3

it introduce copying errors; reordering the data loses valuable information such as which plots were next to each other or what was the time sequence in which measurements were made: see Example 1.5.

For similar reasons, encourage the scientist to present you with the raw data, without making intermediate calculations. The data will be going into a computer in any case, so intermediate calculations do not produce any savings and may well produce errors. The only benefit brought by intermediate calculations is a rough check that certain numbers are the correct order of magnitude.

Example 1.3 (Leafstripe) In an experiment on leafstripe disease in barley, one measurement was apparently the percentage of disease on each plot. A preliminary graph of the data showed one outlier far away from the rest of the data. I asked to see the data for the outlying plot, and was given a collection of pieces of paper like those shown in Figure 1.1. It transpired that the agronomist had taken a random sample of ten quadrats in each plot, had inspected 100 tillers (sideshoots) in each quadrat to see how many were infected, and averaged the ten numbers. Only the average was recorded in the ‘official’ data. For the outlying plot the agronomist rightly thought that he did not need a calculator to add nine zeros to one nonzero number, but he did forget to divide the total by 10. Once I had corrected the average value for this plot, it fell into line with the rest of the data.

Also try to persuade the scientist that data collection is too important to be delegated to junior staff, especially temporary ones. An experiment cannot be better than its data, but a surprising number of good scientists will put much effort into their science while believing that the data can take care of themselves. Unless they really feel part of the team, junior or temporary staff simply do not have the same motivation to record the data carefully, even if they are conscientious. See also Example 1.2.

1.1.4 Data scrutiny

After the experiment is done, the data sheets or data files should be sent to the statistician for analysis. Look over these as soon as possible for obvious anomalies, outliers or evidence of bad practice. Can that number really be a calf’s birthweight? Experienced statisticians

become remarkably good at ‘data sniffing’—looking over a sheet of figures and finding the one or two anomalies. That is how the errors in Example 1.2 were found. Simple tables and graphs can also show up errors: in Example 1.3 the outlier was revealed by a graph of yield in tonnes per hectare against percentage of infected tillers.

Examine the final digits in the data. If the number of significant figures changes at one point, this may indicate a change in the person recording the data or the machine being used. Occasionally it indicates a change such as from weighing in pounds to weighing in kilograms and dividing by 2.205. Any such change is likely to coincide with a change in conditions which is more serious than the appearance of the data. These checks are easier to conduct on paper data than on electronic data, because most spreadsheets give no facility for distinguishing between 29 and 29.00.

Example 1.4 (Kiwi fruit) At an agricultural research station in New Zealand, an instrument called a penetrometer was used to measure the hardness of kiwi fruit. After a preliminary analysis of the data, the statistician examined a graph of residuals and realized that there was something wrong with the data. He looked again at the data sheet, and noticed that two different handwritings had been used. He re-analysed the data, letting the data in one handwriting be an unknown constant multiple of those in the other. The fitted value of the constant was 2.2, indicating that one person had recorded in pounds, the other in kilograms.

Query dubious data while it is still fresh in the scientist’s memory. That way there is a chance that either the data can be corrected or other explanatory information recorded.

Example 1.5 (Rain at harvest) In an experiment whose response was the yield of wheat on each plot, the numbers recorded on the last 12 plots out of a total of 72 were noticeably lower than the others. I asked if there was any reason for this, and was told that it had started to rain during the harvest, with the rain starting when the harvester was about 12 plots from the end. We were therefore able to include an extra variable ‘rain’, whose values were 60 zeros followed by 1, 2, ..., 12. Including ‘rain’ as a covariate in the analysis removed a lot of otherwise unexplained variation.

Example 1.6 (Eucalypts) In a forestry progeny trial in Asia, different families of eucalypts were grown in five-tree plots. After 36 months, a forestry worker measured the diameter of each tree at breast height. In the preliminary examination of the data, the statistician calculated the variance of the five responses in each plot, and found that every plot had exactly the same variance! Challenged on this, the forestry worker admitted that he had measured every tree in the first plot, but thereafter measured just tree 1 in each plot. For trees 2–5 he had added the constant c to the measurements from plot 1, where c was the difference between the diameter at breast height of tree 1 in this plot and the diameter at breast height of tree 1 in plot 1.

In this case, the statistician’s preliminary scrutiny showed that the data were largely bogus.

1.1.5 Analysis

This means calculations with the data. It should be planned at the design stage, because you cannot decide if a design is good until you know how the data will be analysed. Also, this planning enables the experimenter to be sure that she is collecting the relevant data. If necessary, the analysis may be modified in the light of unforeseen circumstances: see Example 1.5.

1.2. The ideal and the reality

5

For a simple design the statistician should, in principle, be able to analyse the data by hand, with a calculator. In practice, it is more sensible to use a reliable statistical computing package. A good package should ask the user to distinguish plot structure from treatment structure, as in Section 1.4; it should be able to handle all the structures given in Section 1.4; and it should automatically calculate the correct variance ratios for experiments like those in Chapters 8 and 10. It is a good idea to do the planned analysis on dummy data *before* the real data arrive, to avoid any unnecessary delay.

Many other statistics books are concerned almost exclusively with analysis. In this book we cover only enough of it to help with the process of designing experiments.

1.1.6 Interpretation

The data analysis will produce such things as analysis-of-variance tables, lists of means and standard errors, P -values and so on. None of these may mean very much to the scientist. It is the statistician's job to interpret the results of the analysis in terms which the scientist can understand, and which are pertinent to his original question.

1.2 The ideal and the reality

Here I discuss a few of the tensions between what the statistician thinks is desirable and what the experimenter wants.

1.2.1 Purpose of the experiment

Why is the experiment being done? If the answer is 'to use an empty greenhouse' or 'to publish another paper', do not put much statistical effort into it. A more legitimate answer is 'to find out about the differences between so-and-so', but even this is too vague for the statistician to be really helpful.

Ideally, the aims of the experiment should be phrased in terms of specific questions. The aim may be to estimate something: for example, 'How much better is Drug *A* than Drug *B*?' This question needs refining: how much of each drug? how administered? to whom? and how will 'better' be measured? For estimation questions we should aim to obtain unbiased estimators with low variance.

On the other hand, the aim may be to test a hypothesis, for example that there is no effective difference between organic and inorganic sources of nitrogen fertilizer. Again the question needs refining: how much fertilizer? applied to what crop? in what sorts of circumstances? is the effect on the yield or the taste or the colour? For hypothesis testing we want high power of detecting differences that are big enough to matter in the science involved.

1.2.2 Replication

This is the word for the number of times that each treatment is tested.

The well-known formula for the variance of the mean of n numbers is σ^2/n , on the assumption that the numbers are a random sample from a population with variance σ^2 . Increasing the replication usually decreases the variance, because it increases the value of n .

On the other hand, increased replication may raise the variance. Typically, a larger number

of experimental units are more variable than a small number, so increasing the replication may increase the value of σ^2 . Sometimes this increase outweighs the increase in n .

Increased replication usually raises power. This is because it usually raises the number of residual degrees of freedom, and certain important families of distribution (such as t) have slimmer tails when they have more degrees of freedom.

The one thing that is almost certain about increased replication is that it increases costs, which the experimenter usually wants to keep down.

1.2.3 Local control

This means dividing the set of experimental units into blocks of alike units: see Chapter 4. It is also called *blocking*.

If it is done well, blocking lowers the variance, by removing some sources of variability from treatment contrasts. If each block is representative rather than homogeneous then blocking has the opposite effect.

Blocking can increase the variance if it forces the design to be non-orthogonal: see Chapter 11.

Because blocking almost always decreases the variance, it usually raises power. However, it decreases the number of residual degrees of freedom, so it can reduce power if numbers are small: see Example 4.15.

Blocking increases the complexity of the design. In turn this not only increases the complexity of the analysis and interpretation but gives more scope for mistakes in procedure during the experiment.

1.2.4 Constraints

The most obvious constraint is cost. Everybody will be pleased if the same results can be achieved for less money. If you can design a smaller, cheaper experiment than the scientist proposes, this is fine if it produces good estimators. On the other hand, it may be impossible to draw clear conclusions from an experiment that is too small, so then the entire cost is wasted. Part of your duty is to warn when you believe that the whole experiment will be wasted.

The availability of the test materials may provide a constraint. For example, in testing new varieties of wheat there may be limited quantities of seed of some or all of the new varieties.

Availability of the experimental units provides a different sort of constraint. There may be competition with other experimenters to use land or bench space. If results are needed by a certain deadline then time limits the number of experimental units. In a clinical trial it is unethical to use far too many patients because this unnecessarily increases the number of patients who do not get the best treatment. On the other hand, it is also unethical to use so few patients that no clear conclusions can be drawn, for then all the patients have been used in vain. Similar remarks apply to experiments on animals in which the animals have to be sacrificed.

If there are natural 'blocks' or divisions among the experimental units these may force constraints on the way that the experiment can be carried out. For example, it may be impossible to have all vaccinations administered by the same nurse.

There are often other constraints imposed by the management of the experiment. For

1.3. An example

7

example, temporary apple-pickers like to work with their friends: it may be unrealistic to expect them each to pick from separate rows of trees.

1.2.5 Choice

Given all the constraints, there are still two fundamentally important choices that have to be made and where the statistician can provide advice.

Which treatments are to be tested? The scientist usually has a clear idea, but questions can still be helpful. Why did he decide on these particular quantities? Why these combinations and not others? Should he consider changing two factors at a time? (see Chapter 5). Does the inclusion of less interesting treatments (such as the boss's favourite) mean that the replication for *all* treatments will be too low?

There is a strong belief in scientific circles that all new treatments should be compared with 'no treatment', which is often called *control*. You should always ask if a control is needed. Scientific orthodoxy says yes, but there are experiments where a control can be harmful. If there is already an effective therapy for a disease then it is unethical to run an experiment comparing a new therapy to 'do nothing'; in this case the treatments should be the new therapy and the one currently in use. In a trial of several pesticides in one field, if there is a 'do nothing' treatment on some plots then the pest may multiply on those plots and then spread to the others. A 'do nothing' treatment is also not useful if this would never be used in practice.

Sometimes it is already known that the 'do nothing' treatment has a very different effect from all the other treatments. Then the experiment may do nothing more than confirm this, as in Examples 3.2 and 6.3. In such cases, it is better to omit the 'do nothing' treatment so that more resources can be devoted to finding out whether there is any difference between the other treatments.

Which experimental units should be used? For example, is it better to use portions of representative farmers' fields or a well-controlled experimental farm? The latter is better if the effect to be detected is likely to be small, or if one of the treatments is sufficiently unknown that it might have disastrous economic or environmental consequences. The former is better for a large confirmatory experiment, before recommending varieties or treatments for use on a wide scale. Similarly, is it better to use 36 heifers from the same herd or 36 bought at the market specifically for this experiment? University students are a convenient source of experimental units for psychologists, but how far can results valid for such students be extrapolated to the general population?

1.3 An example

An example will help to fix ideas.

Example 1.7 (Rye-grass) An experiment was conducted to compare three different cultivars of rye-grass in combination with four quantities of nitrogen fertilizer. Two responses were measured: one was the total weight of dry matter harvested from each plot, and the other was the percentage of water-soluble carbohydrate in the crop.

The three cultivars of rye-grass were called Cropper, Melle and Melba. The four amounts

0	160	240
160	80	80
80	0	160
240	240	0
↑	↑	↑
Cropper	Melba	Melle

160	80	0
0	160	80
240	0	240
80	240	160
↑	↑	↑
Melba	Cropper	Melle

Fig. 1.2. Layout of the rye-grass experiment in Example 1.7

of fertilizer were 0 kg/ha, 80 kg/ha, 160 kg/ha and 240 kg/ha.

The experimental area consisted of two fields, each divided into three strips of land. Each strip consisted of four plots.

Cultivars were sown on whole strips because it is not practicable to sow them in small areas unless sowing is done by hand. In contrast, it is perfectly feasible to apply fertilizers to smaller areas of land, such as the plots. The layout for the experiment is shown in Figure 1.2.

Notice the pattern. Each amount of nitrogen is applied to one plot per strip, and each cultivar is applied to one strip per field. This pattern is the *combinatorial design*.

Notice the lack of pattern. There is no systematic order in the allocation of cultivars to strips in each field, nor any systematic order in the allocation of amounts of nitrogen to plots in each strip. This lack of pattern is the *randomization*.

1.4 Defining terms

Definition An *experimental unit* is the smallest unit to which a treatment can be applied.

Definition A *treatment* is the entire description of what can be applied to an experimental unit.

Although the previous two definitions appear to be circular, they work well enough in practice.

Definition An *observational unit* is the smallest unit on which a response will be measured.

Example 1.6 revisited (Eucalypts) The experimental units were the plots. The observational units should have been the trees.

Example 1.8 (Wheat varieties) The experiment compares different varieties of wheat grown in plots in a field. Here the experimental units are the plots and the treatments are the varieties. We cannot tell what the observational unit is without more information. Probably a plot is the observational unit, but it might be an individual plant. It might even be the whole field.

1.4. Defining terms

9

	whole class	small groups
one hour once per week	✓	✓
12 minutes every day	✓	✓

Fig. 1.3. Four treatments in Example 1.10

Example 1.7 revisited (Rye-grass) Here the treatments are the combinations of cultivars with amounts of fertilizer, so there are twelve treatments. The experimental unit is the plot. The observational unit is probably the plot but might be a plant or a strip.

Example 1.2 revisited (Calf feeding) Here the treatments were different compositions of feed for calves. The calves were not fed individually. They were housed in pens, with ten calves per pen. Each pen was allocated to a certain type of feed. Batches of this type of feed were put into the pen; calves were free to eat as much of this as they liked. Calves were weighed individually.

The experimental units were the pens but the observational units were the calves.

Example 1.9 (Asthma) Several patients take part in an experiment to compare drugs intended to alleviate the symptoms of chronic asthma. For each patient, the drugs are changed each month. From time to time each patient comes into the clinic, where the peak flow rate in their lungs is measured.

Here the treatments are the drugs. An experimental unit is a patient-month combination, so if 30 patients are used for 6 months then there are 180 experimental units. The observational unit is a visit by a patient to the clinic; we do not know how this relates to the patient-months without further information.

Example 1.10 (Mental arithmetic) After calculators became widespread, there was concern that children in primary schools were no longer becoming proficient in mental arithmetic. One suggested remedy was whole-class sessions, where the teacher would call out a question such as ‘5 + 7?’ and children would put up their hands to offer to give the correct answer. An alternative suggestion was to do this in small groups of about four children, to encourage those who were shy of responding in front of the whole class. Another question was: is it better to have these sessions for one hour once a week or for 10–12 minutes every day?

The treatments are the four combinations of group size and timing shown in Figure 1.3. Each treatment can be applied only to a whole class, so the experimental units are classes. However, to measure the effectiveness of the treatments, each child must take an individual test of mental arithmetic after some set time. Thus the observational units are the children.

Example 1.11 (Detergents) In a consumer experiment, ten housewives test new detergents. Each housewife tests one detergent per washload for each of four washloads. She assesses the cleanliness of each washload on a given 5-point scale. Here the 40 washloads are the experimental units and the observational units; the detergents are the treatments.

Example 1.12 (Tomatoes) Different varieties of tomato are grown in pots, with different composts and different amounts of water. Each plant is supported on a vertical stick until it is 1.5 metres high, then all further new growth is wound around a horizontal rail. Groups

of five adjacent plants are wound around the same rail. When the tomatoes are ripe they are harvested and the weight of saleable tomatoes per rail is recorded.

Now the treatment is the variety–compost–water combination. The pots are the experimental units but the rails are the observational units.

These examples show that there are four possible relationships between experimental units and observational units.

- (i) The experimental units and the observational units are the same. This is the most usual situation. It occurs in Example 1.11; in Examples 1.7 and 1.8 if there is one measurement per plot; in Example 1.9 if there is one measurement of peak flow rate in lungs per patient per month.
- (ii) Each experimental unit consists of several observational units. This is usually forced by practical considerations, as in Examples 1.2 and 1.10. Examples 1.7 and 1.8 are of this type if the observational unit is a plant. So is Example 1.9 if the observational unit is a patient-week. This situation is fine so long as the data are analysed properly: see Chapter 8.
- (iii) Each observational unit consists of several experimental units. This would occur in Example 1.9 if each patient had their drugs changed monthly but their peak flow rate measured only every three months. It would also occur in Examples 1.7 and 1.8 if the observational unit were the strip or field respectively. In these cases the measurements cannot be linked to individual treatments so there is no point in conducting such an experiment.

Example 1.12 also appears to be of this form. Because the experiment would be useless if different pots in the same group (allocated to the same rail) had different treatments, in effect it is the group of pots that is the experimental unit, not the individual pot.

In fact, there are some unusual experiments where the response on the observational unit can be considered to be the sum of the (unknown) responses on the experimental units contained within it. However, these are beyond the scope of this book.

- (iv) Experimental units and observational units have a partial overlap, but neither is contained in the other. This case is even sillier than the preceding one.

It is useful to write down the experimental units and the observational units in the experimental protocol. This should draw attention to cases (iii) and (iv) before it is too late to change the protocol.

Definition In cases (i) and (ii) an observational unit will often be called a *plot* for brevity.

This usage is justified by the large amount of early work on experimental design that took place in agricultural research. However, it can be a little disconcerting if the plot is actually a person or half a leaf. It is a useful shorthand in this book, but is not recommended for your conversations with scientists.

Notation In this book, general plots are denoted by lower-case Greek letters, such as α , β , γ , ω . The whole set of plots is denoted by Ω , and the number of plots by N .