PEARSON NEW INTERNATIONAL EDITION

Biostatistical Analysis

Jerrold H. Zar

Fifth Edition

ALWAYS LEARNING

PEARSON

Pearson New International Edition

Biostatistical Analysis

Jerrold H. Zar

Fifth Edition

PEARSON

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsoned.co.uk

© Pearson Education Limited 2014

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without either the prior written permission of the publisher or a licence permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

PEARSON

ISBN 10: 1-292-02404-6
ISBN 13: 978-1-292-02404-2

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Printed in the United States of America

Table of Contents

Data: Types and Presentation

Scientific study involves the systematic collection, organization, analysis, and presentation of knowledge. Many investigations in the biological sciences are quantitative, where knowledge is in the form of numerical observations called *data*. (One numerical observation is a *datum*.*) In order for the presentation and analysis of data to be valid and useful, we must use methods appropriate to the type of data obtained, to the design of the data collection, and to the questions asked of the data; and the limitations of the data, of the data collection, and of the data analysis should be appreciated when formulating conclusions.

The word *statistics* is derived from the Latin for "state," indicating the historical importance of governmental data gathering, which related principally to demographic information (including census data and "vital statistics") and often to their use in military recruitment and tax collecting.†

The term *statistics* is often encountered as a synonym for *data*: One hears of college enrollment statistics (such as the numbers of newly admitted students, numbers of senior students, numbers of students from various geographic locations), statistics of a basketball game (such as how many points were scored by each player, how many fouls were committed), labor statistics (such as numbers of workers unemployed, numbers employed in various occupations), and so on. Hereafter, this use of the word *statistics* will not appear in this text. Instead, it will be used in its other common manner: to refer to the *orderly collection, analysis, and interpretation of data with a view to objective evaluation of conclusions based on the data*.

Statistics applied to biological problems is simply called *biostatistics* or, sometimes, *biometry*‡ (the latter term literally meaning "biological measurement"). Although

*The term *data* is sometimes seen as a singular noun meaning "numerical information." This book refrains from that use.

†Peters (1987: 79) and Walker (1929: 32) attribute the first use of the term *statistics* to a German professor, Gottfried Achenwall (1719–1772), who used the German word *Statistik* in 1749, and the first published use of the English word to John Sinclair (1754–1835) in 1791.

‡The word *biometry*, which literally means "biological measurement," had, since the nineteenth century, been found in several contexts (such as demographics and, later, quantitative genetics; Armitage, 1985; Stigler, 2000), but using it to mean the application of statistical methods to biological information apparently was conceived between 1892 and 1901 by Karl Pearson, along with the name *Biometrika* for the still-important English journal he helped found; and it was first published in the inaugural issue of this journal in 1901 (Snedecor, 1954). The Biometrics Section of the American

From Chapter 1 of *Biostatistical Analysis*, Fifth Edition, Jerrold H. Zar. Copyright © 2010 by Pearson Education, Inc. Publishing as Pearson Prentice Hall. All rights reserved.

the field of statistics has roots extending back hundreds of years, its development began in earnest in the late nineteenth century, and a major impetus from early in this development has been the need to examine biological data.

Statistical considerations can aid in the design of experiments intended to collect data and in the setting up of hypotheses to be tested. Many biologists attempt the analysis of their research data only to find that too few data were collected to enable reliable conclusions to be drawn, or that much extra effort was expended in collecting data that cannot be of ready use in the analysis of the experiment. Thus, a knowledge of basic statistical principles and procedures is important as research questions are formulated *before* an experiment and data collection are begun.

Once data have been obtained, we may organize and summarize them in such a way as to arrive at their orderly and informative presentation. Such procedures are often termed *descriptive statistics*. For example, measurements might be made of the heights of all 13-year-old children in a school district, perhaps determining an average height for each sex. However, perhaps it is desired to make some generalizations from these data. We might, for example, wish to make a reasonable estimate of the heights of all 13-year-olds in the state. Or we might wish to conclude whether the 13-year-old boys in the state are on the average taller than the girls of that age. The ability to make such generalized conclusions, inferring characteristics of the whole from characteristics of its parts, lies within the realm of *inferential statistics*.

1 TYPES OF BIOLOGICAL DATA

A characteristic (for example, size, color, number, chemical composition) that may differ from one biological entity to another is termed a *variable* (or, sometimes, a *variate**), and several different kinds of variables may be encountered by biologists. Because the appropriateness of descriptive or inferential statistical procedures depends upon the properties of the data obtained, it is desirable to distinguish among the principal kinds of data. The classification used here is that which is commonly employed (Senders, 1958; Siegel, 1956; Stevens, 1946, 1968). However, not all data fit neatly into these categories and some data may be treated differently depending upon the questions asked of them.

(a) Data on a Ratio Scale. Imagine that we are studying a group of plants, that the heights of the plants constitute a variable of interest, and that the number of leaves per plant is another variable under study. It is possible to assign a numerical value to the height of each plant, and counting the leaves allows a numerical value to be recorded for the number of leaves on each plant. Regardless of whether the height measurements are recorded in centimeters, inches, or other units, and regardless of whether the leaves are counted in a number system using base 10 or any other base, there are two fundamentally important characteristics of these data.

First, there is a constant size interval between adjacent units on the measurement scale. That is, the difference in height between a 36-cm and a 37-cm plant is the same

---

Statistical Association was established in 1938, successor to the Committee on Biometrics of that organization, and began publishing the *Biometrics Bulletin* in 1945, which transformed in 1947 into the journal *Biometrics*, a journal retaining major importance today. More recently, the term *biometrics* has become widely used to refer to the study of human physical characteristics (including facial and hand characteristics, fingerprints, DNA profiles, and retinal patterns) for identification purposes.

*"Variate" was first used by R. A. Fisher (1925: 5; David, 1995).

as the difference between a 39-cm and a 40-cm plant, and the difference between eight and ten leaves is equal to the difference between nine and eleven leaves.

Second, it is important that there exists a zero point on the measurement scale and that there is a physical significance to this zero. This enables us to say something meaningful about the ratio of measurements. We can say that a 30-cm (11.8-in.) tall plant is half as tall as a 60-cm (23.6-in.) plant, and that a plant with forty-five leaves has three times as many leaves as a plant with fifteen.

Measurement scales having a constant interval size and a true zero point are said to be *ratio scales* of measurement. Besides lengths and numbers of items, ratio scales include weights (mg, lb, etc.), volumes (cc, cu ft, etc.), capacities (ml, qt, etc.), rates (cm/sec, mph, mg/min, etc.), and lengths of time (hr, yr, etc.).

(b) Data on an Interval Scale. Some measurement scales possess a constant interval size but not a true zero; they are called *interval scales*. A common example is that of the two common temperature scales: Celsius (C) and Fahrenheit (F). We can see that the same difference exists between 20°C (68°F) and 25°C (77°F) as between 5°C (41°F) and 10°C (50°F); that is, the measurement scale is composed of equal-sized intervals. But it cannot be said that a temperature of 40°C (104°F) is twice as hot as a temperature of 20°C (68°F); that is, the zero point is arbitrary.* (Temperature measurements on the absolute, or Kelvin [K], scale can be referred to a physically meaningful zero and thus constitute a ratio scale.)

Some interval scales encountered in biological data collection are *circular scales*. Time of day and time of the year are examples of such scales. The interval between 2:00 P.M. (i.e., 1400 hr) and 3:30 P.M. (1530 hr) is the same as the interval between 8:00 A.M. (0800 hr) and 9:30 A.M. (0930 hr). But one cannot speak of ratios of times of day because the zero point (midnight) on the scale is arbitrary, in that one could just as well set up a scale for time of day which would have noon, or 3:00 P.M., or any other time as the zero point. Circular biological data are occasionally compass points, as if one records the compass direction in which an animal or plant is oriented. As the designation of north as 0° is arbitrary, this circular scale is a form of interval scale of measurement.

(c) Data on an Ordinal Scale. The preceding paragraphs on ratio and interval scales of measurement discussed data between which we know numerical differences. For example, if man A weighs 90 kg and man B weighs 80 kg, then man A is known to weigh 10 kg more than B. But our data may, instead, be a record only of the fact that man A weighs more than man B (with no indication of how much more). Thus, we may be dealing with relative differences rather than quantitative differences. Such data consist of an ordering or ranking of measurements and are said to be on an *ordinal* scale of measurement (*ordinal* being from the Latin word for "order"). We may speak of one biological entity being shorter, darker, faster, or more active than another; the sizes of five cell types might be labeled 1, 2, 3, 4, and 5, to denote

*The German-Dutch physicist Gabriel Daniel Fahrenheit (1686–1736) invented the thermometer in 1714 and in 1724 employed a scale on which salt water froze at zero degrees, pure water froze at 32 degrees, and pure water boiled at 212 degrees. In 1742 the Swedish astronomer Anders Celsius (1701–1744) devised a temperature scale with 100 degrees between the freezing and boiling points of water (the so-called "centigrade" scale), first by referring to zero degrees as boiling and 100 degrees as freezing, and later (perhaps at the suggestion of Swedish botanist and taxonomist Carolus Linnaeus [1707–1778]) reversing these two reference points (Asimov, 1982: 177).

their magnitudes relative to each other; or success in learning to run a maze may be recorded as A, B, or C.

It is often true that biological data expressed on the ordinal scale could have been expressed on the interval or ratio scale had exact measurements been obtained (or obtainable). Sometimes data that were originally on interval or ratio scales will be changed to ranks; for example, examination grades of 99, 85, 73, and 66% (ratio scale) might be recorded as A, B, C, and D (ordinal scale), respectively.

Ordinal-scale data contain and convey less information than ratio or interval data, for only relative magnitudes are known. Consequently, quantitative comparisons are impossible (e.g., we cannot speak of a grade of C being half as good as a grade of A, or of the difference between cell sizes 1 and 2 being the same as the difference between sizes 3 and 4). However, we will see that many useful statistical procedures are, in fact, applicable to ordinal data.

(d) Data in Nominal Categories. Sometimes the variable being studied is classified by some qualitative measure it possesses rather than by a numerical measurement. In such cases the variable may be called an *attribute*, and we are said to be dealing with *nominal*, or *categorical*, data. Genetic phenotypes are commonly encountered biological attributes: The possible manifestations of an animal's eye color might be brown or blue; and if human hair color were the attribute of interest, we might record black, brown, blond, or red. As other examples of nominal data (*nominal* is from the Latin word for "name"), people might be classified as male or female, or right-handed or left-handed. Or, plants might be classified as dead or alive, or as with or without fertilizer application. Taxonomic categories also form a nominal classification scheme (for example, plants in a study might be classified as pine, spruce, or fir).

Sometimes, data that might have been expressed on an ordinal, interval, or ratio scale of measurement may be recorded in nominal categories. For example, heights might be recorded as tall or short, or performance on an examination as pass or fail, where there is an arbitrary cut-off point on the measurement scale to separate tall from short and pass from fail.

As will be seen, statistical methods useful with ratio, interval, or ordinal data generally are not applicable to nominal data, and we must, therefore, be able to identify such situations when they occur.

(e) Continuous and Discrete Data. When we spoke previously of plant heights, we were dealing with a variable that could be any conceivable value within any observed range; this is referred to as a *continuous variable*. That is, if we measure a height of 35 cm and a height of 36 cm, an infinite number of heights is possible in the range from 35 to 36 cm: a plant might be 35.07 cm tall or 35.988 cm tall, or 35.3263 cm tall, and so on, although, of course, we do not have devices sensitive enough to detect this infinity of heights. A continuous variable is one for which there is a possible value between any other two values.

However, when speaking of the number of leaves on a plant, we are dealing with a variable that can take on only certain values. It might be possible to observe 27 leaves, or 28 leaves, but 27.43 leaves and 27.9 leaves are values of the variable that are impossible to obtain. Such a variable is termed a *discrete* or *discontinuous variable* (also known as a *meristic variable*). The number of white blood cells in 1 mm^3 of blood, the number of giraffes visiting a water hole, and the number of eggs laid by a grasshopper are all discrete variables. The possible values of a discrete variable generally are consecutive integers, but this is not necessarily so. If the leaves on our plants are always formed in pairs, then only even integers are possible values of the variable. And the ratio of number of wings to number of legs of insects is a discrete variable that may only have the value of 0, 0.3333 . . . , or 0.6666 . . . (i.e., $\frac{0}{6}$, $\frac{2}{6}$, or $\frac{4}{6}$, respectively).*

Ratio-, interval-, and ordinal-scale data may be either continuous or discrete. Nominal-scale data by their nature are discrete.

2 ACCURACY AND SIGNIFICANT FIGURES

Accuracy is the nearness of a measurement to the true value of the variable being measured. *Precision* is not a synonymous term but refers to the closeness to each other of repeated measurements of the same quantity. Figure 1 illustrates the difference between accuracy and precision of measurements.

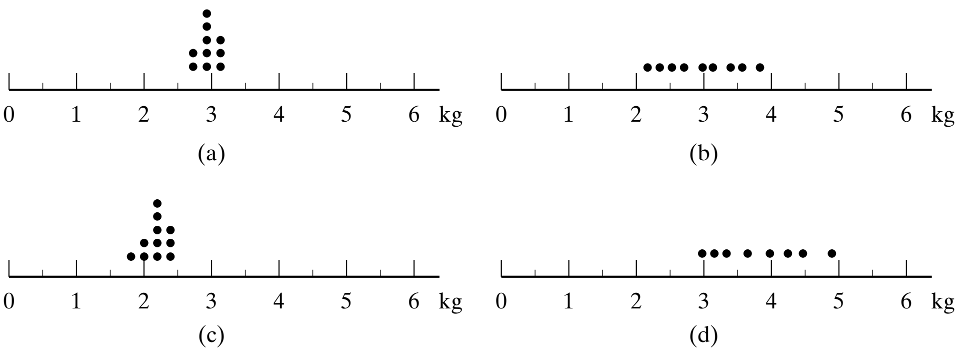

FIGURE 1: Accuracy and precision of measurements. A 3-kilogram animal is weighed 10 times. The 10 measurements shown in sample (a) are relatively accurate and precise; those in sample (b) are relatively accurate but not precise; those of sample (c) are relatively precise but not accurate; and those of sample (d) are relatively inaccurate and imprecise.

Human error may exist in the recording of data. For example, a person may miscount the number of birds in a tract of land or misread the numbers on a heart-rate monitor. Or, a person might obtain correct data but record them in such a way (perhaps with poor handwriting) that a subsequent data analyst makes an error in reading them. We shall assume that such errors have not occurred, but there are other aspects of accuracy that should be considered.

Accuracy of measurement can be expressed in numerical reporting. If we report that the hind leg of a frog is 8 cm long, we are stating the number 8 (a value of a continuous variable) as an estimate of the frog's true leg length. This estimate was made using some sort of a measuring device. Had the device been capable of more accuracy, we might have declared that the leg was 8.3 cm long, or perhaps 8.32 cm long. When recording values of continuous variables, it is important to designate the accuracy with which the measurements have been made. By convention, the value 8 denotes a measurement in the range of 7.50000 . . . to 8.49999 . . . , the value 8.3 designates a range of 8.25000 . . . to 8.34999 . . . , and the value 8.32 implies that the true value lies within the range of 8.31500 . . . to 8.32499 . . . . That is, the reported value is the midpoint of the implied range, and the size of this range is designated by the last decimal place in the measurement. The value of 8 cm implies an ability to

*The ellipsis marks (. . .) may be read as "and so on." Here, they indicate that $\frac{2}{6}$ and $\frac{4}{6}$ are repeating decimal fractions, which could just as well have been written as 0.3333333333333 . . . and 0.6666666666666 . . . , respectively.

determine length within a range of 1 cm, 8.3 cm implies a range of 0.1 cm, and 8.32 cm implies a range of 0.01 cm. Thus, to record a value of 8.0 implies greater accuracy of measurement than does the recording of a value of 8, for in the first instance the true value is said to lie between 7.95000 ... and 8.049999... (i.e., within a range of 0.1 cm), whereas 8 implies a value between 7.50000 ... and 8.49999... (i.e., within a range of 1 cm). To state 8.00 cm implies a measurement that ascertains the frog's limb length to be between 7.99500 ... and 8.00499 ... cm (i.e., within a range of 0.01 cm). Those digits in a number that denote the accuracy of the measurement are referred to as *significant figures*. Thus, 8 has one significant figure, 8.0 and 8.3 each have two significant figures, and 8.00 and 8.32 each have three.

In working with exact values of discrete variables, the preceding considerations do not apply. That is, it is sufficient to state that our frog has four limbs or that its left lung contains thirteen flukes. The use of 4.0 or 13.00 would be inappropriate, for as the numbers involved are exactly 4 and 13, there is no question of accuracy or significant figures.

But there are instances where significant figures and implied accuracy come into play with discrete data. An entomologist may report that there are 72,000 moths in a particular forest area. In doing so, it is probably not being claimed that this is the exact number but an estimate of the exact number, perhaps accurate to two significant figures. In such a case, 72,000 would imply a range of accuracy of 1000, so that the true value might lie anywhere from 71,500 to 72,500. If the entomologist wished to convey the fact that this estimate is believed to be accurate to the nearest 100 (i.e., to three significant figures), rather than to the nearest 1000, it would be better to present the data in the form of *scientific notation*,* as follows: If the number $7.2 \times 10^4 (= 72{,}000)$ is written, a range of accuracy of $0.1 \times 10^4 (= 1000)$ is implied, and the true value is assumed to lie between 71,500 and 72,500. But if 7.20×10^4 were written, a range of accuracy of $0.01 \times 10^4 (= 100)$ would be implied, and the true value would be assumed to be in the range of 71,950 to 72,050. Thus, the accuracy of large values (and this applies to continuous as well as discrete variables) can be expressed succinctly using scientific notation.

Calculators and computers typically yield results with more significant figures than are justified by the data. However, it is good practice—to avoid rounding error—to retain many significant figures until the last step in a sequence of calculations, and on attaining the result of the final step to round off to the appropriate number of figures.

3 FREQUENCY DISTRIBUTIONS

When collecting and summarizing large amounts of data, it is often helpful to record the data in the form of a *frequency table*. Such a table simply involves a listing of all the observed values of the variable being studied and how many times each value is observed. Consider the tabulation of the frequency of occurrence of sparrow nests in each of several different locations. This is illustrated in Example 1, where the observed kinds of nest sites are listed, and for each kind the number of nests observed is recorded. The distribution of the total number of observations among the various categories is termed a *frequency distribution*. Example 1 is a frequency table for nominal data, and these data may also be presented graphically by means of a *bar graph* (Figure 2), where the height of each bar is proportional to the frequency in the class represented. The widths of all bars in a bar graph should be equal so

*The use of scientific notation—by physicists—can be traced back to at least the 1860s (Miller, 2004b).

EXAMPLE 1 The Location of Sparrow Nests: A Frequency Table of Nominal Data

The variable is nest site, and there are four recorded categories of this variable. The numbers recorded in these categories constitute the frequency distribution.

Nest Site	*Number of Nests Observed*
A. Vines	56
B. Building eaves	60
C. Low tree branches	46
D. Tree and building cavities	49

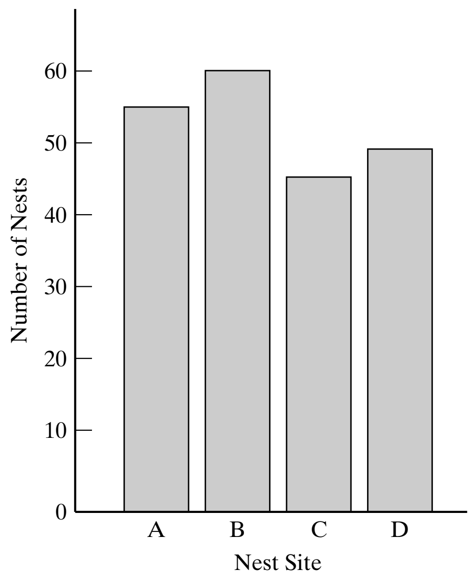

FIGURE 2: A bar graph of the sparrow nest data of Example 1. An example of a bar graph for nominal data.

that the eye of the reader is not distracted from the differences in bar heights; this also makes the area of each bar proportional to the frequency it represents. Also, the frequency scale on the vertical axis should begin at zero to avoid the apparent differences among bars. If, for example, a bar graph of the data of Example 1 were constructed with the vertical axis representing frequencies of 45 to 60 rather than 0 to 60, the results would appear as in Figure 3. Huff (1954) illustrates other techniques that can mislead the readers of graphs. It is good practice to leave space between the bars of a bar graph of nominal data, to emphasize the distinctness among the categories represented.

A frequency tabulation of ordinal data might appear as in Example 2, which presents the observed numbers of sunfish collected in each of five categories, each category being a degree of skin pigmentation. A bar graph (Figure 4) can be prepared for this frequency distribution just as for nominal data.

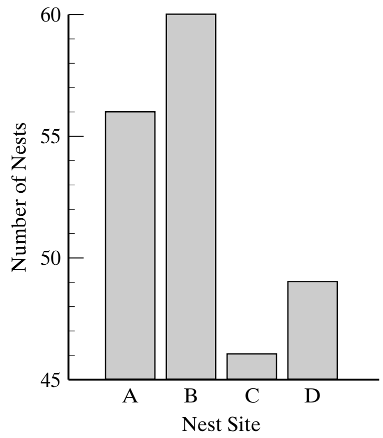

FIGURE 3: A bar graph of the sparrow nest data of Example 1, drawn with the vertical axis starting at 45. Compare this with Figure 1, where the axis starts at 0.

EXAMPLE 2 Numbers of Sunfish, Tabulated According to Amount of Black Pigmentation: A Frequency Table of Ordinal Data

The variable is amount of pigmentation, which is expressed by numerically ordered classes. The numbers recorded for the five pigmentation classes compose the frequency distribution.

Pigmentation Class	*Amount of Pigmentation*	*Number of Fish*
0	No black pigmentation	13
1	Faintly speckled	68
2	Moderately speckled	44
3	Heavily speckled	21
4	Solid black pigmentation	8

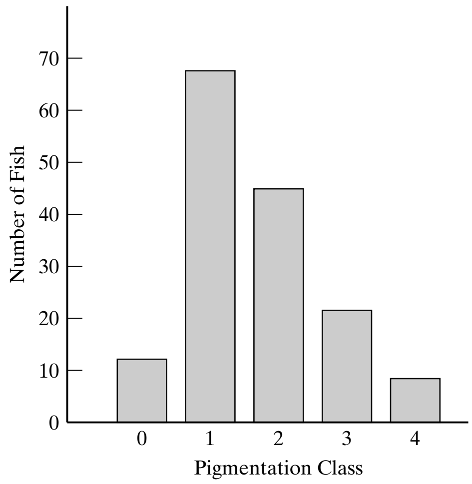

FIGURE 4: A bar graph of the sunfish pigmentation data of Example 2. An example of a bar graph for ordinal data.

In preparing frequency tables of interval- and ratio-scale data, we can make a procedural distinction between discrete and continuous data. Example 3 shows discrete data that are frequencies of litter sizes in foxes, and Figure 5 presents this frequency distribution graphically.

EXAMPLE 3 Frequency of Occurrence of Various Litter Sizes in Foxes: A Frequency Table of Discrete, Ratio-Scale Data

The variable is litter size, and the numbers recorded for the five litter sizes make up frequency distribution.

Litter Size	*Frequency*
3	10
4	27
5	22
6	4
7	1

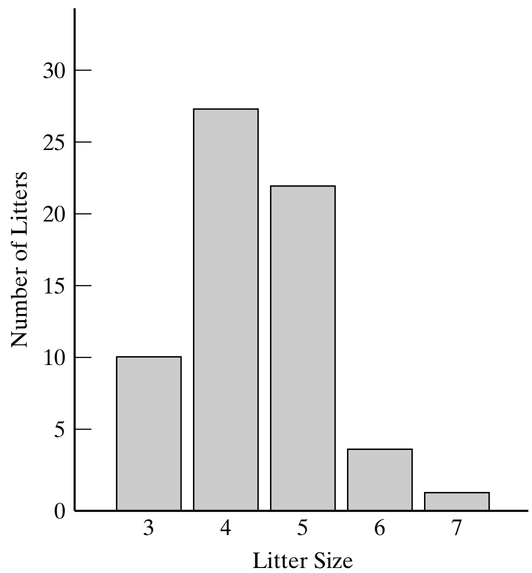

FIGURE 5: A bar graph of the fox litter data of Example 3. An example of a bar graph for discrete, ratio-scale data.

Example 4a shows discrete data that are the numbers of aphids found per clover plant. These data create quite a lengthy frequency table, and it is not difficult to imagine sets of data whose tabulation would result in an even longer list of frequencies. Thus, for purposes of preparing bar graphs, we often cast data into a frequency table by grouping them.

Example 4b is a table of the data from Example 4a arranged by grouping the data into size classes. The bar graph for this distribution appears as Figure 6. Such grouping results in the loss of some information and is generally utilized only to make frequency tables and bar graphs easier to read, and not for calculations performed on the data. There have been several "rules of thumb" proposed to aid in deciding into how many classes data might reasonably be grouped, for the use of too few groups will obscure the general shape of the distribution. But such "rules" or recommendations are only rough guides, and the choice is generally left to good judgment, bearing in mind that from 10 to 20 groups are useful for most biological work. (See also Doane, 1976.) In general, groups should be established that are equal in the size interval of the variable being measured. (For example, the group size interval in Example 4b is four aphids per plant.)

EXAMPLE 4a Number of Aphids Observed per Clover Plant: A Frequency Table of Discrete, Ratio-Scale Data

Number of Aphids on a Plant	*Number of Plants Observed*	*Number of Aphids on a Plant*	*Number of Plants Observed*
0	3	20	17
1	1	21	18
2	1	22	23
3	1	23	17
4	2	24	19
5	3	25	18
6	5	26	19
7	7	27	21
8	8	28	18
9	11	29	13
10	10	30	10
11	11	31	14
12	13	32	9
13	12	33	10
14	16	34	8
15	13	35	5
16	14	36	4
17	16	37	1
18	15	38	2
19	14	39	1
		40	0
		41	1

Total number of observations = 424

Because continuous data, contrary to discrete data, can take on an infinity of values, one is essentially always dealing with a frequency distribution tabulated by groups. If the variable of interest were a weight, measured to the nearest 0.1 mg, a frequency table entry of the number of weights measured to be 48.6 mg would be interpreted to mean the number of weights grouped between 48.5500 . . . and 48.6499 . . . mg (although in a frequency table this class interval is usually written as 48.55–48.65). Example 5 presents a tabulation of 130 determinations of the amount of phosphorus, in milligrams per gram, in dried leaves. (Ignore the last two columns of this table until Section 4.)

EXAMPLE 4b Number of Aphids Observed per Clover Plant: A Frequency Table Grouping the Discrete, Ratio-Scale Data of Example 4a

Number of Aphids on a Plant	*Number of Plants Observed*
0–3	6
4–7	17
8–11	40
12–15	54
16–19	59
20–23	75
24–27	77
28–31	55
32–35	32
36–39	8
40–43	1

Total number of observations = 424

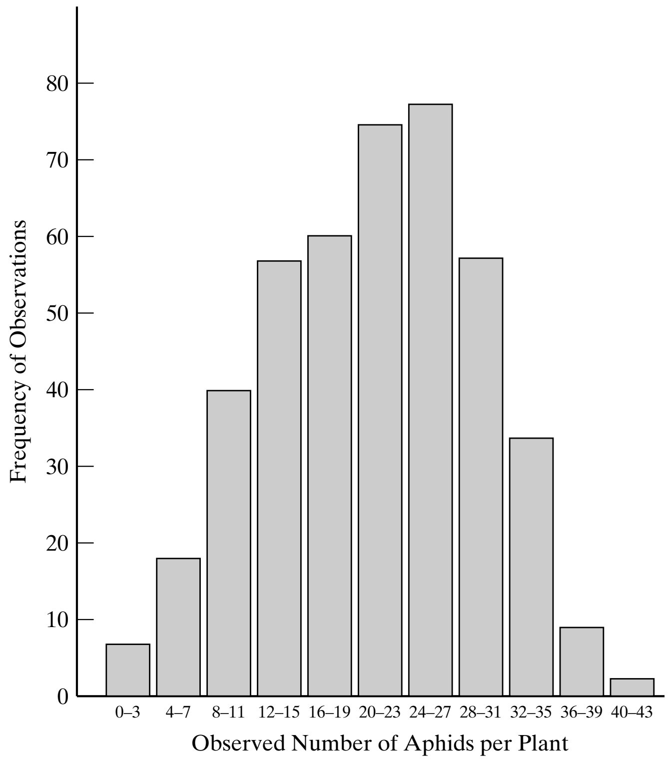

FIGURE 6: A bar graph of the aphid data of Example 4b. An example of a bar graph for grouped discrete, ratio-scale data.

EXAMPLE 5 Determinations of the Amount of Phosphorus in Leaves: A Frequency Table of Continuous Data

		Cumulative frequency	
Phosphorus (mg/g of leaf)	*Frequency (i.e., number of determinations)*	*Starting with Low Values*	*Starting with High Values*
8.15–8.25	2	2	130
8.25–8.35	6	8	128
8.35–8.45	8	16	122
8.45–8.55	11	27	114
8.55–8.65	17	44	103
8.65–8.75	17	61	86
8.75–8.85	24	85	69
8.85–8.95	18	103	45
8.95–9.05	13	116	27
9.05–9.15	10	126	14
9.15–9.25	4	130	4

Total frequency $= 130 = n$

In presenting this frequency distribution graphically, one can prepare a *histogram*,* which is the name given to a bar graph based on continuous data. This is done in Figure 7; note that rather than indicating the range on the horizontal axis, we indicate only the midpoint of the range, a procedure that results in less crowded printing on the graph. Note also that adjacent bars in a histogram are often drawn touching each other, to emphasize the continuity of the scale of measurement, whereas in the other bar graphs discussed they generally are not.

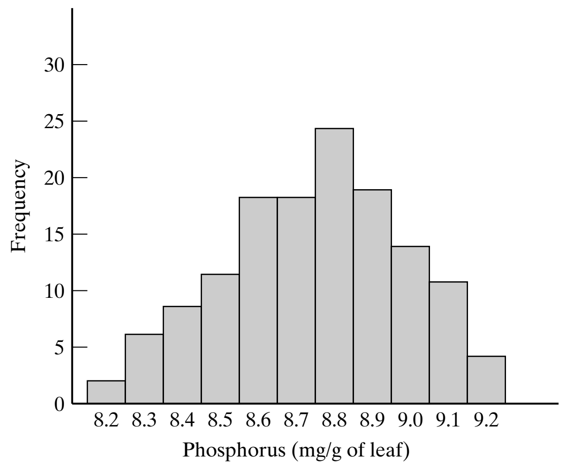

FIGURE 7: A histogram of the leaf phosphorus data of Example 5. An example of a histogram for continuous data.

*The term *histogram* is from Greek roots (referring to a pole-shaped drawing) and was first published by Karl Pearson in 1895 (David 1995).

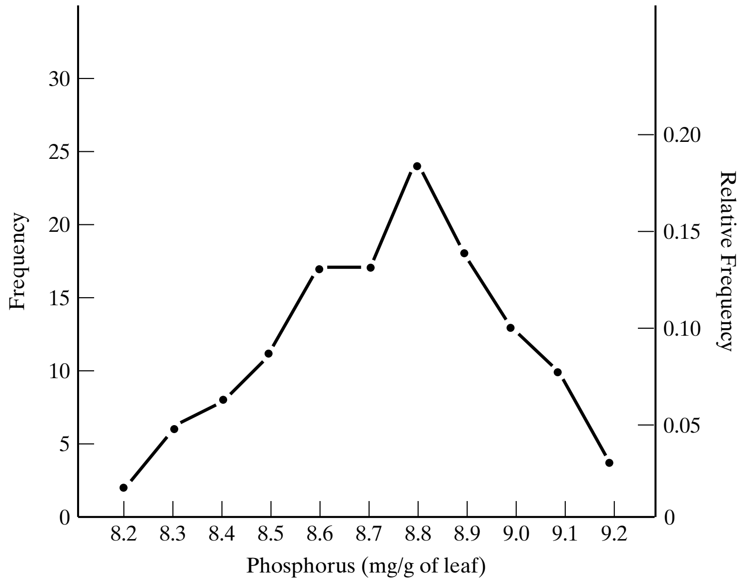

FIGURE 8: A frequency polygon for the leaf phosphorus data of Example 5.

Often a *frequency polygon* is drawn instead of a histogram. This is done by plotting the frequency of each class as a dot (or other symbol) at the class midpoint and then connecting each adjacent pair of dots by a straight line (Figure 8). It is, of course, the same as if the midpoints of the tops of the histogram bars were connected by straight lines. Instead of plotting frequencies on the vertical axis, one can plot *relative frequencies*, or proportions of the total frequency. This enables different distributions to be readily compared and even plotted on the same axes. Sometimes, as in Figure 8, frequency is indicated on one vertical axis and the corresponding relative frequency on the other. (Using the data of Example 5, the relative frequency for 8.2 mg/g is $2/130 = 0.015$, that for 8.3 mg/g is $6/130 = 0.046$, that for 9.2 mg/g is $4/130 = 0.030$, and so on. The total of all the frequencies is n, and the total of all the relative frequencies is 1.)

Frequency polygons are also commonly used for discrete distributions, but one can argue against their use when dealing with ordinal data, as the polygon implies to the reader a constant size interval horizontally between points on the polygon. Frequency polygons should not be employed for nominal-scale data.

If we have a frequency distribution of values of a continuous variable that falls into a large number of class intervals, the data may be grouped as was demonstrated with discrete variables. This results in fewer intervals, but each interval is, of course, larger. The midpoints of these intervals may then be used in the preparation of a histogram or frequency polygon. The user of frequency polygons is cautioned that such a graph is simply an aid to the eye in following trends in frequency distributions, and one should not attempt to read frequencies between points on the polygon. Also note that the method presented for the construction of histograms and frequency polygons requires that the class intervals be equal. Lastly, the vertical axis (e.g., the frequency scale) on frequency polygons and bar graphs generally should begin with zero, especially if graphs are to be compared with one another. If this is not done, the eye may be misled by the appearance of the graph (as shown for nominal-scale data in Figures 2 and 3).

4 CUMULATIVE FREQUENCY DISTRIBUTIONS

A frequency distribution informs us how many observations occurred for each value (or group of values) of a variable. That is, examination of the frequency table of Example 3 (or its corresponding bar graph or frequency polygon) would yield information such as, “How many fox litters of four were observed?”, the answer being 27. But if it is desired to ask questions such as, “How many litters of four or more were observed?”, or “How many fox litters of five or fewer were observed?”, we are speaking of *cumulative frequencies*. To answer the first question, we sum all frequencies for litter sizes four and up, and for the second question, we sum all frequencies from the smallest litter size up through a size of five. We arrive at answers of 54 and 59, respectively.

In Example 5, the phosphorus concentration data are cast into two cumulative frequency distributions, one with cumulation commencing at the low end of the measurement scale and one with cumulation being performed from the high values toward the low values. The choice of the direction of cumulation is immaterial, as can be demonstrated. If one desired to calculate the number of phosphorus determinations less than 8.55 mg/g, namely 27, a cumulation starting at the low end might be used, whereas the knowledge of the frequency of determinations greater than 8.55 mg/g, namely 103, can be readily obtained from the cumulation commencing from the high end of the scale. But one can easily calculate any frequency from a low-to-high cumulation (e.g., 27) from its complementary frequency from a high-to-low cumulation (e.g., 103), simply by knowing that the sum of these two frequencies is the total frequency (i.e., $n = 130$); therefore, in practice it is not necessary to calculate both sets of cumulations.

Cumulative frequency distributions are useful in determining medians, percentiles, and other quantiles. They are not often presented in bar graphs, but *cumulative frequency polygons* (sometimes called *ogives*) are not uncommon. (See Figures 9 and 10.)

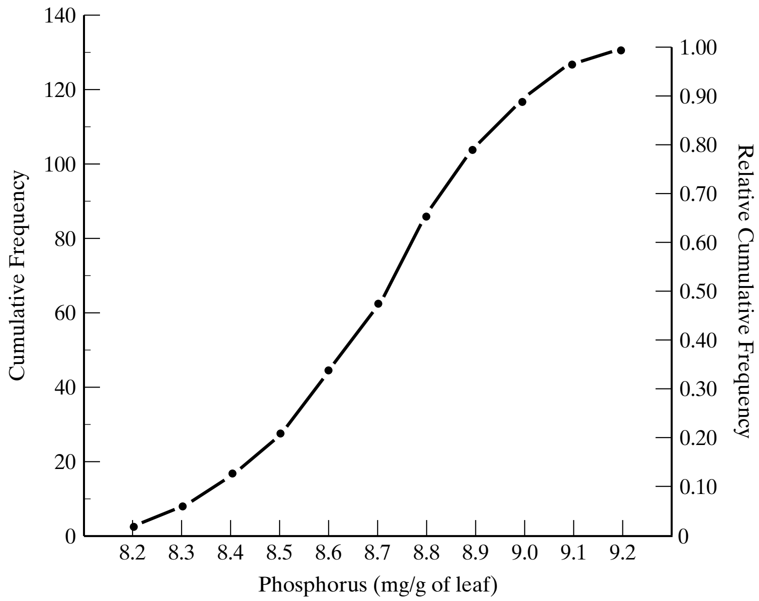

FIGURE 9: Cumulative frequency polygon of the leaf phosphorus data of Example 5, with cumulation commencing from the lowest to the highest values of the variable.

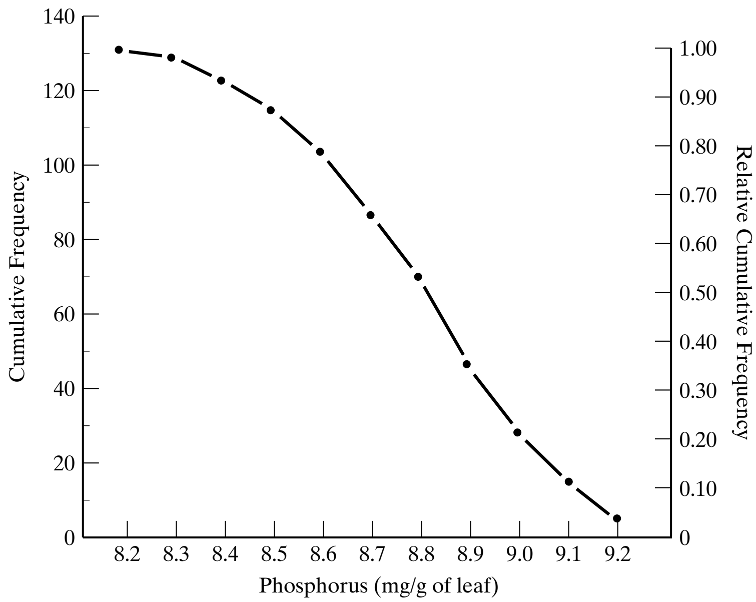

FIGURE 10: Cumulative frequency polygon of the leaf phosphorus data of Example 5, with cumulation commencing from the highest to the lowest values of the variable.

Relative frequencies (proportions of the total frequency) can be plotted instead of (or, as in Figures 9 and 10, in addition to) frequencies on the vertical axis of a cumulative frequency polygon. This enables different distributions to be readily compared and even plotted on the same axes. (Using the data of Example 5 for Figure 9, the relative cumulative frequency for 8.2 mg/g is $2/130 = 0.015$, that for 8.3 mg/g is $8/130 = 0.062$, and so on. For Figure 10, the relative cumulative frequency for 8.2 mg/g is $130/130 = 1.000$, that for 8.3 mg/g is $128/130 = 0.985$, and so on.)

This page intentionally left blank

Populations and Samples

From Chapter 2 of *Biostatistical Analysis*, Fifth Edition, Jerrold H. Zar. Copyright © 2010 by Pearson Education, Inc. Publishing as Pearson Prentice Hall. All rights reserved.

Populations and Samples

The primary objective of a statistical analysis is to infer characteristics of a group of data by analyzing the characteristics of a small sampling of the group. This generalization from the part to the whole requires the consideration of such important concepts as population, sample, parameter, statistic, and random sampling. These topics are discussed in this chapter.

1 POPULATIONS

Basic to statistical analysis is the desire to draw conclusions about a group of measurements of a variable being studied. Biologists often speak of a "population" as a defined group of humans or of another species of organisms. Statisticians speak of a *population* (also called a *universe*) as a group of measurements (not organisms) about which one wishes to draw conclusions. It is the latter definition, the statistical definition of *population*, that will be used throughout this text. For example, an investigator may desire to draw conclusions about the tail lengths of bobcats in Montana. All Montana bobcat tail lengths are, therefore, the population under consideration. If a study is concerned with the blood-glucose concentration in three-year-old children, then the blood-glucose levels in all children of that age are the population of interest.

Populations are often very large, such as the body weights of all grasshoppers in Kansas or the eye colors of all female New Zealanders, but occasionally populations of interest may be relatively small, such as the ages of men who have traveled to the moon or the heights of women who have swum the English Channel.

2 SAMPLES FROM POPULATIONS

If the population under study is very small, it might be practical to obtain all the measurements in the population. If one wishes to draw conclusions about the ages of all men who have traveled to the moon, it would not be unreasonable to attempt to collect all the ages of the small number of individuals under consideration. Generally, however, populations of interest are so large that obtaining all the measurements is unfeasible. For example, we could not reasonably expect to determine the body weight of every grasshopper in Kansas. What can be done in such cases is to obtain a subset of all the measurements in the population. This subset of measurements constitutes a *sample*, and from the characteristics of samples we can draw conclusions about the characteristics of the populations from which the samples came.*

Biologists may sample a population that does not physically exist. Suppose an experiment is performed in which a food supplement is administered to 40 guinea pigs, and the sample data consist of the growth rates of these 40 animals. Then the population about which conclusions might be drawn is the growth rates of all the guinea pigs that conceivably might have been administered the same food supplement under identical conditions. Such a population is said to be “imaginary” and is also referred to as “hypothetical” or “potential.”

3 RANDOM SAMPLING

Samples from populations can be obtained in a number of ways; however, for a sample to be representative of the population from which it came, and to reach valid conclusions about populations by induction from samples, statistical procedures typically assume that the samples are obtained in a *random* fashion. To sample a population randomly requires that each member of the population has an equal and independent chance of being selected. That is, not only must each measurement in the population have an equal chance of being chosen as a member of the sample, but the selection of any member of the population must in no way influence the selection of any other member. Throughout this text, “sample” will always imply “random sample.”†

It is sometimes possible to assign each member of a population a unique number and to draw a sample by choosing a set of such numbers at random. This is equivalent to having all members of a population in a hat and drawing a sample from them while blindfolded. Table 41 from *Appendix: Statistical Tables and Graphs* provides 10,000 random digits for this purpose. In this table, each digit from 0 to 9 has an equal and independent chance of appearing anywhere in the table. Similarly, each combination of two digits, from 00 to 99, is found at random in the table, as is each three-digit combination, from 000 to 999, and so on.

Assume that a random sample of 200 names is desired from a telephone directory having 274 pages, three columns of names per page, and 98 names per column. Entering Table 41 from *Appendix: Statistical Tables and Graphs* at random (i.e., do not always enter the table at the same place), one might decide first to arrive at a random combination of three digits. If this three-digit number is 001 to 274, it can be taken as a randomly chosen page number (if it is 000 or larger than 274, simply skip it and choose another three-digit number, e.g., the next one on the table). Then one might examine the next digit in the table; if it is a 1, 2, or 3, let it denote a page column (if a digit other than 1, 2, or 3 is encountered, it is ignored, passing to the next digit that is 1, 2, or 3). Then one could look at the next two-digit number in the table; if it is from 01 to 98, let it represent a randomly selected name within that column. This three-step procedure would be performed a total of 200 times to obtain the desired random sample. One can proceed in any direction in the random number table: left to right, right to left, upward, downward, or diagonally; but the direction should be decided on before looking at the table. Computers are capable of quickly generating random numbers (sometimes called “pseudorandom” numbers because the number generation is not perfectly random), and this is how Table 41 from *Appendix: Statistical Tables and Graphs* was derived.

---

*This use of the terms *population* and *sample* was established by Karl Pearson (1903).

†This concept of random sampling was established by Karl Pearson between 1897 and 1903 (Miller, 2004a).

Very often it is not possible to assign a number to each member of a population, and random sampling then involves biological, rather than simply mathematical, considerations. That is, the techniques for sampling Montana bobcats or Kansas grasshoppers require knowledge about the particular organism to ensure that the sampling is random. Researchers consult relevant books, periodical articles, or reports that address the specific kind of biological measurement to be obtained.

4 PARAMETERS AND STATISTICS

Several measures help to describe or characterize a population. For example, generally a preponderance of measurements occurs somewhere around the middle of the range of a population of measurements. Thus, some indication of a population "average" would express a useful bit of descriptive information. Such information is called a *measure of central tendency* (also called a *measure of location*).

It is also important to describe how dispersed the measurements are around the "average." That is, we can ask whether there is a wide spread of values in the population or whether the values are rather concentrated around the middle. Such a descriptive property is called a *measure of variability* (or a *measure of dispersion*).

A quantity such as a measure of central tendency or a measure of dispersion is called a *parameter* when it describes or characterizes a population, and we shall be very interested in discussing parameters and drawing conclusions about them. Section 2 pointed out, however, that one seldom has data for entire populations, but nearly always has to rely on samples to arrive at conclusions about populations. Thus, one rarely is able to calculate parameters. However, by random sampling of populations, parameters can be estimated well. An estimate of a population parameter is called a *statistic*.* It is statistical convention to represent population parameters by Greek letters and sample statistics by Latin letters; will demonstrate this custom for specific examples.

The statistics one calculates will vary from sample to sample for samples taken from the same population. Because one uses sample statistics as estimates of population parameters, it behooves the researcher to arrive at the "best" estimates possible. As for what properties to desire in a "good" estimate, consider the following.

First, it is desirable that if we take an indefinitely large number of samples from a population, the long-run average of the statistics obtained will equal the parameter being estimated. That is, for some samples a statistic may underestimate the parameter of interest, and for others it may overestimate that parameter; but in the long run the estimates that are too low and those that are too high will "average out." If such a property is exhibited by a statistic, we say that we have an *unbiased* statistic or an unbiased estimator.

Second, it is desirable that a statistic obtained from any single sample from a population be very close to the value of the parameter being estimated. This property of a statistic is referred to as *precision*,† *efficiency*, or *reliability*. As we commonly secure only one sample from a population, it is important to arrive at a close estimate of a parameter from a single sample.

---

*This use of the terms *parameter* and *statistic* was defined by R. A. Fisher as early as 1922 (Miller, 2004a; Savage, 1976).

†The precision of a sample statistic, as defined here, should not be confused with the precision of a measurement.

Third, consider that one can take larger and larger samples from a population (the largest sample being the entire population). As the sample size increases, a *consistent* statistic will become a better estimate of the parameter it is estimating. Indeed, if the sample were the size of the population, then the best estimate would be obtained: the parameter itself.

5 OUTLIERS

Occasionally, a set of data will have one or more observations that are so different, relative to the other data in the sample, that we doubt they should be part of the sample. For example, suppose a researcher collected a sample consisting of the body weights of nineteen 20-week-old mallard ducks raised in individual laboratory cages, for which the following 19 data were recorded:

1.87, 3.75, 3.79, 3.82, 3.85, 3.87, 3.90, 3.94, 3.96, 3.99,
3.99, 4.00, 4.03, 4.04, 4.05, 4.06, 4.09, 8.97, and 39.8 kilograms.

Visual inspection of these 19 recorded data casts doubt upon the smallest datum (1.87 kg) and the two largest data (8.97 kg and 39.8 kg) because they differ so greatly from the rest of the weights in the sample. Data in striking disagreement with nearly all the other data in a sample are often called *outliers* or *discordant data*, and the occurrence of such observations generally calls for closer examination.

Sometimes it is clear that an outlier is the result of incorrect recording of data. In the preceding example, a mallard duck weight of 39.8 kg is highly unlikely (to say the least!), for that is about the weight of a 12-year-old boy or girl (and such a duck would probably not fit in one of the laboratory cages). In this case, inspection of the data records might lead us to conclude that this body weight was recorded with a careless placement of the decimal point and should have been 3.98 kg instead of 39.8 kg. And, upon interrogation, the research assistant may admit to weighing the eighteenth duck with the scale set to pounds instead of kilograms, so the metric weight of that animal should have been recorded as 4.07 (not 8.97) kg.

Also, upon further examination of the data-collection process, we may find that the 1.87-kg duck was taken from a wrong cage and was, in fact, only 4 weeks old, not 20 weeks old, and therefore did not belong in this sample. Or, perhaps we find that it was not a mallard duck, but some other bird species (and, therefore, did not belong in this sample). Statisticians say a sample is *contaminated* if it contains a datum that does not conform to the characteristics of the population being sampled. So the weight of a 4-week-old duck, or of a bird of a different species, would be a statistical contaminant and should be deleted from this sample.

There are also instances where it is known that a measurement was faulty—for example, when a laboratory technician spills coffee onto an electronic measuring device or into a blood sample to be analyzed. In such a case, the measurements known to be erroneous should be eliminated from the sample.

However, outlying data can also be correct observations taken from an intended population, collected purely by chance. As we shall see, when drawing a random sample from a population, it is relatively likely that a datum in the sample will be around the average of the population and very unlikely that a sample datum will be dramatically far from the average. But sample data very far from the average still may be possible.

It should also be noted that in some situations the examination of an outlier may reveal the effect of a previously unsuspected factor. For example, the 1.87-kg duck might, indeed, have been a 20-week-old mallard but suffering from a genetic mutation or a growth-impeding disease deserving of further consideration in additional research.

In summary, it is not appropriate to discard data simply because they appear (to someone) to be unreasonably extreme. However, if there is a very obvious reason for correcting or eliminating a datum, such as the situations described previously, the incorrect data should be corrected or eliminated. In some other cases questionable data can be *accommodated* in statistical analysis, perhaps by employing statistical procedures that give them less weight or analytical techniques that are *robust* in that they are resistant to effects of discrepant data. And in situations when this cannot be done, dubious data will have to remain in the sample (perhaps encouraging the researcher to repeat the experiment with a new set of data).

The idea of rejecting erroneous data dates back over 200 years; and recommendations for formal, objective methods for such rejection began to appear about 150 years ago. Major discussions of outliers, their origin, and treatment (rejection or accommodation) are those of Barnett and Lewis (1994), Beckman and Cook (1983), and Thode (2002: 123–142).

Measures of Central Tendency

In samples, as well as in populations, one generally finds a preponderance of values somewhere around the middle of the range of observed values. The description of this concentration near the middle is an *average*, or a *measure of central tendency* to the statistician. It is also termed a *measure of location*, for it indicates where, along the measurement scale, the sample or population is located. Various measures of central tendency are useful population parameters, in that they describe an important property of populations. This chapter discusses the characteristics of these parameters and the sample statistics that are good estimates of them.

1 THE ARITHMETIC MEAN

The most widely used measure of central tendency is the *arithmetic mean*,* usually referred to simply as the *mean*,† which is the measure most commonly called an "average."

Each measurement in a population may be referred to as an X_i (read "X sub i") value. Thus, one measurement might be denoted as X_1, another as X_2, another as X_3, and so on. The subscript i might be any integer value up through N, the total number of X values in the population.‡ The mean of the population is denoted by the Greek letter μ (lowercase mu) and is calculated as the sum of all the X_i values divided by the size of the population.

The calculation of the population mean can be abbreviated concisely by the formula

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}. \tag{1}$$

*As an adjective, *arithmetic* is pronounced with the accent on the third syllable. In early literature on the subject, the adjective *arithmetical* was employed.

†The term *mean* (as applied to the arithmetic mean, as well as to the geometric and harmonic means of Section 4) dates from ancient Greece (Walker, 1929: 183), with its current statistical meaning in use by 1755 (Miller, 2004a; Walker, 1929: 176); *central tendency* appeared by the late 1920s (Miller, 2004a).

‡Charles Babbage (1791–1871) (O'Connor and Robertson, 1998) was an English mathematician and inventor who conceived principles used by modern computers—well before the advent of electronics—and who, in 1832, proposed the modern convention of italicizing Latin (also called Roman) letters to denote quantities; nonitalicized letters had already been employed for this purpose for more than six centuries (Miller, 2001).

From Chapter 3 of *Biostatistical Analysis*, Fifth Edition, Jerrold H. Zar. Copyright © 2010 by Pearson Education, Inc. Publishing as Pearson Prentice Hall. All rights reserved.

The Greek letter Σ (capital sigma) means "summation"* and $\sum_{i=}^{N} X$ means "summation of all X_i values from X_1 through X_N." Thus, for example, $\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4$ and $\sum_{i=3}^{5} X_i = X_3 + X_4 + X_5$. Since, in statistical computations, summations are nearly always performed over the entire set of X_i values, this text will assume $\sum X_i$ to mean "sum X_i's over all values of i," simply as a matter of printing convenience, and $\mu = \sum X_i/N$ would therefore designate the same calculation as would $\mu = \sum_{i=1}^{N} X_i/N$.

The most efficient, unbiased, and consistent estimate of the population mean, μ, is the sample mean, denoted as $\overline{X}$ (read as "X bar"). Whereas the size of the population (which we generally do not know) is denoted as N, the size of a sample is indicated by n, and $\overline{X}$ is calculated as

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{or} \quad \overline{X} = \frac{\sum X_i}{n}, \tag{2}$$

which is read "the sample mean equals the sum of all measurements in the sample divided by the number of measurements in the sample."† Example 1 demonstrates the calculation of the sample mean. Note that the mean has the same units of measurement as do the individual observations.

EXAMPLE 1 A Sample of 24 from a Population of Butterfly Wing Lengths

X_i (in centimeters): 3.3, 3.5, 3.6, 3.6, 3.7, 3.8, 3.8, 3.8, 3.9, 3.9, 3.9, 4.0, 4.0, 4.0, 4.0, 4.1, 4.1, 4.1, 4.2, 4.2, 4.3, 4.3, 4.4, 4.5.

$$\sum X_i = 95.0 \text{ cm}$$
$$n = 24$$
$$\overline{X} = \frac{\sum X_i}{n} = \frac{95.0 \text{ cm}}{24} = 3.96 \text{ cm}$$

*Mathematician Leonhard Euler (1707–1783; born in Switzerland, worked mostly in Russia), in 1755, was the first to use Σ to denote summation (Cajori, 1928/9, Vol. II: 61).

†The modern symbols for plus and minus ("+" and "−") appear to have first appeared in a 1456 unpublished manuscript by German mathematician and astronomer Regiomontanus (Johannes Müller, 1436–1476), with Bohemia-born Johann (Johannes) Widman (1562–1498) the first, in 1489, to use them in print (Cajori, 1928/9, Vol. I: 128, 231–232). The modern equal sign ("=") was invented by Welsh physician and mathematician Robert Recorde (1510–1558), who published it in 1557 (though its use then disappeared in print until 1618), and it was well recognized starting in 1631 (Cajori, ibid.: 298; Gullberg, 1997: 107). Recorde also was the first to use the plus and minus symbols in an English work (Miller, 2004b). Using a horizontal line to express division derives from its use, in denoting fractions, by Arabic author Al-Ḥaṣṣâr in the twelfth century, though it was not consistently employed for several more centuries (Cajori, ibid. I: 269, 310). The slash mark ("/"; also known as a solidus, virgule, or diagonal) was recommended to denote division by the English logician and mathematician Augustus De Morgan (1806–1871) in 1845 (ibid. I: 312–313), and the India-born Swiss author Johann Heinhirch Rahn (1622–1676) proposed, in 1659, denoting division by the symbol "÷", which previously was often used by authors as a minus sign (ibid.: 211, 270; Gullberg, 1997: 105). Many other symbols were used for mathematical operations, before and after these introductions (e.g., Cajori, ibid.: 229–245).

If, as in Example 1, a sample contains multiple identical data for several values of the variable, then it may be convenient to record the data in the form of a frequency table, as in Example 2. Then X_i can be said to denote each of k different measurements and f_i can denote the frequency with which that X_i occurs in the sample. The sample mean may then be calculated, using the sums of the products of f_i and X_i, as*

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n}. \tag{3}$$

Example 2 demonstrates this calculation for the same data as in Example 1.

EXAMPLE 2 The Data from Example 1 Recorded as a Frequency Table

X_i (cm)	f_i	$f_i X_i$ (cm)
3.3	1	3.3
3.4	0	0
3.5	1	3.5
3.6	2	7.2
3.7	1	3.7
3.8	3	11.4
3.9	3	11.7
4.0	4	16.0
4.1	3	12.3
4.2	2	8.4
4.3	2	8.6
4.4	1	4.4
4.5	1	4.5
	$\sum f_i = 24$	$\sum f_i X_i = 95.0$ cm

$$k = 13$$

$$\sum_{i=1}^{k} f_i = n = 24$$

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n} = \frac{95.0 \text{ cm}}{24} = 3.96 \text{ cm}$$

$$\begin{aligned} \text{median} &= 3.95 \text{ cm} + \left(\frac{1}{4}\right)(0.1 \text{ cm}) \\ &= 3.95 \text{ cm} + 0.025 \text{ cm} \\ &= 3.975 \text{ cm} \end{aligned}$$

A similar procedure is computing what is called a *weighted mean*, an expression of the average of several means. For example, we may wish to combine the mean of 3.96 cm from the sample of 24 measurements in Example 1 with a mean of 3.78 cm from a sample of 30 measurements and a mean of 4.02 cm from a sample of 15. These three means would be from a total of $24 + 30 + 15 = 69$ data; and if we had all 69 of the data we could sum them and divide the sum by 69 to obtain the overall mean length. However, that overall mean can be obtained without knowing the 69

*Denoting the multiplication of two quantities (e.g., a and b) by their adjacent placement (i.e., ab) derives from practices in Hindu manuscripts of the seventh century (Cajori, 1928/9, Vol. I: 77, 250). Modern multiplication symbols include a raised dot (as in $a \cdot b$), which was suggested in a 1631 posthumous publication of Thomas Harriot (1560?–1621) and prominently adopted in 1698 by the outstanding mathematician Gottfried Wilhelm Leibniz (1646–1716, in what is now Germany); the St. Andrew's cross (as in $a \times b$), which was used in 1631 by English mathematician William Oughtred (1574–1660) though it was not in general use until more than 200 years later; and the letter X, which was used, perhaps by Oughtred, as early as 1618 (Cajori, ibid.: 251; Gullberg, 1997: 104; Miller 2004b). Johann Rahn's 1659 use of an asterisk-like symbol (as in $a * b$) (Cajori, ibid: 212–213) did not persist but resurfaced in electronic computer languages of the latter half of the twentieth century.

individual measurements, by employing Equation 3 with $f_1 = 24$, $X_1 = 3.96$ cm, $f_2 = 30$, $X_2 = 3.78$ cm, $f_3 = 15$, $X_3 = 4.02$ cm, and $n = 69$. This would yield a weighted mean of $\overline{X} = [(24)(3.96\text{ cm}) + (30)(3.78\text{ cm}) + (15)(4.02\text{ cm})]/69 = (268.74\text{ cm})/69 = 3.89$ cm.

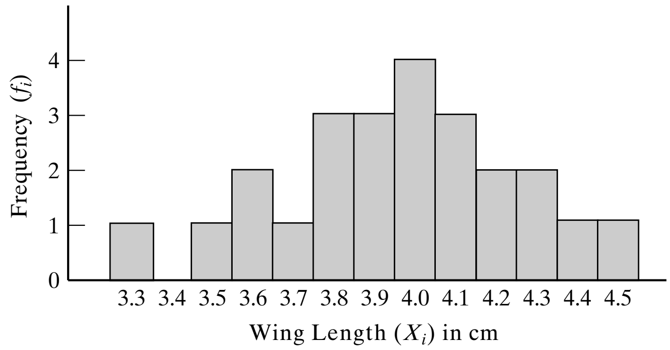

FIGURE 1: A histogram of the data in Example 2. The mean (3.96 cm) is the center of gravity of the histogram, and the median (3.975 cm) divides the histogram into two equal areas.

If data are plotted as a histogram (Figure 1), the mean is the *center of gravity* of the histogram.* That is, if the histogram were made of a solid material, it would balance horizontally with the fulcrum at $\overline{X}$. The mean is applicable to both ratio- and interval-scale data; it should not be used for ordinal data and cannot be used for nominal data.

2 THE MEDIAN

The median is typically defined as the middle measurement in an ordered set of data.† That is, there are just as many observations larger than the median as there are smaller. The sample median is the best estimate of the population median. In a symmetrical distribution (such as Figures 2a and 2b) the sample median is also an unbiased and consistent estimate of μ, but it is not as efficient a statistic as $\overline{X}$ and should not be used as a substitute for $\overline{X}$. If the frequency distribution is asymmetrical, the median is a poor estimate of the mean.

The median of a sample of data may be found by first arranging the measurements in order of magnitude. The order may be either ascending or descending, but ascending order is most commonly used as is done with the samples in Examples 1, 2, and 3. Then, we define the sample median as

$$\text{sample median} = X_{(n+1)/2}. \tag{4}$$

*The concept of the mean as the center of gravity was used by L. A. J. Quetelet in 1846 (Walker, 1929: 73).

†The concept of the median was conceived as early as 1816, by K. F. Gauss; enunciated and reinforced by others, including F. Galton in 1869 and 1874; and independently discovered and promoted by G. T. Fechner beginning in 1874 (Walker, 1929: 83–88, 184). It received its name, in English, from F. Galton in 1882 (David, 1995) and, in French, from A. A. Cournot in 1843 (David, 1998a).

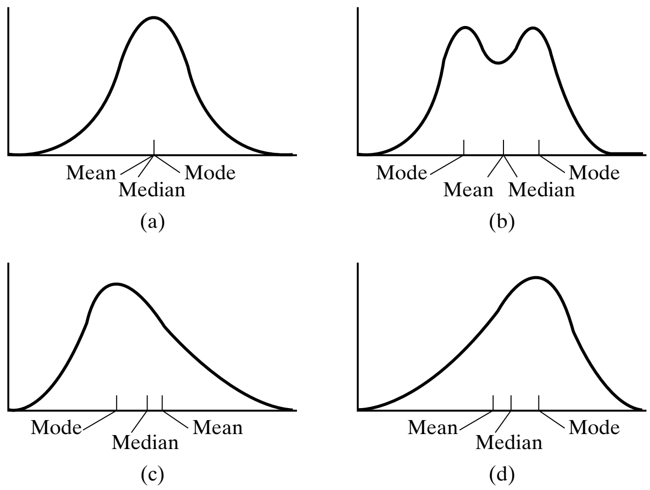

FIGURE 2: Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is asymmetrical and said to be positively skewed, and (d) is asymmetrical and said to be negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution b is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.

EXAMPLE 3 Life Span for Two Species of Birds in Captivity

The data for each species are arranged in order of magnitude

X_i (mo)	X_i (mo)
16	34
32	36
37	38
39	45
40	50
41	54
42	56
50	59
82	69
	91
$n = 9$	$n = 10$
median $= X_{(n+1)/2} = X_{(9+1)/2}$	median $= X_{(n+1)/2} = X_{(10+1)/2}$
$= X_5 = 40$ mo	$= X_{5.5} = 52$ mo
$\overline{X} = 42.11$ mo	$\overline{X} = 53.20$ mo

If the sample size (n) is odd, then the subscript in Equation 4 will be an integer and will indicate which datum is the middle measurement in the ordered sample. For the data of species A in Example 3, $n = 9$ and the sample median is $X_{(\ +1)/2} = X_{(9+1)/2} = X_5 = 40$ mo. If n is even, then the subscript in Equation 4 will be a number midway between two integers. This indicates that there is not a middle value in the ordered list of data; instead, there are two middle values, and the median is defined as the midpoint between them. For the species B data in Example 3, $n = 10$ and $X(n + 1)/2 = X_{(10+1)/2} = X_{5.5}$, which signifies that the median is midway between X_5 and X_6, namely a median of $(50 \text{ mo} + 54 \text{ mo})/2 = 52$ mo.

Note that the median has the same units as each individual measurement. If data are plotted as a frequency histogram (e.g., Figure 1), the median is the value of X that divides the area of the histogram into two equal parts. In general, the sample median is a more efficient estimate of the population median when the sample size is large.

If we find the middle value(s) in an ordered set of data to be among identical observations (referred to as *tied* values), as in Example 1 or 2, a difficulty arises. If we apply Equation 4 to these 24 data, then we conclude the median to be $X_{12.5} = 4.0$ cm. But four data are tied at 4.0 cm, and eleven measurements are less than 4.0 cm and nine are greater. Thus, 4.0 cm does not fit the definition above of the median as that value for which there is the same number of data larger and smaller. Therefore, a better definition of the median of a set of data is that value for which no more than half the data are smaller and no more than half are larger.

When the sample median falls among tied observations, we may interpolate to better estimate the population median. Using the data of Example 2, we desire to estimate a value below which 50% of the observations in the population lie. Fifty percent of the observations in the sample would be 12 observations. As the first 7 classes in the frequency table include 11 observations and 4 observations are in class 4.0 cm, we know that the desired sample median lies within the range of 3.95 to 4.05 cm. Assuming that the four observations in class 4.0 cm are distributed evenly within the 0.1-cm range of 3.95 to 4.05 cm, then the median will be $\left(\frac{1}{4}\right)(0.1 \text{ cm}) = 0.025$ cm into this class. Thus, the median $= 3.95 \text{ cm} + 0.025 \text{ cm} = 3.975$ cm. In general, for the sample median within a class interval containing tied observations,

$$\text{median} = \begin{pmatrix} \text{lower limit} \\ \text{of interval} \end{pmatrix} + \left(\frac{0.5n - \text{cum. freq.}}{\text{no. of observations in interval}}\right)\begin{pmatrix} \text{interval} \\ \text{size} \end{pmatrix}, \quad (5)$$

where "cum. freq." refers to the cumulative frequency of the previous classes.* By using this procedure, the calculated median will be the value of X that divides the area of the histogram of the sample into two equal parts.

The median expresses less information than does the mean, for it does not take into account the actual value of each measurement, but only considers the rank of each measurement. Still, it offers advantages in some situations. For example, extremely high or extremely low measurements ("outliers") do not affect the median as much as they affect the mean (causing the sample median to be called a "resistant" statistic). Distributions that are not symmetrical around the mean (such as in Figures 2c and 2d) are said to be *skewed*.† When we deal with skewed

---

*This procedure was enunciated in 1878 by the German psychologist Gustav Theodor Fechner (1801–1887) (Walker, 1929: 86).

†This term, applied to a distribution and to a curve, was used as early as 1895 by Karl Pearson (Miller, 2004a).

populations and do not want the strong influence of outliers, we may prefer the median to the mean to express central tendency.

Note that in Example 3 the researcher would have to wait 82 months to compute a mean life expectancy for species A and 91 months for species B, whereas the median for species A could be determined in only 40 months and in only 52 months for species B. Also, to calculate a median one does not need to have accurate data for all members of the sample. If, for example, we did not have the first three data for species A accurately recorded, but could state them as "less than 39 months," then the median could have been determined just as readily as if we had all 9 data fully recorded, while calculation of the mean would not have been possible.

The expression "LD fifty" (LD_{50}), used in some areas of biological research, is simply the median lethal dose (and is so named because the median is the 50th percentile).

The median can be determined not only for interval-scale and ratio-scale data, but also for data on an ordinal scale, data for which the use of the mean usually would not be considered appropriate. But neither the median nor the mean is applicable to nominal data.

3 THE MODE

The *mode* is commonly defined as the most frequently occurring measurement in a set of data.* In Example 2, the mode is 4.0 cm. But it is perhaps better to define a mode as a measurement of relatively great concentration, for some frequency distributions may have more than one such point of concentration, even though these concentrations might not contain precisely the same frequencies. Thus, a sample consisting of the data 6, 7, 7, 8, 8, 8, 8, 8, 8, 9, 9, 10, 11, 12, 12, 12, 12, 12, 13, 13, and 14 mm would be said to have two modes: at 8 mm and 12 mm. (Some authors would refer to 8 mm as the "major mode" and call 12 mm the "minor mode.") A distribution in which each different measurement occurs with equal frequency is said to have no mode. If two consecutive values of X have frequencies great enough to declare the X values modes, the mode of the distribution may be said to be the midpoint of these two X's; for example, the mode of 3, 5, 7, 7, 7, 8, 8, 8, and 10 liters is 7.5 liters. A distribution with two modes is said to be *bimodal* (e.g., Figure 2b) and may indicate a combination of two distributions with different modes (e.g., heights of men and women). Modes are often discerned from histograms or frequency polygons; but we should be aware that the shape of such graphs and therefore the appearance of modes, may be influenced by the measurement intervals on the horizontal axis.

The sample mode is the best estimate of the population mode. When we sample a symmetrical unimodal population, the mode is an unbiased and consistent estimate of the mean and median (Figure 2a), but it is relatively inefficient and should not be so used. As a measure of central tendency, the mode is affected by skewness less than is the mean or the median, but it is more affected by sampling and grouping than these other two measures. The mode, but neither the median nor the mean, may be used for data on the nominal, as well as the ordinal, interval, and ratio scales of measurement. In a unimodal asymmetric distribution (Figures 2c and 2d), the median lies about one-third the distance between the mean and the mode.

The mode is not often used in biological research, although it is often interesting to report the number of modes detected in a population, if there are more than one.

*The term *mode* was introduced by Karl Pearson in 1895 (David, 1995).

4 OTHER MEASURES OF CENTRAL TENDENCY

(a) The Geometric Mean. The *geometric mean* is the *n*th root of the product of the *n* data:

$$\overline{X}_G = \sqrt[n]{X_1 X_2 \;\; {}_3 \ldots X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i}\,. \tag{6}$$

Capital Greek pi, Π, means “take the product”[*] in an analogous fashion as Σ indicates “take the sum.” The geometric mean may also be calculated as the antilogarithm of the arithmetic mean of the logarithms of the data (where the logarithms may be in any base); this is often more feasible computationally:

$$\overline{X}_G = \text{antilog}\left(\frac{\log X_1 + \log X_2 + \cdots + \log X_n}{n}\right) = \text{antilog}\,\frac{\sum_{i=1}^{n} \log X_i}{n}. \tag{7}$$

The geometric mean is appropriate to use only for ratio-scale data and only when all of the data are positive (that is, greater than zero). If the data are all equal, then the geometric mean, $\overline{X}_G$, is equal to the arithmetic mean, $\overline{X}$ (and also equal to the harmonic mean described below); if the data are not all equal, then[†] $\overline{X}_G < \overline{X}$.

$\overline{X}_G$ is sometimes used as a measure of location when the data are highly skewed to the right (i.e., when there are many more data larger than the arithmetic mean than there are data smaller than the arithmetic mean).

$\overline{X}_G$ is also useful when dealing with data that represent ratios of change. As an illustration of this, Example 4 considers changes in the size of a population of organisms over four decades. Each of the original data (population size at the end of a decade) is expressed as a ratio, X_i, of the population size to the population size of the previous decade. The geometric mean of those ratios is computed and may be thought of as representing the average rate of growth per decade (which is the same as a constant rate of compound interest). This example demonstrates that the arithmetic mean of those ratios is $\overline{X} = 1.1650$ (i.e., 16.50% growth) per decade. But over the four decades of population change, this mean would have us calculate a final population size of $(10{,}000)(1.1650)(1.1650)(1.1650)(1.1650) = 18{,}421$, which is *not* the population size recorded at the end of the fourth decade. However, using the geometric mean, $\overline{X}_G$, to indicate the average rate of growth, the final population size would be computed to be $(10{,}000)(1.608)(1.608)(1.608)(1.608) = 18{,}156$, which is the fourth-decade population size that was observed.

[*] Use of this symbol to indicate taking the product was introduced by René Descartes (Gullberg, 1997: 105).

[†] The symbols “<” and “>” (meaning “less than” and “greater than”) were inserted by someone else into a 1631 posthumous publication by the English mathematician and astronomer Thomas Harriot (1560?–1621), (Cajori, 1928/9, Vol. I: 199; Gullberg, 1997: 109; Miller, 2004b). The symbols for “less than or equal to” ($\leq$) and “greater than or equal to” ($\geq$) were written as $\leqq$ and $\geqq$ when introduced by the French scientist Pierre Bouguere (1698–1758) in 1734. (Gullberg, 1997: 109).

EXAMPLE 4 The Geometric Mean of Ratios of Change

Decade	*Population Size*	*Ratio of Change* X_i
0	10,000	
1	10,500	$\frac{10{,}500}{10{,}000} = 1.05$
2	11,550	$\frac{11{,}550}{10{,}500} = 1.10$
3	13,860	$\frac{13{,}860}{11{,}550} = 1.20$
4	18,156	$\frac{18{,}156}{13{,}860} = 1.31$

$$\overline{X} = \frac{1.05 + 1.10 + 1.20 + 1.31}{4} = \frac{4.66}{4} = 1.1650$$

and (10,000)(0.1650)(1.650)(1.650)(1.650) = 18,421

But,

$$\overline{X}_G = \sqrt[4]{(1.05)(1.10)(1.20)(1.31)} = \sqrt[4]{1.8157} = 1.1608$$

or

$$\overline{X}_G = \text{antilog}\left[\frac{\log(1.05) + \log(1.10) + \log(1.20) + \log(1.31)}{4}\right]$$

$$= \frac{\text{antilog}(0.0212 + 0.0414 + 0.0792 + 0.1173)}{4} = \frac{\text{antilog}(0.2591)}{4}$$

$$= \text{antilog } 0.0648 = 1.1608$$

and (10,000)(1.1608)(1.1608)(1.1608)(1.1608) = 18,156

(b) The Harmonic Mean. The *harmonic mean* is the reciprocal of the arithmetic mean of the reciprocals of the data:

$$\overline{X}_H = \frac{1}{\frac{1}{n}\sum \frac{1}{X_i}} = \frac{n}{\sum \frac{1}{X_i}}. \tag{8}$$

It may be used for ratio-scale data when no datum is zero. If all of the data are identical, then the harmonic mean, $\overline{X}_H$, is equal to the arithmetic mean, $\overline{X}$ (and equal to the geometric mean, $\overline{X}_G$). If the data are all positive and not identical, then $\overline{X}_H < \overline{X}_G < \overline{X}$.

$\overline{X}_H$ finds use when desiring an average of rates, as described by Croxton, Cowden, and Klein (1967: 182–188). For example, consider that a flock of birds flies from a roosting area to a feeding area 20 km away, flying at a speed of 40 km/hr (which takes 0.5 hr). The flock returns to the roosting area along the same route (20 km), flying at 20 km/hr (requiring 1 hr of flying time). To ask what the average flying speed was, we might employ Equation 2 and calculate the arithmetic mean as $\overline{X}$ = (40 km/hr + 20 km/hr)/2 = 30 km/hr. However, this answer may not be satisfying, because a total of 40 km was traveled in 1.5 hr, indicating a speed of (40 km)/(1.5 hr) = 26. $\overline{X}_G$) is 26.7 km/hr.

EXAMPLE 5 The Harmonic Mean of Rates

$$X_1 = 40 \text{ km/hr}, \; X_2 = 20 \text{ km/hr}$$

$$\overline{X} = \frac{40 \text{ km/hr} + 20 \text{ km/hr}}{2} = \frac{60 \text{ km/hr}}{2} = 30 \text{ km/hr}$$

But

$$\overline{X}_H = \frac{2}{\dfrac{1}{40 \text{ km/hr}} + \dfrac{1}{20 \text{ km/hr}}} = \frac{2}{0.0250 \text{ hr/km} + 0.0500 \text{ hr/km}}$$

$$= \frac{2}{0.075 \text{ hr/km}} = 26.67 \text{ km/hr}$$

(c) The Range Midpoint. The *range midpoint*, or *midrange*, is a measure of location defined as the point halfway between the minimum and the maximum values in the set of data. It may be used with data measured on the ratio, interval, or ordinal scale; but it is not generally a good estimate of location, for it utilizes relatively little information from the data. (However, the so-called mean daily temperature is often reported as the mean of the minimum and maximum and is, therefore, a range midpoint.)

The midpoint of any two symmetrically located percentiles, such as the point midway between the first and third quartiles (i.e., the 25th and 75th percentiles), may be used as a location measure in the same fashion as the range midpoint is used (see Dixon and Massey, 1969: 133–134). Such measures are not as adversely affected by aberrantly extreme values as is the range midpoint, and they may be applied to ratio or interval data. If used with ordinal data, they (and the range midpoint) would be the same as the median.

5 CODING DATA

Often in the manipulation of data, considerable time and effort can be saved if *coding* is employed. Coding is the conversion of the original measurements into easier-to-work-with values by simple arithmetic operations. Generally coding employs a *linear transformation* of the data, such as multiplying (or dividing) or adding (or subtracting) a constant. The addition or subtraction of a constant is sometimes termed a translation of the data (i.e., changing the origin), whereas the multiplication or division by a constant causes an expansion or contraction of the scale of measurement.

EXAMPLE 6 Coding Data to Facilitate Calculations

Sample 1 (Coding by Subtraction: $A = -840$ g)

X_i (g)	coded $X_i = X_i - 840$ g
842	2
844	4
846	6
846	6
847	7
848	8
849	9

$\sum X_i = 5922$ g coded $\sum X_i = 42$ g

$\overline{X} = \frac{5922 \text{ g}}{7} = 846$ g coded $\overline{X} = \frac{42 \text{ g}}{7} = 6$ g

$$\overline{X} = \text{coded } \overline{X} - A$$
$$= 6 \text{ g} - (-840 \text{ g})$$
$$= 846 \text{ g}$$

Sample 2 (Coding by Division: $M = 0\ 001$ liters/ml)

X_i (ml)	coded $X_i = (X_i)(0.001$ liters/ml$) = X_i$ liters
8,000	8.000
9,000	9.000
9,500	9.500
11,000	11.000
12,500	12.500
13,000	13.000

$\sum X_i = 63{,}000$ ml coded $\sum X_i = 63.000$ liters

$\overline{X} = 10{,}500$ ml coded $\overline{X} = 10.500$ liters

$$\overline{X} = \text{coded } \frac{\overline{X}}{M}$$
$$= \frac{10.500 \text{ liters}}{0.001 \text{ liters/ml}}$$
$$= 10{,}500 \text{ ml}$$

The first set of data in Example 6 are coded by subtracting a constant value of 840 g. Not only is each coded value equal to $X_i - 840$ g, but the mean of the coded values is equal to $\overline{X} - 840$ g. Thus, the easier-to-work-with coded values may be used to calculate a mean that then is readily converted to the mean of the original data, simply by adding back the coding constant.

In Sample 2 of Example 6, the observed data are coded by dividing each observation by 1000 (i.e., by multiplying by 0.001).* The resultant mean only needs to be multiplied by the coding factor of 1000 (i.e., divided by 0.001) to arrive at the mean of the original data. As the other measures of central tendency have the same units as the mean, they are affected by coding in exactly the same fashion.

Coding affects the median and mode in the same way as the mean is affected. The widespread use of computers has greatly diminished the need for researchers to utilize coding (although computer software may use it).

*In 1593, mathematician Christopher Clavius (1538–1612, born in what is now Germany but spent most of his life in what is now Italy; also credited with proposing the currently used Gregorian calendar rules regarding leap years; O'Connor and Robertson, 1996) became the first to use a decimal point to separate units from tenths; in 1617, the Scottish mathematician John Napier (1550–1617) used both points and commas for this purpose (Cajori, 1928/9. Vol. I: 322–323), and the comma is still so used in some parts of the world. In some countries a raised dot has been used—a symbol Americans sometimes employ to denote multiplication.

EXERCISES

1. If $X_1 = 3.1$ kg, $X_2 = 3.4$ kg, $X_3 = 3.6$ kg, $X_4 = 7$ kg, and $X_5 = 4.0$ kg, calculate the value of
   (a) $\sum_{i=1}^{4} X_i$,
   (b) $\sum_{i=2}^{4} X_i$,
   (c) $\sum_{i=1}^{5} X_i$,
   (d) $\sum X_i$.

2. **(a)** Calculate the mean of the five weights in Exercise 1.
   (b) Calculate the median of those weights.

3. The ages, in years, of the faculty members of a university biology department are 32.2, 37.5, 41.7, 53.8, 50.2, 48.2, 46.3, 65.0, and 44.8.
   (a) Calculate the mean age of these nine faculty members.
   (b) Calculate the median of the ages.
   (c) If the person 65.0 years of age retires and is replaced on the faculty with a person 46.5 years old, what is the new mean age?
   (d) What is the new median age?

4. Consider the following frequency tabulation of leaf weights (in grams):

X_i	f_i
1.85–1.95	2
1.95–2.05	1
2.05–2.15	2
2.15–2.25	3
2.25–2.35	5
2.35–2.45	6
2.45–2.55	4
2.55–2.65	3
2.65–2.75	1

   Using the midpoints of the indicated ranges of X_i,
   (a) Calculate the mean leaf weight using Equation 2, and
   (b) Calculate the mean leaf weight using Equation 3.
   (c) Calculate the median leaf weight using Equation 4, and
   (d) Calculate the median using Equation 5.
   (e) Determine the mode of the frequency distribution.

5. A fruit was collected from each of eight lemon trees, with the intent of measuring the calcium concentration in the rind (grams of calcium per 100 grams of dry rind). The analytical method used could only detect a concentration of at least 0.80 g/100 g of dry weight. Six of the eight concentrations were measured to be 1.02, 0.98, 0.91, 0.84, 0.87, 1.04 g/100 g of dry weight, and two of the concentrations were known to be less than 0.80 g/100 g of dry weight. What is the median of this sample of eight data?

ANSWERS TO EXERCISES

1. **(a)** 13.8 kg; **(b)** 10.7 kg; **(c)** 17.8 kg; **(d)** 17.8 kg.
2. **(a)** 3.56 kg; **(b)** 3.6 kg.
3. **(a)** 46.63 yr; **(b)** 46.3 yr; **(c)** 44.58 yr; **(d)** 46.3 yr.
4. **(a)** 2.33 g; **(b)** 2.33 g; **(c)** 2.4 g; **(d)** 2.358 g; **(e)** 2.4 g.
5. 0.89 g/100 g.

Measures of Variability and Dispersion

In addition to a description of the central tendency of a set of data, it is generally desirable to have a description of the *variability*, or of the *dispersion*,* of the data. A measure of variability (or measure of dispersion, as it is often called) is an indication of the spread of measurements around the center of the distribution. Measurements that are concentrated around the center of a distribution of data have low variability (low dispersion), whereas data that are very spread out along the measurement scale have high variability (high dispersion). Measures of variability of a population are population parameters, and sample measures of variability are statistics that estimate those parameters.

1 THE RANGE

The difference between the highest and lowest measurements in a group of data is termed the *range*.† If sample measurements are arranged in increasing order of magnitude, as if the median were about to be determined, then

$$\text{sample range} = X_n - X_1, \tag{1}$$

which is

$$\text{sample range} = \text{largest } X - \text{smallest } X.$$

Sample 1 in Example 1 is a hypothetical set of ordered data in which $X_1 = 1.2$ g and $X_n = 2.4$ g. Thus, the range may be expressed as 1.2 to 2.4 g, or as 2.4 g $-$ 1.2 g $=$ 1.2 g. Note that the range has the same units as the individual measurements. Sample 2 in Example 1 has the same range as Sample 1.

*The statistical use of this term first appeared in an 1876 publication by Francis Galton (David, 1998a).

†This statistical term dates from an 1848 paper by H. Lloyd (David, 1995). It was already used by the Greek astronomer Hipparchus as a measure of dispersion in the second century B.C.E. (David, 1998b).

From Chapter 4 of *Biostatistical Analysis*, Fifth Edition, Jerrold H. Zar. Copyright © 2010 by Pearson Education, Inc. Publishing as Pearson Prentice Hall. All rights reserved.

EXAMPLE 1 Calculation of Measures of Dispersion for Two Hypothetical Samples of 7 Insect Body Weights

Sample 1

X_i (g)	$X_i - \overline{X}$ (g)	$\lvert X_i - \overline{X}\rvert$ (g)	$(X_i - \overline{X})^2$ (g^2)
1.2	−0.6	0.6	0.36
1.4	−0.4	0.4	0.16
1.6	−0.2	0.2	0.04
1.8	0.0	0.0	0.00
2.0	0.2	0.2	0.04
2.2	0.4	0.4	0.16
2.4	0.6	0.6	0.36
$\sum X_i = 12.6$ g	$\sum (X_i - \overline{X}) = 0.0$ g	$\sum \lvert X_i - \overline{X}\rvert = 2.4$ g	$\sum (X_i - \overline{X})^2 = 1.12\ g^2$

= sum of squared deviations from the mean
= "sum of squares"

$$n = 7;\ \overline{X} = \frac{\sum X_i}{n} = \frac{12.6\text{ g}}{7} = 1.8\text{ g}$$

$$\text{range} = X_7 - X_1 = 2.4\text{ g} - 1.2\text{ g} = 1.2\text{ g}$$

$$\text{interquartile range} = Q_3 - Q_1 = 2.2\text{ g} - 1.4\text{ g} = 0.8\text{ g}$$

$$\text{mean deviation} = \frac{\sum \lvert X_i - \overline{X}\rvert}{n} = \frac{2.4\text{ g}}{7} = 0.34\text{ g}$$

$$\text{variance} = s^2 = \frac{\sum (X_i - \overline{X})^2}{n - 1} = \frac{1.12\text{ g}^2}{6} = 0.1867\text{ g}^2$$

$$\text{standard deviation} = s = \sqrt{0.1867\text{ g}^2} = 0.43\text{ g}$$

Sample 2

X_i (g)	$X_i - \overline{X}$ (g)	$\lvert X_i - \overline{X}\rvert$ (g)	$(X_i - \overline{X})^2$ (g^2)
1.2	−0.6	0.6	0.36
1.6	−0.2	0.2	0.04
1.7	−0.1	0.1	0.01
1.8	0.0	0.0	0.00
1.9	0.1	0.1	0.01
2.0	0.2	0.2	0.04
2.4	0.6	0.6	0.36
$\sum X_i = 12.6$ g	$\sum (X_i - \overline{X}) = 0.0$ g	$\sum \lvert X_i - \overline{X}\rvert = 1.8$ g	$\sum (X_i - \overline{X})^2 = 0.82\ g^2$

= sum of squared deviations from the mean
= "sum of squares"

$$n = 7;\ \overline{X} = \frac{\sum X_i}{n} = \frac{12.6\text{ g}}{7} = 1.8\text{ g}$$

$$\text{range} = X_7 - X_1 = 2.4\text{ g} - 1.2\text{ g} = 1.2\text{ g}$$

$$\text{interquartile range} = Q_3 - Q_1 = 2.0\text{ g} - 1.6\text{ g} = 0.4\text{ g}$$

$$\text{mean deviation} = \frac{\sum |X_i - \overline{X}|}{n} = \frac{1.8\text{ g}}{7} = 0.26\text{ g}$$

$$\text{variance} = s^2 = \frac{\sum (X_i - \overline{X})^2}{n - 1} = \frac{0.82\text{ g}^2}{6} = 0.1367\text{ g}^2$$

$$\text{standard deviation} = s = \sqrt{0.1367\text{ g}^2} = 0.37\text{ g}$$

The range is a relatively crude measure of dispersion, inasmuch as it does not take into account any measurements except the highest and the lowest. Furthermore, it is unlikely that a sample will contain both the highest and lowest values in the population, so the sample range usually underestimates the population range; therefore, it is a biased and inefficient estimator. Nonetheless, it is considered useful by some to present the sample range as an estimate (although a poor one) of the population range. For example, taxonomists are often concerned with having an estimate of what the highest and lowest values in a population are expected to be. Whenever the range is specified in reporting data, however, it is usually a good practice to report another measure of dispersion as well. The range is applicable to ordinal-, interval-, and ratio-scale data.

2 DISPERSION MEASURED WITH QUANTILES

Because the sample range is a biased and inefficient estimate of the population range, being sensitive to extremely large and small measurements, alternative measures of dispersion may be desired. Just as the median is the value above and below which lies half the set of data, one can define measures, called *quantiles*, above or below which lie other fractional portions of the data.

For example, if the data are divided into four equal parts, we speak of *quartiles*. One-fourth of all the ranked observations are smaller than the first quartile, one-fourth lie between the first and second quartiles, one-fourth lie between the second and third quartiles, and one-fourth are larger than the third quartile. The second quartile is identical to the median. As with the median, the first and third quartiles might be one of the data or the midpoint between two of the data. The first quartile, Q_1, is

$$Q_1 = X_{(n+1)/4}; \tag{2}$$

if the subscript, $(n + 1)/4$, is not an integer or half-integer, then it is rounded up to the nearest integer or half-integer. The second quartile is the median, and the subscript on X for the third quartile, Q_3, is

$$n + 1 - (\text{subscript on } X \text{ for } Q_1, \text{ after any rounding}). \tag{3}$$

For species A, $n = 9$, $(n + 1)/4 = 2.5$, and $Q_1 = X_{2.5} = 34.5$ mo; and $Q_3 = X_{10-2.5} = X_{7.5} = 46$ mo. For species B, $n = 10$, $(n + 1)/4 = 2.75$ (which we round up to 3), and $Q_1 = X_3 = 38$ mo, and $Q_3 = X_{11-3} = X_8 = 59$ mo.

The distance between Q_1 and Q_3, the first and third quartiles (i.e., the 25th and 75th percentiles), is known as the *interquartile range* (or *semiquartile range*):

$$\text{interquartile range} = Q_3 - Q_1. \tag{4}$$

One may also encounter the *semi-interquartile range*:

$$\text{semi-interquartile range} = \frac{Q_3 - Q_1}{2}, \tag{5}$$

also known as the *quartile deviation*.*

If the distribution of data is symmetrical, then 50% of the measurements lie within one quartile deviation above and below the median. For Sample 1 in Example 1, $Q_1 = 1.4$ g, $Q_3 = 2.2$ g, and the interquartile range is 2.2 g − 1.4 g = 0.8 g. And for Sample 2, $Q_1 = 1.6$ g, $Q_3 = 2.0$ g, and the interquartile range is 2.0 g − 1.6 g = 0.4 g.

Similarly, values that partition the ordered data set into eight equal parts (or as equal as n will allow) are called *octiles*. The first octile, $\mathcal{O}_1$, is

$$\mathcal{O}_1 = X_{(n+1)/8}; \tag{6}$$

and if the subscript, $(n + 1)/8$, is not an integer or half-integer, then it is rounded up to the nearest integer or half-integer. The second, fourth, and sixth octiles are the same as quartiles; that is, $\mathcal{O}_2 = Q_1$, $\mathcal{O}_4 = Q_2 =$ median and $\mathcal{O}_6 = Q_3$. The subscript on X for the third octile, $\mathcal{O}_3$, is

$$2(\text{subscript on } X \text{ for } Q_1) - \text{subscript on } X \text{ for } \mathcal{O}_1; \tag{7}$$

the subscript on X for the fifth octile, $\mathcal{O}_5$, is

$$n + 1 - \text{subscript on } X \text{ for } \mathcal{O}_3; \tag{8}$$

and the subscript on X for the seventh octile, $\mathcal{O}_7$, is

$$n + 1 - \text{subscript on } X \text{ for } \mathcal{O}_1. \tag{9}$$

For species A, $n = 9$, $(n + 1)/8 = 1.5$ and $\mathcal{O}_1 = X_{1.5} = 35$ mo; $2(2.5) - 1.5 = 3.5$, so $\mathcal{O}_3 = X_{3.5} = 38$ mo; $n + 1 - 3.5 = 6.5$, so $\mathcal{O}_5 = X_{6.5} = 41.5$ mo; and $n + 1 - 1.5 = 8.5$, so $\mathcal{O}_7 = 61$. For species B, $n = 10$, $(n + 1)/8 = 1.25$ (which we round up to 1.5) and $\mathcal{O}_1 = X_{1.5} = 35$ mo; $2(3) - 1.5 = 4.5$, so $\mathcal{O}_3 = X_{4.5} = 39.5$ mo; $n + 1 - 4.5 = 6.5$, so $\mathcal{O}_5 = X_{6.5} = 41.5$ mo; and $n + 1 - 1.5 = 9.5$, so $\mathcal{O}_7 = 44.5$ mo.

Besides the median, quartiles, and octiles, ordered data may be divided into fifths, tenths, or hundredths by quantities that are respectively called *quintiles, deciles*, and *centiles* (the latter also called *percentiles*). Measures that divide a group of ordered data into equal parts are collectively termed *quantiles*.† The expression "LD_{50}," used in some areas of biological research, is simply the 50th percentile of the lethal doses, or the median lethal dose. That is, 50% of the experimental subjects survived this dose, whereas 50% did not. Likewise, "LC_{50}" is the median lethal concentration, or the 50th percentile of the lethal concentrations.

Instead of distance between the 25th and 75th percentiles, distances between other quantiles (e.g., 10th and 90th percentiles) may be used as a dispersion measure. Quantile-based measures of dispersion are valid for ordinal-, interval-, or ratio-scale data, and they do not exhibit the bias and inefficiency of the range.

*This measure was proposed in 1846 by L. A. J. Quetelet (1796–1874); Sir Francis Galton (1822–1911) later called it the "quartile deviation" (Walker, 1929: 84) and, in 1882, used the terms "quartile" and "interquartile range" (David, 1995).

†Sir Francis Galton developed the concept of percentiles, quartiles, deciles, and other quantiles in writings from 1869 to 1885 (Walker, 1929: 86–87, 177, 179). The term *quantile* was introduced in 1940 by M. G. Kendall (David, 1995).

3 THE MEAN DEVIATION

As is evident from the two samples in Example 1, the range conveys no information about how clustered about the middle of the distribution the measurements are. As the mean is so useful a measure of central tendency, one might express dispersion in terms of deviations from the mean. The sum of all deviations from the mean, that is, $\sum(X_i - \overline{X})$, will always equal zero, however, so such a summation would be useless as a measure of dispersion (as seen in Example 1).

Using the absolute values of the deviations from the mean eliminates the negative signs of the deviations, and summing those absolute values results in a quantity that is an expression of dispersion about the mean. Dividing this quantity by n yields a measure known as the *mean deviation*, or *mean absolute deviation*,* of the sample; this measure has the same units as do the data. In Example 1, Sample 1 is more variable (or more dispersed, or less concentrated) than Sample 2. Although the two samples have the same range, the mean deviations, calculated as

$$\text{sample mean deviation} = \frac{\sum |X_i - \overline{X}|}{n}, \tag{10}$$

express the differences in dispersion.† A different kind of mean deviation can be defined by using the sum of the absolute deviations from the median instead of from the mean.

Mean deviations are seldom encountered, because their utility is far less than that of the statistics in Sections 4 and 5.

4 THE VARIANCE

Another method of eliminating the negative signs of deviations from the mean is to square the deviations. The sum of the squares of the deviations from the mean is often simply called the *sum of squares*, abbreviated SS, and is defined as follows:‡

$$\text{population SS} = \sum(X_i - \mu)^2 \tag{11}$$

$$\text{sample SS} = \sum(X_i - \overline{X})^2. \tag{12}$$

It can be seen from the above two equations that as a measure of variability, or dispersion, the sum of squares considers how far the X_i's deviate from the mean. In

---

*The term *mean deviation* is apparently due to Karl Pearson (1857–1936) (Walker, 1929: 55) and *mean absolute deviation*, in 1972, to D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey (David, 1995).

†Karl Weierstrass, in 1841, was the first to denote the absolute value of a quantity by enclosing it within two vertical lines (Cajori, 1928/9, Vol. II: p. 123); that is, $|a| = a$ and $|-a| = a$.

‡The modern notation using raised numerals as exponents was introduced by René Descartes in 1637, and many other kinds of notation for exponents were employed before and after that (Cajori, 1928/9, Vol. I: 358; Gullberg, 1997: 134). An 1845 notation of Augustus De Morgan, $a \wedge b$ to indicate a^b (Cajori, ibid.: 358), has reemerged in modern computer use. Nicolas Chuquet (1445–1488) was the first to use negative exponents, and Nicole (also known as Nicolaus) Oresme (1323–1382) was the first to use fractional exponents, though neither of these French mathematicians employed the modern notation of Isaac Newton (1642–1727), the colossal English mathematician, physicist, and astronomer (Cajori, ibid.: 91, 102, 354–355):

$$X^{-a} = \frac{1}{X^a}; \quad X^{\frac{1}{a}} = \sqrt[a]{X}.$$

Using parentheses or brackets to group quantities dates from the mid-sixteenth century, though it was not common mathematical notation until more than two centuries later (ibid.: 392).

Sample 1 of Example 1, the sample mean is 1.8 g and it is seen (in the last column) that

$$\begin{aligned}\text{Sample SS} &= (1.2 - 1.8)^2 + (1.4 - 1.8)^2 + (1.6 - 1.8)^2 + (1.8 - 1.8)^2 \\ &\quad + (2.0 - 1.8)^2 + (2.2 - 1.8)^2 + (2.4 - 1.8)^2 \\ &= 0.36 + 0.16 + 0.04 + 0.00 + 0.04 + 0.16 + 0.36 \\ &= 1.12\end{aligned}$$

(where the units are grams^2).* The sum of squares may also be visualized as a measure of the average extent to which the data deviate from each other, for (using the same seven data from Sample 1 in Example 1):

$$\begin{aligned}\text{SS} &= [(1.2 - 1.4)^2 + (1.2 - 1.6)^2 + (1.2 - 1.8)^2 + (1.2 - 2.0)^2 \\ &\quad + (1.2 - 2.2)^2 + (1.2 - 2.4)^2 + (1.4 - 1.6)^2 + (1.4 - 1.8)^2 \\ &\quad + (1.4 - 2.0)^2 + (1.4 - 2.2)^2 + (1.4 - 2.4)^2 + (1.6 - 1.8)^2 \\ &\quad + (1.6 - 2.0)^2 + (1.6 - 2.2)^2 + (1.6 - 2.4)^2 + (1.8 - 2.0)^2 \\ &\quad + (1.8 - 2.2)^2 + (1.8 - 2.4)^2 + (2.0 - 2.2)^2 + (2.0 - 2.4)^2 \\ &\quad + (2.2 - 2.4)^2]/7 \\ &= [0.04 + 0.16 + 0.36 + 0.64 + 1.00 + 1.44 + 0.04 + \cdots + 0.04 + 0.16 \\ &\quad + 0.04]/7 \\ &= 7.84/7 = 1.12\end{aligned}$$

(again in grams^2).

The mean sum of squares is called the *variance* (or *mean square*,† the latter being short for *mean squared deviation*), and for a population is denoted by σ^2 ("sigma squared," using the lowercase Greek letter):

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}. \tag{14}$$

The best estimate of the population variance, σ^2, is the sample variance, s^2:

$$s^2 = \frac{\sum(X_i - \overline{X})^2}{n - 1}. \tag{15}$$

If, in Equation 14, we replace μ by $\overline{X}$ and N by n, the result is a quantity that is a biased estimate of σ^2 in that it underestimates σ^2. Dividing the sample sum of squares

*Owing to an important concept in statistics, known as *least squares*, the sum of squared deviations from the mean is smaller than the sum of squared deviations from any other quantity (e.g., the median). Indeed, if Equation 12 is applied using some quantity in place of the mean, the resultant "sum of squares" would be

$$SS + nd^2, \tag{13}$$

where d is the difference between the mean and the quantity used. For the population sum of squares (defined in Equation 11), the relationship would be $SS + Nd^2$.

†The term *mean square* dates back at least to an 1875 publication of Sir George Biddel Airy (1801–1892), Astronomer Royal of England (Walker, 1929: 54). The term *variance* was introduced in 1918 by English statistician Sir Ronald Aylmer Fisher (1890–1962) (ibid.: 189; David, 1995).

by $n - 1$ (called the *degrees of freedom*,* often abbreviated DF), rather than by n, yields an unbiased estimate, and it is Equation 15 that should be used to calculate the sample variance.

If all observations in a sample are equal, then there is no variability (that is, no dispersion) and $s^2 = 0$. And s^2 becomes increasingly large as the amount of variability, or dispersion, increases. Because s^2 is a mean sum of squares, it can never be a negative quantity.

The variance expresses the same type of information as does the mean deviation, but it has certain very important mathematical properties relative to probability and hypothesis testing that make it superior. Thus, the mean deviation is very seldom encountered in biostatistical analysis.

The calculation of s^2 can be tedious for large samples, but it can be facilitated by the use of the equality

$$\text{sample SS} = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}. \tag{16}$$

This formula is equivalent to Equation 12 but is much simpler to work with. Example 2 demonstrates its use to obtain a sample sum of squares.

Because the sample variance equals the sample SS divided by DF,

$$s^2 = \frac{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}{n - 1}. \tag{17}$$

This last formula is often referred to as a "working formula," or "machine formula," because of its computational advantages. There are, in fact, two major advantages in calculating SS by Equation 16 rather than by Equation 12. First, fewer computational steps are involved, a fact that decreases chance of error. On many calculators the summed quantities, $\sum X_i$ and $\sum X_i^2$, can both be obtained with only one pass through the data, whereas Equation 12 requires one pass through the data to calculate $\overline{X}$ and at least one more pass to calculate and sum the squares of the deviations, $X_i - \overline{X}$. Second, there may be a good deal of rounding error in calculating each $X_i - \overline{X}$, a situation that leads to decreased accuracy in computation, but that is avoided by the use of Equation 16.†

For data recorded in frequency tables,

$$\text{sample SS} = \sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{n}, \tag{18}$$

*Given the sample mean ($\overline{X}$) and sample size (n) in Example 1, *degrees of freedom* means that the data could have been weights different from those shown, but when any six (i.e., $n - 1$) of the seven weights are specified, then the seventh weight is also known. The term was first used, though in a different context, by Ronald Aylmer Fisher in 1922 (David, 1955).

†Computational formulas advantageous on calculators may not prove accurate on computers (Wilkinson and Dallal, 1977), largely because computers may use fewer significant figures. (Also see Ling, 1974.) Good computer programs use calculation techniques designed to help avoid rounding errors.

where f_i is the frequency of observations with magnitude X_i. But with a calculator or computer it is often faster to use Equation 18 for the individual observations, disregarding the class groupings.

The variance has square units. If measurements are in grams, their variance will be in grams squared, or if the measurements are in cubic centimeters, their variance will be in terms of cubic centimeters squared, even though such squared units have no physical interpretation.

EXAMPLE 2 "Machine Formula" Calculation of Variance, Standard Deviation, and Coefficient of Variation (These are the data of Example 1)

Sample 1

X_i (g)	X_i^2 (g^2)
1.2	1.44
1.4	1.96
1.6	2.56
1.8	3.24
2.0	4.00
2.2	4.84
2.4	5.76

$$\sum X_i = 12.6 \text{ g} \qquad \sum X_i^2 = 23.80 \text{ g}^2$$

$$n = 7$$

$$\overline{X} = \frac{12.6 \text{ g}}{7} = 1.8 \text{ g}$$

$$\begin{aligned} \text{SS} &= \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} \\ &= 23.80 \text{ g}^2 - \frac{(12.6 \text{ g})^2}{7} \\ &= 23.80 \text{ g}^2 - 22.68 \text{ g}^2 \\ &= 1.12 \text{ g}^2 \end{aligned}$$

$$\begin{aligned} s^2 &= \frac{\text{SS}}{n - 1} \\ &= \frac{1.12 \text{ g}^2}{6} = 0.1867 \text{ g}^2 \end{aligned}$$

$$s = \sqrt{0.1867 \text{ g}^2} = 0.43 \text{ g}$$

$$V = \frac{s}{\overline{X}} = \frac{0.43 \text{ g}}{1.8 \text{ g}} = 0.24 = 24\%$$

Sample 2

X_i (g)	X_i^2 (g^2)
1.2	1.44
1.6	2.56
1.7	2.89
1.8	3.24
1.9	3.61
2.0	4.00
2.4	5.76

$$\sum X_i = 12.6 \text{ g} \qquad \sum X_i^2 = 23.50 \text{ g}^2$$

$$n = 7$$

$$\overline{X} = \frac{12.6 \text{ g}}{7} = 1.8 \text{ g}$$

$$\begin{aligned} \text{SS} &= 23.50 \text{ g}^2 - \frac{(12.6 \text{ g})^2}{7} \\ &= 0.82 \text{ g}^2 \end{aligned}$$

$$s^2 = \frac{0.82 \text{ g}^2}{6} = 0.1367 \text{ g}^2$$

$$s = \sqrt{0.1367 \text{ g}^2} = 0.37 \text{ g}$$

$$V = \frac{0.37 \text{ g}}{1.8 \text{ g}} = 0.21 = 21\%$$

5 THE STANDARD DEVIATION

The *standard deviation** is the positive square root[†] of the variance; therefore, it has the same units as the original measurements. Thus, for a population,

$$\sigma = \sqrt{\frac{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{N}}{N}}, \tag{19}$$

And for a sample,[‡]

$$s = \sqrt{\frac{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}{n - 1}}. \tag{20}$$

Examples 1 and 2 demonstrate the calculation of s. This quantity frequently is abbreviated SD, and on rare occasions is called the *root mean square deviation* or *root mean square*. Remember that the standard deviation is, by definition, always a nonnegative quantity.[§]

---

*It was the great English statistician Karl Pearson (1857–1936) who coined the term *standard deviation* and its symbol, σ, in 1893, prior to which this quantity was called the *mean error* (Eells, 1926; Walker, 1929: 54–55, 183, 188). In early literature (e.g., by G. U. Yule in 1919), it was termed *root mean square deviation* and acquired the symbol s, and (particularly in the fields of education and psychology) it was occasionally computed using deviations from the median (or even the mode) instead of from the mean (Eells, 1926).

[†]The square root sign ($\surd$) was introduced by Silesian-born Austrian mathematician Christoff Rudolff (1499–1545) in 1525; by 1637 René Descartes (1596–1650) combined this with a vinculum (a horizontal bar placed above quantities to group them as is done with parentheses or brackets) to obtain the symbol $\sqrt{}$, but Gottfried Wilhelm Leibniz (1646–1716) preferred $\surd(\quad)$, which is still occasionally seen (Cajori, 1928/9, Vol. I: 135, 208, 368, 372, 375).

[‡]The sample s is actually a slightly biased estimate of the population σ, in that on the average it is a slightly low estimate, especially in small samples. But this fact is generally considered to be offset by the statistic's usefulness. Correction for this bias is sometimes possible (e.g., Bliss, 1967: 131; Dixon and Massey, 1969: 136; Gurland and Tripathi, 1971; Tolman, 1971), but it is rarely employed.

[§]It can be shown that the median of a distribution is never more than one standard deviation away from the mean (μ); that is,

$$|\text{median} - \mu| \le \sigma \tag{21}$$

(Hotelling and Solomon, 1932; O'Cinneide, 1990; Page and Murty, 1982; Watson, 1994). This is a special case, where $p = 50$, of the relationship

$$\mu - \sigma\sqrt{\frac{1 - p/100}{p/100}} \le X_p \le \mu + \sigma\sqrt{\frac{p/100}{1 - p/100}}, \tag{22}$$

where X_p is the pth percentile of the distribution (Dharmadhikari, 1991). Also, Page and Murty (1982) have shown these population-parameter relationships between the standard deviation and the range and between the standard deviation and the mean, median, and mode:

$$\text{range}/\sqrt{2n} \le \sigma \le \text{range}/2\,; \tag{4.22a}$$

$$|\text{mode} - \mu| \le \sigma\sqrt{n/m} \text{ and } |\text{mode} - \text{median}| \le \sigma(n/m), \tag{4.22b}$$

where m is the number of data at the modal value.

6 THE COEFFICIENT OF VARIATION

The *coefficient of variation** or *coefficient of variability*, is defined as

$$V = \frac{s}{\overline{X}} \quad \text{or} \quad V = \frac{s}{\overline{X}} \cdot 100\%. \tag{23}$$

As $s/\overline{X}$ is generally a small quantity, it is frequently multiplied by 100% in order to express V as a percentage. (The coefficient of variation is often abbreviated as CV.)

As a measure of variability, the variance and standard deviation have magnitudes that are dependent on the magnitude of the data. Elephants have ears that are perhaps 100 times larger than those of mice. If elephant ears were no more variable, relative to their size, than mouse ears, relative to their size, the standard deviation of elephant ear lengths would be 100 times as great as the standard deviation of mouse ear lengths (and the variance of the former would be $100^2 = 10{,}000$ times the variance of the latter). The sample coefficient of variation expresses sample variability relative to the mean of the sample (and is on rare occasion referred to as the "relative standard deviation"). It is called a measure of *relative variability* or *relative dispersion*.

Because s and $\overline{X}$ have identical units, V has no units at all, a fact emphasizing that it is a relative measure, divorced from the actual magnitude or units of measurement of the data. Thus, had the data in Example 2 been measured in pounds, kilograms, or tons, instead of grams, the calculated V would have been the same. The coefficient of variation of a sample, namely V, is an estimate of the coefficient of variation of the population from which the sample came (i.e., an estimate of σ/μ). The coefficient of variation may be calculated only for ratio scale data; it is, for example, not valid to calculate coefficients of variation of temperature data measured on the Celsius or Fahrenheit temperature scales. Simpson, Roe, and Lewontin (1960: 89–95) present a good discussion of V and its biological application, especially with regard to zoomorphological measurements.

7 INDICES OF DIVERSITY

For nominal-scale data there is no mean or median or ordered measurements to serve as a reference for discussion of dispersion. Instead, we can invoke the concept of *diversity*, the distribution of observations among categories. Consider that sparrows are found to nest in four different types of location (vines, eaves, branches, and cavities). If, out of twenty nests observed, five are found at each of the four locations, then we would say that there was great diversity in nesting sites. If, however, seventeen nests were found in cavities and only one in each of the other three locations, then we would consider the situation to be one of very low nest-site diversity. In other words, observations distributed evenly among categories display high diversity, whereas a set of observations where most of the data occur in very few of the categories is one exhibiting low diversity.

A large number of diversity measures have been introduced, especially for ecological data (e.g., Brower, Zar, and von Ende, 1998: 177–184; Magurran, 2004), a few of which are presented here.

*The term *coefficient of variation* was introduced by the statistical giant Karl Pearson (1857–1936) in 1896 (David, 1995). In early literature the term was variously applied to the ratios of different measures of dispersion and different measures of central tendency (Eells, 1926).

Among the quantitative descriptions of diversity available are those based on a field known as *information theory*.* The underlying considerations of these measures can be visualized by considering *uncertainty* to be synonymous with diversity. If seventeen out of twenty nest sites were to be found in cavities, then one would be relatively certain of being able to predict the location of a randomly encountered nest site. However, if nests were found to be distributed evenly among the various locations (a situation of high nest-site diversity), then there would be a good deal of uncertainty involved in predicting the location of a nest site selected at random. If a set of nominal scale data may be considered to be a random sample, then a quantitative expression appropriate as a measure of diversity is that of Shannon (1948):

$$H' = -\sum_{i=1}^{k} p_i \log\, p_i \tag{24}$$

(often referred to as the Shannon-Wiener diversity index or the Shannon-Weaver index). Here, k is the number of categories and p_i is the proportion of the observations found in category i. Denoting n to be sample size and f_i to be the number of observations in category i, then $p_i = f_i/n$; and an equivalent equation for H' is

$$H' = \frac{n \log n \; - \; \sum_{i=1}^{k} f_i \log\, f_i}{n}, \tag{25}$$

a formula that is easier to use than Equation 24 because it eliminates the necessity of calculating the proportions (p_i). Published tables of $n \log n$ and $f_i \log f_i$ are available (e.g., Brower, Zar, and von Ende, 1998: 181; Lloyd, Zar, and Karr, 1968). Any logarithmic base may be used to compute H'; bases 10, e, and 2 (in that order of commonness) are the most frequently encountered. A value of H' (or of any other measure of this section except evenness measures) calculated using one logarithmic base may be converted to that of another base; Table 1 gives factors for doing this for bases 10, e, and 2. Unfortunately, H' is known to be an underestimate of the diversity in the sampled population (Bowman et al., 1971). However, this bias decreases with increasing sample size. Ghent (1991) demonstrated a relationship between H' and testing hypotheses for equal abundance among the k categories.

The magnitude of H' is affected not only by the distribution of the data but also by the number of categories, for, theoretically, the maximum possible diversity for a set of data consisting of k categories is

$$H'_{\max} = \log k. \tag{26}$$

Therefore, some users of Shannon's index prefer to calculate

$$J' = \frac{H'}{H'_{\max}} \tag{27}$$

instead of (or in addition to) H', thus expressing the observed diversity as a proportion of the maximum possible diversity. The quantity J' has been termed *evenness* (Pielou, 1966) and may also be referred to as *homogeneity* or *relative diversity*. The measure

*Claude Elwood Shannon (1916–2001) founded what he first called "a mathematical theory of communication" and has become known as "information theory."

TABLE 1: Multiplication Factors for Converting among Diversity Measures (H, H', H_{max}, or H'_{max}) Calculated Using Different Logarithmic Bases*

To convert to:	To convert from:		
	Base 2	**Base *e***	**Base 10**
Base 2	1.0000	1.4427	3.3219
Base *e*	0.6931	1.0000	2.3026
Base 10	0.3010	0.4343	1.0000

For example, if $H' = 0.255$ using base 10; H' would be $(0.255)(3.3219) = 0.847$ using base 2.

*The measures J and J' are unaffected by change in logarithmic base.

$1 - J'$ may then be viewed as a measure of *heterogeneity*; it may also be considered a measure of *dominance*, for it reflects the extent to which frequencies are concentrated in a small number of categories. The number of categories in a sample (k) is typically an underestimate of the number of categories in the population from which the sample came, because some categories (especially the rarer ones) are likely to be missed in collecting the sample. Therefore, the sample evenness, J', is typically an overestimate of the population evenness. (That is, J' is a biased statistic.) Example 3 demonstrates the calculation of H' and J'.

If a set of data may not be considered a random sample, then Equation 24 (or 25) is not an appropriate diversity measure (Pielou, 1966). Examples of such

EXAMPLE 3 Indices of Diversity for Nominal Scale Data: The Nesting Sites of Sparrows

Category (i)	**Observed Frequencies (f_i)**
	Sample 1
Vines	5
Eaves	5
Branches	5
Cavities	5

$$H' = \frac{n\log n - \sum f_i \log f_i}{n} = [20\log 20 - (5\log 5 + 5\log 5 + 5\log 5 + 5\log 5)]/20$$

$$= [26.0206 - (3.4949 + 3.4949 + 3.4949 + 3.4949)]/20$$

$$= 12.0410/20 = 0.602$$

$$H'_{max} = \log 4 = 0.602$$

$$J' = \frac{0.602}{0.602} = 1.00$$

Sample 2	
Vines	1
Eaves	1
Branches	1
Cavities	17

$$H' = \frac{n \log n - \sum f_i \log f_i}{n} = [20 \log 20 - (1 \log 1 + 1 \log 1 + 1 \log 1 + 17 \log 17)]/20$$
$$= [26.0206 - (0 + 0 + 0 + 20.9176)]/20$$
$$= 5.1030/20 = 0.255$$
$$H'_{\max} = \log 4 = 0.602$$
$$J' = \frac{0.255}{0.602} = 0.42$$

Sample 3	
Vines	2
Eaves	2
Branches	2
Cavities	34

$$H' = \frac{n \log n - \sum f_i \log f_i}{n} = [40 \log 40 - (2 \log 2 + 2 \log 2 + 2 \log 2 + 34 \log 34)]/40$$
$$= [64.0824 - (0.6021 + 0.6021 + 0.6021 + 52.0703)]/40$$
$$= 10.2058/40 = 0.255$$
$$H'_{\max} = \log 4 = 0.602$$
$$J' = \frac{0.255}{0.602} = 0.42$$

situations may be when we have, in fact, data composing an entire population, or data that are a sample obtained nonrandomly from a population. In such a case, one may use the information-theoretic diversity measure of Brillouin (1962: 7–8):*

$$H = \frac{\log\left(\dfrac{n!}{\prod_{i=1}^{k} f_i!}\right)}{n}, \tag{28}$$

*The notation $n!$ is read as "n factorial" and signifies the product $(n)(n-1)(n-2)\cdots(2)(1)$. It was proposed by French physician and mathematician Christian Kramp (1760–1826) around 1798; he originally called this function *faculty* ("facultés" in French) but in 1808 accepted the term *factorial* ("factorielle" in French) used by Alsatian mathematician Louis François Antoine Arbogast (1759–1803) (Cajori, 1928/9, Vol. II: 72; Gullberg, 1997: 106; Miller, 2004a; O'Connor and Robertson, 1997). English mathematician Augustus De Morgan (1806–1871) decried the adoption of this symbol as a "barbarism" because it introduced into mathematics a symbol that already had an established meaning in written language, thus giving "the appearance of expressing surprise or admiration" in a mathematical result (Cajori, ibid.: 328).

where Π (capital Greek pi) means to take the product, just as Σ means to take the sum. Equation 28 may be written, equivalently, as

$$H = \frac{\log \dfrac{n!}{f_1!f_2!\ldots f_k!}}{n} \tag{29}$$

or as

$$H = \frac{(\log n! - \sum \log f_i!)}{n}. \tag{30}$$

Table 40 gives logarithms of factorials to ease this calculation. Other such tables are available, as well (e.g., Brower, Zar, and von Ende 1998: 183; Lloyd, Zar, and Karr, 1968; Pearson and Hartly, 1966: Table 51).* Ghent (1991) discussed the relationship between H and the test of hypotheses about equal abundance among k categories.

The maximum possible Brillouin diversity for a set of n observations distributed among k categories is

$$H_{\max} = \frac{\log n! - (k - d)\log c! - d\log(c + 1)!}{n}, \tag{35}$$

where c is the integer portion of n/k, and d is the remainder. (For example, if $n = 17$ and $k = 4$, then $n/k = 17/4 = 4.25$ and $c = 4$ and $d = 0.25$.) The Brillouin-based evenness measure is, therefore,

$$J = \frac{H}{H_{\max}}, \tag{36}$$

with $1 - J$ being a dominance measure. When we consider that we have data from an entire population, k is a population measurement, rather than an estimate of one, and J is not a biased estimate as is J'.

For further considerations of these and other diversity measures, see Brower, Zar, and von Ende (1998: Chapter 5B) and Magguran (2004: 100–121).

8 CODING DATA

Coding data may facilitate statistical computations of measures of central tendency. Such benefits are even more apparent when calculating SS, s^2, and s, because of the

*For moderate to large n (or f_i), "Stirling's approximation" is excellent:

$$n! = \sqrt{2\pi n}(n/e)^n = \sqrt{2\pi}\sqrt{n}e^{-n}n^n, \tag{31}$$

of which this is an easily usable derivation:

$$\log n! = (n + 0.5)\log n - 0.434294n + 0.399090. \tag{32}$$

An approximation with only half the error of the above is

$$n! = \sqrt{2\pi}\left(\frac{n + 0.5}{e}\right)^{n+0.5} \tag{33}$$

and

$$\log n! = (n + 0.5)\log(n + 0.5) - 0.434294(n + 0.5) + 0.399090. \tag{34}$$

This is named for James Stirling, who published something similar to the latter approximation formula in 1730, making an arithmetic improvement in the approximation earlier known by Abraham de Moivre (Kemp, 1989; Pearson, 1924; Walker, 1929: 16).

labor, and concomitant chances of error, involved in the unwieldy squaring of large or small numbers.

When data are coded by adding or subtracting a constant (call it A), the measures of dispersion of Sections 1 through 5 are not changed from what they were for the data before coding. This is because these measures are based upon deviations, and deviations are not changed by moving the data along the measurement scale (e.g., the deviation between 1 and 10 is the same as the deviation between 11 and 20). Sample 1 in Example 4 demonstrates this.

However, when coding by multiplying by a constant (call it M), the measures of dispersion are affected, for the magnitudes of the deviations will be changed. With such coding, the range, mean deviation, and standard deviation are changed by a factor of M, in the same manner as the arithmetic mean and the median are, whereas the sum of squares and variance are changed in accordance with the square of the coding constant (i.e., M^2), and the coefficient of variance is not affected. This is demonstrated in Sample 2 of Example 4.

A coded datum is described as

$$[X_i] = MX_i + A. \tag{37}$$

EXAMPLE 4 Coding Data to Facilitate the Calculation of Measures of Dispersion

Sample 1 (Coding by Subtraction: $A = -840$ g)

Without Coding X_i		*Using Coding* $[X_i]$	
X_i (g)	X_i^2 (g^2)	$[X_i]$ (g)	$[X_i]^2$ (g^2)
842	708,964	2	4
843	710,649	3	9
844	712,336	4	16
846	715,716	6	36
846	715,716	6	36
847	717,409	7	49
848	719,104	8	64
849	720,801	9	81
$\sum X_i = 6765$ g	$\sum X_i^2 = 5{,}720{,}695$ g^2	$\sum[X_i] = 45$ g	$\sum[X_i]^2 = 295$ g^2

$$s^2 = \frac{5720695 \text{ g}^2 - \dfrac{(6765 \text{ g})^2}{8}}{7}$$
$$= 5.98 \text{ g}^2$$
$$s = 2.45 \text{ g}$$
$$\overline{X} = 845.6 \text{ g}$$
$$V = \frac{s}{\overline{X}} = \frac{2.45 \text{ g}}{845.6 \text{ g}}$$
$$= 0.0029 = 0.29\%$$

$$[s^2] = \frac{295 \text{ g}^2 - \dfrac{(45 \text{ g})^2}{8}}{7}$$
$$= 5.98 \text{ g}^2$$
$$[s] = 2.44 \text{ g}$$
$$[\overline{X}] = 5.6 \text{ g}$$

Sample 2 (Coding by Division: $M = 0.01$)			
Without Coding X_i		*Using Coding $[X_i]$*	
X_i (sec)	X_i^2 (sec^2)	$[X_i]$ (sec)	$[X_i]^2$ (sec^2)
800	640,000	8.00	64.00
900	810,000	9.00	81.00
950	902,500	9.50	90.25
1100	1,210,000	11.00	121.00
1250	1,562,500	12.50	156.25
1300	1,690,000	13.00	169.00
$\sum X_i = 6300$ sec	$\sum X_i^2 = 6{,}815{,}000$ sec^2	$\sum [X_i] = 63.00$ sec	$\sum [X_i]^2 = 681.50$ sec^2

Without coding:

$$s^2 = \frac{6815000 \text{ sec}^2 - \dfrac{(6300 \text{ sec})^2}{6}}{5}$$

$$= 40{,}000 \text{ sec}^2$$

$$s = 200 \text{ sec}$$

$$\overline{X} = 1050 \text{ sec}$$

$$V = 0.19 = 19\%$$

Using coding:

$$[s^2] = \frac{681.50 \text{ sec}^2 - \dfrac{(63.00 \text{ sec})^2}{6}}{5}$$

$$= 4 \text{ sec}^2$$

$$[s] = 2.00 \text{ sec}$$

$$[\overline{X}] = 10.50 \text{ sec}$$

$$[V] = 0.19 = 19\%$$

EXERCISES

1. Five body weights, in grams, collected from a population of rodent body weights are

66.1, 77.1, 74.6, 61.8, 71.5.

(a) Compute the "sum of squares" and the variance of these data using Equations 12 and 15, respectively.
(b) Compute the "sum of squares" and the variance of these data by using Equations 16 and 17, respectively.

2. Consider the following data, which are a sample of amino acid concentrations (mg/100 ml) in arthropod hemolymph:

240.6, 238.2, 236.4, 244.8, 240.7, 241.3, 237.9.

(a) Determine the range of the data.
(b) Calculate the "sum of squares" of the data.
(c) Calculate the variance of the data.
(d) Calculate the standard deviation of the data.
(e) Calculate the coefficient of variation of the data.

3. The following frequency distribution of tree species was observed in a random sample from a forest:

Species	*Frequency*
White oak	44
Red oak	3
Shagbark hickory	28
Black walnut	12
Basswood	2
Slippery elm	8

(a) Use the Shannon index to express the tree species diversity.
(b) Compute the maximum Shannon diversity possible for the given number of species and individuals.
(c) Calculate the Shannon evenness for these data.

4. Assume the data in Exercise 3 were an entire population (e.g., all the trees planted around a group of buildings).
(a) Use the Brillouin index to express the tree species diversity.
(b) Compute the maximum Brillouin diversity possible for the given number of species and individuals.
(c) Calculate the Brillouin evenness measure for these data.

ANSWERS TO EXERCISES

1. **(a)** SS $= 156.028\ \text{g}^2$, $s^2 = 39.007\ \text{g}^2$; **(b)** same as (a).

2. **(a)** Range $= 236.4$ mg/100 ml to 244.8 mg/100 ml $= 8.4$ mg/100 ml; **(b)** SS $= 46.1886$ (mg/100 ml)^2; **(c)** $s^2 = 7.6981$ (mg/100 ml)^2; **(d)** $s = 2.77$ mg/100 ml; **(e)** $V = 0.0115 = 1.15\%$.
3. $k = 6, n = 97$; **(a)** $H' = 0.595$; **(b)** $H'_{\text{max}} = 0.778$; **(c)** $J' = 0.76$.
4. $k = 6, n = 97$; **(a)** $H = 0.554$; **(b)** $c = 16$, $d = 0.1667$, $H_{\text{max}} = 0.741$; **(c)** $J = 0.75$.

This page intentionally left blank

Probabilities

Everyday concepts of "likelihood," "predictability," and "chance" are formalized by that branch of mathematics called *probability*. Although earlier work on the subject was done by writers such as Giralamo Cardano (1501–1576) and Galileo Galilei (1564–1642), the investigation of probability as a branch of mathematics sprang in earnest from 1654 correspondence between two great French mathematicians, Blaise Pascal (1623–1662) and Pierre Fermat (1601–1665). These two men were stimulated by the desire to predict outcomes in the games of chance popular among the French nobility of the mid-seventeenth century; we still use the devices of such games (e.g., dice and cards) to demonstrate the basic concepts of probability.*

A thorough discourse on probability is well beyond the scope and intent of this text, but aspects of probability are of biological interest and considerations of probability theory underlie the many procedures for statistical hypothesis testing discussed. Therefore, this chapter will introduce probability concepts that bear the most pertinence to biology and biostatistical analysis.

Worthwhile presentations of probability specifically for the biologist are found in Batschelet (1976: 441–474); Eason, Coles, and Gettinby (1980: 395–414); and Mosimann (1968).

1 COUNTING POSSIBLE OUTCOMES

Suppose a phenomenon can occur in any one of k different ways, but in only one of those ways at a time. For example, a coin has two sides and when tossed will land

*The first published work on the subject of probability and gaming was by the Dutch astronomer, physicist, and mathematician Christiaan (also known as Christianus) Huygens (1629–1695), in 1657 (Asimov, 1982: 138; David, 1962: 113, 133). This, in turn, aroused the interest of other major minds, such as Jacob (also known as Jacques, Jakob, and James) Bernoulli (1654–1705, whose 1713 book was the first devoted entirely to probability), several other members of the remarkable Bernoulli family of Swiss mathematicians, and others such as Abraham de Moivre (1667–1754), Pierre Rémond de Montmort (1678–1719), and Pierre-Simon Laplace (1749–1827) of France. The term *probability* in its modern mathematical sense was used as early as 1718 by de Moivre (Miller, 2004a). For more detailed history of the subject, see David (1962) and Walker (1928: 5–13).

From Chapter 5 of *Biostatistical Analysis*, Fifth Edition, Jerrold H. Zar. Copyright © 2010 by Pearson Education, Inc. Publishing as Pearson Prentice Hall. All rights reserved.

with either the "head" side (H) up or the "tail" side (T) up, but not both. Or, a die has six sides and when thrown will land with either the 1, 2, 3, 4, 5, or 6 side up.* We shall refer to each possible outcome (i.e., H or T with the coin; or 1, 2, 3, 4, 5, or 6 with the die) as an *event*.

If something can occur in any one of k_1 different ways and something else can occur in any one of k_2 different ways, then the number of possible ways for both things to occur is $k_1 \times k_2$. For example, suppose that two coins are tossed, say a silver one and a copper one. There are two possible outcomes of the toss of the silver coin (H or T) and two possible outcomes of the toss of the copper coin (H or T). Therefore, $k_1 = 2$ and $k_2 = 2$ and there are $(k_1)(k_2) = (2)(2) = 4$ possible outcomes of the toss of both coins: both heads, silver head and copper tail, silver tail and copper head, and both tails (i.e., H,H; H,T; T,H; T,T).

Or, consider tossing of a coin together with throwing a die. There are two possible coin outcomes ($k_1 = 2$) and six possible die outcomes ($k_2 = 6$), so there are $(k_1)(k_2) = (2)(6) = 12$ possible outcomes of the two events together:

H,1; H,2; H,3; H,4; H,5; H,6; T,1; T,2; T,3; T,4; T,5; T,6.

If two dice are thrown, we can count six possible outcomes for the first die and six for the second, so there are $(k_1)(k_2) = (6)(6) = 36$ possible outcomes when two dice are thrown:

1,1; 1,2; 1,3; 1,4; 1,5; 1,6; 2,1; 2,2; 2,3; 2,4; 2,5; 2,6;
3,1; 3,2; 3,3; 3,4; 3,5; 3,6; 4,1; 4,2; 4,3; 4,4; 4,5; 4,6;
5,1; 5,2; 5,3; 5,4; 5,5; 5,6; 6,1; 6,2; 6,3; 6,4; 6,5; 6,6.

The preceding counting rule is extended readily to determine the number of ways more than two things can occur together. If one thing can occur in any one of k_1 ways, a second thing in any one of k_2 ways, a third thing in any of k_3 ways, and so on, through an nth thing in any one of k_n ways, then the number of ways for all n things to occur together is

$$(k_1)(k_2)(k_3)\cdots(k_n).$$

Thus, if three coins are tossed, each toss resulting in one of two possible outcomes, then there is a total of

$$(k_1)(k_2)(k_3) = (2)(2)(2) = 8$$

possible outcomes for the three tosses together:

H,H,H; H,H,T; H,T,H; H,T,T; T,H,H, T,H,T; T,T,H; T,T,T.

Similarly, if three dice are thrown, there are $(k_1)(k_2)(k_3) = (6)(6)(6) = 6^3 = 216$ possible outcomes; if two dice and three coins are thrown, there are

*What we recognize as metallic coins originated shortly after 650 B.C.E.—perhaps in ancient Lydia (located on the Aegean Sea in what is now western Turkey). From the beginning, the obverse and reverse sides of coins have had different designs, in earliest times with the obverse commonly depicting animals and, later, deities and rulers (Sutherland, 1992). Dice have long been used for both games and religion. They date from nearly 3000 years B.C.E., with the modern conventional arrangement of dots on the six faces of a cubic die (1 opposite 6, 2 opposite 5, and 3 opposite 4) becoming dominant around the middle of the fourteenth century B.C.E. (David, 1962: 10). Of course, the arrangement of the numbers 1 through 6 on the six faces has no effect on the outcome of throwing a die.

$(k_1)(k_2)(k_3)(k_4)(k_5) = (6)(6)(2)(2)(2) = (6^2)(2^3) = 288$ outcomes; and so on. Example 1 gives two biological examples of counting possible outcomes.

EXAMPLE 1 Counting Possible Outcomes

(a) A linear arrangement of three deoxyribonucleic acid (DNA) nucleotides is called a triplet. A nucleotide may contain any one of four possible bases: adenine (A), cytosine (C), guanine (G), and thymine (T). How many different triplets are possible?

As the first nucleotide in the triplet may be any one of the four bases (A; C; G; T), the second may be any one of the four, and the third may be any one of the four, there is a total of

$$(k_1)(k_2)(k_3) = (4)(4)(4) = 64 \text{ possible outcomes;}$$

that is, there are 64 possible triplets:

A, A, A; A, A, C; A, A, G; A, A, T;
A, C, A; A, C, C; A, C, G; A, C, T;
A, G, A; A, G, C; A, G, G; A, G, T;
and so on.

(b) If a diploid cell contains three pairs of chromosomes, and one member of each pair is found in each gamete, how many different gametes are possible?

As the first chromosome may occur in a gamete in one of two forms, as may the second and the third chromosomes,

$$(k_1)(k_2)(k_3) = (2)(2)(2) = 2^3 = 8.$$

Let us designate one of the pairs of chromosomes as "long," with the members of the pair being L_1 and L_2; one pair as "short," indicated as S_1 and S_2; and one pair as "midsized," labeled M_1 and M_2. Then the eight possible outcomes may be represented as

$$\begin{array}{llll} L_1, M_1, S_1; & L_1, M_1, S_2; & L_1, M_2, S_1; & L_1, M_2, S_2; \\ L_2, M_1, S_1; & L_2, M_1, S_2; & L_2, M_2, S_1; & L_2, M_2, S_2. \end{array}$$

2 PERMUTATIONS

(a) Linear Arrangements. A *permutation** is an arrangement of objects in a specific sequence. For example, a horse (H), cow (C), and sheep (S) could be arranged linearly in six different ways: H,C,S; H,S,C; C,H,S; C,S,H; S,H,C; S,C,H. This set of outcomes may be examined by noting that there are three possible ways to fill the first position in the linear order; but once an animal is placed in this position, there are only two ways to fill the second position; and after animals are placed in the first two positions, there is only one possible way to fill the third position. Therefore, $k_1 = 3$, $k_2 = 2$, and $k_3 = 1$, so that by the method of counting of Section 1 there are $(k_1)(k_2)(k_3) = (3)(2)(1) = 6$ ways to align these three animals. We may say that there are six permutations of three distinguishable objects.

*The term *permutation* was invented by Jacob Bernoulli in his landmark posthumous 1713 book on probability (Walker, 1929: 9).

In general, if there are n linear positions to fill with n objects, the first position may be filled in any one of n ways, the second may be filled in any one of $n - 1$ ways, the third in any one of $n - 2$ ways, and so on until the last position, which may be filled in only one way. That is, the filling of n positions with n objects results in ${}_nP_n$ permutations, where

$$ {}_nP_n = n(n - 1)(n - 2) \cdots (3)(2)(1). \tag{1} $$

This equation may be written more simply in *factorial* notation as

$$ {}_nP_n = n!, \tag{2} $$

where "n factorial" is the product of n and each smaller positive integer; that is,

$$ n! = n(n - 1)(n - 2) \cdots (3)(2)(1). \tag{3} $$

Example 2 demonstrates such computation of the numbers of permutations.

EXAMPLE 2 The Number of Permutations of Distinct Objects

In how many sequences can six photographs be arranged on a page?

$$ {}_nP_n = 6! = (6)(5)(4)(3)(2)(1) = 720 $$

(b) Circular Arrangements. The numbers of permutations considered previously are for objects arranged on a line. If objects are arranged on a circle, there is no "starting position" as there is on a line, and the number of permutations is

$$ {}_nP'_n = \frac{n!}{n} = (n - 1)!. \tag{4} $$

(Observe that the notation ${}_nP'_n$ is used here for circular permutations to distinguish it from the symbol ${}_nP_n$ used for linear permutations.)

Referring again to a horse, a cow, and a sheep, there are ${}_nP'_n = \frac{n!}{n} = (n - 1)! = (3 - 1)! = 2! = 2$ distinct ways in which the three animals could be seated around a table, or arranged around the shore of a pond:

	H				H	
			or			
S		C		C		S

In this example, there is an assumed orientation of the observer, so clockwise and counterclockwise patterns are treated as different. That is, the animals are observed arranged around the top of the table, or observed from above the surface of the pond. But either one of these arrangements would look like the other one if observed from under the table or under the water; and if we did not wish to count the results of these two mirror-image observations as different, we would speak of there being one possible permutation, not two. For example, consider each of the preceding two diagrams to represent three beads on a circular string, one bead in the shape of a horse, one in the shape of a cow, and the other in the shape of a sheep. The two arrangements of H, C, and S shown are not really different, for there is no specific way of viewing the circle; one of the two arrangements turns into the other if the circle is turned over. If $n > 2$ and the orientation of the circle is not specified, then the number of permutations of n objects on a circle is

$$_nP''_n = \frac{n!}{2n} = \frac{(n - 1)!}{2}. \quad (5)$$

(c) Fewer than *n* Positions. If one has n objects, but fewer than n positions in which to place them, then there would be considerably fewer numbers of ways to arrange the objects than in the case where there are positions for all n. For example, there are $_4P_4 = 4! = (4)(3)(2)(1) = 24$ ways of placing a horse (H), cow (C), sheep (S), and pig (P) in four positions on a line. However, there are only twelve ways of linearly arranging these four animals two at a time:

H,C; H,S; H,P; C,H; C,S; C,P; S,H; S,C; S,P; P,H; P,C; P,S.

The number of linear permutations of n objects taken X at a time is*

$$_nP_X = \frac{n!}{(n - X)!}. \quad (6)$$

For the preceding example,

$$_4P_2 = \frac{4!}{(4 - 2)!} = \frac{4!}{2!} = \frac{(4)(3)(2)(1)}{(2)(1)} = 12.$$

Equation 2 is a special case of Equation 6, where $X = n$; it is important to know that 0! is defined to be 1.†

If the arrangements are circular, instead of linear, then the number of them possible is

$$_nP'_X = \frac{n!}{(n - X)!X}. \quad (7)$$

So, for example, there are only $4!/[(4 - 2)!2] = 6$ different ways of arranging two out of our four animals around a table:

H	H	H	C	C	S
C	S	P	S	P	P

for C seated at the table opposite H is the same arrangement as H seated across from C, S seated with H is the same as H with S, and so on. Example 3 demonstrates this further. Equation 4 is a special case of Equation 7, where $X = n$; and recall that 0! is defined as 1.

EXAMPLE 3 The Number of Permutations of *n* Objects Taken *X* at a Time: In How Many Different Ways Can a Sequence of Four Slides Be Chosen from a Collection of Six Slides?

$$_nP_X = {}_6P_4 = \frac{6!}{(6 - 4)!} = \frac{6!}{2!} = \frac{(6)(5)(4)(3)(2)(1)}{(2)(1)}$$
$$= (6)(5)(4)(3) = 360$$

*Notation in the form of $_nP_X$ to indicate permutations of n items taken X at a time was used prior to 1869 by Harvey Goodwin (Cajori, 1929: 79).

†Why is 0! defined to be 1? In general, $n! = n[(n - 1)!]$; for example, $5! = 5(4!)$, $4! = 4(3!)$, $3! = 3(2!)$, and $2! = 2(1!)$. Thus, $1! = 1(0!)$, which is so only if $0! = 1$.

If $n > 2$, then for every circular permutation viewed from above there is a mirror image of that permutation, which would be observed from below. If these two mirror images are not to be counted as different (e.g., if we are dealing with beads of different shapes or colors on a string), then the number of circular permutations is

$$_nP''_X = \frac{n!}{2(n - X)!X}. \tag{8}$$

(d) If Some of the Objects Are Indistinguishable. If our group of four animals consisted of two horses (H), a cow (C), and a sheep (S), the number of permutations of the four animals would be twelve:

H,H,C,S; H,H,S,C; H,C,H,S; H,C,S,H; H,S,H,C; H,S,C,H;
C,H,H,S; C,H,S,H; C,S,H,H; S,H,H,C; S,H,C,H; S,C,H,H.

If n_i represents the number of like individuals in category i (in this case the number of animals in species i), then in this example $n_1 = 2$, $n_2 = 1$, and $n_3 = 1$, and we can write the number of permutations as

$$_nP_{n_1,n_2,n_3} = \frac{n!}{n_1!n_2!n_3!} = \frac{4!}{2!1!1!} = 12.$$

If the four animals were two horses (H) and two cows (C), then there would be only six permutations:

H,H,C,C; C,C,H,H; H,C,H,C; C,H,C,H; H,C,C,H; C,H,H,C.

In this case, $n = 4, n_1 = 2$, and $n_2 = 2$, and the number of permutations is calculated to be $_nP_{n_1,n_2} = n!/(n_1!n_2!) = 4!/(2!2!) = (4)(3)(2)/[(2)(2)] = 6$.

In general, if n_1 members of the first category of objects are indistinguishable, as are n_2 of the second category, n_3 of the third category, and so on through n_k members of the kth category, then the number of different permutations is

$$_nP_{n_1,n_2,\cdots,n_k} = \frac{n!}{n_1!n_2!\cdots n_k!} \text{ or } \frac{n!}{\prod_{i=1}^{k} n_i!}, \tag{9}$$

where the capital Greek letter pi (Π) denotes taking the product just as the capital Greek sigma (Σ) indicates taking the sum. This is shown further in Example 4.

EXAMPLE 4 Permutations with Categories Containing Indistinguishable Members

There are twelve potted plants, six of one species, four of a second species, and two of a third species. How many different linear sequences of species are possible (for example, if arranging the pots on a shelf)?

$$
\begin{aligned}
_nP_{n_1,n_2,n_3} &= \frac{n}{\prod n_i!} \\
= {}_{12}P_{6,4,2} &= \frac{12!}{6!4!2!} \\
&= \frac{(12)(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}{(6)(5)(4)(3)(2)(1)(4)(3)(2)(1)(2)(1)} = 13{,}860.
\end{aligned}
$$

Note that the above calculation could have been simplified by writing

$$\frac{12!}{6!4!2!} = \frac{(12)(11)(10)(9)(8)(7)6!}{6!(4)(3)(2)(2)} = \frac{(12)(11)(10)(9)(8)(7)}{(4)(3)(2)(2)} = 13{,}860.$$

Here, "(1)" is dropped; also, "6!" appears in both the numerator and denominator, thus canceling out.

3 COMBINATIONS

In Section 2 we considered groupings of objects where the sequence within the groups was important. In many instances, however, only the components of a group, not their arrangement within the group, are important. We saw that if we select two animals from among a horse (H), cow (C), sheep (S), and pig (P), there are twelve ways of arranging the two on a line:

H,C; H,S; H,P; C,H; C,S; C,P; S,H; S,C; S,P; P,H; P,C; P,S.

However, some of these arrangements contain exactly the same kinds of animals, only in different order (e.g., H,C and C,H; H,S and S,H). If the groups of two are important to us, but not the sequence of objects within the groups, then we are speaking of *combinations*,* rather than permutations. Designating the number of combinations of n objects taken X at a time as ${}_nC_X$, we have†

$$ {}_nC_X = \frac{{}_nP_X}{X!} = \frac{n!}{X!(n-X)!}. \tag{10}$$

So for the present example, $n = 4$, $\overline{X} = 2$, and

$$ {}_4C_2 = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = \frac{(4)(3)(2)(1)}{(2)(1)(2)(1)} = \frac{(4)(3)}{2} = 6,$$

the six combinations of the four animals taken two at a time being

H,C; H,S; H,P; C,S; C,P; S,P.

Example 5 demonstrates the determination of numbers of combinations for another set of data.

It may be noted that

$$ {}_nC_n = 1, \tag{11}$$

meaning that there is only one way of selecting all n items; and

$$ {}_nC_1 = n, \tag{12}$$

indicating that there are n ways of selecting n items one at a time. Also,

$$ {}_nC_X = {}_nC_{n-X}, \tag{13}$$

*The word *combination* was used in this mathematical sense by Blaise Pascal (1623–1662) in 1654 (Smith, 1953: 528).

†Notation in the form of ${}_nC_X$ to indicate combinations of n items taken X at a time was used by G. Chrystal in 1899 (Cajori, 1929: 80).

EXAMPLE 5 Combinations of n Objects Taken X at a Time

Of a total of ten dogs, eight are to be used in a laboratory experiment. How many different combinations of eight animals may be formed from the ten?

$$\begin{aligned} {}_nC_X = {}_{10}C_8 &= \frac{10!}{8!(10-8)!} = \frac{10!}{8!2!} = \frac{(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}{(8)(7)(6)(5)(4)(3)(2)(1)(2)(1)} \\ &= 45. \end{aligned}$$

It should be noted that the above calculations with factorials could have been simplified by writing

$${}_{10}C_8 = \frac{10!}{8!2!} = \frac{(10)(9)8!}{8!2!} = \frac{(10)(9)}{2} = 45,$$

so that "8!" appears in both the numerator and denominator, thus canceling each other out.

which means that if we select X items from a group of n, we have at the same time selected the remaining $n - X$ items; that is, an exclusion is itself a selection. For example, if we selected two out of five persons to write a report, we have simultaneously selected three of the five to refrain from writing. Thus,

$${}_5C_2 = \frac{5!}{2!(5-2)!} = \frac{5!}{2!3!} = 10 \quad \text{and} \quad {}_5C_{5-2} = {}_5C_3 = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = 10,$$

meaning that there are ten ways to select two out of five persons to perform a task and ten ways to select three out of five persons to be excluded from that task. This question may be addressed by applying Equation 9, reasoning that we are asking how many distinguishable arrangements there are of two writers and three nonwriters: ${}_5P_{2,3} = 5!/(2!3!) = 10$.

The product of combinatorial outcomes may also be employed to address questions such as in Example 4. This is demonstrated in Example 6.

EXAMPLE 6 Products of Combinations

This example provides an alternate method of answering the question of Example 4.

There are twelve potted plants, six of one species, four of a second species, and two of a third. How many different linear sequences of species are possible?

There are twelve positions in the sequence, which may be filled by the six members of the first species in this many ways:

$${}_{12}C_6 = \frac{12!}{(12-6)!6!} = 924.$$

The remaining six positions in the sequence may be filled by the four members of the second species in this many ways:

$${}_6C_4 = \frac{6!}{(6-4)!4!} = 15.$$

And the remaining two positions may be filled by the two members of the third species in only one way:

$$ {}_2C_2 = \frac{2!}{(2 - 2)!2!} = 1. $$

As each of the ways of filling positions with members of one species exists in association with each of the ways of filling positions with members of each other species, the total different sequences of species is

$$ (924)(15)(1) = 13{,}860. $$

From Equation 10 it may be noted that, as ${}_nC_X = {}_nP_X/X!$,

$$ {}_nP_X = X!\,{}_nC_X. \tag{14} $$

It is common mathematical convention to indicate the number of combinations of n objects taken X at a time as $\binom{n}{X}$ instead of ${}_nC_X$, so for the problem at the beginning of Section 3 we could have written*

$$ \binom{n}{X} = \binom{4}{2} = \frac{4!}{2!(4 - 2)!} = 6. $$

Binomial coefficients, take this form.

4 SETS

A *set* is a defined collection of items. For example, a set may be a group of four animals, a collection of eighteen amino acids, an assemblage of twenty-five students, or a group of three genetic traits. Each item in a set is termed an *element*. If a set of animals includes these four elements: horse (H), cow (C), sheep (S), and pig (P), and a second set consists of the elements P, S, H, and C, then we say that the two sets are *equal*, as they contain exactly the same elements. The sequence of elements within sets is immaterial in defining equality or inequality of sets.

If a set consisted of animals H and P, it would be declared a *subset* of the above set (H, C, S, P). A subset is a set, all of whose elements are elements of a larger set.† Therefore, the determination of combinations of X items taken from a set of n items (Section 3) is really the counting of possible subsets of items from the set of n items.

In an experiment (or other phenomenon that yields results to observe), there is a set (usually very large) of possible outcomes. Let us refer to this set as the *outcome set*.‡

Each element of the set is one of the possible outcomes of the experiment. For example, if an experiment consists of tossing two coins, the outcome set consists of four elements: H,H; H,T; T,H; T,T, as these are all of the possible outcomes.

A subset of the outcome set is called an *event*. If the outcome set were the possible rolls of a die: 1, 2, 3, 4, 5, 6, an event might be declared to be "even-numbered rolls" (i.e., 2, 4, 6), and another event might be defined as "rolls greater than 4"

*This parenthetical notation for combinations was introduced by Andreas von Ettingshausen in 1826 (Miller, 2004c). Some authors have used a symbol in the form of C_X^n (or ${}^nC_X$) instead of ${}_nC_X$ for combinations and P_X^n (or ${}^nP_X$) instead of ${}_nP_X$ for permutations; those symbols will not be used in this text, in order to avoid confusing n with an exponent.

†Utilizing the terms *set* and *subset* in this fashion dates from the last half of the nineteenth century (Miller, 2004a).

‡Also called the *sample space*.

(i.e., 5, 6). In tossing two coins, one event could be "the two coins land differently" (i.e., T,H; H,T), and another event could be "heads do not appear" (i.e., T,T). If the two events in the same outcome set have some elements in common, the two events are said to intersect; and the *intersection* of the two events is that subset composed of those common elements. For example, the event "even-numbered rolls" of a die (2, 4, 6) and the event "rolls greater than 4" (5, 6) have an element in common (namely, the roll 6); therefore 6 is the intersection of the two events. For the events "even-numbered rolls" (2, 4, 6) and "rolls less than 5" (1, 2, 3, 4), the intersection subset consists of those elements of the events that are both even-numbered and less than 5 (namely, 2, 4).*

If two events have no elements in common, they are said to be *mutually exclusive*, and the two sets are said to be *disjoint*. The set that is the intersection of disjoint sets contains no elements and is often called the *empty set* or the *null set*. For example, the events "odd-numbered rolls" and "even-numbered rolls" are mutually exclusive and there are no elements common to both of them.

If we ask what elements are found in either one event or another, or in both of them, we are speaking of the *union* of the two events. The union of the events "even-numbered rolls" and "rolls less than 5" is that subset of the outcome set that contains elements found in either set (or both sets), namely 1, 2, 3, 4, 6.†

Once a subset has been defined, all other elements in the outcome set are said to be the *complement* of that subset. So, if an event is defined as "even-numbered rolls" of a die (2, 4, 6), the complementary subset consists of "odd-numbered rolls" (1, 3, 5). If subset is "rolls less than 5" (1, 2, 3, 4), the complement is the subset consisting of rolls 5 or greater (5, 6).

The above considerations may be presented by what are known as *Venn diagrams*,‡ shown in Figure 1.

The rectangle in this diagram denotes the outcome set, the set of all possible outcomes from an experiment or other producer of observations. The circle on the

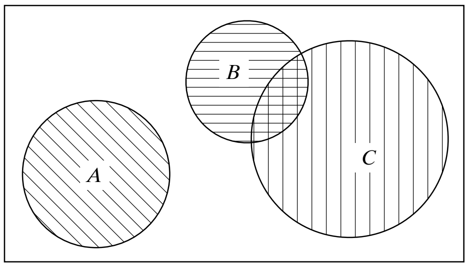

FIGURE 1: A Venn diagram showing the relationships among the outcome set represented by the rectangle and the subsets represented by circles *A*, *B*, and *C*. Subsets *B* and *C* intersect, with no intersection with *A*.

*The term *intersection* had been employed in this manner by 1909 (Miller, 2004a). The mathematical symbol for intersection is "∩", first used by Italian mathematician Giuseppe Peano (1858–1932) in 1888 (Miller, 2004a); so, for example, the intersection of set A (consisting of 2, 4, 6) and set B (consisting of 5, 6) is set $A \cap B$ (consisting of 6).

†The term *union* had been employed in this way by 1912 (Miller, 2004a). The mathematical symbol for union is "∪", first used by Giuseppe Peano in 1888 (Miller, 2004a); so, for example, if set A is composed of even-numbered rolls of a die (2, 4, 6), and set B is odd-numbered rolls (1, 3, 5), the union of the two sets, namely $A \cup B$, is 2, 4, 6, 1, 3, 5.

‡Named for English mathematical logician John Venn (1834–1923), who in 1880 greatly improved and popularized the diagrams (sometimes called "Euler diagrams") devised by Leonhard Euler (1707–1783) (Gullberg, 1997: 242; O'Connor and Robertson, 2003).

left represents a subset of the outcome set that we shall refer to as event A, the circle in the center signifies a second subset of the outcome set that we shall refer to as event B, and the circle on the right depicts a third subset of the outcome set that we shall call event C. If, for example, an outcome set (the rectangle) is the number of vertebrate animals in a forest, subset A might be animals without legs (namely, snakes), subset B might be mammals, and subset C might be flying animals. Figure 1 demonstrates graphically what is meant by union, intersection, mutually exclusive, and complementary sets: The union of B and C (the areas with any horizontal or vertical shading) represents all birds and mammals; the intersection of B and C (the area with both horizontal and vertical shading) represents flying mammals (i.e., bats); the portion of C with only vertical shading represents birds; A is mutually exclusive relative to the union of B and C, and the unshaded area (representing all other vertebrates—namely, amphibians and turtles) is complementary to A, B, and C (and is also mutually exclusive of A, B, and C).

5 PROBABILITY OF AN EVENT

We shall define the *relative frequency* of an event as the proportion of the total observations of outcomes that event represents. Consider an outcome set with two elements, such as the possible results from tossing a coin (H; T) or the sex of a person (male; female). If n is the total number of coin tosses and f is the total number of heads observed, then the relative frequency of heads is f/n. Thus, if heads are observed 52 times in 100 coin tosses, the relative frequency is $52/100 = 0.52$ (or 52%). If 275 males occur in 500 human births, the relative frequency of males is $f/n = 275/500 = 0.55$ (or 55%). In general, we may write

$$\text{relative frequency of an event} = \frac{\text{frequency of that event}}{\text{total number of all events}} = \frac{f}{n}. \tag{15}$$

The value of f may, of course, range from 0 to n, and the relative frequency may, therefore, range from 0 to 1 (or 0% to 100%). A biological example is given as Example 7.

EXAMPLE 7 Relative Frequencies

A sample of 852 vertebrate animals is taken randomly from a forest. The sampling was done *with replacement*, meaning that the animals were taken one at a time, returning each one to the forest before the next one was selected. This is done to prevent the sampling procedure from altering the relative frequency in the sampled population. If the sample size is very small compared to the population size, replacement is not necessary. (Recall that random sampling assumes that each individual animal is equally likely to become a part of the sample.)

Vertebrate Subset	*Number*	*Relative Frequency*
amphibians	53	$53/852 = 0.06$
turtles	41	$41/852 = 0.05$
snakes	204	$204/852 = 0.24$
birds	418	$418/852 = 0.49$
mammals	136	$136/852 = 0.16$
total	852	1.00

The *probability* of an event is the likelihood of that event expressed either by the relative frequency observed from a large number of data or by knowledge of the system under study. In Example 7 the relative frequencies of vertebrate groups have been observed from randomly sampling forest animals. If, for the sake of the present example, we assume that each animal has the same chance of being caught as part of our sample (an unrealistic assumption in nature), we may estimate the probability, P, that the next animal captured will be a snake ($P = 0.24$). Or, using the data of the preceding paragraph, we can estimate that the probability that a human birth will be a male is 0.55, or that the probability of tossing a coin that lands head side up is 0.52. A probability may sometimes be predicted on the basis of knowledge about the system (e.g., the structure of a coin or of a die, or the Mendelian principles of heredity). If we assume that there is no reason why a tossed coin should land "heads" more or less often than "tails," we say there is an equal probability of each outcome: $P(H) = \frac{1}{2}$ and $P(T) = \frac{1}{2}$ states that "the probability of heads is 0.5 and the probability of tails is 0.5."

Probabilities, like relative frequencies, can range from 0 to 1. A probability of 0 means that the event is impossible. For example, in tossing a coin, $P(\text{neither H nor T}) = 0$, or in rolling a die, $P(\text{number} > 6) = 0$. A probability of 1 means that an event is certain. For example, in tossing a coin, $P(\text{H or T}) = 1$; or in rolling a die, $P(1 \leq \text{number} \leq 6) = 1$.*

6 ADDING PROBABILITIES

(a) If Events Are Mutually Exclusive. If two events (call them A and B) are mutually exclusive (e.g., legless vertebrates and mammals are disjoint sets in Figure 1), then the probability of either event A or event B is the sum of the probabilities of the two events:

$$P(A \text{ or } B) = P(A) + P(B). \tag{16}$$

For example, if the probability of a tossed coin landing head up is $\frac{1}{2}$ and the probability of its landing tail up is $\frac{1}{2}$, then the probability of either head or tail up is

$$P(H \text{ or } T) = P(H) + P(T) = \frac{1}{2} + \frac{1}{2} = 1. \tag{17}$$

And, for the data in Example 7, the probability of selecting, at random, a reptile would be $P(\text{turtle or snake}) = P(\text{turtle}) + P(\text{snake}) = 0.05 + 0.24 = 0.29$.

This rule for adding probabilities may be extended for more than two mutually exclusive events. For example, the probability of rolling a 2 on a die is $\frac{1}{6}$, the probability of rolling a 4 is $\frac{1}{6}$, and the probability of rolling a 6 is $\frac{1}{6}$; so the probability of rolling an even number is

$$\begin{aligned} P(\text{even number}) = P(2 \text{ or } 4 \text{ or } 6) &= P(2) + P(4) + P(6) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}. \end{aligned}$$

*A concept related to probability is the *odds* for an event, namely the ratio of the probability of the event occurring and the probability of that event not occurring. For example, if the probability of a male birth is 0.55 (and, therefore, the probability of a female birth is 0.45), then the odds in favor of male births are 0.55/0.45, expressed as "11 to 9."

And, for the data in Example 7, the probability of randomly selecting a reptile or amphibian would be $P(\text{turtle}) + P(\text{snake}) + P(\text{amphibian}) = 0.05 + 0.24 + 0.06 = 0.35$.

(b) If Events Are Not Mutually Exclusive. If two events are not mutually exclusive—that is, they intersect (e.g., mammals and flying vertebrates are not disjoint sets in Figure 1)—then the addition of the probabilities of the two events must be modified. For example, if we roll a die, the probability of rolling an odd number is

$$\begin{aligned} P\ (\text{odd number}) = P(1 \text{ or } 3 \text{ or } 5) &= P(1) + P(3) + P(5) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}; \end{aligned}$$

and the probability of rolling a number less than 4 is

$$\begin{aligned} P\ (\text{number} < 4) = P(1 \text{ or } 2 \text{ or } 3) &= P(1) + P(2) + P(3) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}. \end{aligned}$$

The probability of rolling either an odd number or a number less than 4 obviously is *not* calculated by Equation 16, for that equation would yield

$$\begin{aligned} &P(\text{odd number or number} < 4) \\ &\stackrel{?}{=} P(\text{odd}) + P(\text{number} < 4) \\ &= P[(1 \text{ or } 3 \text{ or } 5) \text{ or } (1 \text{ or } 2 \text{ or } 3)] \\ &= [P(1) + P(3) + P(5)] + [P(1) + P(2) + P(3)] \\ &= \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) + \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) = 1, \end{aligned}$$

and that would mean that we are certain ($P = 1$) to roll either an odd number or a number less than 4, which would mean that a roll of 4 or 6 is impossible!

The invalidity of the last calculation is due to the fact that the two elements (namely 1 and 3) that lie in both events are counted twice. The subset of elements consisting of rolls 1 and 3 is the intersection of the two events and its probability needs to be subtracted from the preceding computation so that $P(1 \text{ or } 3)$ is counted once, not twice. Therefore, for two intersecting events, A and B, the probability of either A or B is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \quad (18)$$

In the preceding example,

$$\begin{aligned} &P(\text{odd number or number} < 4) \\ &\quad = P(\text{odd number}) + P(\text{number} < 4) \\ &\qquad - P(\text{odd number and number} < 4) \\ &\quad = P[(1 \text{ or } 3 \text{ or } 5) \text{ or } (1 \text{ or } 2 \text{ or } 3)] - P(1 \text{ or } 3) \\ &\quad = [P(1) + P(3) + P(5)] + [P(1) + P(2) + P(3)] - [P(1) + P(3)] \\ &\quad = \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) + \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) - \left(\frac{1}{6} + \frac{1}{6}\right) = \frac{4}{6} = \frac{2}{3}. \end{aligned}$$

It may be noted that Equation 16 is a special case of Equation 18, where $P(A \text{ and } B) = 0$. Example 8 demonstrates these probability calculations with a different set of data.

EXAMPLE 8 Adding Probabilities of Intersecting Events

A deck of playing cards is composed of 52 cards, with thirteen cards in each of four suits called clubs, diamonds, hearts, and spades. In each suit there is one card each of the following thirteen denominations: ace (A), 2, 3, 4, 5, 6, 7, 8, 9, 10, jack (J), queen (Q), king (K). What is the probability of selecting at random a diamond from the deck of 52 cards?

The event in question (diamonds) is a subset with thirteen elements; therefore,

$$P(\text{diamond}) = \frac{13}{52} = \frac{1}{4} = 0.250.$$

What is the probability of selecting at random a king from the deck?

The event in question (king) has four elements; therefore,

$$P(\text{king}) = \frac{4}{52} = \frac{1}{13} = 0.077.$$

What is the probability of selecting at random a diamond or a king?

The two events (diamonds and kings) intersect, with the intersection having one element (the king of diamonds); therefore,

$$\begin{aligned} P(\text{diamond or king}) &= P(\text{diamond}) + P(\text{king}) - P(\text{diamond and king}) \\ &= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} \\ &= \frac{16}{52} = \frac{4}{13} = 0.308. \end{aligned}$$

If three events are not mutually exclusive, the situation is more complex, yet straightforward. As seen in Figure 2, there may be three two-way intersections, shown with vertical shading (A and B; A and C; and B and C), and a three-way

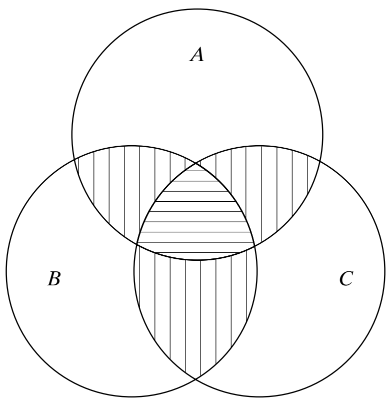

FIGURE 2: A Venn diagram showing three intersecting sets: A, B, and C. Here there are three two-way intersections (vertical shading) and one three-way intersection (horizontal shading).

intersection, shown with horizontal shading (A and B and C). If we add the probabilities of the three events, A, B, and C, as $P(A) + P(B) + P(C)$, we are adding the two-way intersections twice. So, we can subtract $P(A \text{ and } B)$, $P(A \text{ and } C)$, and $P(B \text{ and } C)$. Also, the three-way intersection is added three times in $P(A) + P(B) + P(C)$, and subtracted three times by subtracting the three two-way intersections; thus, $P(A \text{ and } B \text{ and } C)$ must be added back into the calculation. Therefore, for three events, not mutually exclusive,

$$\begin{aligned} P(A \text{ or } B \text{ or } C) = {} & P(A) + P(B) + P(C) \\ & - P(A \text{ and } B) - P(A \text{ and } C) - P(B \text{ and } C) \\ & + P(A \text{ and } B \text{ and } C). \end{aligned} \tag{19}$$

7 MULTIPLYING PROBABILITIES

If two or more events intersect (as A and B in Figure 1 and A, B, and C in Figure 2), the probability associated with the intersection is the product of the probabilities of the individual events. That is,

$$P(A \text{ and } B) = [P(A)][P(B)], \tag{20}$$

$$P(A \text{ and } B \text{ and } C) = [P(A)][P(B)][P(C)], \tag{21}$$

and so on.

For example, the probability of a tossed coin landing heads is $\frac{1}{2}$. If two coins are tossed, the probability of *both* coins landing heads is

$$P(H,H) = [P(H)][P(H)] = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{4}\right) = 0.25.$$

This can be verified by examining the outcome set:

H,H; H,T; T,H; T,T,

where $P(H,H)$ is one outcome out of four equally likely outcomes. The probability that 3 tossed coins will land heads is

$$P(H,H,H) = [P(H)][P(H)][P(H)] = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{8}\right) = 0.125.$$

Note, however, that if one or more coins have already been tossed, the probability that the next coin toss (of the same or a different coin) will be heads is simply $\frac{1}{2}$.

8 CONDITIONAL PROBABILITIES

There are occasions when our interest will be in determining a *conditional probability*, which is the probability of one event with the stipulation that another event also occurs. An illustration of this, using a deck of 52 playing cards (as described in Example 8), would be the probability of selecting a queen, given that the card is a spade. In general, a conditional probability is

$$P(\text{event } A, \text{ given event } B) = \frac{P(A \text{ and } B \text{ jointly})}{P(B)}, \tag{22}$$

which can also be calculated as

$$P(\text{event } A, \text{ given event } B) = \frac{\text{frequency of events } A \text{ and } B \text{ jointly}}{\text{frequency of event } B}. \tag{23}$$

So, the probability of randomly selecting a queen, with the specification that the card is a spade, is (using Equation 22)

$$P(\text{queen, given it is a spade}) = \frac{P(\text{queen of spades})}{P(\text{spade})}$$
$$= (1/52)/(13/52) = 0.02/0.25 = 0.08,$$

which (by Equation 23) would be calculated as

$$P(\text{queen, given it is a spade}) = \frac{\text{frequency of queen of spades}}{\text{frequency of spades}}$$
$$= 1/13 = 0.8.$$

Note that this conditional probability is quite different from the probability of selecting a spade, given that the card is a queen, for that would be (by Equation 23)

$$P(\text{spade, given it is a queen}) = \frac{\text{frequency of queen of spades}}{\text{frequency of queens}}$$
$$= 1/4 = 0.25.$$

EXERCISES

1. A person may receive a grade of either high (H), medium (M), or low (L) on a hearing test, and a grade of either good (G) or poor (P) on a sight test.
   (a) How many different outcomes are there if both tests are taken?
   (b) What are these outcomes?

2. A menu lists three meats, four salads, and two desserts. In how many ways can a meal of one meat, one salad, and one dessert be selected?

3. If an organism (e.g., human) has 23 pairs of chromosomes in each diploid cell, how many different gametes are possible for the individual to produce by assortment of chromosomes?

4. In how many ways can five animal cages be arranged on a shelf?

5. In how many ways can 12 different amino acids be arranged into a polypeptide chain of five amino acids?

6. An octapeptide is known to contain four of one amino acid, two of another, and two of a third. How many different amino-acid sequences are possible?

7. Students are given a list of nine books and told that they will be examined on the contents of five of them. How many combinations of five books are possible?

8. The four human blood types below are genetic phenotypes that are mutually exclusive events. Of 5400 individuals examined, the following frequency of each blood type is observed. What is the relative frequency of each blood type?

Blood Type	*Frequency*
O	2672
A	2041
B	486
AB	201

9. An aquarium contains the following numbers of tropical freshwater fishes. What is the relative frequency of each species?

Species	*Number*
Paracheirodon innesi, neon tetra	11
Cheirodon axelrodi, cardinal tetra	6
Pterophyllum scalare, angelfish	4
Pterophyllum altum, angelfish	2
Pterophyllum dumerilii, angelfish	2
Nannostomus marginatus, one-lined pencilfish	2
Nannostomus anomalus golden pencilfish	2

10. Use the data of Exercise 8, assuming that each of the 5400 has an equal opportunity of being encountered.
 (a) Estimate the probability of encountering a person with type A blood.
 (b) Estimate the probability of encountering a person who has either type A or type AB blood.

11. Use the data of Exercise 9, assuming that each individual fish has the same probability of being encountered.
 (a) Estimate the probability of encountering an angelfish of the species *Pterophyllum scalare*.
 (b) Estimate the probability of encountering a fish belonging to the angelfish genus *Pterophyllum*.

12. Either allele A or a may occur at a particular genetic locus. An offspring receives one of its alleles from each of its parents. If one parent possesses alleles A and a and the other parent possesses a and a:
 (a) What is the probability of an offspring receiving an A and an a?
 (b) What is the probability of an offspring receiving two a alleles?
 (c) What is the probability of an offspring receiving two A alleles?

13. In a deck of playing cards (see Example 8 for a description),
 (a) What is the probability of selecting a queen of clubs?
 (b) What is the probability of selecting a black (i.e., club or spade) queen?
 (c) What is the probability of selecting a black face card (i.e., a black jack, queen, or king)?

14. A cage contains six rats, two of them white (W) and four of them black (B); a second cage contains four rats, two white and two black; and a third cage contains five rats, three white and two black. If one rat is selected randomly from each cage,
 (a) What is the probability that all three rats selected will be white?
 (b) What is the probability that exactly two of the three will be white?
 (c) What is the probability of selecting at least two white rats?

15. A group of dogs consists of three brown males, two brown females, four white males, four white females, five black males, and four black females. What is the probability of selecting at random
 (a) A brown female dog?
 (b) A female dog, if the dog is brown?
 (c) A brown dog, if the dog is a female?

ANSWERS TO EXERCISES

1. **(a)** $(3)(2) = 6$; **(b)** H,G H,P M,G M,P L,G L,P.

2. $(3)(4)(2) = 24$.

3. $2^{23} = 8{,}388{,}608$.

4. ${}_5P_5 = 5! = 120$.

5. ${}_{12}P_5 = 12!/7! = 95{,}040$.

6. ${}_8P_{4,2,2} = 8!/[4!2!2!] = 420$.

7. ${}_9C_5 = 9!/(5!4!) = 126$.

8. O: 0.49; A: 0.38; B: 0.09; AB: 0.04.

9. $n = 29$; 0.38, 0.21, 0.14, 0.07, 0.07, 0.07, 0.07.

10. **(a)** $P = 0.38$; **(b)** $P = 0.38 + 0.04 = 0.42$.

11. **(a)** $P = 4/29 = 0.14$;
(b) $P = 4/29 + 2/29 + 2/29 = 0.28$.

12. **(a)** $P = \left(\frac{1}{2}\right)(1) = \left(\frac{1}{2}\right) = 0.5$; **(b)** $P = \left(\frac{1}{2}\right)(1) = \left(\frac{1}{2}\right) = 0.5$; **(c)** $P = \left(\frac{1}{2}\right)(0) = 0$.

13. **(a)** $P = \left(\frac{1}{13}\right)\left(\frac{1}{4}\right) = \frac{1}{52} = 0.019$; **(b)** $P = \left(\frac{1}{4} + \frac{1}{4}\right)\left(\frac{1}{13}\right) = \frac{1}{26} = 0.038$;
(c) $P = \left(\frac{1}{2}\right)\left(\frac{3}{13}\right) = \frac{3}{26} = 0.12$.

14. **(a)** $P(\text{all 3 white}) = [P(W)][P(W)][P(W)] = \left(\frac{2}{6}\right)\left(\frac{2}{4}\right)\left(\frac{3}{5}\right) = \frac{12}{120} = 0.10$; **(b)** $P(\text{2 white}) = [P(W)][P(W)][P(B)] + [P(W)][P(B)][P(W)] + [P(B)][P(W)][P(W)] = \left(\frac{2}{6}\right)\left(\frac{2}{4}\right)\left(\frac{2}{5}\right) + \left(\frac{2}{6}\right)\left(\frac{2}{4}\right)\left(\frac{3}{5}\right) + \left(\frac{4}{6}\right)\left(\frac{2}{4}\right)\left(\frac{3}{5}\right) = \frac{8}{120} + \frac{12}{120} + \frac{24}{120} = \frac{44}{120} = 0.37$.
(c) $P(\text{2 or 3 white}) = 0.10 + 0.37 = 0.47$.

15. **(a)** $P = 3/22 = 0.14$; **(b)** $P = 2/5 = 0.40$;
(c) $P = 3/10 = 0.30$.

The Normal Distribution

From Chapter 6 of *Biostatistical Analysis*, Fifth Edition, Jerrold H. Zar. Copyright © 2010 by
Pearson Education, Inc. Publishing as Pearson Prentice Hall. All rights reserved.

The Normal Distribution

Commonly, a distribution of interval- or ratio-scale data is observed to have a preponderance of values around the mean with progressively fewer observations toward the extremes of the range of values. If n is large, the frequency polygons of many biological data distributions are "bell-shaped"* and look something like Figure 1.

Figure 1 is a frequency curve for a *normal distribution.*† Not all bell-shaped curves are normal; although biologists are unlikely to need to perform calculations with this equation, it can be noted that a *normal distribution* is defined as one in which height of the curve at X_i is as expressed by the relation:

$$Y_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_i-\mu)^2/2\sigma^2}. \tag{1}$$

The height of the curve, Y_i, is referred to as the *normal density*. It is not a frequency, for in a normally distributed population of continuous data the frequency of occurrence of a measurement *exactly* equal to X_i (e.g., exactly equal to 12.5000 cm, or exactly equal to 12.50001 cm) is zero. Equation 1 contains two mathematical constants:

*Comparing the curve's shape to that of a bell has been traced as far back as 1872 (Stigler, 199: 405).

†The normal distribution is sometimes called the *Gaussian distribution*, after [Johann] Karl Friedrich Gauss (1777–1855), a phenomenal German mathematician contributing to many fields of mathematics and for whom the unit of magnetic induction ("gauss") is named. Gauss discussed this distribution in 1809, but the influential French mathematician and astronomer Pierre-Simon Laplace (1749–1827) mentioned it in 1774, and it was first announced in 1733 by mathematician Abraham de Moivre (1667–1754; also spelled De Moivre and Demoivre), who was born in France but emigrated to England at age 21 (after three years in prison) to escape religious persecution as a Protestant (David, 1962: 161–178; Pearson, 1924; Stigler, 1980; Walker, 1934). This situation has been cited as an example of "Stigler's Law of Eponymy," which states that "no scientific discovery is named after its original discoverer" (Stigler, 1980). The distribution was first used, by de Moivre, to approximate a binomial distribution (Stigler, 1999: 407). The adjective *normal* was first used for the distribution by Charles S. Peirce in 1873, and by Wilhelm Lexis and Sir Francis Galton in 1877 (Stigler, 1999: 404–415); Karl Pearson recommended the routine use of that term to avoid "an international question of priority" although it "has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another 'abnormal' " (Pearson, 1920).