
Data presentation

Sami L. Bahna, MD, DrPH,* and Jerry W. McLarty, PhD†

INTRODUCTION

Organization and summarization of data are prerequisites for appropriate analysis, presentation, and interpretation. Effective data presentation can be used not only to summarize and communicate findings but also to aid in quality control, data cleaning, and error identification and to help determine the analytical strategy. The presentation process depends on the type of data, the complexity of the information, and the objective of the display.

TYPES OF DATA

Data collected during an experiment or trial can take on a constant value or have varying values. In the case of *constant data*, the observed data have a fixed value (eg, number of eyes or the number of fingers on one hand). Constant data usually form the basis for experimental design rather than observed data. *Variable data* take on varying values that can be classified as qualitative or quantitative.

Categorical Data

Data that can be assigned to one of a set of discrete, usually predefined, categories are termed *categorical data*. In analysis of categorical data, the categories may be used as factors in experimental design. For example, a study may involve the numbers of males and females who have been enrolled in a clinical trial. In this case, *male* and *female* are predefined categories, the numbers of each are the observational data, and the analysis may involve comparison of the number of males vs females. Categorical data may also be treated as observational data and can be used in analyses such as logistical regression. Consider as an example the measurement of B-type natriuretic peptide along with the observation of the heart failure classification of each subject. Logistical regression can be used to define whether there is a linear relationship between B-type natriuretic peptide level and clinical expression of heart failure.

It is not uncommon to create categories from noncategorical data. In place of using age (a quantitative variable) in a pediatric study, one may wish to assign subjects to the categories of neonate, infant, child, and adolescent. Although this may be helpful for clinical purposes, it changes the way this variable can be used in data analysis.

Categorical data are of 3 fundamental types, nominal, ordinal, and interval, based on the relationship among the assigned categories.

Nominal data are categorical data that can be described only by a name, for example, allergy manifestation (rhinitis, asthma, urticaria, or eczema), race or ethnic origin (white, black, Hispanic, or Asian), and state of residence (Alabama, California, Florida, and so on). The categories are not related to each other with respect to value or order. Although their listing does not have to follow a particular order, they may for convenience be arranged alphabetically or according to some other designation, such as frequency of occurrence.

Ordinal (or ranked) data are categorical data that follow a certain order, for example, disease severity (mild, moderate, or severe) or faculty rank (instructor, assistant professor, associate professor, or professor). Ordering does not necessarily imply equal intervals among ranks. The categories may exist as named or numbered categories. For analytical convenience, named categories may be assigned numbers. Numbered categories or numbering schemes, such as 1 for mild, 2 for moderate, and 3 for severe, appropriately designate ranking, but a score of 2 is not necessarily twice the effect as a score of 1, and the difference in disease severity between scores 1 and 2 may not be the same as the difference between scores 2 and 3.

Interval data are categorical data that have a predefined ordering and a constant interval or effect relating the categories. The months of the year represent an interval scale because each is represented by (approximately) 30 days. Assigning quantitative data to 1 of 4 quartiles representing the range of the data, or age to equal bins of 5 years, would also represent interval scales.

Quantitative Data

Quantitative data are data that are numerical and exist over a range of values. The number representing a quantitative observation can be continuous or discrete.

Continuous data are numeric measurements on variables that can take on any value and in which the measurements include a high degree of precision. Data of this type commonly include a fractional component, such as the measurement of cardiac output as 4.36 L/min. Data without fractional components can also be considered continuous if the number of potential values is relatively high, such as B-type natriuretic peptide, which can range from less than 100 ng/L to more than 10,000 ng/L.

Discrete data are those that can take on only certain values rather than any value. These are usually represented as an integral value. Implied in this definition is that the range of values is relatively wide. Although interval (categorical) data

Affiliations: * Department of Pediatrics, Allergy and Immunology Section, Louisiana State University Health Sciences Center, Shreveport, Louisiana; †Department of Medicine, Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Louisiana.

Disclosures: Authors have nothing to disclose.

Received for publication January 7, 2009; Received in revised form February 19, 2009; Accepted for publication February 27, 2009.

as described in the previous section could also fit in this definition, interval data are generally limited to a small number of categories, typically less than 5 to 10 categories. In practice, discrete data most commonly represent an integral approximation of a continuous variable, such as age when represented in years or heart rate when represented as beats per minute. Age and heart rate could easily be measured to a much higher precision than their integral values, but the number of age or heart rate values that can be observed is sufficiently large to assume the characteristics of quantitative data. When this is the case, we can apply parametric statistics to these data.

METHODS OF DATA PRESENTATION

Presentation of data may take 3 forms, often combined: text, tables, and graphs. A good rule of thumb, as described by van Belle, is to “use sentence structure (ie, text) for displaying 2 to 5 numbers, tables for displaying more numerical information, and graphs for complex relationships.”

Text Presentation

Describing the data in a text form is the most common but not always the best method. It is usually the simplest but has a suboptimal readability rate and does not have the visual impact of the other 2 forms. Its comprehension can be enhanced by constructing short sentences and using simple, yet grammatically correct, language.

Tabular Presentation

Tabular presentations are common and popular. Tables can include informative details with a visual effect better than the text's. A table's readability and visual impact depend on its design. Tables can be *simple*, representing 1 variable, or *compound*, representing 2 or more variables. Complex tables with numerous variables have low readability rates.

The principal variable in the table (usually the first column) may be categorized as follows: an *array*, where qualitative variables may be listed in any preferred way, but quantitative variables should be categorized in an ascending or descending order, or a *frequency distribution*, in which distributing the observations in class intervals of a reasonable number to reveal a trend or the type of distribution.

Recommendations on the use of frequency distributions include the following:

- The class intervals should allow the inclusion of every possible observation by having a definite beginning and an end to avoid overlap. For example, a group of children 1 to <15 years of age may be classified as 1 to 4 years, 5 to 9 years, and 10 to 14 years; and not 1 to 5 years, 5 to 10 years, and 10 to 15 years. Accordingly, a child who just turned 5 years will be in the second category not the first.
- Avoid open-end intervals (eg, <5 years or >10 years). Such intervals are more likely to be associated with errors and may result in erroneous computations or graphic presentation.

- Choose a number of categories that can reveal trends and facilitate appropriate analysis. A category that has a low frequency may be combined with another.
- Any categories with low frequencies that cannot be combined with the others may be combined together in one category as “other.”
- Minimize the construction of intervals of unequal width. A high frequency may give a wrong impression of importance to a category merely because it is wider than the others. In such situations, adjusted frequencies (according to the width of the category) should be considered. In certain situations, however, the use of unequal intervals of classification would be appropriate, for example, classification according to developmental age: infancy (0 to <1 year), childhood (1 to <12 years), adolescence (12 to <18 years), and adulthood (≥ 18 years).
- The frequencies may be expressed as absolute numbers, percentages (relative frequencies), or preferably both. The inclusion of actual numbers is particularly necessary when the frequency is low.
- In presenting computed values, rounding the number and choosing the place of the decimal point follow common sense (ie, according to the desired degree of precision or the relative value of the fraction compared with the whole number).
- Cumulative frequencies can be categorized in an ascending or descending manner. The former starts with a “less than” category that includes the lowest frequency, whereas the descending cumulative frequency starts with a “greater than” category that includes the total number.

Table Composition

The informative value of the table depends on its construction. The following principles should be helpful:

- The title should be concise yet sufficiently descriptive to stand by itself without requiring the reader to search the text for clarifications. Many prefer reading tables rather than text.
- The rows and columns should have clear headings and subheadings (if applicable) and must include the relevant unit of measure (eg, number, percentage, lb, kg, U/mL, mg/dL; the latter should not be expressed as mg%).
- A footnote below the table should be for clarifications or spelling of abbreviations that have been indicated by specific symbols in the table.
- If the table includes data from previous publications, the source should be included as a footnote.
- Tables being prepared for publication need to follow instructions provided by the publisher.

Graphic Presentation

Graphs, also referred to as figures, charts, or diagrams, have the quickest, strongest, and most long-lasting visual impact but often lack precision. Their impact can be enhanced by inserting selected data (eg, the number of subjects in each group). Graphs are excellent for demonstrating trends or

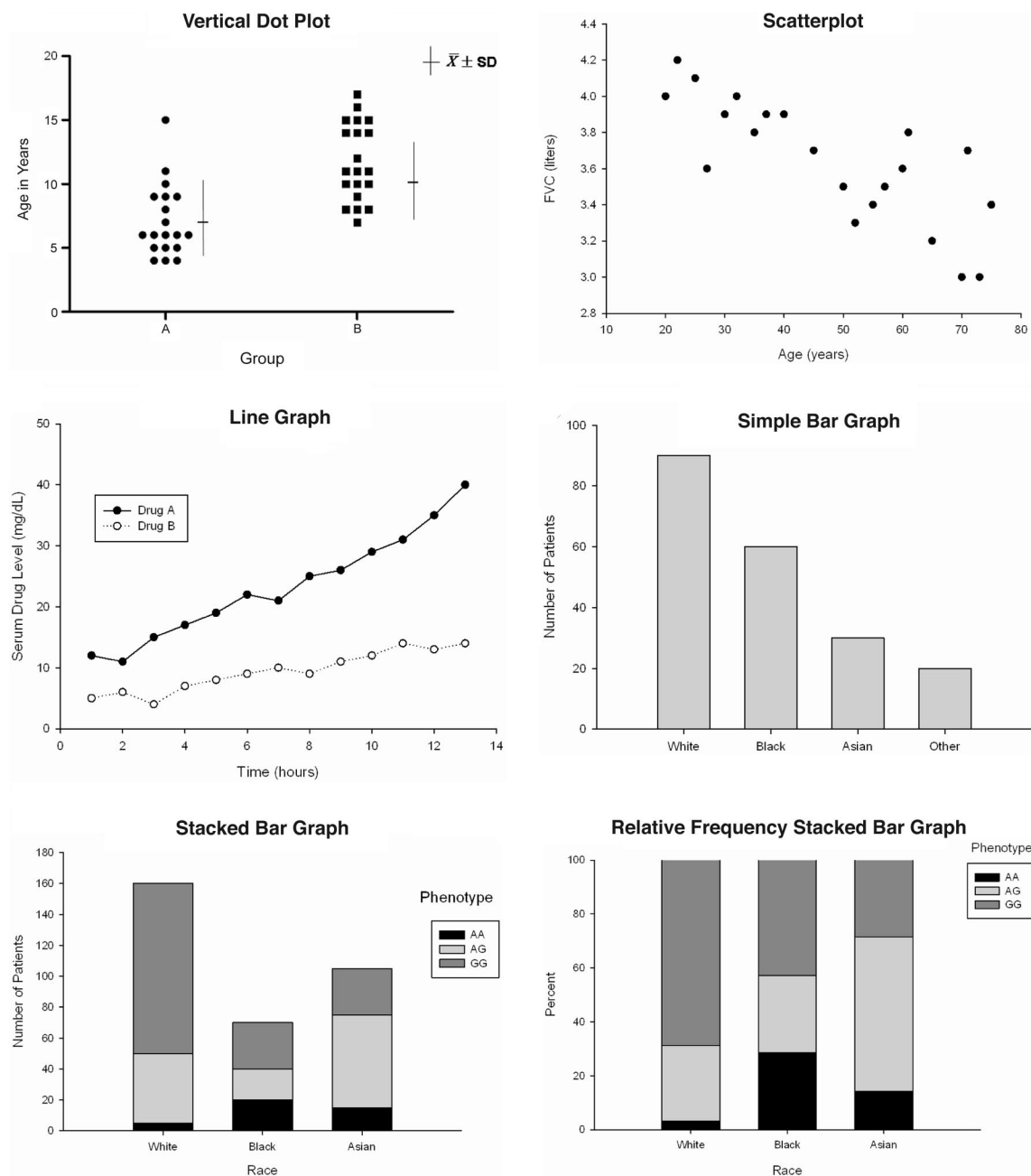


Figure 1. Examples of common graphic presentations.

comparisons. Information displayed in a graph is often derived from tabulated data. Graphs may be used as a substitute for tables or in addition to tables to highlight certain data selected from a table and are especially good when relationships among variables are complex.

Most graphs require a vertical scale (y-axis) and a horizontal scale (x-axis); one may represent numerical values

and the other may represent quantitative or qualitative variables. Equal intervals on an arithmetic scale should represent equal differences in value. The graph's title should be sufficiently descriptive yet concise. Whereas a table's title is placed at the top, the graph's title has been traditionally placed below the graph, at least in publications.

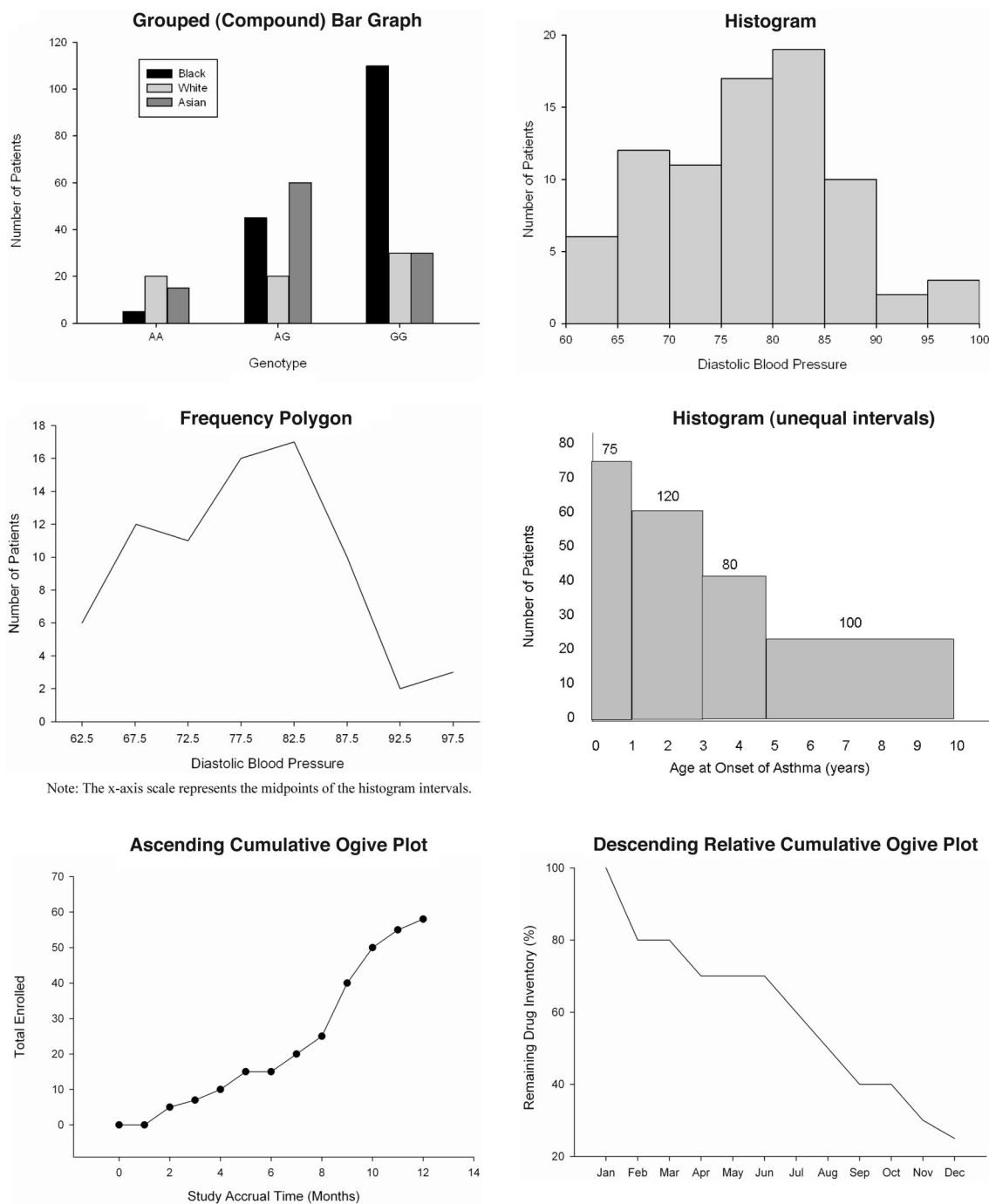


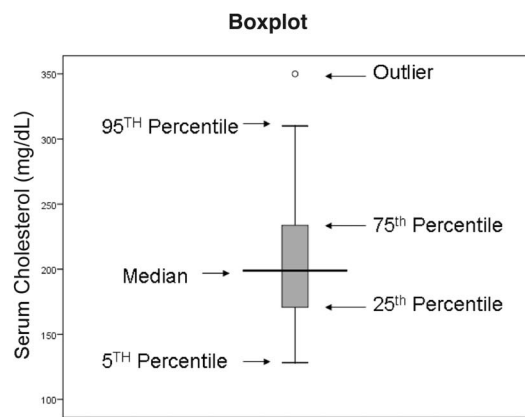
Fig. 1—Continued

Types of Graphs

The advances in computer technology have markedly facilitated the production of a large variety of graphic presentations, including complex ones. Several software packages are commercially available and continuously updated. The most

common basic types are briefly discussed herein. Examples of common graphs are presented in Figure 1.

The *dot graph* has the advantage of individual representation of each observation in a linear fashion and thus shows the range, pattern of spread, and any outlying observations. The



Note that there are several methods of drawing the “whiskers”, the vertical lines above and below the box. Some use the interquartile range (IQR), 75th percentile - 25th percentile, and the first point 1.5 IQR above (and below) the box as the ends of the whiskers. Outliers are more than 1.5 IQR from the box.

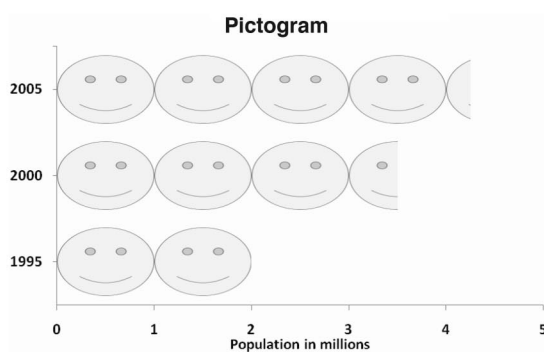
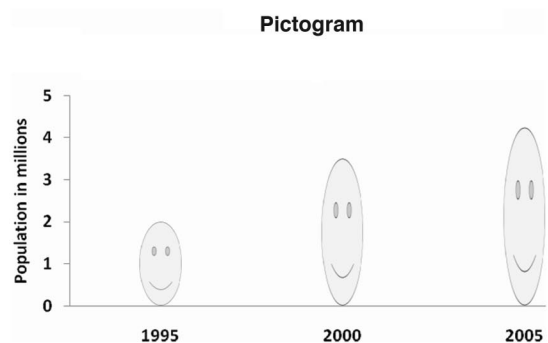
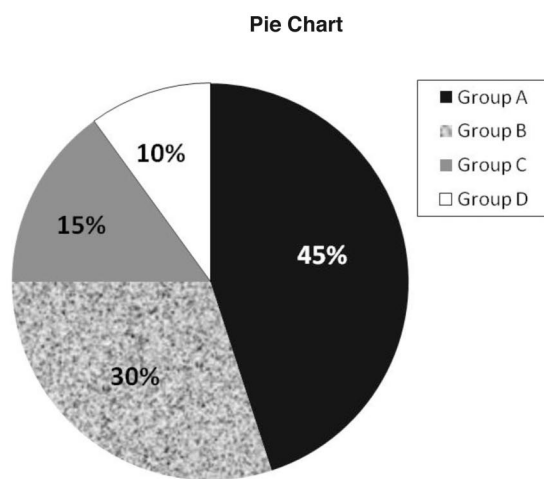
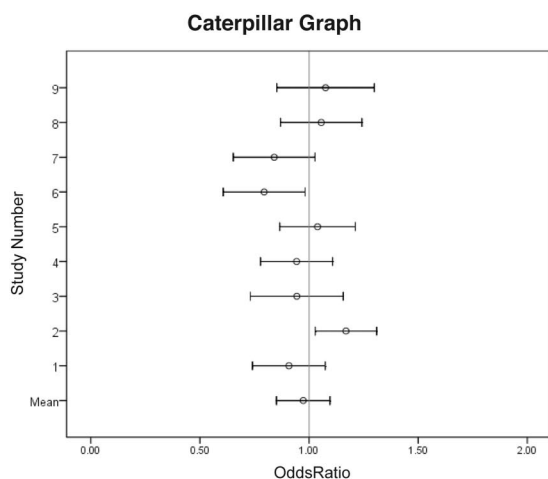
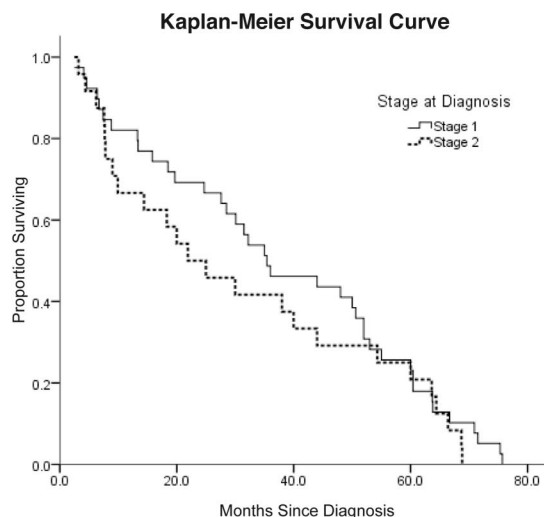


Fig. 1—Continued

mean may be added as a horizontal short line and the standard deviation (SD) or standard error (SE) by vertical lines. Alternative to the mean, the median and the lower quartile (25th percentile) and upper quartile (75th percentile) can be displayed. A dot plot is suitable for a small number of observa-

tions and can be used for presentation of multiple groups. Too many observations may produce a clutter and overlapping of too many dots.

The *scatter diagram* is a dot graph where each dot represents 2 measurements for each individual observation (eg,

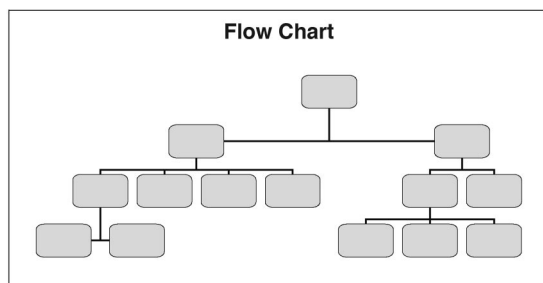


Fig. 1—Continued

height and weight). Hence, it gives a picture of correlation. A regression line or idealized curve can be superimposed on a scatter diagram.

The *line graph* is good for revealing a relationship or trend (ie, the change of numeric variable on the vertical axis in relation to an ordinal or a quantitative variable on the horizontal axis). It can be a single line or multiple, each representing a separate set of data. The line graph can allow presentation of the trend of 2 variables that have different units. In this case, the first variable is assigned the left vertical axis and the second variable is assigned the right vertical axis.

The *bar graph* is commonly used for qualitative or discrete quantitative variables. The height of the bar represents the value and in case of a mean may be combined with a vertical line representing the SD or SE. All tables and graphs should specify whether SD or SE are shown. It is a common mistake not to make this clear. The bars must be of the same width, and the distance in between can be half or equal to the bar's width. In a complex bar graph where multiple measurements are being represented, the bars of each set can be next to each other without spacing. In certain situations, the bar can be divided into parts representing the share of the bar's components, whether as absolute numbers or percentages. If the bar labels are lengthy, the axes may be switched.

The *histogram* is actually a set of bars next to each other without spacing to represent a continuous quantitative vari-

able. The horizontal axis represents the numerical variable and the vertical axis represents the frequency (absolute or relative). It differs from the bar graph in that the width of the bars may vary and each should reflect the width of the respective category and that the vertical axis represents counts of values not values themselves. Hence, the area, not the height, of the bar represents the frequency in that category. Histograms are useful to identify extreme values and to identify the underlying distribution of the measures (eg, normal, log normal), which are important to the choice of statistical analytical technique.

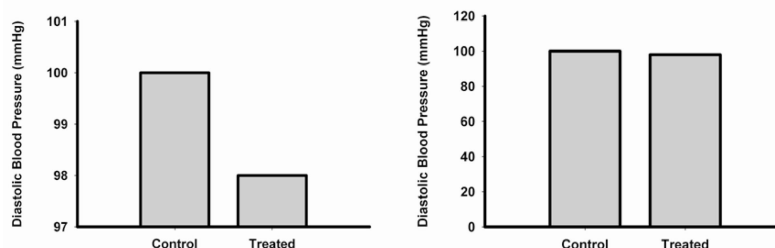
The *frequency polygon* is practically a line graph connecting the midpoints of the tops of the bars of a histogram. Hence, it can be used for multiple sets of data without plotting the histograms. It can also demonstrate the shape of distribution (eg, normality, skewness, or kurtosis) as addressed in detail in the preceding article on descriptive statistics.

The *box graph* (or box plot) is useful in presenting a measure of central tendency together with a certain spread of dispersion, for example, the mean with spreads of ± 1 SD and/or ± 2 SDs or the median (50th percentile) with spreads of certain percentiles (eg, 25th and 75th and/or 10th and 90th). The 25th and 75th percentiles are also called first (lower) and third (upper) quartiles, respectively, and the length of the box is called the interquartile range. Most commonly, medians and interquartile ranges are used for box plots. The figure annotation should specify what the symbols mean. It has the advantage of excluding any outlying or extreme observations (eg, below the first quartile or above the fourth quartile). Box plots also indicate presence or lack of symmetry in the distribution.

An *ogive* is a line graph representing cumulative frequencies, either ascending (begins with the lowest category) or descending (begins with the total of all categories).

Caterpillar plots are being increasingly used with the surge of meta-analysis studies. The data are gathered from previously published trials selected according to certain criteria to

Misleading Axis Scales*



*Note: the graph on the left is scaled such that the difference between the treated and control groups is exaggerated.

Figure 2. Example demonstrating the misleading visual effect of using an inappropriate scale.

maximize the homogeneity and soundness of the methods. It is a systematic review by experts who survey the literature to identify and characterize the quality of data to be included then calculate odds ratios (or relative risks) with a 95% confidence interval (CI). In addition, an overall odds ratio (displayed at the bottom of the graph) is estimated, taking into consideration a “weight” assigned to each study.

In the caterpillar chart, the conclusion about the findings of each trial is determined by the location of its respective odds ratio together with its whole 95% CI in relation to a line representing an odds ratio of 1.0, with a decreasing scale to the left and an increasing scale to the right. Overriding indicates nonsignificance.

Survivorship or *actuarial graphs* represent changes or probability of occurrence of an event over time relative to the starting point. A well-known example is the Kaplan-Meier graph, which is frequently used to demonstrate the natural course of a disease outcome or to compare different interventional methods on the long-term outcome.

A *pie graph* shows the relative distribution of a total circle into sections (slices of a pie). The magnitude of each component is commonly displayed numerically as a percentage (plus absolute numbers) inside or next to each slice of the circle. Multiple pie charts can be used to compare different sets of data.

Pictograms represent the variable as a picture the size of which reflects its approximate magnitude in contrast to the size of another similar picture. Another method can be representation by a number of pictures of a constant size. Each figurine represents a fixed amount, and a part of a figurine represents a fraction of that amount. Pictograms markedly lack precision and are usually used for commercial advertisements.

A *flow diagram* is used to demonstrate the stages of a study’s method or outcome, either in a vertical or a horizontal direction and often combined.

Graph Composition

The type of data being presented and the main objective of the presentation would determine the most appropriate choice of the following:

- Type of graph.
- The measurement scale regarding spacing.
- The presence of a scale break. If a scale break is used, the 2 points on both sides of the break should not be connected.
- The starting point of the scale (ie, zero or another value).
- The choice of an arithmetic or logarithmic scale. Many biological variables follow logarithmic or exponential distributions and may best be represented on a logarithmic scale.
- Clear labeling of the scales, including identification and units (eg, number, percentage, kg, mg/dL).
- A legend may be needed to supplement the labeling and is best placed in the right upper quarter of the graph.
- A reference line can be included horizontally across the entire graph. A reference range can be included as 2 lines or as a lightly shaded area that does not hide the data.
- The choice of scale is an important issue. Small differences can be made to appear large by expanding the scale of the vertical axis to encompass a small range of values. Conversely, important differences can be masked by compressing the scale to cover too large a range (Figure 2). Where practical, it is good to have a scale starting with the value 0 or at least the typical range of numbers that could be observed. It is also important with a series of graphs to make the scales cover a similar range so that the graphs are visually comparable.

REFERENCES

Included in a comprehensive list at the end of the Supplement.

Requests for reprints should be addressed to:

Sami L. Bahna, MD, DrPH

Department of Pediatrics

Allergy and Immunology Section

Louisiana State University Health Sciences Center

1501 Kings Highway

Shreveport, LA 71103-3932

E-mail: SBAHNA@LSUHSC.EDU