4

Principles of Research Design

4.1 Desirable Properties of Research Design

We discussed in chapter 1 how a fundamental step in scientific inquiry is to ask, exactly, "What is the question?" Explicit formulation of the question is essential, because it determines what we do in the design of the study that is supposed to answer it. This might appear trivial, but much experience with student-designed studies shows that insufficient critical thought is given to (a) stating the question exactly and (b) designing the work explicitly to answer the question.

Just what is it that we need to do in science? Offhand, one might think that we will want to compare a "treatment" with an untreated "control" and that's all. It turns out that ensuring that the treatment and control differ in just the one aspect that tests the question, that the treatment is effective and unbiased, that the measurements we collect from treated and control units are precise and accurate, and that the results are widely applicable, as well as accessible to available methods and tests, is considerably more demanding. There is no "correct" experimental design or statistical analysis; both depend on the question being investigated. Once we really know our question, however, we can more effectively look for appropriate ways to answer the question.

Whether we propose to do a sampling survey for comparative studies, long-term monitoring, perturbation studies, or manipulative experiments, certain characteristics are desirable in the design of a research plan. These characteristics include[1]

1. good estimation of treatment effects,
2. good estimation of random variation,
3. absence of bias,
4. precision and accuracy,

1. I am tempted to add that the research question should be interesting. Many of us focus too narrowly; if we seek the underlying generalities, even when dealing with local, everyday questions, our work will be more interesting to more people, and the consequences of our results will reach farther. This matter of interest is important, but I did not add this idea to the list simply because "interesting" is such a value-laden concept that it seemed too subjective.

5. wide range of applicability, and
6. simplicity in execution and analysis

To incorporate these desirable characteristics, a variety of research design options are available. The options for design focus on three different parts of research studies: design of treatments (how the treatments relate to each other], *design* of layout (how the treatments are assigned to experimental units), and design of response (how to assure an appropriate response by the experimental units to the treatments). An excellent extended discussion of the topics of this chapter is given in Mead (1988).

4.2 Design of Treatments

The design of treatments merits more thought than it is often given, because the treatments define the way we pose the question and how we carry out the test. There are many ways to design treatments; this section details only a few key approaches.

As an example, I borrow an experiment discussed by Urquhart (1981). Water from different localities in arid regions often differs markedly in chemical composition, and unknown differences in chemical content could affect plant responses. The experiment therefore addressed the question of whether irrigation using water from different sources led to different growth of plants. Chrysanthemums were selected as the assay organism, and water was obtained from 24 different sites and included distilled water, tap water, brackish water, and water from sulfur springs. The mums were grown in 360 pots in a greenhouse. Pots were placed on 3 benches, 24 groups of 5 pots each on each bench. Each treatment (water source) was allocated at random to a group of 5 pots on each bench, with an additional random assignment for each bench. The experiment could be run with one plant per pot, or more than one. The dependent variable to be measured as the response to treatments was height of the plants after 7 weeks of growth

Some More Statistical Terms

Statisticians, as you have no doubt noticed, use certain everyday terms (normal, mean, significant, parameter, and error, among others) in specialized ways. Before we examine the design of treatments, layout, and response, we should review some other familiar terms that statisticians use with specialized meaning:

Experimental unit: element or amount of experimental material to which a treatment is applied.

Factor: set of treatments of a single type applied to experimental units.

Interaction: differences among levels of one factor within levels of another factor.

Level of a factor: particular treatment from a graded set of treatments that make up the factor.

Main effect: differences among levels of one factor, averaging levels of other factors.

Population: a well-defined set of items about which we seek inferences.

Treatment: distinctive feature, classification, or manipulation that defines or can be applied to experimental units.

Unstructured Treatment Designs

If the 24 water samples were merely a random sample of the different kinds of water available for irrigation, we could refer to the experiment as having an *unstructured random* treatment design. If we had been dealing with comparisons of defined fertilizer formulations on mum growth, we would have an *unstructured fixed* treatment design. The importance of the fixed or random status is that, as already noted, these models lead to slightly different methods of statistical analysis. Actually, unstructured designs are used less often than structured designs, because we more often select the treatments with more specific purposes.

Structured Treatment Designs

Factorial Treatments

If we thought that the relative concentration of some key chemical in the water was important, we could run an experiment in which we watered mums with 4 dilutions of original water samples. To make the experiment feasible, we would pick 6 out of the 24 sources of water; these treatments would yield a set of data that would be conveniently shown in a table with 6 rows for the sources, and 4 columns for the dilutions. We have already encountered this sort of design in chapter 3, when discussing ANOVA. Such an experiment is referred to as having a *factorial treatment* design. In this case statisticians call the treatments *factors*, for obscure historical reasons.

These experiments require much effort in execution (note that we reduced the number of water sources in our example to make it feasible) and in analysis (see Sokal and Rohlf 1995, chap. 12). Factorial treatment designs, however, provide the opportunity to closely examine the significance of dose effects of the factors and of interactions among the factors manipulated — powerful and desirable features.

Nested Treatments

If the 24 water samples were known to come from sites that could be classified into, say, 4 regions, then we would have set up a *nested* or *grouped random* treatment design, in which comparisons among groups would test regional differences. These designs have also been referred to as *hierarchical*, to highlight that one variable, water chemistry in our case, is grouped at a different (and lower) level than the other variable, region. The effects of waters of different chemistry are compared within each region; the effects of waters from different sites are compared, naturally enough, by among-site comparisons.

Nested designs in general are less desirable than the cross-classified designs discussed in section 3.1 (table 3.6). One reason for this is that interactions between the higher and the nested variable are not separable in nested designs. In our water chemistry experiment, for instance, interpretation is limited in that within each region we can compare water chemistry among only those sites that happen to be located in the geo-

graphical region; this may not be an entirely satisfactory analysis, because it may well be, for example, that hard waters predominate in one region but not in another.

Nesting can occur at different levels. The regions might be one level of nesting. We could nest at a lower level if we needed to know within-pot variation. To do this lower level nesting, we would grow four plants in each pot, instead of one plant. This design would provide information as to how variable is the growth of plants within pots subjected to each of the treatments on each bench.

In nested treatment designs, the highest level of classification can be random or fixed, but the nested level of classification is usually random. For example, in the lower level of nesting, the four plants would be selected at random before planting.

Nested designs are usually the result of a shortage of subjects or some other limitation on experimental units. For example, suppose we were zookeepers concerned with keeping our rare New Guinean cassowaries free of lice. We wish to find out whether a topical application of a quick-acting, easily degradable pyrethrin insecticide reduces number of lice per feather in males, females, and young. We could run a treatment design that assigns a dose of pyrethrin to randomly chosen replicate males, to females, and to young birds. Even better would be to employ a factorial design, in which we add levels of dose as the factor. Both these designs require the availability of numerous cassowaries.

It is far more likely, since cassowaries are rare, that our zoo has only a pair and its single young. This shortage may force the choice of nested treatments. We apply a dose of pyrethrins topically to one area on each bird and use another area of the same bird as the control: the pyrethrin and control treatments are nested within a bird. If we select feathers randomly within each area before we count lice per feather, the nested treatments are random.

It may not always be evident whether we have a cross-classified or a nested treatment design. To help clarify the notion, we can lay out the cassowary/lice experiment as a cross-classified design (fig. 4.1, top) and as a nested design (bottom). In the cross-classified design it is clear that a common set of treatments (pyrethrins, P, and controls, *C)* are applied to k replicates (birds) of three types of cassowary (male, female, and juvenile). In the nested version, we have *k* replicates of the three cassowary types, and we apply the pyrethrins and control treatment to each bird. Because the birds may differ in ways that we are not aware of, the treatments (pyrethrin and control) are particular (nested) to each bird. In actuality, most nested design comes about because we lack replicates. and only one experimental unit might be available.

Some of the drawbacks of nested designs emerge in figure 4.1. We are unable to evaluate a possible interaction between the insecticide treatment and sex or age of cassowaries, because interactions between variables can be quantified only in cross-classified treatments. In the nested treatments, we compare the difference between the two treatments *within* each bird only, rather than among a random sample of cassowaries. We might also be concerned that there is a correlation between treatments, either because the lice in the control area are af-

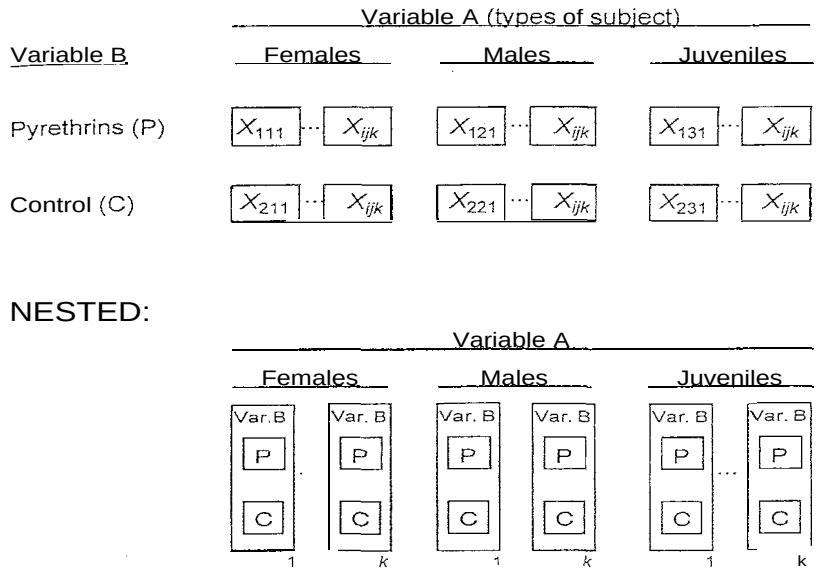

Fig. 4.1 Illustration of a cross-classified (top) and a nested [bottom) design for the cassowary/lice experiment.

fected by the pyrethrins in the treated area, or because the host bird influences both nested treatments unduly (what if this one bird is inordinately fond of dust baths?).

Gradient Treatments

If we knew that the water samples in our chrysanthemum experiment differed in concentration of a known substance (salt, nitrate, molybdenum, etc.), the responses of the plants could be related to that specific characteristic. The design of an experiment to assess the response of the experimental units to a gradient of a treatment variable is called a gradient (or regression) treatment design. The resulting data would be analyzed by regressions of the appropriate model.

This kind of design could be used more often than it is, particularly if we deal with comparative research approaches rather than strictly manipulative approaches. For example, we might be interested in how much the nitrogen that enters estuaries affects the concentration of chlorophyll in estuarine water. Since enriching estuaries by experimentally adding nitrogen is impractical, and in some places illegal, we might have to content ourselves with comparing chlorophyll concentrations in a series of estuaries subject to different rates of nitrogen enrichment. We cannot really fix the rate of nitrogen supply to the experimental units (the estuaries). We can, however, select a range of estuaries with a range of nitrogen loading rates and use this as the gradient treatment whose effect on the dependent variable is assessed by regression analysis.

4.3 Design of Layout

There are myriad ways to lay out studies, that is, to apply treatments to experimental units. This topic has received much attention and is often referred to as experimental design. Here I take the liberty of calling this layout design, because experimental design more appropriately might refer to all three components of research design (treatments, layout, and response).

Principles of Layout Design

To try to make some sense of the bewildering diversity of designs, we will first focus on a few principles underlying layout design, including randomization, replication, and stratification. Balancing, confounding, and splitting of plots are other basic, more complicated, but perhaps less important principles (at least in my experience), so I leave it to the interested reader to find out about these additional principles in the additional readings at the end of the chapter. All these principles of layout design deal with how we might assign treatments to experimental units so as to assess the influence of the treatments on dependent variables. Once we have learned something about the principles, we will briefly examine a few selected designs in the following section to see how the principles are applied.

In this section we will use terms and ideas already broached in our discussion of statistical analysis in chapter 3. There we reviewed the methods of analyzing data; here we go over options for layouts that would produce data amenable to the kinds of analyses described in chapter 3.

Randomization

Randomization is the assignment of treatments to experimental units so as to reduce bias. It is designed to control (reduce or eliminate) for any sort of bias. Suppose that we plan to do an experiment to assess the effect of fertilization with 0, 5, or 10 mg nitrogen per week on growth of lettuce plants in a greenhouse during winter. If the heat source is at one end of the greenhouse, we might suspect that there could be a bias; that is, plants grown nearer the heater will do better. If we place the plants that receive one or another nitrogen dose at either end of the greenhouse, the bias provided by the heat might confuse our results. Actually, such gradients might exist in any experiment, and many of the biases surely present will be unknown to us. Therefore, it is always a good precaution to assign the treatments to experimental units at random, hence nullifying as much as possible any biases that might be present. The fundamental objective of randomization is to ensure that each treatment is equally likely to be assigned to any given experimental unit. In our experiment, this means that each fertilization treatment applied to lettuce plants is equally likely to be located in any position along the axis of the greenhouse.

Randomization can be achieved by use of random number tables available in most statistical textbooks or random numbers produced by many

computers. If neither of these is available, most of us grizzled experimentalists have at one time or another appealed to a certain time-tested ploy, using the last digits of phone numbers in telephone directories. In relatively simple experiments, we can randomize fairly readily. If we wanted to grow 9 lettuce plants in a row oriented along the axis of the greenhouse, we would number each of the pots, up to 9. We then could use the series of random numbers, which could be

5 2 9 5 6 0 2 8 0 1 4 9 3 6 7 8, and so on

We would then allocate each of our three treatments (call them doses 1, 2, 3) to pots. For example, treatment dose 1 would be applied to pot position 5, dose 2 to position 2, dose 3 to position 9. We would continue with dose 1 applied to position 6 (since position 5 was already occupied), and so on, until we had completed the number of pots to be given each treatment.

It should be evident that if the experimental design is more complicated, the randomization may become more elaborate. For example, if we want to grow the lettuce plants in several parallel rows, we need to randomize position assignment within each row.

Replication

Replication[2] is the assigning of more than one experimental unit to a treatment combination (in the case of a manipulative experiment) or classification (in the case of comparisons). Replication has several functions. First and foremost, replication provides a way to control for random variation—recall from chapter 1 that a hallmark of empirical science is the principle of controlled observations. Replication makes possible the isolation of effects of treatments by controlling for variation caused by chance effects. Replication is the only way we can measure within-group variation of the dependent variable we are studying. Replication allows us to obtain a more representative measure of the population we wish to make inferences about, since the larger the sample, the more likely we are to get an estimate of the population.[3] It also generally improves the precision of estimates of the variable being studied.

Although at first thought replication appears to be a simple notion, it is a matter of considerable subtlety. There are different ways to obtain multiple samples, including *external replication*, *internal replication*, *subsampling*, and *repeated measures* (fig. 4.2). Each of these has different properties and applications; immediately below we examine the procedures and properties of alternative ways to obtain multiple samples.

2. The term ***replication*** has been used in other ways in experimentation. Some use it to describe the initial similarity of experimental units; others, to say that a response by the dependent variable can be reliably repeated after repeated application of the treatment These are unfortunate and confusing uses that should be discouraged.

3. This generalization may not be true if, as we increase n, we begin-to include values for some other, different population. Larger n is hence not always desirable.

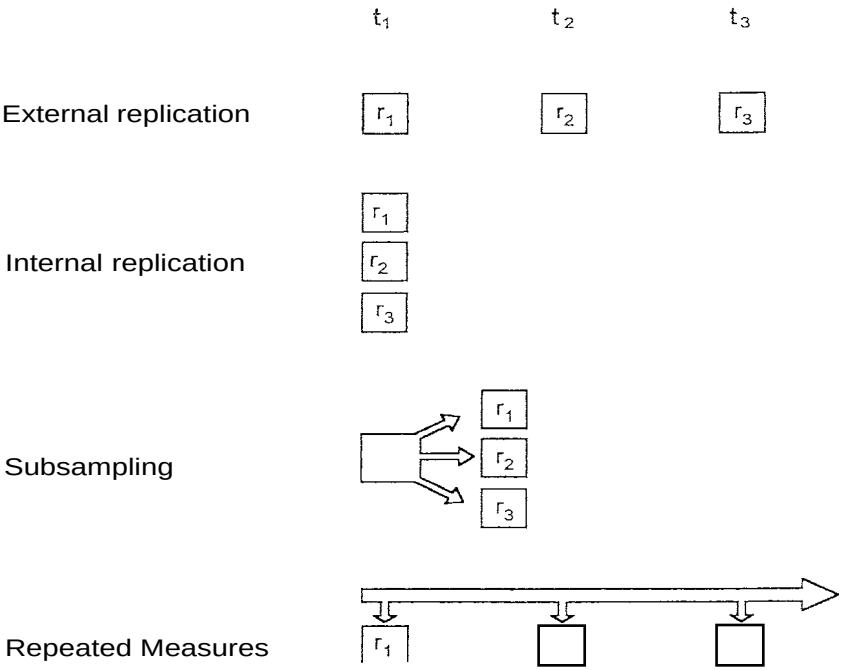

Fig. 4.2 Different ways to obtain multiple measurements. "r" refer to the specific replicate, "t" refers to measurements done at different times.

External Replication Suppose we are interested in measuring the nitrate content of water in a lake. We know that there is bound to be some variability in nitrate content of water over the lake, so we plan to take more than one sample; that is, we want *replicate samples*, so that we might then calculate a mean value that represents the whole lake.

We could obtain a sample of water at a given time (t_1; see fig. 4.2, top), and then measure its nitrate content. To get more than one sample, we could return at times t_2 and t_3 and collect more water in which to measure nitrate. We thus have three samples of nitrate from the lake. These are indeed replicates, but the variation that they would include reflects not only the variation of nitrate over the lake, but also the variation that could have occurred over the time interval (t_1 to t_3) through which they were collected. This method of obtaining replication, called *external replication*, confounds the contribution to variation due to time with the variation that is the subject of the study. If variation through time can be assumed to be modest, this procedure works well. There may be circumstances where it is necessary, for logistic or other reasons, to use external replicates.

Internal Replication A better way to obtain replicates is to collect independent samples as contemporaneously as possible (fig. 4.2, second row). This procedure, called *internal replication*, provides samples that capture the variation of interest, without confounding the results with the potential effects of passage of time.

Clearly, these internal and external replications are extremes in a continuum; it is the rare study in which replicetes are taken synchronously, and there is always some spatial separation to taking samples or treating experimental units alike. Decisions as to type of replication depend on whether the effects of time are likely to be important relative to the variation to be measured. Much depends on the system and its variation. For example, it may very well be that in a large lake, with one vessel available, it may take days to sample widely spaced stations, and time becomes a potentially more important factor to worry about; over the course of days, winds may change or a storm may alter nutrient content of the water. Alternatively, if samples taken only some meters apart are as variable as those taken many kilometers apart, then the sampling can be nearly contemporaneous. Logistics of sampling, spatial and temporal scales of the measurements, and inherent variability of the system studied therefore affect how we can carry out replication in any study.

Subsampling If at any given time we went to a site within our lake and collected a large carboy of water, brought it to the laboratory, subdivided the contents into aliquots, and performed nitrate measurements on each aliquot, we would also have multiple samples. These are *subsamples*, however, replicates not of the variation in the lake but of the water that was collected in the carboy (and probably made more homogeneous yet by mixing in the carboy). In general, variability among subsamples is smaller, naturally enough, than variability among replicates.

The relative homogeneity of subsamples may be useful if, for example, we want to assess the variability of our analytical procedure to measure nitrate (orany other variable). For that purpose we expressly want to start with samples of water that are as similar as possible, and see what variation is introduced by the analytical procedure by itself.

Hurlbert (1984) argued that it is important not to confuse true replication with subsampling or repeated measurements. A survey of published papers in environmental science showed that 26% of the studies committed "pseudoreplication," that is, used subsamples from an experimental unit to calculate the random error term with which to compare the treatment effects. That may sound too abstract; let us examine an example.

Suppose we have a comparative study in which we are trying to determine whether maple leaves decompose more rapidly when lying on sediments at a depth of 1 m compared to a depth of 10 m. Say we are in a hurry and place all of 8 bags of leaves at one site at 1 m, and 8 more at another site where the depth is 10 m. We come back 1 month later, harvest the bags, weigh the leaf material left, calculate the variation from the 8 bags, and do a statistical analysis, in this case, a one-way ANOVA with n = 8. If the F test shows that the differences between sites relative to within bags are significantly high, we can correctly infer that the decay rates between the two sites differ. If, on the other hand, we conclude that the results show that there are significant differences betrveen the *1* m and *10 m* depths, not only are we committing pseudoreplication, but we are also wrong. Since the bags were not randomly allotted to sites at each of the depths, we have no way to examine whether the differences in decay are related to depth or if similar differences could have been obtained at

any two stations, regardless of depth. In this example, the bags are more like subsamples than true replicates; differences among the bags measure the variability within a small site.

Pseudoreplication occurs if we push data into inappropriate statistical tests. It is not a problem in the science itself. In the example of the preceding paragraph, if we were content to make conclusions simply about the specific depths or stations, rather than about sites and depths in general, there is no problem of pseudoreplication.

Repeated Measurements A special case of multiple measurements that is common in animal research is when measurements are repeatedly done on the same experimental unit over the course of time. The variation captured by series of such measurements reflects effects of time (as in external replication), plus the effect of repeated or prolonged exposure of the experimental unit to the treatment. Unless the experimenter can be assured that there are no such cumulative effects (a most difficult task), repeated measures are not a good way to achieve replication. Repeated measures are more suited to detect cumulative effects of treatments on processes such as learning, memory, or tolerance.

In the case of our water sample, for example, repeated measurements could be used to estimate the time during which the sample still remains a good estimate of field conditions. Such data could also be used to assess rate of loss of nitrate to microbial action in the sample bottle, under whatever conditions the bottle was held.

In the experiment designed to find relief for our lousy cassowaries, we might want to see if there is indeed a progressive reduction of lice per bird as a result of the insecticide treatment. We could repeat the measurements of lice per feather in both areas of the three birds, perhaps once a week for several weeks. This sampling would address the issue of cumulative effects following treatment. This sampling would also answer the question of whether the insecticide reduced lice in the control area as well as in the treated area. You might note that this is not exactly a repeated measure design, since at the different times we could collect and count lice on a different set of feathers. This just shows that sometimes designs are hard to classify into simple categories.

Similarly, although we have discussed different categories for obtaining multiple measurements, in reality these categories are less clear cut than they might appear, and often create much confusion. For example, it should be evident that there is a continuum between internal and external replication, depending on the temporal and spatial scales of the samples and system under study. There is also a continuum between external replicates and repeated measures, since, for example, we might repeatedly sample vegetation biomass in a parcel subject to a fertilization treatment. If we measure vegetation cover in a nondestructive fashion, we are obtaining data similar to that of a behaviorist recording activity of one animal subject to a given treatment. The issue here is not to be too concerned with types of replicates, but rather to decide what is the most appropriate way to assess variation within a set of experimental units treated alike, for whatever scientific question being asked.

How Many Replicates?

The preceding paragraphs address the issues of replication as a way to, first, estimate random variation and, second, obtain representative estimates of variables we wish to study. A third function of replication is to control variation. Replication can increase .the precision of estimates of means, for example. This becomes evident when we consider the variance of a mean, $s_{\bar{Y}}^2 = s^2/n$: our estimates of variation are proportional to $1/n$, where n is the number of replicates. As it turns out, however, this is an oversimplification, since variances do not in reality decrease indefinitely. As n increases, we are necessarily sampling larger and larger proportions of values in perhaps different populations, and there is usually increased heterogeneity as n increases, which may result in larger variances. We discussed this topic in section 2.6, addressing tests of hypotheses. So, the question is, how many replicates are necessary and sufficient? There is, in fact, a large subfield of statistics that addresses the choice of sample size.

One quick way to roughly ascertain if the level of replication in a study might be suitable is to plot a graph of variance versus n. If we already have some measurements, the variance may be calculated by picking (at random from among the n replicates) 2, 3, 4, . . . , n values and calculating s. We then plot this versus n. A reasonable sample size for further study is that n beyond which the variance seems to become relatively stable as n increases. Unfortunately, such graphs more often than not are done after the study is completed, when there is no chance to modify the design. More sophisticated versions of the s^2/n approach are the basis for procedures given in statistical texts for determining sample size. The values of n required by statistical analyses of sample size are almost invariably higher than most researchers can expect to obtain. This is muck like the well-known example that by the principles of aerodynamics, bumblebees cannot possibly fly. Bumblebees nonetheless fly—and researchers go on advancing science, even though the replicate numbers they use should not in theory enable them to evaluate their results.

It has been my experience that, at least in environmental sciences, number of replicates is, in the end, restricted by practical considerations of logistics and available resources. We often find ourselves choosing sample sizes as large as possible given the situation, and the number is usually fewer than might be desirable based on calculations of sample size. In actuality, in most fields the numbers of replicates are relatively low, and the adequacy of sample size largely goes unevaluated. Kareiva and Anclersen (1988) found that 45% of ecological studies used replication of no more than 2. They also found that up to 20 replicates seemed feasible if plots were of smaller size, but replication was invariably less than 5 if the experiment involved plots larger than one meter in diameter.

Some important pieces of research, however, have been unreplicated manipulations. The experiment that motivated Sir R. A. Fisher to develop much of statistics, the celebrated Rothamsted fertilizer trials (see chap-

ter frontispiece), was unreplicated.[4] Many key experiments that gave rise to new directions in environmental science were also unreplicated. Lack of replication did not prevent the Hubbard Brook whole-watershed deforestation experiment (Bormann and Likens 1979), the Canadian Experimental Lakes Area whole-lake fertilization experiments (Schindler 1987), or IRONEX, the km^2-scale iron enrichment in the Pacific Ocean (Martin et al. 1994) from making major contributions to environmental science (Carpenter et al. 1995). In these studies the effects of the manipulations were sufficiently clear that there was little doubt as to the effects of the treatment, and statistical analysis was not required. Unreplicated studies cannot, therefore, be disregarded; we merely need to make sure that we choose our treatments, layout, and response well enough that if the effect is there, it will be evident. In addition, newer statistical approaches promise better ways to scrutinize results of unreplicated studies (Matson and Carpenter 1990).

As discussed in chapter 3, it is good to be wary of studies with very large numbers of samples or replicates. Statistical comparisons based on large numbers of observations might turn out to yield statistically significant differences (because so many degrees of freedom are involved), even though the actual differences found are so tiny that under practical circumstances the differences might be undetectable or unimportant (e.g., see fig. 10.2, bottom).

It is also good to be wary of studies involving very few samples. Comparisons based on a few observations lack power, as discussed in chapter 2. Large and important differences might not be declared "significant" if replicates are few. Of course, a larger number of replicates might be unfeasible, in which case the researcher has to sharpen the experimental design as much as possible, applying design principles that are described next.

Stratification

Replicate experimental units, or sampling sites, have to be laid out or located over space. Inevitably, there will be differences in many variables from one site to another. The effects of all these differences will be reflected in the variability of the measurements taken from the replicate units, and since we aim to compare treatment variation with variation associated with random error, it is desirable to minimize the variation attributable to random variation. We might be able to reduce the undesirable random variability if we could isolate the contribution to variation attributable to gradients in other variables that might or might not be of interest but are known to vary in our area of study. Once we can remove the effect of the known variables, we have a better estimate of random variation and can better compare the effects of the treatment, the independent variable that is of interest in the experiment, relative to random variation.

4. Even the most capable can err — Hurlbert (1984) points out that Fisher himself committed pseudoreplication in a first analysis of data from a manure experiment with potatoes. Fisher subsequently omitted the offending data, but never acknowledged the slip.

Controls

I have hardly mentioned this essential part of experimentation since chapter 1. Now we have some more concepts and terminology that allow us to make a few more distinctions. The term controls is used in a variety of ways in relation to experimentation:

- Control treatments allow evaluation of manipulative treatments, by controlling for procedure effects and temporal changes.
- Replication and randomization control for random effects and bias.
- Interspersion of treatments controls for regular spatial variation among experimental units.
- Regulation of the physical environment or experimental material (by the experimenter) confers better control on the experiment.

It would be preferable to reserve the term controls for only the first type of use, the one most meaningful for statistical tests. The second and third meanings are only extensions of the first, to help us understand the functions of replication, randomization, and stratification. The last meaning has the least to recommend it. It may derive from the maxim "hold constant all variables but the one of interest" but is unfortunate because in a true experiment the adequacy of a control treatment is not necessarily related to the degree to which the researcher restricts the conditions under which the experiment is run.

To isolate the effects of the "nuisance" variables we lay out "blocks" or "strata" (hence stratification), within which these variables are more or less constant. In such stratified designs, if we have j blocks or strata, we can therefore remove a term β_j from ε_{ij}, the term describing random variation:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

We discuss stratified experimental designs further below

Basic Experimental Layouts

While there may be as many layout designs as experiments, there are a few essential layouts that illustrate the fundamental principles.

Randomized Layouts

The simplest way to lay out an experiment (i.e., to assign treatments to experimental units) is to randomize. The top two layouts in figure 4.3 show randomized layouts in a case where we have a linear or a square formation of experimental units. The experimental units could be rows of plots in a field, ponds, aquaria on a laboratory bench, and so on. Two treatments (shown by black and white boxes) are assigned at random to the 8 or 12 experimental units.

In experiments in which the number of replicates is small (fewer than 4–6), a randomized layout may segregate replicates subjected to one treatment from those given the other treatment (top row, fig. 4.4) leading to possible biases. With just three replicates, for example, there is 10% chance that the first three replicates will receive one of the treatments. Thus, a completely randomized layout might not always provide the best layout de-

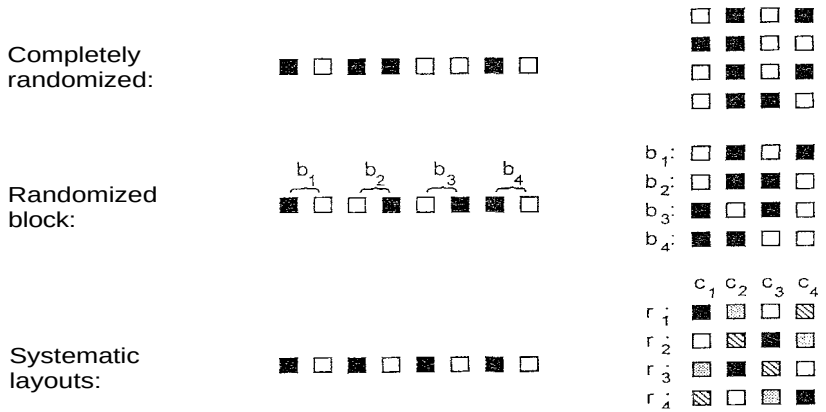

Fig. 4.3 Layout of experimental units in rows *(left)* or in squares *(right)* for completely randomized, randomized block, and systematic layout designs.

sign. Completely randomized layouts such as that shown on the top right of figure 4.3 are less subject to the problem of segregated treatments.

The data from layouts such as the linear arrangement (fig. 4.3, top left) and the square (middle right) are good candidates for analysis by a one-way ANOVA or equivalent nonparametric tests.

Randomized Blocks

We can lag out experimental units by restricting randomization in one direction; this is called *blocking* by statisticians (fig. 4.3, middle). We discussed this concept above as stratification. In the linear arrangement of units (middle left), we can set up blocks b_1, b_2, b_3, and b_4 and assign treatments randomly within each block (hence the name of this layout). Similarly, we can set up blocks in the square formation (middle right)

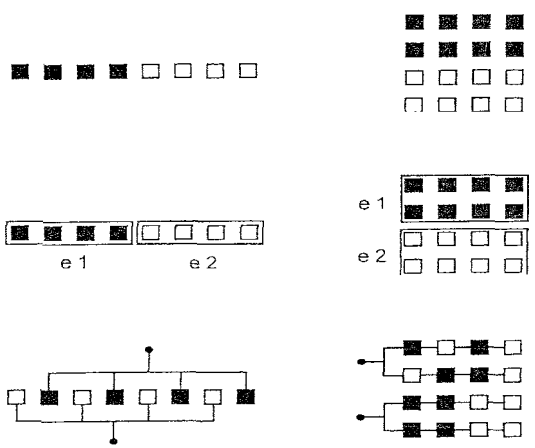

Fig. 4.4 Less desirable layouts. The e refers to emplacements that hold groups of experimental units.

and assign treatments randomly within blocks. Blocking reduces the possibility of chance segregation of treatments, in addition to preventing preexisting gradients and nondemonic intrusions from obscuring real treatment effects or prompting spurious ones.

The data resulting from a layout such as that of the middle left panel of figure 4.3 can be analyzed by a paired *t* test, a one-way unreplicated ANOVA, or equivalent nonparametric methods. The square arrangement of the middle right panel would yield data amenable to a two-way replicated ANOVA.

Systematic Layouts

Another way to restrict randomization in linear layouts is to intersperse or alternate the treatments systematically (fig. 4.3, bottom left). This is acceptable in many situations and might be better than a completely randomized layout, because this layout prevents segregation in linear arrangements of experimental units. Data from a systematic linear layout would be analyzed by unpaired *t* test or nonparametric tests.

One special case of systematic restriction of randomization in which the stratification is done in two directions (rows and columns) is called a *Latin square* (fig. 4.3, bottom right). Latin squares are useful for a variety of situations. They have been especially valuable in agricultural work. For instance, if fertilizer trials are done on a sloping field in which there is a strong prevailing wind perpendicular to the slope, a Latin square design offers the chance to stratify experimental units in the two directions and hence improve estimation of the fertilizer effect. Another situation in which Latin squares are appropriate is for tests under various classifications. One example is wear in automobile tires of different brands. If we are testing four brands of tires, we put one of each brand in each wheel position. We have to have four cars (or kinds of vehicles) to use in the test. We assign tire brands to each vehicle such that all brands appear in all four wheel positions. This layout allows us to examine performance of all four tire brands in each vehicle in each wheel position.

Latin square layouts are analyzed with a three-way ANOVA, in which rows, columns, and crops are the three components of the variation:

Latin but Not Creek Squares

In 1782 the Swiss mathematician Leonhard Euler gave a lecture to the Zeeland Scientific Society of Holland, in which he posed the following problem. The Emperor was coming to visit a certain garrison town. In the town there were six regiments, with six ranks of officers. It occurred to the garrison commander to choose 36 officers and arrange them in a square formation, so that the Emperor could inspect one of each of the six ranks of officers, and one officer from each regiment, from any side of the arrangernent. Euler assigned Latin letters (as he called them) to the ranks of officers and Creek letters to the regiments. He solved the Latin square and showed that the Creek square could not be solved simultaneously. Euler was correct, although his proof was flawed, but in any case he gave the name to the experimental layout: Latin squares.

Adapted from Pearce (*1965*).

$$Y_{ij} = \mu + r_i + c_i + t_{k(ij)} + \varepsilon_{ij}$$

where r and c stand for rows and columns and t is the treatment effect. The bracketed ij shows that the kth observation is only one within each row and column. Latin square layouts are useful but restrictive: the number of replicates must be equal to the number of treatments, and to obtain replicates we need to run more than one square. Mead (1988) provides details of the analysis.

4.4 Response Design

In any experiment or sampling, many kinds of responses by the experimental or sampling unit could be recorded. If we were testing the success of an antibiotic, we could record presence or absence of colonies of the bacterium in the agar in petri dishes to which the antibiotic and bacteria were added. We could also count the number of colonies per petri dish. We could also measure the area of petri dish covered by colonies. All these would in a way assess the action of the antibiotic.

The first question to ask about a response measurement is whether it relevantly answers the question posed in the experiment. Is presence or absence a sufficient response? Do we want a quantitative response? If we were testing different doses we might, but we could also be interested in a presence or absence response at various doses, to identify the threshold of action of the antibiotic. Are the responses meaningful in terms of the question? For example, area of colony might be just a response by a few resistant cells that grew rapidly after the antibiotic eliminated other cells; if so, area is not the best response to use to evaluate effectiveness of the antibiotic.

The data resulting from the different measurements would differ in the kind of analysis to be used. The presence and absence measurements

Some Undesirable Experimental Layouts

We have already noted that layouts such as those of figure 4.4 are undesirable. The obvious reason is that effects of position of the units might be confounded with treatment effects. But there are more subtle features to notice. At times the units are placed on different fields, ponds, receptacles, aquaria, and so on. This may be necessary, for example, if we are doing studies of responses of different bacterial clones to pressure and we can afford only two hyperbaric chambers. When possible, this is to be avoided, because our replicates become subsamples by this layout.

At other times we might apply treatments to units but in a way that links the units (fig. 4.4, bottom left). If we are adding two kinds of substrates in a test to see which increases yield of a fungal antibiotic, we might have a source for each of the substrates. This might be inevitable, but this delivery to the layout makes the units less independent of each other. We need to watch out for other links among the units in our layout (fig. 4.4, bottom right). If we are testing two different diets for cultivation ot mussels and we have only one seawater source, we might set up seawater connections as shown in the upper case in the bottom right panel. That physical link could make the treatments less distinct and deprive the units of independence. The worst situation is one I saw during a visit to an aquaculture facility. In that instance, not only was there a link from one unit to the next by flowing seawater, but one treatment was upstream of the other (bottom case in the bottom right panel).

would give *discrete responses, binomially distributed data,* or *Poisson-distributed data;* all would be analyzed by a nonparametric method. The binomially distributed data would yield counts that are initially discrete but that could be averaged if the experiment had replicates, yielding a *continuous response.* The continuous data would be analyzable using ANOVA, perhaps after some transformation because of the origin of the data as counts. The Poisson-distributed data would give continuous measurement data amenable to ANOVA analysis.

It is often the case that we measure a response not from the experimental unit but from what Urquhart (1981) calls the *evaluation unit.* For example, we may use a sample of blood from a toucan of a given species to assess its genetic similarity to another species. We might measure the diameter of a sample of a few eggs from a female trout to evaluate whether diet given to the trout affected reproduction. We take measurements from evaluation units, but we want to make inferences about experimental units. We need to make sure, therefore, that the evaluation units aptly represent the experimental units.

In some cases, we might want to make repeated evaluations of the response of the evaluation unit. When measuring the length of a wiggling fish, perhaps more than one measurement might be warranted; we might want to do several repetitions of a titration, just to make sure that no demonic intrusions affect our measurements. In a way, we discussed this above with the issue of subsampling. As in that case, these repetitions are used only to improve our measurement, rather than to increase significance of tests.

4.5 Sensible Experimental Design

I have emphasized concepts in chapters 1–3 because too many of us simply go to a statistics book and find a design and analysis that seem to fit our data. We then make the data fit that Procrustean bed,[5] often inappropriately, and rush on to comment on the results. Designing research and analyzing results will be better with some consultation with a knowledgeable person.

The first question that will be asked by a statistician is, "What is the question being asked?" I cannot overemphasize the importance of keeping—at all times—firmly in mind the specific question or questions we wish to answer. All else in design and analysis in science stems from *the questions.*

Then the statistician will delve into two themes—what are the best treatments and layout, and what constraints there might be on the experimental units. These echo the three parts of experimental design we just reviewed: treatments, layout, and response.

> [T]o adopt arrangements that we suspect are bad, simply because *[of* statistical demands] is to force our behavior into the Procrustean bed of a mathematical theory. Our object is the design of individual experiments that will work well: good [statistical] properties are concepts that help us doing this, but the exact fulfillment of . . . mathematical conditions is not the ultimate aim.
>
> *D. R. Cox (1958)*

5. Procrustes, in Greek mythology, was a cruel highwayman who owned a rather long bed. He forced passersby to fit the bed by stretching them. Procrustes is also said to have had a rather short bed; to make his captives fit that bed, he sawed off their legs. Theseus eventually dispatched Procrustes using Procrustes' own methods.