CHAPTER 8

Managing and Curating Data

Data are the raw material of scientific studies. However, we rarely devote the same amount of time and energy to data organization, management, and curation that we devote to data collection, analysis, and publication. This is an unfortunate state of affairs; it is far easier to use the original (raw) data themselves to test new hypotheses than it is to extract the original data from the published literature. A basic requirement of any scientific study is that it be repeatable by others. Analyses cannot be reconstructed unless the original data are properly organized and documented, safely stored, and made available. Moreover, it is a legal requirement that any data collected as part of a project that is supported by public funds (e.g., federal or state grants and contracts) be made available to other scientists and to the public. Data sharing is legally mandated because the granting agency, not the investigator, technically owns the data. Most granting agencies allow ample time (usually one or more years) for investigators to publish their results before the data have to be made available publicly.[1] However, the Freedom of Information Act (5 U.S.C. § 552) allows anyone in the United States to request publicly funded data from scientists at any time. For all of these reasons, we encourage you to manage your data as carefully as you collect, analyze, and publish them.

---

[1] This policy is part of an unwritten gentleman's agreement between granting agencies and the scientific community. This agreement works because the scientific community so far has demonstrated its willingness to make data available freely and in a timely fashion. Ecologists and environmental scientists generally are slower to make data available than are scientists working in more market-driven fields (such as biotechnology and molecular genetics) where there is greater public scrutiny. Similarly, provisions for data access and standards for metadata production by the ecological and environmental community have lagged behind those of other fields. It could be argued that the lack of a GenBank for ecologists has slowed the progress of ecological and environmental sciences. However, ecological data are much more heterogeneous than gene sequences, so it has proven difficult to design a single data structure to facilitate rapid data sharing.

The First Step: Managing Raw Data

In the laboratory-based bench sciences, the bound lab notebook is the traditional medium for recording observations. In contrast, ecological and environmental data are recorded in many forms. Examples include sets of observations written in notebooks, scrawled on plastic diving slates or waterproof paper, or dictated into voice recorders; digital outputs of instruments streamed directly into data loggers, computers, or palm-pilots; and silver-oxide negatives or digital satellite images. Our first responsibility with our data, therefore, is to transfer them rapidly from their heterogeneous origins into a common format that can be organized, checked, analyzed, and shared.

Spreadsheets

In Chapter 5, we illustrated how to organize data in a **spreadsheet**. A spreadsheet is an electronic page[2] in which each row represents a single observation (e.g., the unit of study, such as an individual leaf or feather, organism, or island), and each column represents a single measured or observed variable. The entry in each **cell** of the spreadsheet, or row-by-column pair, is the value for the unit of study (the row) of the measurement or observation of the variable (the column).[3] All types of data used by ecologists and environmental scientists can be stored in spreadsheets, and the majority of experimental designs (Chapter 7) can be represented in spreadsheet formats.

Field data should be entered into spreadsheets as soon as possible after collection. In our research, we try to transfer the data from our field notes or instruments into spreadsheets on the same day that we collect it. There are several important reasons to transcribe your data promptly. First, you can determine quickly if necessary observations or experimental units were missed due to researcher oversight or instrument failure. If it is necessary to go back and collect additional observations, this must be done very quickly or the new data won't be comparable. Second, your data sheets usually will include marginal notes and observations, in addition to the numbers you record. These marginal notes make sense when you take them, but they have a short half-life and rapidly become incomprehensible (and sometimes illegible). Third, once data are entered into a spreadsheet, you then have two copies of them; it is less likely you will misplace or lose both copies of your valuable results. Finally, rapid organization of data may promote timely analysis and publication; unorganized raw data certainly won't.

Data should be proofread as soon as possible after entry into a spreadsheet. However, proofreading usually does not catch all errors in the data, and once the data have been organized and documented, further checking of the data is necessary (discussed later in the chapter).

[2] Although commercial spreadsheet software can facilitate data organization and management, it is not appropriate to use commercial software for the archival electronic copy of your data. Rather, archived data should be stored as ASCII (American Standard Code for Information Exchange) text files, which can be read by any machine without recourse to commercial packages that may not exist in 50 years.

[3] We illustrate and work only with two-dimensional (row, column) spreadsheets (also called **flat files**). Regardless of whether data are stored in two- or three-dimensional formats, pages within electronic workbooks must be exported as flat files for analysis with statistical software. Although spreadsheet software often contains statistical routines and random number generators, we do not recommend the use of these programs for analysis of scientific data. Spreadsheets are fine for generating simple summary statistics, but their random number generators and statistical algorithms may not be reliable for more complex analyses (Knüsel 1998; McCullough and Wilson 1999).

Metadata

At the same time that you enter the data into a spreadsheet, you should begin to construct the **metadata** that must accompany the data. Metadata are "data about data" and describe key attributes of the dataset. Minimal metadata[4] that should accompany a dataset include:

- Name and contact information for the person who collected the data
- Geographic information about where the data were collected
- The name of the study for which the data were collected
- The source of support that enabled the data to be collected
- A description of the organization of the data file, which should include:
  - A brief description of the methods used to collect the data
  - The types of experimental units
  - The units of measurement or observation of each of the variables
  - A description of any abbreviations used in the data file
  - An explicit description of what data are in columns, what data are in rows, and what character is used to separate elements within rows

[4] Three "levels" of metadata have been described by Michener et al. (1997) and Michener (2000). Level I metadata include a basic description of the dataset and the structure of the data. Level II metadata include not only descriptions of the dataset and its structure, but also information on the origin and location of the research, information on the version of the data, and instructions on how to access the data. Finally, Level III metadata are considered auditable and publishable. Level III metadata include all Level II metadata plus descriptions of how data were acquired, documentation of quality assurance and quality control (QA/QC), a full description of any software used to process the data, a clear description of how the data have been archived, a set of publications associated with the data, and a history of how the dataset has been used.

(e.g., tabs, spaces, or commas) and between columns (e.g., a "hard return").[5]

An alternative approach, more common in the laboratory sciences, is to create the metadata *before* the data are ever collected. A great deal of clear thinking can accompany the formal laying out and writing up of the methods, including plot design, intended number of observations, sample space, and organization of the data file. Such a priori metadata construction also facilitates the writing of the Methods section of the final report or publication of the study. We also encourage you to draw blank plots and tables that illustrate the relationships among variables that you propose to examine.

Regardless of whether you choose to write up your metadata before or after conducting your study, we cannot overemphasize their importance. Even though it is time-consuming to assemble metadata, and they would seem to contain information that is self-evident, we guarantee that you will find them invaluable. You may think that you will never forget that Nemo Swamp was where you collected the data for your opus on endangered orchids. And of course you can still recall all the back roads you drove to get there. But time will erode those memories as the months and years pass, and you move on to new projects and studies.

We all have experiences looking through lists of computer file names (NEMO, NEMO03, NEMO03NEW), trying to find the one we need. Even if we can find the needed data file, we may still be out of luck. Without metadata, there is no way to recollect the meaning of the alphabet soup of abbreviations and acronyms running across the top row of the file. Reconstructing datasets is frustrating, inefficient, and time-consuming, and impedes further data analysis. Undocumented datasets are nearly useless to you and to anyone else. Data without metadata cannot be stored in publicly-accessible data archives, and have no life beyond the summary statistics that describe them in a publication or report.

The Second Step: Storing and Curating the Data

Storage: Temporary and Archival

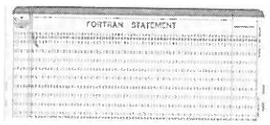
Once the data are organized and documented, the dataset—a combination of data and metadata—should be stored on permanent media. The most permanent medium, and the only medium acceptable as truly archival, is acid-free paper. It is good practice to print out a copy of the raw data spreadsheet and its accompanying metadata onto acid-free paper using a laser printer.[6] This copy should be stored somewhere that is safe from damage. Electronic storage of the original dataset also is a good idea, but you should not expect electronic media to last more than 5–10 years.[7] Thus, electronic storage should be used primarily for working copies. If you intend the electronic copies to last more than a few years, you will have to copy the datasets onto newer electronic media on a regular basis. You must recognize that data storage has real costs, in space (where the data are stored), time (to maintain and transfer datasets among media), and money (because time and space are money).

Curating the Data

Most ecological and environmental data are collected by researchers using funds obtained through grants and contracts. Usually, these data technically are owned by the granting agency, and they must be made available to others within a reasonably short period of time. Thus, regardless of how they are stored, your datasets must be maintained in a state that makes them usable not only by you, but also by the broader scientific community and other interested individuals and parties. Like museum and herbarium collections of animals and plants, datasets must be curated so they are accessible to others. Multiple datasets gen-

[5] Ecologists and environmental scientists are developing standards for metadata. These include the Content Standards for Digital Geospatial Metadata (FGDC 1998) for spatially-organized data (e.g., data from Geographic Information Systems, or GIS); standard descriptors for ecological and environmental data on soils (Boone et al. 1999); and Classification and Information Standards for data describing types of vegetation (FGDC 1997). Many of these have been incorporated within the Ecological Metadata Language, or EML, developed in late 2002 (http://knb.ecoinformatics.org/software/eml/).

[6] Inks from antique dot-matrix printers (and typewriters) last longer than those of inkjet printers, but the electrostatic bonding of ink to paper by laser printers lasts longer than either. True Luddites prefer Higgins Aeterna India Ink (with a tip o' the hat to Henry Horn).



Computer punch card

[7] Few readers will remember floppy disks of any size (8", 5-1/4", 3-1/2"), much less paper tapes, punch cards (illustrated), or magnetic tape reels. As we write, the ubiquitous CD-ROMs are being replaced rapidly by DVDs, and we expect to have to purchase our vinyl record collections again for the second time in as many decades. Although the actual lifespan of most magnetic media is on the order of years to decades, and that of most optical media (CDs, DVDs) is on the order of decades-to-centuries (only years in the humid tropics where fungi eat the glue between the laminated layers of material and etch the disk), the capitalist marketplace replaces them all on an approximately 5-year cycle. Thus, although we still have on our shelves readable 5-1/4" floppy disks with data from the 1980s and 1990s, it is nearly impossible in 2004 to find, much less buy, a disk drive with which to read them. And even if we could find one, the current operating systems of our computers wouldn't recognize it as usable hardware. A good resource for obsolete equipment and software is the Computer History Museum in Mountain View, California: http://www.computerhistory.org/.

erated by a project or research group can be organized into electronic data catalogs. These catalogs often can be accessed and searched from remote locations using the internet.

At the time that we are writing this book, there are no established standards for curation of ecological and environmental data. Most data catalogs are maintained on servers connected to the World Wide Web,[8] and can be found using search engines that are available widely. Like data storage, data curation is costly; it has been estimated that data management and curation should account for approximately 20% of the total cost of a research project (Michener and Haddad 1992). Unfortunately, although granting agencies require that data be made available, they have been reluctant to fund the full costs of data management and curation. Researchers themselves rarely budget these costs in their grant proposals. When budgets inevitably are cut, data management and curation costs often are the first items to be dropped.

The Third Step: Checking the Data

Before beginning to analyze a dataset, you must carefully examine it for outliers and errors.

The Importance of Outliers

Outliers are recorded values of measurements or observations that are outside the range of the bulk of the data (Figure 8.1). Although there is no standard level of "outside the range" that defines an outlier, it is common practice to consider values beyond the upper or lower deciles (90th percentile, 10th percentile) of a

[8] Many scientists consider posting data on the World Wide Web to be a means of permanently archiving data. This is illusory. First, it is simply a transfer of responsibility from you to a computer system manager (or other information technology professional). By placing your electronic archival copy on the Web, you imply a belief that regular backups are made and maintained by the system manager. Every time a server is upgraded, the data have to be copied from the old server to the new one. Most laboratories or departments do not have their own systems managers, and the interests of computing centers in archiving and maintaining Web pages and data files do not necessarily parallel those of individual investigators. Second, server hard disks fail regularly (and often spectacularly). Last, the Web is neither permanent nor stable. GOPHER and LYNX have disappeared, FTP is being replaced by HTTP, and HTML, the current language of the web, is already being phased out in favor of (the not entirely compatible) XML. All of these changes to the functionality of the World Wide Web and the accessibility of files stored within it have occurred within 10 years. It often is easier to recover data from notebooks that were hand-written in the nineteenth century than it is to recover data from Web sites that were digitally "archived" in the 1990s!

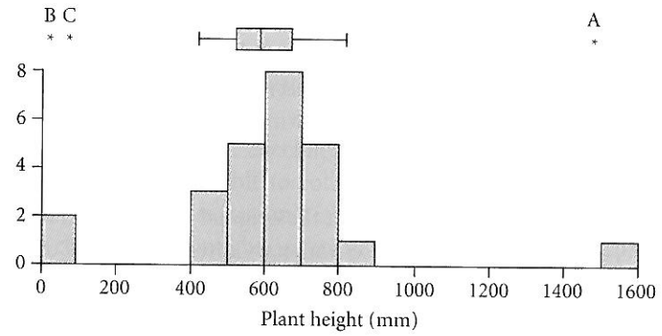


Figure 8.1 Histogram and box plot of measurements of plant height for samples of the carnivorous pitcher plant *Darlingtonia californica* (sample size $N = 25$; data from Table 8.1). The histogram shows the frequency of plants in size class intervals of 100 mm. Above the histogram is the box plot of the same data. The vertical line in the box indicates the median of the distribution. The box encompasses 50% of the data, and the whiskers encompass 90% of the data. A, B, and C indicate the three extreme data points and correspond to the labeled rows in Table 8.1 Basic histograms and box plots allow you to rapidly identify extreme values in your data.

distribution to be potential outliers, and those smaller than the 5th percentile or larger than the 95th percentile to be possible extreme values. Many statistical packages highlight these values on box plots (as we do with stars and letters in Figure 8.1) to indicate that you should scrutinize them carefully. Some of these packages also have interactive tools, described in the next section, that allow for location of these values within the spreadsheet.

It is important to identify outliers, because they can have dramatic effects on your statistical tests. Outliers and extreme values increase the variance in the data. Inflated variances decrease the power of the test and increase the chance of a Type II error (failure to reject a false null hypothesis; see Chapter 4). For this reason, some researchers automatically delete outliers and extreme values prior to conducting their analyses. This is bad practice! There are only two reasons for justifiably discarding data: (1) the data are in error (e.g., they were entered incorrectly in the field notebook); (2) the data no longer represent valid observations from your original sample space (e.g. one of your dune plots was overrun by all-terrain vehicles). Deleting observations simply because they are "messy" is laundering or doctoring the results and legitimately could be considered scientific fraud.

Outliers are more than just noise. Outliers can reflect real biological processes, and careful thought about their meaning can lead to new ideas and

hypotheses.[9] Moreover, some data points will appear as outliers only because the data are being forced into a normal distribution. Data from non-normal distributions such as the log-normal or exponential often seem to be extreme, but these data fall in place after they are appropriately transformed (see below).

Errors

Errors are recorded values that do not represent the original measurements or observations. Some, but not all, errors are also outliers. Conversely, not all outliers in your data are necessarily errors. Errors can enter the dataset in two ways: as errors in collection or as errors in transcription. Errors in collection are data that result from broken or miscalibrated instruments, or from mistaken data entry in the field. Errors in collection can be difficult to identify. If you or your field assistant were recording the data and wrote down an incorrect value, the error rarely can be corrected.[10] Unless the error is recognized immediately during data transcription, it will probably remain undetected and contribute to the variation of the data.

[9] Although biologists are trained to think about means and averages, and to use statistics to test for patterns in the central tendency of data, interesting ecology and evolution often happen in the statistical tails of a distribution. Indeed, there is a whole body of statistics dedicated to the study and analysis of extreme values (Gaines and Denny 1993), which may have long-lasting impacts on ecosystems (e.g., Foster et al. 1998). For example, isolated populations (peripheral isolates, *sensu* Mayr 1963) that result from a single founder event may be sufficiently distinct in their genetic makeup that they would be considered statistical outliers relative to the original population. Such peripheral isolates, if subject to novel selection pressures, could become reproductively isolated from their parent population and form new species (Schluter 1996). Phylogenetic reconstruction has suggested that peripheral isolates can result in new species (e.g., Green et al. 2002), but experimental tests of the theory have not supported the model of speciation by founder effect and peripheral isolates (Mooers et al.1999).

[10] To minimize collection errors in the field, we often engage in a call and response dialogue with each other and our field assistants:

> *Data Collector: "This is Plant 107. It has 4 leaves, no phyllodes, no flowers."*
> *Data Recorder: "Got it, 4 leaves no phyllodes or flowers for 107."*
> *Data Collector: "Where's 108?"*
> *Data Recorder: " Died last year. 109 should be about 10 meters to your left. Last year it had 7 leaves and an aborted flower. After 109, there should be only one more plant in this plot."*

Repeating the observations back to one another and keeping careful track of the replicates helps to minimize collection errors. This "data talk" also keeps us alert and focused during long field days.

Errors resulting from broken or miscalibrated instruments are easier to detect (once it is determined that the instrument is broken or after it has been recalibrated). Unfortunately, instrument error can result in the loss of a large amount of data. Experiments that rely on automated data collection need to have additional procedures for checking and maintaining equipment to minimize data loss. It may be possible to collect replacement values for data resulting from equipment failures, but only if the failures are detected very soon after the data are collected. Potential equipment failures are another reason to transcribe rapidly your raw data into spreadsheets and evaluate their accuracy.

Errors in transcription result primarily from mistyping values into spreadsheets. When these errors show up as outliers, they can be checked against the original field data sheets. Probably the most common transcription error is the misplaced decimal point, which changes the value by an order of magnitude. When errors in transcription do not show up as outliers, they can remain undetected unless spreadsheets are proofread very carefully. Errors in transcription are less common when data are transferred electronically from instruments directly into spreadsheets. However, transmission errors can occur and can result in "frame-shift" errors, in which one value placed in the wrong column results in all of the subsequent values being shifted into incorrect columns. Most instruments have built-in software to check for transmission errors and report them to the user in real-time. The original data files from automated data-collection instruments should never be discarded from the instrument's memory until after spreadsheets have been checked carefully for errors.

Missing Data

A related, but equally important, issue is the treatment of missing data. Be very careful in your spreadsheets and in your metadata to distinguish between measured values of 0 and missing data (replicates for which no observations were recorded). Missing data in a spreadsheet should be given their own designation (such as the abbreviation "NA" for "not available") rather than just left as blank cells. Be careful because some software packages may treat blank cells as zeroes (or insert 0s in the blanks), which will ruin your analyses.

Detecting Outliers and Errors

We have found three techniques to be especially useful for detecting outliers and errors in datasets: calculating column statistics, checking ranges and precision of column values; and graphical exploratory data analysis. Other, more complex methods are discussed by Edwards (2000). We use as our example for outlier detection the measurement of heights of pitcher-leaves of *Darlingtonia californica*

(the cobra lily, or California pitcher plant).[11] We collected these data, reported in Table 8.1, as part of a study of the growth, allometry, and photosynthesis of this species. In this example of only 25 observations, the outliers are easy to spot simply by scanning the columns of numbers. Most ecological datasets, however, will have far more observations and it will be harder to find the unusual values in the data file.

COLUMN STATISTICS The calculation of simple column statistics within the spreadsheet is a straightforward way to identify unusually large or small values in the dataset. Measures of location and spread, such as the column mean, median, standard deviation, and variance give a quick overview of the distribution of the values in the column. The minimum and maximum values of the column may indicate suspiciously large or small values. Most spreadsheet software packages have functions that calculate these values. If you enter these functions as the last six rows of the spreadsheet (Table 8.1), you can use them as a first check on the data. If you find an extreme value and determine that it is a true error, the spreadsheet will automatically update the calculations when you replace it with the correct number.

CHECKING RANGE AND PRECISION OF COLUMN VALUES Another way to check the data is to use spreadsheet functions to ensure that values in a given column are within reasonable boundaries or reflect the precision of measurement. For example, *Darlingtonia* pitchers rarely exceed 900 mm in height. Any measurement much above this value would be suspect. We could look over the dataset carefully to see if any values are extremely large, but for real datasets with hundreds or thousands of observations, manual checking is inefficient and inaccurate. Most spreadsheets have logical functions that can be used for value checking and that can automate this process. Thus, you could use the statement:

If (value to be checked) > (maximum value), record a "1"; otherwise, record a "0"

to quickly check to see if any of the values of height exceed an upper (or lower) limit.



Darlingtonia californica

[11] *Darlingtonia californica* is a carnivorous plant in the family Sarraceniaceae that grows only in serpentine fens of the Siskiyou Mountains of Oregon and California and the Sierra Nevada Mountains of California. Its pitcher-shaped leaves normally reach 800 mm in height, but occasionally exceed 1 m. These plants feed primarily on ants, flies, and wasps.

TABLE 8.1 **Ecological data set illustrating typical extreme values *Darlingtonia californica***

	Plant #	Height	Mouth	Tube
	1	744	34.3	18.6
	2	700	34.4	20.9
	3	714	28.9	19.7
	4	667	32.4	19.5
	5	600	29.1	17.5
	6	777	33.4	21.1
	7	640	34.5	18.6
	8	440	29.4	18.4
	9	715	39.5	19.7
	10	573	33.0	15.8
A	**11**	**1500**	**33.8**	**19.1**
	12	650	36.3	20.2
	13	480	27.0	18.1
	14	545	30.3	17.3
	15	845	37.3	19.3
	16	560	42.1	14.6
	17	450	31.2	20.6
	18	600	34.6	17.1
	19	607	33.5	14.8
	20	675	31.4	16.3
	21	550	29.4	17.6
B	**22**	**5.1**	**0.3**	**0.1**
	23	534	30.2	16.5
	24	655	35.8	15.7
C	**25**	**65.5**	**3.52**	**1.77**
Mean		611.7	30.6	16.8
Median		607	33.0	18.1
Standard Deviation		265.1	9.3	5.1
Variance		70271	86.8	26.1
Minimum		5.1	0.3	0.1
Maximum		1500	42.1	21.1

The data are morphological measurements of individual pitcher plants. Pitcher height (in mm), pitcher mouth diameter (in mm), and pitcher tube diameter (in mm) were recorded for 25 individuals sampled at Days Gulch. Rows with extreme values in Figure 8.1 are shown in boldface and designated as A, B, and C. Summary statistics for each variable are given at the bottom of the table. It is crucial to identify and evaluate extreme values in a dataset. Such values greatly inflate variance estimates, and it is important to ascertain whether those observations represent measurement error, atypical measurements or simply natural variation of the sample.

This same method can be used to check the precision of the values. We measured height using a measuring tape marked in millimeters, and routinely recorded measurements to the nearest whole millimeter. Thus, no value in the spreadsheet should be more precise than a single millimeter (i.e., no value should have fractions of millimeters entered). The logical statement to check this is:

> If (value to be checked has any value other than 0 after the decimal point), record a "1"; otherwise, record a "0"

There are many similar checks on column values that you can imagine, and most can be automated using logical functions within spreadsheets.

Range checking on the data in Table 8.1 reveals that the recorded height of Plant 11 (1500 mm) is suspiciously large because this observation exceeds the 900-mm threshold. Moreover the recorded heights of Plant 22 (5.1 mm) and Plant 25 (65.5 mm) were misentered because the decimal values are greater than the precision of our measuring tapes. This audit would lead us to re-examine our field data sheets or perhaps return to the field to measure the plants again.

GRAPHICAL EXPLORATORY DATA ANALYSIS Scientific pictures—graphs—are one of the most powerful tools you can use to summarize the results of your study (Tufte 1986, 1990; Cleveland 1985, 1993). They also can be used to detect outliers and errors, and to illuminate unexpected patterns in your data. The use of graphics in advance of formal data analysis is called **graphical exploratory data analysis** or graphical EDA, or simply **EDA** (e.g., Tukey 1977; Ellison 2001). Here, we focus on the use of EDA for detecting outliers and extreme values. The use of EDA to hunt for new or unexpected patterns in the data ("data mining" or more pejoratively, "data dredging") is evolving into its own cottage industry (Smith and Ebrahim 2002).[12]

[12] The statistical objection to data dredging is that if we create enough graphs and plots from a large, complex dataset, we will surely find some relationships that turn out to be statistically significant. Thus, data dredging undermines our calculation of *P*-values and increases the chances of incorrectly rejecting the null hypothesis (Type I error). The philosophical objection to EDA is that you should already know in advance how your data are going to be plotted and analyzed. As we emphasized in Chapters 6 and 7, laying out the design and analysis ahead of time goes hand-in-hand with having a focused research question. On the other hand, EDA is essential for detecting outliers and errors. And there is no denying that you can uncover interesting patterns just by casually plotting your data. You can always return to the field to test these patterns with an appropriate experimental design. Good statistical packages facilitate EDA because graphs can be constructed quickly and easily and viewed on your computer screen.


```
  0   06
  1
  2
  3
  4   458
H 5   34567
M 6   00145577
H 7   01148
  8   4
  9
 10
 11
 12
 13
 14
 15   0
```

Figure 8.2 Stem-and-leaf plot. The data are measurements of the heights (in mm) of 25 individuals of *Darlingtonia californica* (Table 8.1). In this plot, the first column of numbers on the left—the "stem"—represent the "hundreds" place, and each "leaf" to its right represents the "tens" place of each recorded value. For example, the first row illustrates the observations 00x and 06x, or an observation less than 10 (the value 5.1 of Plant 22 of Table 8.1), and another observation between 60 and 70 (the value 65.5 of Plant 25). The second row illustrates observations 44x, 45x, and 48x, which correspond to values 440 (Plant 8), 450 (Plant 17), and 480 (Plant 13). The last row illustrates observation 150x, corresponding to value 1500 (Plant 11). All observations in column 1 of Table 8.1 are included in this display. The median (607) is in row 7, which is labeled with the letter M. The lower quartile (545), also known as a **hinge**, is in row 6, which is labeled with the letter H. The upper quartile or hinge (700) is in row 8, also labeled H. Like histograms and box plots, stem-and-leaf plots are a diagnostic tool for visualizing your data and detecting extreme data points.

There are three indispensable types of graphs for detecting outliers and extreme values: **box plots**, **stem-and-leaf plots**, and **scatterplots**. The first two are best used for univariate data (plotting the distribution of a single variable), whereas scatterplots are used for bivariate or multivariate data (plotting relationships among two or more variables). You have already seen many examples of univariate data in this book: the 50 observations of spider tibial spine length (see Figure 2.6 and Table 3.1), the presence or absence of *Rhexia* in Massachusetts towns (Chapter 2), and the lifetime reproductive output of orange-spotted whirligig beetles (Chapter 1) are all sets of measurements of a single variable. To illustrate outlier detection and univariate plots for EDA, we use the first column of data in Table 8.1—pitcher height.

Our summary statistics in Table 8.1 suggest that there may be some problems with the data. The variance is large relative to the mean, and both the minimum and the maximum value look extreme. A box plot (top of Figure 8.1) shows that two points are unusually small and one point is unusually large. We can identify these unusual values in more detail in a stem-and-leaf plot (Figure 8.2), which is a variant of the more familiar histogram (Chapter 1; Tukey 1977). The stem-and-leaf plot shows clearly that the two small values are an order of magnitude (i.e., approximately tenfold) smaller than the mean or median, and that the one large value is more than twice as large as the mean or median.

How unusual are these values? Some *Darlingtonia* pitchers are known to exceed a meter in height, and very small pitchers can be found as well. Perhaps these three observations simply reflect the inherent variability in the size of cobra lily pitchers. Without additional information, there is no reason to think that these values are in error or are not part of our sample space, and it would be inappropriate to remove them from the dataset prior to further statistical analysis.

We do have additional information, however. Table 8.1 shows two other measurements we took from these same pitchers: the diameter of the pitcher's tube, and the diameter of the opening of the pitcher's mouth. We would expect all of these variables to be related to each other, both because of previous research on *Darlingtonia* (Franck 1976) and because the growth of plant (and animal) parts tends to be correlated (Niklas 1994).[13] A quick way explore how these variables co-vary with one another is to plot all possible pair-wise relationships among them (Figure 8.3).

The **scatterplot matrix** shown in Figure 8.3 provides additional information with which to decide whether our unusual values are outliers. The two scatterplots in the top row of Figure 8.3 suggest that the unusually tall plant in Table 8.1 (Point A; Plant 11) is remarkably tall for its tube and mouth diameters. The plots of mouth diameter versus tube diameter, however, show that neither of these measurements are unusual for Plant 11. Collectively, these results suggest an error either in recording the plant's height in the field or in transcribing its height from the field datasheet to the computer spreadsheet.

[13] Imagine a graph in which you plotted the logarithm of pitcher length on the *x*-axis and the logarithm of tube diameter on the *y*-axis. This kind of allometric plot reveals the way that the shape of an organism may change with its size. Suppose that small cobra lilies were perfect scaled-down replicas of large cobra lilies. In this case, increases in tube diameter are synchronous with increases in pitcher length, and the slope of the line (β_1) = 1 (see Chapter 9). The growth is isometric, and the small plants look like miniatures of the large plants. On the other hand, suppose that small pitchers had tube diameters that were relatively (but not absolutely) larger than the tube diameters of large plants. As the plants grow in size, tube diameters increase relatively slowly compared to pitcher lengths. This would be a case of negative allometry, and it would be reflected in a slope value $\beta_1 < 1$. Finally, if there is positive allometry, the tube diameter increases faster with increasing pitcher length ($\beta_1 > 1$). These patterns are illustrated by Franck (1976). A more familiar example is the human infant, in which the head starts out relatively large, but grows slowly, and exhibits negative allometry. Other body parts, such as fingers and toes, exhibit positive allometry and grow relatively rapidly.

Allometric growth (either positive or negative) is the rule in nature. Very few organisms grow isometrically, with juveniles appearing as miniature replicas of adults. Organisms change shape as they grow, in part because of basic physiological constraints on metabolism and uptake of materials. Imagine that (as a simplification) the shape of an organism is a cube with a side of length *L*. The problem is that, as organisms increase in size, volume increases as a cubic function (L^3) of length (L), but surface area increases only as a square function (L^2). Metabolic demands of an organism are proportional to volume (L^3), but the transfer of material (oxygen, nutrients, waste products) is proportional to surface area (L^2). Therefore, structures that function well for material transfer in small organisms will not get the job done in larger organisms. This is especially obvious when we compare species of very different body size. For example, the cell membrane of a tiny microorganism does a fine job transferring oxygen by simple diffusion, but a mouse or a human requires a vascularized lung, circulatory system, and a specialized transport molecule (hemoglobin) to accomplish the same thing. Both within and between species, patterns of shape, size, and morphology often reflect the necessity of boosting surface area to meet the physiological demands of increasing volume. See Gould (1977) for an extended discussion of the evolutionary significance of allometric growth.

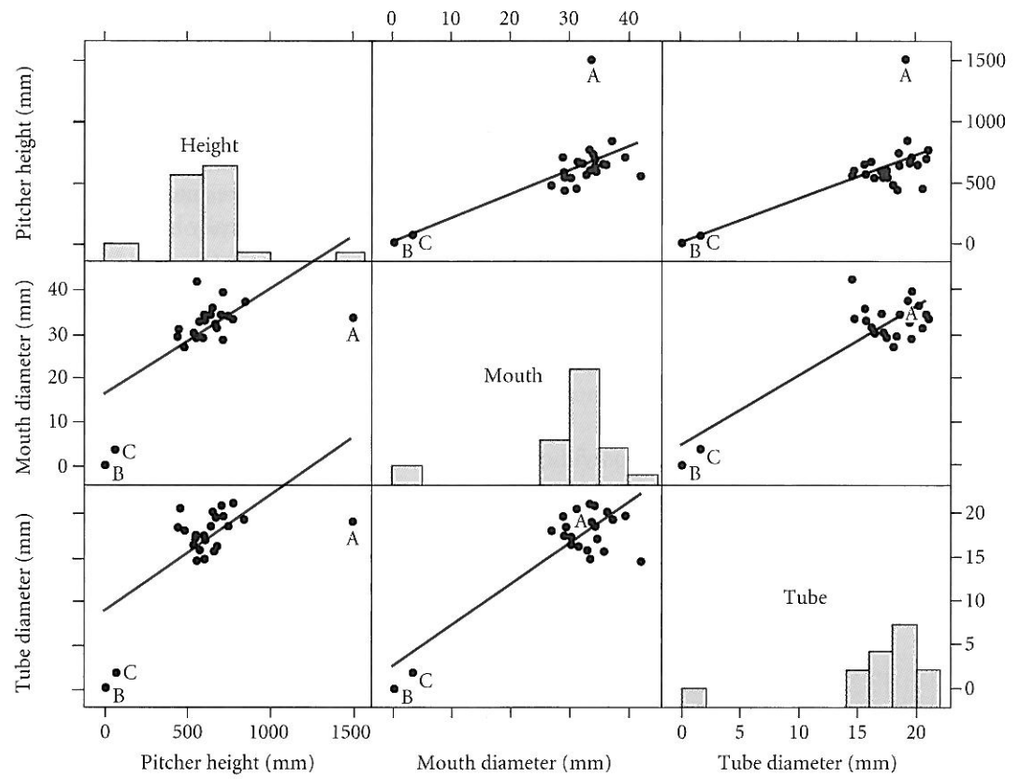


Figure 8.3 Scatterplot matrix illustrating the relationship between the height, mouth diameter, and tube diameter of 25 *D. californica* individuals (Table 8.1). The extreme values labeled in Figure 8.1 and Table 8.1 are also labeled here. The diagonal panels in this figure illustrate the histograms for each of the 3 morphological variables. In the off-diagonal panels, scatterplots illustrate the relationship between the two variables indicated in the diagonal; the scatterplots above the diagonal are mirror images of those below the diagonal. For example, the scatterplot at the lower left illustrates the relationship between pitcher height (on the *x*-axis) and tube diameter (on the *y*-axis), whereas the scatterplot at the upper right illustrates the relationship between tube diameter (on the *x*-axis) and pitcher height (on the *y*-axis). The lines in each scatterplot are the best-fit linear regressions between the two variables (see Chapter 9 for more discussion of linear regression).

In contrast, the short plants are not unusual in any regard except for being short. Not only are they short, but they also have small tubes and small mouths. If we remove Plant 11 (the unusually tall plant) from the dataset (Figure 8.4), the relationships among all three variables of the short plants are not unusual relative to the rest of the population, and there is no reason based on the data themselves to suspect that the values for these plants were entered incorrectly. Nonetheless, these two plants are very small (could we really have measured a

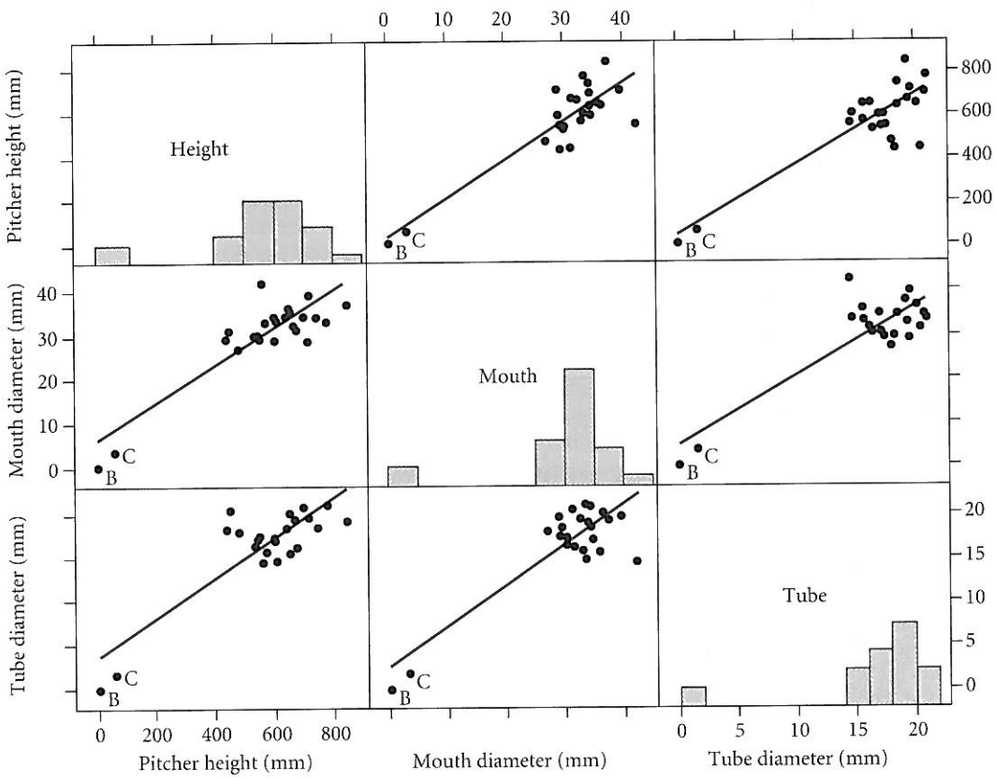


Figure 8.4 Scatterplot matrix illustrating the relationship between pitcher height, mouth diameter, and tube diameter. The data and figure layout are identical to Figure 8.3. However, the outlier Point A (Table 8.1) has been eliminated, which changes some of these relationships. Because Point A was a large outlier for pitcher height (height = 1500 mm), the scale of this variable (x-axis in the left-hand column) now ranges from only 0 to 800 mm, compared to 0 to 1500 mm in Figure 8.3. It is not uncommon to see correlations and statistical significance change based on the inclusion or exclusion of a single extreme data value.

5mm-tall plant?) and we should re-check our datasheets and perhaps return to the field to re-measure these plants.

Creating an Audit Trail

The process of examining a dataset for outliers and errors is a special case of the generic **quality assurance and quality control** process (abbreviated **QA/QC**). As with all other operations on a spreadsheet or dataset, the methods used for outlier and error detection should be documented in a QA/QC section of the metadata. If it is necessary to remove data values from the spreadsheet, a description of the values removed should be included in the metadata, and a new spreadsheet should be created that contains the corrected data. This spreadsheet also should be stored on permanent media and given a unique identifier in the data catalog.

The process of subjecting data to QA/QC and modifying data files leads to the creation of an **audit trail**. An audit trail is a series of files and their associated documentation that enables other users to recreate the process by which the final, analyzed dataset was created. The audit trail is a necessary part of the metadata for any study. If you are preparing an environmental impact report or other document that is used in legal proceedings, the audit trail may be part of the legal evidence supporting the case.[14]

The Final Step: Transforming the Data

You will often read in a scientific paper that the data were "transformed prior to analysis". Miraculously, after the transformation, the data "met the assumptions" of the statistical test being used and the analysis proceeded apace. Little else is usually said about the data transformations, and you may wonder why the data were transformed at all.

But first, what is a transformation? By a transformation, we simply mean a mathematical function that is applied to all of the observations of a given variable:

$$Y^* = f(Y) \tag{8.1}$$

[14] Historians of science have reconstructed the development of scientific theories by careful study of inadvertent audit trails—marginal notes scrawled on successive drafts of handwritten manuscripts and different editions of monographs. The use of word processors and spreadsheets has all but eliminated marginalia and yellow pads, but does allow for the creation and maintenance of formal audit trails. Other users of your data and future historians of science will thank you for maintaining an audit trail of your data.

Y represents the original variable, Y^* is the transformed variable, and f is a mathematical function that is applied to the data. Most transformations are fairly simple algebraic functions, subject to the requirement that they are **continuous monotonic functions**.[15] Because they are monotonic, transformations do not change the rank order of the data, but they do change the relative spacing of the data points, and therefore affect the variance and shape of the probability distribution (Chapter 2).

There are two legitimate reasons to transform your data before analysis. First, transformations can be useful (although not strictly necessary) because the patterns in the transformed data may be easier to understand and communicate than patterns in the raw data. Second, transformations may be necessary so that the analysis is valid—this is the "meeting the assumptions" reason that is used most frequently in scientific papers and discussed in biometry textbooks.

Data Transformations as a Cognitive Tool

Transformations often are useful for converting curves into straight lines. Linear relationships are easier to understand conceptually, and often have better statistical properties (see Chapter 9). When two variables are related to each other by multiplicative or exponential functions, the logarithmic transformation is one of the most useful data transformations (see Footnote 13). A classic ecological example is the species–area relationship: the relationship between species number and island or sample area (Preston 1962, MacArthur and Wilson 1967). If we measure the number of species on an island and plot it against the area of the island, the data often follow a simple **power function:**

$$S = cA^z \tag{8.2}$$

where S is the number of species, A is island area, and c and z are constants that are fitted to the data. For example, the number of species of plants recorded from each of the Galápagos Islands (Preston 1962) seems to follow a power relationship (Table 8.2). First, note that island area ranges over three orders of magnitude, from less than 1 to nearly 7500 km^2. Similarly, species richness spans two orders of magnitude, from 7 to 325.

[15] A continuous function is a function $f(X)$ such that for any two values of the random variable X, x_i and x_j, that differ by a very small number ($|x_i - x_j| < \delta$), then $|f(x_i) - f(x_j)| < \varepsilon$, another very small number. A monotonic function is a function $f(X)$ such that for any two values of the random variable X, x_i and x_j, if $x_i < x_j$ then $f(x_i) < f(x_j)$. A continuous monotonic function has both of these properties.

TABLE 8.2 **Species richness on 17 of the Galápagos Islands**

Island	Area (km^2)	No. of species	$\log_{10}$ (Area)	$\log_{10}$ (Species)
Albemarle (Isabela)	5824.9	325	3.765	2.512
Charles (Floreana)	165.8	319	2.219	2.504
Chatham (San Cristóbal)	505.1	306	2.703	2.486
James (Santiago)	525.8	224	2.721	2.350
Indefatigable (Santa Cruz)	1007.5	193	3.003	2.286
Abingdon (Pinta)	51.8	119	1.714	2.076
Duncan (Pinzón)	18.4	103	1.265	2.013
Narborough (Fernandina)	634.6	80	2.802	1.903
Hood (Española)	46.6	79	1.669	1.898
Seymour	2.6	52	0.413	1.716
Barringon (Santa Fé)	19.4	48	1.288	1.681
Gardner	0.5	48	–0.286	1.681
Bindloe (Marchena)	116.6	47	2.066	1.672
Jervis (Rábida)	4.8	42	0.685	1.623
Tower (Genovesa)	11.4	22	1.057	1.342
Wenman (Wolf)	4.7	14	0.669	1.146
Culpepper (Darwin)	2.3	7	0.368	0.845
Mean	526.0	119.3	1.654	1.867
Standard Deviation	1396.9	110.7	1.113	0.481
Standard Error of the Mean	388.8	26.8	0.270	0.012

Area (originally given in square miles) has been converted to square kilometers, and island names from Preston (1962) are retained, with modern island names given in parentheses. The last two columns give the logarithm of area and species richness, respectively. Mathematical transformations such as logarithms are used to better meet the assumptions of parametric analyses (normality, linearity, and constant variances) and they are used as a diagnostic tool to aid in the identification of outliers and extreme data points. Because both the response variable (species richness) and the predictor variable (island area) are measured on a continuous scale, a regression model is used for analysis (see Table 7.1).

If we plot the raw data—the number of species S as a function of area A (Figure 8.5)—we see that most of the data points are clustered to the left of the figure (as most islands are small). As a first pass at the analysis, we could try fitting a straight line to this relationship:

$$S = \beta_0 + \beta_1 A \tag{8.3}$$

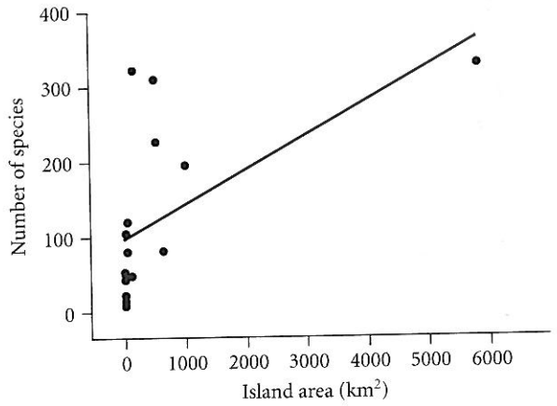


Figure 8.5 Plot of plant species richness as a function of Galápagos Island area using the raw data (Columns 2 and 3) of Table 8.2. The line shows the best-fit linear regression line (see Chapter 9). Although a linear regression can be fit to any pair of continuous variables, the linear fit to these data is not good: there are too many negative outliers at small island areas, and the slope of the line is dominated by Albermarle, the largest island in the dataset. In many cases, a mathematical transformation of the *X* variable, the *Y* variable, or both will improve the fit of a linear regression.

In this case, β_0 represents the intercept of the line, and β_1 is its slope (see Chapter 9). However, the line does not fit the data very well. In particular, notice that the slope of the regression line seems to be dominated by the datum for Albermarle, the largest island in the dataset. In Chapter 7, we warned of precisely this problem that can arise with outlier data points that dominate the fit of a regression line (Figure 7.3). The line fit to the data does not capture well the relationship between species richness and island area.

If species richness and island area are related exponentially (Equation 8.2), we can transform this equation by taking logarithms of both sides:

$$\log(S) = \log(cA^z) \tag{8.4}$$

$$\log(S) = \log(c) + z\log(A) \tag{8.5}$$

This transformation takes advantage of two properties of logarithms. First, the logarithm of a product of two numbers equals the sum of their logarithms:

$$\log(ab) = \log(a) + \log(b) \tag{8.6}$$

Second, the logarithm of a number raised to a power equals the power times the logarithm of the number:

$$\log(a^b) = b\log(a) \tag{8.7}$$

We can rewrite Equation 8.4, denoting logarithmically-transformed values with an asterisk (*):

$$S^* = c^* + zA^* \tag{8.8}$$

Thus, we have taken an exponential equation (Equation 8.2) and transformed it into a linear equation (Equation 8.8). When we plot the logarithms of the data, the relationship between species richness and island area is now much clearer (Figure 8.6), and the coefficients have a simple interpretation. The value for *z* in Equation 8.2 and 8.8, which equals the slope of the line in Figure 8.6, is 0.331; this means that every time we increase A^* (the logarithm of island area) by one unit (that is, by a factor of 10 as we used $\log_{10}$ to transform our variables in Table 8.2, and $10^1 = 10$), we increase species richness by 0.331 units (that is, by a factor of approximately 2 because $10^{0.331}$ equals 2.14). Thus, we can say that a

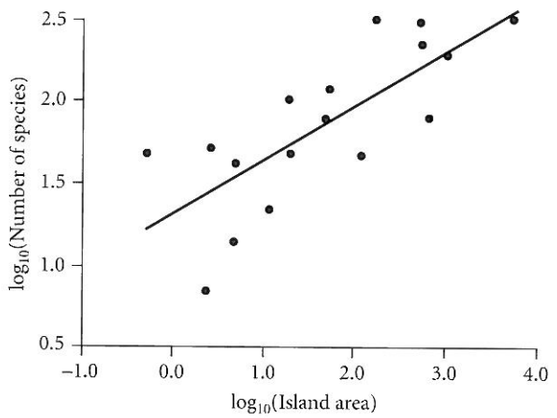


Figure 8.6 Plot of the logarithm of plant species richness as a function of the logarithm of Galápagos Island area (Columns 4 and 5 of Table 8.2). The line is the best-fit linear regression line (see Chapter 9). Compared to the linear plot in Figure 8.5, this regression fits the data considerably better: the largest island in the data set no longer looks like an outlier, and the linearity of the fit is improved.

10-fold increase in island area results in a doubling of the number of species present on the Galápagos Islands.[16]

Other transformations can be used to convert non-linear relationships to linear ones. For example, the cube-root transformation ($\sqrt[3]{Y}$) is appropriate for measures of mass or volume (Y^3) that are allometrically related to linear measures of body size or length (Y; see Footnote 13). In studies that examine relationships between two measures of masses or volumes (Y^3), such as comparisons of brain mass and body mass, both the X and Y variables are logarithmically transformed. The logarithmic transformation reduces variation in data that may range over several orders of magnitude (for a detailed exposition of the brain-body mass relationship, see Allison and Cicchetti 1976, and Edwards 2000).

[16] The history of the species-area relationship illustrates the dangers of **reification**: the conversion of an abstract concept into a material thing. The power function ($S = cA^z$) has formed the basis for several important theoretical models of the species-area relationship (Preston 1962; MacArthur and Wilson 1967; Harte et al. 1999). It also has been used to argue for making nature reserves as large as possible so that they contain the greatest possible number of species. This led to a long debate on whether a single large or several small preserves would protect the most species (e.g., Willis 1984; Simberloff and Abele 1984). However, Lomolino and Weiser (2001) have recently proposed that the species-area relationship has an asymptote, in which case a power function is not appropriate. But this proposal has itself been challenged on both theoretical and empirical grounds (Williamson et al. 2001). In other words, there is no consensus that the power function always forms the basis for the species-area relationship.

The power function has been popular because it appears to provide a good fit to many species-area datasets. However, a detailed statistical analysis of 100 published species-area relationships (Connor and McCoy 1979) found that the power function was the best-fitting model in only half of the datasets. Although island area is usually the single strongest predictor of species number, area typically accounts for only half of the variation in species richness (Boecklen and Gotelli 1984). As a consequence, the value of the species-area relationship for conservation planning is limited because there will be much uncertainty associated with species richness forecasts. Moreover, the species-area relationship can be used to predict only the number of species present, whereas most conservation management strategies will be concerned with the identity of the resident species. The moral of the story is that data can always be fit to a mathematical function, but we have to use statistical tools to evaluate whether the fit is reasonable or not. Even if the fit of the data is acceptable, this result, by itself, is rarely a strong test of a scientific hypothesis, because there are usually alternative mathematical models that can fit the data just as well.

Data Transformations because the Statistics Demand It

All statistical tests require that the data fit certain mathematical assumptions. For example, data to be analyzed using analysis of variance (Chapters 7 and 10) must meet two assumptions:

1. The data must be **homoscedastic**—that is, the residual variances of all treatment groups need to be approximately equal to each other.
2. The **residuals**, or deviations from the means of each group, must be normal random variables.

Similarly, data to be analyzed using regression or correlation (Chapter 9) also should have normally-distributed residuals that are uncorrelated with the independent variable.

Mathematical transformations of the data can be used to meet these assumptions. It turns out that common data transformations often address both assumptions simultaneously. In other words, a transformation that equalizes variances (Assumption 1) often normalizes the residuals (Assumption 2).

Five transformations are used commonly with ecological and environmental data: the logarithmic transformation, the square-root transformation, the angular (or arcsine) transformation, the reciprocal transformation, and the Box-Cox transformation.

THE LOGARITHMIC TRANSFORMATION The **logarithmic transformation** (or **log transformation**) replaces the value of each observation with its logarithm:[17]



John Napier

[17] Logarithms were invented by the Scotsman John Napier (1550–1617). In his epic tome *Mirifici logarithmorum canonis descriptio* (1614), he sets forth the rationale for using logarithms (from the English translation of 1616):

Seeing there is nothing (right well-beloved Students of the Mathematics) that is so troublesome to mathematical practice, nor that doth more molest and hinder calculators, than the multiplications, divisions, square and cubical extractions of great numbers, which besides the tedious expense of time are for the most part subject to many slippery errors, I began therefore to consider in my mind by what certain and ready art I might remove those hindrances.

As illustrated in Equations 8.4–8.8, the use of logarithms allows multiplication and division to be carried out by simple addition and subtraction. Series of numbers that are multiplicative on a linear scale turn out to be additive on a logarithmic scale (see Footnote 2, Chapter 3).

$$Y^* = \log(Y) \tag{8.9}$$

The logarithmic transformation (most commonly using the natural logarithm or the logarithm to the base e)[18] often equalizes variances for data in which the mean and the variance are positively correlated. A positive correlation means that groups with large averages will also have large variances (in ANOVA) or that the magnitude of the residuals is correlated with the magnitude of the independent variable (in regression). Univariate data that are positively skewed (skewed to the right) often have a few large outlier values. With a log-transformation, these outliers often are drawn into the mainstream of the distribution, which becomes more symmetrical. Datasets in which means and variances are positively correlated also tend to have outliers with positively-skewed residuals. A logarithmic transformation often addresses both problems simultaneously.

Note that the logarithm of 0 is not defined: regardless of base b, there is no number for which $a^b = 0$. One way around this problem is to add 1 to each observation before taking its logarithm (as $\log(1) = 0$, again regardless of base b). However, this is not a useful or appropriate solution if the dataset (or especially if one treatment group) contains many zeros.[19]

THE SQUARE-ROOT TRANSFORMATION The **square-root transformation** replaces the value of each observation with its square root:

$$Y^* = \sqrt{Y} \tag{8.10}$$

[18] Logarithms can be taken with respect to many "bases," and Napier did not specify a particular base in his *Mirifici*. In general, for a value a and a base b we can write

$$\log_b a = X$$

which means that $b^X = a$. We have already used the logarithm "base 10" in our example of the species-area relationship in Table 8.2; for the first line in Table 8.2, the $\log_{10}(7498) = 3.875$ because $10^{3.875} = 7498$ (allowing for round-off error). A change in one $\log_{10}$ unit is a power of 10, or an order of magnitude. Other logarithmic bases used in the ecological literature are 2 (the octaves of Preston (1962) and used for data that increase by powers of 2), and e, the base of the natural logarithm, or approximately 2.71828… e is a transcendental number, whose value was proven by Leonhard Euler (1707–1783) to equal

$$\lim_{n\to\infty}(1+\frac{1}{n})^n$$

The $\log_e$ was first referred to as a "natural logarithm" by the mathematician Nicolaus Mercator (1620–1687), who should not to be confused with the map-maker Gerardus Mercator (1512–1594). It is often written as ln instead of $\log_e$.

This transformation is used most frequently with count data – such as the number of caterpillars per milkweed or the number of *Rhexia* per town. In Chapter 2 we showed that such data often follow a Poisson distribution, and we pointed out that the mean and variance of a Poisson random variable are equal (both equal the Poisson rate parameter λ). Thus, for a Poisson random variable, the mean and variance vary identically. Taking square-roots of Poisson random variables yields a variance that is independent of the mean. Because the square root of 0 itself equals 0, the square-root transformation does not transform data values that are equal to 0. Thus, to complete the transformation, you should add some small number to the value prior to taking its square root. Adding 1/2 (0.5) to each value is suggested by Sokal and Rohlf (1995), whereas 3/8 (0.325) is suggested by Anscombe (1948).

THE ARCSINE OR ARCSINE-SQUARE ROOT TRANSFORMATION The **arcsine**, **arcsine-square root**, or **angular transformation** replaces the value of each observation with the arcsin of the square root of the value:

$$Y^* = \text{arcsine}\sqrt{Y} \tag{8.11}$$

This transformation is used principally for proportions (and percentages), which are distributed as binomial random variables. In Chapter 3, we noted that the mean of a binomial distribution = np and its variance = $np(1-p)$, where p is the probability of success and n is the number of trials. Thus, the variance is a direct function of the mean (the variance = $(1-p)$ times the mean). The arcsine transformation (which is simply the inverse of the sine function) removes this dependence. Because the sine function yields values only between –1 and +1, the inverse sine function can be applied only to data whose values fall

[19] Adding a constant before taking logarithms can cause problems for estimating population variability (McArdle et al. 1990), but it is not a serious issue for most parametric statistics. However, there are some subtle philosophical issues in how zeroes in the data are treated. By adding a constant to the data, you are implying that 0 represents a measurement error. Presumably, the true value is some very small number, but, by chance, you measured 0 for that replicate. However, if the value is a true zero, the most important source of variation may be between the absence and presence of the quantity measured. In this case, it would be more appropriate to use a discrete variable with two categories to describe the process. The more zeros in the dataset (and the more that are concentrated in one treatment group), the more likely it is that 0 represents a qualitatively different condition than a small positive measurement. In a population time series, a true zero would represent a population extinction rather than just an inaccurate measurement.

between −1 and +1: $-1 \leq Y_i \leq +1$. Therefore, this transformation is appropriate only for data that are expressed as proportions (such as p, the proportion or probability of success in a binomial trial). Two caveats to note with the arcsine transformation. First, if your data are percentages (0 to 100 scale), they must first be converted to proportions (0 to 1.0 scale). Second, the arcsine function gives transformed data in units of radians, not degrees.

THE RECIPROCAL TRANSFORMATION The **reciprocal transformation** replaces the value of each observation with its reciprocal:

$$Y^* = 1/Y \tag{8.12}$$

It is used most commonly for data that record rates, such as number of offspring per female. Rate data often appear hyperbolic when plotted as a function of the variable in the denominator. For example if you plot the number of offspring per female on the y-axis and number of females in the population on the x-axis, the resulting curve may resemble a hyperbola, which decreases steeply at first, then more gradually as X increases. These data are generally of the form

$$aXY = 1$$

(where X here is number of females and Y is number of offspring per female), which can be re-written as a hyperbola

$$1/Y = aX$$

Transforming Y to its reciprocal $1/Y$ results in a new relationship

$$Y^* = 1/(1/Y) = aX$$

which is more amenable to linear regression.

THE BOX-COX TRANSFORMATION We use the logarithmic, square root, reciprocal and other transformations to reduce variance and skew in the data and to create a transformed data series that has an approximately normal distribution. The final transformation is the **Box-Cox transformation**, or generalized power transformation. This transformation is actually a family of transformations, expressed by the equation

$$\begin{aligned} Y^* &= (Y^{\lambda} - 1)/\lambda \quad (\text{for } \lambda \neq 0) \\ Y^* &= \log_e(Y) \quad (\text{for } \lambda = 0) \end{aligned} \tag{8.13}$$

where λ is the number that maximizes the **log-likelihood function**:

$$L = -\frac{\nu}{2}\log_e(s_T^2) + (\lambda - 1)\frac{\nu}{n}\sum_{i=1}^{n}\log_e Y \tag{8.14}$$

where ν is the degrees of freedom, n is the sample size, and s_T^2 is the variance of the transformed values of Y (Box and Cox 1964). The value of λ that results when Equation 8.14 is maximized is used in Equation 8.13 to provide the closest fit of the transformed data to a normal distribution. Equation 8.14 must be solved iteratively (trying different values of λ until L is maximized) with computer software.

Certain values of λ correspond to the transformations we have already described. When $\lambda = 1$, Equation 8.13 results in a linear transformation (a shift operation; see Figure 2.7), when $\lambda = 1/2$, the result is the square-root transformation, when $\lambda = 0$, the result is the natural logarithmic transformation, and when $\lambda = -1$, the result is the reciprocal transformation. Before going to the trouble of maximizing Equation 8.14, you should try transforming your data using simple arithmetic transformations. If your data are right-skewed, try using the more familiar transformations from the series $1/\sqrt{Y}, \sqrt{Y}, \ln(Y), 1/Y$. If your data are left-skewed, try Y^2, Y^3, etc. (Sokal and Rohlf 1995).

Reporting Results: Transformed or Not?

Although you may transform the data for analysis, you should report the results in the original units. For example, the species–area data in Table 8.2 would be analyzed using the log-transformed values, but in describing island size or species richness, you would report them in their original units, **back-transformed**. Thus, the average island size is

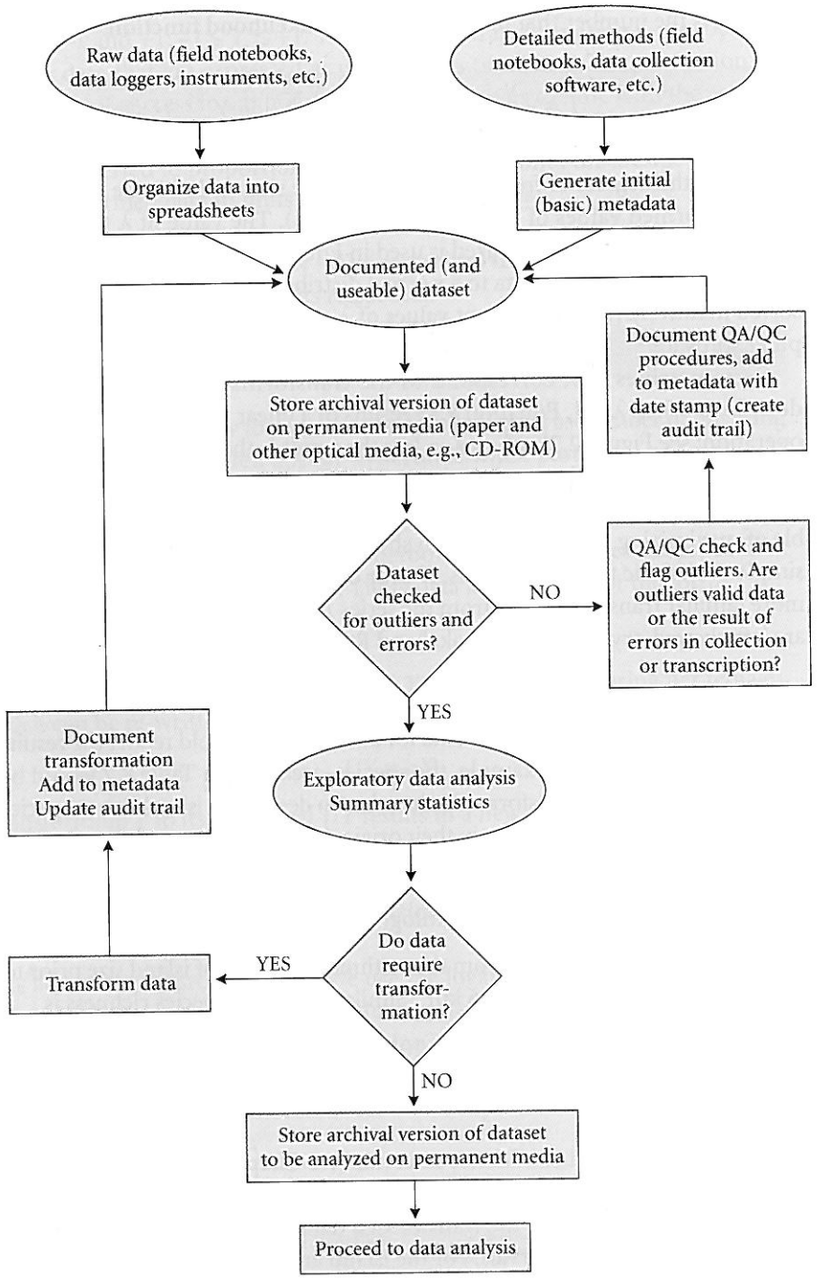
$$\text{antilog}(\overline{\log A}) = \text{antilog}(1.654) = 45.110 \text{ km}^2$$

Note that this is very different from the arithmetic average of island size prior to transformation, which equals 526 km^2. Similarly, average species richness is

$$\text{antilog}(\overline{\log S}) = 10^{1.867} = 73.6 \text{ species}$$

which differs from the arithmetic average of species richness prior to transformation, 119.3.

You would construct confidence intervals (see Chapter 3) in a similar manner, by taking antilogarithms of the confidence limits constructed using the standard errors of the means of the transformed data. This will normally result in asymmetric confidence intervals. For the island area data in Table 8.2, the stan-



◀ **Figure 8.7** The data management flowchart. This chart outlines the steps for managing, curating, and storing data. These steps should all be taken before formal data analysis begins. Documentation of metadata is especially important to ensure that the dataset will still be accessible and usable in the distant future.

dard error of the mean of $\log(A) = 0.270$. For $n = 17$, the necessary value from the t distribution, $t_{0.5[17]} = 2.110$. Thus, the lower bound of the 95% confidence interval (using Equation 3.16) $= 1.654 - 2.110 \times 0.270 = 1.084$. Similarly, the upper bound of the 95% confidence interval $= 1.654 + 2.110 \times 0.270 = 2.224$. On a logarithmic scale these two values form a symmetrical interval around the mean of 1.654, but when they are back-transformed, the interval is no longer symmetrical. The antilog of $1.084 = 10^{1.084} = 12.13$, whereas the antilog of $2.224 = 10^{2.224} = 167.49$, which is not symmetrical around the back-transformed mean of 45.11.

The Audit Trail Redux

The process of transforming data also should be added to your audit trail. Just as with data QA/QC, the methods used for data transformation should be documented in the metadata. Rather than writing over your data in your original spreadsheet with the transformed values, you should create a new spreadsheet that contains the transformed data. This spreadsheet also should be stored on permanent media, given a unique identifier in your data catalog, and added to the audit trail.

Summary: The Data Management Flow Chart

Organization, management, and curation of your dataset are essential parts of the scientific process, and must be done before you begin to analyze your data. This overall process is summarized in our data management flow chart (Figure 8.7).

Well-organized datasets are a lasting contribution to the broader scientific community, and their widespread sharing enhances collegiality and increases the pace of scientific progress. The free and public availability of data is a requirement of many granting agencies. Data collected with public funds distributed by organizations and agencies in the United States may be requested at any time by anyone under the provisions of the Freedom of Information Act. Adequate funds should be budgeted to allow for necessary and sufficient data management.

Data should be organized and computerized rapidly following collection, documented with sufficient metadata necessary to reconstruct the data collection, and archived on permanent media. Datasets should be checked carefully

for errors in collection and transcription, and for outliers that are nonetheless valid data. Any changes or modifications to the original ("raw") data resulting from error- and outlier-detection procedures should be documented thoroughly in the metadata; the altered files should not replace the raw data file. Rather, they should be stored as new (modified) data files. An audit trail should be used to track subsequent versions of datasets and their associated metadata.

Graphical exploratory data analysis and calculation of basic summary statistics (e.g., measures of location and spread) should precede formal statistical analysis and hypothesis testing. A careful examination of preliminary plots and tables can indicate whether or not the data need to be transformed prior to further analysis. Data transformations also should be documented in the metadata. Once a final version of the dataset is ready for analysis, it too should be archived on permanent media along with its complete metadata and audit trail.

PART III

Data Analysis

