Visit www.abacon.com/miller
for related data sets

BEYOND STATISTICS

A Practical Guide to Data Analysis

Benjamin Miller

Beyond Statistics

A Practical Guide to Data Analysis

Benjamin Miller
Salem State College

Allyn and Bacon
Boston London Toronto Sydney Tokyo Singapore

A Pearson Education Company
160 Gould Street
Needham Heights, MA 02494

Internet: www.abacon.com

ISBN 0-205-31744-8

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1   05 04 03 02 01 00

To the memory of Paul David Miller

Preface

What is data analysis?

Data analysis is more than statistics. It is the entire process of finding out what happened, of turning data into conclusions and questions for future research. This process includes entering data into a computer, checking the data for errors and omissions, deciding what statistics we want to compute, what graphs we want to draw, what tests we want to perform (if any), and figuring out how to arrange our data to make it possible for the computer to do these things. Until they have done it, most people have no idea how much work is involved in these non-statistical aspects of data analysis. That's where this book comes in.

About the book

The first three chapters provide some important background and introduce some important terminology. Each of the remaining chapters deals with a different aspect of data analysis: data entry, data cleaning, arranging (and rearranging) data, creating new variables, combining data, displaying data, and collapsing data. Each of these chapters (4 - 10) explains why, when, and how a particular set of data techniques is used. The chapters include many illustrations, using small sets of (mostly) made-up data. Between some of the chapters are larger-scale demonstrations, using real data. In all cases the data are provided on the Allyn & Bacon web site (www.abacon.com/miller), and you should replicate the demonstration using your statistical software. Finally, the later chapters include both pencil-and-paper and computer exercises.

About statistics

The purpose of this book is to help you apply your knowledge of statistics to the kinds of problems you are likely to encounter in research. Although we will talk about statistics, this book is not about statistics; it is about data. If you have taken (or are now taking) an introductory course in statistics, you understand central tendency, variability, probability, hypothesis testing, and the like, but you may not have a very good idea what to do with a boxful (or a diskful) of data. How do you get data into a computer? What do you do with them once they're in? This book will show you.

About statistical software

To some people statistical software makes perfect sense and to others it is a perfect nightmare. In my experience, trouble with statistical software usually comes from asking "What should I do?" or "How do I work this?" rather than "What do I want to do?" Remember that the computer (and the statistical software it is running) is only a tool; you are the brains of the operation. It is up to you to say, clearly and precisely and in English, what you want to do. If you do this, it is not hard to find the right menu or command in your particular statistical software.

This book will help you with the process of thinking about your research questions and figuring out how to get answers to them out of your data using statistical software. The book is not, however, a software manual. Thinking about what to do is the same for everyone, but ultimately the question of how to do it depends on what software you are using. We will talk about software in general, but for help with your particular software you'll want a manual or an expert.

Because different readers will be using different software, all the examples use an imaginary program that understands commands given in a quasi-English known as pseudocode. Computer scientists use pseudocode when they are concerned with showing what a program does rather than how it is written. Pseudocode uses a minimum of special terms or syntax; it attempts to convey, tersely but plainly, what we are asking the computer to do.

The key to data analysis is not remembering a thousand rules for a thousand situations, but having a few good habits. I hope this book will help you learn those habits.

■ ■ ■

My father taught me at a tender age to be methodologically orthodox and pedagogically radical, and I hope this book lives up to that credo. Meanwhile I have many wise and sympathetic people to thank.

A backpacking companion is the ultimate captive audience, and Alec Bodkin provided miles of listening and encouragement when I was trying to figure out what it was I needed to write. A Summer faculty writing workshop sponsored by Salem State College was the perfect opportunity to begin putting something on paper. I am indebted to my workshop colleagues – Joe Buttner, Kristine Doll, Eileen Margerum, Eric Metchik, Nancy Schultz, and J.D. Scrimgeour – for their careful reading, their bulls-eye questions and their creative suggestions. Norman Miller, Sue Regan, and Jeffrey Adams read later drafts and offered many invaluable suggestions. My brother Paul died before I could properly thank him for his many contributions, but in substance and style the finished book bears the imprint of his sharp eye and keen judgment. Mindy Clawson worked magic to turn a big mess of words and figures into a book, improving it substantially in the process. Becky Pascal, at Allyn & Bacon, was unfailingly helpful and patient. Very patient.

Contents

1

Orientation

Chapter Outline

1. It is important to distinguish between variables, values, and constants.

2. There are many distinctions among different types of variables. Some of the most useful distinctions are:

   a. Independent and dependent variables.

   b. Discrete and continuous variables.

   c. Categorical and quantitative variables.

3. There are four broad categories of questions we can ask about data:

   a. Questions about the distribution of a single variable.

   b. Questions about differences between groups or conditions.

   c. Questions about the association between categorical variables.

   d. Questions about the relation between quantitative variables.

4. There are several principles that make data analysis easier, more fun, and less error-prone.

   a. Frame questions in terms of data rather than in terms of software.

   b. Look at the data before computing anything; be alert for errors, outliers, skewed distributions, and so on.

   c. Don't do categorical things with quantitative variables or quantitative things with categorical variables.

ONE OF THE MOST significant problems that we encounter in learning to analyze data is knowing what to do in different data situations. The statistics text offers us false security: If Chapter 8 is about t-tests and Chapter 9 is about analysis of variance, we can be reasonably confident in choosing a t-test for the Chapter 8 problems and in choosing analysis of variance for the Chapter 9 problems. But when the exam comes along and the different kinds of problems are mixed together we often make mistakes. Real data analysis is more like the exam than like the problems at the end of the chapter.

This chapter offers a solution to the what-should-I-do problem in the form of a simple framework for talking about data and asking questions about data.

1.1 Variables, values, and constants

variables & constants

Let's begin with some important – and probably familiar – terminology. A *variable* is, not surprisingly, something that varies. A *constant*, on the other hand, is something that does not vary. Temperature is something that varies in a pretty obvious way, so temperature is a variable. But in Mammoth Cave the temperature is 54°F year round, so temperature there is a constant. There's no contradiction here; the point is that we can say what's a variable and what's a constant only within a specific context.

Suppose, for example, you are interested in the (possible) relation between obesity and hypertension. You go out and find a few hundred people and weigh them and measure their blood pressure. You now have a set of data containing two variables: weight and blood pressure. But you realize that weight alone doesn't tell the whole story: someone 5'2" tall who weighs 200 pounds is probably obese, but someone 6'2" tall who weighs 200 pounds is probably not. Height is a variable that you care about, but since you didn't measure it it's not a variable in your data. Accordingly, you might repeat your study, selecting only people who are, say, 5'10" tall. Now your data set contains the same two variables as before (weight and blood pressure), but height is no longer an extraneous variable; in this new context it is a constant.

Any variable can be turned into a constant simply by not allowing it to vary, and that's about all there is to know about constants. Variables, on the other hand, are what research is all about, and there are a few things to know about variables. Let's look at some (hypothetical) blood pressure data:

```
name      weight    blood pressure
Ed        153       130
Ted       175       120
Ned       240       110
Fred      192       150
Jed       132       170
Hubert    147       140
...
```

variables & values

Each row contains the name, weight and blood pressure measurements of a particular subject, or *case*. Each column contains a *variable*, something that varies from case to case. Each item is a *value* of a particular variable for a particular case. Thus `Jed` is the value of the variable `name` for the fifth case; `240` is the value of the variable `weight` in the case of Ned; and `120` is the value of the variable `blood pressure` in the case of Ted. The difference between a variable and a value is very important because, as we will see in the next section, we make a number of distinctions among variables in terms of the kind of values they have.

1.2 Kinds of variables

Independent variables correspond to groups, conditions, predictors.

Dependent variables correspond to results.

There are many distinctions among variables. In this section we will look at three that are particularly important: independent/dependent, discrete/continuous, and categorical/-quantitative.

Independent and dependent variables. This is a familiar if not always easy distinction. For example, if we give group A an anti-anxiety drug and group B a placebo and then measure everybody's anxiety levels, the *treatment* variable (drug/placebo) is an independent variable and the *outcome* variable (anxiety) is a dependent variable. One way to know which is which is to notice that we (sometimes) manipulate the independent variable (in this case by deciding which drug(s) to use), but we never manipulate the dependent variable. Another way is to think of the value of the dependent variable as depending on the value of the independent variable; in this case how much anxiety someone has may depend on which treatment group we put her in. Which variable is independent and which is dependent is usually fairly clear, but not always. If we ask people their political party and who they voted for in the last election, vote is more likely to depend on party than the other way around. But what if we ask people their position on abortion rights and their position on affirmative action? Here there does not seem to be a plausible independent/dependent distinction.

Discrete and continuous variables. The engine in your car can have 3 cylinders or 4 cylinders but it can't have 3.1 cylinders. Number of cylinders is a discrete variable: between two adjacent values (e.g. 3 and 4 cylinders) on the scale there cannot be any intermediate value (e.g. 3.5 cylinders). By contrast, your engine can get 35 or 36 miles per gallon, or it can get 35.1 mpg, or 35.11 or 35.111 or 35.1111... Miles per gallon is a continuous variable: no matter how close two measurements are, it is always possible for a third measurement to fall between them.[1] Notice that the discrete/continuous distinction gets blurry at the edges: If you weigh people using a digital scale that rounds weights to the nearest pound, then you are measuring something (weight) that is inherently continuous using a scale of measurement that is artificially discrete. You can't turn a discrete variable into a continuous one, but you can turn a continuous variable into a discrete one. As you will see, we often do this in analyzing data.

Discrete variables count things and events.

Continuous variables measure.

Categorical and quantitative variables. The distinction between *categorical* and *quantitative* variables is important because what we can do with one kind is so different from what we can do with the other. Height, weight, number of cavities, shoe size, age, and income are all quantitative variables; the kind of measuring they do is to quantify. Some quantities can be measured on continuous scales (height, weight) and others are discrete (number of cavities), but in both cases we are measuring a quantity of something. Quantitative variables ask *How many?* (discrete) or *How much?* (continuous), while categorical variables ask *Which?*

Categorical variables classify.

Quantitative variables count and measure.

There are two kinds of categorical variable. Political party, sex, zip code, species, and marital status are all *nominal* categorical variables. Nominal, from the Latin *nomen* (name) refers to the fact that the values of such variables are names (of categories); the kind of measuring they do is to categorize, or classify. You are either female or male; married or not; you belong to one political party or another, and so on. You can't be something in between, and one category isn't more or less than another.

Nominal categorical variables name.

On the other hand, variables such as position on the bestseller list, critics' ratings of movies or restaurants (★, ★★, ★★★, etc.), or finishing place in a race are *ordinal* categorical variables. These variables measure by categorizing, but the categories they use can be ordered. The values of ordinal variables are not only names of categories, they are ranks or comparative judgments. For example, if my self-help book, *Living with Boring Neighbors: A Survivor's Guide,* is no. 3 on the bestseller list, we know something about how well it is selling, but only relatively. It might be selling much better than the no. 4 book, or only a little better; it might have sold a million copies last week, or a dozen. It is very important to keep in mind that ordinal variables express *only* order. Even though ranks are expressed by numbers, these are not numbers that count or measure in the usual sense, and you can't do quan-

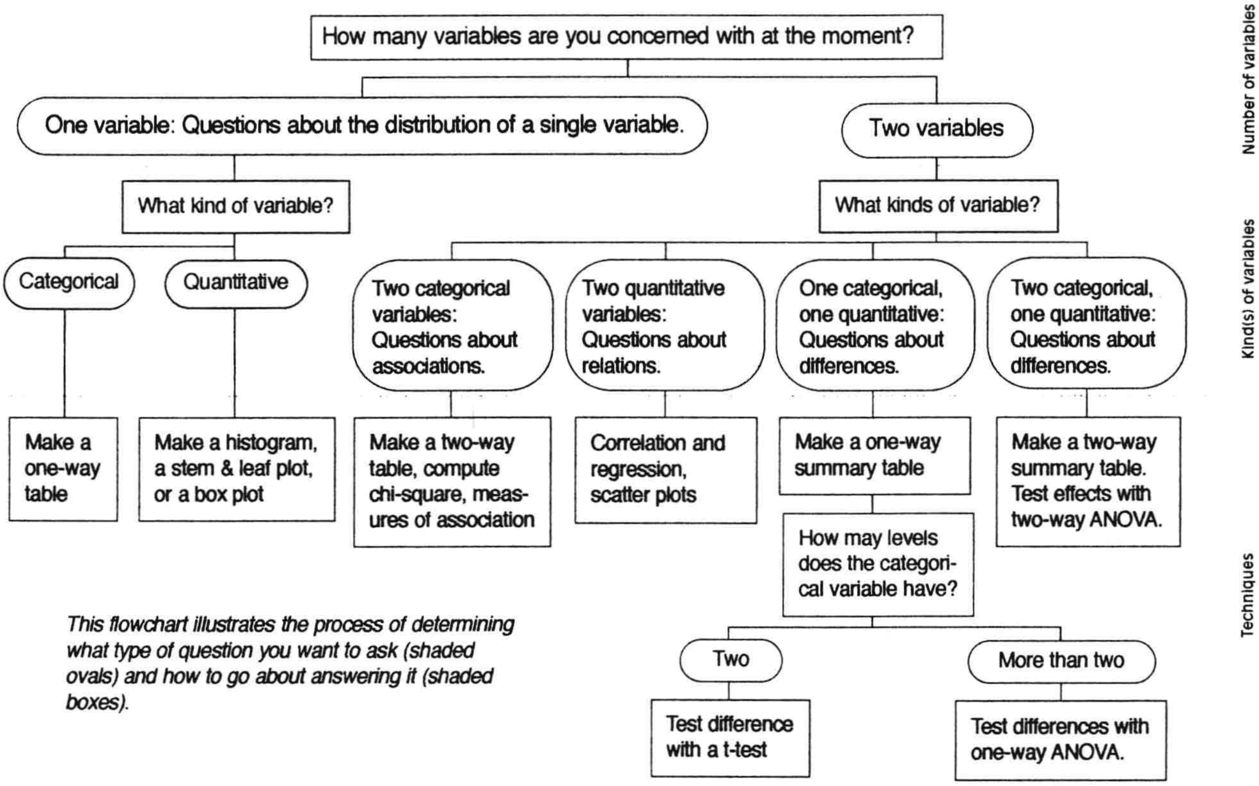
Ordinal categorical variables order, rank or rate.

[1] Discrete variables are sometimes referred to as counting variables, because their values are always integers (counting numbers), and continuous variables are sometimes referred to as measuring variables, but these terms are ambiguous and potentially misleading. For some purposes, counting is legitimately considered a kind of measurement.

titative things — such as computing a mean[2] — with these numbers.

1.3 Kinds of questions

Most research questions fall into one of four broad categories. Determining which category we're working in goes a long way toward deciding what kinds of procedures are (and aren't) appropriate. The flow chart below summarizes the relations between types of data, types of questions, and types of procedures. The tables and graphs mentioned in this section are discussed in detail in chapters 8 and 9.

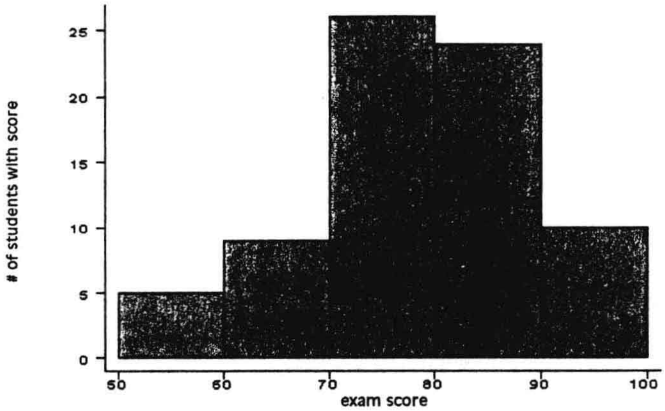

How many variables are you concerned with at the moment?
One variable: Questions about the distribution of a single variable.
Two variables
What kind of variable?
What kinds of variable?
Categorical
Quantitative
Two categorical variables: Questions about associations.
Two quantitative variables: Questions about relations.
One categorical, one quantitative: Questions about differences.
Two categorical, one quantitative: Questions about differences.
Make a one-way table
Make a histogram, a stem & leaf plot, or a box plot
Make a two-way table, compute chi-square, measures of association
Correlation and regression, scatter plots
Make a one-way summary table
Make a two-way summary table. Test effects with two-way ANOVA.
How may levels does the categorical variable have?
Two
More than two
Test difference with a t-test
Test differences with one-way ANOVA.
Number of variables
Kind(s) of variables
Techniques

This flowchart illustrates the process of determining what type of question you want to ask (shaded ovals) and how to go about answering it (shaded boxes).

This chart is not everything there is to know about analyzing data. Rather, it will point you in the right direction when you have some data and don't know what to do with them.

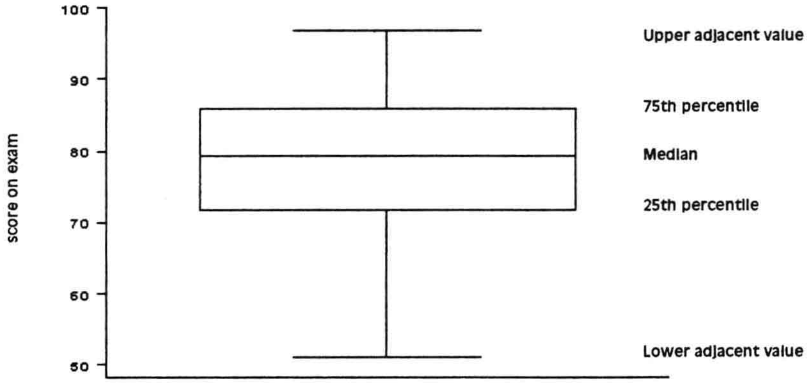
Questions about a distribution. Often we have questions about the distribution of a single variable, either categorical or quantitative. For example, a teacher who has given a multiple-choice exam has some questions about how the scores are distributed: what is the mean? the standard deviation? the range? He may want to know the shape of the distribution (is it skewed? flat? normal? multimodal?) and whether there are any extremely high or low scores (outliers) that may be pulling the mean up or down. In addition to computing the various descriptive statistics mentioned, one way to answer some of these questions is to plot a *histogram*, or *grouped frequency distribution*.

[2] There is no practical difficulty in computing the mean of an ordinal variable, but interpreting the result is another story. The problem is that the intervals between ranks do not necessarily represent the same amount of difference between the things ranked. The difference between $1 and $2 is the same amount of money as the difference between $2 and $3, and so on. But the difference betwen first place and second place in a race might be .02 seconds while the difference between second and third place might be 5 seconds. The mean is defined as the value of μ such that $\sum(X-\mu)=0$, and this definition is violated by variables that do not have equal intervals.



a histogram, or grouped frequency distribution

Histograms show the number of cases (here, students) falling within defined ranges of the variable (test score). The pattern of bar heights is the shape of the distribution of test scores. Another way to see the shape of the distribution is with a *box-and-whisker plot* (described in detail in Chapter 9).



a box-and-whisker plot

The teacher might also want to know the distribution of answers to a particular multiple-choice question, such as:

17. Napoleon Bonaparte was born in
    a. Corsica
    b. Corfu
    c. Sardinia
    d. Paris
    e. Jersey City

Knowing what percentage of his students chose each answer will tell him something about what they learned and didn't learn. For this he would use a frequency table:

```
Question 17
     choice  |     Freq.     Percent
-------------+----------------------
           A |        52       69.33
           B |         8       10.67
           C |        14       18.67
           D |        12       16.00
           E |         1        1.33
-------------+----------------------
       Total |        75      100.00
```

A frequency table shows the number (frequency) of scores at each level of the independent variable.

Or consider a pediatrician measuring a child's height. To determine whether the child is growing at an appropriate rate she needs to be able to compare the child's height to a distribution of the heights of a large number of children of the same age and sex. If she can assume the distribution is normal (there is a test for this) then she can turn the child's height into a standard (z) score and then into a percentile. If the child's height is at the 40th percentile but was at the 80th percentile a year ago, clearly the child's growth has slowed; something may be wrong.

Finally, suppose a gubernatorial candidate running on an education reform platform argues that the school day must be lengthened. To strengthen his rhetoric he points out that the average school day length is only five and a half hours. Because you've learned not to believe anything he says you seek to verify (or discredit) his claim. You call a random sample of 100 school districts around the state and ask them all how long their school day is. Your data will allow you to compute a t-test of the hypothesis that the average school day length is 5.5 hours.

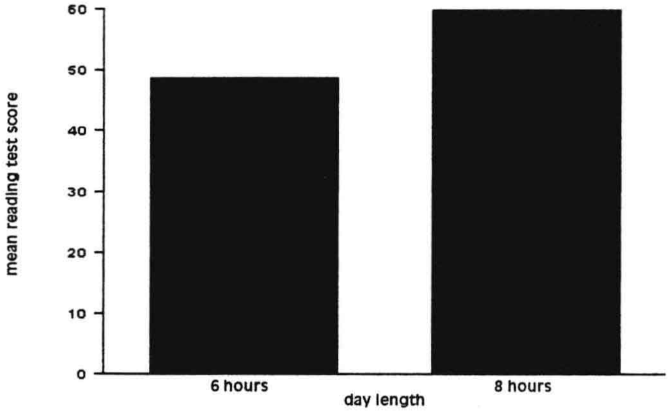
Questions about differences. A great deal of research asks questions about differences between groups or conditions (values, or levels, of a categorical independent variable) in terms of some quantitative dependent measure. For example, if we look at the difference in reading test scores (quantitative dependent variable) between students taught in a six-hour school day and students taught in an eight-hour day (levels of a categorical independent variable), we can measure the differences with means and standard deviations and present these in a summary table, or table of means.

A summary table shows the mean, s.d., and frequency of the scores at each level of the independent variable.

```
            | Summary of reading test score
day length  | Mean      s.d.    N
------------+----------------------
   6 hours  | 48.4      27.9    10
   8 hours  | 59.7      31.5    10
------------+----------------------
     Total  | 54.1      29.5    20
```

We can also present the difference in a bar graph, which conveys less information but makes the point more clearly:

A bar graph shows the mean (y-axis) of the scores at each level of the independent variable (x-axis).



We may use a two-sample t-test to evaluate the possibility that the difference arose by chance (i.e. through sampling error). If we can reject this possibility, we may be able to predict that a school with a six-hour day could raise test scores by lengthening the day to eight hours, but we could not predict that the school could raise test scores even more by lengthening the day to ten hours because this level of the independent variable (day length) wasn't sampled in the experiment.

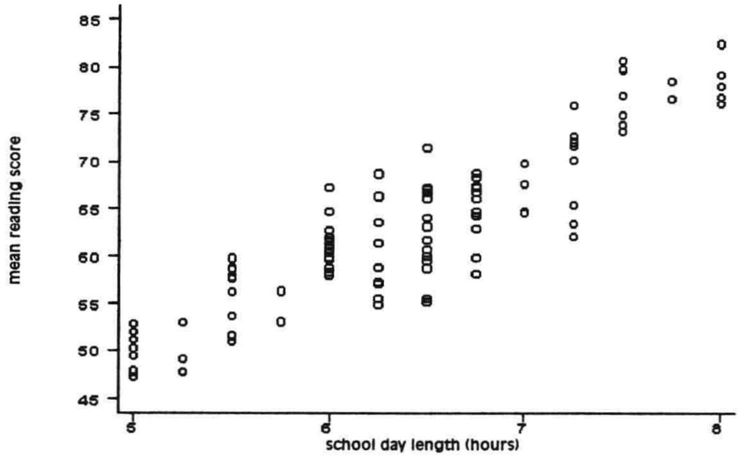
Questions about associations. A huge amount of research asks about whether two (or more) categorical variables are associated with one another. For example, we may have some data on school day length (six hours or less and more than six hours) and reading test scores (at or below median and above median). Notice that although day length and test scores are both inherently quantitative, here they have been measured using categorical variables. We would cross-tabulate day length (short/long) and test scores (high/low):

	test scores	
day length	≤ median	> median
≤ 6 hours	60	40
> 6 hours	45	55

A cross-tabulation shows the number of cases in each combination of two categorical variables.

A table shows the number (frequency) of cases (here, school districts) that fall into each of the combinations of values of two categorical values. This table shows that schools with long days are more likely than short-day schools to have high test scores. If necessary we could use a chi-square test to try to rule out the possibility that our result came about by chance. This would allow us to predict that other long-day schools have a certain likelihood of having high test scores.

Questions about relations. In the previous section I use *association*, somewhat arbitrarily, to refer to a contingency between two categorical variables, but when the same question concerns two quantitative variables I will, again somewhat arbitrarily, use the term *relation*. For example, if we randomly sample 100 school districts and record their school day lengths and reading test scores, we may examine the relation between these variables using a scatter plot:



A scatter plot shows each case's values on two quantitative variables.

The pattern of the data points indicates a fairly strong linear relation between day length and reading scores. We can measure its strength and test its reliability using methods of correlation and regression, which may allow us to predict that a longer school day will raise test scores.

Each of the methods used in the preceding examples was an appropriate choice for the kind of data that were available and for the question being asked. The importance of these four basic categories – distributions, differences, associations, relations – comes from the guidance they offer in data analysis. Identifying what kind of data you have and what kind of question(s) you are asking will help you think straight about what you want to do.

1.4 Some principles of data analysis

There are a few general principles that you should keep in mind throughout this book and whenever you work with data. They will make your life easier and your work better.

Principle 1: Have a destination. Most basic data activities are not software specific. If you know what you want to do, chances are any statistical software will do it; if you don't, no software will help you. For example, any software worth talking about will let you make a new variable that is the sum of two existing variables. The specific way of doing this will vary from package to package; for example, here are three commands that create a new variable (newvar) that is the sum of two existing variables (x1 and x2):

SAS:	newvar = x1 + x2
SPSS:	compute newvar = x1 + x2
Stata:	generate newvar = x1 + x2

You will find that these variations don't matter if you approach data analysis from the perspective of what you want to do rather than from the perspective of how the software works.

I can't say this strongly enough. The reason many people find software manuals confusing and unhelpful is that they don't quite know what they want to do. At this point in your data analysis career it is crucial that you give most of your attention and effort to thinking about what needs to be done, and why, in a particular data situation; worry about the software only enough to get your work done. I've met plenty of software whizzes who couldn't analyze data, but I've never met a competent data analyst who couldn't reasonably quickly figure out how to use a new software package to accomplish a particular task.

Principle 2: Look at the data. Always look at the data — and I mean always — before you run tests or compute summary statistics. There may be something unanticipated or erroneous in your data that you should know about before you start blindly computing things. Carpenters live by the rule "Measure twice, cut once." The same idea applies to us: "Think twice, compute once." Why is this so important? After all, the computer is doing all the work, quickly, and we're not wasting any two-by-fours. What's the harm in a few extra computations?

Extra computations sometimes entail a subtle scientific risk. Suppose we hypothesize that taking ginkgo pills improves memory. We recruit a random sample of 20 people and randomly divide them into two groups of ten. We give all subjects a memory test, then give group A ginkgo and group B a placebo for a month. We then give all subjects another memory test, comparable to the first, and subtract the first score (pretest) from the second score (posttest) to see how much memory improvement there was. Our data look like this:

Ginkgo group	Placebo group
-5	-6
-9	-5
10	-1
7	-9
-1	1
81	10
1	4
-1	7
-6	-1
4	0

If we plunge blithely into testing our hypothesis without looking at the data, we will find that the mean memory scores for the ginkgo and placebo groups are 8.1 and 0, respectively.