Radiology

Seema S. Sonnad, PhD

Index terms:
Data analysis
Statistical analysis

Published online before print
10.1148/radiol.2253012154
Radiology 2002; 225:622–628

[1] From the Department of Surgery, University of Michigan Medical Center, Ann Arbor. Received January 14, 2002; revision requested March 2; revision received May 20; accepted June 14. **Address correspondence to** the author, Department of Surgery, University of Pennsylvania Health System, 4 Silverstein, 3400 Spruce St, Philadelphia, PA 19104-4283 (e-mail: *seema.sonnad@uphs.upenn.edu*).

Describing Data: Statistical and Graphical Methods[1]

An important step in any analysis is to describe the data by using descriptive and graphic methods. The author provides an approach to the most commonly used numeric and graphic methods for describing data. Methods are presented for summarizing data numerically, including presentation of data in tables and calculation of statistics for central tendency, variability, and distribution. Methods are also presented for displaying data graphically, including line graphs, bar graphs, histograms, and frequency polygons. The description and graphing of study data result in better analysis and presentation of data.

A primary goal of statistics is to collapse data into easily understandable summaries. These summaries may then be used to compare sets of numbers from different sources or to evaluate relationships among sets of numbers. Later articles in this series will discuss methods for comparing data and evaluating relationships. The focus of this article is on methods for summarizing and describing data both numerically and graphically. Options for constructing measures that describe the data are presented first, followed by methods for graphically examining your data. While these techniques are not methodologically difficult, descriptive statistics are central to the process of organizing and summarizing anything that can be presented as numbers. Without an understanding of the key concepts surrounding calculation of descriptive statistics, it is difficult to understand how to use data to make comparisons or draw inferences, topics that will be discussed extensively in future articles in this series.

In this article, five properties of a set of numbers will be discussed. *(a)* Location or central tendency: What is the central or most typical value seen in the data? *(b)* Variability: To what degree are the observations spread or dispersed? *(c)* Distribution: Given the center and the amount of spread, are there specific gaps or concentrations in how the data cluster? Are the data distributed symmetrically or are they skewed? *(d)* Range: How extreme are the largest and smallest values of the observations? *(e)* Outliers: Are there any observations that do not fit into the overall pattern of the data or that change the interpretation of the location or variability of the overall data set?

The following tools are used to assess these properties: *(a)* summary statistics, including means, medians, modes, variances, ranges, quartiles, and tables; and *(b)* plotting of the data with histograms, box plots, and others. Use of these tools is an essential first step to understand the data and make decisions about succeeding analytic steps. More specific definitions of these terms can be found in the Appendix.

DESCRIPTIVE STATISTICS

Frequency Tables

One of the steps in organizing a set of numbers is counting how often each value occurs. An example would be to look at diagnosed prostate cancers and count how often in a 2-year period cancer is diagnosed as stage A, B, C, or D. For example, of 236 diagnosed cancers, 186 might be stage A, 42 stage B, six stage C, and two stage D. Because it is easier to understand these numbers if they are presented as percentages, we say 78.8% (186 of 236) are stage A, 17.8% (42 of 236) are stage B, 2.5% (six of 236) are stage C, and 0.9% (two of 236) are stage D. This type of calculation is performed often, and two definitions are important. The frequency of a value is the number of times that value occurs in a given data set. The relative frequency of a value is the proportion of all observations in the data set with that value. Cumulative frequency is obtained by adding relative frequency for one

TABLE 1
Frequencies, Relative Frequencies, and Cumulative Frequencies of Cancer Staging Distribution

Cancer Stage	Frequency	Cumulative Frequency	Percentage	Relative Frequency (proportion)
A	186	.788	78.8	.79
B	42	.967	17.9	.18
C	6	.992	2.5	.025
D	2	1.0	0.9	.009

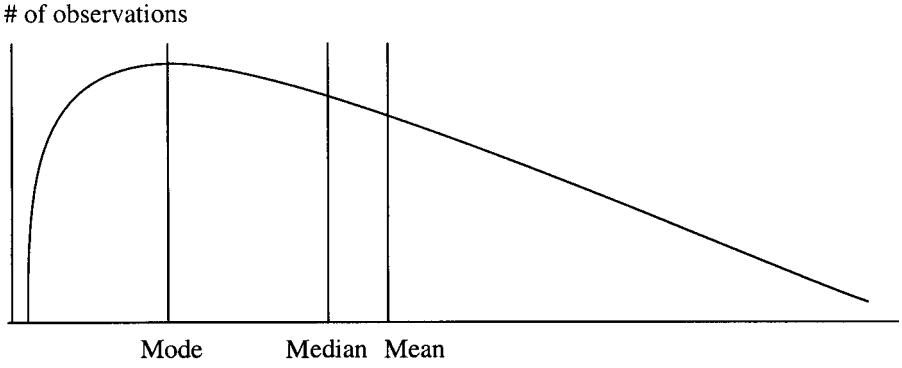

Figure 1. Line graph shows the mean, median, and mode of a skewed distribution. In a distribution that is not symmetric, such as this one, the mean (arithmetic average), the median (point at which half of the data lie above and half lie below), and the mode (most common value in the data) are not the same.

value at a time. The cumulative frequency of the first value would be the same as the relative frequency and that of the first two values would be the sum of their relative frequencies, and so on. Frequency tables appear regularly in *Radiology* articles and other scientific articles. An example of a frequency table, including both frequencies and percentages, is shown in Table 1. In the section of this article about graphic methods, histograms are presented. They are a graphic method that is analogous to frequency tables.

Measurement of the Center of the Data

The three most commonly used measures of the center of the data are the mean, median, and mode. The mean (often referred to as the average) is the most commonly used measure of center. It is most often represented in the literature as $\bar{x}$. The mean is the sum of the values of the observations divided by the number of observations. The median is the midpoint of the observations, when arranged in order. Half of the observations in a data set lie below the median and half lie above the median. The mode is the most frequent value. It is the value that occurs most commonly in the data set. In a data set like the one in Figure 1, the mean, median, and mode will be different numbers. In a perfectly normal distribution, they will all be the same number. A normal distribution is a commonly occurring symmetric distribution, which is defined by the familiar bell-shaped curve, that includes a set percentage of data between the center and each standard deviation (SD) unit.

When a median is used, the data are arranged in order and the middle observation is selected. If the number of observations is even, there will be a middle pair of values, and the median will be the point halfway between those two values. The numbers must be ordered when the median is selected, and each observation must be included. With a data set of 4, 4, 5, 5, 6, 6, 8, and 9, the median is 5.5. However, if values represented by the data were listed rather than listing the value of each observation, the median would be erroneously reported as 6. Finally, if there are 571 observations, there is a method to avoid counting in from the ends. Instead, the following formula is used: If there are n observations, calculate $(n + 1)/2$. Arrange the observations from smallest to largest, and count $(n + 1)/2$ observations up from the bottom. This gives the median. In real life, many statistical programs (including Excel; Microsoft, Redmond, Wash), will give the mean, median, and mode, as well as many other descriptive statistics for data sets.

To look for the mode, it is helpful to create a histogram or bar chart. The most common value is represented by the highest bar and is the mode. Some distributions may be bimodal and have two values that occur with equal frequency. When no value occurs more than once, they could all be considered modes. However, that does not give us any extra information. Therefore, we say that these data do not have a mode. The mode is not used often because it may be far from the center of the data, as in Figure 1, or there may be several modes, or none. The main advantage of the mode is that it is the only measure that makes sense for variables in nominal scales. It does not make sense to speak about the median or mean race or sex of radiologists, but it does make sense to speak about the most frequent (modal) race (white) or sex (male) of radiologists. The median is determined on the basis of order information but not on the basis of the actual values of observations. It does not matter how far above or below the middle a value is, but only that it is above or below.

The mean comprises actual numeric values, which may be why it is used so commonly. A few exceptionally large or small values can significantly affect the mean. For example, if one patient who received contrast material before computed tomography (CT) developed a severe life-threatening reaction and had to be admitted to the intensive care unit, the cost for care of that patient might be several hundred thousand dollars. This would make the mean cost of care associated with contrast material–enhanced CT much higher than the median cost because without such an episode costs might only be several hundred dollars. For this reason, it may be most appropriate to use the median rather than the mean for describing the center of a data set if the data contain some very large or very small outlier values or if the data are not centered (Fig 1).

"Skewness" is another important term. While there is a formula for calculating skew, which refers to the degree to which a distribution is asymmetric, it is not commonly used in data analysis. Data that are skewed right (as seen in Fig 1) are common in biologic studies because many measurements involve variables that have

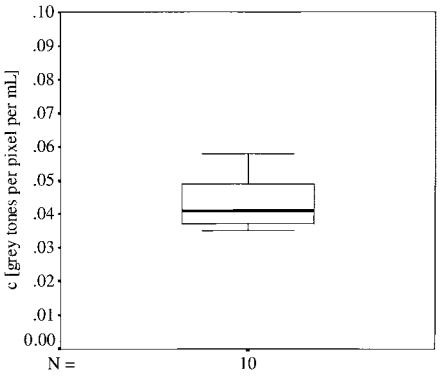

Figure 2. Box plot demonstrates low variation. A high-variation box plot would be much taller. The horizontal line is the median, the ends of the box are the upper and lower quartiles, and the vertical lines are the full range of values in the data. (Reprinted, with permission, from reference 1.)

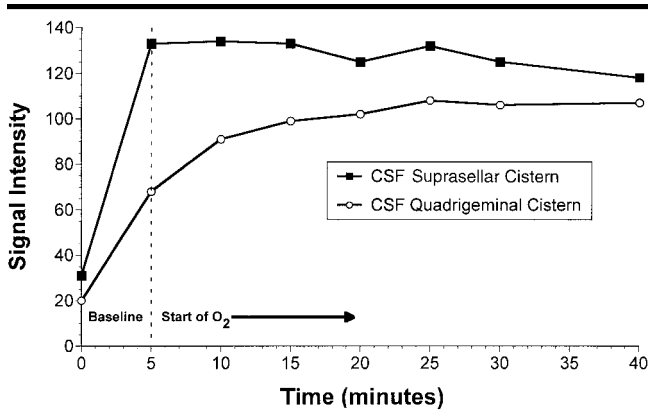

Figure 3. Line graph shows change in MR signal intensity in the cerebrospinal fluid *(CSF)* collected during oxygen inhalation over time. Signal intensity in the quadrigeminal plate cistern increases more gradually, and equilibration is reached at 15–20 minutes after the start of oxygen inhalation. (Reprinted, with permission, from reference 3.)

a natural lower boundary but no definitive upper boundary. For example, hospital length of stay can be no shorter than 1 day or 1 hour (depending on the units used by a given hospital), but it could be as long as several hundred days. The latter would result in a distribution with more values below some cutoff and then a few outliers that create a long "tail" on the right.

TABLE 2
Intra- and Interobserver Variability in MR Reading: Automated versus Manual Method

Segmented Volume and Tumor Histologic Type	Manual Method		Automated Method	
	Intraobserver	Interobserver	Intraobserver	Interobserver
Brain				
Meningioma	0.42 ± 0.03	4.93 ± 1.75	0.36 ± 0.45	1.84 ± 0.65
Low-grade glioma	1.79 ± 1.53	6.31 ± 2.85	1.44 ± 1.33	2.71 ± 1.68
Tumor				
Meningioma	1.58 ± 0.98	7.08 ± 2.18	0.66 ± 0.72	2.66 ± 0.38
Low-grade glioma	2.08 ± 0.78	13.61 ± 2.21	2.06 ± 1.73	2.97 ± 1.58

Note.—Data are the mean coefficient of variation percentage plus or minus the SD. (Adapted and reprinted, with permission, from reference 2.)

Measuring Variability in the Data

Measures of center are an excellent starting point in summarizing data, but they usually do not "tell the full story" and can be misleading if there is no information about the variability or spread of the data. An adequate summary of a set of data requires both a measure of center and a measure of variability. Just as with the center, there are several options for measuring variability. Each measure of variability is most often associated with one of the measures of center. When the median is used to describe the center, the variability and general shape of the data distribution are described by using percentiles. The *x*th percentile is the value at which *x* percent of the data lie below that percentile and the rest lie above it; therefore, the median is also the 50th percentile. As seen in the box plot (Fig 2), the 25th and 75th percentiles, also known as the lower and upper quartiles, respectively, are often used to describe data (1).

Although the percentiles and quartiles used for creating box plots are useful and simple, they are not the most common measures of spread. The most common measure is the SD. The SD (and the related variance) is used to describe spread around the center when the center is expressed as a mean. The formula for variance would be written as

$$\frac{\Sigma\ (\text{obs} - \text{mean})^2}{\text{no. of obs}},$$

where Σ represents a summation of all the values, and obs means observations. Squaring of the differences results in all positive values. Then the SD is

$$\sqrt{\frac{\Sigma\ (\text{obs} - \text{mean})^2}{\text{no. of obs}}},$$

the square root of the variance. It is helpful to think of the SD as the average distance of observations from the mean. Because it is a distance, it is always positive. Large outliers affect the SD drastically, just as they do the mean. Occasionally, the coefficient of variation—the SD or mean multiplied by 100 to get a percentage value—is used. This can be useful if the interest is in the percentage variation rather than the absolute value in numeric terms. Kaus and colleagues (2) presented an interesting table in their study. Table

TABLE 3
Standard Scores and Corresponding Percentiles

Standard Score	Percentile
−3.0	0.13
−2.5	0.62
−2.0	2.27
−1.5	6.68
−1.0	15.87
−0.5	30.85
0.0	50.00
0.5	69.15
1.0	84.13
1.5	93.32
2.0	97.73
2.5	99.38
3.0	99.87

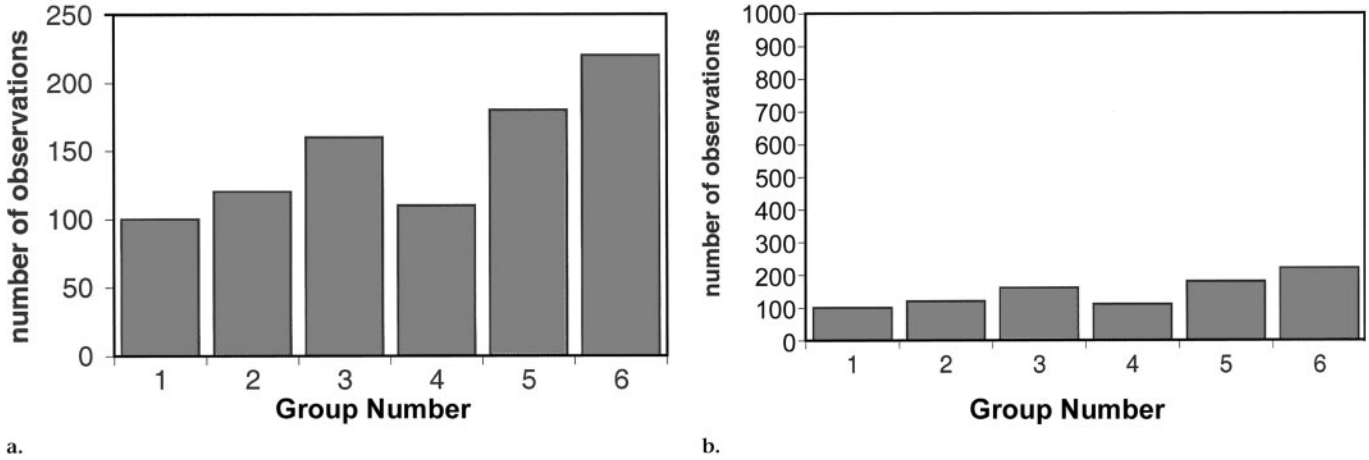

Figure 4. Bar graphs. **(a)** Use of a scale with a maximum that is only slightly higher than the highest value in the data shows the differences between the groups more clearly than does **(b)** use of a scale with a maximum that is much higher than the highest value.

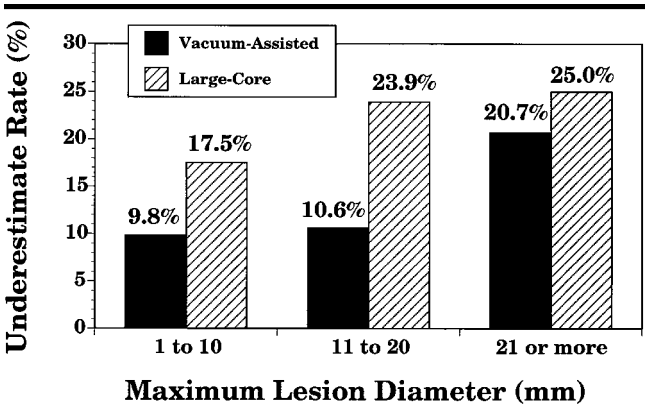

Figure 5. Bar graph shows the underestimation rate for lesion size with vacuum-assisted and large-core needle biopsy. Underestimation rates were lower with the vacuum-assisted device. It is helpful to label the bars with the value. (Reprinted, with permission, from reference 4.)

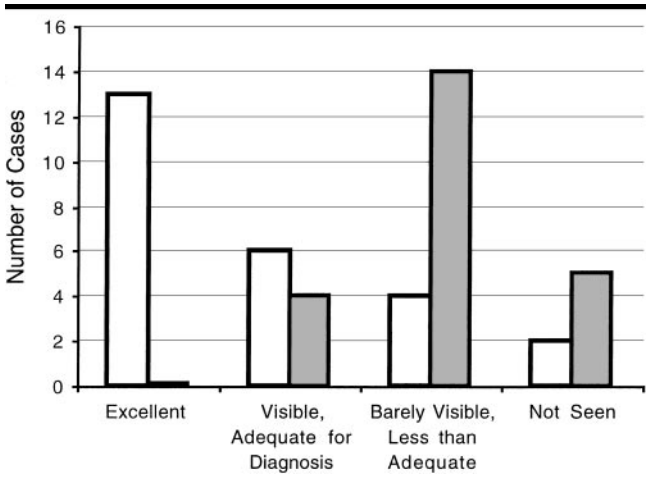

Figure 6. Bar graph shows the diagnostic adequacy of high-spatial-resolution MR angiography with a small field of view and that with a large field of view for depiction of the segmental renal arteries. (Reprinted, with permission, from reference 5.)

2 shows their data on the intra- and interobserver variability in the reading of magnetic resonance (MR) images with automated and manual methods.

Normal Distribution and Standard Scores

Once a mean and SD are calculated, they can be used to further simplify description of data or to create statistics that can be compared across data sets. One common measure is the standard score. With the standard score, data are assumed to come from a normal distribution, a symmetric bell-shaped distribution. In a normal distribution, 68% of all observations are within 1 SD of the mean (34% above the mean and 34% below). Another 27% are between 1 and 2 SDs; therefore, 95% of all observations are within 2 SDs of the mean. A total of 99.7% are within 3 SDs of the mean; therefore, any normal curve (or histogram that represents a large set of data drawn from a normal distribution) is about 6 SDs wide. Two SDs from the mean is often used as a cutoff for the assignment of values as outliers. This convention is related to the common use of 95% confidence intervals and the selection of confidence for testing of a hypothesis as 95% (these concepts will be defined in a future article). The use of 2 SDs from the mean in normally distributed data ensures that 95% of the data are included. As can be seen, the SD is the common measure of variability for data from normal distributions. These data can be expressed as standard scores, which are a measure of SD units from the mean. The standard score is calculated as $(OV - M)/SD$, where OV is the observation value and M is the mean. A standard score of 1 corresponds to the 84th percentile of data from a normal distribution. Standard scores are useful because if the data are drawn from a normal distribution, regardless of the original mean and SD, each standard score corresponds to a specific percentile. Table 3 shows percentiles that correspond to some standard scores.

GRAPHICAL METHODS FOR DATA SUMMARY

Line and Bar Graphs

It is often easier to understand and interpret data when they are presented graphically rather than descriptively or as

a table. Line graphs are most often used to show the behavior of one variable over time. Time appears on the horizontal axis and the variable of interest on the vertical axis. Figure 3 is an example of a line graph. The variable is MR signal intensity in the cerebrospinal fluid collected during oxygen inhalation. It is apparent from this graph that signal intensity within the quadrigeminal plate cistern increases more gradually than that within the suprasellar cistern, which was the conclusion drawn by the authors (3).

When you look at graphs, it is important to examine both the horizontal and vertical axis scales. Selection of the scale to be used may influence how the reader interprets the graph. This can be seen in Figure 4, which depicts bar graphs.

Bar graphs are another common way to display data. They are used for comparing the value of multiple variables. In many cases, multiple bar graphs will appear in the same figure, such as in Figures 5 and 6. Figure 5 shows the diagnostic underestimation rate for three categories of lesion size with vacuum-assisted and large-core needle biopsies. This bar chart is presented with a percentage rate on the y axis and the categories on the x axis (4). Figure 6 is similar and shows the diagnostic adequacy of high-spatial-resolution MR angiography with a small field of view compared with that with a large field of view for depiction of the segmental renal arteries. In this case, the y axis represents the number of cases and the percentages appeared in the figure caption (5). Bars may be drawn either vertically, as in Figures 5 and 6, or horizontally. They may be drawn to be touching each other or to be separate. It is important that the width of the bars remains consistent so that one bar does not appear to represent more occurrences because it has a larger area.

TABLE 4
Data for Body Weight of 10 Patients Used to Construct Stem and Leaf Plot

Volunteer No./ Age (y)/Sex	Body Weight (kg)	Height (cm)	ΔQ_F (gray tones per pixel)*	c (gray tones per pixel per milliliter)
1/27/F	76	178	4.7	0.049
2/32/F	61	173	3.3	0.035
3/31/M	83	181	3.8	0.040
4/28/M	85	180	4.0	0.042
5/44/M	103	189	4.6	0.048
6/32/M	72	179	3.4	0.036
7/23/M	78	174	5.5	0.058
8/26/F	72	179	4.8	0.050
9/28/M	80	180	3.5	0.037
10/23/F	74	177	3.6	0.038

Note.—The mean values ± SEM were as follows: age, 29 years ± 2; body weight, 78 kg ± 3; height, 179 cm ± 1; ΔQ_F, 4.1 gray tones per pixel ± 0.2; c ($\Delta Q_F/V$), 0.043 gray tones per pixel per milliliter ± 0.003. (Adapted and reprinted, with permission, from reference 1).

* Ninety-five milliliters of saline solution was instilled.

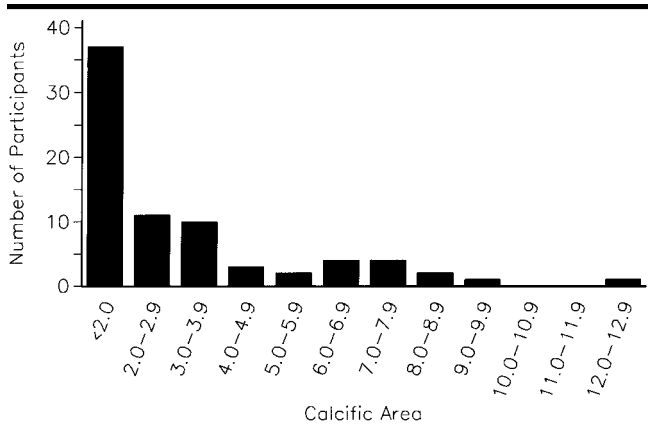

Figure 7. Histogram represents distribution of calcific area. Labels on the x axis indicate the calcific area in square millimeters for images with positive findings. (Reprinted, with permission, from reference 6.)

Histograms and Frequency Polygons

Histograms look somewhat like bar charts, but they serve a different purpose. Rather than displaying occurrence in some number of categories, a histogram is intended to represent a sampling distribution. A sampling distribution is a representation of how often a variable would have each of a given set of values if many samples were to be drawn from the total population of that variable. The bars in a histogram appear in a graph where the y axis is frequency and the x axis is marked in equal units. Remember that a bar chart does not have x-axis units. When a program automatically creates a histogram, it creates x-axis units of equal size. The histogram is analogous to the frequency table discussed earlier. Figure 7 shows a histogram, although it was called a bar chart in the original figure caption (6). This figure represents the distribution of calcific area among participants and has equal units on the x axis, which makes it a histogram.

Frequency polygons are a less used alternative to histograms. Figure 8a shows a histogram that might represent the distribution of the mean partition coefficients in a number of healthy individuals. The line in the figure is created by connecting the middle of each of the histogram bars. The figure represented by that line is called a frequency polygon. The frequency polygon is shown in Figure 8b. In Figure 8a, both the histogram and the frequency polygon are shown on the same graph. Most often, one or the other appears but not both.

Stem and Leaf Plots

Another graphic method for representing distribution is known as the stem and leaf plot, which is useful for relatively small data sets, as seen in many radiologic investigations. With this approach, more of the information from the original data is preserved than is preserved with the histogram or frequency polygon, but a graphic summary is still provided. To make a stem plot, the first digits of the data values represent the stems. Stems appear vertically with a vertical line to their right, and the digits are sorted into ascending order. Then the second digit of each value occurs as a leaf to the right of the proper stem. These leaves should also be sorted into ascend-

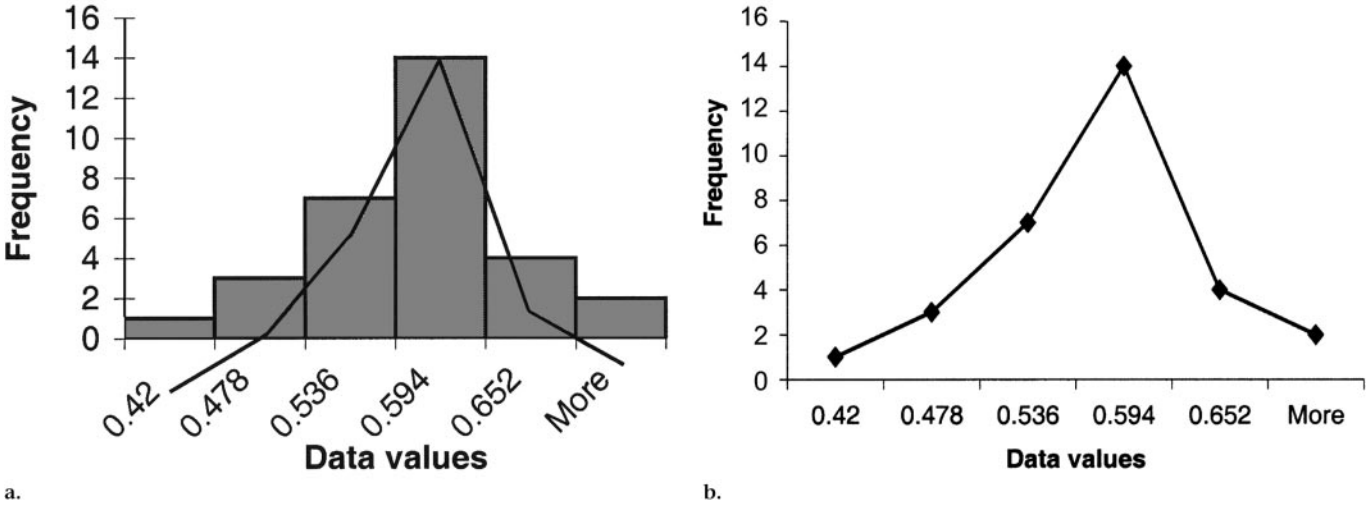

Figure 8. Representative histogram and frequency polygon constructed from hypothetical data. Alternate ways of showing the distribution of a set of data are shown **(a)** with both the histogram and the frequency polygon depicted or **(b)** with only the frequency polygon depicted.

STEP 1	STEP 2	
6	6	1
7	7	2 2 4 6 8
8	8	0 3 5
9	9	
10	10	3

Figure 9. Stem and leaf plot constructed from data in Table 4. Step one shows construction of the stem, and step 2 shows construction of the leaves. These steps result in a sideways histogram display of the data distribution, which preserves the values of the data used to construct it. The number *9* appears on the stem as a place saver; the lack of digits to the right of this stem number indicates that no values began with this number in the original data.

ing order. In a simple example, Heverhagen and colleagues (1) used the body weight in kilograms of 10 patients. The original data appear in Table 4. Figure 9 shows the making of a stem plot from these data. The stem and leaf plot looks like a histogram with horizontal bars made of numbers. The plot's primary advantage is that all of the actual values of the observations are retained. Stem and leaf plots do not work well with very large data sets because there are too many leaves, which makes it difficult both to read and to fit on a page.

Box Plots

The final graphic method presented in this article is the box plot. The box plot shows the distribution of the data and is especially useful for comparing distributions graphically. It is created from a set of five numbers: the median, the 25th percentile or lower quartile, the 75th percentile or upper quartile, the minimum data value, and the maximum data value. The horizontal line in the middle of the box is the median of the measured values, the upper and lower sides of the box are the upper and lower quartiles, and the bars at the end of the vertical lines are the data minimum and maximum values.

Tombach and colleagues (7) used box plots in their study of renal tolerance of a gadolinium chelate to show changes over time in the distribution of values of serum creatinine concentration and creatinine clearance across different groups of patients. Some of these box plots are shown in Figure 10. They clearly demonstrate the changing distribution and the difference in change between patient groups.

In conclusion, a statistical analysis typically requires statistics in addition to a measure of location and a measure of variability. However, the plotting of data to see their general distribution and the computing of measures of location and spread are the first steps in being able to determine the interesting relationships that exist among data sets. In this article, methods have been provided for calculating the data center with the mean, median, or mode and for calculating data spread. In addition, several graphic methods were explained that are useful both when exploring and when presenting data. By starting with describing and graphing of study data, better analysis and clear presentation of data will result; therefore, descriptive and graphic methods will improve communication of important research findings.

APPENDIX

The following is a list of the common terms and definitions related to statistical and graphic methods of describing data.

Coefficient of variation.—SD divided by the mean and then multiplied by 100%.

Descriptive statistics.—Statistics used to summarize a body of data. Contrasted with inferential statistics.

Frequency distribution.—A table that shows a body of data grouped according to numeric values.

Frequency polygon.—A graphic method of presenting a frequency distribution.

Histogram.—A bar graph that represents a frequency distribution.

Inferential statistics.—Use of sample statistics to infer characteristics about the population.

Mean.—The arithmetic average for a group of data.

Median.—The middle item in a group of data when the data are ranked in order of magnitude.

Mode.—The most common value in any distribution.

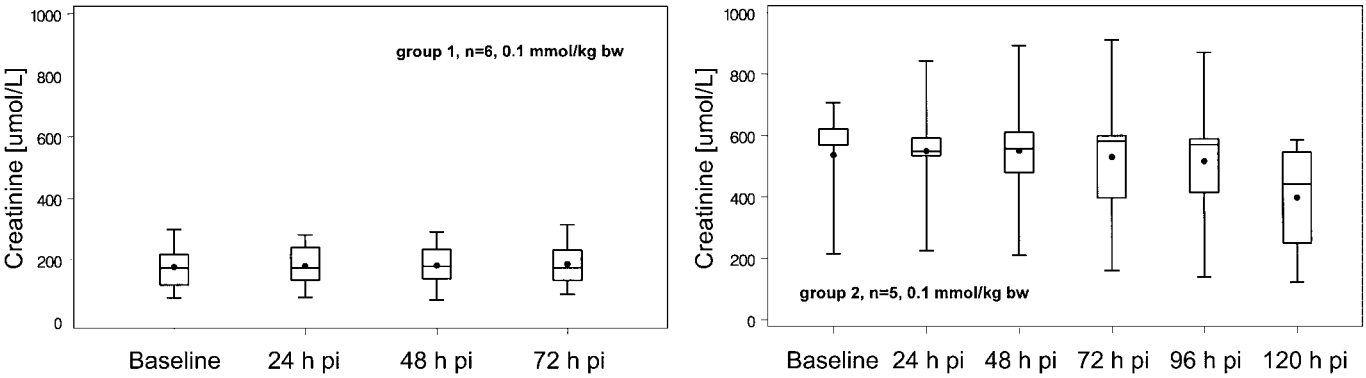

Figure 10. Box plots present follow-up data for serum creatinine concentration. Left: Within 72 hours after injection of a gadolinium chelate (group 1 baseline creatinine clearance, <80 mL/min [<0.50 mL/sec]). Right: Within 120 hours after injection (group 2 baseline creatinine clearance, <30 mL/min [<0.50 mL/sec]). (Reprinted, with permission, from reference 7.)

Nominal data.—Data with items that can only be classified into groups. The groups cannot be ranked.

Normal distribution.—A bell-shaped curve that describes the distribution of many phenomena. A symmetric curve with the highest value in the center and with set amounts of data on each side with the mathematical property that the logarithm of its probability density is a quadratic function of the standardized error.

Percentage distribution.—A frequency distribution that contains a column listing the percentage of items in each class.

Quartile.—Value below which 25% (lower quartile) or 75% (upper quartile) of data lie.

Sample.—A subset of the population that is usually selected randomly. Measures that summarize a sample are called sample statistics.

Sampling distribution.—The distribution actually seen (often represented with a histogram) when data are drawn from an underlying population.

Standard deviation.—A measure of dispersion, the square root of the average squared deviation from the mean.

Variance.—The average squared deviation from the mean, or the square of the SD.

References

1. Heverhagen JT, Muller D, Battmann A, et al. MR hydrometry to assess exocrine function of the pancreas: initial results of noninvasive quantification of secretion. Radiology 2001; 218:61–67.
2. Kaus MR, Warfield SK, Nabavi A, Black PM, Jolesz FA, Kikinis R. Automated segmentation of MR images of brain tumors. Radiology 2001; 218:586–591.
3. Deliganis AV, Fisher DJ, Lam AM, Maravilla KR. Cerebrospinal fluid signal intensity increase on FLAIR MR images in patients under general anesthesia: the role of supplemental O(2). Radiology 2001; 218: 152–156.
4. Jackman RJ, Burbank F, Parker SH, et al. Stereotatic breast biopsy of nonpalpable lesions: determinants of ductal carcinoma in situ underestimation rates. Radiology 2001; 218:497–502.
5. Fain SB, King BF, Breen JF, Kruger DG, Riederer SJ. High-spatial-resolution contrast-enhanced MR angiography of the renal arteries: a prospective comparison with digital subtraction angiography. Radiology 2001; 218:481–490.
6. Bielak LF, Sheedy PF II, Peyser PA. Automated segmentation of MR images of brain tumors. Radiology 2001; 218:224–229.
7. Tombach B, Bremer C, Reimer P, et al. Renal tolerance of a neutral gadolinium chelate (gadobutrol) in patients with chronic renal failure: results of a randomized study. Radiology 2001; 218:651–657.