of lines of computer script or code required to implement a model. Crudity and sophistication are harder to recognize; they represent the conceptual depth, or the amount of *hidden* complexity, involved in a model or statistical approach. For example, a computer model that picks random numbers to determine when individuals give birth and die and keeps track of the total population size, for particular values of the birth and death rates and starting population size, is simple and crude. Even simpler, but far more sophisticated, is the mathematical theory of random walks (Okubo, 1980) which describes the same system but — at the cost of challenging mathematics — predicts its behavior for *any* birth and death rates and any starting population sizes. A statistical model that searches at random for the line that minimizes the sum of squared deviations of the data is crude and simple; the theory of linear models, which involves more mathematics, does the same thing in a more powerful and general way. Computer programs, too, can be either crude or sophisticated. One can pick numbers from a binomial distribution by virtually flipping the right number of coins and seeing how many come up heads, or by using numerical methods that arrive at the same result far more efficiently. A simple R command like `rbinom`, which picks random binomial deviates, hides a lot of complexity.

The value of sophistication is generality, simplicity, and power; its costs are opacity and conceptual and mathematical difficulty. In this book, I will take advantage of many of R's sophisticated tools for optimization and random number generation (since in this context it's more important to have these tools available than to learn the details of how they work), but I will avoid many of its sophisticated statistical tools, so that you can learn from the ground up how statistical models really work and make your models work the way you want them to rather than being constrained by existing frameworks. Having reinvented the wheel, however, we'll briefly revisit some standard statistical frameworks like generalized linear models and see how they can solve some problems more efficiently.

1.4 FRAMEWORKS FOR STATISTICAL INFERENCE

This section will explore three different ways of drawing statistical conclusions from data — frequentist, Bayesian, and likelihood-based. While the differences among these frameworks are sometimes controversial, most modern statisticians know them all and use whatever tools they need to get the job done; this book will teach you the details of those tools, and the distinctions among them.

To illustrate the ideas I'll draw on a seed predation data set from Duncan and Duncan (2000) that quantifies how many times seeds of two different species disappeared (presumably taken by seed predators, although we can't be sure) from observation stations in Kibale National Park, Uganda. The two species (actually the smallest- and largest-seeded species of a set of eight species) are *Polyscias fulva* (`pol`: seed mass < 0.01 g) and *Pseudospondias microcarpa* (`psd`: seed mass ≈ 50 g).

1.4.1 Classical frequentist

Classical statistics, which are part of the broader *frequentist* paradigm, are the kind of statistics typically presented in introductory statistics classes. For a specific experimental procedure (such as drawing cards or flipping coins), you calculate the probability of a particular outcome, which is defined as *the long-run average frequency of that outcome in a sequence of repeated experiments*. Next you calculate a *p-value*, defined as the probability of that outcome *or any more extreme outcome* given a specified null hypothesis. If this so-called *tail probability* is small, then you reject the null hypothesis: otherwise, you fail to reject it. But you don't accept the alternative hypothesis if the tail probability is large, you just fail to reject the null hypothesis.

The frequentist approach to statistics (due to Fisher, Neyman and Pearson) is useful and very widely used, but it has some serious drawbacks — which are repeatedly pointed out by proponents of other statistical frameworks (Berger and Berry, 1988). It relies on the probability of a series of outcomes that didn't happen (the tail probabilities), and which depend on the way the experiment is defined; its definition of probability depends on a series of hypothetical repeated experiments that are often impossible in any practical sense; and it tempts us to construct straw-man null hypotheses and make convoluted arguments about why we have failed to reject them. Probably the most criticized aspect of frequentist statistics is their reliance on p-values, which when misused (as frequently occurs) are poor tools for scientific inference. It seems to be human nature to abuse p-values, acting as though alternative hypotheses (which are usually what we're really interested in) are "true" if we can reject the null hypothesis with $p < 0.05$ and "false" if we can't. In fact, when the null hypothesis is true we still find $p \leq 0.05$ one time in twenty (we falsely reject the null hypothesis 5% of the time, by definition). If $p > 0.05$ the null hypothesis could still be false but we have insufficient data to reject it. We could also reject the null hypothesis, in cases where we have lots of data, even though the results are biologically insignificant — that is, if the estimated effect size is ecologically irrelevant (e.g. a 0.01% increase in plant growth rate with a 30°C increase in temperature). More fundamentally, if we use a so-called *point null hypothesis* (such as "the slope of the relationship between plant productivity and temperature is zero"), common sense tells us that the null hypothesis *must* be false, because it can't be exactly zero — which makes the p value into a statement about whether we have enough data to detect a non-zero slope, rather than about whether the slope is actually different from zero. Working statisticians will tell you that it is better to focus on estimating the values of biologically meaningful parameters and finding their confidence limits rather than worrying too much about whether p is greater or less than 0.05 (Yoccoz, 1991; Johnson, 1999; Osenberg et al., 2002) — although Stephens et al. (2005) remind us that hypothesis testing can still be useful.

Looking at the seed data, we have the following 2 × 2 table:

	pol	psd
any taken (t)	26	25
none taken	184	706
total (N)	210	731

If t_i is the number of times that species i seeds disappear and N_i is the total number of observations of species i then the observed proportions of the time that seeds disappeared for each species are (pol) $t_1/N_1 = 0.124$ and (psd) $t_2/N_2 = 0.034$. The overall proportion taken (which is not the average of the two proportions since there are different total numbers of observations for each species) is $(t_1 + t_2)/(N_1 + N_2) = 0.054$. The ratio of the predation probabilities (proportion for pol/proportion for psd) is $0.124/0.034 = 3.62$. The ecological question we want to answer is "is there differential predation on the seeds on these two species?" (Given the sample sizes and the size of the observed difference, what do you think? Do you think the answer is likely to be statistically significant? How about biologically significant? What assumptions or preconceptions does your answer depend on?)

A frequentist would translate this biological question into statistics as "what is the probability that I would observe a result this extreme, or more extreme, given the sampling procedure?" More specifically, "what proportion of possible outcomes would result in observed ratios of proportions greater than 3.62, *or* smaller than $1/3.62 = 0.276$?" (Figure 1.1). Fisher's exact test (`fisher.test` in R) calculates this probability, as a one-tailed test (proportion of outcomes with ratios greater than 3.62) or a two-tailed test (proportion with ratios greater than 3.62 or less than its reciprocal, 0.276); the two-tailed answer in this case is 5.26×10^{-6}. According to Fisher's original interpretation, this number represents the strength of evidence against the null hypothesis, or (loosely speaking) for the alternative hypothesis — that there is a difference in seed predation rates. According to the Neyman-Pearson decision rule, if we had set our acceptance cutoff at $\alpha = 0.05$, we

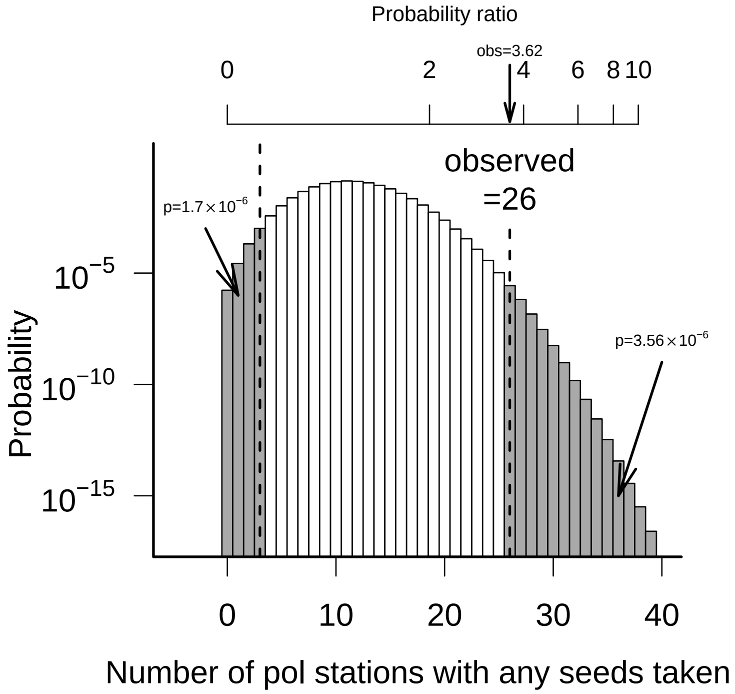

Figure 1.1 Classical frequentist analysis. Fisher's exact test calculates the probability of a given number of `pol` stations having seeds taken under the null hypothesis that both species have the same predation probability. The total probability that as many or more `pol` stations had seeds taken, *or* that the difference was more extreme in the other direction, is the two-tailed frequentist p-value ($3.56 \times 10^{-6} + 1.70 \times 10^{-6} = 5.26 \times 10^{-6}$). The top axis shows the equivalent in seed predation probability ratios. (*Note*: I put the y-axis on a log scale because the tails of the curve are otherwise too small to see, even though this change means that the area under the curve no longer represents the total probability.)

could conclude that there was a *statistically significant* difference in predation rates.

We needn't fixate on p-values: the R command for Fisher's test, `fisher.test`, also tells us the 95% confidence limits for the difference between rates*. In terms of probability ratios, this example gives (2.073, 6.057), which as expected does not include 1. Do you think a range of a 110% to a 510% increase in seed predation probability† is significant?

*R expresses the difference in predation rates in terms of the *odds ratio* — if there are t_1 seeds taken and $N_1 - t_1$ seeds not taken for species 1, then the odds of a seed being taken are $t_1/(N_1 - t_1)$ and the odds ratio between the species is $(t_1/(N_1 - t_1))/(t_2/(N_2 - t_2))$. The odds ratio and its logarithm (the *logit* or log-odds ratio) have nice statistical properties.

†These values are the confidence limits on the probability ratios, minus 1, converted into

1.4.2 Likelihood

Most of the book will focus on frequentist statistics, but not the standard version that you may be used to. Most modern statistics uses an approach called *maximum likelihood estimation*, or approximations to it. For a particular statistical model, maximum likelihood finds the set of parameters (e.g. seed removal rates) *that makes the observed data* (e.g. the particular outcomes of predation trials) *most likely to have occurred.* Based on a model for both the deterministic and stochastic aspects of the data, we can compute the *likelihood* (the probability of the observed outcome) given a particular choice of parameters. We then find the set of parameters that makes the likelihood as large as possible, and take the resulting *maximum likelihood estimates* (MLEs) as our best guess at the parameters. So far we haven't assumed any particular definition of probability of the parameters. We could decide on confidence limits by choosing a likelihood-based cutoff, for example by saying that any parameters that make the probability of the observed outcomes at least 1/10 as likely as the maximum likelihood are "reasonable". For mathematical convenience, we often work with the logarithm of the likelihood (the *log-likelihood*) instead of the likelihood; the parameters that give the maximum log-likelihood also give the maximum likelihood. On the log scale, statisticians have suggested a cutoff of 2 log-likelihood units (Edwards, 1992), meaning that we consider any parameter reasonable that is at least $e^{-2} \approx 1/7.4 = 14\%$ as likely as the maximum likelihood.

However, most modelers add a frequentist interpretation to likelihoods, using a mathematical proof that says that, across the hypothetical repeated trials of the frequentist approach, the distribution of the negative logarithm of the likelihood itself follows a χ^2 ("chi-squared") distribution[*]. This fact means that we can set a cut-off for differences in log-likelihoods based on the 95th percentile of the χ^2 distribution, which corresponds to 1.92 log-likelihood units, or parameters that lower the likelihood by a factor of 6.82. The theory says that the estimated value of the parameter will fall farther away than that from the true value only 5% of the time in a long series of repeated experiments. This rule is called the *Likelihood Ratio Test* (LRT)[†]. We will see that it lets us both estimate confidence limits for parameters and choose between competing models.

Bayesians also use the likelihood — it is part of the recipe for com-

approximate percentages: for example, a probability ratio of 1.1 would represent a 10% increase in predation.

[*] This result holds in the *asymptotic* case where we have lots of data, which happens less than we would like — but we often gloss over the fact of limited data and use it anyway.

[†] The difference between log-likelihoods is equivalent to the ratio of likelihoods.

puting the posterior distribution — but they take it as a measure of the information we can gain from the data, without saying anything about what the distribution of the likelihood would be in repeated trials.

How would one apply maximum likelihood estimation to the seed predation example? Lumping all the data from both species together at first, and assuming that (1) all observations are independent of each other and (2) the probability of at least one seed being taken is the same for all observations, it follows that the number of times at least one seed is removed is *binomially* distributed (we'll get to the formulas in Chapter 4). Now we want to know how the probability of observing the data (the likelihood, $\mathcal{L}$) depends on the probability p_s that at least one seed was taken from a particular station by a predator*, and what value of p_s maximizes the likelihood. The likelihood $\mathcal{L}$ is the probability that seeds were taken in 51 out of the total of 941 observations. This probability varies as a function of p_s (Figure 1.2): for $p_s = 0.05$, $\mathcal{L} = 0.048$, while for $p_s = 0.04$, $\mathcal{L}$ is only 6.16×10^{-3}. As it turns out, the MLE for the probability that seeds were taken in any one trial (p_s) is exactly what we'd expect—51/941, or 0.054—and the likelihood is $\mathcal{L} = 0.057$. (This likelihood is small, but it just means that the probability of any *particular* outcome — seeds being taken in 51 trials rather than 50 or 52 — is small.)

To answer the questions that really concern us about the different predation probabilities for different species, we need to allow different probabilities for each species, and see how much better we can do (how much higher the likelihood is) with this more complex model. Now we take the separate values for each species (26 out of 210 and 25 out of 731) and, for a per-observation probability for each species, compute the likelihoods of each species' data and multiply them (see Chapter 4 for basic probability calculations), or add the log-likelihoods. If I define the model in terms of the probability for psd and the ratio of the probabilities, I can plot a *likelihood profile* for the maximum likelihood I can get for a given value of the ratio (Figure 1.3).

The conclusions from this frequentist, maximum-likelihood analysis are essentially identical to those of the classical frequentist (Fisher's exact test) analyses. The maximum-likelihood estimate equals the observed ratio of the probabilities, 3.62; the confidence limits are (2.13, 6.16), which do not include 1; and the LRT-based p-value for rejecting the null hypothesis that the probabilities are the same is 3.83×10^{-6}.

*One of the most confusing things about maximum likelihood estimation is that there are so many different probabilities floating around. The likelihood $\mathcal{L}$ is the probability of observing the complete data set (i.e., Prob(seeds were taken 51 times out of 941 observations)); p_s is the probability that seeds were taken in any given trial; and the frequentist p-value is the probability, given a particular value of p_s, that seeds were taken 51 *or more* times out of 941 observations.

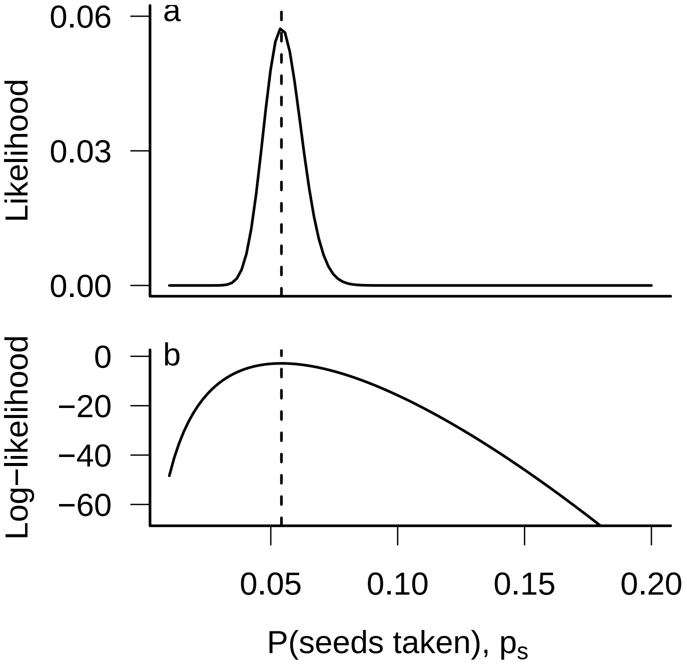

Figure 1.2 Likelihood and log-likelihood curves for predation probability p. Both curves have their maxima at the same point ($p = 0.054$). Log-likelihoods are based on natural ($\log_e$ or ln) logarithms.

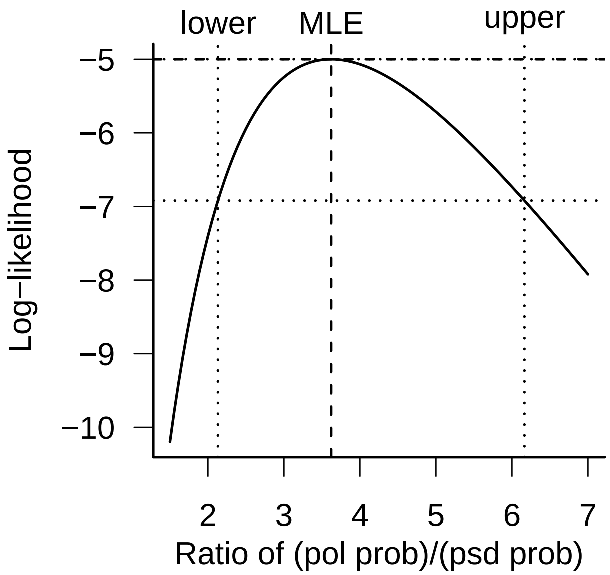

Figure 1.3 Likelihood curve for the ratio of the predation probabilities, showing the maximum likelihood estimate and 95% confidence limits. The null value (ratio equal to 1) is just below the lower limit of the graph.

Likelihood and classical frequentist analysis share the same philosophical underpinnings. Likelihood analysis is really a particular flavor of frequentist analysis, one that focuses on writing down a likelihood model and then testing for significant differences in the likelihood ratio rather than applying frequentist statistics directly to the observed outcomes. Classical analyses are usually easier because they are built into common statistics packages, and they may make fewer assumptions than likelihood analyses (for example, Fisher's test is exact while the LRT is only valid for large data sets), but likelihood analyses are often better matched with ecological questions.

1.4.3 Bayesian

Frequentist statistics assumes that there is a "true" state of the world (e.g. the difference between species in predation probability) which gives rise to a distribution of possible experimental outcomes. The Bayesian framework says instead that the experimental outcome — what we actually saw happen — is the truth, while the parameter values or hypotheses have probability distributions. The Bayesian framework solves many of the conceptual problems of frequentist statistics: answers depend on what we actually saw and not on a range of hypothetical outcomes, and we can legitimately make statements about the probability of different hypotheses or parameter values.

The major fly in the ointment of Bayesian statistics is that in order to make it work we have to specify our *prior beliefs* about the probability of different hypotheses, and these prior beliefs actually affect our answers! One hard-core frequentist ecologist says "Bayesianism means never having to say you're wrong" (Dennis, 1996). It is indeed possible to cheat in Bayesian statistics by setting unreasonably strong priors*. The standard solution to the problem of subjectivity is to assume you are completely ignorant before the experiment (setting a *flat prior*, or "letting the data speak for themselves"), although for technical reasons this isn't always possible. For better or worse, Bayesian statistics operates in the same way as we typically do science: we down-weight observations that are too inconsistent with our current beliefs, while using those in line with our current beliefs to strengthen and sharpen those beliefs (statisticians are divided on whether this is good or bad).

The big advantages of Bayesian statistics, besides their ease of interpretation, come (1) when we actually have data from prior observations we

*But if you really want to cheat with statistics you can do it in any framework!

want to incorporate; (2) in complex models with missing data and several layers of variability; (3) when we are trying to make (e.g.) management decisions based on our data (the Bayesian framework makes it easier to incorporate the effect of unlikely but catastrophic scenarios in decision-making). The only big disadvantage (besides the problem of priors) is that problems of small to medium complexity are actually harder with Bayesian approaches than with frequentist approaches — at least in part because most statistical software is geared toward classical statistics.

How would a Bayesian answer our question about predation rates? First of all, they would say (without looking at the data) that the answer is "yes" — the true difference between predation rates is certainly not zero. (This discrepancy reflects the difference in perspective between frequentists, who believe that the true value is a fixed number and uncertainty lies in what you observe [or might have observed], and Bayesians, who believe that observations are fixed numbers and the true values are uncertain.) Then they might define a parameter, the ratio of the two proportions, and ask questions about the *posterior distribution* of that parameter—our best estimate of the probability distribution given the observed data and some prior knowledge of its distribution (see Chapter 4). What is the mode (most probable value) of that distribution? What is its expected value, or mean? What is the *credible interval*, which is the interval with equal probability cutoffs below and above the mean within which 95% of the probability falls?

The Bayesian answers, in a nutshell: using a flat prior distribution, the mode is 3.48 (near the observed proportion of 3.62). The mean is 3.87, slightly larger than the mode since the posterior probability density is slightly asymmetric — the density is skewed to the right (Figure 1.4)*. The 95% credible interval, from 2.01 to 6.01, doesn't include 1, so a Bayesian would say that there was good evidence against the hypothesis: even more strongly, they could say that the probability that the predation ratio is greater than 1 is 0.998 (the probability that it is less than 1 is 0.002).

If the details of Bayesian statistics aren't perfectly clear at this point, don't worry. We'll explore Bayes' Rule and revisit Bayesian statistics in future chapters.

In this example all three statistical frameworks have given very similar answers, but they don't always. Ecological statisticians are still hotly debating which framework is best, or whether there is a single best framework. While it is important to be clear on the differences among the approaches,

*While Figure 1.1 showed the probability of each possible discrete outcome (number of seeds taken), Figure 1.4 shows a posterior probability *density* of a continuous parameter, i.e. the relative probability that the parameter lies in a particular range. Chapter 4 will explain this distinction more carefully.

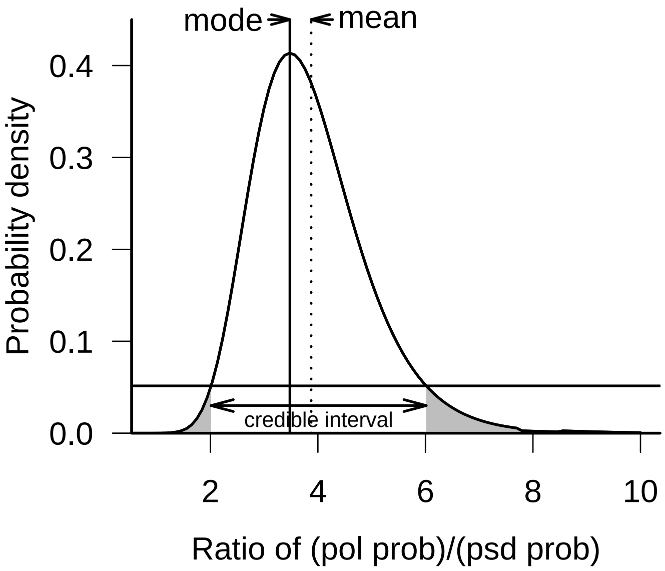

Figure 1.4 Bayesian analysis of seed predation. We calculate the probability density of the ratio of proportions of seeds taken being equal to some particular value, based on our prior (flat, assuming perfect ignorance — and in this case *improper* because it doesn't integrate to 1 [Chapter 4]) and on the data. The most probable value is the mode; the expected value is the mean. The gray shaded areas contain 5% of the area under the curve and cut off at the same height (probability density); the range between them is therefore the 95% credible interval.

knowing what question each is trying to answer, statisticians commonly move back and forth among them. My own approach is eclectic, agreeing with the advice of Crome (1997) and Stephens et al. (2005) to try to understand the strengths and weaknesses of several different approaches and use each one as appropriate.

We will revisit these frameworks in more detail later. Chapter 4 will cover Bayes' rule, which underpins Bayesian statistics; Chapters 6 and 7 will return to a much more detailed look at the practical details of maximum likelihood and Bayesian analysis. (Textbooks like Dalgaard (2003) cover classical frequentist approaches very well.)

1.5 FRAMEWORKS FOR COMPUTING

In order to construct your own models, you will need to learn some of the basics of statistical computing. There are many computer languages and modeling tools with built-in statistical libraries (MATLAB, Mathematica) and several statistics packages with serious programming capabilities (SAS, IDL). We will use a system called R that is both a statistics package and a computing language.

1.5.1 What is R?

R's developers call it a "language and environment for statistical computing and graphics". This awkward phrase gets at the idea that R is more than just a statistics package. R is closest in spirit to other higher-level modeling languages like MATLAB or MathCAD. It is a dialect of the S computing language, which was written at Bell Labs in the 1980s as a research tool in statistical computing. MathSoft, Inc. (now Insightful Corporation) bought the rights to S and developed it into S-PLUS, a commercial package with a graphical front-end. In the 1990s two New Zealand statisticians, Ross Ihaka and Robert Gentleman, re-wrote S from scratch, again as a research project. The re-written (and free) version became immensely popular and is now maintained by an international "core team" of about a dozen well-respected statisticians and computer scientists.

1.5.2 Why use R?

R is an extremely powerful tool. It is a full-fledged modern computer language with sophisticated data structures; it supports a wide range of computations and statistical procedures; it can produce graphics ranging from