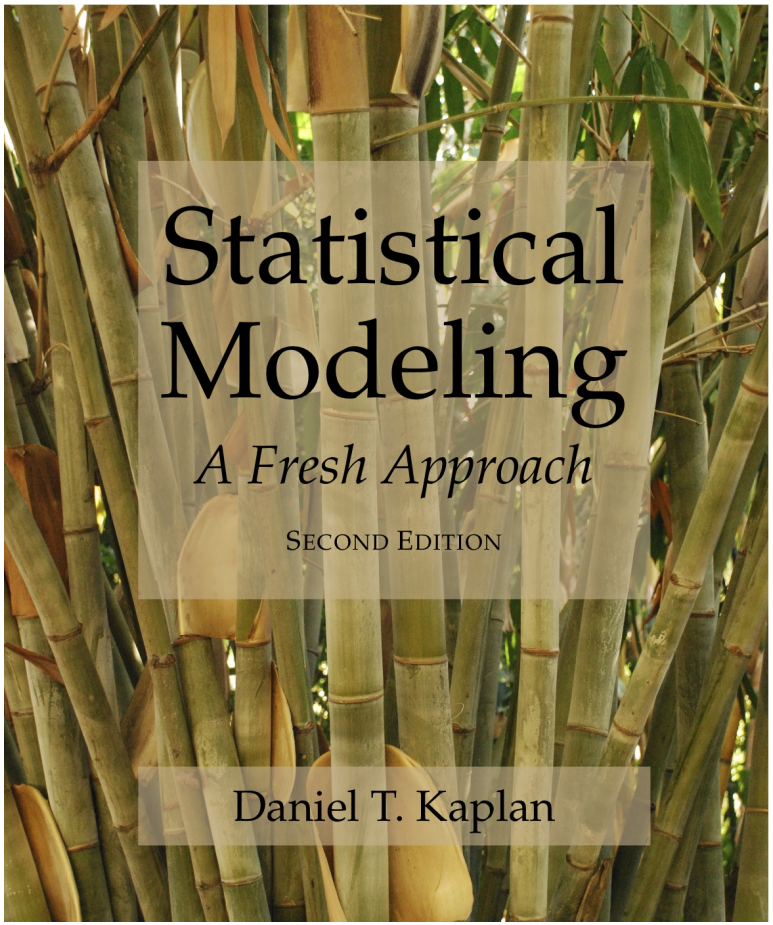


1 Introduction

All models are wrong. Some models are useful. — George Box

Art is a lie that tells the truth. — Pablo Picasso

This book is about **statistical modeling**, two words that themselves require some definition.

"Modeling" is a process of asking questions. "Statistical" refers in part to data — the statistical models you will construct will be rooted in data. But it refers also to a distinctively modern idea: that you can measure what you *don't* know and that doing so contributes to your understanding.

There is a saying, "A person with a watch knows the time. A person with two watches is never sure." The statistical point of view is that it's better not to be sure. With two watches you can see how they disagree with each other. This provides an idea of how precise the watches are. You don't know the time exactly, but knowing the precision tells you something about what you don't know. The non-statistical certainty of the person with a single watch is merely an uninformed self-confidence: the single watch provides no indication of what the person doesn't know.

The physicist Ernest Rutherford (1871-1937) famously said, "If your experiment needs statistics, you ought to have done a better experiment." In other words, if you can make a good enough watch, you need only one: no statistics. This is bad advice. Statistics never hurts. A person with two watches that agree perfectly not only knows the time, but has evidence that the watches are working at high precision. Sensibly, the official world time is based on an average of many atomic clocks. The individual clocks are fantastically precise; the point of averaging is to know when one or more of the clocks is drifting out of precision.

Why "statistical modeling" and not simply "statistics" or "data analysis?" Many

people imagine that data speak for themselves and that the purpose of statistics is to extract the information that the data carry. Such people see data analysis as an objective process in which the researcher should, ideally, have no influence. This can be true when very simple issues are involved; for instance, how precise is the average of the atomic clocks used to set official time or what is the difference in time between two events? But many questions are much more complicated; they involve many variables and you don't necessarily know what is doing what to what.[2]

The conclusions you reach from data depend on the specific questions you ask. Like it or not, the researcher plays an active and creative role in constructing and interrogating data. This means that the process involves some subjectivity. But this is not the same as saying anything goes. Statistical methods allow you to make objective statements about how the data answer your questions. In particular, the methods help you to know if the data show anything at all.

The word "modeling" highlights that your goals, your beliefs, and your current state of knowledge all influence your analysis of data. The core of the scientific method is the formation of hypotheses that can be tested and perhaps refuted by experiment or observation. Similarly, in statistical modeling, you examine your data to see whether they are consistent with the hypotheses that frame your understanding of the system under study.

Example 1.1: Grades A student is applying to law school. The schools she applies to ask for her class rank, which is based on the average of her college course grades.

A simple statistical issue concerns the precision of the grade-point average. This isn't a question of whether the average was correctly computed or whether the grades were accurately recorded. Instead, imagine that you could send two essentially identical students to essentially identical schools. Their grade-point averages might well differ, reflecting perhaps the grading practices of their different instructors or slightly different choices of subjects or random events such as illness or mishaps or the scheduling of classes. One way to think about this is that the students' grades are to some extent random, contingent on factors that are unknown or perhaps irrelevant to the students' capabilities.

How do you measure the extent to which the grades are random? There is no practical way to create "identical" students and observe how their grades differ. But you can look at the variation in a single student's grades — from class to class — and use this as an indication of the size of the random influence in each grade. From this, you can calculate the likely range of the random influences on the overall grade-point average.

Statistical models let you go further in interpreting grades. It's a common belief that there are easy- and hard-grading teachers and that a grade reflects not just the student's work but the teacher's attitude and practices. Statistical modeling provides a way to use data on grades to see whether teachers grade differently and to correct for these differences between teachers. Doing this involves some subtlety, for example taking into account the possibility that strong students take different courses than weaker students.

Example 1.2: Nitrogen Fixing Plants All plants need nitrogen to grow. Since nitrogen is the primary component of air, there is plenty around. But it's hard for plants to get nitrogen from the air; they get it instead from the soil. Some plants, like alder and soybean, support nitrogen-fixing bacteria in nodules on the plant roots. The plant creates a hospitable environment for the bacteria; the bacteria, by fixing nitrogen in the soil, create a good environment for the plant. In a word, symbiosis.

Biologist Michael Anderson is interested in how genetic variation in the bacteria influences the success with which they fix nitrogen. One can imagine using this information to breed plants and bacteria that are more effective at fixing nitrogen and thereby reducing the need for agricultural fertilizer.

Anderson has an promising early result. His extensive field studies indicate that different genotypes of bacteria fix nitrogen at different rates. Unfortunately, the situation is confusing since the different genotypes tend to populate different areas with different amounts of soil moisture, different soil temperatures, and so on. How can he untangle the relative influences of the genotype and the other environmental factors in order to decide whether the variation in genotype is genuinely important and worth further study?

Example 1.3: Sex Discrimination A large trucking firm is being audited by the government to see if the firm pays wages in a discriminatory way. The audit finds wage discrepancies between men and women for "office and clerical workers" but not for other job classifications such as technicians, supervisors, sales personnel, or "skilled craftworkers." It finds no discrepancies based on race.

A simple statistical question is whether the observed difference in average wages for men and women office and clerical workers is based on enough data to be reliable. In answering this question, it actually makes a difference what other groups the government auditors looked at when deciding to focus on sex discrimination in office and clerical workers.

Further complicating matters are the other factors that contribute to people's wages: the kind of job they have, their skill level, their experience. Statistical models can be used to quantify how these contributions and how they connect to one another. For instance, it turns out that men on average tend to have more job experience than women, and some or all of the men's higher average wages might be due to this.

Models can help you decide whether this potential explanation is plausible. For instance, if you see that both men's and women's wages increase with experience in the same way, you might be more inclined to believe that job experience is a legitimate factor rather than just a mask for discrimination.

1.1 Models and their Purposes

Many of the toys you played with as a child are models: dolls, balsa-wood airplanes with wind-up propellers, wooden blocks, model trains. But so are many serious objects of the adult world: architectural plans, bank statements, train schedules, the results of medical diagnostic tests, the signals transmitted by a telephone, the equations of physics, the genetic sequences used by biologists. There are too many to list.

What all models have in common is this:

> *A model is a representation for a particular purpose.*

A model might be a physical object or it might be an idea, but it always stands for something else: it's a representation. Dolls stand for babies and animals, architectural plans stand for buildings and bridges, a white blood-cell count stands for the function of the immune system.

When you create a model, you have (or ought to have) a purpose in mind. Toys are created for the entertainment and (sometimes) edification of children. The various kinds of toys — dolls, blocks, model airplanes and trains — have a form that serves this purpose. Unlike the things they represent, the toy versions are small, safe, and inexpensive.

Models always leave things out and get some things — many things — wrong. Architectural plans are not houses; you can't live in them. But they are easy to transport, copy, and modify. That's the point. Telephone signals — unlike the physical sound waves that they represent — can be transported over long distances and even stored. A train schedule tells you something important but it obviously doesn't reproduce every aspect of the trains it describes; it doesn't carry passengers.

Statistical models revolve around data. But even so, they are first and foremost models. They are created for a purpose. The intended use of a model should shape the appropriate form of the model and determines the sorts of data that can properly be used to build the model.

There are three main uses for statistical models. They are closely related, but distinct enough to be worth enumerating.

Description. Sometimes you want to describe the range or typical values of a quantity. For example, what's a "normal" white blood cell count? Sometimes you want to describe the relationship between things. Example: What's the relationship between the price of gasoline and consumption by automobiles?

Classification or prediction. You often have information about some observable traits, qualities, or attributes of a system you observe and want to draw conclusions about other things that you can't directly observe. For instance, you know a patient's white blood-cell count and other laboratory measurements and want to diagnose the patient's illness.

Anticipating the consequences of interventions. Here, you intend to do something: you are not merely an observer but an active participant in the system. For example, people involved in setting or debating public policy have to deal with questions like these: To what extent will increasing the tax on gasoline reduce consumption? To what extent will paying teachers more increase student performance?

The appropriate form of a model depends on the purpose. For example, a model that diagnoses a patient as ill based on an observation of a high number of white blood cells can be sensible and useful. But that same model could give absurd predictions about intervention: Do you really think that lowering the white blood cell count by bleeding a patient will make the patient better?

To anticipate correctly the effects of an intervention you need to get the direction of cause and effect correct in your models. But for a model used for classification or prediction, it may be unnecessary to represent causation correctly. Instead, other issues, e.g., the reliability of data, can be the most important. One of the thorniest issues in statistical modeling — with tremendous consequences for science, medicine, government, and commerce — is how you can legitimately draw conclusions about interventions from models based on data collected without performing these interventions.

1.2 Observation and Knowledge

How do you know what you know? How did you find it out? How can you find out what you don't yet know? These are questions that philosophers have addressed

for thousands of years. The views that they have expressed are complicated and contradictory.

From the earliest times in philosophy, there has been a difficult relationship between knowledge and observation. Sometimes philosophers see your knowledge as emerging from your observations of the world, sometimes they emphasize that the way you see the world is rooted in your innate knowledge: the things that are obvious to you.

This tension plays out on the pages of newspapers as they report the controversies of the day. Does the death penalty deter crime? Does increased screening for cancer reduce mortality?

Consider the simple, obvious argument for why severe punishment deters crime. Punishments are things that people don't like. People avoid what they don't like. If crime leads to punishment, then people will avoid committing crime.

Each statement in this argument seems perfectly reasonable, but none of them is particularly rooted in observations of actual and potential criminals. It's artificial — a learned skill — to base knowledge such as "people avoid punishment" on observation. It might be that this knowledge was formed by our own experiences, but usually the only explanation you can give is something like, "that's been my experience" or give one or two anecdotes.

When observations contradict opinions — opinions are what you think you know — people often stick with their opinions. Put yourself in the place of someone who believes that the death penalty really does deter crime. You are presented with accurate data showing that when a neighboring state eliminated the death penalty, crime did not increase. So do you change your views on the matter? A skeptic can argue that it's not just punishment but also other factors that influence the crime rate, for instance the availability of jobs. Perhaps a generally improving economic condition in the other state kept the crime rate steady even at a time when society is imposing lighter punishments.

It's difficult to use observation to inform knowledge because relationships are complicated and involve multiple factors. It isn't at all obvious how people can discover or demonstrate causal relationships through observation. Suppose one school district pays teachers well and another pays them poorly. You observe that the first district has better student outcomes than the second. Can you legitimately conclude that teacher pay accounts for the difference? Perhaps something else is at work: greater overall family wealth in the first district (which is what enabled them to pay teachers more), better facilities, smaller classes, and so on.

Historian Robert Hughes concisely summarized the difficulty of trying to use observation to discover causal relationships. In describing the extensive use of hanging in 18th and 19th century England, he wrote, "One cannot say whether public hanging did terrify people away from crime. Nor can anyone do so, until we can count crimes that were never committed." [3, p.35] To know whether

hanging did deter crime, you would need to observe a **counterfactual**, something that didn't actually happen: the crimes in a world without hanging. You can't observe counterfactuals. So you need somehow to generate observations that give you data on what happens for different levels of the causal variable.

A modern idea is the **controlled experiment**. In its simplest ideal form, a controlled experiment involves changing one thing — teacher pay, for example — while *holding everything else constant*: family wealth, facilities, etc.

The experimental approach to gaining knowledge has great success in medicine and science. For many people, experiment is the essence of all science. But experiments are hard to perform and sometimes not possible at all. How do you hold everything else constant? Partly for this reason, you rarely see reports of experiments when you read the newspaper, unless the article happens to be about a scientific discovery.

Scientists pride themselves on recording their observations carefully and systematically. Laboratories are filled with high-precision instrumentation. The quest for precision culminates perhaps in the physicist's fundamental quantities. For instance, the mass of the electron is reported as $9.10938215 \pm 0.00000045 \times 10^{-31}$ kg. The precision is about 1 part in twenty million.

Contrast this extreme precision with the humble speed measurements from a policeman's radar gun (perhaps a couple of miles or kilometers per hour — one part in 50) or the weight indicated on a bathroom scale (give or take a kilogram or a couple of pounds — about one part in 100 for an adult).

All such observations and measures are the stuff of **data**, the records of observations. Observations do not become data by virtue of high precision or expensive instrumentation or the use of metric rather than traditional units. For many purposes, data of low precision is used. An ecologist's count of the number of mating pairs of birds in a territory is limited by the ability to find nests. A national census of a country's population, conducted by the government can be precise to only a couple of percent. The physicists counting neutrinos in huge observatories buried under mountains to shield them from extraneous events waits for months for their results and in the end the results are precise to only one part in two.

The precision that is needed in data depends on the purpose for which the data will be used. The important question for the person using the data is whether the precision, whatever it be, is adequate for the purpose at hand. To answer this question, you need to know how to measure precision and how to compare this to a standard reflecting the needs of your task. The scientist with expensive instrumentation and the framer of social policy both need to deal with data in similar ways to understand and interpret the precision of their results.

It's common for people to believe that conclusions drawn from data apply to certain areas — science, economics, medicine — but aren't terribly useful in other areas. In teaching, for example, almost all decisions are based on "experience"

rather than observation. Indeed, there is often strong resistance to making formal observations of student progress as interfering with the teaching process.

This book is based on the idea that techniques for drawing valid conclusions from observations — data — are valuable for two groups of people. The first group is scientists and others who routinely need to use statistical methods to analyze experimental and other data.

The second group is everybody else. All of us need to draw conclusions from our experiences, even if we're not in a laboratory. It's better to learn how to do this in valid ways, and to understand the limitations of these ways, than to rely on an informal, unstated process of opinion formation. It may turn out that in any particular area of interest there are no useful data. In such situations, you won't be able to use the techniques. But at least you will know what you're missing. You may be inspired to figure out how to supply it or to recognize it when it does come along, and you'll be aware of when others are misusing data.

As you will see, the manner in which the data are collected plays a central role in what sorts of conclusions can be legitimately made; data do not always speak for themselves. You will also see that strongly supported statements about causation are difficult to make. Often, all you can do is point to an "association" or a "correlation," a weaker form of statement.

Statistics is sometimes loosely described as the "science of data." This description is apt, particularly when it covers both the collection and analysis of data, but it does not mean much until you understand what data are. That's the subject of the next chapter.

1.3 The Main Points of this Book

- Statistics is about variation. Describing and interpreting variation is a major goal of statistics.
- You can create empirical, mathematical descriptions not only of a single trait or variable but also of the relationships between two or more traits. Empirical means based on measurements, data, observations.
- Models let you split variation into components: "explained" versus "unexplained." How to measure the size of these components and how to compare them to one another is a central aspect of statistical methodology. Indeed, this provides a definition of statistics:

  Statistics is the explanation of variation in the context of what remains unexplained.

- By collecting data in ways that require care but are quite feasible, you can estimate how reliable your descriptions are, e.g., whether it's plausible that you should see similar relationships if you collected new data. This notion of reliability is very narrow and there are some issues that depend critically on the context in which the data were collected and the correctness of assumptions that you make about how the world works.
- Relationships between pairs of traits can be studied in isolation only in special circumstances. In general, to get valid results it is necessary to study entire systems of traits simultaneously. Failure to do so can easily lead to conclusions that are grossly misleading.
- Descriptions of relationships are often **subjective** — they depend on choices that you, the modeler, make. These choices are generally rooted in your own beliefs about how the world works, or the theories accepted as plausible within some community of inquiry.
- If data are collected properly, you can get an indication of whether the data are consistent or inconsistent with your subjective beliefs or — and this is important — whether you don't have enough data to tell either way.
- Models can be used to check out the sensitivity of your conclusions to different beliefs. People who disagree in their views of how the world works often may not be able to reconcile their differences based on data, but they will be able to decide objectively whether their own or the other party's beliefs are reasonable given the data.
- Notwithstanding everything said above about the strong link between your prior, subjective beliefs and the conclusions you draw from data, by collecting data in a certain context — experiments — you can dramatically simplify the interpretation of the results. It's actually possible to remove the dependence on identified subjective beliefs by intervening in the system under study experimentally.

This book takes a different approach than most statistics texts. Many people want statistics to be presented as a kind of automatic, algorithmic way to process data. People look for mathematical certainty in their conclusions. After all, there are right-or-wrong answers to the mathematical calculations that people (or computers) perform in statistics. Why shouldn't there be right-or-wrong answers to the conclusions that people draw about the world?

The answer is that there can be, but only when you are dealing with narrow circumstances that may not apply to the situations you want to study. An insistence on certainty and provable correctness often results in irrelevancy.

The point of view taken in this book is that it is better to be useful than to be provably certain. The objective is to introduce methods and ideas that can help you deal with drawing conclusions about the real world from data. The methods and ideas are meant to guide your reasoning; even if the conclusions you draw are not guaranteed by proof to be correct, they can still be more useful than the alternative, which is the conclusions that you draw without data, or the

conclusions you draw from simplistic methods that don't honor the complexity of the real system.

1.4 Introduction to Computation with R

Modern statistics is done on the computer. There was a time, 60 years ago and before, when computation could only be done by hand or using balky mechanical calculators. The methods of applied statistics developed during this time reflected what could be done using such calculators, not necessarily what was best for illuminating the system under study. These methods took on a life of their own — they became the operational definition of statistics. They continue to be taught today, using electronic calculators or personal computers or even just using paper and pencil. For the old statistical methods, computers are merely a labor saving device.

But not for modern statistics. The statistical methods at the core of this book cannot be applied in a authentic and realistic way without powerful computers. Thirty years ago, many of the methods could not be done at all unless you had access to the resources of a government agency or a large university. But with the revolutionary advances in computer hardware and numerical algorithms over the last half-century, modern statistical calculations can be performed on an ordinary home computer or laptop. (Even a cell phone has phenomenal computational power, often besting the mainframes of thirty years ago.) Hardware and software today pose no limitation; they are readily available.

Each chapter of this book includes a section on computational technique. Many readers will be completely new to the use of computers for scientific and statistial work, so the first chapters cover the foundations, techniques that are useful for many different aspects of computation. Working through the early chapters is essential for developing the skills that will be used later in actual statistical work. It will take a few hours, but this investment will pay off handsomely.

Chances are, you use a computer almost every day: for email, word-processing, managing your music or your photograph collection, perhaps even using a spreadsheet program for accounting. The software you use for such activities makes it easy to get started. Possibly you have never even looked at an instruction manual or used the "help" features on your computer.

When you use a word processor or email, the bulk of what you enter into the computer — the content of your documents and email — is without meaning to the computer. This is not at all to say that it is meaningless. Your documents and letters are intended for human readers; most of the work you do is directed so that the recipients can understand them. But the computer doesn't need to understand what you write in order to format it, print it, or transmit it over the Internet.

When doing scientific and statistical computing, things are different. What you enter into the computer is instructions to the computer to perform calculations and re-arrangements of data. Those instructions have to be comprehensible to the computer. If they make no sense or if they are inconsistent or ill formed, the computer won't be able to carry out your instructions. Worse, if the instructions make sense in some formal way but don't convey your actual intentions, the computer will perform some operation but the result will mislead you.

The difficulty with using software for mathematics and statistics is in making sure that your instructions make sense and do what you want them to do. This difficulty is not a matter of bad software design; it's intrinsic to the problem of communicating your intentions to the computer. The same difficulty would arise in word processing if the computer had to make sense of your writing, rejecting it when a claim is unconvincing or when a sentence is ambiguous. Statistical computing pioneer John Chambers refers to the "Prime Directive" of software[4]: "to program in such a way that computations can be understood and trusted."

Much of the design of software for scientific and statistical work is oriented around the difficulty of communicating intentions. A popular approach is based on the computer mouse: the program provides a list of possible operations — like the keys on a calculator — and lets the user choose which operation to apply to some selected data. This style of user interface is employed, for example, in spreadsheet software, letting users add up columns of numbers, make graphs, etc. The reason this style is popular is that it can make things extremely easy ... so long as the operation that you want has been included in the software. But things get very hard if you need to construct your own operation and it can be difficult to understand or trust the operations performed by others.

Another style of scientific computation — the one used in this book — is based on language. Rather than selecting an option with a mouse, you construct a **command** that conveys both the operation that you want and the data to which you want to apply that operation. There are dramatic advantages to this language-based style of computation:

- It lets you **connect** computations to one another, so that the output of one operation can become the input to another.
- It lets you **repeat** the operation on new or modified data, allowing you to automate tedious tasks and, importantly, to verify the correctness of your computations on data where you already know the answer.
- It lets you **accumulate** the results of previous operations, treating those results as new data.
- It lets you **document** concisely what was done so that you can demonstrate that what you said you did is what you actually did. In this way, you or others can repeat the analysis later if necessary to confirm your results.

- It lets you **modify** the computation in a controlled way to correct it or to vary some aspect of it while holding other aspects exactly the same as in the original.

In order to use the language-based approach, you will need to learn a few principles of the language itself: some vocabulary, some syntax, some grammar. This is much, much easier for the computer language than for a natural language like English or Chinese; it will take you only a couple of hours before you are fluent enough to do useful computations. In addition to letting you perform scientific computations in ways that use the computer and your own time and effort effectively, the principles that you will learn are broadly applicable to many computer systems and can provide significant insight even to how to use mouse-based interfaces.

1.4.1 The Choice of Software

The software package used in this book is called R. The R package provides an environment for doing statistical and scientific computation at a professional level. It was designed for statistics work, but suitable for other forms of scientific calculations and the creation of high-quality scientific graphics.[5]

There are several other major software packages widely used in statistics. Among the leaders are SPSS, SAS, and Stata. Each of them provides the computational power needed for statistical modeling. Each has its own advantages and its own devoted group of users.

One reason for the choice of R is that it offers a command-based computing environment. That makes it much easier to write about computing and also reveals better the structure of the computing process.[6] R is available for free and works on the major types of computers, e.g., Windows, Macintosh, and Unix/Linux. The RStudio software lets you work with a complete R system using an ordinary web browser or on your own computer.

In making your own choice, the most important thing is this: *choose something*! Readers who are familiar with SPSS, SAS, or STATA can use the information in each chapter's computational technique section to help them identify the facilities to look for in those packages.

Another form of software that's often used with data is the spreadsheet. Examples are Excel and Google Spreadsheets. Spreadsheets are effective for entering data and have nice facilities for formatting tables. The visual layout of the data seems to be intuitive to many people. Many businesses use spreadsheets and they are widely taught in high schools. Unfortunately, they are very difficult to use for statistical analyses of any sophistication. Indeed, even some very elementary statistical tasks such as making a histogram are difficult with spreadsheets and the results are usually unsatisfactory from a graphical point of view. Worse,

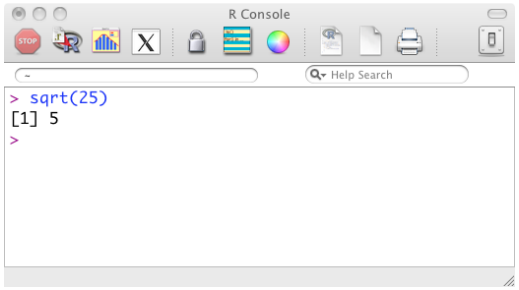


Figure 1.1: The R command console.

spreadsheets can be very hard to use reliably. There are lots of opportunities to make mistakes that will go undetected. As a result, despite the popularity of spreadsheets, I encourage you to reserve them for data entry and consider other software for the analysis of your data.

1.4.2 The R Command Console

Depending on your circumstances, you may prefer to install R as software on your own computer, or use a web-based ("cloud computing") service such as RStudio that runs through a web browser. If you are using this book with a course, your instructor may have set up a system for you. If you are on your own, follow the set-up instructions available on-line at `www.r-project.org`.

Depending on whether you run R on your own computer or in the cloud, you will start R either by clicking on an icon in the familiar way or by logging in to the web service. In either case, you will see a **console panel** in which you will be typing your commands to R. It will look something like Figure 1.1.

The R console gives you great power to carry out statistical and other mathematical and scientific operations. To use this power, you need to learn a little bit about the syntax and meaning of R commands. Once you have learned this, operations become simple to perform.

1.4.3 Invoking an Operation

People often think of computers as *doing* things: sending email, playing music, storing files. Your job in using a computer is to tell the computer *what* to do. There are many different words used to refer to the "what": a procedure, a task, a function, a routine, and so on. I'll use the word **computation**. Admittedly, this is a bit circular, but it is easy to remember: computers perform computations.

Complex computations are built up from simpler computations. This may seem

obvious, but it is a powerful idea. An **algorithm** is just a description of a computation in terms of other computations that you already know how to perform. To help distinguish between the computation as a whole and the simpler parts, it is helpful to introduce a new word: an **operator** performs a computation.

It's helpful to think of the computation carried out by an operator as involving four parts:

1. The name of the operator
2. The input arguments
3. The output value
4. Side effects

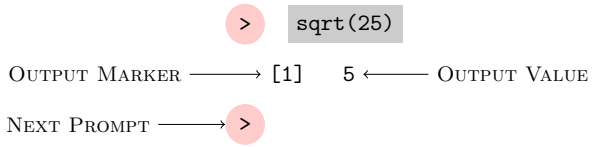
A typical operation takes one or more **input arguments** and uses the information in these to produce an **output value**. Along the way, the computer might take some action: display a graph, store a file, make a sound, etc. These actions are called **side effects**.

To tell the computer to perform a computation — call this **invoking an operation** or giving a **command** — you need to provide the name and the input arguments in a specific format. The computer then returns the output value. For example, the command `sqrt(25)` invokes the square root operator (named `sqrt`) on the argument 25. The output from the computation will, of course, be 5.

The syntax or invoking an operation consists of the operator's name, followed by round parentheses. The input arguments go inside the parentheses.

The software program that you use to invoke operators is called an **interpreter**. (The interpreter is the program you are running when you start R.) You enter your commands as a dialog between you and the interpreter. To start, the interpreter prints a prompt, after which you type your command:

When you press "Enter," the interpreter reads your command and performs the computation. For commands such as this one, the interpreter will print the output value from the computation:



The dialog continues as the interpreter prints another prompt and waits for your further command.

To save space, I'll usually show just the give-and-take from one round of the dialog:

```
> sqrt(25)
[1] 5
```

(Go ahead! Type `sqrt(25)` after the prompt in the R interpreter, press "enter," and see what happens.)

Often, operations involve more than one argument. The various arguments are separated by commas. For example, here is an operation named `seq` that produces a sequence of numbers:

```
> seq(3, 10)
[1]  3  4  5  6  7  8  9 10
```

The first argument tells where to start the sequence, the second tells where to end it.

The order of the arguments is important. Here is the sequence produced when 10 is the first argument and 3 the second:

```
> seq(10, 3)
[1] 10  9  8  7  6  5  4  3
```

For some operators, particularly those that have many input arguments, some of the arguments can be referred to by name rather than position. This is particularly useful when the named argument has a sensible default value. For example, the `seq` operator can be instructed how big a jump to take between successive items in the sequence. This is accomplished using an argument named `by`:

```
> seq(3, 10, by = 2)
[1] 3 5 7 9
```

Depending on the circumstances, all four parts of a operation need not be present. For example, the `date` operation returns the current time and date; no input arguments are needed.

```
> date()
[1] "Tue Aug 23 16:52:29 2011"
```

Note that even though there are no arguments, the parentheses are still used. Think of the pair of parentheses as meaning, "Do this."

Naming and Storing Values

Often the value returned by an operation will be used later on. Values can be stored for later use with the **assignment operator**. This has a different syntax that reminds the user that a value is being stored. Here's an example of a simple assignment:

```
> x = 16
```

This command has stored the value 16 under the name `x`. The syntax is always the same: an equal sign (`=`) with a name on the left and a value on the right.

Such stored values are called **objects**. Making an assignment to an object defines the object. Once an object has been defined, it can be referred to and used in later computations.

Notice that an assignment operation does not return a value or display a value. Its sole purpose is to have the side effects of defining the object and thereby storing a value under the object's name.

To refer to the value stored in the object, just use the object's name itself. For instance:

```
> x
[1] 16
```

Doing a computation on the value stored in an object is much the same:

```
> sqrt(x)
[1] 4
```

You can create as many objects as you like and give them names that remind you of their purpose. Some examples: `wilma`, `ages`, `temp`, `dog.houses`, `foo3`. There are some rules for object names:

- Use only letters and numbers and the two punctuation marks "dot" (`.`) and "underscore" (`_`).
- Do NOT use spaces anywhere in the name.
- A number or underscore cannot be the first character in the name.
- Capital letters are treated as distinct from lower-case letters. The objects named `wilma` and `Wilma` are different.

For the sake of readability, keep object names short. But if you really must have an object named something like `agesOfChildrenFromTheClinicalTrial`, feel free.

Objects can store all sorts of things, for example a sequence of numbers:

```
> x = seq(1, 7)
```

When you assign a new value to an existing object, as just done to `x`, the former value of that object is erased from the computer memory. The former value of `x` was 16, but after the above assignment command it is

```
> x
```



```
[1] 1 2 3 4 5 6 7
```

The value of an object is changed only *via* the assignment operator. Using an object in a computation does not change the value. For example, suppose you invoke the square-root operator on `x`:

```
> sqrt(x)
[1] 1.00 1.41 1.73 2.00 2.24 2.45 2.65
```

The square roots have been returned as a value, but this doesn't change the value of `x`:

```
> x
[1] 1 2 3 4 5 6 7
```

If you want to change the value of `x`, you need to use the assignment operator:

```
> x = sqrt(x)
> x
[1] 1.00 1.41 1.73 2.00 2.24 2.45 2.65
```

Aside. 1.1 Assignment vs Algebra

An assignment command like `x=sqrt(x)` can be confusing to people who are used to algebraic notation. In algebra, the equal sign describes a relationship between the left and right sides. So, $x = \sqrt{x}$ tells us about how the quantity x and the quantity $\sqrt{x}$ are related. Students are usually trained to "solve" such relationships, going through a series of algebraic steps to find values for x that are consistent with the mathematical statement. (For $x = \sqrt{x}$, the solutions are $x = 0$ and $x = 1$.) In contrast, the assignment command `x=sqrt(x)` is a way of replacing the previous values stored in `x` with new values that are the square root of the old ones.

Connecting Computations

The brilliant thing about organizing operators in terms of input arguments and output values is that the output of one operator can be used as an input to another. This lets complicated computations be built out of simpler ones.

For example, suppose you have a list of 10000 voters in a precinct and you want to select a random sample of 20 of them for a survey. The `seq` operator can be used to generate a set of 10000 choices. The `sample` operator can be used to select some of these choices at random.

One way to connect the computations is by using objects to store the intermediate outputs.

```
> choices = seq(1, 10000)
> sample(choices, 20)
```

```
 [1] 7196 5352 6734 2438 9370 4713 6481 3549  593 3294 7815 6967 5097 3820
[15] 9170 6637 8801 2304 9919 4102
```

You can also pass the output of an operator *directly* as an argument to another operator. Here's another way to accomplish exactly the same thing as the above.

```
> sample(seq(1, 10000), 20)
 [1]   10 3029 9507 1037 9562 4057  786 7542 6190 2248 9237 3089 3216 8473
[15] 4951 8816 4473 7651 6449 8676
```

Numbers and Arithmetic

The language has a concise notation for arithmetic that looks very much like the traditional one:

```
> 7 + 2
[1] 9
> 3 * 4
[1] 12
> 5/2
[1] 2.5
> 3 - 8
[1] -5
> -3
[1] -3
> 5^2
[1] 25
```

Arithmetic operators, like any other operators, can be connected to form more complicated computations. For instance,

```
> 8 + 4/2
[1] 10
```

To a human reader, the command `8+4/2` might seem ambiguous. Is it intended to be `(8+4)/2` or `8+(4/2)`? The computer uses unambiguous rules to interpret the expression, but it's a good idea for you to use parethesis so that you can make sure that what you intend is what the computer carries out:

```
> (8 + 4)/2
[1] 6
```

Traditional mathematical notation uses superscripts and radicals to indicate exponentials and roots, e.g., 3^2 or $\sqrt{3}$ or $\sqrt[3]{8}$. This special typography doesn't work well with an ordinary keyboard, so R and most other computer languages uses a different notation:


```
> 3^2
[1] 9
> sqrt(3)
[1] 1.73
> 8^(1/3)
[1] 2
```

There is a large set of mathematical functions: exponentials, logs, trigonometric and inverse trigonometric functions, etc. Some examples:

Traditional	Computer
e^2	`exp(2)`
$\log_e(100)$	`log(100)`
$\log_{10}(100)$	`log10(100)`
$\log_2(100)$	`log2(100)`
$\cos(\frac{\pi}{2})$	`cos(pi/2)`
$\sin(\frac{\pi}{2})$	`sin(pi/2)`
$\tan(\frac{\pi}{2})$	`tan(pi/2)`
$\cos^{-1}(-1)$	`acos(-1)`

Numbers can be written in **scientific notation**. For example, the "universal gravitational constant" that describes the gravitational attraction between masses is 6.67428×10^{-11} (with units meters-cubed per kilogram per second squared). In the computer notation, this would be written `G=6.67428e-11`. The Avogradro constant, which gives the number of atoms in a mole, is $6.02214179 \times 10^{23}$ per mole, or `6.02214178e23`.

The computer language does not directly support the recording of units. This is unfortunate, since in the real world numbers often have units and the units matter. For example, in 1999 the Mars Climate Orbiter crashed into Mars because the design engineers specified the engine's thrust in units of pounds, while the guidance engineers thought the units were newtons.

Computer arithmetic is accurate and reliable, but it often involves very slight rounding of numbers. Ordinarily, this is not noticeable. However, it can become apparent in some calculations that produce results that are zero. For example, mathematically $\sin(\pi) = 0$, however the computer does not duplicate this mathematical relationship exactly:

```
> sin(pi)
[1] 1.22e-16
```

Whether a number like this is properly interpreted as "close to zero," depends on the context and, for quantities that have units, on the units themselves. For instance, the unit "parsec" is used in astronomy in reporting distances between stars. The closest star to the sun is Proxima, at a distance of 1.3 parsecs. A

distance of 1.22×10^{-16} parsecs is tiny in astronomy but translates to about 2.5 meters — not so small on the human scale.

In statistics, many calculations relate to probabilities which are always in the range 0 to 1. On this scale, `1.22e-16` is very close to zero.

There are two "special" numbers. `Inf` stands for ∞, as in

```
> 1/0
[1] Inf
```

`NaN` stands for "not a number," and is the result when a numerical operation isn't defined, for instance

```
> 0/0
[1] NaN
```

Aside. 1.2 Complex Numbers

Mathematically oriented readers will wonder why R should have any trouble with a computation like $\sqrt{-9}$; the result is the imaginary number $3i$. R works with complex numbers, but you have to tell the system that this is what you want to do. To calculate $\sqrt{-9}$, use `sqrt(-9+0i)`.

Types of Objects

Most of the examples used so far have dealt with numbers. But computers work with other kinds of information as well: text, photographs, sounds, sets of data, and so on. The word **type** is used to refer to the kind of information.

Modern computer languages support a great variety of types. It's important to know about the types of data because operators expect their input arguments to be of specific types. When you use the wrong type of input, the computer might not be able process your command.

For the purpose of starting with R, it's important to distinguish among three basic types:

numeric The numbers of the sort already encountered.

data frames Collections of data more or less in the form of a spreadsheet table. The Computation Technique section in Chapter 2 introduces the operators for working with data frames.

character Text data.

You indicate character data to the computer by enclosing the text in double quotation marks. For example:

```
> filename = "swimmers.csv"
```


There is something a bit subtle going on in the above command, so look at it carefully. The purpose of the command is to create an object, named `filename`, that stores a little bit of text data. Notice that the name of the object is not put in quotes, but the text characters are.

Whenever you refer to an object name, make sure that you don't use quotes, for example:

```
> filename
[1] "swimmers.csv"
```

If you make a command with the object name in quotes, it won't be treated as referring to an object. Instead, it will merely mean the text itself:

```
> "filename"
[1] "filename"
```

Similarly, if you omit the quotation marks from around text, the computer will treat it as if it were an object name and will look for the object of that name. For instance, the following command directs the computer to look up the value contained in an object named `swimmers.csv` and insert that value into the object `filename`.

```
> filename = swimmers.csv
Error in try(swimmers.csv) : object 'swimmers.csv' not found
```

As it happens, there was no object named `swimmers.csv` because it had not been defined by any previous assignment command. So, the computer generated an error.

For the most part, you will not need to use very many operators on text data; you just need to remember to include text, such as file names, in quotation marks, `"like this"`.

Reading Questions

1. How can a model be useful even if it is not exactly correct?

2. Give an example of a model used for classification.

3. Often we describe personalities as "patient," "kind," "vengeful," etc. How can these descriptions be used as models for prediction?

4. Give three examples of models that you use in everyday life. For each, say what is the purpose of the model and in what ways the representation differs from the real thing.

5. Make a sensible statement about how precisely these quantities are typically measured:

- The speed of a car.
- Your weight.
- The national unemployment rate.
- A person's intelligence.

6. Give an example of a controlled experiment. What quantity or quantities have been varied and what has been held constant?

7. Using one of your textbooks from another field, pick an illustration or diagram. Briefly describe the illustration and explain how this is a model, in what ways it is faithful to the system being described and in what ways it fails to reflect that system.

For exercises, see `www.mosaic-web.org/StatisticalModeling/Exercises`.