The online version of this article, along with updated information and services, is located on the World Wide Web at:

http://circ.ahajournals.org/cgi/content/full/114/1/76

Subscriptions: Information about subscribing to Circulation is online at
http://circ.ahajournals.org/subsriptions/

Permissions: Permissions & Rights Desk, Lippincott Williams & Wilkins, 351 West Camden Street, Baltimore, MD 21202-2436. Phone 410-5280-4050. Fax: 410-528-8550. Email:
journalpermissions@lww.com

Reprints: Information about reprints can be found online at
http://www.lww.com/static/html/reprints.html

Descriptive Statistics and Graphical Displays

Martin G. Larson, SD

*S*tatistics is a broad mathematical discipline dealing with techniques for the collection, analysis, interpretation, and presentation of numerical data. Data are information used for reasoning, discussion, or calculation; data are the foundation of modern scientific inference. Data may be obtained by a formal sampling procedure, by recording responses to experimental conditions, or by observing a process repeatedly over time. Once data are collected, statistical analysis typically begins by calculating *descriptive statistics*—numbers that characterize features of those specific data—and by presenting the descriptive statistics in tables or graphs. In contrast, *inferential statistics*—statistics for making inferences about the populations from which data are sampled—is a related, broader category of statistical analysis. Forthcoming articles in this series will cover topics in inferential statistics. In this article, we consider elementary descriptive statistics. Similar material can be found in standard statistical textbooks or online.[1–4]

Types of Variables

The characteristics of interest in a research study are called *variables*, measurable quantities that vary among individuals (for our purposes, an individual may be a person, animal, place, or thing) or within individuals over time. By contrast, *parameters* are not actual measurements or attributes of individuals but are quantities that define a statistical model.

Variables are classified as *discrete* or *continuous*: Discrete variables can assume only certain values (fixed and readily countable), whereas continuous variables can assume an infinite number of values. Examples of discrete variables commonly encountered in cardiovascular research include species, strain, racial/ethnic group, sex, education level, treatment group, hypertension status, and New York Heart Association class. Corresponding examples of continuous variables include age, height, weight, blood pressure, measures of cardiac structure and function, blood chemistries, and survival time.

Discrete variables (also called categorical variables) are divided into 2 subtypes: *nominal* (unordered) and *ordinal* (ordered). Nominal variables take values such as yes/no, human/dog/mouse, female/male, treatment A/B/C; a nominal variable that takes only 2 possible values is called binary. One may apply numbers as labels for nominal categories, but there is no natural ordering. Ordinal variables take naturally or-dered values such as New York Heart Association class (I, II, III, or IV), hypertension status (optimal, normal, high-normal, or hypertensive), or education level (less than high school, high school, college, graduate school). Ordering among the categories is meaningful, but spacing between categories may be arbitrary.

In contrast with discrete variables, continuous variables have fixed intervals between adjacent values. Consequently, they can be manipulated mathematically, taking sums and differences, for example, whereas discrete variables cannot be manipulated that way.

Descriptive Statistics

The application of statistics to problems in cardiovascular research typically begins by defining the *population* of interest with respect to time, place, and other features. As examples, the population might be all people in the United States at mid-year 2000, all cases of acute myocardial infarction in the United States during the year 2000, or all cardiac myocytes in those acute myocardial infarction cases. Practical reasons usually dictate that investigators collect information on a *sample*, a defined subset of the population, and that they use the sample to represent the larger population.

Data analysis begins with calculation of descriptive statistics for the research variables. These statistics summarize various aspects about the data, giving details about the sample and providing information about the population from which the sample was drawn. Each variable's type determines the nature of descriptive statistics that one calculates and the manner in which one reports or displays those statistics.

Frequency statistics are the main descriptive statistics used with discrete variables. These include *absolute frequencies* (raw counts) for each category of the discrete variable, *relative frequencies* (proportions or percentages of the total number of observations), and *cumulative frequencies* for successive categories of ordinal variables.

Consider data on body mass index (BMI, kg/m^2) collected for 3480 participants in the Framingham Offspring Study sixth examination, 1995 through 1998. Height was measured to the nearest quarter inch and weight to the nearest pound; each was converted to metric units. Thus, for practical purposes, raw BMI was a continuous variable. To illustrate data for an ordinal variable, BMI values were collapsed into

---

TABLE 1.   Distribution of BMI in Framingham Offspring Participants, Sixth Examination

BMI, kg/m^2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<18.5	13	0.4	13	0.4
18.5 to <25.0	1044	30.0	1057	30.4
25.0 to <30.0	1447	41.6	2504	72.0
30.0 to <35.0	678	19.5	3182	91.4
35.0 to <40.0	192	5.5	3374	97.0
≥40.0	106	3.1	3480	100.0

TABLE 2.   Descriptive Statistics for BMI (kg/m^2) in Framingham Offspring Participants, Sixth Examination

Sample Size, n	3480	Variance	26.5
Sum	97 190.4	Standard deviation	5.1
Mean	27.9	Skewness	1.1
Mode	26.4	Kurtosis	2.0
Q1	24.4	Minimum	16.6
Q2, median	27.2	Maximum	54.3
Q3	30.6	Range: maximum−minimum	37.7
		IQR=Q3−Q1	6.2

ordinal categories based on US standards[5]: <18.5 underweight, [18.5, 25.0) normal weight, [25.0, 30.0) overweight, [30.0, 35.0) class I obesity, [35.0, 40.0] class II obesity, and [≥40.0) class III obesity. The notation [a,b) denotes a set of values x, such that a≤x<b. The frequency distribution of BMI categories is shown in Table 1. Note that few participants were underweight: only 13 of 3480 (0.4%). Another 1044 (30.0%) were normal weight, 147 (41.6%) were overweight, and 678 (19.5%) had class I obesity, with 192 (5.5%) and 106 (3.1%) classified as having class II or III obesity, respectively. Referring to cumulative statistics, one sees that 30.4% were underweight or normal weight but that 28.0% (ie, 100%−72.0%) were obese. When showing simple descriptive frequency statistics, displaying numbers to 1 decimal place usually provides sufficient detail; because showing excessive decimal places is distracting, rounding to integers may be preferred, especially for values 10% or higher.

Descriptive statistics for continuous variables fall into 3 general classes, namely: *location statistics* (eg, *mean, median, mode, quantiles*), *dispersion statistics* (eg, *variance, standard deviation, range, interquartile range*), and *shape statistics* (eg, *skewness, kurtosis*). The *mean* is the simple arithmetic average of all values. The sample mean is represented by the symbol $\bar{x}$ (read "x bar"), and it is calculated as $\bar{x}=\Sigma x/N$, where x represents the value of an individual observation, Σ represents summation over all observations, and N is the number of observations with nonmissing values. The *median* is defined as the middle value among the ordered values, such that half of the ordered values are below and half are above the median. Formally, when N is an odd number, the median is the middle ordered value, at position $(N+1)/2$, but when N is an even number, the median is calculated as the average of the 2 middle ordered values, at positions $N/2$ and $(N+2)/2$. The third location statistic, the *mode*, is defined as the most frequent value or values in the data; for example, the mode of {1, 1, 2, 3, 3, 3, 4} is 3, but it is not necessarily unique: the values {1, 1, 1, 3, 3, 3, 4} have 2 modes, 1 and 3. To illustrate these statistics, consider the continuous BMI data from the Framingham sample: the mean is 27.9 kg/m^2, the median is 27.2 kg/m^2, and the mode is 26.4 kg/m^2 (see Table 2). One should not report data with more significant digits than is required by the specific research purpose or justified by the sample size and measurement accuracy.

The mean, median, and mode are called *measures of central tendency*; ie, they provide information about the center of a distribution of values. The median is not strongly

affected by outliers or by extreme changes to a small portion of the observations—it is *robust*. However, the mean is sensitive (not robust) to those conditions. The mode is robust to outliers, but it may be affected by data collection operations, such as rounding or digit preference, that alter data precision.

Additional location statistics called *quantiles* combine aspects of ordered data and cumulative frequencies. The *p*-th quantile (0≤p≤1) is defined such that a proportion p of the ranked data values are below x(p) and a proportion (1−p) are above x(p). When 100p is an integer, the quantiles are called *percentiles*. Thus, the median, or 0.50 quantile, is the 50th percentile, the 0.99 quantile is the 99th percentile, and so forth. Three specific percentiles are widely used in descriptive statistics, namely, when 100p is an integer multiple of 25; these percentiles are called *quartiles* and given special labels: Q1=first quartile (25th percentile, 0.25 quantile); Q2=second quartile (50th percentile, 0.50 quantile), also known as the median; and Q3=third quartile (75th percentile, 0.75 quantile). In the BMI data from the Framingham sample, the quartiles are Q1=24.4, Q2=27.2, and Q3=30.6; the 5th and 10th percentiles are 21.0 and 22.1, whereas the 90th and 95th percentiles are 34.4 and 37.6, respectively. Various software packages use slightly different algorithms for estimating quantiles and may produce different results for any specific data set.

In contrast with location statistics, *dispersion statistics* provide information about the variability of the data about the measures of central tendency. One simple dispersion statistic, the *range*, is the difference between the maximum and minimum observed values: that is, range(x)=maximum(x)−minimum(x). Note that the range is a single value. In the Framingham BMI data, minimum and maximum values are 16.6 and 54.3, so the range of BMI is 37.7 kg/m^2 (see Table 2). Dispersion about the mean typically is quantified by the *variance* or the *standard deviation*. The variance is defined as the average of squared deviations from the mean. Therefore, it cannot be negative. For a sample of data, the variance is represented by the symbol s^2, and it is defined by the formula $s^2=\Sigma(x-\bar{x})^2/(N-1)$. Its units are the square of the original units of x. To obtain a dispersion statistic with the same units as x, one uses the standard deviation, defined as the square root of the variance. Thus, the sample standard deviation is $s=\sqrt{s^2}$. The standard deviation may be regarded as the average deviation from the mean. If all observed values are similar, the standard deviation (and variance) will be lower than if the values are more spread out. In the Framingham
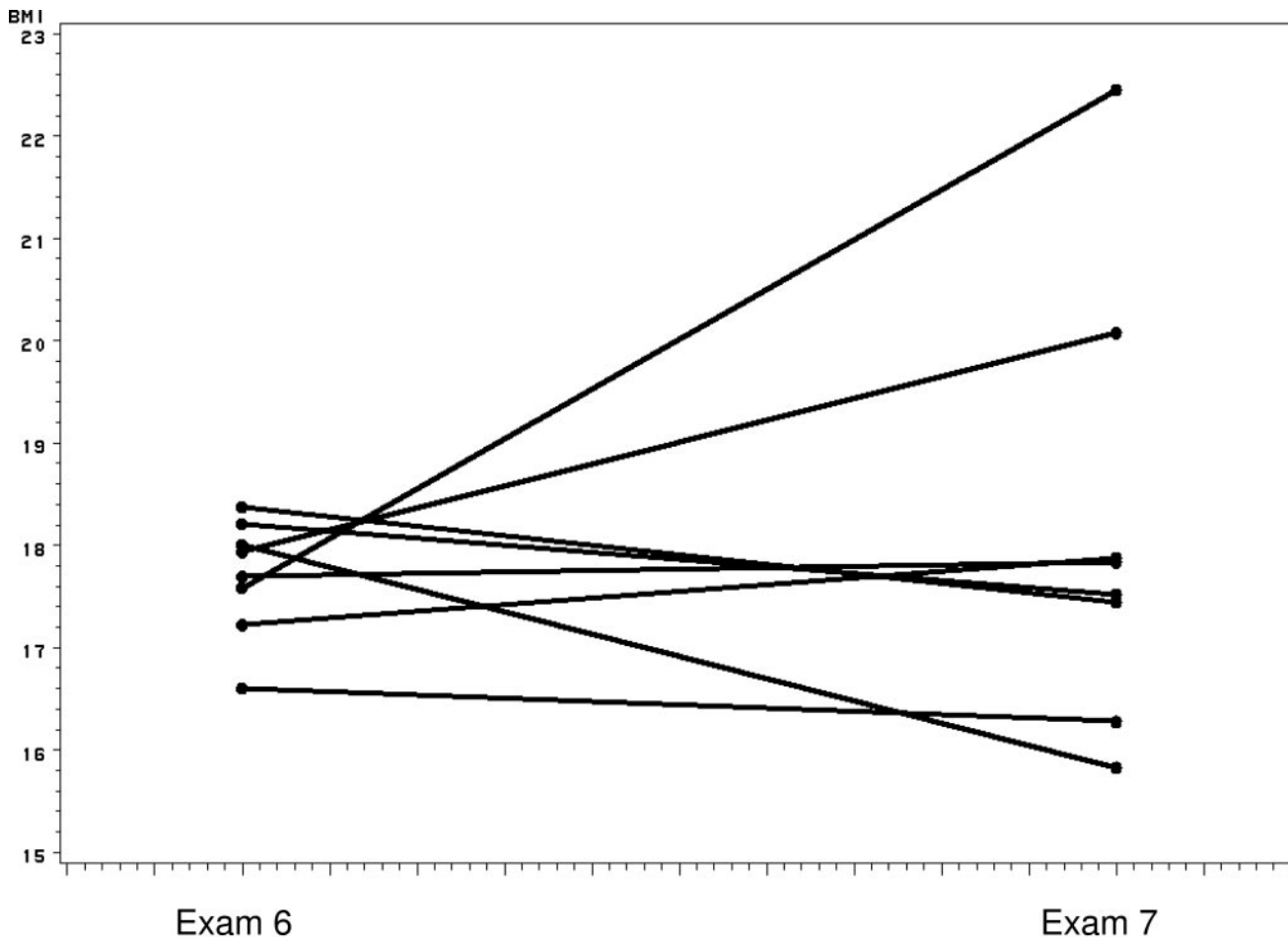
Figure 1. Connected dot plot for BMI across exams. Data are displayed for 8 underweight women at the Framingham Offspring sixth examination and corresponding BMI values at the seventh examination roughly 3 years later.

BMI data (Table 2), the variance is 26.5 $(kg/m^2)^2$ and the standard deviation is 5.1 kg/m^2.

The *interquartile range*, abbreviated IQR, is another commonly used dispersion statistic. It is a single number, defined as IQR=Q3−Q1. Whereas variance and standard deviation are affected (increased) by the presence of extreme observations, the IQR is not; it is robust. The interquartile range for the BMI data is 6.2 kg/m^2 (Table 2).

One general rule of thumb classifies values as *mild outliers* if $x<$Q1−1.5*IQR or $x>$Q3+1.5*IQR, but as *extreme outliers* if $x<$Q1−3*IQR or $x>$Q3+3*IQR. With this rule for the BMI data, mild outliers would be <15.1 or >39.9, and extreme outliers would be <5.8 or >49.2 kg/m^2; there were no BMI outliers with low values, but on the upper end, there were 97 (2.8%) mild outliers and 10 (0.3%) extreme outliers. It is important to note that the term "outlier" does not mean "incorrect," and it does not imply measurement or recording/data entry error (although these should be checked to verify correctness of extreme outliers). Validated outlier observations should be retained in analyses, although secondary analyses to assess sensitivity of major results to outliers may be conducted without them.

Two commonly used shape statistics are *skewness* and *kurtosis*. The skewness coefficient for a sample of data indicates whether the data distribution is symmetric (skewness=0) or has a more pronounced tail in 1 direction than the other (left tail, skewness<0; right tail, skewness>0). For data with skewness=0, the mean and median are equal, but a right- (left-) skewed distribution has its mean value greater (less than) the median. Kurtosis is a measure of the "peakedness" of a distribution. A gaussian distribution (also called "normal") with a bell-shaped frequency curve has kurtosis=0.[1,4] Positive kurtosis indicates a sharper peak with longer/fatter tails and relatively more variability due to extreme deviations; in contrast, a negative kurtosis coefficient indicates broader shoulders with shorter/thinner tails. In the BMI data (Table 2), the skewness coefficient of 1.08 tells us that the distribution is not symmetric: It is skewed to the right, and the kurtosis of 2.03 tells us that the distribution has a sharp peak and long tails.

If there are no outliers and especially if the distribution is symmetric, the mean and standard deviation are excellent measures of location and dispersion, whereas the median and interquartile range may be more appropriate if outliers or strong skewness is present. Yet, there is no hard and fast rule. When estimating length of stay or costs associated with a medical condition, skewed data and outliers are common, but
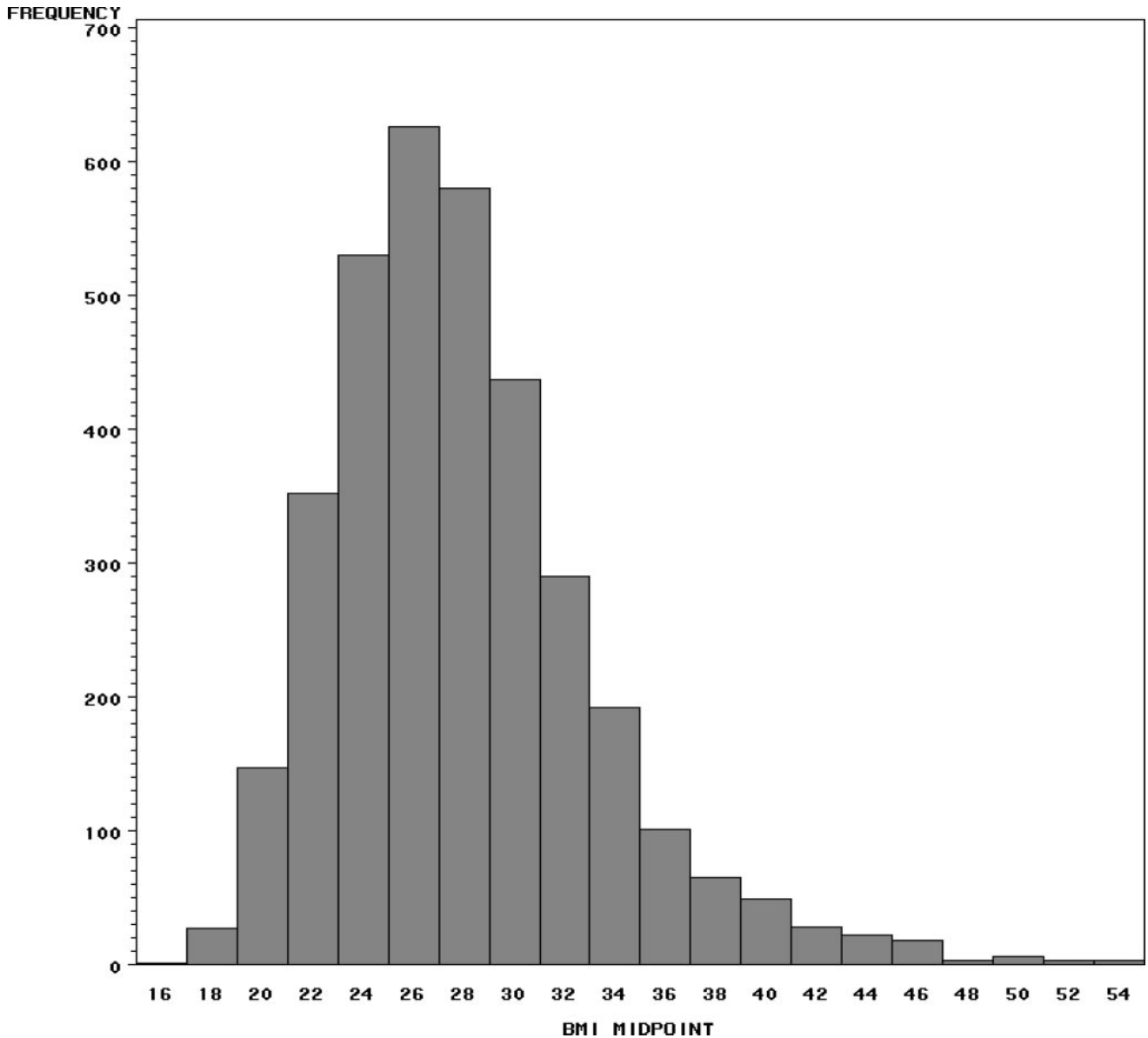
Figure 2. Histogram for BMI (kg/m^2). Data are displayed for 3480 participants at the Framingham Offspring sixth examination.

means are more appropriate than the medians for planning and administrative purposes. Also, in large samples, outliers are likely to occur, even in gaussian distributions, simply by chance; but if the distribution is reasonably smooth and symmetric, without large gaps between ordered values, the mean and standard deviation are appropriate. On the other hand, with strong skewness and if the ratio of maximum/minimum is high, say >10, as seen for measurements of many biological markers, then median and quartiles are preferred, or data transformation, such as taking logarithms, might be applied before descriptive and inferential statistical analyses.

Graphs

Graphical displays complement tabular presentations of descriptive statistics. Generally, graphs are better suited than

tables for identifying patterns in the data, whereas tables are better for providing large amounts of data with a high degree of numerical detail. The *dot plot* is a simple graph that is used mainly with small data sets to show individual values of sample data in 1 dimension. Individual data values are plotted along an axis (usually vertically) and location statistics may be added by using bars or special symbols. A variant called the *connected dot plot* may be used to display the time course of data for individuals when the same variable is measured repeatedly on each subject. The connected dot plot in Figure 1 shows BMI at exams 6 and 7 (roughly 3 years later) for 8 women who were classified as underweight at examination 6. This plot demonstrates heterogeneity of BMI measurements over time: Substantial BMI changes (therefore, weight changes) occurred for 3 women between exams 6 and 7, with 2 gaining and 1 losing weight, but the others had modest
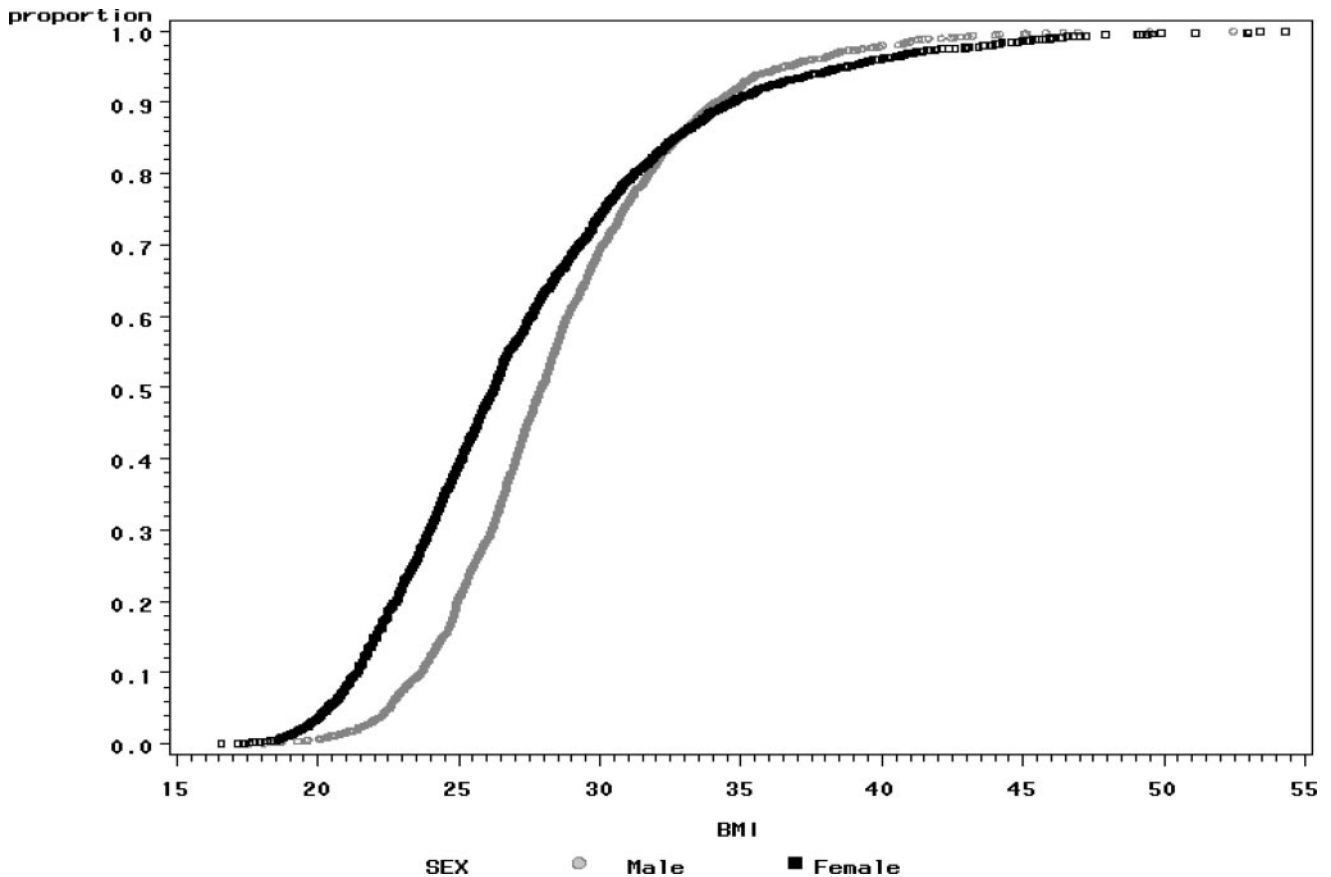
Figure 3. Cumulative frequency plots for BMI (kg/m^2) by sex (1637 men, 1843 women) among participants at the Framingham Offspring sixth examination.

weight changes between exams. Note that dot plots work best when the number of observations is small, such that individual values are discernible. Also, dots should be connected only if multiple observations have been made on the same subjects, objects, or process.

A *histogram* is a plot of the data distribution that shows the number, or relative frequency, of observations that fall into various disjoint categories (called bins) separated by fixed intervals. The choices of number of bins and interval widths are arbitrary. Many statistical programs have defaults of ≈ 10 bins and define the interval width roughly by (range/number of bins). The histogram in Figure 2 shows the distribution of BMI for 3480 participants at Framingham Offspring examination 6 in intervals of width of 2.0 kg/m^2 (lower limit 15, upper limit 55). The distribution is neither bell shaped nor symmetric but is peaked and has a long right tail (positive skewness and kurtosis). About half of the BMI values are between 24 and 31 (recall Q1=24.4 and Q3=30.6 kg/m^2), the median is 27.2, and the mean is 27.9 kg/m^2 (because of right skewness, mean>median). A small number of individuals have extreme values with BMI >50 kg/m^2.

As an alternative to the histogram, one might use a *cumulative frequency plot* to display the distribution of data values. Cumulated frequency (as a proportion) is plotted against ordered values of a variable. Sample quantiles (eg, Q1, Q2, and Q3) can be read directly from the graph as the

value at which the cumulative frequency attains a specified quantile. The cumulative frequency plot in Figure 3 shows BMI separately for men and women. Clearly, the BMI distribution is shifted to the right for men relative to women, such that most quantiles are higher in men than in women (eg, Q1=23.4 in women but 25.5 in men and Q3=30.2 in women but 30.9 in men), except toward the upper end of ordered values (eg, 90th percentiles=34.7 in women compared with 34.1 in men).

Another commonly used graph is called a *box-and-whisker plot* (also known simply as boxplot). This type of graph displays values of quartiles (Q1, Q2, Q3) by a rectangular box. The ends of the box correspond to Q1 and Q3, such that the length of the box is the interquartile range (IQR=Q3−Q1). There is a line drawn inside the box at the median, Q2, and there is a "+" symbol plotted at the mean. Traditionally, "whiskers" (thin lines) extend out to, at most, 1.5 times the box length from both ends of the box: they connect all values outside the box that are not >1.5 IQR away from the box, and they must end at an observed value. Beyond the whiskers are outliers, identified individually by symbols such as circles or asterisks. Alternative versions of boxplot details appear in various software packages, such as the whiskers extending to the 5th and 95th percentiles.

The boxplot in Figure 4 shows descriptive statistics for left ventricular mass according to BMI categories for women in
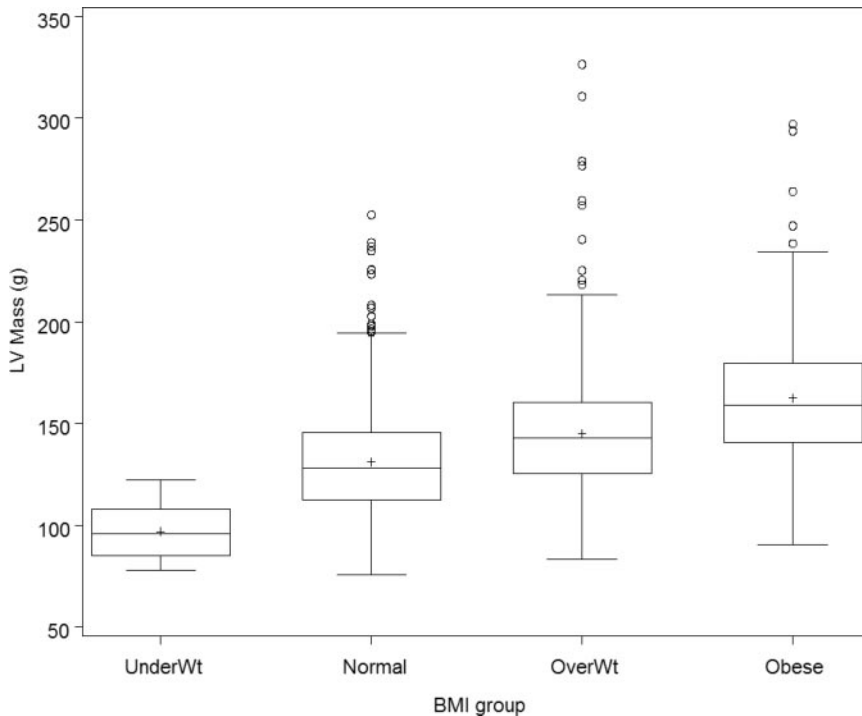
Figure 4. Box-and-whisker plot of left ventricular mass by BMI category in Women. Participants at the Framingham Offspring sixth examination were classified as follows: n=9 underweight (BMI <18.5 kg/m^2), n=712 normal weight (BMI 18.5 to <25.0 kg/m^2), n=645 overweight (BMI 25.0 to <30.0 kg/m^2), and n=477 obese (BMI ≥30.0 kg/m^2).

the Framingham Offspring study at examination 6. Note that there is a clear progression from lower to higher left ventricular mass associated with increasing BMI category, both in medians and means. In addition, the spread of the data values increases across the BMI categories, modestly for IQR but markedly for the whiskers and extreme values.

Concluding Remarks

This introductory article focuses on elementary descriptive statistics and on graphs designed to display simple aspects of research data. Because of space limitations, it does not cover some commonly used graphs (eg, *circle diagrams*, which show counts or proportions for discrete variables, and *bar charts*, often used to show incidence rates or mean values for 2 or more groups). Dot plots or box-and-whisker plots provide more detail about sample values than a bar chart with mean and error bars, and they generally should be preferred to the bar chart. This article also does not cover several graphs that have specific applications (eg, *scatterplots, survival plots, Bland-Altman plots*) that are part of a well-stocked

statistical toolkit, because those will be illustrated in forthcoming articles.

Disclosures

None.

References

1. Zar JH. *Biostatistical Analysis.* Upper Saddle River, NJ: Prentice Hall; 1999.
2. D'Agostino RB Sr, Sullivan LM, Beiser A. *Introductory Applied Biostatistics.* St. Paul, Minn: Minnesota Brooks/Cole; 2003.
3. Wikipedia. Descriptive Statistics. Available at: http://en.wikipedia.org/wiki/Descriptive_statistics. Accessed January 26, 2006.
4. National Institute of Standards and Technology, Information Technology Library. NIST/SEMATECH e-Handbook of Statistical Methods. Available at: http://www.itl.nist.gov/ div898/handbook/ (accessed January 26, 2006) and http://www.itl.nist.gov/div898/ handbook/ eda/ section3/eda35b.htm (accessed January 26, 2006).
5. National Heart, Lung and Blood Institute. Classification of Overweight and Obesity by BMI, Waist Circumference, and Associated Disease Risks. Available at: http://www.nhlbi.nih.gov/ health/public/heart/ obesity/lose_wt/bmi_dis.htm. Accessed January 26, 2006.