1

Fundamentals

The hardest part of any statistical work is getting started – and one of the hardest things about getting started is choosing the right kind of statistical analysis. The choice depends on the nature of your data and on the particular question you are trying to answer. The truth is that there is no substitute for experience; the way to know what to do, is to have done it properly lots of times before.

The key is to understand what kind of **response** variable you have got, and to know the nature of your **explanatory** variables. The response variable is the thing you are working on; it is the variable whose variation you are attempting to understand. This is the variable that goes on the *y* axis of the graph (the ordinate). The explanatory variable goes on the *x* axis of the graph (the abscissa); you are interested in the extent to which variation in the response variable is associated with variation in the explanatory variable. A continuous measurement is a variable like height or weight that can take any real numbered value. A categorical variable is a **factor** with two or more **levels**: gender is a factor with two levels (male and female), and a rainbow might be a factor with seven levels (red, orange, yellow, green, blue, indigo, violet).

It is essential, therefore, that you know:

- which of your variables is the response variable;
- which are the explanatory variables;
- are the explanatory variables continuous or categorical, or a mixture of both;
- what kind of response variable have you got – is it a continuous measurement, a count, a proportion, a time-at-death or a category?

These simple keys will then lead you to the appropriate statistical method.

1. The explanatory variables

(a)	All explanatory variables continuous	**Regression**
(b)	All explanatory variables categorical	**Analysis of variance (Anova)**
(c)	Explanatory variables both continuous and categorical	**Analysis of covariance (Ancova)**

Statistics: An Introduction using R M. J. Crawley
 ISBNs: 0-470-02298-1 (PBK); 0-470-02297-3 (PPC)

2. The response variable

(a)	Continuous	**Normal regression, Anova or Ancova**
(b)	Proportion	**Logistic regression**
(c)	Count	**Log linear models**
(d)	Binary	**Binary logistic analysis**
(e)	Time-at-death	**Survival analysis**

There are some key ideas that need to be understood from the outset. We cover these here before getting into any detail about different kinds of statistical model.

Everything Varies

If you measure the same thing twice you will get two different answers. If you measure the same thing on different occasions you will get different answers because the thing will have aged. If you measure different individuals, they will differ for both genetic and environmental reasons (nature and nurture). Heterogeneity is universal: spatial heterogeneity means that places always differ and temporal heterogeneity means that times always differ.

Because everything varies, finding that things vary is simply not interesting. We need a way of discriminating between variation that is scientifically interesting, and variation that just reflects background heterogeneity. That is why we need statistics. It is what this whole book is about.

The key concept is the amount of variation that we would expect to occur by chance alone, when nothing scientifically interesting was going on. If we measure bigger differences than we would expect by chance, we say that the result is statistically significant. If we measure no more variation than we might reasonably expect to occur by chance alone, then we say that our result is not statistically significant. It is important to understand that this is not to say that the result is not important. Non-significant differences in human life span between two drug treatments may be massively important (especially if you are the patient involved). Non-significance is not the same as 'not different'. The lack of significance may simply be due to the fact that our replication is too low.

On the other hand, when nothing really **is** going on, then we want to know this. It makes life much simpler if we can be reasonably sure that there is no relationship between y and x. Some students think that 'the only good result is a significant result'. They feel that their study has somehow failed if it shows that 'A has no significant effect on B'. This is an understandable failing of human nature, but it is not good science. The point is that we want to know the truth, one way or the other. We should try not to care too much about the way things turn out. This is not an amoral stance, it just happens to be the way that science works best. Of course, it is hopelessly idealistic to pretend that this is the way that scientists really behave. Scientists often hope passionately that a particular experimental result will turn out to be statistically significant, so that they can have a paper published in Nature and get promoted, but that doesn't make it right.

Significance

What do we mean when we say that a result is significant? The normal dictionary definitions of significant are 'having or conveying a meaning' or 'expressive; suggesting or implying deeper or unstated meaning' but in statistics we mean something very specific indeed. We mean that 'a result was unlikely to have occurred by chance'. In particular, we mean 'unlikely to have occurred by chance if the null hypothesis was true'. So there are two elements to it: we need to be clear about what we mean by 'unlikely', and also what exactly we mean by the 'null hypothesis'. Statisticians have an agreed convention about what constitutes 'unlikely'. They say that an event is unlikely if it occurs less than 5% of the time. In general, the 'null hypothesis' says that 'nothing's happening' and the alternative says 'something **is** happening'.

Good and Bad Hypotheses

Karl Popper was the first to point out that a good hypothesis is one that is capable of **rejection**. He argued that **a good hypothesis is a falsifiable hypothesis**. Consider the following two assertions.

1. There are vultures in the local park.
2. There are no vultures in the local park.

Both involve the same essential idea, but one is refutable and the other is not. Ask yourself how you would refute option 1. You go out into the park and you look for vultures, but you don't see any. Of course, this doesn't mean that there aren't any. They could have seen you coming, and hidden behind you. No matter how long or how hard you look, you cannot refute the hypothesis. All you can say is 'I went out and I didn't see any vultures'. One of the most important scientific notions is that **absence of evidence is not evidence of absence**. Option 2 is fundamentally different. You reject hypothesis 2 the first time that you see a vulture in the park. Until the time that you **do** see your first vulture in the park, you work on the assumption that the hypothesis is true. But if you see a vulture, the hypothesis is clearly false, so you reject it.

Null Hypotheses

The null hypothesis says 'nothing's happening'. For instance, when we are comparing two sample means, the null hypothesis is that the means of the two samples are the same. Again, when working with a graph of y against x in a regression study, the null hypothesis is that the slope of the relationship is zero, i.e. y is not a function of x, or y is independent of x. The essential point is that the null hypothesis is falsifiable. We reject the null hypothesis when our data show that the null hypothesis is sufficiently unlikely.

p Values

A p value is an estimate of the probability that a particular result, or a result more extreme than the result observed, could have occurred by chance, if the null hypothesis were true. In short, the p value is a measure of the credibility of the null hypothesis. If

something is very unlikely to have occurred by chance, we say that it is statistically significant, e.g. $p < 0.001$. For example, in comparing two sample means, where the null hypothesis is that the means are the same, a low p value means that the hypothesis is unlikely to be true and the difference is statistically significant. A large p value (e.g. $p = 0.23$) means that there is no compelling evidence on which to reject the null hypothesis. Of course, saying 'we do not reject the null hypothesis' and 'the null hypothesis is true' are two quite different things. For instance, we may have failed to reject a false null hypothesis because our sample size was too low, or because our measurement error was too large. Thus, p values are interesting, but they don't tell the whole story; effect sizes and sample sizes are equally important in drawing conclusions.

Interpretation

It should be clear by this point that we can make two kinds of mistakes in the interpretation of our statistical models:

- we can reject the null hypothesis when it is true, or
- we can accept the null hypothesis when it is false.

These are referred to as **Type I** and **Type II** errors respectively. Supposing we knew the true state of affairs (which, of course, we seldom do), then in tabular form:

	Actual situation	
Null hypothesis	*True*	*False*
Accept	Correct decision	Type II
Reject	Type I	Correct decision

Statistical Modelling

The object is to determine the values of the parameters in a specific model that lead to the best fit of the model to the data. The data are sacrosanct, and they tell us what actually happened under a given set of circumstances. It is a common mistake to say 'the data were fitted to the model' as if the data were something flexible, and we had a clear picture of the structure of the model. On the contrary, what we are looking for is the minimal adequate model to describe the data. The model is fitted to the data, not the other way around. The best model is the model that produces the least unexplained variation (the **minimal residual deviance**), subject to the constraint that all the parameters in the model should be statistically significant.

You have to specify the model. It embodies your mechanistic understanding of the factors involved, and of the way that they are related to the response variable. We want the model to be **minimal** because of the principle of parsimony, and **adequate** because there is no point in retaining an inadequate model that does not describe a significant fraction of the variation in the data. It is very important to understand that there is not just

one model; this is one of the common implicit errors involved in traditional regression and Anova, where the same models are used, often uncritically, over and over again. In most circumstances, there will be a large number of different, more or less plausible models that might be fitted to any given set of data. Part of the job of data analysis is to determine which, if any, of the possible models are adequate and then, out of the set of adequate models, which is the minimal adequate model. In some cases there may be no single best model and a set of different models may all describe the data equally well (or equally poorly if the variability is great).

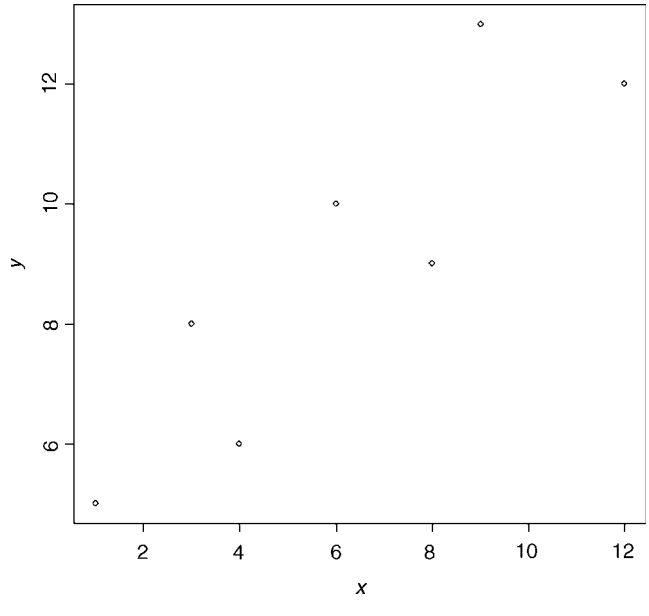
Maximum Likelihood

What exactly do we mean when we say that the parameter values should afford the 'best fit of the model to the data'? The convention we adopt is that our techniques should lead to **unbiased, variance minimizing estimators**. We define 'best' in terms of **maximum likelihood**. This notion is likely to be unfamiliar, so it is worth investing some time to get a feel for it. This is how it works.

- Given the data,
- and given our choice of model,
- what values of the parameters of that model make the observed data most likely?

Here are the data: y is the response variable and x is the explanatory variable. Because both x and y are continuous variables, the appropriate model is regression.

```
x <- c(1,3,4,6,8,9,12)
y <- c(5,8,6,10,9,13,12)
plot(x,y)
```

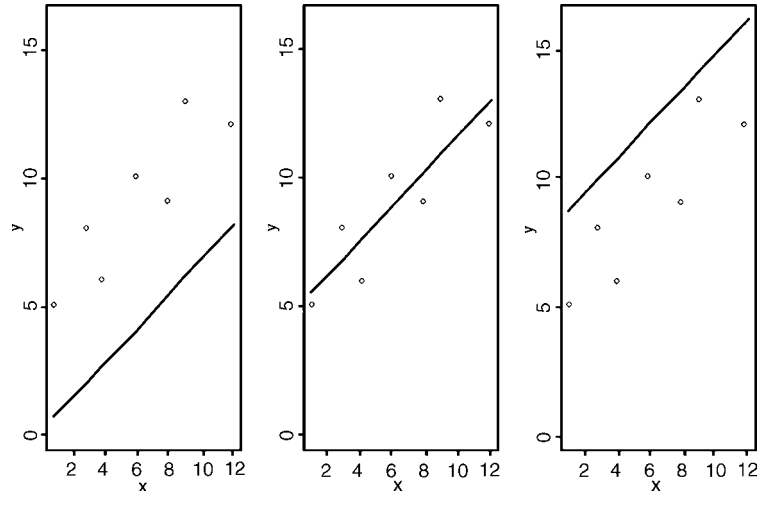


Now we need to select a regression model to describe these data from the vast range of possible models available. Let's choose the simplest model, the straight line

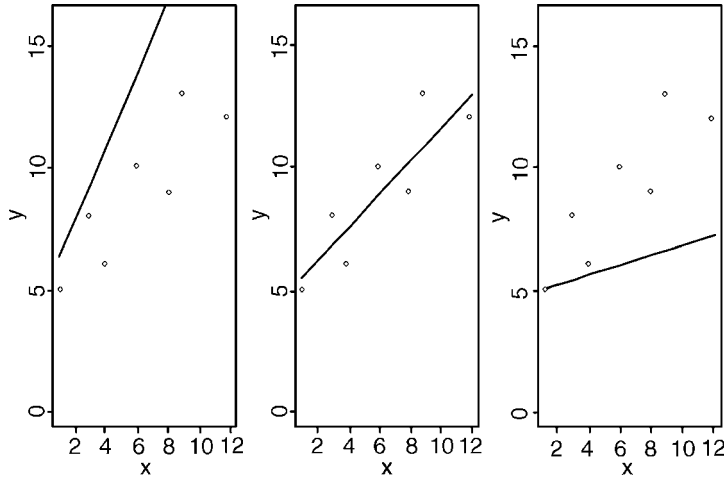
$$y = a + bx.$$

This is a two-parameter model; the first parameter, a, is the intercept (the value of y when x is 0) and the second, b, is the slope (the change in y associated with unit change in x). The response variable y, is a linear function of the explanatory variable x. Now suppose that we knew that the slope was 0.68, then the maximum likelihood question can be applied to the intercept a.

If the intercept were 0 (left-hand graph, below), would the data be likely? The answer of course, is no. If the intercept were 8 (right-hand graph) would the data be likely? Again, the answer is obviously no. The maximum likelihood estimate of the intercept is shown in the central graph (its value turns out to be 4.827).



We could have a similar debate about the slope. Suppose we knew that the intercept was 4.827, then would the data be likely if the graph had a slope of 1.5 (left graph, below)?



The answer, of course, is no. What about a slope of 0.2 (right graph)? Again, the data are not at all likely if the graph has such a gentle slope. The maximum likelihood of the data given the model is obtained with a slope of 0.679 (centre graph). This is not how the procedure is actually carried out, but it makes the point that we judge the model on the basis of how likely the data would be if the model were correct. In practice of course, both parameters are estimated simultaneously.

Experimental Design

There are only two key concepts:

- replication, and
- randomization.

You replicate to increase reliability. You randomize to reduce bias. If you replicate thoroughly and randomize properly, you will not go far wrong.

There are a number of other issues whose mastery will increase the likelihood that you analyse your data the right way rather than the wrong way:

- the principle of parsimony,
- the power of a statistical test,
- controls,
- spotting pseudoreplication and knowing what to do about it,
- the difference between experimental and observational data (non-orthogonality).

It does not matter very much if you cannot do your own advanced statistical analysis. If your experiment is properly designed, you will often be able to find somebody to help you with the statistics. However, if your experiment is not properly designed, or not thoroughly randomized, or lacking adequate controls, then no matter how good you are at statistics, some (or possibly even all) of your experimental effort will have been wasted. No amount of high-powered statistical analysis can turn a bad experiment into a good one. R is good, but not that good.

The Principle of Parsimony (Occam's Razor)

One of the most important themes running through this book concerns model simplification. The principle of parsimony is attributed to the 14th century English Nominalist philosopher William of Occam who insisted that, given a set of equally good explanations for a given phenomenon, then **the correct explanation is the simplest explanation**. It is called Occam's razor because he 'shaved' his explanations down to the bare minimum. In statistical modelling, the principle of parsimony means that:

- models should have as few parameters as possible,
- linear models should be preferred to non-linear models,
- experiments relying on few assumptions should be preferred to those relying on many,

- models should be pared down until they are *minimal adequate*,
- simple explanations should be preferred to complex explanations.

The process of model simplification is an integral part of hypothesis testing in R. In general, a variable is retained in the model only if it causes a significant increase in deviance when it is removed from the current model. Seek simplicity, then distrust it.

In our zeal for model simplification, we must be careful not to throw the baby out with the bathwater. Einstein made a characteristically subtle modification to Occam's razor. He said: 'A model should be as simple as possible. But no simpler'.

Observation, Theory and Experiment

There is no doubt that the best way to solve scientific problems is through a thoughtful blend of observation, theory and experiment. In most real situations, however, there are constraints on what can be done, and on the way things can be done, which mean that one or more of the trilogy has to be sacrificed. There are lots of cases, for example, where it is ethically or logistically impossible to carry out manipulative experiments. In these cases it is doubly important to ensure that the statistical analysis leads to conclusions that are as critical and as unambiguous as possible.

Controls

No controls, no conclusions.

Replication: It's the *n*'s that Justify the Means

The requirement for replication arises because if we do the same thing to different individuals we are likely to get different responses. The causes of this heterogeneity in response are many and varied (genotype, age, gender, condition, history, substrate, microclimate, and so on). The object of replication is to increase the reliability of parameter estimates, and to allow us to quantify the variability that is found within the same treatment. To qualify as replicates, the repeated measurements:

- must be independent,
- must not form part of a time series (data collected from the same place on successive occasions are not independent),
- must not be grouped together in one place (aggregating the replicates means that they are not spatially independent),
- must be of an appropriate spatial scale.

Ideally, one replicate from each treatment ought to be grouped together into a block, and each treatment repeated in many different blocks. Repeated measures (e.g. from the same individual or the same spatial location) are not replicates (this is probably the commonest cause of pseudoreplication in statistical work).

How Many Replicates?

The usual answer is 'as many as you can afford'. An alternative answer is 30. A very useful rule of thumb is this: a sample of 30 or more is a big sample, but a sample of less than 30 is a small one. The rule doesn't always work, of course: 30 would be derisively small as a sample in an opinion poll, for instance. In other circumstances, it might be impossibly expensive to repeat an experiment as many as 30 times. Nevertheless, it is a rule of great practical utility, if only for giving you pause as you design your experiment with 300 replicates that perhaps this might really be a bit over the top – or when you think you could get away with just five replicates this time.

There are ways of working out the replication necessary for testing a given hypothesis (these are explained below). Sometimes we know little or nothing about the variance or the response variable when we are planning an experiment. Experience is important. So are pilot studies. These should give an indication of the variance between initial units before the experimental treatments are applied, and also of the approximate magnitude of the responses to experimental treatment that are likely to occur. Sometimes it may be necessary to reduce the scope and complexity of the experiment, and to concentrate the inevitably limited resources of manpower and money on obtaining an unambiguous answer to a simpler question. It is immensely irritating to spend three years on a grand experiment, only to find at the end of it that the response is only significant at $p = 0.08$. A reduction in the number of treatments might well have allowed an increase in replication to the point where the same result would have been unambiguously significant.

Power

The power of a test is the probability of rejecting the null hypothesis when it is false. It has to do with Type II errors: β is the probability of accepting the null hypothesis when it is false. In an ideal world, we would obviously make β as small as possible, but there is a snag. The smaller we make the probability of committing a Type II error, the greater we make the probability of committing a Type I error, and rejecting the null hypothesis when, in fact, it is correct. A compromise is called for. Most statisticians work with $\alpha = 0.05$ and $\beta = 0.2$. Now the power of a test is defined as $1 - \beta = 0.8$ under the standard assumptions. This is used to calculate the sample sizes necessary to detect a specified difference when the error variance is known (or can be guessed at). Suppose that for a single sample the size of the difference you want to detect is ∂ and the variance in the response is s^2 (e.g. known from a pilot study or extracted from the literature), then you will need n replicates to reject the null hypothesis with power = 80%:

$$n \approx \frac{8 \times s^2}{\partial^2}.$$

This is a reasonable rule of thumb, but you should err on the side of caution by having larger, not smaller samples than these. Suppose that the mean is close to 20, and the variance is 10, but we want to detect a 10% change (i.e. $\partial = \pm 2$) with probability 0.8, then $n = 8 \times 10/2^2 = 20$.

Here is the built-in function power.t.test in action for the case just considered. We need to specify that the type is ''one sample'', the power we want to obtain is 0.8, the difference to be detected (called delta) is 2.0, and the standard deviation (sd) is $\sqrt{10}$

```
power.t.test(type = "one.sample",power = 0.8,sd = sqrt(10),delta = 2)
```

```
     One-sample t test power calculation

              n = 21.62146
          delta = 2
             sd = 3.162278
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Other power functions available in R include power.anova.test and power.prop.test

Randomization

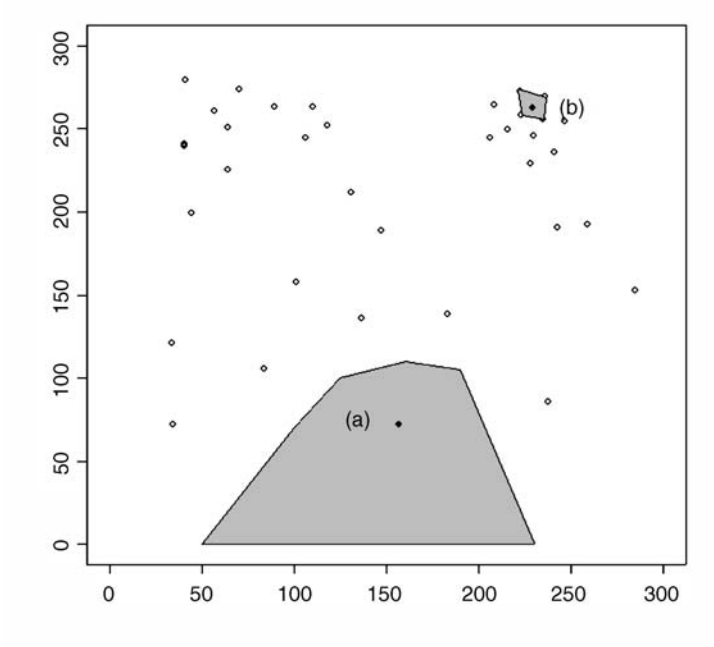
Randomization is something that everybody says they do, but hardly anybody does properly. Take a simple example. How do I select one tree from a forest of trees, on which to measure photosynthetic rates? I want to select the tree at random in order to avoid bias. For instance, I might be tempted to work on a tree that had accessible foliage near to the ground, or a tree that was close to the lab, or a tree that looked healthy, or a tree that had nice insect-free leaves, and so on. I leave it to you to list the biases that would be involved in estimating photosynthesis on any of those trees. One common way of selecting a 'random' tree is to take a map of the forest and select a random pair of coordinates (say 157m east of the reference point, and 68m north). Then pace out these coordinates and, having arrived at that particular spot in the forest, select the nearest tree to those coordinates. But is this really a randomly selected tree?

If it was randomly selected, then it would have exactly the same chance of being selected as every other tree in the forest. Let us think about this. Look at the figure below which shows a plan of the distribution of trees on the ground. Even if they were originally planted out in regular rows, accidents, tree-falls, and heterogeneity in the substrate would soon lead to an aggregated spatial distribution of trees. Now ask yourself how many different random points would lead to the selection of a given tree. Start with tree (a). This will be selected by any points falling in the large shaded area.

Now consider tree (b). It will only be selected if the random point falls within the tiny area surrounding that tree. Tree (a) has a much greater chance of being selected than tree (b), and so the nearest tree to a random point is not a randomly selected tree. In a spatially heterogeneous woodland, isolated trees and trees on the edges of clumps will always have a higher probability of being picked than trees in the centre of clumps.

The answer is that to select a tree at random, every single tree in the forest must be numbered (all 24 683 of them, or whatever), and then a random number between 1 and 24 683 must be drawn out of a hat. There is no alternative. Anything less than that is not randomization.

Now ask yourself how often this is done in practice, and you will see what I mean when I say that randomization is a classic example of 'do as I say, and not do as I do'. As



an example of how important proper randomization can be, consider the following experiment that was designed to test the toxicity of five contact insecticides by exposing batches of flour beetles to the chemical on filter papers in Petri dishes. The animals walk about and pick up the poison on their feet. The *Tribolium* culture jar was inverted, flour and all, into a large tray, and beetles were collected as they emerged from the flour. The animals were allocated to the five chemicals in sequence; four replicate Petri dishes were treated with the first chemical, and ten beetles were placed in each Petri dish. Do you see the source of bias in this procedure?

It is entirely plausible that flour beetles differ in their activity levels (gender differences, differences in body weight, age, etc.). The most active beetles might emerge first from the pile of flour. These beetles all end up in the treatment with the first insecticide. By the time we come to finding beetles for the last replicate of the fifth pesticide, we may be grubbing round in the centre of the pile, looking for the last remaining *Tribolium*. This matters, because the amount of pesticide picked up by the beetles will depend upon their activity levels. The more active the beetles, the more chemical they pick up, and the more likely they are to die. Thus, the failure to randomize will bias the result in favour of the first insecticide because this treatment received the most active beetles.

What we should have done is this. Fill $5 \times 4 = 20$ Petri dishes with ten beetles each, adding one beetle to each Petri dish in turn. Then allocate a treatment (one of the five pesticides) to each Petri dish at random, and place the beetles on top of the pre-treated filter paper. We allocate Petri dishes to treatments most simply by writing a treatment number of a slip of paper, and placing all 20 pieces of paper in a bag. Then draw one piece of paper from the bag. This gives the treatment number to be allocated to the Petri dish in question. All of this may sound absurdly long-winded but, believe me, it is vital.

The recent trend towards 'haphazard' sampling is a cop-out. What it means is that 'I admit that I didn't randomize, but you have to take my word for it that this did not introduce any important bias'. You can draw your own conclusions.

Strong Inference

One of the most powerful means available to demonstrate the accuracy of an idea is an experimental confirmation of a prediction made by a carefully formulated hypothesis. There are two essential steps to the protocol of strong inference (Platt 1964):

- formulate a clear hypothesis, and
- devise an acceptable test.

Neither one is much good without the other. For example, the hypothesis should not lead to predictions that are likely to occur by other extrinsic means. Similarly, the test should demonstrate unequivocally whether the hypothesis is true or false.

A great many scientific experiments appear to be carried out with no particular hypothesis in mind at all, but simply to see what happens. While this approach may be commendable in the early stages of a study, such experiments tend to be weak as an end in themselves, because there will be such a large number of equally plausible explanations for the results. Without contemplation there will be no testable predictions; without testable predictions there will be no experimental ingenuity; without experimental ingenuity there is likely to be inadequate control; in short, equivocal interpretation. The results could be due to myriad plausible causes. Nature has no stake in being understood by scientists. We need to work at it. Without replication, randomization and good controls we shall make little progress.

Weak Inference

The phrase weak inference is used (often disparagingly) to describe the interpretation of observational studies and the analysis of so-called 'natural experiments'. It is silly to be disparaging about these data, because they are often the only data that we have. The aim of good statistical analysis is to obtain the maximum information from a given set of data, bearing the limitations of the data firmly in mind.

Natural experiments arise when an event (often assumed to be an unusual event, but frequently without much justification of what constitutes unusualness) occurs that is like an experimental treatment (a hurricane blows down half of a forest block; a landslide creates a bare substrate; a stock market crash produces lots of suddenly poor people, etc). Hairston (1989) said: 'The requirement of adequate knowledge of initial conditions has important implications for the validity of many natural experiments. Inasmuch as the ''experiments'' are recognized only when they are completed, or in progress at the earliest, it is impossible to be certain of the conditions that existed before such an ''experiment'' began. It then becomes necessary to make assumptions about these conditions, and any conclusions reached on the basis of natural experiments are thereby weakened to the point of being hypotheses, and they should be stated as such' (Hairston 1989).

How Long to Go On?

Ideally, the duration of an experiment should be determined in advance, lest one falls prey to one of the twin temptations:

- to stop the experiment as soon as a pleasing result is obtained;
- to keep going with the experiment until the 'right' result is achieved (the 'Gregor Mendel effect').

In practice, most experiments probably run for too short a period, because of the idiosyncrasies of scientific funding. This short-term work is particularly dangerous in medicine and the environmental sciences, because the kind of short-term dynamics exhibited after pulse experiments may be entirely different from the long-term dynamics of the same system. Only by long-term experiments of both the pulse and the press kind, will the full range of dynamics be understood. The other great advantage of long-term experiments is that a wide range of patterns (e.g. 'kinds of years') is experienced.

Pseudoreplication

Pseudoreplication occurs when you analyse the data as if you had more degrees of freedom than you really have. There are two kinds of pseudoreplication:

- temporal pseudoreplication, involving repeated measurements from the same individual, and
- spatial pseudoreplication, involving several measurements taken from the same vicinity.

Pseudoreplication is a problem because one of the most important assumptions of standard statistical analysis is **independence of errors**. Repeated measures through time on the same individual will have non-independent errors because peculiarities of the individual will be reflected in all of the measurement made on it (the repeated measures will be temporally correlated with one another). Samples taken from the same vicinity will have non-independent errors because peculiarities of the location will be common to all the samples (e.g. yields will all be high in a good patch and all be low in a bad patch).

Pseudoreplication is generally quite easy to spot. The question to ask is how many degrees of freedom for error does the experiment really have? If a field experiment appears to have lots of degrees of freedom, it is probably pseudoreplicated. Take an example from pest control of insects on plants. There are 20 plots, ten sprayed and ten unsprayed. Within each plot there are 50 plants. Each plant is measured five times during the growing season. Now this experiment generates $20 \times 50 \times 5 = 5000$ numbers. There are two spraying treatments, so there must be 1 degree of freedom for spraying and 4998 degrees of freedom for error. Or must there? Count up the replicates in this experiment. Repeated measurements on the same plants (the five sampling occasions) are certainly not replicates. The 50 individual plants within each quadrat are not replicates either. The reason for this is that conditions within each quadrat are quite likely to be unique, and so all 50 plants will experience more or less the same unique set of conditions, irrespective of the spraying treatment they receive. In fact, there are ten replicates in this experiment. There are ten sprayed plots and ten unsprayed plots, and each plot will yield only one independent datum to the response variable (the proportion of leaf area consumed by

insects, for example). Thus, there are nine degrees of freedom within each treatment, and $2 \times 9 = 18$ degrees of freedom for error in the experiment as a whole. It is not difficult to find examples of pseudoreplication on this scale in the literature (Hurlbert 1984). The problem is that it leads to the reporting of masses of spuriously significant results (with 4998 degrees of freedom for error, it is almost impossible not to have significant differences). The first skill to be acquired by the budding experimenter is the ability to plan an experiment that is properly replicated.

There are various things that you can do when your data are pseudoreplicated:

- average away the pseudoreplication and carry out your statistical analysis on the means,
- carry out separate analyses for each time period,
- use proper time series analysis or mixed effects models.

Initial Conditions

Many otherwise excellent scientific experiments are spoiled by a lack of information about initial conditions. How can we know if something has changed if we don't know what it was like to begin with? It is often implicitly assumed that all the experimental units were alike at the beginning of the experiment, but this needs to be demonstrated rather than taken on faith. One of the most important uses of data on initial conditions is as a check on the efficiency of randomization. For example, you should be able to run your statistical analysis to demonstrate that the individual organisms were not significantly different in mean size at the beginning of a growth experiment. Without measurements of initial size, it is always possible to attribute the end result to differences in initial conditions. Another reason for measuring initial conditions is that the information can often be used to improve the resolution of the final analysis through analysis of covariance (see Chapter 10).

Orthogonal Designs and Non-orthogonal Observational Data

The data in this book fall into two distinct categories. In the case of planned experiments, all of the treatment combinations are equally represented and, barring accidents, there are no missing values. Such experiments are said to be *orthogonal*. In the case of observational studies, however, we have no control over the number of individuals for which we have data, or over the combinations of circumstances that are observed. Many of the explanatory variables are likely to be correlated with one another, as well as with the response variable. Missing treatment combinations are commonplace, and the data are said to be non-orthogonal. This makes an important difference to our statistical modelling because, in orthogonal designs, the deviance that is attributed to a given factor is constant, and does not depend upon the order in which that factor is removed from the model. In contrast, with non-orthogonal data, we find that the deviance attributable to a given factor does depend upon the order in which the factor is removed from the model. We must be careful, therefore, to judge the significance of factors in non-orthogonal studies, when they are removed from the maximal model (i.e. from the model including all the other factors and interactions with which they might be confounded). Remember, for non-orthogonal data, order matters.