Chapter 1

Variation

If there were no variation, if every observation were predictable, a mere repetition of what had gone before, there would be no need for statistics.

In this chapter, you'll learn what statistics is all about, variation and its potential sources, and how to use *R* to display the data you've collected. You'll start to acquire additional vocabulary, including such terms as accuracy and precision, mean and median, and sample and population.

1.1 VARIATION

We find physics extremely satisfying. In high school, we learned the formula $S = VT$, which in symbols relates the distance traveled by an object to its velocity multiplied by the time spent in traveling. If the speedometer says 60 mph, then in half an hour, you are certain to travel exactly 30 mi. Except that during our morning commute, the speed we travel is seldom constant, and the formula not really applicable. Yahoo Maps told us it would take 45 minutes to get to our teaching assignment at UCLA. Alas, it rained and it took us two and a half hours.

Politicians always tell us the best that can happen. If a politician had spelled out the worst-case scenario, would the United States have gone to war in Iraq without first gathering a great deal more information?

In college, we had Boyle's law, $V = KT/P$, with its tidy relationship between the volume V, temperature T and pressure P of a perfect gas. This is just one example of the perfection encountered there. The problem was we could never quite duplicate this (or any other) law in the Freshman Physics' laboratory. Maybe it was the measuring instruments, our lack of familiarity with the equipment, or simple measurement error, but we kept getting different values for the constant K.

By now, we know that variation is the norm. Instead of getting a fixed, reproducible volume V to correspond to a specific temperature T and pressure P, one ends up with a *distribution* of values of V instead as a result of errors in measurement. But we also know that with a large enough representative *sample* (defined later in this chapter), the center and shape of this distribution are reproducible.

Introduction to Statistics Through Resampling Methods and R, Second Edition. Phillip I. Good.

Here's more good and bad news: Make astronomical, physical, or chemical measurements and the only variation appears to be due to observational error. Purchase a more expensive measuring device and get more precise measurements and the situation will improve.

But try working with people. Anyone who spends any time in a schoolroom—whether as a parent or as a child, soon becomes aware of the vast differences among individuals. Our most distinct memories are of how large the girls were in the third grade (ever been beat up by a girl?) and the trepidation we felt on the playground whenever teams were chosen (not right field again!). Much later, in our college days, we were to discover there were many individuals capable of devouring larger quantities of alcohol than we could without noticeable effect. And a few, mostly of other nationalities, whom we could drink under the table.

Whether or not you imbibe, we're sure you've had the opportunity to observe the effects of alcohol on others. Some individuals take a single drink and their nose turns red. Others can't seem to take just one drink.

Despite these obvious differences, scheduling for outpatient radiology at many hospitals is done by a computer program that allots exactly 15 minutes to each patient. Well, I've news for them and their computer. Occasionally, the technologists are left twiddling their thumbs. More often the waiting room is overcrowded because of routine exams that weren't routine or where the radiologist wanted additional X-rays. (To say nothing of those patients who show up an hour or so early or a half hour late.)

The majority of effort in *experimental design*, the focus of Chapter 6 of this text, is devoted to finding ways in which this variation from individual to individual won't swamp or mask the variation that results from differences in treatment or approach. It's probably safe to say that what distinguishes statistics from all other branches of applied mathematics is that it is devoted to characterizing and then accounting for *variation* in the observations.

Consider the Following Experiment

You catch three fish. You heft each one and estimate its weight; you weigh each one on a pan scale when you get back to the dock, and you take them to a chemistry laboratory and weigh them there. Your two friends on the boat do exactly the same thing. (All but Mike; the chemistry professor catches him in the lab after hours and calls campus security. This is known as missing data.)

The 26 weights you've recorded (3 × 3 × 3–1 when they nabbed Mike) differ as result of measurement error, observer error, differences among observers, differences among measuring devices, and differences among fish.

1.2 COLLECTING DATA

The best way to observe variation is for you, the reader, to collect some data. But before we make some suggestions, a few words of caution are in order: 80% of the effort in any study goes into data collection and preparation for data collection. Any effort you don't expend initially goes into cleaning up the resulting mess. Or, as my carpenter friends put it, "measure twice; cut once."

We constantly receive letters and emails asking which statistic we would use to rescue a misdirected study. We know of no magic formula, no secret procedure known only to statisticians with a PhD. The operative phrase is GIGO: garbage in, garbage out. So think carefully before you embark on your collection effort. Make a list of possible sources of variation and see if you can eliminate any that are unrelated to the objectives of your study. If midway through, you think of a better method—don't use it.* Any inconsistency in your procedure will only add to the undesired variation.

1.2.1 A Worked-Through Example

Let's get started. Suppose we were to record the time taken by an individual to run around the school track. Before turning the page to see a list of some possible sources of variation, test yourself by writing down a list of all the factors you feel will affect the individual's performance. Obviously, the running time will depend upon the individual's sex, age, weight (for height and age), and race. It also will depend upon the weather, as I can testify from personal experience.

Soccer referees are required to take an annual physical examination that includes a mile and a quarter run. On a cold March day, the last time I took the exam in Michigan, I wore a down parka. Halfway through the first lap, a light snow began to fall that melted as soon as it touched my parka. By the third go around the track, the down was saturated with moisture and I must have been carrying a dozen extra pounds. Needless to say, my running speed varied considerably over the mile and a quarter.

As we shall see in the chapter on analyzing experiments, we can't just add the effects of the various factors, for they often interact. Consider that Kenyan's dominate the long-distance races, while Jamaicans and African-Americans do best in sprints.

The sex of the observer is also important. Guys and stallions run a great deal faster if they think a maiden is watching. The equipment the observer is using is also important: A precision stopwatch or an ordinary wrist watch? (See Table 1.1.)

Before continuing with your reading, follow through on at least one of the following data collection tasks or an equivalent idea of your own as we will be using the data you collect in the very next section:

1. **a.** Measure the height, circumference, and weight of a dozen humans (or dogs, or hamsters, or frogs, or crickets).

   b. Alternately, date some rocks, some fossils, or some found objects.

* On the other hand, we strongly recommend you do a thorough review *after* all your data have been collected and analyzed. You can and should learn from experience.

Table 1.1 Sources of Variation in Track Results

Test subject	Observer	Environment	Track
Sex	Sex	Wind	Surface
Age	Measuring device	Rain	Length
Weight/height	Experience	Sun	Condition
Experience		Temperature	
Race			
Current Health			
Desire			

2. Time some tasks. Record the times of 5–10 individuals over three track lengths (say, 50 m, 100 m, and a quarter mile). Since the participants (or trial subjects) are sure to complain they could have done much better if only given the opportunity, record at least two times for each study subject. (Feel free to use frogs, hamsters, or turtles in place of humans as runners to be timed. Or to replaces foot races with knot tying, bandaging, or putting on a uniform.)
3. Take a survey. Include at least three questions and survey at least 10 subjects. All your questions should take the form "Do you prefer A to B? Strongly prefer A, slightly prefer A, indifferent, slightly prefer B, strongly prefer B." For example, "Do you prefer Britney Spears to Jennifer Lopez?" or "Would you prefer spending money on new classrooms rather than guns?"

Exercise 1.1: Collect data as described in one of the preceding examples. Before you begin, write down a complete description of exactly what you intend to measure and how you plan to make your measurements. Make a list of all potential sources of variation. When your study is complete, describe what deviations you had to make from your plan and what additional sources of variation you encountered.

1.3 SUMMARIZING YOUR DATA

Learning how to adequately summarize one's data can be a major challenge. Can it be explained with a single number like the median or mean? The *median* is the middle value of the observations you have taken, so that half the data have a smaller value and half have a greater value. Take the observations 1.2, 2.3, 4.0, 3, and 5.1. The observation 3 is the one in the middle. If we have an even number of observations such as 1.2, 2.3, 3, 3.8, 4.0, and 5.1, then the best one can say is that the median or midpoint is a number (any number) between 3 and 3.8. Now, a question for you: what are the median values of the measurements you made in your first exercise?

Hopefully, you've already collected data as described in the preceding section; otherwise, face it, you are behind. Get out the tape measure and the scales. If you conducted time trials, use those data instead. Treat the observations for each of the three distances separately.

If you conducted a survey, we have a bit of a problem. How does one translate "I would prefer spending money on new classrooms rather than guns" into a number a computer can add and subtract? There is more one way to do this, as we'll discuss in what follows under the heading, "Types of Data." For the moment, assign the number 1 to "Strongly prefer classrooms," the number 2 to "Slightly prefer classrooms," and so on.

1.3.1 Learning to Use R

Calculating the value of a statistic is easy enough when we've only one or two observations, but a major pain when we have 10 or more. As for drawing graphs—one of the best ways to summarize your data—many of us can't even draw a straight line. So do what I do: let the computer do the work.

We're going to need the help of a programming language R that is specially designed for use in computing statistics and creating graphs. You can download that language without charge from the website http://cran.r-project.org/. Be sure to download the kind that is specific to your model of computer and operating system.

As you read through the rest of this text, be sure to have R loaded and running on your computer at the same time, so you can make use of the R commands we provide.

R is an *interpreter*. This means that as we enter the lines of a typical program, we'll learn on a line-by-line basis whether the command we've entered makes sense (to the computer) and be able to correct the line if we've made a typing error.

When we run R, what we see on the screen is an arrowhead

```
>
```

If we type 2 + 3 after and then press the enter key, we see

```
[1] 5.
```

This is because R reports numeric results in the form of a vector. In this example, the first and only element in this vector takes the value 5.

To enter the observations 1.2, 2.3, 4.0, 3 and 5.1, type

```
ourdata = c(1.2, 2.3, 4.0, 3, 5.1)
```

If you've never used a programming language before, let us warn you that R is very inflexible. It won't understand (or, worse, may misinterpret) both of the following:

```
ourdata = c(1.2 2.3 4.0 3 5.1)
ourdata = (1.2, 2.3, 4.0, 3, 5.1)
```

If you did type the line correctly, then typing median (ourdata) afterward will yield the answer 3 after you hit the enter key.

```
ourdata = c(1.2 2.3 4.0 3 5.1)
Error: syntax error
ourdata = c(1.2, 2.3, 4.0, 3, 5.1)
median(ourdata)
[1] 3
```

R Functions

median() is just one of several hundred built-in R functions.

You must use parentheses when you make use of an R function and you must spell the function name correctly.

```
> Median()
Error: could not find function “Median”
> median(Ourdata)
Error in median(Ourdata) : object ‘Ourdata’ not found
```

The median may tell us where the center of a distribution is, but it provides no information about the variability of our observations, and variation is what statistics is all about. Pictures tell the story best.*

The one-way *strip chart* (Figure 1.1) reveals that the *minimum* of this particular set of data is 0.9 and the *maximum* is 24.8. Each vertical line in this strip chart corresponds to an observation. Darker lines correspond to multiple observations. The *range* over which these observations extend is 24.8—0.9 or about 24.

Figure 1.2 shows a combination *box plot* (top section) and one-way strip chart (lower section). The “box” covers the middle 50% of the sample extending from the 25th to the 75th percentile of the distribution; its length is termed the *interquartile range*. The bar inside the box is located at the median or 50th *percentile* of the sample.

A weakness of this figure is that it’s hard to tell exactly what the values of the various percentiles are. A glance at the *box and whiskers plot* (Figure 1.3) made with R suggests the median of the classroom data described in Section 1.5 is about 153 cm, and the interquartile range (the “box”) is close to 14 cm. The minimum and maximum are located at the ends of the “whiskers.”

To illustrate the use of R to create such graphs, in the next section, we’ll use some data I gathered while teaching mathematics and science to sixth graders.

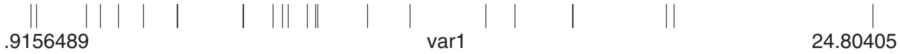


Figure 1.1 Strip chart.

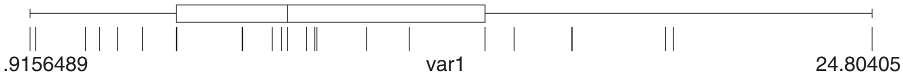


Figure 1.2 Combination *box plot* (top section) and one-way strip chart.

* The R code you’ll need to create graphs similar to Figures 1.1–1.3 is provided in Section 1.4.1.

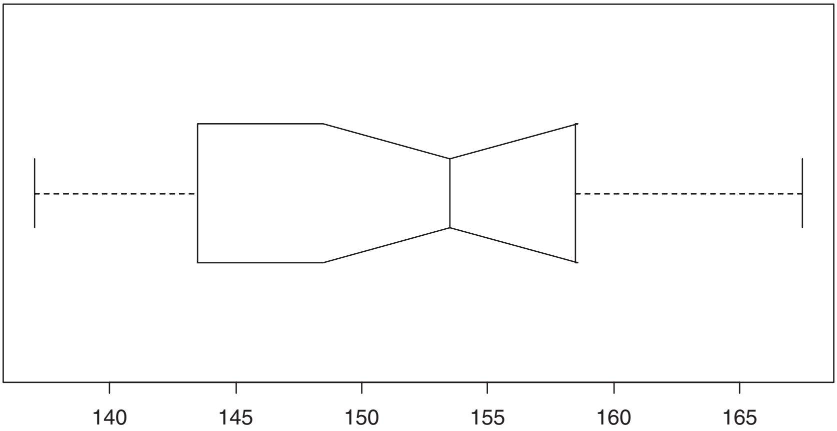


Figure 1.3 Box and whiskers plot of the classroom data.

1.4 REPORTING YOUR RESULTS

Imagine you are in the sixth grade and you have just completed measuring the heights of all your classmates.

Once the pandemonium has subsided, your instructor asks you and your team to prepare a report summarizing your results.

Actually, you have two sets of results. The first set consists of the measurements you made of you and your team members, reported in centimeters, 148.5, 150.0, and 153.0. (Kelly is the shortest incidentally, while you are the tallest.) The instructor asks you to report the minimum, the median, and the maximum height in your group. This part is easy, or at least it's easy once you look the terms up in the glossary of your textbook and discover that *minimum* means smallest, *maximum* means largest, and *median* is the one in the middle. Conscientiously, you write these definitions down—they could be on a test.

In your group, the minimum height is 148.5 cm, the median is 150.0 cm, and the maximum is 153.0 cm.

Your second assignment is more challenging. The results from all your classmates have been written on the blackboard—all 22 of them.

```
141, 156.5, 162, 159, 157, 143.5, 154, 158, 140, 142,
150, 148.5, 138.5, 161, 153, 145, 147, 158.5, 160.5,
167.5, 155, 137
```

You copy the figures neatly into your notebook computer. Using R, you store them in classdata using the command,

```
classdata = c(141, 156.5, 162, 159, 157, 143.5, 154,
158, 140, 142, 150, 148.5, 138.5, 161, 153, 145, 147,
158.5, 160.5, 167.5, 155, 137)
```

Next, you brainstorm with your teammates. Nothing. Then John speaks up—he's always interrupting in class. "Shouldn't we put the heights in order from smallest to largest?"

"Of course," says the teacher, "you should always begin by ordering your observations."

```
➢ sort(classdata)
[1] 137.0 138.5 140.0 141.0 142.0 143.5 145.0 147.0
148.5 150.0 153.0 154.0
[13] 155.0 156.5 157.0 158.0 158.5 159.0 160.5 161.0
162.0 167.5
```

In R, when the resulting output takes several lines, the position of the output item in the data set is noted at the beginning of the line. Thus, 137.0 is the first item in the ordered set classdata, and 155.0 is the 13th item.

"I know what the minimum is," you say—come to think of it, you are always blurting out in class, too, "137 millimeters, that's Tony."

"The maximum, 167.5, that's Pedro, he's tall," hollers someone from the back of the room.

As for the median height, the one in the middle is just 153 cm (or is it 154)? What does R say?

```
➢ median(classdata)
```

It is a custom among statisticians, honored by R, to report the median as the value midway between the two middle values, when the number of observations is even.

1.4.1 Picturing Data

The preceding scenario is a real one. The results reported here, especially the pandemonium, were obtained by my sixth-grade homeroom at St. John's Episcopal School in Rancho Santa Marguerite CA. The problem of a metric tape measure was solved by building their own from string and a meter stick.

My students at St. John's weren't through with their assignments. It was important for them to build on and review what they'd learned in the fifth grade, so I had them draw pictures of their data. Not only is drawing a picture fun, but pictures and graphs are an essential first step toward recognizing patterns.

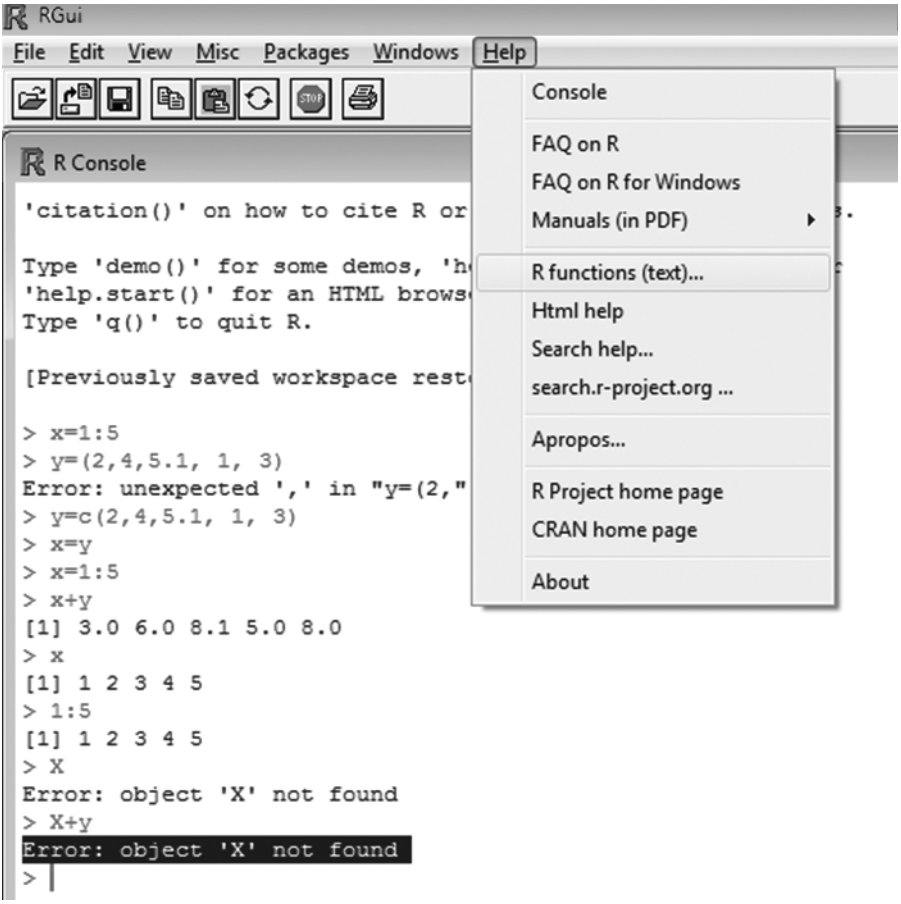
Begin by constructing both a strip chart and a box and whiskers plot of the classroom data using the R commands

```
➢ stripchart(classdata)
```

and

```
➢ boxplot(classdata)
```

All R plot commands have options that can be viewed via the R HELP menu. For example, Figure 1.4 was generated with the command



```
'citation()' on how to cite R or

Type 'demo()' for some demos, 'h
'help.start()' for an HTML brows
Type 'q()' to quit R.

[Previously saved workspace rest

> x=1:5
> y=(2,4,5.1, 1, 3)
Error: unexpected ',' in "y=(2,"
> y=c(2,4,5.1, 1, 3)
> x=y
> x=1:5
> x+y
[1] 3.0 6.0 8.1 5.0 8.0
> x
[1] 1 2 3 4 5
> 1:5
[1] 1 2 3 4 5
> X
Error: object 'X' not found
> X+y
Error: object 'X' not found
>
```

Figure 1.4 Getting help from R with using R.

```
➤ boxplot(classdata, notch=TRUE, horizontal =TRUE)
```

Generate a strip chart and a box plot for one of the data sets you gathered in your initial assignment. Write down the values of the median, minimum, maximum, 25th and 75th percentiles that you can infer from the box plot. Of course, you could also obtain these same values directly by using the R command, **quantile**(classdata**),** which yields all the desired statistics.

```
0%      25%     50%     75%     100%
137.000 143.875 153.500 158.375 167.500
```

One word of caution: R (like most statistics software) yields an excessive number of digits. Since we only measured heights to the nearest centimeter, reporting the 25th percentile as 143.875 suggests far more precision in our measurements than what actually exists. Report the value 144 cm instead.

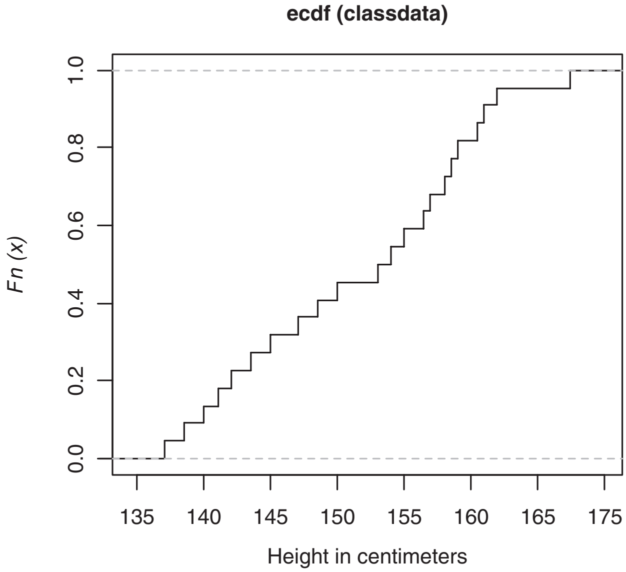


Figure 1.5 Cumulative distribution of heights of sixth-grade class.

A third way to depict the distribution of our data is via the histogram:

```
➢ hist(classdata)
```

To modify a histogram by increasing or decreasing the number of bars that are displayed, we make use of the "breaks" parameter as in

```
➢ hist(classdata, breaks = 4)
```

Still another way to display your data is via the *cumulative distribution function* **ecdf()**. To display the cumulative distribution function for the classdata, type

```
➢ plot(ecdf(classdata), do.points = FALSE, verticals =
  TRUE, xlab = "Height in Centimeters")
```

Notice that the *X*-axis of the cumulative distribution function extends from the minimum to the maximum value of your data. The *Y*-axis reveals that the probability that a data value is less than the minimum is 0 (you knew that), and the probability that a data value is less than the maximum is 1. Using a ruler, see what *X*-value or values correspond to 0.5 on the *Y*-scale (Figure 1.5).

Exercise 1.2: What do we call this *X*-value(s)?

Exercise 1.3: Construct histograms and cumulative distribution functions for the data you've collected.

1.4.2 Better Graphics

To make your strip chart look more like the ones shown earlier, you can specify the use of a vertical line as the character to be used in plotting the points:

```
➢ stripchart(classdata,pch = “|”)
```

And you can create a graphic along the lines of Figure 1.2, incorporating both a box plot and strip chart, with these two commands

```
➢ boxplot(classdata,horizontal = TRUE,xlab = “classdata”)
➢ rug(classdata)*
```

The first command also adds a label to the *x*-axis, giving the name of the data set, while the second command adds the strip chart to the bottom of the box plot.

1.5 TYPES OF DATA

Statistics such as the minimum, maximum, median, and percentiles make sense only if the data are *ordinal*, that is, if the data can be ordered from smallest to largest. Clearly height, weight, number of voters, and blood pressure are ordinal. So are the answers to survey questions, such as “How do you feel about President Obama?”

Ordinal data can be subdivided into metric and nonmetric data. *Metric* data or measurements like heights and weights can be added and subtracted. We can compute the mean as well as the median of metric data. (Statisticians further subdivide metric data into observations such as time that can be measured on a *continuous* scale and counts such as “buses per hour” that take *discrete* values.)

But what is the average of “he’s destroying our country” and “he’s no worse than any other politician?” Such preference data are ordinal, in that the data may be ordered, but they are *not* metric.

In order to analyze ordinal data, statisticians often will impose a metric on the data—assigning, for example, weight 1 to “Bush is destroying our country” and weight 5 to “Bush is no worse than any other politician.” Such analyses are suspect, for another observer using a different set of weights might get quite a different answer.

The answers to other survey questions are not so readily ordered. For example, “What is your favorite color?” Oops, bad example, as we can associate a metric wavelength with each color. Consider instead the answers to “What is your favorite breed of dog?” or “What country do your grandparents come from?” The answers to these questions fall into nonordered categories. Pie charts and bar charts are used to display such *categorical* data, and contingency tables are used to analyze it. A scatter plot of categorical data would not make sense.

Exercise 1.4: For each of the following, state whether the data are metric and ordinal, only ordinal, categorical, or you can’t tell:

a. Temperature
b. Concert tickets
c. Missing data
d. Postal codes.

[*] The rug() command is responsible for the tiny strip chart or rug at the bottom of the chart. Sometimes, it yields a warning message that can usually be ignored.

1.5.1 Depicting Categorical Data

Three of the students in my class were of Asian origin, 18 were of European origin (if many generations back), and one was part American Indian. To depict these categories in the form of a pie chart, I first entered the categorical data:

```
➢ origin = c(3,18,1)
➢ pie(origin)
```

The result looks correct, that is, *if* the data are in front of the person viewing the chart. A much more informative diagram is produced by the following R code:

```
➢ origin = c(3,18,1)
➢ names(origin) = c("Asian","European","Amerind")
➢ pie (origin, labels = names(origin))
```

All the graphics commands in R have many similar options; use R's help menu shown in Figure 1.4 to learn exactly what these are.

A pie chart also lends itself to the depiction of ordinal data resulting from surveys. If you performed a survey as your data collection project, make a pie chart of your results, now.

1.6 DISPLAYING MULTIPLE VARIABLES

I'd read but didn't quite believe that one's arm span is almost exactly the same as one's height. To test this hypothesis, I had my sixth graders get out their tape measures a second time. They were to rule off the distance from the fingertips of the left hand to the fingertips of the right while the student they were measuring stood with arms outstretched like a big bird. After the assistant principal had come and gone (something about how the class was a little noisy, and though we were obviously having a good time, could we just be a little quieter), they recorded their results in the form of a two-dimensional scatter plot.

They had to reenter their height data (it had been sorted, remember), and then enter their armspan data:

```
➢ classdata = c(141, 156.5, 162, 159, 157, 143.5,
154, 158, 140, 142, 150, 148.5, 138.5, 161, 153, 145,
147, 158.5, 160.5, 167.5, 155, 137)
➢ armspan = c(141, 156.5, 162, 159, 158,
143.5, 155.5, 160, 140, 142.5, 148, 148.5, 139,
160, 152.5, 142, 146.5, 159.5, 160.5, 164, 157,
137.5)
```

This is trickier than it looks, because unless the data are entered in exactly the same order by student in each data set, the results are meaningless. (We told you that 90% of the problem is in collecting the data and entering the data in the computer for analysis. In another text of mine, *A Manager's Guide to the Design and Conduct of Clinical Trials*, I recommend eliminating paper forms completely and entering all data directly into the computer.*) Once the two data sets have been read in, creating a scatterplot is easy:

```
➢ height = classdata
➢ plot(height, armspan)
```

Notice that we've renamed the vector we called classdata to reveal its true nature as a vector of heights.

Such plots and charts have several purposes. One is to summarize the data. Another is to compare different samples or different populations (girls versus boys, my class versus your class). For example, we can enter gender data for the students, being careful to enter the gender codes in the same order in which the students' heights and arm spans already have been entered:

```
➢ sex = c(“b”,rep(“g”,7),“b”,rep(“g”,6),rep(“b”,7))
```

Note that we've introduced a new R function, **rep**(), in this exercise to spare us having to type out the same value several times. The first student on our list is a boy; the next seven are girls, then another boy, six girls, and finally seven boys. R requires that we specify non-numeric or *character data* by surrounding the elements with quote signs. We then can use these gender data to generate side-by-side box plots of height for the boys and girls.

```
➢ sexf = factor(sex)
➢ plot(sexf,height)
```

The R-function **factor ()** tells the computer to treat gender as a categorical variable, one that in this case takes two values “b” and “g.” The **plot()** function will not work until character data have been converted to factors.

The primary value of charts and graphs is as aids to critical thinking. The figures in this specific example may make you start wondering about the uneven way adolescents go about their growth. The exciting thing, whether you are a parent or a middle-school teacher, is to observe how adolescents get more heterogeneous, more individual with each passing year.

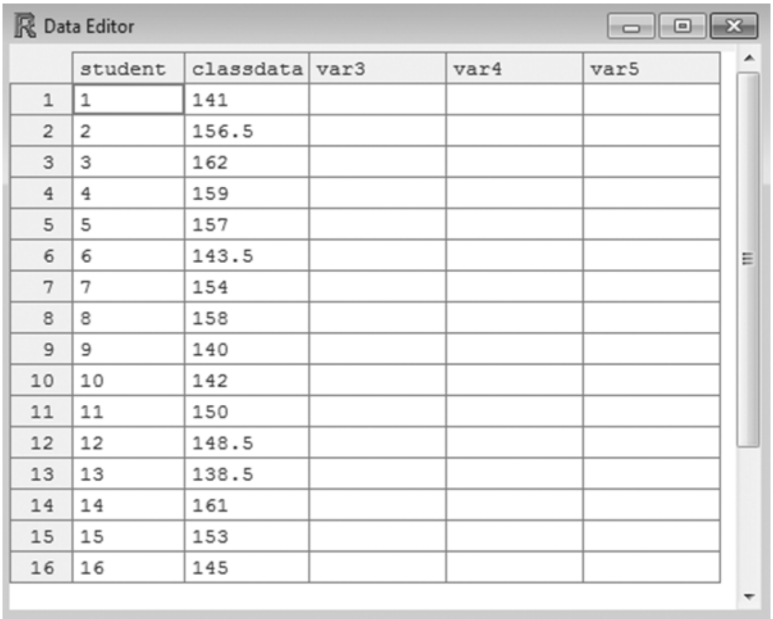
Exercise 1.5: Use the preceding R code to display and examine the indicated charts for my classroom data.

Exercise 1.6: Modify the preceding R code to obtain side-by-side box plots for the data you've collected.

1.6.1 Entering Multiple Variables

We've noted several times that the preceding results make sense *only* if the data are entered in the same order by student for each variable that you are interested in. R provides a simple way to achieve this goal. Begin by writing,

* We'll discuss how to read computer data files using R later in this chapter.



	student	classdata	var3	var4	var5
1	1	141			
2	2	156.5			
3	3	162			
4	4	159			
5	5	157			
6	6	143.5			
7	7	154			
8	8	158			
9	9	140			
10	10	142			
11	11	150			
12	12	148.5			
13	13	138.5			
14	14	161			
15	15	153			
16	16	145			

Figure 1.6 The R edit() screen.

```
➢ student = 1:length(height)
➢ classdata = data.frame(student, height)
➢ classdata = edit(classdata)
```

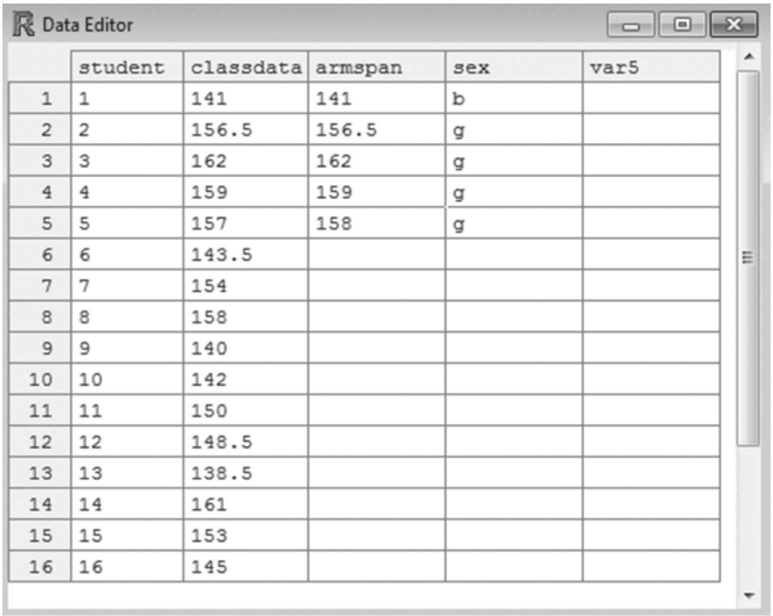
The upper left corner of R's screen will then resemble Figure 1.6. While still in edit mode, we add data for arm length and sex until the screen resembles Figure 1.7.

To rename variables, simply click on the existing name. Enter the variable's name, then note whether the values you enter for that variable are to be treated as numbers or characters.

1.6.2 From Observations to Questions

You may want to formulate your theories and suspicions in the form of questions: Are girls in the sixth-grade taller on the average than sixth-grade boys (not just those in my sixth-grade class, but in all sixth-grade classes)? Are they more homogenous, that is, less variable, in terms of height? What is the average height of a sixth grader? How reliable is this estimate? Can height be used to predict arm span in sixth grade? Can it be used to predict the arm spans of students of any age?

You'll find straightforward techniques in subsequent chapters for answering these and other questions. First, we suspect, you'd like the answer to one really big question: Is statistics really much more difficult than the sixth-grade exercise we just completed? No, this is about as complicated as it gets.



	student	classdata	armspan	sex	var5
1	1	141	141	b	
2	2	156.5	156.5	g	
3	3	162	162	g	
4	4	159	159	g	
5	5	157	158	g	
6	6	143.5			
7	7	154			
8	8	158			
9	9	140			
10	10	142			
11	11	150			
12	12	148.5			
13	13	138.5			
14	14	161			
15	15	153			
16	16	145			

Figure 1.7 The R edit() screen after entering additional observations.

1.7 MEASURES OF LOCATION

Far too often, we find ourselves put on the spot, forced to come up with a one-word description of our results when several pages, or, better still, several charts would do. "Take all the time you like," coming from a boss, usually means, "Tell me in ten words or less."

If you were asked to use a single number to describe data you've collected, what number would you use? One answer is "the one in the middle," the *median* that we defined earlier in this chapter. The median is the best statistic to use when the data are *skewed*, that is, when there are unequal numbers of small and large values. Examples of skewed data include both house prices and incomes.

In most other cases, we recommend using the *arithmetic mean* rather than the median.[*] To calculate the mean of a sample of observations by hand, one adds up the values of the observations, then divides by the number of observations in the sample. If we observe 3.1, 4.5, and 4.4, the arithmetic mean would be 12/3 = 4. In symbols, we write the mean of a sample of n observations, X_i with $i = 1, 2, \ldots, n$ as

$$(X_1 + X_2 + \cdots + X_n)/n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$ [†]

[*] Our reason for this recommendation will be made clear in Chapter 5.

[†] The Greek letter Σ is pronounced sigma.

Is adding a set of numbers and then dividing by the number in the set too much work? To find the mean height of the students in Dr. Good's classroom, use R and enter

```
➤ mean(height).
```

A playground seesaw (or teeter-totter) is symmetric in the absence of kids. Its midpoint or median corresponds to its center of gravity or its mean. If you put a heavy kid at one end and two light kids at the other so that the seesaw balances, the mean will still be at the pivot point, but the median is located at the second kid.

Another population parameter of interest is the most frequent observation or *mode*. In the sample 2, 2, 3, 4 and 5, the mode is 2. Often the mode is the same as the median or close to it. Sometimes it's quite different and sometimes, particularly when there is a mixture of populations, there may be several modes.

Consider the data on heights collected in my sixth-grade classroom. The mode is at 157.5 cm. But aren't there really two modes, one corresponding to the boys, the other to the girls in the class? As you can see on typing the command

```
➤ hist(classdata, xlab = "Heights of Students in Dr.
  Good's Class (cms)")
```

a histogram of the heights provides evidence of two modes. When we don't know in advance how many subpopulations there are, modes serve a second purpose: to help establish the number of subpopulations.

> **Exercise 1.7:** Compare the mean, median, and mode of the data you've collected (Figure 1.8).

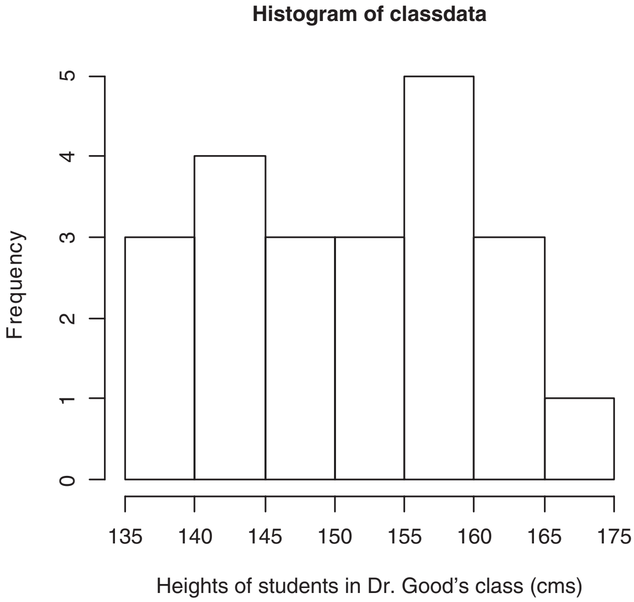


Figure 1.8 Histogram of heights of sixth-grade students.

Exercise 1.8: A histogram can be of value in locating the modes when there are 20 to several hundred observations, because it groups the data. Use R to draw histograms for the data you've collected.

1.7.1 Which Measure of Location?

The arithmetic mean, the median, and the mode are examples of sample *statistics*. Statistics serve three purposes:

1. Summarizing data
2. Estimating population parameters
3. Aids to decision making.

Our choice of one statistic rather than another depends on the use(s) to which it is to be put.

The Center of a Population

Median. The value in the middle; the halfway point; that value which has equal numbers of larger and smaller elements around it.

Arithmetic Mean or Arithmetic Average. The sum of all the elements divided by their number or, equivalently, that value such that the sum of the deviations of all the elements from it is zero.

Mode. The most frequent value. If a population consists of several subpopulations, there may be several modes.

For summarizing data, graphs—box plots, strip plots, cumulative distribution functions, and histograms are essential. If you're not going to use a histogram, then for samples of 20 or more, be sure to report the number of modes.

We always recommend using the median in two instances:

1. If the data are ordinal but not metric.
2. When the distribution of values is highly skewed with a few very large or very small values.

Two good examples of the latter are incomes and house prices. A recent *LA Times* featured a great house in Beverly Hills at US$80 million. A house like that has a large effect on the mean price of homes in an area. The median house price is far more representative than the mean, even in Beverly Hills.

The weakness of the arithmetic mean is that it is too easily biased by extreme values. If we eliminate Pedro from our sample of sixth graders—he's exceptionally tall for his age at 5′7″—the mean would change from 151.6 to 3167/21 = 150.8 cm. The median would change to a much lesser degree, shifting from 153.5 to 153 cm.

Because the median is not as readily biased by extreme values, we say that the median is more *robust* than the mean.

*1.7.2 The Geometric Mean

The *geometric mean* is the appropriate measure of location when we are expressing changes in percentages, rather than absolute values. The geometric mean's most common use is in describing bacterial and viral populations.

Here is another example: If in successive months the cost of living was 110, 105, 110, 115, 118, 120, 115% of the value in the base month, set

```
➢ ourdata = c(1.1,1.05,1.1,1.15,1.18,1.2,1.15)
```

The geometric mean is given by the following R expression:

```
➢ exp(mean(log(ourdata))
```

Just be sure when you write expressions as complicated as the expression in the line above, that the number of right parentheses (matches the number of left parentheses).

For estimation: In deciding which *sample statistic* to use in estimating the corresponding *population parameter*, we must distinguish between precision and accuracy. Let us suppose Robin Hood and the Sheriff of Nottingham engage in an archery contest. Each is to launch three arrows at a target 50 m (half a soccer pitch) away. The Sheriff launches first and his three arrows land one atop the other in a dazzling display of shooting *precision*. Unfortunately all three arrows penetrate and fatally wound a cow grazing peacefully in the grass nearby. The Sheriff's *accuracy* leaves much to be desired.

1.7.3 Estimating Precision

We can show mathematically that for very large samples, the sample median and the population median will almost coincide. The same is true for large samples and the mean. Alas, "large" may mean larger than we can afford to examine. With small samples, the accuracy of an estimator is always suspect.

With most of the samples we encounter in practice, we can expect the value of the sample median and virtually any other estimator to vary from sample to sample. One way to find out for small samples how *precise* a method of estimation is would be to take a second sample the same size as the first and see how the estimator varies between the two. Then a third, and fourth, . . . , say, 20 samples. *But a large sample will* always *yield more precise results than a small one*. So, if we'd been able to afford it, the sensible thing would have been to take 20 times as large a sample to begin with.*

Still, there is an alternative. We can treat our sample as if it were the original population and take a series of *bootstrap samples* from it. The variation in the value

* Of course, there is a point at which each additional observation will cost more than it yields in information. The bootstrap described here will also help us to find the "optimal" sample size.

of the estimator from bootstrap sample to bootstrap sample will be a measure of the variation to be expected in the estimator had we been able to afford to take a series of samples from the population itself. The larger the size of the original sample, the closer it will be in composition to the population from which it was drawn, and the more accurate this measure of precision will be.

1.7.4 Estimating with the Bootstrap

Let's see how this process, called bootstrapping, would work with a specific set of data. Once again, here are the heights of the 22 students in Dr. Good's sixth grade class, measured in centimeters and ordered from shortest to tallest:

```
137.0 138.5 140.0 141.0 142.0 143.5 145.0 147.0 148.5
150.0 153.0 154.0 155.0 156.5 157.0 158.0 158.5 159.0
160.5 161.0 162.0 167.5
```

Let's assume we record each student's height on an index card, 22 index cards in all. We put the cards in a big hat, shake them up, pull one out, and make a note of the height recorded on it. We *return the card to the hat* and repeat the procedure for a total of 22 times until we have a second sample, the same size as the original. Note that we may draw Jane's card several times as a result of using this method of *sampling with replacement*.

Our first bootstrap sample, arranged in increasing order of magnitude for ease in reading, might look like this:

```
138.5 138.5 140.0 141.0 141.0 143.5 145.0 147.0 148.5
150.0 153.0 154.0 155.0 156.5 157.0 158.5 159.0 159.0
159.0 160.5 161.0 162.
```

Several of the values have been repeated; not surprising, as we are sampling with replacement, treating the original sample as a stand-in for the much larger population from which the original sample was drawn. The minimum of this bootstrap sample is 138.5, higher than that of the original sample; the maximum at 162.0 is less than the original, while the median remains unchanged at 153.5.

```
137.0 138.5 138.5 141.0 141.0 142.0 143.5 145.0 145.0
147.0 148.5 148.5 150.0 150.0 153.0 155.0 158.0 158.5
160.5 160.5 161.0 167.5
```

In this second bootstrap sample, again we find repeated values—quick, what are they? This time the minimum, maximum, and median are 137.0, 167.5, and 148.5, respectively.

Two bootstrap samples cannot tell us very much. But suppose we were to take 50 or a hundred such samples. Figure 1.9 shows a one-way strip plot of the medians of 50 bootstrap samples taken from the classroom data. These values provide an

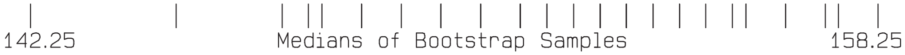


Figure 1.9 One-way strip plot of the medians of 50 bootstrap samples taken from the classroom data.

insight into what might have been had we sampled repeatedly from the original population.

Quick question: What is that population? Does it consist of all classes at the school where I was teaching? All sixth-grade classes in the district? All sixth-grade classes in the state? The school was Episcopalian, and the population from which its pupils were drawn is typical of middle- to upper-class Southern California, so perhaps the population was all sixth-grade classes in Episcopalian schools in middle- to upper-class districts of Southern California.

Exercise 1.9: Our original question, you'll recall, is which is the least variable (or, equivalently, the most precise) estimate: mean or median? To answer this question, at least for several samples, let us apply the bootstrap, first to our classroom data and then to the data we collected in Exercise 1.1. You'll need the following R listing:

```
#Comments to R code all start with the pound sign #
#This program selects 100 bootstrap samples from your
#data and then produces a boxplot of the results.
#first, we give a name, urdata, to the observations
#in our original sample
➢  urdata = c( , . . .)
#Record group size
➢  n = length(urdata)
#set number of bootstrap samples
➢  N = 100
#create a vector of length N in which to store the
#results
➢  stat = numeric(N)
#Set up a loop using the for() function to generate a
#series of bootstrap samples
#All the commands between the brackets { } will be
#repeated N times in this example; the first time i =
#1, the next time i = i + 1.
➢  for (i in 1:N){
#bootstrap sample counterparts to observed samples
#are denoted with "B"
+  urdataB = sample (urdata, n, replace = T)
+  stat[i] = mean(urdataB)
+  }
➢  boxplot (stat)
```

1.8 SAMPLES AND POPULATIONS

If it weren't for person-to-person variation, it really would be easy to find out what brand of breakfast cereal people prefer or which movie star they want as their leader. Interrogate the first person you encounter on the street and all will be revealed. As things stand, either we must pay for and take a total census of everyone's view (the cost of the 2003 recall election in California pushed an already near-bankrupt state one step closer to the edge) or take a sample and learn how to extrapolate from that sample to the entire population.

In each of the data collection examples in Section 1.2, our observations were limited to a sample from a population. We measured the height, circumference, and weight of a dozen humans (or dogs, or hamsters, or frogs, or crickets), but not all humans or dogs or hamsters. We timed some individuals (or frogs or turtles) in races but not all. We interviewed some fellow students but not all.

If we had interviewed a different set of students, would we have gotten the same results? Probably not. Would the means, medians, interquartile ranges, and so forth have been similar for the two sets of students? Maybe, *if* the two samples had been large enough and similar to each other in composition.

If we interviewed a sample of women and a sample of men regarding their views on women's right to choose, would we get similar answers? Probably not, as these samples were drawn from completely different populations (different, i.e., with regard to their views on women's right to choose.) If we want to know how the citizenry as a whole feels about an issue, we need to be sure to interview both men and women.

In every statistical study, two questions immediately arise:

1. How large should my sample be?

2. How can I be sure this sample is representative of the population in which my interest lies?

By the end of Chapter 6, we'll have enough statistical knowledge to address the first question, but we can start now to discuss the second.

After I deposited my ballot in a recent election, I walked up to the interviewer from the *LA Times* who was taking an exit poll and offered to tell her how I'd voted. "Sorry," she said, "I can only interview every ninth person."

What kind of a survey wouldn't want my views? Obviously, a survey that wanted to ensure that shy people were as well represented as boisterous ones and that a small group of activists couldn't bias the results.*

One sample we would all insist be representative is the jury.[†] The Federal Jury Selection and Service Act of 1968 as revised[‡] states that citizens cannot be disqualified from jury duty "on account of race, color, religion, sex, national origin or economic status."[§] The California Code of Civil Procedure, section 197, tells us *how* to get a representative sample. First, you must be sure your sample is taken from the appropriate population. In California's case, the "list of registered voters and the Department of Motor Vehicles list of licensed drivers and identification card holders . . . shall be considered inclusive of a representative cross section of the population." The Code goes on to describe how a table of random numbers or a computer could be used to make the actual selection. The bottom line is that to obtain a random, representative sample:

* To see how surveys could be biased deliberately, you might enjoy reading Grisham's *The Chamber*.

[†] Unless of course, we are the ones on trial.

[‡] 28 U.S.C.A. x1861 et. seq (1993).

[§] See 28 U.S.C.A. x1862 (1993).

- Each individual (or item) in the eligible population must have an equal probability of being selected.
- No individual (item) or class of individuals may be discriminated against.

There's good news and bad news. The bad news is that any individual sample may not be representative. You can flip a coin six times and every so often it will come up heads six times in a row.* A jury may consist entirely of white males. The good news is that as we draw larger and larger samples, the samples will resemble more and more closely the population from which they are drawn.

> **Exercise 1.10:** For each of the three data collection examples of Section 1.2, describe the populations you would hope to extend your conclusions to and how you would go about ensuring that your samples were representative in each instance.

1.8.1 Drawing a Random Sample

Recently, one of our clients asked for help with an audit. Some errors had been discovered in an invoice they'd submitted to the government for reimbursement. Since this client, an HMO, made hundreds of such submissions each month, they wanted to know how prevalent such errors were. Could we help them select a sample for analysis?

We could, but we needed to ask the client some questions first. We had to determine what the population was from which the sample would be taken and what constituted a *sampling unit*.

Were they interested in all submissions or just some of them? The client told us that some submissions went to state agencies and some to Federal, but for audit purposes, their sole interest was in certain Federal submissions, specifically in submissions for reimbursement for a certain type of equipment. Here too, a distinction needed to be made between custom equipment (with respect to which there was virtually never an error) and more common off-the-shelf supplies. At this point in the investigation, our client breathed a sigh of relief. We'd earned our fee, it appeared, merely by observing that instead of 10,000 plus potentially erroneous claims, the entire population of interest consisted of only 900 or so items.

(When you read earlier that 90% of the effort in statistics was in designing the experiment and collecting the data, we meant exactly that.)

Our client's staff, like that of most businesses, was used to working with an electronic spreadsheet. "Can you get us a list of all the files in spreadsheet form?" we asked.

They could and did. The first column of the spreadsheet held each claim's ID. The second held the date. We used the spreadsheet's sort function to sort all the

* Once you've completed the material in the next chapter, you'll be able to determine the probability of six heads in a row.

claims by date, and then deleted all those that fell outside the date range of interest. Next, a new column was inserted, and in the top cell (just below the label row) of the new column, we put the Excel command **rand**(). We copied this command all the way down the column.

A series of numbers between 0 and 1 was displayed down the column. To lock these numbers in place, we went to the Tools menu, clicked on "options," and then on the calculation tab. Next, we made sure that Calculation was set to manual and there was no check mark opposite "recalculate before save."

Now, we resorted the data based on the results of this column. Beforehand, we'd decided there would be exactly 35 claims in the sample, so we simply cut and pasted the top 35 items.

With the help of R, we can duplicate this entire process with a single command:

```
> randsamp = sample(Data).
```

*1.8.2 Using Data That Are Already in Spreadsheet Form

Use Excel's "Save As" command to save your spreadsheet in csv format. Suppose you have saved it in c:/mywork/Data.dat. Then use the following command to bring the data into R:

```
> Data = read.table("c:/mywork/Data.dat", sep=",")
```

1.8.3 Ensuring the Sample Is Representative

Exercise 1.11: We've already noted that a random sample might not be representative. By chance alone, our sample might include men only, or African-Americans but no Asians, or no smokers. How would you go about ensuring that a random sample is representative?

1.9 SUMMARY AND REVIEW

In this chapter, you learned R's syntax and a number of R commands with which to

- Manipulate data and create vectors of observations (c, edit, sort, numeric, factor)
- Perform mathematical functions (exp,log)
- Compute statistics (mean, median, quantile)
- Create graphs (boxplot, hist, pie, plot, plotCDF, stripchart, rug).

You learned that these commands have parameters such as xlab and ylab that allow you to create more attractive graphs

- Control program flow (for)
- Select random samples (sample)
- Read data from tables (read.table).

The best way to summarize and review the statistical material we've covered so far is with the aid of three additional exercises.

> **Exercise 1.12:** Make a list of all the *italicized* terms in this chapter. Provide a definition for each one along with an example.
>
> **Exercise 1.13:** The following data on the relationship of performance on the LSATs to GPA are drawn from a population of 82 law schools. We'll look at this data again in Chapters 3 and 4.

```
LSAT = c(576,635,558,578,666,580,555,661,651,605,653,5
75,545,572,594)
GPA = c(3.39,3.3,2.81,3.03,3.44,3.07,3,3.43,3.36,3.13,
3.12,2.74,2.76,2.88,2.96)
```

Make box plots and histograms for both the LSAT score and GPA. Tabulate the mean, median, interquartile range, and the 95th and 5th percentiles for both variables.

> **Exercise 1.14:** I have a theory that literally all aspects of our behavior are determined by our birth order (oldest/only, middle, and youngest) including clothing, choice of occupation, and sexual behavior. How would you go about collecting data to prove or disprove some aspect of this theory?