Reprinted from: Scheiner, S. M., & J. Gurevitch. 1993. Design and analysis of ecological experiments. Chapman and Hall, New York.

2

Exploratory Data Analysis and Graphic Display

Aaron M. Ellison

2.1 Introduction

You have designed your experiment, collected the data, and are now confronted with a tangled mass of information that must be analyzed, presented, and published. Turning this heap of raw spaghetti into an elegant *fettucine alfredo* will be immensely easier if you can visualize the message buried in your data. Data graphics, the visual "display [of] measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color" (Tufte, 1983:9) provide the means for this visualization.

Graphics serve two general functions in the context of data analysis. First, graphics are a tool one can use to explore patterns in data prior to formal statistical analysis (Exploratory Data Analysis, or EDA sensu Tukey, 1977). Second, graphics communicate large amounts of information clearly, concisely, and rapidly, and illuminate complex relationships within datasets. Graphic EDA yields rough sketches to help guide you to appropriate, often counterintuitive formal statistical analyses. In contrast to EDA, presentation graphics are final illustrations suitable for publication. Presentation graphics of high quality can leave a lasting impression on readers or audiences, while vague, sloppy, or overdone graphics easily can obscure valuable information and engender confusion. Ecological researchers should view EDA and sound presentation graphic techniques as an essential component of data analysis, presentation, and publication.

This chapter provides an introduction to graphic EDA, and some guidelines for clear presentation graphics. More detailed discussions of these and related topics can be found in texts by Tukey (1977), Tufte (1983, 1990), and Cleveland (1985). These techniques are illustrated for univariate, bivariate, and classified quantitative (ANOVA) data sets that exemplify types of data sets encountered commonly in ecological research. Sample data sets are described briefly in Section

2.3; formal analyses of three of the illustrated data sets can be found in Chapters 13 (univariate data set), 8 (predator-prey data set), and 3 (ANOVA data set). You may find some of the graphics types presented unfamiliar or puzzling, but consider them seriously as alternatives for the more common bar charts, histograms, pie charts, etc. The majority of these graphs cannot be produced by SAS or SAS/Graph, the statistical package used in many chapters in this volume. Therefore, for each example, I describe in detail how it was constructed. All figures in this chapter were produced with an IBM PS/2-70 computer and a PostScript laser printer. I produced Figs. 2.8 and 2.9 with S-plus *DOS version 2.0* (Becker et al., 1988; Chambers and Hastie, 1992), and the remainder (save Fig. 2.1) using SYGRAPH *version 5.01* (the graphics module of SYSTAT: Wilkinson, 1990). Examples of SYGRAPH code used to construct the figures are given in Appendix 2.1.

2.1.1 Guiding Principles

The question or hypothesis guiding the experimental design should guide the decision as to which graphics are appropriate for exploring or illustrating the dataset. Sketching a mock graph, without data points, *prior* to beginning the experiment usually will clarify experimental design and alternative outcomes. This procedure also clarifies a priori hypotheses that will prevent inappropriately considering a posteriori hypotheses (suggested by EDA) as a priori ones. Often the simplest graph, without frills, is the best. However, graphs do not have to be simple-minded, conveying only a single type of information, and they need not be assimilated in a single glance. Tufte (1983) and Cleveland (1985) provide numerous examples of graphs that require detailed inspection before they reveal their messages. Besides the aesthetic and cognitive interest they provoke, complex graphs that are data and information rich can save publication costs and time in presentations. Regardless of the complexity of your illustrations, you should adhere to the following four guidelines in EDA and production graphics:

1. Underlying patterns of interest should be illuminated, while not compromising the integrity of the data.
2. The data structure should be maintained, so that readers can reconstruct the data from the figure.
3. Figures should have a high data:ink ratio and no chartjunk—"graphical paraphernalia routinely added to every display" (Tufte, 1983:107), including excessive shading, grid lines, ticks, special effects, and unnecessary three dimensionality.
4. Figures should not distort, exaggerate, or censor the data.

With the increasing availability of hardware and software able to digitize information directly out of published sources, adherence to these guidelines has

become increasingly important. Gurevitch (Chapter 17; Gurevitch et al., 1992), for example, relied extensively on information gleaned by digitizing data from many different published figures to explore common ecological effects across many experiments via meta-analysis. Readers will be better able to compare published data sets that are represented clearly and accurately.

2.2 Graphic Approaches

2.2.1 Exploratory Data Analysis (EDA)

Tukey (1977) established many of the principles of EDA, and his book is an indispensable guide to EDA techniques. You should view EDA as a first pass through your dataset prior to formal statistical analysis. EDA is particularly appropriate when there is a large amount of variability in the data (low signal-to-noise ratio) and when treatment effects are not immediately apparent. You can then proceed to explore, through formal analysis, the patterns illuminated with graphic EDA.

Since EDA is designed to illuminate underlying pattern in noisy data, it is imperative that the underlying data structure not be obscured or hidden completely in the process. Also, as EDA is the predecessor to formal analysis, it should not be time consuming. Personal computer-based packages such as SYSTAT, S-plus, and Sigma-plot permit rapid, *interactive* graphic construction with little of the effort needed for formal analysis. Finally, EDA should lead you to appropriate formal analyses and models. A common use of EDA is to determine if the raw data satisfy the assumptions of the statistical tests suggested by the experimental design (see Sections 2.3.1 and 2.3.4). Violation of assumptions revealed by EDA may lead to use of different statistical models from those you had intended to employ a priori. For example, Antonovics and Fowler (1985) found unanticipated effects of planting position in their studies of plant competitive interactions in hexagonal planting arrays. These results led to a new appreciation for neighborhood interactions in plant assemblages (e.g., Czárán and Bartha, 1992).

2.2.2 Production Graphics

Graphics are an essential medium of communication in scientific literature and at seminars and meetings. In a small amount of space or time, it is imperative to get out the message and fix it clearly and memorably in the audience's mind. Numerous authors have investigated and analyzed how individuals perceive different types of graphs, and what make 'good' and 'bad' graphs from a cognitive perspective (reviewed concisely by Wilkinson, 1990; and in depth by Cleveland, 1985). It is not my intention to review this material; rather, through example, I hope to change the way we as ecologists display our data to maximize the amount of information communicated while minimizing distraction.

Cleveland (1985) presented a hierarchy of graphic elements used to construct data graphics that satisfy the guidelines suggested in Section 2.1.1 (Fig. 2.1). Although there is no simple way to distinguish 'good' graphics from 'bad' graphics, we can derive general principles from Cleveland's ranking. First, color, shading, and other chartjunk effects do not as a rule enhance the information content of graphs. They may look snazzy in a seminar, but they lack substance,

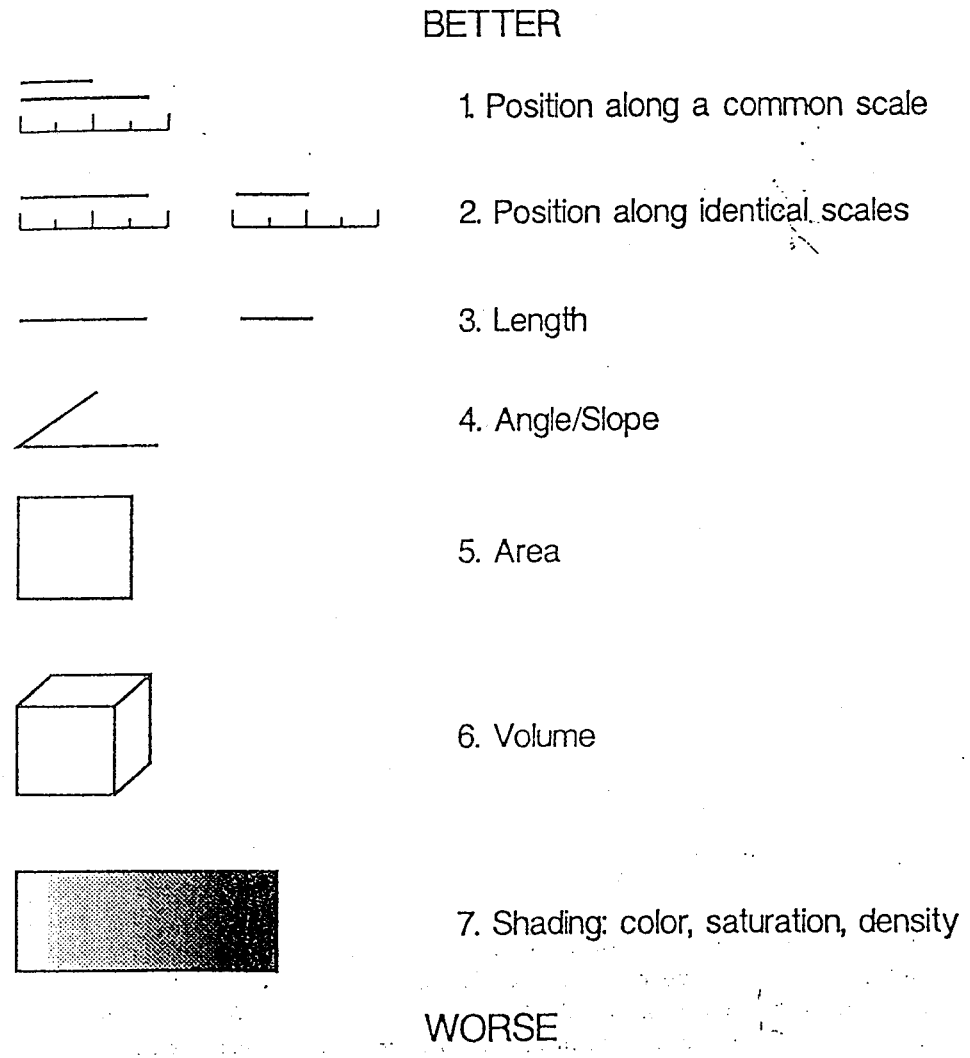


Figure 2.1. Ordering of graphic features according to their relative accuracy in representing quantitative variation (after Cleveland, 1985).

and use a lot of ink. Second, three-dimensional graphs that are mere extensions of two-dimensional ones (e.g., ribbon charts, three-dimensional histograms, or pie charts) not only do not increase the information content available, but often obscure the message (a dramatic, if unfortunate set of examples can be found in Benditt, 1992). These graphics, common in business presentations and increasingly rife at scientific meetings, violate all of the suggested guidelines. Finally, more dimensions often are used than are needed; e.g., "areas" and lines where a point would do. Simken and Hastie (1987) discuss exceptions to Cleveland's graphic hierarchy. In general, when designing graphics, adhere to the Shaker maxim: form follows function.

High-quality graphical elements can be assembled into effective graphic displays of data (Cleveland, 1985). First, emphasize the data. Lines drawn through data points should not hide the points themselves. Second, data points should never lie on axes themselves, as the axes can obscure data points. If, for example, there are many points that would fall along a 0 line, then extend that axis beyond 0 (Fig. 2.6). Third, reference lines, if needed (which they rarely are) should be deemphasized relative to the data. This can be accomplished with different line types (variable thicknesses; dotted, dashed, or solid, etc.) or shading. Fourth, overlapping data symbols or data sets should be clearly distinguishable. You can increase data visibility and minimize overlap by varying symbol size or position, separating datasets to be compared onto multiple plots, or changing from arithmetic to logarithmic scales. Exemplars include the jitter plot, which avoids overlap of identical values (Fig. 2.3B), and spreading of responses to categories across an axis (Fig. 2.12D). Fifth, the plot must be easily readable following reduction for publication or when projected as a slide to a seminar audience. Finally, Cleveland recommends using a full rectangular plot frame, not the more common bottom axis/left axis only combination seen in many papers. This, together with tick marks *outside* the plot frame (1) emphasize the data and (2) help the reader accurately place individual data points. Tufte (1983) disagrees, as the extra axes are an excessive use of ink and convey no information. Examples in this chapter illustrate most of these possibilities. In the final analysis, many of these rules reflect not only insight into cognitive perception, but also aesthetic judgments by you, the author.

From the above discussion, we could ask, isn't all this too much trouble? Should we dispense with graphs altogether in favor of tables? Because of their conciseness, graphics are almost always preferable in oral presentations. Graphs illustrate more clearly relationships among variables, and can display rapidly multivariate information. However, where exact values are important (as in final publications), tables are more precise. The need for precise tables has been obviated by the increasing availability of digitizing software. When presenting data graphically, however, you must present unbiased and uncensored data. A discussion of what data should be provided, in either graphs or tables, follows in Section 2.4.

2.3 Examples

2.3.1 Univariate Data: Frequency (Density) Distributions

Distributions of height, biomass, or other size metrics are often the primary descriptor of populations or communities. As an example of size distributions, I use a data set containing the number of leaf nodes of 75 *Ailanthus altissima* plants. The experimental design and formal analysis of these data are given in Chapter 13.

With univariate data, two questions are paramount: (1) how are the data distributed (including summary statistics such as the mean, variance, median) and (2) are the data normally distributed or can they be transformed to make them amenable to parametric analyses? Investigators often explore these questions via histograms or normality plots.

A histogram is an example of a *density* plot; that is, what is illustrated in each bar is the frequency, or density, of the values occurring in the dataset between the lower bound and the upper bound of each bar. Histograms are commonly confused with bar charts (see Section 2.3.4). The latter are used to illustrate some summary measure (often the mean, sum, or percent) of all the values within a given treatment category. Histograms of the *Ailanthus* data are shown in Fig. 2.2.

For three reasons a histogram is not the best method for answering the above two questions. First, the raw data are hidden. In this example, there are 75 plants, which have been divided into 12 biomass groups, or *bins* (Fig. 2.2A). It is impossible to know, for example, if the third bar (range 12–14 nodes) contains 10 observations of 12 nodes, 10 observations of 14 nodes, or any other of the possible combinations of 12–14 nodes into 10 observations. Second, the division into 12 bins is arbitrary; it was the default of the graphics program. One could just as easily use 24 or 6 bins, both of which change the apparent shape of the distribution (Figs. 2.2B,C) without conveying additional information. Third, summary statistics cannot be computed from the data illustrated in the histogram. Thus, a histogram does not enable one to answer key questions about univariate data. In addition, histograms fall low on Cleveland's hierarchy of graphic primitives. Bars in a histogram use vertical lines, horizontal lines, and shading in concert to present information embodied in the single point indicated by the top of the bar.

Tukey (1977) introduced the stem-and-leaf diagram as the simplest alternative to the histogram (Fig. 2.3A). The main advantage of the stem-and-leaf diagram is that the raw data are presented in toto. Summary statistics can be derived easily from or incorporated into the figure. Nevertheless, stem-and-leaf diagrams suffer visually from one of the same drawbacks as histograms: the number of bins is arbitrary. Two other alternatives to histograms are jitter plots (Fig. 2.3B) and dit plots (Fig. 2.3C). These two figures preserve the underlying data structure (all

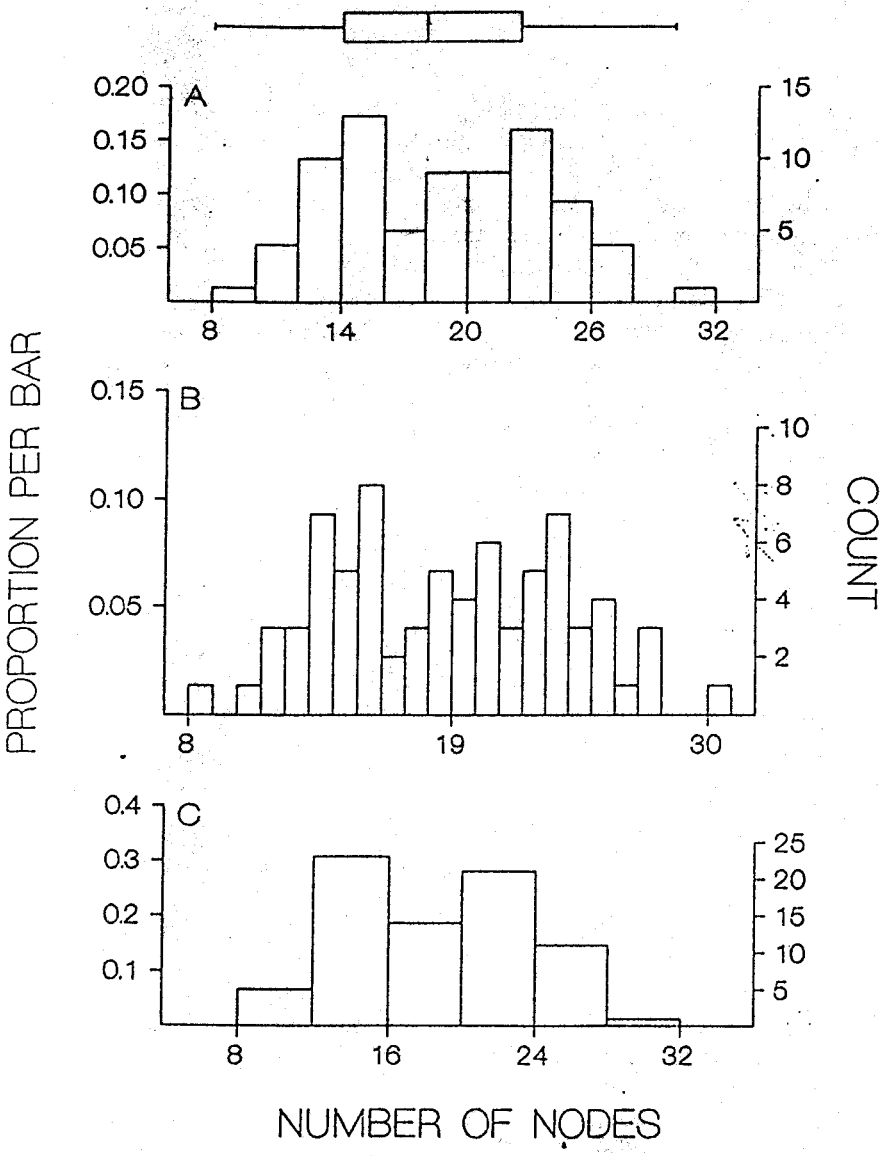


Figure 2.2. Histograms of the number of nodes per plant of 75 surviving *Ailanthus altissima* individuals grown in a 5 × 20 plant rectangular array. Each bar represents the frequency or count (right axis) of observations within the bounds indicated by the ticks on the *x*-axis, and the proportion of the total sample (left axis) represented by each bar. The three plots illustrate the variation in histogram presentation obtained by changing the bin width: (A) default (bin width=4); (B) bin width=2; (C) bin width = 8. At the top of the figure, a box plot (see Fig. 2.4 for construction details) illustrates summary statistics and a better indication of the true data distribution.

values are presented), do not use arbitrary bins, and can be constructed quickly without additional preparation (e.g., sorting) of the data set. Both plots permit rapid assessment of density patterns and are simple to understand.

Stem-and-leaf plots and the density diagrams presented in Fig. 2.3 can be used as simple alternatives to histograms. However, these plots do not convey clearly some of the information that ecologists may want to communicate, and it is difficult to compare the information in two or more of these plots. I suggest the box-and-whisker plot (Tukey, 1977), often called simply a box plot, as a presentation alternative to the univariate histogram (Figs. 2.2 and 2.4A). An advantage of the box plot is that it provides more summary statistical information than a histogram—it includes medians, quartiles, ranges, and outliers (extreme variates)—in much less space and with much less ink. Box plot construction is not dependent on arbitrary bins, so these plots do not exaggerate or distort the data distribution. By notching the box plot (Fig. 2.12E), one can easily add confidence intervals so that plots of several distributions can be compared easily.

Wilkinson (1990; Haber and Wilkinson 1982) developed the fuzzygram (Fig. 2.4B), another alternative to the histogram. Fuzzygrams are histograms with probability distributions superimposed on each bar. Consequently, fuzzygrams present not only the data, but also some estimation of how realistically they represent the actual population distribution. Such a presentation is particularly useful in concert with results derived from sensitivity analyses (Ellison and Bedford, 1991) or resampling methods (Efron, 1982; Dixon, Chapter 13). Haber and Wilkinson (1982) discuss, from a cognitive perspective, the merits of fuzzygrams and other density plots relative to traditional histograms. Histograms (Fig. 2.2), stem-and-leaf plots (Fig. 2.3A), dit plots (Fig. 2.3C), and fuzzygrams (Fig. 2.4B) can indicate possible bimodality in the data. Bimodal data, observed commonly in plant ecology, are obscured by box plots and jittered density diagrams.

Probability plots are common features of most statistical packages, and provide a visual estimate of whether or not the data fit a given distribution. The most common probability plot is the normal probability plot (Fig. 2.5A). Here, the observed values are plotted against their expected values if the data came from a normal distribution; if the data are derived from an approximately normal distribution, the points will fall along a relatively straight diagonal line. There are also numerical statistical tests for normality (e.g., Sokal and Rohlf, 1981; Zar, 1984). If, for biological reasons, the investigator believes the data come from a population with a known distribution different from a normal one, it is similarly possible to construct probability plots for other distribution functions (Fig. 2.5B).

2.3.2 Bivariate Data: Examining Relationships Between Variables

Ecological experiments often explore relationships between two or more continuous variables. Two general questions related to bivariate data can be addressed

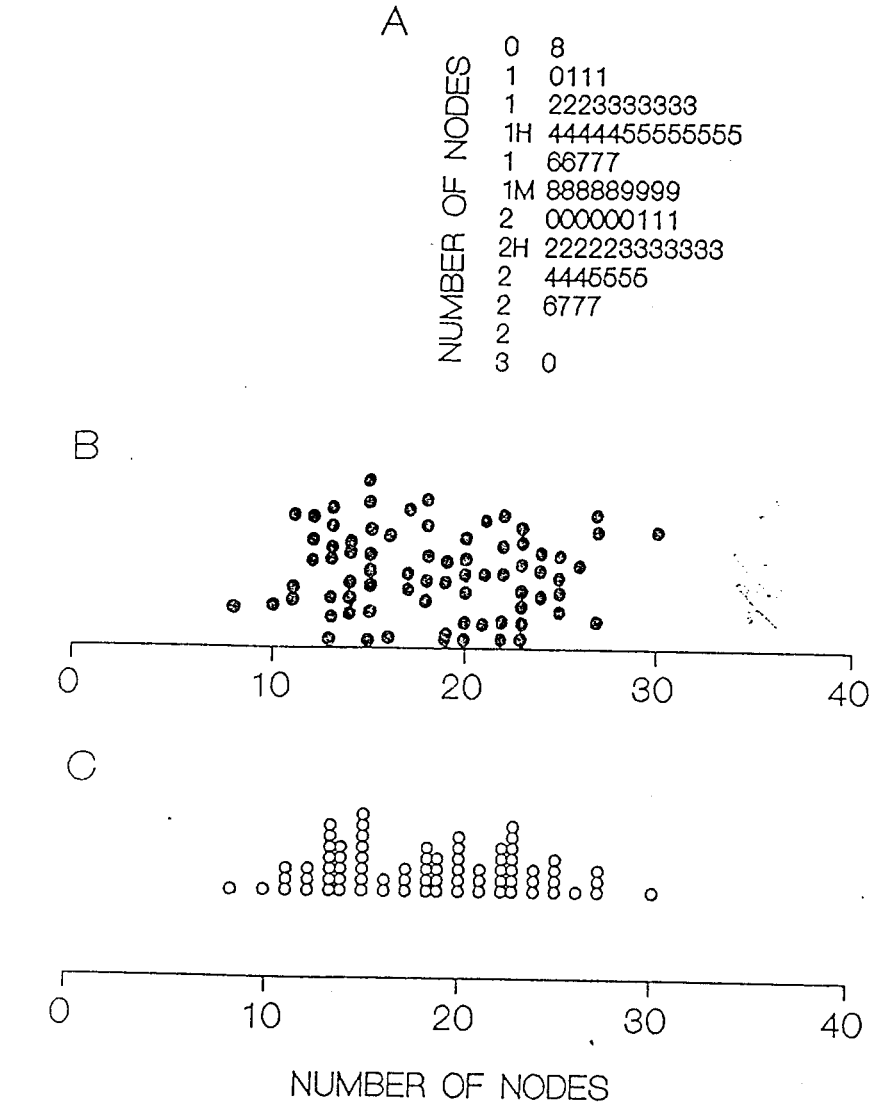


Figure 2.3. Alternative density plots that convey more information than a histogram. (A) A stem-and-leaf plot. In this plot, each line is a *stem*, and the each datum on a stem is a *leaf*. The label for the stem is the first digit (*starting part*) of the number, followed by the value of the leaf. On the first line, the starting part is 0 and the only leaf is 8, indicating a value of 08 nodes. On the second line, the starting part is 1, and there are four leaves, indicating four data points: 10, 11, 11, and 11 nodes. The location of the sample median (M) and upper and lower quartiles (H) are also marked on this plot. (B) A jittered density plot. Each point is placed along the horizontal scale at the exact location of its value. To keep points with equal value from overlapping, they are located at random heights above the x-axis. (C) A dit plot. Each point indicates an individual observation, stacked up the y-axis at its location along the x-axis. In essence, a dit plot is a stem-and-leaf plot with symbols substituted for leaves.

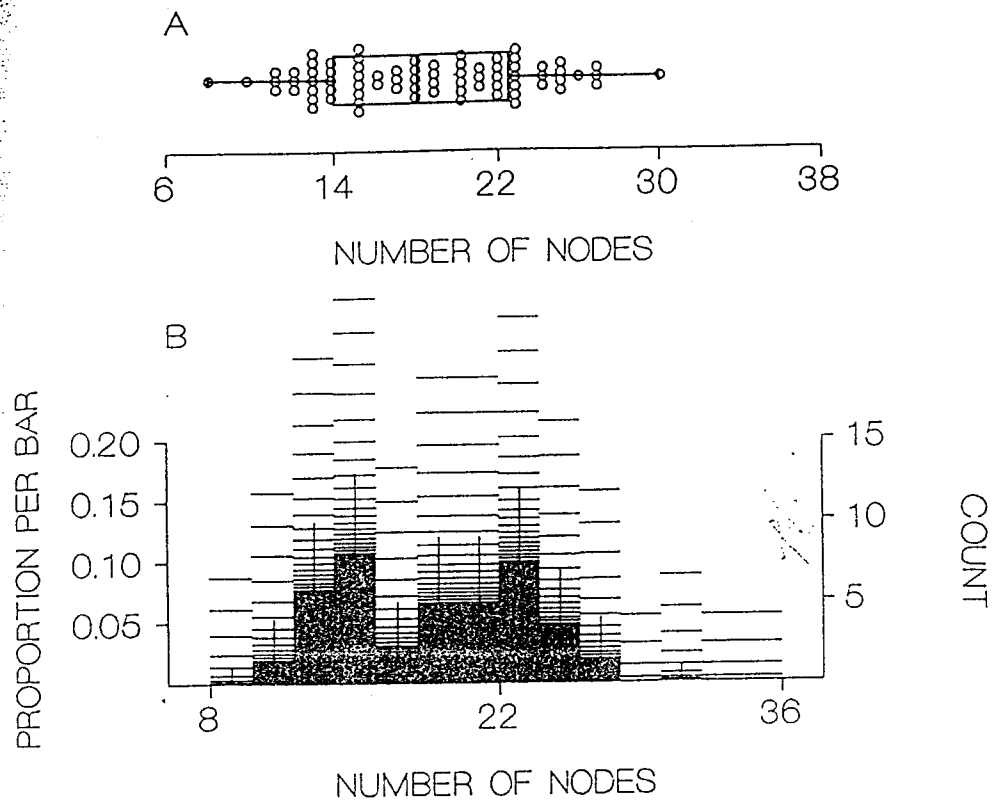


Figure 2.4. Information-rich production alternatives to histograms. (A) A box-and-whisker plot. The vertical line in the center of the box plot indicates the *sample median*. The left and right vertical sides of the box indicate respectively the location of the 25th and 75th percentile of the data (*lower and upper quartiles*, or *hinges*). The absolute value of the distance between the hinges (obtained by subtracting the value of the lower quartile from the value of the upper quartile) is the *hspread*. The whiskers of the box extend *to the last point* occurring between each hinge and its *inner fence*, a distance 1.5 *hspreads* from the hinge. Two kinds of outliers can be distinguished on a box plot. Points occurring between 1.5 hspreads and 3 hspreads (the outer fence) are indicated by an asterisk (see Fig. 2.12E). Points occurring greater beyond the outer fence are indicated by an open circle. The various summary statistics are clearly seen in relation to the raw data, which are overlain on this box plot as a symmetric dit plot. The distance encompassed by the whiskers includes ≈90% of the data (Norusis, 1990). (B) A fuzzygram (Wilkinson, 1990). This plot is a standard histogram (counts and proportions of each bin indicated by the height of the vertical line), with a probability distribution superimposed on each bar. The shading of the bars is based on a gray-scale distribution according to the probability that the ith observation will occur in that region: $P_i = P(p_i > \pi_i)$, where $p_i = n_i/n$ is the sample estimate of π_i (the expected proportion of a sample of n values from a continuous distribution to fall in the ith bin of the histogram. The more likely that $p_i > \pi_i$, the lighter the bar. Consequently, for large sample sizes, the bars will appear in sharp focus, while for small counts, the bars will be fuzzy. See Haber and Wilkinson (1982) for a discussion of the cognitive perception of fuzzygrams.

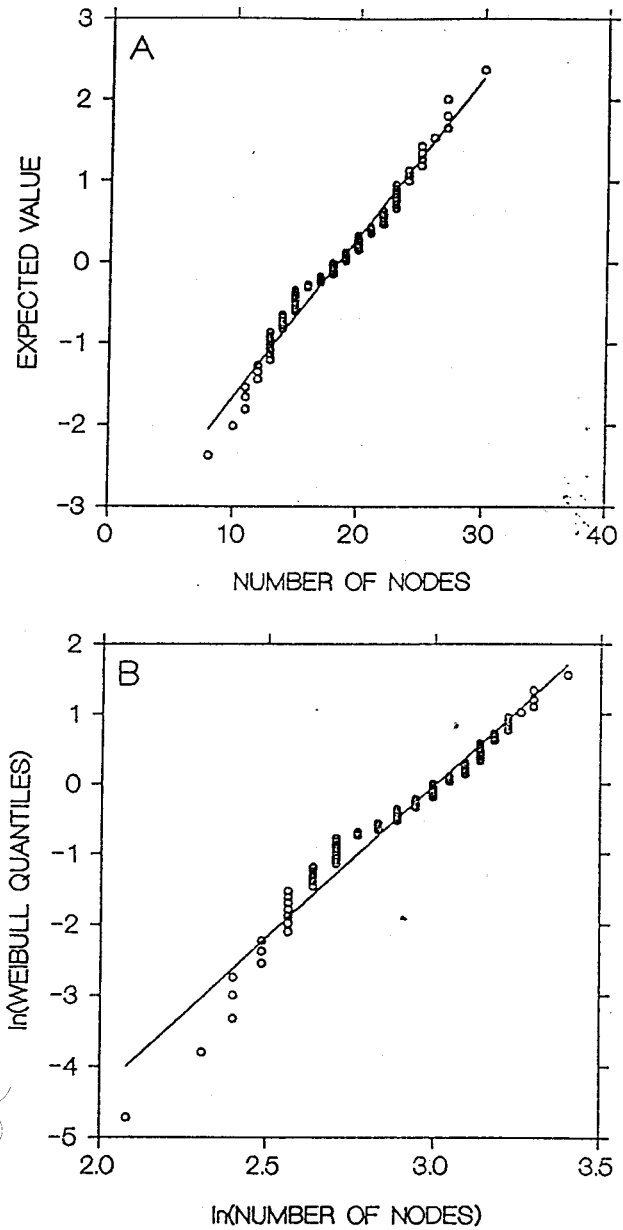


Figure 2.5. Probability plots of the *Ailanthus* data. (A) A normal probability plot. (B) A probability plot with the predicted values coming from a Weibull distribution: $f(y) = 1 - \exp[(-y/s)^t]$, where s is a spread parameter and t is a shape parameter. In this probability plot, the slope of the line is an estimate of $1/t$, and the intercept is an estimate of $\ln(s)$. See Gnanadesikan (1977) for a general discussion of probability plots.

with graphical EDA: (1) what is the general relationship between the two variables and (2) what point(s) is (are) outliers—those points that affect disproportionately the apparent relationship between the two variables? The answers to these questions lead, in formal analyses, to investigations of the strength and significance of the relationship (Chapters 6, 8, 9, and 10). Scatterplots and generalized smoothing routines are illustrated here for exploring and presenting bivariate data. Extensions of these techniques to multivariate data are presented in Section 2.3.3.

Bivariate data sets can be grouped into two types: those where there is a priori knowledge as to which variable ought to be considered independent, leading one to consider formal regression models (Chapters 8 and 9), and those where such a priori knowledge is lacking, leading one to examine correlation coefficients, and subsequent a posteriori analyses. The functional response of *Notonecta glauca*, a predatory aquatic hemipteran, presented experimentally with varying numbers of the isopod *Asellus aquaticus* is used to illustrate the first type of data set; these data are described in detail in Chapter 8. For the latter type of data, I use a dataset consisting of the height, diameter at breast height (dbh), and distance to nearest neighbor of 41 trees in a 625 m^2 plot within an ≈75-year-old mixed hardwood stand in South Hadley, Massachusetts (A. M. Ellison, unpublished data). Data sets of this type are commonly used to construct forestry yield tables (e.g., Tritton and Hornbeck, 1982), and have been used to infer competitive interactions among trees (e.g., Weller, 1987) and forest successional dynamics (e.g., Horn et al., 1989).

For both exploration and presentation, scatterplots are the most straightforward way of displaying bivariate data (Fig. 2.6A). However, scatterplots are merely a display, they do not necessarily reveal pattern. Figure 2.6A illustrates clearly this idea. Three functional response curves (Holling, 1966; Juliano, Chapter 8) could be fit to these data, and it is not clear from the scatterplot itself which one would best fit the data. EDA is particularly useful for dealing with these data, which show high variability and no obvious best relationship between the two variables.

Recent computer-intensive innovations in smoothing techniques (reviewed by Efron and Tibshirani, 1991) have expanded the palette of smoothers developed by Tukey (1977). Basically, to construct a smoothed curve through the data, a best-fit line is constructed through a subset of the data, local to each point along the x-axis. This process is repeated for each point, and a smooth line is constructed by connecting up the intersections of each local regression line. The result of this process, using lowess (robust *lo*cally weighted reg*ression*: Cleveland, 1979; Efron and Tibshirani, 1991), is shown for the predator-prey data in Fig. 2.6B. In this case, 50% of the data were used to construct each segment of the smoothed curve. That is, to construct the first segment, the response data from $0 \leq N_0 \leq 50$ were used; to construct the second segment, the response data from $1 \leq N_0 \leq 51$ were used, etc. The apparent type III functional response observed in the smoothed curve is supported by the formal analysis of these data (Chapter 8). The lack of underlying assumptions about the distribution and variance of the data and the ability to elucidate pattern from within very noisy data are two advantages of

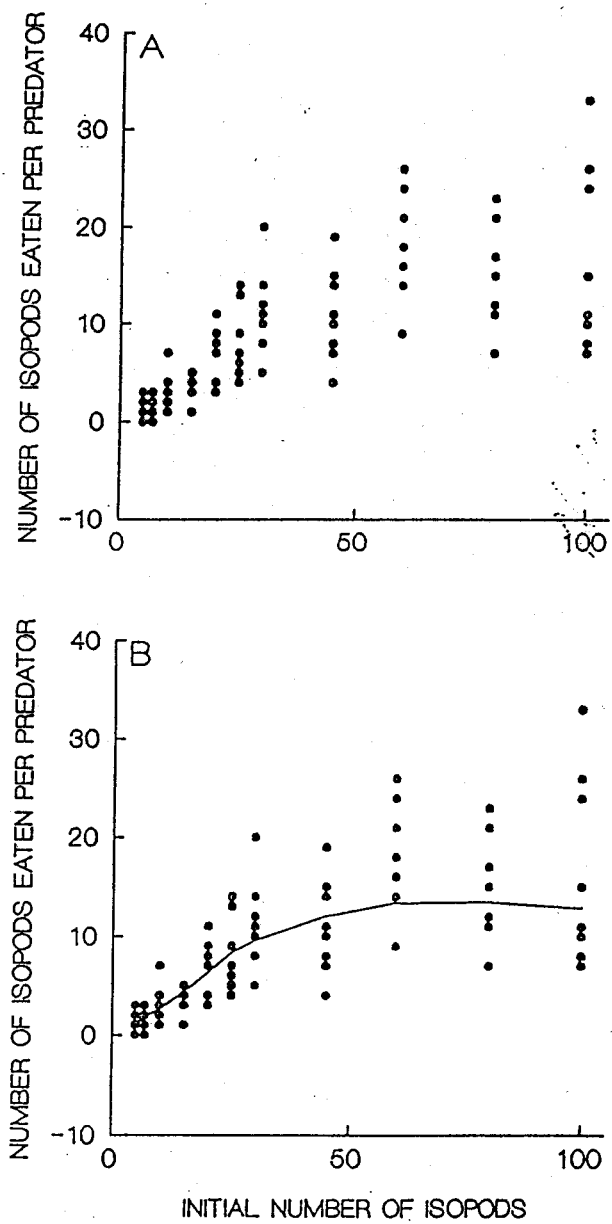


Figure 2.6. Scatterplots of the functional response of *Notonecta* to varying levels of *Asellus*. (A) A simple scatterplot showing the raw data. (B) A scatterplot with a lowess smooth fitted to the data. Note the apparent type III functional response revealed by the smoother (see Chapter 8).

smoothing over traditional regression techniques. Disadvantages of smoothing are that relative weighting of data used for each segment needs to be specified in advance, usually with little or no rational basis for the decision. Moreover, statistical comparison of different smoothed curves is virtually impossible. Finally, with the exception of Tukey's (1977) 3R smoother based on medians, virtually all smoothed curves require sophisticated software (e.g., SYSTAT, Stata, Statistica, Minitab, and S-Plus; see Ellison, 1992 for a review) and fast computers for accurate construction.

Smoothers are used appropriately only when there is clear a priori knowledge of an independent variable and a corresponding dependent variable or variables. When this is not the case, other exploratory techniques are more appropriate for examining relationships between variables. In addition, smoothing does not provide information about potential outliers in the data set. For examining correlations between variables, and to search a posteriori for outliers, influence plots and convex hulls are useful exploratory tools.

A scatterplot of the relationship between tree height and stem diameter (A. M. Ellison, unpublished data) is illustrated in Fig. 2.7A. The raw data are shown, and there appears to be an apparent outlier (a 30-m-tall tree with a dbh > 70 cm). In an influence plot of these data (Fig. 2.7B), the size of each point becomes directly proportional to the magnitude of the change its removal would have on the Pearson correlation coefficient (r) between the two variables. By overlaying a bivariate 50% confidence ellipse, it becomes obvious that outlying points have greater influence on r than do points within the ellipse.

In an influence plot of the logarithmically transformed data (Fig. 2.7C) the apparent outliers have all but disappeared (the large outlier in Fig. 2.7B now has an influence on r of only .01), and the data are better distributed for formal analysis. Fig. 2.7D supports this notion. The outer ellipse is a 95% confidence ellipse centered on the sample (dbh and height) means, with the ellipses' major and minor axes equal in length to the unbiased *sample* standard deviations of height and dbh, respectively. The orientation of the ellipse is determined by the sample covariance. All of the points, save the apparent outlier, fall within this confidence ellipse. For comparison, the inner ellipse is a 95% confidence ellipse with axes computed from the standard errors of the means of each variable and centered on the sample centroid—a graphic illustration of the real difference between the standard deviation and the standard error (see Section 2.4).

Convex hulls and subsequent peeled convex hulls (Barnett, 1976) are useful exploratory tools when the distribution underlying the data is not normal or not known. Convex hulls illustrate order in bivariate or multivariate data, and are used to distinguish distinct groups, outliers, and general shapes of multivariate distributions (a detailed discussion is given in Barnett, 1976). Peeled convex hulls are essentially bivariate smoothers. Figure 2.8 illustrates a convex hull and a subsequent peel around the same data set illustrated in Fig. 2.7. The initial hull (Fig. 2.8A) describes the boundaries of the data—it encompasses the full range

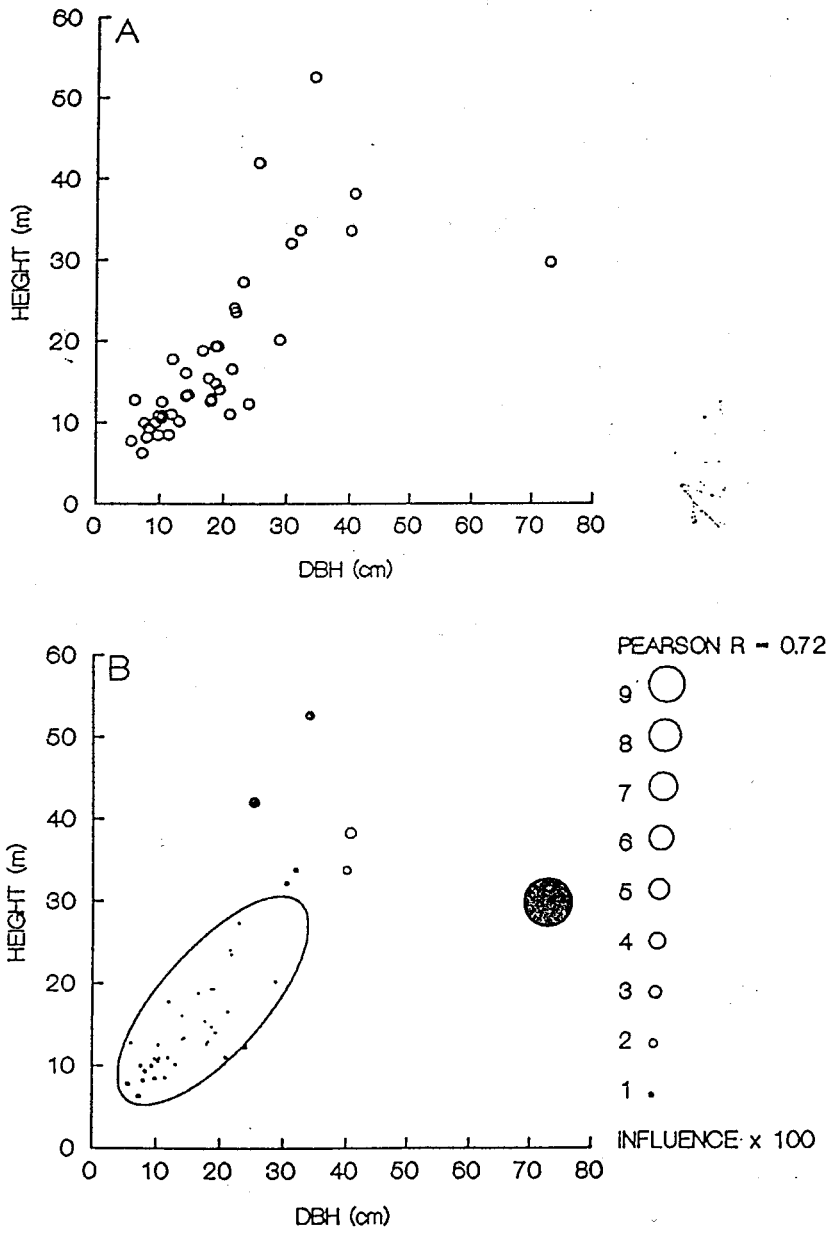
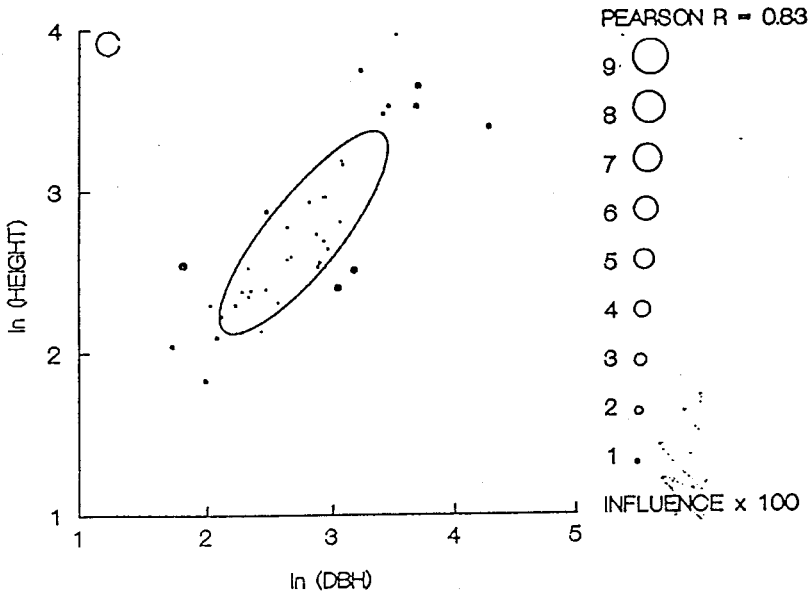


Figure 2.7. Scatterplots of tree diameter vs tree height for 41 trees in a mixed hardwood stand. (A) The raw data. (B) An influence plot, where the size of each point is directly proportional to the magnitude of its influence on *r*. Shading of the points indicates the direction of the influence (open circles have a positive influence on *r*, solid circles a negative influence). In this case, the putative outlier is shown as a large solid point



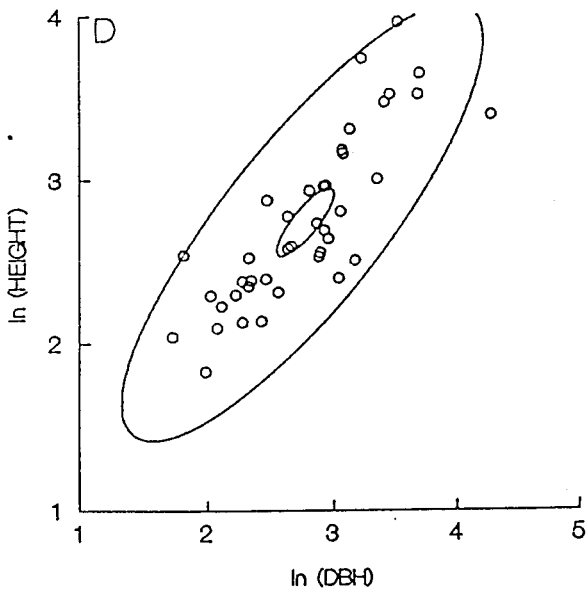


Figure 2.7.—Continued (influence × 100 = 11). Removal of this point alone, therefore, would increase the value of *r* from 0.72 to 0.83. A 50% bivariate confidence ellipse is overlain on the figure. (C) An influence plot of the data following log transformation. (D) Two different 95% confidence ellipses, the outer constructed based on the variables' standard deviations, and the inner constructed based on the standard errors of the means of the variables.

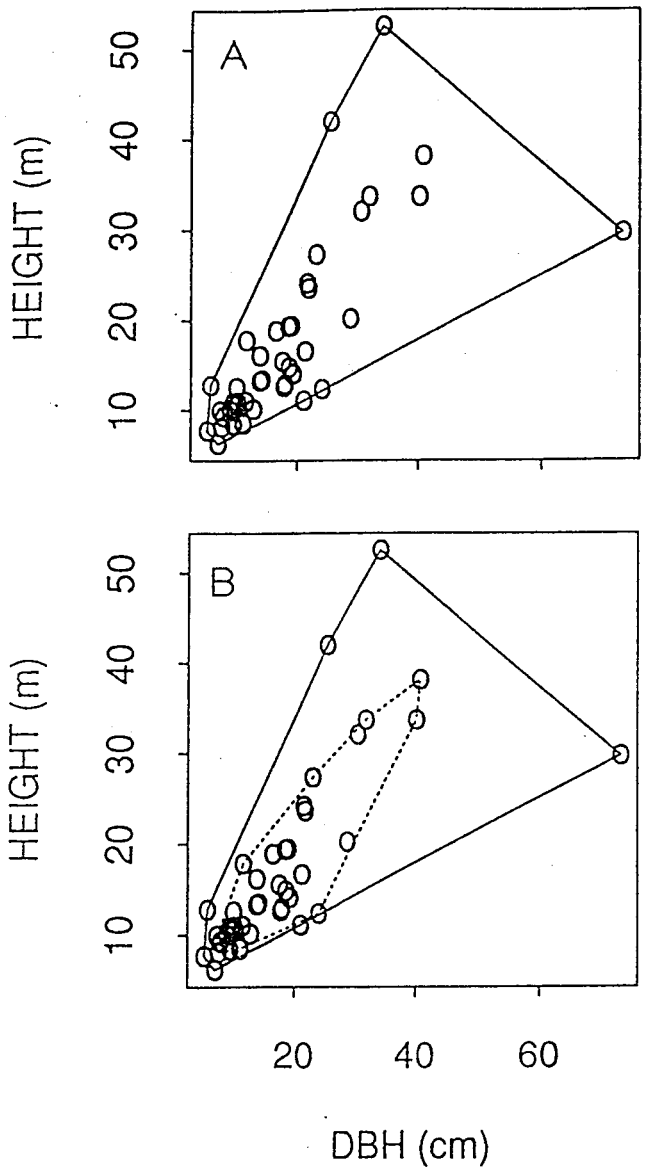


Figure 2.8. A convex hull (A and B, solid line) and a depth 2 peel (B, dotted line) around the tree size data. The hull is constructed by determining which points are furthest from the centroid of the data, and joining those points to form a polygon enveloping the other points. To peel the hull, all the points that lie on the initial convex hull are deleted, and a new convex hull is constructed for the remaining points.

of variation in the data set. The peeled hull, referred to as "peeled to depth 2" (Fig. 2.8B) includes all but the most extreme values of the dataset (compare the points outside the peeled hull of Fig. 2.8B to the points with strong influence on r in Fig. 2.7B). This process can be repeated ad infinitum, but normally does not proceed beyond depth 3. This is analogous to Tukey's (1977) running median (3R) smoother, extended in two dimensions. Like smoothers, convex hulls are constructed most easily with pencil and paper, or fast, interactive computer software (S-Plus). Convex hulls are useful for highlighting pattern within noisy data, and make no assumptions about the underlying distribution of the data.

Bivariate plots suitable for EDA are also suitable for final presentation. In preparing these plots for publication, however, there are several conventions often observed in the literature that should be dropped in favor of clarity of presentation. First, it is common in scatterplots to always start each axis at the origin (0,0). In fact, closely adhering to the actual range of the data when scaling axes is far more useful and informative than always including 0, especially if the extreme value of either variable is $<< 0$ or $>> 0$. Restricting the values on the axes to just beyond the extreme values of the data improves clarity and highlights pattern. Axis breaks do not always help, and changing the relative scaling after an axis break usually hinders accurate perception of the data, and can stymie future digitizers.

2.3.3 Extensions of Bivariate Techniques to Multivariate Data Sets

For data sets that include a number of continuous variables, it may not be clear which, if any, pair(s) of variables should be subjected to bivariate correlation or regression analysis, or whether you need to resort to multivariate techniques (Chapter 9). Three-dimensional plots (e.g., Fig. 2.11A) are often used to examine and illustrate higher dimensional data. While aesthetically pleasing, and easy to produce with current graphic software, accurate interpretation and digitizing of these graphs depend on the perspective and orientation of the plot.

The scatterplot matrix, whose origins are shrouded in mystery, provides an alternative exploratory and presentation tool for higher dimensional data. A symmetrical scatterplot matrix of the tree data is shown in Fig. 2.9. This is simply a plot of all possible bivariate combinations of the variables in the dataset. Plots above the diagonal have x- and y-axes transposed relative to those below the diagonal, which frees the investigator from preconceived notions of "dependent" and "independent" variables. One can, of course, apply the bivariate exploratory techniques described above to each of the scatterplots within the matrix. The possible addition of density plots of each variable along the diagonal gives the investigator a simultaneous feel for the distribution of individual variables (Ellison and Bedford, 1991). The final construction provides an information-rich, but rapidly comprehensible picture of the overall dataset.

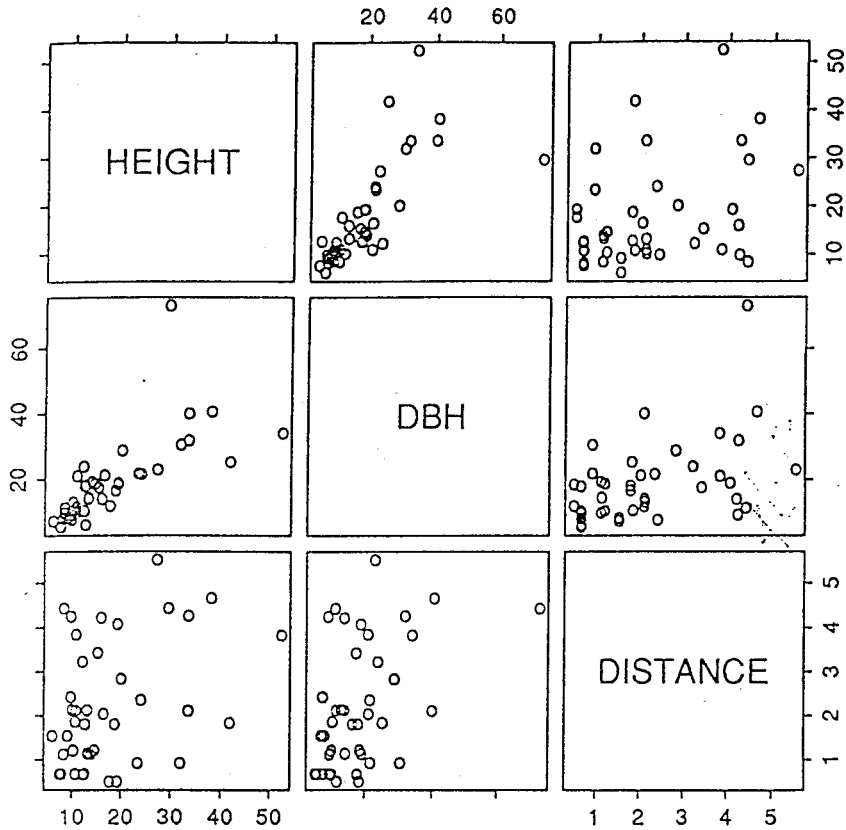


Figure 2.9. A scatterplot matrix of the tree size data. This plot illustrates bivariate relationships between all possible combinations of variables in a multivariate dataset. The variable name in the boxes along the diagonal corresponds to *x*-axis variable of plots below the diagonal, and *y*-axis variables above the diagonal.

2.3.4 Classified Quantitative Data: Alternatives to Bars and Pies

Classified quantitative data are common in many experimental situations. This type of data set consists of responses of a given parameter to discrete treatments. Such experiments may be analyzed by ANOVA (Chapters 3 and 4), and the results expressed in terms of the significance of treatment effects and/or interaction effects. Data from these types of experiments often are not explored prior to formal analysis, although the univariate techniques described in Section 2.3.1 are appropriate for examining the data structure of individual treatment groups. The exception to this generalization are common tests of the critical assumptions of ANOVA: homoscedasticity (variances among treatment groups are equal) and

normal distribution of residuals within treatment groups. In particular, failure to test for homoscedasticity is one of the most common statistical errors (Fowler, 1990b), and heteroscedastic (unequal variances) data can complicate or compromise results obtained from ANOVA (Sokal and Rohlf, 1981).

To illustrate EDA and graphical presentation of classified quantitative data, I use data from Potvin (Tables 3.2 and 3.3) that examine effects of genotype (the classifying variable) on fresh mass of *Plantago major*, and the interaction effects of bench position and genotype on stem dry mass of *Helianthus annuus* grown in a latin square design. In each of these data sets, there is only one response variable: plant mass. More complex data sets include responses of several variables to multiple levels of a given treatment. As an example of this latter type of data set, I use data from Ellison et al. (1993). We measured a number of growth and morphological characteristics of *Nepsera aquatica* (an herbaceous species of disturbed areas in tropical wet forests) in response to varying light levels (2, 20, and 40% of full sunlight).

Spread (some measure of variance) vs level (mean, median, etc.) plots (Norusis, 1990) are a rapid, graphic way to examine the within- and between-treatment group variances, and give clues as to appropriate data transformations to bring heteroscedastic data into line. Norusis (1990), modifying the technique of Box et al. (1978), suggests plotting the natural log of the interquartile distance (i.e., the hspread; Fig. 2.4A) vs the natural log of the median for *each* treatment group. An appropriate transformation of the data to remove dependency of the spread on the level is then given as 1 minus the slope of the linear regression line fit to the spread vs level plot. Figure 2.10A illustrates a spread vs level plot for Potvin's *Plantago* data. Note that the raw data are not homoscedastic; the variance increases with the mean. Following Norusis (1990) and Box et al. (1978), the slope of the regression line for this plot is 1.71, suggesting that the data be transformed by raising each observation to the -0.71 power. After such a transformation, the spread vs level plot (Fig. 2.10B) illustrates that the strict dependency of spread on level no longer exists, and the data are somewhat more suitable for ANOVA (the variances are no longer correlated with the mean, although they are still not equalized). Plant size data are often subject to logarithmic transformations to equalize variances within treatment groups. A log transformation of these data does about as well as the negative exponential transform in equalizing these variances (Table 2.1). Box and Cox (1964) and Zar (1984) provide detailed methods on determining the "best" transformation to be used on heteroscedastic data. Such transformations may not make biological sense, but keep in mind that the role of transformations is to bring your data in line with the assumptions and requirements of the statistical model(s) you are testing.

Graphic EDA can also be used to examine interaction effects in data. An example is illustrated in Fig. 2.11 for Potvin's *Helianthus* data. In this experiment, Potvin illustrates how position on a greenhouse bench interacts with genotype to determine plant mass. The top figure illustrates the relative small size of genotype

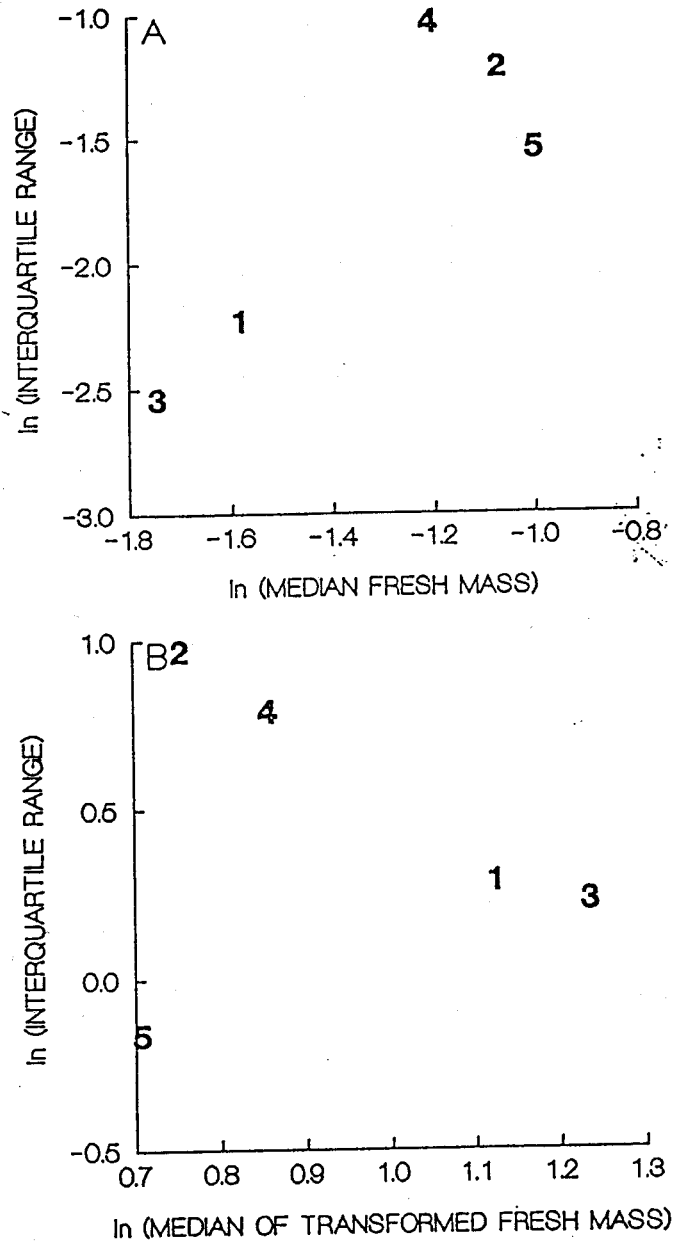


Figure 2.10. Spread vs. level plots of the *Plantago* data. Values plotted on (A) are ln(interquartile distance) on the *y*-axis vs ln(median plant mass) on the *x*-axis of seven replicate individuals of each of five genotypes. Genotype number is indicated on the plot. (B) Spread vs level plot of data following a negative exponential transformation (Norusis, 1990). See text and Table 1.1 for further explanation.

Table 2.1. Variance (s^2) of $n = 7$ Observations Per Genotype of *Plantago* Fresh Mass[a]

Genotype	Mean	Variance		
		Untransformed	Log transformation	Negative exponential transformation ($y^{-0.71}$)
1	0.198	0.006	0.179	1.245
2	0.309	0.034	0.440	1.798
3	0.109	0.008	0.151	0.710
4	0.298	0.029	0.354	1.302
5	0.412	0.039	0.196	0.392

[a]Variances are shown prior to transformation, following transformation by natural logarithms, and following transformation by the negative exponential suggested by the spread vs level plot (Fig. 2.10).

A and the relatively large size of genotype E. Although a scatterplot matrix might have made this pattern clearer, there is no real point to plotting row × column, or row × genotype, or column × genotype when the point is to illustrate the row × column *interaction* effect on genotype. The lower figure, a contour plot of the top one, illustrates the clear "hot-spot" in the upper left corner of the bench. As interaction effects often involve visualizing data in more than two dimensions, you can use many of the techniques normally applied to multivariate data in the exploration of interactions.

Classified quantitative data are presented poorly in the ecological literature. These problems are illustrated with the data of Ellison et al. on resource allocation and morphological responses to light by *Nepsera* (Fig. 2.12). The most common ways of presenting classified quantitative data are bar charts, separated or stacked (Figs. 2.12A,B), and pie charts (Fig. 2.12C). Separated bar charts (Fig. 2.12A), where a single bar represents the results of a single treatment, suffer from the same problems as histograms. The bars themselves use a lot of ink—horizontal lines, vertical lines, shading of bars of arbitrary width—to convey information about only a single point at the top of the bar (compare Fig. 2.12A with 2.12D). Stacked bar charts (Fig. 2.12B), where treatment groups are divided into subsets and the groups are compared against one another, are virtually unintelligible and never should be used. In this example, the percent allocation to leaves, roots, and stems sum to roughly 100% (allowing for error and missing values). Figure 2.12A (bars side-by-side) at least clearly illustrates the relative allocation to each part. It is not so simple, on the other hand, to determine the relative allocation in Figure 2.12B. Because we use 0 as our reference point, the first guess would be that the allocation to roots in 2% light is ≈70%, and that to stems 100%, when clearly this cannot be true. However, it is difficult to determine visually the beginning point of any of the stacked segments beyond the lowest one. Although measures of variance can be placed clearly on side-by-side bar charts, error bars cannot be placed on stacked bar charts (see Section 2.4). Shadings, hatching,

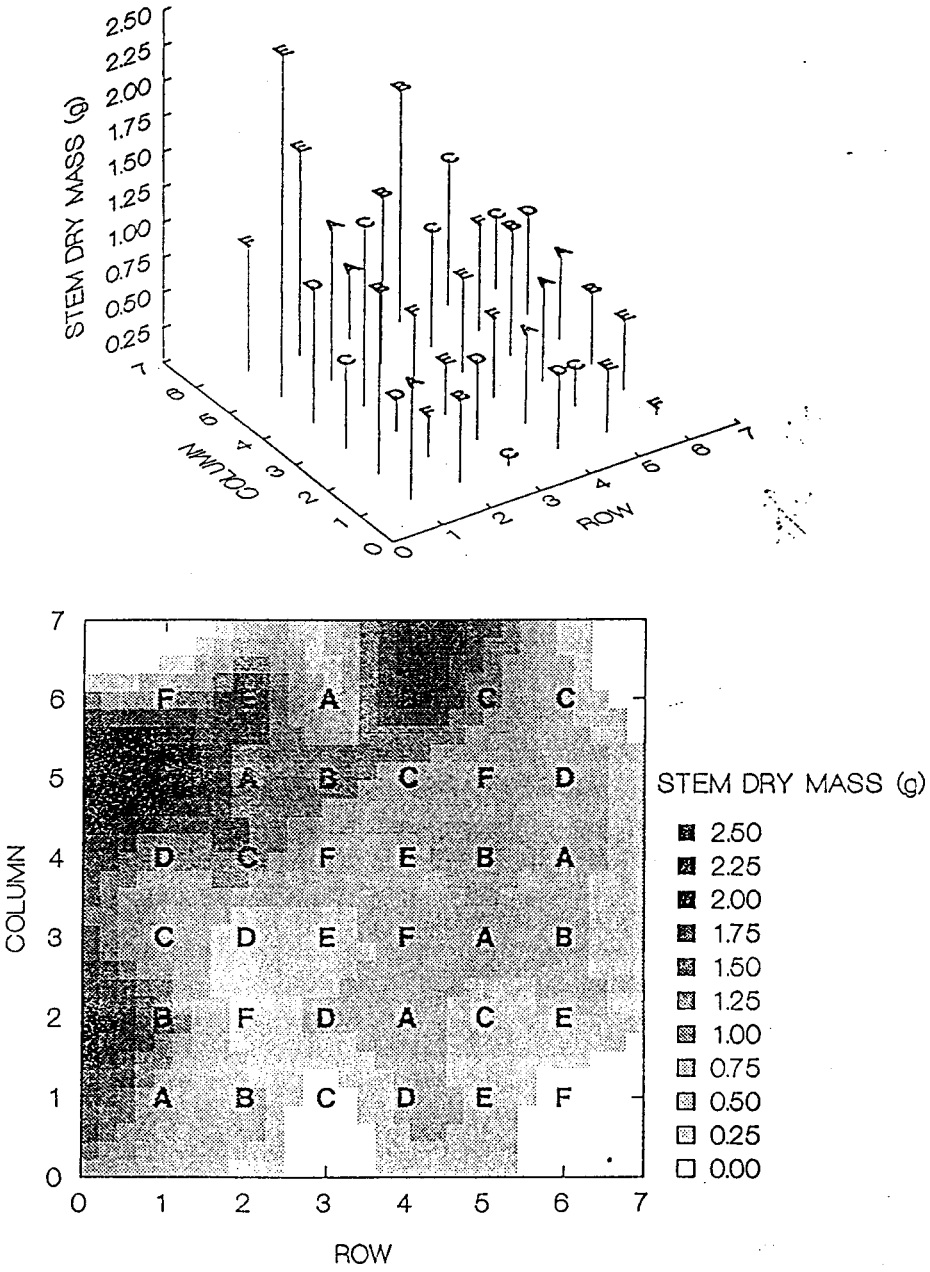


Figure 2.11. Two ways of visualizing the effect of bench position and genotype on stem dry weight of *Helianthus*. The top figure is a three-dimensional scatterplot, with genotype letter (A-F) as the plotting symbol. The addition of sticks connecting each point to its

and other chartjunk used in bar charts also can interfere with accurate perception of the data and decrease the data:ink ratio. Pies share all of the problems of stacked bar charts, and none of the advantages of side-by-side bar charts. I can think of no cases in which a pie chart should be used.

There are several alternatives to bar charts and pie charts. Plots in which the mean value of the response variable is plotted as a single point, along with some measure of error, clearly illustrate the same data as in a bar chart with greater clarity and less data ink (Fig. 2.12D). Sets of box plots better illustrate the underlying data structure and convey more information with less ink and confusion (Fig. 2.12E). These box plots have been "notched" (McGill, et al. 1978) to show 95% confidence intervals. Polar category plots (with or without error bars; the latter are shown in Fig. 2.12F) are the minimalist alternative to bar charts, and are a visually comparable substitute for pie charts. These polar category plots illustrate the response of eight measured variables to the three light environments and clearly convey overall differences between treatment groups.

2.4 A Word about Error Bars

Any reported parameter must include a measure of the reliability of that parameter as well as the sample size. For example, sample means, whether reported graphically or in tables, must be accompanied by the sample size and some estimator of the variance. Error bars on graphs must be correctly identified. Three kinds of error bars are seen commonly in the ecological literature: standard deviations, standard errors, and *n*% confidence intervals. Note that strictly speaking, the first is the *sample standard deviation*. The second, more properly referred to as the *standard error of the mean*, is an estimate of the accuracy of the estimate of the mean. We compute it as the standard deviation of a distribution of means of samples of identical sizes from the underlying population (see Zar, 1984:31 for a complete description). Thus, calling error bars simply *standard deviation* bars confounds the two. Measures of error are used to calculate *n*% confidence intervals. We can compute easily confidence intervals of normally distributed data from the standard error of the mean (Sokal and Rohlf, 1981). For other distributions, approximations of confidence intervals can be computed using bootstraps, jackknifes, or other resampling techniques (Efron, 1982; Dixon, Chapter 13). *All* of these measures require information about sample size, which must be reported to ensure accurate interpretation of results.

Figure 2.11.—Continued position on the *x*-*y* plane permits more accurate perception of the true height along the *z*-axis of each point. The lower figure is a contour plot, with intensity of shading indicating the biomass at a particular row × column location on the bench. These contours were determined by a negative exponential smoothing routine, where the influence of neighboring values decreases exponentially with distance. Shading density increases with biomass.

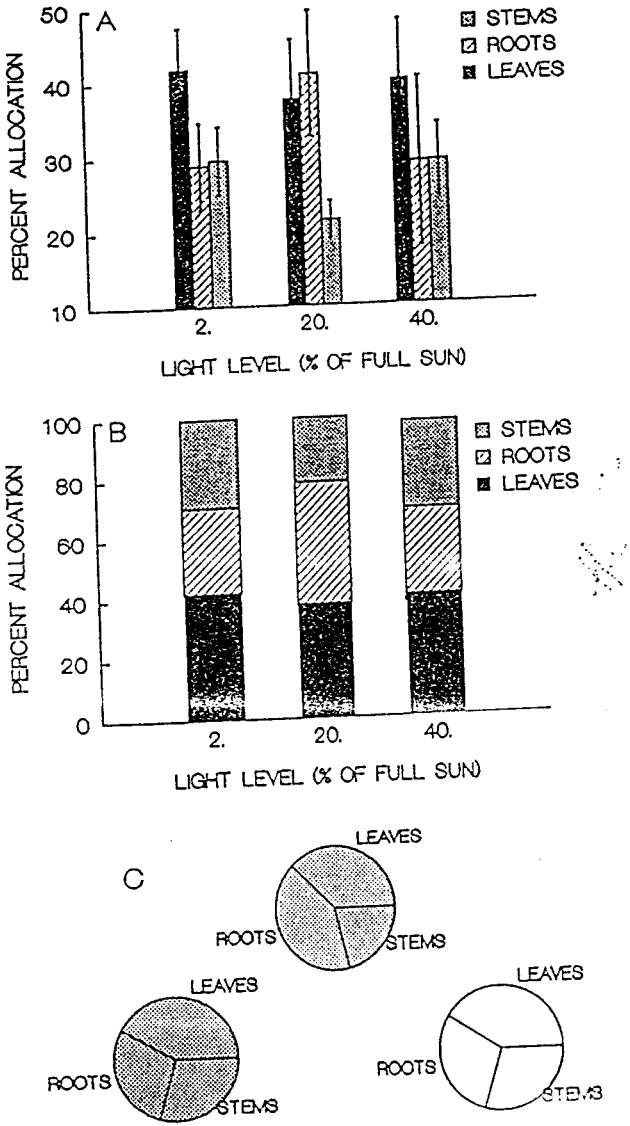


Figure 2.12. Six alternative presentations for presenting classified quantitative data. Data presented are from an experiment examining the effect of three different light levels (2, 20, and 40% of full sun) on growth, resource allocation, and morphology of *Nepsera aquatica*. Each treatment consisted of 20 individually potted plants, harvested after 6 months of growth (Ellison et al., 1993). (A) A side-by-side bar chart illustrating percent allocation to leaves, roots, and stems by plants in each light treatment. Height of the bar indicates mean percent allocation, and error bars indicate 1 standard deviation of the mean. (B) A stacked bar chart illustrating the same data. (C) Pie charts illustrating the relative resource allocation in the three light environments (dark shading: 2% light; intermediate shading: 20% light; no shading: 40% light). Note that it is not possible to place error bars on stacked bar charts or pie charts. (D) Simple category plot of the data illustrated in

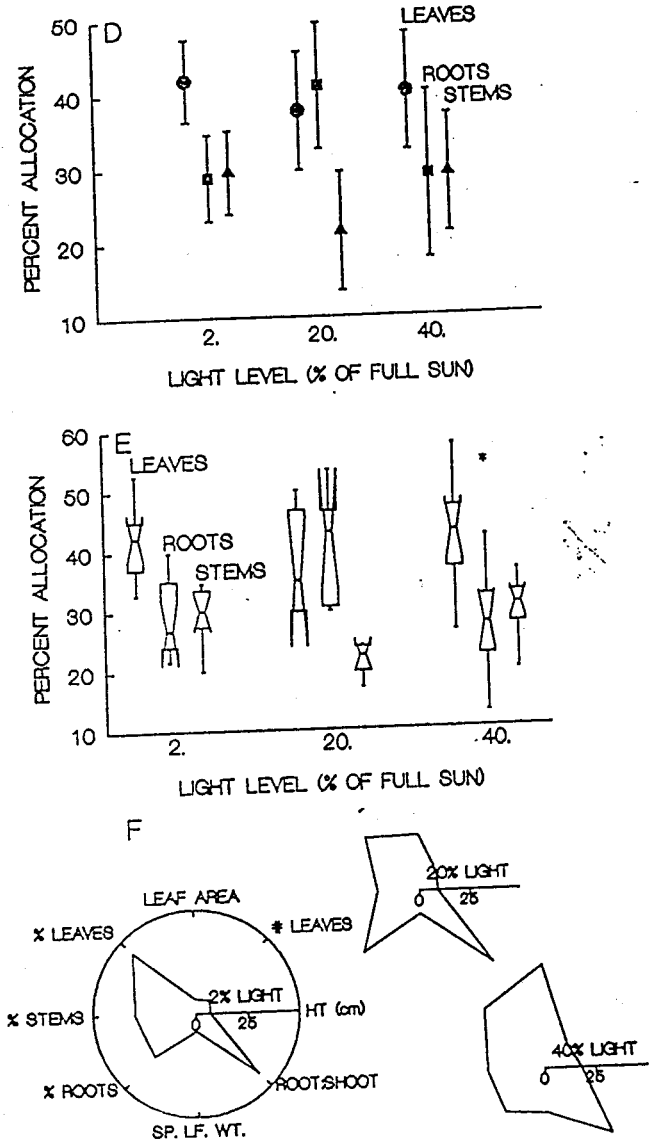


Figure 2.12.—Continued Fig. 2.12A. Each point represents the mean percent allocation to leaves (*circles*), roots (*squares*), and stems (*triangles*); error bars are 1 standard deviation. (E) Notched box plots of the data. Box plot construction as in Fig. 2.4A. Plots are "notched" to illustrate 95% confidence intervals. Where the box reaches full width on either side of the median indicates the limits of the confidence interval. (F) Polar projections of category plots (also known as star plots) of the response of eight measured parameters to the three light treatments. The radius of the circle is equivalent to the *y*-axis of a rectangular plot; the distance from the center of the circle to each vertex of the polygon is the mean response of each variable to the treatment. Variables are arranged equidistantly around the perimeter of the circle (equivalent to the *x*-axis of a rectangular plot). One

In general, error bars are useful only when they convey information about confidence intervals. Typically, in the ecological literature, means are plotted along with error bars illustrating one standard error of the mean. For suitably large n, or for samples from a normal distribution, one standard error bar approximates a 68% confidence interval. This conveys little information of interest, since we are accustomed to thinking either in terms of 50, 90, 95, or 99% confidence intervals. Further, most ecological samples are small, or the underlying data distributions are unknown. In those cases, error bars representing one standard error of the mean convey no useful information at all. In keeping with the guidelines for graphical display presented at the beginning of the chapter, I suggest that sample standard deviations or 95% confidence intervals be your error bars of choice. Two-tiered error bars (Cleveland, 1985) that display both quantities are an excellent compromise. Meta-analysis (Gurevitch, Chapter 17) requires sample standard deviations, and if they are reported together with sample size, permit rapid calculation of confidence intervals, standard errors, or most other measures of variation. In the end, the choice of error bar reported lies with you. It is most important that they be identified accurately. Note that if you transformed your data prior to analysis, your calculated standard deviation will be symmetrical relative only with respect to the transformed mean. If you present the results back-transformed (as is common practice), the error bars should be asymmetric.

2.5 Concluding Remarks

Ecologists traditionally have used a limited palette of graphic elements and techniques for exploring and presenting our data. We need to refocus our vision to grasp new or unfamiliar graphic elements and techniques that will permit clear communication of our data. We can now use available computer hardware and software with expanded EDA and presentation capabilities to display our results accurately, concisely, and in aesthetically pleasing ways (Ellison, 1992). We will improve our comprehension and appreciation of data by using many of the graphic techniques presented in this chapter, just as we can increase our appreciation of the diversity of pasta entrées with a trip to a fine Italian restaurant.

Acknowledgments

I thank the late Deborah Rabinowitz for introducing me to EDA and data-rich graphic techniques. Phil Dixon, Steve Juliano, and Catherine Potvin generously

Figure 2.12.—Continued obtains a picture of the overall response of the plant to each light treatment by constructing a polygon whose vertices are equal to the value of the response variable. Different shapes in the different light treatments indicate overall treatment effects. For this type of plot to be effective, all data must be similarly scaled; for this plot, root:shoot ratio (g g^{-1}) was multiplied by 10^2, and specific leaf weight (g cm^{-2}) was multiplied by 10^4. Leaf area is in cm^2, and is a measure of total leaf area per plant.

shared data from their respective chapters. The data on tree size was collected by the 1992 population ecology class at Mount Holyoke College. The work on *Nepsera* was supported by NSF Grant BSR-8605106 to Julie Denslow. Technical support personnel at Systat, Inc. and Statistical Sciences, Inc. helped immensely with final graphics production. Phil Dixon, Elizabeth Farnsworth, Jessica Gurevitch, Catherine Potvin, Sam Scheiner, and one anonymous reviewer provided constructive reviews of early drafts of this chapter that resulted in a much-improved final version. Hardware for graphics production was provided by the BioCIS grant from IBM Corporation. Additional support was provided by NSF Grant BSR-9107195, and the Internet.

Appendix 2.1

SYGRAPH program code to create Figs. 2.2, 2.5, and 2.12.

```
USE '<\PATH\FILENAME.FILETYPE>'                    /specifies data file/
SYGRAPH                                            /executes Sygraph/
OUTPUT=PRINTER                                     /directs output to printer/
MODE PRINTER=POS1/LPT1                             /defines printer port/

TYPE=STROKE                                        /chooses character set/
CS=1.1                                             /specifies character size/
THICK=1.2                                          /specifies line thickness/

                                                   /Fig. 2.5/
BEGIN                                              /To place multiple plots on a single page, the
                                                   commands must be bracketed by BEGIN and END
                                                   statements. Everything between these two
                                                   keywords will appear on a single page./

ORIGIN=1.125IN,4.75IN                              /location of 1st plot on the page/
PPLOT NODES/NORM,SHORT,                            /normal probability plot/
  SMOOTH=LINEAR,STICK,
  HEIGHT=3IN,WIDTH=3IN,
  XLABEL='NUMBER OF NODES',
  SYMBOL=2,FILL=0,SIZE=.75
WRITE 'A'/HEIGHT=10PT,                             /plot label/
  WIDTH=10PT,X=0.1IN,Y=2.75IN

ORIGIN=1.125IN,0.75IN                              /location of 2nd plot on the page/
PPLOT NODES/WEIBULL,SHORT,                         /Weibull probability plot/
  SMOOTH=LINEAR,STICK,
  HEIGHT=3IN,WIDTH=3IN,
  XLABEL='ln(NUMBER OF NODES)',
  YLABEL='ln(WEIBULL QUANTILES)',
  SYMBOL=2,FILL=0,SIZE=.75
WRITE 'B'/HEIGHT=10PT,                             /plot label/
  WIDTH=10PT,X=0.1IN,Y=2.75IN
END                                                /end of page/
```

```
                                                              /Fig. 2.2/
USE '<\PATH\FILENAME.FILETYPE>'                          /choose new data file/
BEGIN                                                    /start page production/
ORIGIN=0.125IN,7.85IN                          /location of box plot on the page/
BOX NODES/XLABEL=' ',                                /construct box plot; note that
  AXES=0,SCALE=0,WIDTH=4IN,                            plot widths, and data ranges
  HEIGHT=2IN,MIN=6,MAX=34                          match for all plots on this page/

ORIGIN=0.125IN,6.25IN                                    /location of Fig. 2.2A/
DENSITY NODES/HIST,                                      /construct histogram/
  HEIGHT=3IN,WIDTH=4IN,
  XLABEL=' ',YLABEL=' '
WRITE 'A'/X=0.1IN,Y=1.35IN,
  HEIGHT=10PT,WIDTH=10PT

ORIGIN=0.125IN,3.25IN                                    /location of Fig. 2.2B/
DENSITY NODES/HIST,BWIDTH=1,                     /construct histogram; note the
  HEIGHT=3IN,WIDTH=4IN,                              bin width is specified with
  XLABEL=' ',YLABEL=' '                                the Bwidth option/
WRITE 'B'/X=0.1IN,Y=2.1IN,                               /plot label/
  HEIGHT=10PT,WIDTH=10PT

ORIGIN=0.125IN,0.25IN                                    /location of Fig. 2.2C/
DENSITY NODES/HIST,BWIDTH=4,                 /construct histogram; specify bin width/
  HEIGHT=3IN,WIDTH=4IN,
  XLABEL=' ',YLABEL=' '
WRITE 'C'/X=0.1IN,Y=1.35IN,
  HEIGHT=10PT,WIDTH=10PT

WRITE 'NUMBER OF NODES'/                          /x-axis label; specifies
  Y=-0.75IN,X=.8IN,                               location of label origin,
  HEIGHT=13PT,WIDTH=13PT                            and character size/
WRITE 'PROPORTION PER BAR'/                              /left y-axis label/
  Y=2.75IN,X=-.85IN,
  HEIGHT=13PT,WIDTH=13PT,
  ANGLE=90                                    /specifies 90° label rotation/
WRITE 'COUNT'/X=4.625IN,                                 /right y-axis label/
  Y=4.635IN,HEIGHT=13PT,
  WIDTH=13PT,ANGLE=270
END
```

```
                                                          /Fig. 2.12/
USE '<\PATH\FILE1.FILETYPE>'                          /select first data file/
BEGIN                                              /start 1st page production/

ORIGIN=0.75IN,6IN                                  /location of Fig. 2.12A/
BAR LEAVES,ROOTS,STEMS*TREAT/                       /construct bar chart/
  FILL=1,4,7,ERROR=SDLVS,                           /choose fill pattern/
  SDROOTS,SDSTEMS,AXES=2,                           /specify error bars/
  YMIN=10,YMAX=50,                                  /choose y-axis range/
  XLABEL='LIGHT LEVEL                                /axis labels/
          (% OF FULL SUN)',
  YLABEL='PERCENT ALLOCATION',
  WIDTH=2.75IN,HEIGHT=1.85IN,          /size of plot - consistent for all plots on page/
  LEGEND=2.35IN,1.35IN                 /location of legend with respect to plot/
WRITE 'A'/X=0.1IN,Y=1.75IN,
  HEIGHT=8PT,WIDTH=8PT

ORIGIN=0.75IN,3.25IN                               /location of Fig. 2.12B/
BAR LEAVES,ROOTS,STEMS*TREAT/                       /construct bar chart/
  FILL=1,4,7,STACK,                    /use STACK option for stacked bars/
  AXES=2,YMIN=0,YMAX=100,
  XLABEL='LIGHT LEVEL
          (% OF FULL SUN)',
  YLABEL='PERCENT ALLOCATION',
  WIDTH=2.75IN,HEIGHT=1.85IN,
  LEGEND=2.35IN,1.35IN
WRITE 'B'/X=0.1IN,Y=1.75IN,
  HEIGHT=8PT,WIDTH=8PT

CS=3.2                                 /new character size to make pie chart
                                       labels conform to other plots on page/
ORIGIN=0.75IN,0.5IN                                /location of 1st pie chart/
PIE AVG2*PARTS/SORT,                  /construct 1st pie chart - 2% light/
  WIDTH=0.75IN,HEIGHT=0.75IN,
  FILL= 0.35                          /specifies density of pie's shading/
WRITE 'C'/X=0.1IN,Y=1.75IN,
  HEIGHT=8PT,WIDTH=8PT

ORIGIN=1.75IN,1.375IN                              /location of 2nd pie chart/
PIE AVG20*PARTS/SORT,                 /construct 2nd pie chart - 20% light/
  WIDTH=0.75IN,HEIGHT=0.75IN,
  FILL= 0.1

ORIGIN=2.75IN,0.5IN                                /location of 3rd pie chart/
PIE AVG40*PARTS/SORT,                 /construct 3rd pie chart - 40% light/
  WIDTH=0.75IN,HEIGHT=0.75IN,
  FILL= 0                                          /no shading/

END                                                /end of 1st page/
```

```
USE '<\PATH\FILE2.FILETYPE>'                          /choose new file/
BEGIN                                                 /begin 2nd page production/
ORIGIN=0.75IN,6IN                          /Fig. 2.12D is actually a composite of 3 overlaid
                                             plots. Plots for roots, stems, and leaves are
                                             constructed separately, placed at the same
                                           y-position (height) on the page, but the leaves plot is
                                           moved left, and the stem plot is moved right relative
                                           to the roots plot. Note the change in origin
                                             location to accomplish this overlay./

CPLOT ROOTS*TREAT/SYMBOL=7,                           /plot roots vs light level/
  FILL=1,AXES=2,YMIN=10,YMAX=50,
  ERROR=SDROOTS,WIDTH=2.75IN,
  HEIGHT=1.85IN,SIZE=1.75,
  XLABEL='LIGHT LEVEL
          (% OF FULL SUN)'
  YLABEL='PERCENT ALLOCATION',

ORIGIN=0.625IN,6IN                                    /move left 1/8"/
CPLOT LEAVES*TREAT/SYMBOL=2,                          /plot leaves vs light level/
  FILL=1,AXES=0,YMIN=10,
  YMAX=50,ERROR=SDLVS,
  XLABEL=' ',YLABEL=' ',                  /no axes or axis labels to preserve overlay/
  WIDTH=2.75IN,HEIGHT=1.85IN,
  SCALE=0,SIZE=1.75

ORIGIN=0.875IN,6IN                        /move right 1/8" relative to roots plot/
CPLOT STEMS*TREAT/SYMBOL=3,
  FILL=1,AXES=0,YMIN=10,
  YMAX=50,ERROR=SDLVS,
  XLABEL=' ',YLABEL=' ',
  WIDTH=2.75IN,HEIGHT=1.85IN,
  SCALE=0,SIZE=1.75
WRITE 'D'/X=0.1IN,Y=1.75IN,
  HEIGHT=8PT,WIDTH=8PT

WRITE 'LEAVES'/X=1.9375IN,                /legend needs to be added on explicitly/
  Y=1.8IN,HEIGHT=6PT,WIDTH=6PT
WRITE 'ROOTS'/X=2.0625IN,
  Y=1.435IN,HEIGHT=6PT,WIDTH=6PT
WRITE 'STEMS'/X=2.1876IN,
  Y=1.3IN,HEIGHT=6PT,WIDTH=6PT
```

```
USE '<\PATH\FILE3.FILETYPE>'                                  /choose new file/
ORIGIN=0.75IN,3.25IN                                          /location of Fig. 2.12E/
BOX PERCENT*TREAT/AXES=2,          /To construct this figure, dummy x-values were
  XLABEL='LIGHT LEVEL          placed in the data file to accomplish the offsetting
          (% OF FULL SUN)',      done with multiple plots in Fig. 2.12D. This was
  YLABEL='PERCENT ALLOCATION',     done so that the 95% confidence intervals
  WIDTH=2.75IN,HEIGHT=1.85IN,      specified with the Notch option would be
  SCALE=-2,NOTCH,MIN=10                constructed correctly./
WRITE 'E'/X=0.1IN,Y=1.75IN,
  HEIGHT=8PT,WIDTH=8PT

WRITE 'LEAVES'/X=0.1875IN,          /Again, the legend was placed on manually, as
  Y=1.625IN,HEIGHT=6PT,WIDTH=6PT      were the x-axis labels. Tick marks were
WRITE 'ROOTS'/X=0.375IN,                                  later erased by hand/
  Y=1.135IN,HEIGHT=6PT,WIDTH=6PT
WRITE 'STEMS'/X=0.5625IN,
  Y=0.9475IN,HEIGHT=6PT,WIDTH=6PT
WRITE '2.'/X=0.375IN,Y=-0.125IN,
  HEIGHT=6PT,WIDTH=6PT
WRITE '20.'/X=1.3125IN,Y=-0.125IN,
  HEIGHT=6PT,WIDTH=6PT
WRITE '40.'/X=2.3125IN,Y=-0.125IN,
  HEIGHT=6PT,WDITH=6PT

USE '<\PATH\FILE4.FILETYPE>' /Choose new file. Data files for polar category plots
                          needs to be arranged so that each star's 'vertex' is
                          an individual case. To construct this plot, four
                          variables are needed in the file: treat$ - the label for
                          each vertex; avg2 - the response of the given
                          parameter (=label) in 2% light; avg20 - the response
                          in 20% light; avg40 - the response in 40% light. As
                          with Fig. 2.12D, Fig. 2.12F is a composite of 3 plots./

CS=1.8                                  /character size set to conform to previous plots/
ORIGIN=0.75IN,0.5IN
WRITE 'F'/X=0.1IN,Y=1.75IN,
  HEIGHT=8PT,WIDTH=8PT

ORIGIN=0.6IN,.55IN
CPLOT AVG2*TREATS/AXES=2,                          /plot of response in 2% light/
  LINE=1,POLAR,NSORT,SCALE=2,          /the polar option gives the projection/
  TICK=2,YMIN=0,YMAX=50,               /the line option connects the vertices/
  SIZE=0,WIDTH=1.25IN,
  XLABEL=' ',YLABEL='2% LIGHT',
  HEIGHT=1.25IN

ORIGIN=2IN,1.25IN                                  /plot of response in 20% light/
CPLOT AVG20*TREATS/AXES=-2,
  LINE=1,POLAR,NSORT,SCALE=-2,
  TICK=2,YMIN=0,YMAX=50,
  SIZE=0,WIDTH=1.25IN,
  HEIGHT=1.25IN,XLABEL=' ',
  YLABEL='20% LIGHT'

ORIGIN=2.75IN,.15IN                                /plot of response in 40% light/
CPLOT AVG40*TREATS/AXES=-2,
  LINE=1,POLAR,NSORT,SCALE=-2,
  TICK=2,YMIN=0,YMAX=50,
  SIZE=0,WIDTH=1.25IN,
  HEIGHT=1.25IN,XLABEL=' ',
  YLABEL='40% LIGHT'
END                                                          /end 2nd page/
```