CAMBRIDGE

Cambridge University Press
978-0-521-86214-1 - Statistical Principles for the Design of Experiments
R. Mead, S. G. Gilmour and A. Mead
Excerpt
More information

Part I

Overture

Cambridge University Press
978-0-521-86214-1 - Statistical Principles for the Design of Experiments
R. Mead, S. G. Gilmour and A. Mead
Excerpt
More information

1

Introduction

1.1 Why a statistical theory of design?

The need to develop statistical theory for designing experiments stems, like the need for statistical analysis of numerical information, from the inherent variability of experimental results. In the physical sciences, this variability is frequently small and, when thinking of experiments at school in physics and chemistry, it is usual to think of 'the correct result' from an experiment. However, practical experience of such experiments makes it obvious that the results are, to a limited extent, variable, this variation arising as much from the complexities of the measurement procedure as from the inherent variability of experimental material. As the complexity of the experiment increases, and the differences of interest become relatively smaller, then the precision of the experiment becomes more important. An important area of experimentation within the physical sciences, where precision of results and hence the statistical design of experiments is important, is the optimisation and control of industrial chemical processes.

Whereas the physical sciences are thought of as exact, it is quite obvious that biological sciences are not. Most experiments on plants or animals use many plants or animals because it is clear that the variation between plants, or between animals, is very large. It is impossible, for example, to predict quantitatively the exact characteristics of one plant from the corresponding characteristics of another plant of the same species, age and origin.

Thus, no medical research worker would make confident claims for the efficacy of a new drug merely because a single patient responded well to the drug. In the field of market research, no newspaper would publish an opinion poll based on interviews with only two people, but would require a sample of at least 500, together with information about the method of selection of the sample. In a drug trial, the sample of patients would often be quite small, possibly between 20 and 100, but for a final trial before the release of the drug, as many as 2000–3000 patients might be used to detect unexpected side effects of the drug. In psychological experiments, the number of subjects used might be only 8–12. In agricultural experiments, there may be 20–100 plots of land, each with a crop grown on it. In a laboratory experiment hundreds of plants may be treated and examined individually. Or just six cows may be examined while undergoing various diets, with measurements taken frequently and in great detail.

The size of an experiment will vary according to the type of experimental method and the objective of the experiment. One of the important statistical ideas of experimental design is the choice of the size of an experiment. Another is the control of the use of experimental material. It is of little value to use large numbers of patients in the comparison of two drugs,

Cambridge University Press
978-0-521-86214-1 - Statistical Principles for the Design of Experiments
R. Mead, S. G. Gilmour and A. Mead
Excerpt
More information

if all the patients given one drug are male, aged between 20 and 30, and all the patients given the other drug are female, aged 50–65. Any reasonably sceptical person would doubt claims made about the relative merits of the two drugs from such a trial. This example may seem trivially obvious, but the scientific literature in medicine and many other disciplines shows that many examples of badly planned (or unplanned) experiments occur.

And this is just the beginning of statistical design theory. From avoiding foolish experiments, we can go on to plan improvements in precision for experiments. We can consider the choice of experiments as part of research strategy and can, for example, discuss the relative merits of many small experiments or a few large experiments. We can consider how to design experiments when our experimental material is generally heterogeneous, but includes groups of similar experimental units. Thus, if we are considering the effects of applying different chemicals on the properties of different geological materials, then these may be influenced by the environment from which they are taken, as well as by the chemical treatment applied. However, we may have only two or three samples from some environments, but as many as ten samples from other environments; how then do we decide which chemicals to apply to different samples so that we can compare six different chemical treatments?

1.2 History, computers and mathematics

If we consider the history of statistical experimental design, then most of the developments have been in biological disciplines, in particular in agriculture, and also in medicine and psychology. There is therefore an inevitable agricultural bias to any discussion of experimental design. Many of the important principles of experimental design were developed in the 1920s and 1930s, in particular by R. A. Fisher. The practical manifestation of these principles was very much influenced by the calculating capacity then available. Had the computational facilities which we now enjoy been available when the main theory of experimental design was being developed it is quite possible that the whole subject of design would have developed very differently. Whether or not this belief is valid, it is certainly true that a view of experimental design today must differ from that of the 1930s, or even the 1970s. The principles have not changed, but the principles are often forgotten, and only the practical manifestation of the principles retained; these practical applications do require rethinking.

The influence of the computer is one stimulus to reassessing experimental design. Another cause for concern in the development of experimental design is the tendency for increasingly formal mathematical ideas to supplant the statistical ideas. Thus the fact that a particularly elegant piece of mathematics can be used to demonstrate the existence of groups of designs, allocating treatments to blocks of units in a particular way, begs the statistical question of whether such designs would ever be practically useful.

Although our backgrounds have been in mathematics, we believe that the presentation of statistical design theory has tended to be quite unnecessarily mathematical, and we will hope to demonstrate the important ideas of statistical design without excessive mathematical encumbrance. The language of statistical theory, like that of physics, is mathematical and there will be sections of the book where those with a mathematical education beyond school level will find a use for their mathematical expertise. However, even in these sections, which we believe should be included because they will improve the understanding of statistical theory of those readers able to appreciate the mathematical demonstrations, there are intuitive explanations of the theory at a less advanced mathematical level.

1.3 The influence of analysis on design

To write a book solely about the theory of experimental design, excluding all mention of the analysis of data, would be impossible. Any experimenter must know how he or she intends to analyse the experimental data before the experiment to yield the data is designed. If not, how can the experimenter know whether the form of information which is to be collected can be used to answer the questions which prompted the experiment? And just as crucially, whether the amount of information from the experiment is not only sufficient but is not excessive.

Thus, consider again the medical trial to compare two drugs. Suppose the experimenter failed to think about the analysis and argued that one of the drugs was well known, while the other was not; in the controlled experiment to compare them, there are available 40 patients. Since a lot is already known about drug A, let it be given to one patient, and let drug B be given to the other 39. When the data on the response to the drugs are obtained, the natural analysis is to compare the mean responses to the drugs, and to consider the difference $(\bar{y}_A - \bar{y}_B)$. To test the strength of evidence for a real difference in the effects of the two drugs, we need the standard error of $(\bar{y}_A - \bar{y}_B)$, which will be

$$\sigma(1/1 + 1/39)^{1/2} = 1.013\sigma.$$

However, if the experimenter had considered this analysis of data before designing the experiment, he would have realised that the effectiveness of his experiment depended on making the standard error of $(\bar{y}_A - \bar{y}_B)$ within his experiment as small as possible (to use the previous knowledge about the effects of drug A requires that the experimental conditions are identical to those in which the previous experience was gained). This is achieved by allocating 20 patients to drug A and 20 to drug B. This gives a standard error of

$$\sigma(1/20 + 1/20)^{1/2} = 0.316\sigma,$$

showing that equal allocation to the two groups reduces the standard error by a factor of more than 3. Of course, it would not be necessary to consider the standard error formally to guess that equal allocation of patients gives the most precise answer. In a sense, it would be intuitively surprising if any unequal division were to be more efficient than the equal division. Prior thought of what is to be done with the results of the experiment should lead to avoiding a foolish design.

Nevertheless, although analysis is integral to any consideration of design, this book is about the design of experiments, and will not be concerned with the analysis of data, except insofar as this is essential to the understanding of the design principles discussed. The general theory of the analysis of data from designed experiments, using the method of least squares estimation for linear models is developed in Chapter 4. Some particular examples of this general theory will appear in later chapters for some examples of experimental designs. There will also be some discussion, particularly in Chapters 9, 11 and 18, of other methods and models for analysis; these methods are related to, but different from, the general method described in Chapter 4.

We shall assume throughout the book that the analysis of data from experiments will be achieved through the use of computer programs. There is available a wide range of programs for the analysis of experimental data. Some of the programs apply the simplest forms of the general method which we describe in Chapter 4, but there are many other analysis programs available, and it is important that an experimenter should understand enough about different methods of analysis to be able to detect when the method used in a particular program is appropriate to his situation.

Thus, not only is it possible to derive the algebraic formulae necessary to define the calculations required for the analysis of data from any experiment, but the computer programs should make it possible for an experimenter, without the mathematical skills necessary to derive the algebraic formulae, to understand the statistical basis for the interpretation of the calculations. The need, in the earlier development of statistical methods, to be able to derive the algebraic form of analysis has made it easier to discuss some of the important design principles we shall discuss later. In this way, the earlier lack of advanced computing power may be regarded as a blessing, though there must be many erstwhile technical assistants to statisticians who would find that hard to accept! A caveat to the advice to use the general computer programs to analyse all experimental data must be to remember that much of our understanding of design theory was stimulated by algebraic necessity because of the lack of computers. Consequently, we should search for increased understanding both through the use of computers and through algebraic manipulation.

1.4 Separate consideration of units and treatments

Throughout this book, we shall emphasise the need to think separately about the properties of the experimental units available for the experiment, and about the choice of experimental treatments. Far too often, the two aspects are confused at an early stage, and an inefficient or useless experiment results. Thus, an experimenter considering a comparison of the effects of different diets on the growth of rats may observe that the rats available for experimentation come from litters with between five and ten animals per litter. Having heard about the randomised block design, in which each treatment must occur in each block, he decides that he must use litters as blocks, and therefore will have blocks of five units. He further decides that he should consequently have five treatments, and chooses his treatments on this premise, rather than considering how many treatments are required for the objectives of his experiment.

Similarly, a microbiologist investigating the growth of salmonella as affected by five different temperatures, four different media and two independent different chemical additives, may decide that a factorial set of 80 treatments is necessary. Recognising the possibility of day-to-day variation and also knowing about the randomised block design he tries to squeeze the preparation of 80 units into a single day, where only 40 can be efficiently managed, and consequently suffers a much higher error variance for his observations than would normally be expected.

In both of these examples, which though plausible are fictitious, we believe that separate consideration of properties of units and choice of treatments, together with an appreciation of the possibilities of modern statistical design, would lead to better experiments. Thus, in the rat growth experiment, on considering the experimental material available, the experimenter might recognise that each litter forms a natural block, and that he has one block of five units, one of six units, two of seven units, and one of ten units; each block could be reduced in size while still retaining its natural block quality. For treatments, he has a control and three additions to the basic diet, each of which he wishes to try at two concentrations; a total of seven treatments. It is quite simple to construct a sensible design to compare seven treatments in blocks of five, six, seven, seven and ten units, respectively, and this problem will be discussed further in Chapter 7.

For the microbiologist, again the solution of his problem requires only the application of standard statistical design theory, though because the problem does not fit into any neat mathematical classification, the reader of most books on experimental design might be forgiven for thinking that the problem was impossible or at best required very complex mathematical theory. The problem of using blocks of 40 units with a set of 80 treatments in a $5 \times 4 \times 2 \times 2$ factorial structure is a very simple example of confounding to be met again in Chapter 15.

In summary, to achieve good experimental design, the experimenter should think first about the experimental units, and should make sure that he understands the likely patterns of variation. He should also consider the set of treatments appropriate to the objectives of his experiment. If the set of treatments fits simply with the structure of units using a standard design, then the experiment is already designed. If not, then a design should be constructed to fit the requirements of units and treatments, either by the experimenter alone or with the assistance of a statistician. The natural pattern of units and treatments should not be deformed in a Procrustean fashion to fit a standard design. It may perhaps reassure the reader to point out that, for most experiments (80 or 90%), a standard design will be appropriate. To counterbalance the note of reassurance, it is obvious from the experimental scientific literature that frequently an inappropriate simple standard design has been used when an appropriate but slightly more complex standard design exists and should have been used.

1.5 The resource equation

A major principle of design arises from the consideration of degrees of freedom (df) for treatment comparisons. These will be defined more formally in Chapter 2, but in essence df count the number of independent comparisons which can be made. The total information in an experiment involving N experimental units may be represented by the total variation based on $(N - 1)$ df. In the general experimental situation, this total variation is divided into three components, each serving a different function:

(a) the treatment component, T, being the sum of df corresponding to the questions to be asked;

(b) the blocking component, B, representing all the environment effects that are to be included in the fitted model, usually $(b - 1)$ df for the b blocks, but other alternatives are $(r - 1) + (c - 1)$ for row and column designs, or more generally, all effects in multiple control or multiple covariate designs. It may also be desirable to allow a few df for unforeseen contingencies such as missing observations;

(c) the error component, E, being used to estimate σ^2 for the purpose of calculating standard errors for treatment comparison, and, more generally, for inference about treatments.

We can express this division of information as *the resource equation* in terms of the df used in each component T, B and E:

$$T + B + E = N - 1.$$

Now, to obtain a good estimate of error, it is necessary to have at least 10 df. Many statisticians would take 12 or 15 df as their preferred lower limit, although there are situations where statisticians have yielded to pressure from experimenters and agreed to a design with only 8 or even 6 df. The basis for the decision is most simply seen by examining the 5% point of the t-distribution, and using sufficient df so that increasing the error df makes very little difference to the significance point, and hence to the interpretation.

To return to the implications for design, it is necessary to have at least 10 df for error, to estimate σ^2. In choosing the number of treatment question df, we should not normally allow E to fall below 10. But equally, if E is allowed to be large, say greater than 20, then the experimenter is wasting resources by using too much resource on estimating s^2, and not asking enough questions. This is a failure of experiments which occurs much more frequently than allowing E to fall too low, and it is just as much a failure of design. Often, the way of increasing the number of treatment question df is through adding an additional treatment factor, so that the previously planned treatments are all applied at each of two levels of some other factor. The additional factor may be an additional interference treatment, or it could be a treatment factor representing different environments. Various approaches to the choice of treatment factors will be discussed in Chapter 12 and subsequent chapters. The basic economic advantage of factorial experiments arises from asking many questions simultaneously and to insist, as many scientists do, on asking only one or two questions in each experiment, is to waste resources on a grand scale.

One final comment on this question of efficient use of resources. It is, of course, possible to keep the value of E in the region of 10–20 by doing several small experiments, rather than one rather larger experiment in which the total number of experimental units is much less than the sum of the units in the separate small experiments. This is inefficient because of the need to estimate σ^2 for each separate experiment. Experimenters should identify clearly the questions they wish the experiment to cover, and they should also consider carefully if they are asking enough questions to use the experimental resources efficiently.

Cambridge University Press
978-0-521-86214-1 - Statistical Principles for the Design of Experiments
R. Mead, S. G. Gilmour and A. Mead
Excerpt
More information

2

Elementary ideas of blocking: the randomised complete block design

2.1 Controlling variation between experimental units

In an experiment to compare different treatments, each treatment must be applied to several different units. This is because the responses from different units vary, even if the units are treated identically. If each treatment is only applied to a single unit, the results will be ambiguous – we will not be able to distinguish whether a difference in the responses from two units is caused by the different treatments, or is simply due to the inherent differences between the units.

The simplest experimental design to compare t treatments is that in which treatment 1 is applied to n_1 units, treatment 2 to n_2 units, . . . and treatment t to n_t units. In many experiments the numbers of units per treatment, $n_1, n_2, \ldots, n_t$, will be equal, but this is not necessary, or even always desirable. Some treatments may be of greater importance than others, in which case more information will be needed about them, and this will be achieved by increasing the replication for these treatments. This design, in which the only recognisable difference between units is the treatments which are applied to those units, is called a *completely randomised design*.

However, the ambiguity, when each treatment is only applied to a single unit, is not always removed by applying each treatment to multiple units. One treatment might be ‘lucky’ in the selection of units to which it is to be applied, and another might be ‘unlucky’. There is no foolproof method of overcoming the vagaries of allocating treatments to units, since the responses, if the same treatment were applied to all units, are not known before the experiment. However, most experimenters have some idea about which units are likely to behave similarly, and such ideas can be used to control the allocation of treatments to units in an attempt to make the allocation more ‘fair’. Essentially each group of ‘similar’ units should include roughly equal numbers of units for each treatment. This control is referred to as *blocking*, and the simplest and most frequently used design resulting from the idea of blocking is the randomised block design, more correctly called the *randomised complete block design*, in which each block contains a single replicate of each of the treatments.

The term blocking arises from the introduction of a controlled treatment allocation in agricultural crop experiments. Here the experimental unit is a small area of land, often called a plot, typically between 2 m^2 and 20 m^2, depending on the crop and on the treatments. The set of units or plots for an experiment might be arranged as in Figure 2.1(a). In general we might expect that adjacent plots would behave more similarly than those further apart. Of course, the experimenter might have more specific knowledge about the plots than this. For example, there might be a slope from left to right, so that the plots down the left hand side might be

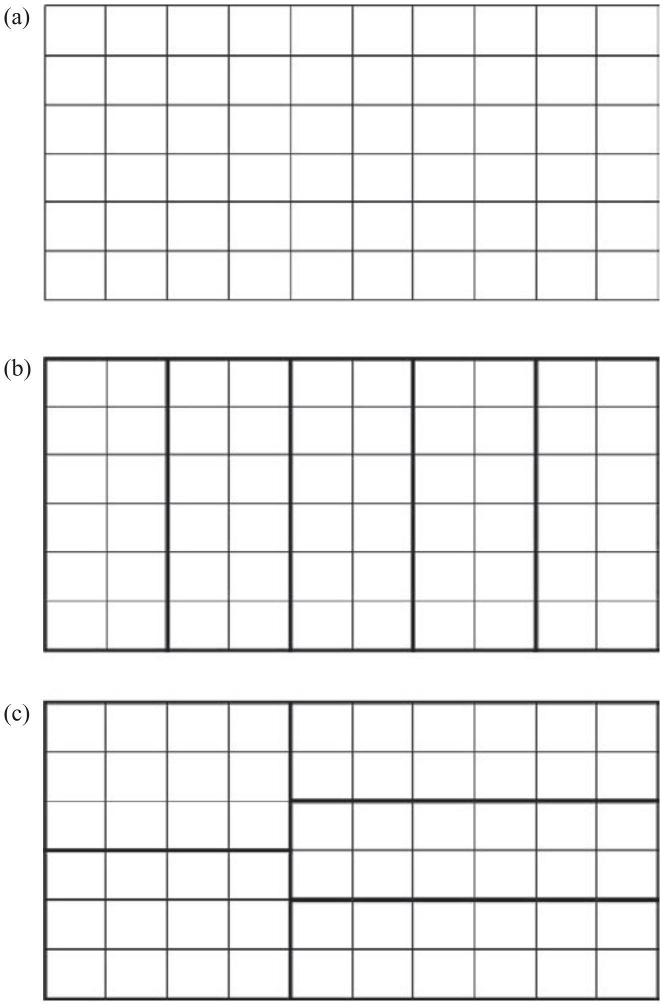

Figure 2.1 Possible blocking plans (b) and (c) for a set of 60 plots (a).

expected to behave similarly, and differently from those on the right hand side, with a trend in response from left to right. However, either from general or particular considerations, if the plots are grouped together in sets of, say, 12, so that the plots in a set might be expected to behave similarly, then the resulting grouping might look like Figure 2.1(b) for a trend from left to right or (c) for a trend from top to bottom. Such groupings are naturally called blocks of plots, and the term has become standard for groupings of units in disciplines other than agronomy.

The patterns of variation on which blocking systems can be based are very numerous. But the most important source of information in determining blocks of units must always be the specialist knowledge of the experimenter about his experimental material. Some examples of blocking systems are discussed here, to help the experimenter to identify the appropriate choice of blocks for their particular experiment.