***BIOSTATISTICS\_Daniel\_C02***

p33

**C02\_***DESCRIPTIVE\_STATISTICS*

**CHAPTER OVERVIEW**

This chapter introduces a set of basic procedures and statistical measures for describing data. Data generally consist of an extensive number of measurements or observations that are too numerous or complicated to be understood through simple observation.Therefore, this chapter introduces several techniques including the construction of tables, graphical displays, and basic statistical computations that provide ways to condense and organize information into a set of descriptive measures and visual devices that enhance the understanding of complex data.

Este capítulo presenta un conjunto de procedimientos básicos y medidas estadísticas para la descripción de datos. Los datos generalmente consisten en un gran número de mediciones u observaciones que son demasiado numerosas o complejas para comprenderse mediante la simple observación. Por lo tanto, este capítulo presenta diversas técnicas, incluyendo la construcción de tablas, representaciones gráficas y cálculos estadísticos básicos que permiten condensar y organizar la información en un conjunto de medidas descriptivas y recursos visuales que facilitan la comprensión de datos complejos.

**TOPICS**

**2.1** INTRODUCTION  
   
**2.2** THE ORDERED ARRAY  
   
**2.3** GROUPED DATA:THE FREQUENCY DISTRIBUTION  
   
**2.4** DESCRIPTIVE STATISTICS: MEASURES OF CENTRALTENDENCY  
   
**2.5** DESCRIPTIVE STATISTICS: MEASURES OF DISPERSION  
   
**2.6** SUMMARY

**LEARNING OUTCOMES**

After studying this chapter, the student will

1. understand how data can be appropriately organized and displayed.
2. understand how to reduce data sets into a few useful, descriptive measures.
3. be able to calculate and interpret measures of central tendency, such as the mean, median, and mode.
4. be able to calculate and interpret measures of dispersion, such as the range, variance, and standard deviation.

**2.1 INTRODUCTION**

In Chapter 1 we stated that the taking of a measurement and the process of counting yield numbers that contain information. The objective of the person applying the tools of statistics to these numbers is to determine the nature of this information. This task is made much easier if the numbers are organized and summarized. When measurements of a random variable are taken on the entities of a population or sample, the resulting values are made available to the researcher or statistician as a mass of unordered data. Measurements that have not been organized, summarized, or otherwise manipulated are called *raw data*. Unless the number of observations is extremely small, it will be unlikely that these raw data will impart much information until they have been put into some kind of order.

En el Capítulo 1, afirmamos que la toma de medidas y el proceso de conteo producen números que contienen información. El objetivo de quien aplica las herramientas estadísticas a estos números es determinar la naturaleza de esta información. Esta tarea se facilita mucho si los números están organizados y resumidos. Cuando se toman medidas de una variable aleatoria en las entidades de una población o muestra, los valores resultantes se ponen a disposición del investigador o estadístico como una masa de datos desordenados. Las mediciones que no se han organizado, resumido ni manipulado de otro modo se denominan datos brutos. A menos que el número de observaciones sea extremadamente pequeño, será improbable que estos datos brutos proporcionen mucha información hasta que se hayan ordenado de algún modo.

In this chapter we learn several techniques for organizing and summarizing data so that we may more easily determine what information they contain. The ultimate in summarization of data is the calculation of a single number that in some way conveys important information about the data from which it was calculated. Such single numbers that are used to describe data are called *descriptive measures.* After studying this chapter you will be able to compute several descriptive measures for both populations and samples of data.

En este capítulo, aprenderemos diversas técnicas para organizar y resumir datos, lo que nos permitirá determinar con mayor facilidad su contenido. El resumen de datos, en su máxima expresión, consiste en calcular un único número que, de alguna manera, transmite información importante sobre los datos a partir de los cuales se calculó. Estos números únicos que se utilizan para describir datos se denominan medidas descriptivas. Tras estudiar este capítulo, podrá calcular diversas medidas descriptivas tanto para poblaciones como para muestras de datos.

The purpose of this chapter is to equip you with skills that will enable you to manipulate the information—in the form of numbers—that you encounter as a health sciences professional. The better able you are to manipulate such information, the better understanding you will have of the environment and forces that generate the information.

El propósito de este capítulo es brindarle habilidades que le permitirán manipular la información —en forma de números— con la que se encuentra como profesional de las ciencias de la salud. Cuanto mejor maneje dicha información, mejor comprenderá el entorno y las fuerzas que la generan.

**2.2 THE ORDERED ARRAY**

A first step in organizing data is the preparation of an ordered array. An *ordered array* is a listing of the values of a collection (either population or sample) in order of magnitude from the smallest value to the largest value. If the number of measurements to be ordered is of any appreciable size, the use of a computer to prepare the ordered array is highly desirable.

Un primer paso para organizar datos es preparar una matriz ordenada. Una matriz ordenada es una lista de los valores de una colección (ya sea una población o una muestra) ordenados por magnitud, desde el valor más pequeño hasta el más grande. Si el número de mediciones a ordenar es considerable, es muy recomendable usar una computadora para preparar la matriz ordenada.

An ordered array enables one to determine quickly the value of the smallest meas- urement, the value of the largest measurement, and other facts about the arrayed data that might be needed in a hurry. We illustrate the construction of an ordered array with the data discussed in Example 1.4.1.

Una matriz ordenada permite determinar rápidamente el valor de la medida más pequeña, el valor de la medida más grande y otros datos sobre los datos de la matriz que podrían necesitarse con urgencia. Ilustramos la construcción de una matriz ordenada con los datos analizados en el Ejemplo 1.4.1.

**EXAMPLE 2.2.1**

Table 1.4.1 contains a list of the ages of subjects who participated in the study on smoking cessation discussed in Example 1.4.1. As can be seen, this unordered table requires considerable searching for us to ascertain such elementary information as the age of the youngest and oldest subjects.

La Tabla 1.4.1 contiene una lista de las edades de los sujetos que participaron en el estudio sobre el abandono del hábito tabáquico, analizado en el Ejemplo 1.4.1. Como puede observarse, esta tabla desordenada requiere una búsqueda considerable para determinar información tan básica como la edad de los sujetos más jóvenes y mayores.

**Solution:** Table 2.2.1 presents the data of Table 1.4.1 in the form of an ordered array. By referring to Table 2.2.1 we are able to determine quickly the age of the youngest subject (30) and the age of the oldest subject (82). We also readily note that about one-third of the subjects are 50 years of age or younger.

**Computer Analysis** If additional computations and organization of a data set have to be done by hand, the work may be facilitated by working from an ordered array. If the data are to be analyzed by a computer, it may be undesirable to prepare an ordered array, unless one is needed for reference purposes or for some other use. A computer does not need its user to first construct an ordered array before entering data for the construction of frequency distributions and the performance of other analyses. However, almost all computer statistical packages and spreadsheet programs contain a routine for sorting data in either an ascending or descending order. See Figure 2.2.1, for example.

**2.3 GROUPED DATA: THE FREQUENCY DISTRIBUTION**

Although a set of observations can be made more comprehensible and meaningful by means of an ordered array, further useful summarization may be achieved by grouping the data. Before the days of computers one of the main objectives in grouping large data sets was to facilitate the calculation of various descriptive measures such as percentages and averages. Because computers can perform these calculations on large data sets with- out first grouping the data, the main purpose in grouping data now is summarization. One must bear in mind that data contain information and that summarization is a way of making it easier to determine the nature of this information.

Aunque un conjunto de observaciones puede hacerse más comprensible y significativo mediante una matriz ordenada, se puede lograr un resumen más útil agrupando los datos. Antes de la era de las computadoras, uno de los principales objetivos al agrupar grandes conjuntos de datos era facilitar el cálculo de diversas medidas descriptivas, como porcentajes y promedios. Dado que las computadoras pueden realizar estos cálculos en grandes conjuntos de datos sin agruparlos primero, el propósito principal al agrupar datos ahora es el resumen. Hay que tener en cuenta que los datos contienen información y que el resumen es una forma de facilitar la determinación de la naturaleza de esta información.

To group a set of observations we select a set of contiguous, nonoverlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals. These intervals are usually referred to as *class intervals.*

Para agrupar un conjunto de observaciones, seleccionamos un conjunto de intervalos contiguos que no se superponen, de modo que cada valor del conjunto de observaciones pueda ubicarse en uno solo de los intervalos. Estos intervalos suelen denominarse intervalos de clase.

One of the first considerations when data are to be grouped is how many intervals to include. Too few intervals are undesirable because of the resulting loss of information. On the other hand, if too many intervals are used, the objective of summarization will not be met. The best guide to this, as well as to other decisions to be made in grouping data, is your knowledge of the data. It may be that class intervals have been determined by precedent, as in the case of annual tabulations, when the class intervals of previous years are maintained for comparative purposes. A commonly followed rule of thumb states that there should be no fewer than five intervals and no more than 15. If there are fewer than five intervals, the data have been summarized too much and the information they contain has been lost. If there are more than 15 intervals, the data have not been summarized enough.

Una de las primeras consideraciones al agrupar datos es cuántos intervalos incluir. Es indeseable tener muy pocos intervalos debido a la pérdida de información resultante. Por otro lado, si se utilizan demasiados intervalos, no se cumplirá el objetivo del resumen. La mejor guía para esto, así como para otras decisiones que deben tomarse al agrupar datos, es su conocimiento de los datos. Es posible que los intervalos de clase se hayan determinado por precedentes, como en el caso de las tabulaciones anuales, cuando los intervalos de clase de años anteriores se mantienen para fines comparativos. Una regla general comúnmente seguida establece que no debe haber menos de cinco intervalos ni más de 15. Si hay menos de cinco intervalos, los datos se han resumido demasiado y se ha perdido la información que contienen. Si hay más de 15 intervalos, los datos no se han resumido lo suficiente.

Those who need more specific guidance in the matter of deciding how many class intervals to employ may use a formula given by Sturges (1). This formula gives *k* = 1 + 3.3221log10 *n*2, where *k* stands for the number of class intervals and *n* is the number of values in the data set under consideration. The answer obtained by applying *Sturges’s rule* should not be regarded as final, but should be considered as a guide only. The number of class intervals specified by the rule should be increased or decreased for convenience and clear presentation.

Quienes necesiten una guía más específica para decidir cuántos intervalos de clase emplear pueden usar la fórmula de Sturges (1). Esta fórmula da k = 1 + 3,3221 log⁻⁻⁻⁻⁻⁻, donde k representa el número de intervalos de clase y n el número de valores en el conjunto de datos considerado. El resultado obtenido al aplicar la regla de Sturges no debe considerarse definitivo, sino solo orientativo. El número de intervalos de clase especificado por la regla debe aumentarse o disminuirse para mayor comodidad y claridad en la presentación.

Suppose, for example, that we have a sample of 275 observations that we want to group. The logarithm to the base 10 of 275 is 2.4393. Applying Sturges’s formula gives *k* = 1 + 3.32212.43932 M 9. In practice, other considerations might cause us to use eight or fewer or perhaps 10 or more class intervals.

Supongamos, por ejemplo, que tenemos una muestra de 275 observaciones que queremos agrupar. El logaritmo en base 10 de 275 es 2,4393. Aplicando la fórmula de Sturges, obtenemos k = 1 + 3,32212,43932 M 9. En la práctica, otras consideraciones podrían llevarnos a utilizar ocho o menos, o quizás diez o más intervalos de clase.

Another question that must be decided regards the width of the class intervals. Class intervals generally should be of the same width, although this is sometimes impossible to accomplish. This width may be determined by dividing the range by *k*, the number of class intervals. Symbolically, the class interval width is given by

Otra cuestión que debe resolverse se refiere a la amplitud de los intervalos de clase. Los intervalos de clase generalmente deben tener la misma amplitud, aunque a veces esto es imposible de lograr. Esta amplitud puede determinarse dividiendo el rango entre k, el número de intervalos de clase. Simbólicamente, la amplitud del intervalo de clase se expresa mediante



where *R* (the range) is the difference between the smallest and the largest observation in the data set. As a rule this procedure yields a width that is inconvenient for use. Again, we may exercise our good judgment and select a width (usually close to one given by Equation 2.3.1) that is more convenient.

Donde R (el rango) es la diferencia entre la observación más pequeña y la más grande del conjunto de datos. Por lo general, este procedimiento produce un ancho poco práctico. De nuevo, podemos usar nuestro buen juicio y seleccionar un ancho (generalmente cercano al dado por la Ecuación 2.3.1) que sea más conveniente.

There are other rules of thumb that are helpful in setting up useful class intervals. When the nature of the data makes them appropriate, class interval widths of 5 units, 10 units, and widths that are multiples of 10 tend to make the summarization more com- prehensible. When these widths are employed it is generally good practice to have the lower limit of each interval end in a zero or 5. Usually class intervals are ordered from smallest to largest; that is, the first class interval contains the smaller measurements and the last class interval contains the larger measurements. When this is the case, the lower limit of the first class interval should be equal to or smaller than the smallest measure- ment in the data set, and the upper limit of the last class interval should be equal to or greater than the largest measurement.

Existen otras reglas generales útiles para establecer intervalos de clase útiles. Cuando la naturaleza de los datos las hace apropiadas, los anchos de intervalo de clase de 5 unidades, 10 unidades y anchos que son múltiplos de 10 tienden a hacer que el resumen sea más comprensible. Cuando se emplean estos anchos, generalmente es una buena práctica que el límite inferior de cada intervalo termine en cero o 5. Por lo general, los intervalos de clase se ordenan del más pequeño al más grande; es decir, el primer intervalo de clase contiene las mediciones más pequeñas y el último intervalo de clase contiene las mediciones más grandes. Cuando este es el caso, el límite inferior del primer intervalo de clase debe ser igual o menor que la medición más pequeña en el conjunto de datos, y el límite superior del último intervalo de clase debe ser igual o mayor que la medición más grande.

Most statistical packages allow users to interactively change the number of class intervals and/or the class widths, so that several visualizations of the data can be obtained quickly. This feature allows users to exercise their judgment in deciding which data dis- play is most appropriate for a given purpose. Let us use the 189 ages shown in Table 1.4.1 and arrayed in Table 2.2.1 to illustrate the construction of a frequency distribution.

La mayoría de los paquetes estadísticos permiten a los usuarios modificar interactivamente el número de intervalos de clase o su ancho, lo que permite obtener rápidamente varias visualizaciones de los datos. Esta función permite a los usuarios decidir qué visualización de datos es la más adecuada para un propósito determinado. Utilicemos las 189 edades que se muestran en la Tabla 1.4.1 y que se muestran en la Tabla 2.2.1 para ilustrar la construcción de una distribución de frecuencias.

**EXAMPLE 2.3.1**

We wish to know how many class intervals to have in the frequency distribution of the data. We also want to know how wide the intervals should be.

**Solution:**

To get an idea as to the number of class intervals to use, we can apply Sturges’s rule to obtain

*k* = 1 + 3.3221log1892

= 1 + 3.32212.27646182 L9

Now let us divide the range by 9 to get some idea about the class interval width. We have



It is apparent that a class interval width of 5 or 10 will be more con- venient to use, as well as more meaningful to the reader. Suppose we decide on 10. We may now construct our intervals. Since the smallest value in Table 2.2.1 is 30 and the largest value is 82, we may begin our intervals with 30 and end with 89. This gives the following intervals:

30–39

40–49

50–59

60–69

70–79

80–89

We see that there are six of these intervals, three fewer than the number suggested by Sturges’s rule.

It is sometimes useful to refer to the center, called the *midpoint,* of a class interval. The midpoint of a class interval is determined by obtaining the sum of the upper and lower limits of the class interval and dividing by 2. Thus, for example, the midpoint of the class interval 30–39 is found to be130 + 392>2 = 34.5.

A veces resulta útil referirse al centro, llamado punto medio, de un intervalo de clase. El punto medio de un intervalo de clase se determina sumando los límites superior e inferior del intervalo de clase y dividiéndolo entre 2. Así, por ejemplo, el punto medio del intervalo de clase 30–39 es 130 + 392 > 2 = 34,5.

When we group data manually, determining the number of values falling into each class interval is merely a matter of looking at the ordered array and counting the num- ber of observations falling in the various intervals. When we do this for our example, we have Table 2.3.1.

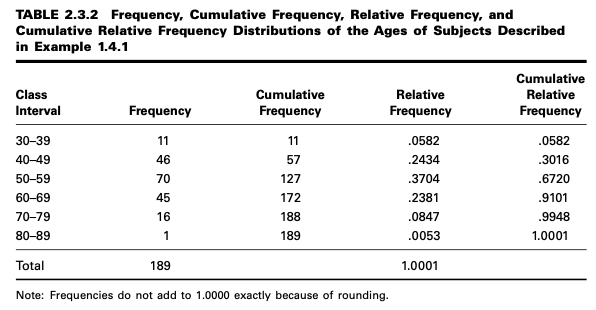
Al agrupar los datos manualmente, determinar el número de valores que corresponden a cada intervalo de clase es simplemente observar la matriz ordenada y contar el número de observaciones que corresponden a los distintos intervalos. Al hacer esto en nuestro ejemplo, obtenemos la Tabla 2.3.1.

A table such as Table 2.3.1 is called a *frequency distribution.* This table shows the way in which the values of the variable are distributed among the specified class intervals. By consulting it, we can determine the frequency of occurrence of values within any one of the class intervals shown.

Una tabla como la Tabla 2.3.1 se denomina distribución de frecuencias. Esta tabla muestra cómo se distribuyen los valores de la variable entre los intervalos de clase especificados. Al consultarla, podemos determinar la frecuencia de ocurrencia de los valores dentro de cualquiera de los intervalos de clase mostrados.

**Relative Frequencies** It may be useful at times to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this infor- mation by dividing the number of values in the particular class interval by the total num- ber of values. If, in our example, we wish to know the proportion of values between 50 and 59, inclusive, we divide 70 by 189, obtaining .3704. Thus we say that 70 out of 189, or 70􏰉189ths, or .3704, of the values are between 50 and 59. Multiplying .3704 by 100 gives us the percentage of values between 50 and 59. We can say, then, that 37.04 percent of the subjects are between 50 and 59 years of age. We may refer to the proportion of values falling within a class interval as the *relative frequency of occurrence* of values in that inter- val. In Section 3.2 we shall see that a relative frequency may be interpreted also as the probability of occurrence within the given interval. This probability of occurrence is also called the *experimental probability* or the *empirical probability*.

Frecuencias relativas A veces puede ser útil conocer la proporción, en lugar del número, de valores que caen dentro de un intervalo de clase particular. Obtenemos esta información dividiendo el número de valores en el intervalo de clase particular por el número total de valores. Si, en nuestro ejemplo, deseamos saber la proporción de valores entre 50 y 59, inclusive, dividimos 70 por 189, obteniendo .3704. Por lo tanto, decimos que 70 de 189, o 70s 189, o .3704, de los valores están entre 50 y 59. Multiplicar .3704 por 100 nos da el porcentaje de valores entre 50 y 59. Podemos decir, entonces, que el 37.04 por ciento de los sujetos tienen entre 50 y 59 años de edad. Podemos referirnos a la proporción de valores que caen dentro de un intervalo de clase como la frecuencia relativa de ocurrencia de valores en ese intervalo. En la Sección 3.2 veremos que una frecuencia relativa también puede interpretarse como la probabilidad de ocurrencia dentro del intervalo dado. Esta probabilidad de ocurrencia también se denomina probabilidad experimental o probabilidad empírica.



In determining the frequency of values falling within two or more class intervals, we obtain the sum of the number of values falling within the class intervals of interest. Similarly, if we want to know the relative frequency of occurrence of values falling within two or more class intervals, we add the respective relative frequencies. We may sum, or *cumulate,* the frequencies and relative frequencies to facilitate obtaining information regarding the frequency or relative frequency of values within two or more contiguous class intervals. Table 2.3.2 shows the data of Table 2.3.1 along with the *cumulative fre- quencies,* the *relative frequencies,* and *cumulative relative frequencies*.

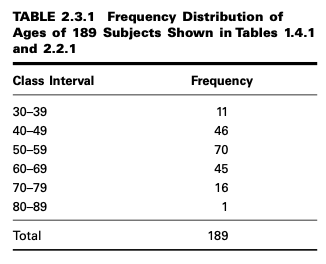
Para determinar la frecuencia de valores dentro de dos o más intervalos de clase, se obtiene la suma del número de valores dentro de los intervalos de clase de interés. De igual forma, si se desea conocer la frecuencia relativa de ocurrencia de valores dentro de dos o más intervalos de clase, se suman las frecuencias relativas respectivas. Se pueden sumar o acumular las frecuencias y las frecuencias relativas para facilitar la obtención de información sobre la frecuencia o la frecuencia relativa de valores dentro de dos o más intervalos de clase contiguos. La Tabla 2.3.2 muestra los datos de la Tabla 2.3.1 junto con las frecuencias acumuladas, las frecuencias relativas y las frecuencias relativas acumuladas.

Suppose that we are interested in the relative frequency of values between 50 and 79. We use the cumulative relative frequency column of Table 2.3.2 and subtract .3016 from .9948, obtaining .6932.

We may use a statistical package to obtain a table similar to that shown in Table 2.3.2. Tables obtained from both MINITAB and SPSS software are shown in Figure 2.3.1.

**The Histogram** We may display a frequency distribution (or a relative frequency distribution) graphically in the form of a *histogram,* which is a special type of bar graph.

When we construct a histogram the values of the variable under consideration are represented by the horizontal axis, while the vertical axis has as its scale the frequency (or relative frequency if desired) of occurrence. Above each class interval on the hori- zontal axis a rectangular bar, or cell, as it is sometimes called, is erected so that the height corresponds to the respective frequency when the class intervals are of equal width. The cells of a histogram must be joined and, to accomplish this, we must take into account the true boundaries of the class intervals to prevent gaps from occurring betweenthe cells of our graph.

Al construir un histograma, los valores de la variable en cuestión se representan en el eje horizontal, mientras que el eje vertical tiene como escala la frecuencia (o frecuencia relativa, si se desea) de ocurrencia. Sobre cada intervalo de clase en el eje horizontal se erige una barra rectangular, o celda, como a veces se le llama, cuya altura corresponde a la frecuencia respectiva cuando los intervalos de clase tienen el mismo ancho. Las celdas de un histograma deben estar unidas y, para lograrlo, debemos tener en cuenta los límites reales de los intervalos de clase para evitar que se produzcan espacios entre las celdas de nuestro gráfico.

The level of precision observed in reported data that are measured on a continuous scale indicates some order of rounding. The order of rounding reflects either the reporter’s personal preference or the limitations of the measuring instrument employed. When a fre- quency distribution is constructed from the data, the class interval limits usually reflect the degree of precision of the raw data. This has been done in our illustrative example.

El nivel de precisión observado en los datos reportados, medidos en una escala continua, indica cierto orden de redondeo. Este orden refleja la preferencia personal del reportero o las limitaciones del instrumento de medición empleado. Cuando se construye una distribución de frecuencias a partir de los datos, los límites del intervalo de clase suelen reflejar el grado de precisión de los datos brutos. Esto se ha hecho en nuestro ejemplo ilustrativo.

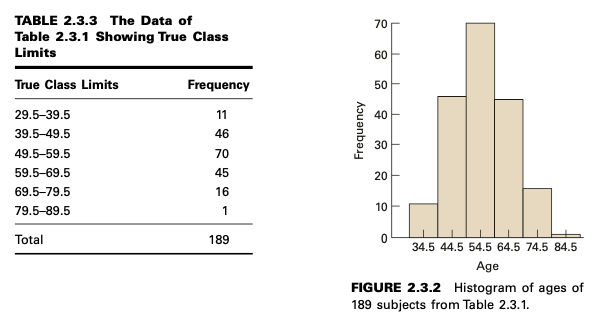
We know, however, that some of the values falling in the second class interval, for exam- ple, when measured precisely, would probably be a little less than 40 and some would be a little greater than 49. Considering the underlying continuity of our variable, and assum- ing that the data were rounded to the nearest whole number, we find it convenient to think of 39.5 and 49.5 as the true limits of this second interval. The true limits for each of the class intervals, then, we take to be as shown in Table 2.3.3.

Sabemos, sin embargo, que algunos de los valores que caen en el segundo intervalo de clase, por ejemplo, cuando se miden con precisión, probablemente serían un poco menores que 40 y algunos serían un poco mayores que 49. Considerando la continuidad subyacente de nuestra variable, y suponiendo que los datos se redondearon al número entero más cercano, nos parece conveniente pensar en 39.5 y 49.5 como los límites verdaderos de este segundo intervalo. Los límites verdaderos para cada uno de los intervalos de clase, entonces, los tomamos como se muestra en la Tabla 2.3.3.

If we construct a graph using these class limits as the base of our rectangles, no gaps will result, and we will have the histogram shown in Figure 2.3.2. We used MINITAB to construct this histogram, as shown in Figure 2.3.3.

We refer to the space enclosed by the boundaries of the histogram as the *area* of the histogram. Each observation is allotted one unit of this area. Since we have 189 observa- tions, the histogram consists of a total of 189 units. Each cell contains a certain propor- tion of the total area, depending on the frequency. The second cell, for example, contains 46/189 of the area. This, as we have learned, is the relative frequency of occurrence of val- ues between 39.5 and 49.5. From this we see that subareas of the histogram defined by the cells correspond to the frequencies of occurrence of values between the horizontal scale boundaries of the areas. The ratio of a particular subarea to the total area of the histogram is equal to the relative frequency of occurrence of values between the corresponding points on the horizontal axis.

Nos referimos al espacio encerrado por los límites del histograma como el área del histograma. A cada observación se le asigna una unidad de esta área. Dado que tenemos 189 observaciones, el histograma consta de un total de 189 unidades. Cada celda contiene una cierta proporción del área total, dependiendo de la frecuencia. La segunda celda, por ejemplo, contiene 46/189 del área. Esta, como hemos aprendido, es la frecuencia relativa de aparición de valores entre 39,5 y 49,5. A partir de esto, vemos que las subáreas del histograma definidas por las celdas corresponden a las frecuencias de aparición de valores entre los límites de escala horizontales de las áreas. La relación de una subárea particular con el área total del histograma es igual a la frecuencia relativa de aparición de valores entre los puntos correspondientes en el eje horizontal.

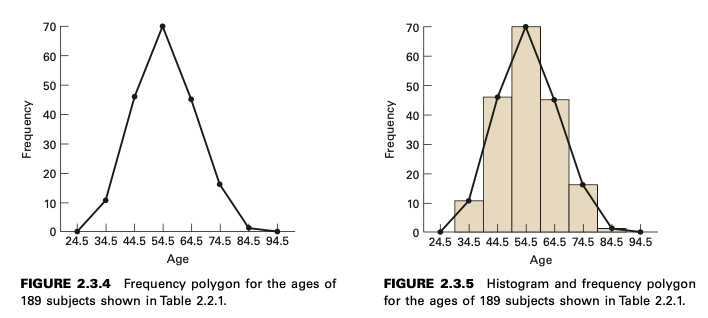


**The Frequency Polygon** A frequency distribution can be portrayed graphi- cally in yet another way by means of a *frequency polygon,* which is a special kind of line graph. To draw a frequency polygon we first place a dot above the midpoint of each class interval represented on the horizontal axis of a graph like the one shown in Figure 2.3.2. The height of a given dot above the horizontal axis corresponds to the frequency of the relevant class interval. Connecting the dots by straight lines produces the frequency polygon. Figure 2.3.4 is the frequency polygon for the age data in Table 2.2.1.

El Polígono de Frecuencias. Una distribución de frecuencias puede representarse gráficamente de otra manera mediante un polígono de frecuencias, que es un tipo especial de gráfico lineal. Para dibujar un polígono de frecuencias, primero colocamos un punto sobre el punto medio de cada intervalo de clase representado en el eje horizontal de un gráfico como el que se muestra en la Figura 2.3.2. La altura de un punto dado sobre el eje horizontal corresponde a la frecuencia del intervalo de clase relevante. Al conectar los puntos con líneas rectas se obtiene el polígono de frecuencias. La Figura 2.3.4 es el polígono de frecuencias para los datos de edad de la Tabla 2.2.1.

Note that the polygon is brought down to the horizontal axis at the ends at points that would be the midpoints if there were an additional cell at each end of the corre- sponding histogram. This allows for the total area to be enclosed. The total area under the frequency polygon is equal to the area under the histogram. Figure 2.3.5 shows the frequency polygon of Figure 2.3.4 superimposed on the histogram of Figure 2.3.2. This figure allows you to see, for the same set of data, the relationship between the two graphic forms.

Nótese que el polígono se reduce hasta el eje horizontal en los extremos, en puntos que serían los puntos medios si hubiera una celda adicional en cada extremo del histograma correspondiente. Esto permite delimitar el área total. El área total bajo el polígono de frecuencias es igual al área bajo el histograma. La Figura 2.3.5 muestra el polígono de frecuencias de la Figura 2.3.4 superpuesto al histograma de la Figura 2.3.2. Esta figura permite ver, para el mismo conjunto de datos, la relación entre las dos formas gráficas.



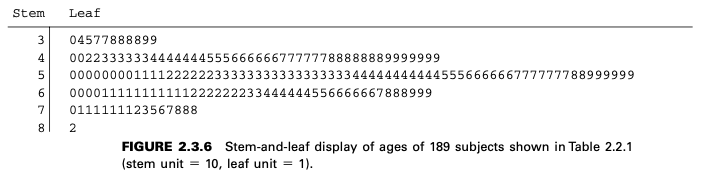
**Stem-and-Leaf Displays** Another graphical device that is useful for represent- ing quantitative data sets is the *stem-and-leaf display*. A stem-and-leaf display bears a strong resemblance to a histogram and serves the same purpose. A properly constructed stem-and-leaf display, like a histogram, provides information regarding the range of the data set, shows the location of the highest concentration of measurements, and reveals the presence or absence of symmetry. An advantage of the stem-and-leaf display over the his- togram is the fact that it preserves the information contained in the individual measure- ments. Such information is lost when measurements are assigned to the class intervals of a histogram. As will become apparent, another advantage of stem-and-leaf displays is the fact that they can be constructed during the tallying process, so the intermediate step of preparing an ordered array is eliminated.

Visualizaciones de tallo y hojas Otro dispositivo gráfico útil para representar conjuntos de datos cuantitativos es la visualización de tallo y hojas. Una visualización de tallo y hojas se parece mucho a un histograma y cumple la misma función. Una visualización de tallo y hojas bien construida, al igual que un histograma, proporciona información sobre el rango del conjunto de datos, muestra la ubicación de la mayor concentración de mediciones y revela la presencia o ausencia de simetría. Una ventaja de la visualización de tallo y hojas sobre el histograma es que conserva la información contenida en las mediciones individuales. Dicha información se pierde cuando las mediciones se asignan a los intervalos de clase de un histograma. Como se hará evidente, otra ventaja de las visualizaciones de tallo y hojas es que se pueden construir durante el proceso de recuento, por lo que se elimina el paso intermedio de preparar una matriz ordenada.

To construct a stem-and-leaf display we partition each measurement into two parts.

The first part is called the *stem,* and the second part is called the *leaf*. The stem consists of one or more of the initial digits of the measurement, and the leaf is composed of one or more of the remaining digits. All partitioned numbers are shown together in a single display; the stems form an ordered column with the smallest stem at the top and the largest at the bottom. We include in the stem column all stems within the range of the data even when a measurement with that stem is not in the data set. The rows of the display con- tain the leaves, ordered and listed to the right of their respective stems. When leaves con- sist of more than one digit, all digits after the first may be deleted. Decimals when pres- ent in the original data are omitted in the stem-and-leaf display. The stems are separated from their leaves by a vertical line. Thus we see that a stem-and-leaf display is also an ordered array of the data.

Stem-and-leaf displays are most effective with relatively small data sets. As a rule they are not suitable for use in annual reports or other communications aimed at the gen- eral public. They are primarily of value in helping researchers and decision makers under- stand the nature of their data. Histograms are more appropriate for externally circulated publications. The following example illustrates the construction of a stem-and-leaf display.



**EXAMPLE 2.3.2**

Let us use the age data shown in Table 2.2.1 to construct a stem-and-leaf display.

**Solution:**

Since the measurements are all two-digit numbers, we will have one-digit stems and one-digit leaves. For example, the measurement 30 has a stem of 3 and a leaf of 0. Figure 2.3.6 shows the stem-and-leaf display for the data.

The MINITAB statistical software package may be used to construct stem-and-leaf displays. The MINITAB procedure and output are as shown in Figure 2.3.7. The increment subcommand specifies the distance from one stem to the next. The numbers in the leftmost output column of Figure 2.3.7 provide information regarding the number of observations (leaves) on a given line and above or the number of observations on a given line and below. For example, the number 57 on the second line shows that there are 57 observations (or leaves) on that line and the one above it. The number 62 on the fourth line from the top tells us that there are 62 observations on that line and all the ones below. The number in parentheses tells us that there are 70 observations on that line. The parentheses mark the line con- taining the middle observation if the total number of observations is odd or the two middle observations if the total number of observations is even.

The + at the end of the third line in Figure 2.3.7 indicates that the fre- quency for that line (age group 50 through 59) exceeds the line capacity, and that there is at least one additional leaf that is not shown. In this case, the frequency for the 50–59 age group was 70. The line contains only 65 leaves, so the + indicates that there are five more leaves, the number 9, that are not shown. s

One way to avoid exceeding the capacity of a line is to have more lines. This is accomplished by making the distance between lines shorter, that is, by decreasing the widths of the class intervals. For the present example, we may use class interval widths of 5, so that the distance between lines is 5. Figure 2.3.8 shows the result when MINITAB is used to produce the stem-and-leaf display.

**EXERCISES**

**2.3.1** In a study of the oral home care practice and reasons for seeking dental care among individuals on renal dialysis, Atassi (A-1) studied 90 subjects on renal dialysis. The oral hygiene status of all subjects was examined using a plaque index with a range of 0 to 3 10 = no soft plaque deposits, 3 = an abundance of soft plaque deposits). The following table shows the plaque index scores for all 90 subjects.

**(a)** Use these data to prepare:

A frequency distribution

A relative frequency distribution

A cumulative frequency distribution

A cumulative relative frequency distribution A histogram

A frequency polygon

**(b)** What percentage of the measurements are less than 2.00?

**(c)** What proportion of the subjects have measurements greater than or equal to 1.50?

**(d)** What percentage of the measurements are between 1.50 and 1.99 inclusive?

**(e)** How many of the measurements are greater than 2.49?

**(f)** What proportion of the measurements are either less than 1.0 or greater than 2.49?

**(g)** Someone picks a measurement at random from this data set and asks you to guess the value. What would be your answer? Why?

**(h)** Frequency distributions and their histograms may be described in a number of ways depending on their shape. For example, they may be symmetric (the left half is at least approximately a mirror image of the right half), skewed to the left (the frequencies tend to increase as the meas- urements increase in size), skewed to the right (the frequencies tend to decrease as the measure- ments increase in size), or U-shaped (the frequencies are high at each end of the distribution and small in the center). How would you describe the present distribution?

**2.3.2** Janardhan et al. (A-2) conducted a study in which they measured incidental intracranial aneurysms (IIAs) in 125 patients. The researchers examined postprocedural complications and concluded that IIAs can be safely treated without causing mortality and with a lower complications rate than pre- viously reported. The following are the sizes (in millimeters) of the 159 IIAs in the sample.

**(a)** Use these data to prepare:

A frequency distribution

A relative frequency distribution

A cumulative frequency distribution

A cumulative relative frequency distribution A histogram

A frequency polygon

**(b)** What percentage of the measurements are between 10 and 14.9 inclusive?

**(c)** How many observations are less than 20?

**(d)** What proportion of the measurements are greater than or equal to 25?

**(e)** What percentage of the measurements are either less than 10.0 or greater than 19.95?

**(f)** Refer to Exercise 2.3.1, part h. Describe the distribution of the size of the aneurysms in this sample.

2.3.3 Hoekema et al. (A-3) studied the craniofacial morphology of patients diagnosed with obstructive sleep apnea syndrome (OSAS) in healthy male subjects. One of the demographic variables the researchers collected for all subjects was the Body Mass Index (calculated by dividing weight in kg by the square of the patient’s height in cm). The following are the BMI values of 29 OSAS subjects.

**(a)** Use these data to construct:

A frequency distribution

A relative frequency distribution

A cumulative frequency distribution

A cumulative relative frequency distribution A histogram

A frequency polygon

**(b)** What percentage of the measurements are less than 30?

**(c)** What percentage of the measurements are between 40.0 and 49.99 inclusive?

**(d)** What percentage of the measurements are greater than 34.99?

**(e)** Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

**(f)** How many of the measurements are less than 40?

**2.3.4** David Holben (A-4) studied selenium levels in beef raised in a low selenium region of the United States. The goal of the study was to compare selenium levels in the region-raised beef to selenium levels in cooked venison, squirrel, and beef from other regions of the United States. The data below are the selenium levels calculated on a dry weight basis in mg>100 g for a sample of 53 region- raised cattle.

**(a)** Use these data to construct: A frequency distribution

A relative frequency distribution

A cumulative frequency distribution

A cumulative relative frequency distribution A histogram

A frequency polygon

**(b)** Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

**(c)** How many of the measurements are greater than 40? **(d)** What percentage of the measurements are less than 25?

2.3.5 The following table shows the number of hours 45 hospital patients slept following the adminis- tration of a certain anesthetic.

**(a)** Use these data to construct:

A frequency distribution

A relative frequency distribution

A histogram

A frequency polygon

**(b)** Describe these data relative to symmetry and skewness as discussed in Exercise 2.3.1, part h.

2.3.6 The following are the number of babies born during a year in 60 community hospitals

**(a)** From these data construct:

A frequency distribution

A relative frequency distribution

A frequency polygon

**(b)** Describe these data relative to symmetry and skewness as discussed in Exercise 2.3.1, part h.

**2.3.7** In a study of physical endurance levels of male college freshman, the endurance scores based on several exercise routines were collected.

**(a)** From these data construct:

A frequency distribution

A relative frequency distribution

A frequency polygon

A histogram

**(b)** Describe these data relative to symmetry and skewness as discussed in Exercise 2.3.1, part h

**2.3.8** The following are the ages of 30 patients seen in the emergency room of a hospital on a Friday night. Construct a stem-and-leaf display from these data. Describe these data relative to symme- try and skewness as discussed in Exercise 2.3.1, part h.  
   
**2.3.9** The following are the emergency room charges made to a sample of 25 patients at two city hos- pitals. Construct a stem-and-leaf display for each set of data. What does a comparison of the two displays suggest regarding the two hospitals? Describe the two sets of data with respect to sym- metry and skewness as discussed in Exercise 2.3.1, part h.

2.3.10 Refer to the ages of patients discussed in Example 1.4.1 and displayed in Table 1.4.1.

**(a)** Use class interval widths of 5 and construct:

A frequency distribution

A relative frequency distribution

A cumulative frequency distribution

A cumulative relative frequency distribution A histogram

A frequency polygon

**(b)** Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

2.3.11 The objectives of a study by Skjelbo et al. (A-5) were to examine (a) the relationship between chloroguanide metabolism and efficacy in malaria prophylaxis and (b) the mephenytoin metabo- lism and its relationship to chloroguanide metabolism among Tanzanians. From information pro- vided by urine specimens from the 216 subjects, the investigators computed the ratio of unchanged *S*-mephenytoin to *R*-mephenytoin (*S/R* ratio). The results were as follows:

**(a)** From these data construct the following distributions: frequency, relative frequency, cumula- tive frequency, and cumulative relative frequency; and the following graphs: histogram, frequency polygon, and stem-and-leaf plot.

**(b)** Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

**(c)** The investigators defined as poor metabolizers of mephenytoin any subject with an *S/R* mepheny- toin ratio greater than .9. How many and what percentage of the subjects were poor metabolizers?

**(d)** How many and what percentage of the subjects had ratios less than .7? Between .3 and .6999 inclusive? Greater than .4999?

**2.3.12** Schmidt et al. (A-6) conducted a study to investigate whether autotransfusion of shed mediastinal blood could reduce the number of patients needing homologous blood transfusion and reduce the amount of transfused shows the heights in homologous blood if fixed transfusion criteria were used. The following table centimeters of the 109 subjects of whom 97 were males.

**(a)** For these data construct the following distributions: frequency, relative frequency, cumulative frequency, and cumulative relative frequency; and the following graphs: histogram, frequency poly- gon, and stem-and-leaf plot.

**(b)** Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

**(c)** How do you account for the shape of the distribution of these data? **(d)** How tall were the tallest 6.42 percent of the subjects?

**(e)** How tall were the shortest 10.09 percent of the subjects?

**2.4 DESCRIPTIVE STATISTICS: MEASURES OF CENTRAL TENDENCY**

Although frequency distributions serve useful purposes, there are many situations that require other types of data summarization. What we need in many instances is the ability to summarize the data by means of a single number called a *descriptive measure*. Descriptive measures may be computed from the data of a sample or the data of a population. To distinguish between them we have the following definitions:

Si bien las distribuciones de frecuencia son útiles, existen muchas situaciones que requieren otros tipos de resumen de datos. En muchos casos, lo que necesitamos es la capacidad de resumir los datos mediante un único número llamado medida descriptiva. Las medidas descriptivas pueden calcularse a partir de los datos de una muestra o de una población. Para distinguirlas, tenemos las siguientes definiciones:

**DEFINITIONS**

**1. A descriptive measure computed from the data of a sample is called a *statistic*.**

**2. A descriptive measure computed from the data of a population is called a *parameter*.**

Several types of descriptive measures can be computed from a set of data. In this chapter, however, we limit discussion to *measures of central tendency* and *measures of dispersion.* We consider measures of central tendency in this section and measures of disper- sion in the following one.

Se pueden calcular varios tipos de medidas descriptivas a partir de un conjunto de datos. Sin embargo, en este capítulo, nos limitamos a las medidas de tendencia central y de dispersión. En esta sección, consideramos las medidas de tendencia central y en la siguiente, las de dispersión.

In each of the measures of central tendency, of which we discuss three, we have a single value that is considered to be typical of the set of data as a whole. Measures of central tendency convey information regarding the average value of a set of values. As we will see, the word *average* can be defined in different ways.

En cada una de las medidas de tendencia central, de las cuales analizaremos tres, tenemos un único valor que se considera típico del conjunto de datos. Las medidas de tendencia central proporcionan información sobre el valor promedio de un conjunto de valores. Como veremos, el término promedio puede definirse de diferentes maneras.

The three most commonly used measures of central tendency are the *mean,* the *median,* and the *mode*.

**Arithmetic Mean** The most familiar measure of central tendency is the arith- metic mean. It is the descriptive measure most people have in mind when they speak of the “average.” The adjective *arithmetic* distinguishes this mean from other means that can be computed. Since we are not covering these other means in this book, we shall refer to the arithmetic mean simply as the *mean*. The mean is obtained by adding all the values in a population or sample and dividing by the number of values that are added.

**EXAMPLE 2.4.1**

We wish to obtain the mean age of the population of 189 subjects represented in Table 1.4.1.

**Solution:** We proceed as follows:

48+35+46+ Á +73+66

mean age = 189 = 55.032 s

The three dots in the numerator represent the values we did not show in order to save space.

**General Formula for the Mean** It will be convenient if we can generalize the procedure for obtaining the mean and, also, represent the procedure in a more com- pact notational form. Let us begin by designating the random variable of interest by the capital letter *X*. In our present illustration we let *X* represent the random variable, age. Specific values of a random variable will be designated by the lowercase letter *x*. To dis- tinguish one value from another, we attach a subscript to the *x* and let the subscript refer to the first, the second, the third value, and so on. For example, from Table 1.4.1 we have

*x*1 = 48, *x*2 = 35, . . . , *x*189 = 66

In general, a typical value of a random variable will be designated by *xi* and the final value, in a finite population of values, by *xN*, where *N* is the number of values in the population. Finally, we will use the Greek letter m to stand for the population mean. We may now write the general formula for a finite population mean as follows:



The symbol g instructs us to add all values of the variable from the first to the last.

This symbol g, called the *summation sign,* will be used extensively in this book. When from the context it is obvious which values are to be added, the symbols above and below g will be omitted.

**The Sample Mean** When we compute the mean for a sample of values, the pro- cedure just outlined is followed with some modifications in notation. We use *x* to desig- nate the sample mean and *n* to indicate the number of values in the sample. The sample mean then is expressed as

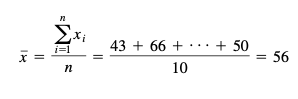


**EXAMPLE 2.4.2**

In Chapter 1 we selected a simple random sample of 10 subjects from the population of subjects represented in Table 1.4.1. Let us now compute the mean age of the 10 subjects in our sample.

**Solution:**

We recall (see Table 1.4.2) that the ages of the 10 subjects in our sam- ple were *x*1 = 43,*x*2 = 66,*x*3 = 61,*x*4 = 64,*x*5 = 65,*x*6 = 38,*x*7 = 59, *x*8 = 57,*x*9 = 57,*x*10 = 50. Substitution of our sample data into Equa- tion 2.4.2 gives



**Properties of the Mean** The arithmetic mean possesses certain properties, some desirable and some not so desirable. These properties include the following:

1. Uniqueness. For a given set of data there is one and only one arithmetic mean.
2. Simplicity. The arithmetic mean is easily understood and easy to compute.
3. Since each and every value in a set of data enters into the computation of the mean, it is affected by each value. Extreme values, therefore, have an influence on the mean and, in some cases, can so distort it that it becomes undesirable as a measure central tendency.

As an example of how extreme values may affect the mean, consider the following situation. Suppose the five physicians who practice in an area are surveyed to deter- mine their charges for a certain procedure. Assume that they report these charges: $75, $75, $80, $80, and $280. The mean charge for the five physicians is found to be $118, a value that is not very representative of the set of data as a whole. The single atypical value had the effect of inflating the mean.

**Median** The median of a finite set of values is that value which divides the set into two equal parts such that the number of values equal to or greater than the median is equal to the number of values equal to or less than the median. If the number of values is odd, the median will be the middle value when all values have been arranged in order of magnitude. When the number of values is even, there is no single middle value. Instead there are two middle values. In this case the median is taken to be the mean of these two middle values, when all values have been arranged in the order of their magnitudes. In other words, the median observation of a data set is the 1*n* + 12>2th one when the observation have been ordered. If, for example, we have 11 observations, the median is the 111 + 12>2 = 6th ordered observation. If we have 12 observations the median is the 112 + 12>2 = 6.5th ordered observation and is a value halfway between the 6th and 7th ordered observations.

**EXAMPLE 2.4.3**

Let us illustrate by finding the median of the data in Table 2.2.1.

**Solution:** The values are already ordered so we need only to find the two middle values. The middle value is the 1*n* + 12>2 = 1189 + 12>2 = 190>2 = 95th one. Counting from the smallest up to the 95th value we see that it is 54. Thus the median age of the 189 subjects is 54 years. s

**EXAMPLE 2.4.4**

We wish to find the median age of the subjects represented in the sample described in Example 2.4.2.

**Solution:** Arraying the 10 ages in order of magnitude from smallest to largest gives 38, 43, 50, 57, 57, 59, 61, 64, 65, 66. Since we have an even number of ages, there is no middle value. The two middle values, however, are 57 and 59. The median, then, is 157 + 592>2 = 58. s

**Properties of the Median** Properties of the median include the following:

1. Uniqueness. As is true with the mean, there is only one median for a given set of data.
2. Simplicity. The median is easy to calculate.
3. It is not as drastically affected by extreme values as is the mean

**The Mode** The mode of a set of values is that value which occurs most frequently. If all the values are different there is no mode; on the other hand, a set of values may have more than one mode.

**EXAMPLE 2.4.5**

Find the modal age of the subjects whose ages are given in Table 2.2.1.

**Solution:** A count of the ages in Table 2.2.1 reveals that the age 53 occurs most frequently (17 times). The mode for this population of ages is 53. s

For an example of a set of values that has more than one mode, let us consider a laboratory with 10 employees whose ages are 20, 21, 20, 20, 34, 22, 24, 27, 27, and 27. We could say that these data have two modes, 20 and 27. The sample consisting of the values 10, 21, 33, 53, and 54 has no mode since all the values are different.

The mode may be used for describing qualitative data. For example, suppose the patients seen in a mental health clinic during a given year received one of the following diagnoses: mental retardation, organic brain syndrome, psychosis, neurosis, and person- ality disorder. The diagnosis occurring most frequently in the group of patients would be called the modal diagnosis.

An attractive property of a data distribution occurs when the mean, median, and mode are all equal. The well-known “bell-shaped curve” is a graphical representation of a distribution for which the mean, median, and mode are all equal. Much statistical infer- ence is based on this distribution, the most common of which is the normal distribution. The normal distribution is introduced in Section 4.6 and discussed further in subsequent chapters. Another common distribution of this type is the *t*-distribution, which is intro- duced in Section 6.3.

**Skewness** Data distributions may be classified on the basis of whether they are symmetric or asymmetric. If a distribution is symmetric, the left half of its graph (his- togram or frequency polygon) will be a mirror image of its right half. When the left half and right half of the graph of a distribution are not mirror images of each other, the dis- tribution is asymmetric.

Las distribuciones de datos se pueden clasificar según su simetría o asimetría. Si una distribución es simétrica, la mitad izquierda de su gráfico (histograma o polígono de frecuencias) será una imagen especular de su mitad derecha. Cuando las mitades izquierda y derecha del gráfico de una distribución no son imágenes especulares entre sí, la distribución es asimétrica.

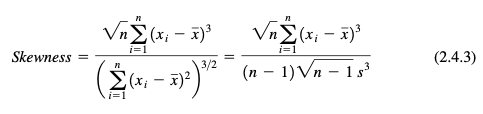
**DEFINITION**

**If the graph (histogram or frequency polygon) of a distribution is asymmetric, the distribution is said to be *skewed*. If a distribution is not symmetric because its graph extends further to the right than to the left, that is, if it has a long tail to the right, we say that the distri- bution is *skewed to the right* or is *positively skewed*. If a distribution is not symmetric because its graph extends further to the left than to the right, that is, if it has a long tail to the left, we say that the distribu- tion is *skewed to the left* or is *negatively skewed*.**

Si el gráfico (histograma o polígono de frecuencias) de una distribución es asimétrico, se dice que la distribución está sesgada. Si una distribución no es simétrica porque su gráfico se extiende más a la derecha que a la izquierda, es decir, si tiene una cola larga a la derecha, decimos que la distribución está sesgada a la derecha o es positivamente sesgada. Si una distribución no es simétrica porque su gráfico se extiende más a la izquierda que a la derecha, es decir, si tiene una cola larga a la izquierda, decimos que la distribución está sesgada a la izquierda o es negativamente sesgada.

A distribution will be skewed to the right, or positively skewed, if its mean is greater than its mode. A distribution will be skewed to the left, or negatively skewed, if its mean is less than its mode. Skewness can be expressed as follows:

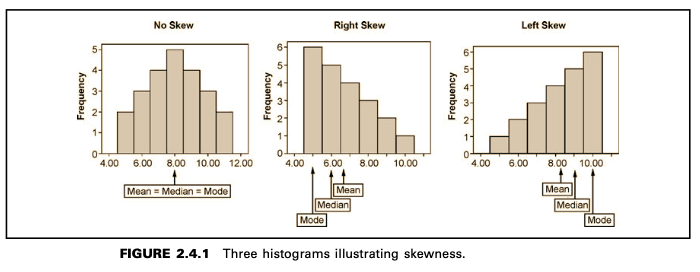
Una distribución estará sesgada hacia la derecha, o positivamente sesgada, si su media es mayor que su moda. Una distribución estará sesgada hacia la izquierda, o negativamente sesgada, si su media es menor que su moda. La asimetría se puede expresar de la siguiente manera:



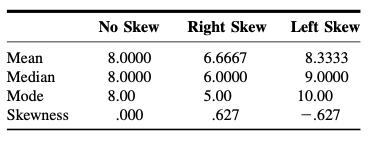
In Equation 2.4.3, *s* is the standard deviation of a sample as defined in Equation 2.5.4. Most computer statistical packages include this statistic as part of a standard printout. A value of skewness 􏰃 0 indicates positive skewness and a value of skewness 􏰄 0 indi- cates negative skewness. An illustration of skewness is shown in Figure 2.4.1.

**EXAMPLE 2.4.6**

Consider the three distributions shown in Figure 2.4.1. Given that the histograms repre- sent frequency counts, the data can be easily re-created and entered into a statistical pack- age. For example, observation of the “No Skew” distribution would yield the following data: 5, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11. Values can be



obtained from the skewed distributions in a similar fashion. Using SPSS software, the following descriptive statistics were obtained for these three distributions



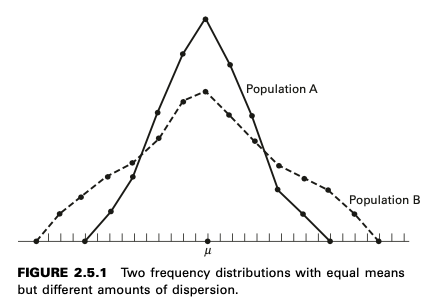
**2.5 DESCRIPTIVE STATISTICS: MEASURES OF DISPERSION**

The *dispersion* of a set of observations refers to the variety that they exhibit. A measure of dispersion conveys information regarding the amount of variability present in a set of data. If all the values are the same, there is no dispersion; if they are not all the same, dispersion is present in the data. The amount of dispersion may be small when the val- ues, though different, are close together. Figure 2.5.1 shows the frequency polygons for two populations that have equal means but different amounts of variability. Population B, which is more variable than population A, is more spread out. If the values are widely scattered, the dispersion is greater. Other terms used synonymously with dispersion include *variation, spread,* and *scatter*.

La dispersión de un conjunto de observaciones se refiere a la variedad que exhiben. Una medida de dispersión transmite información sobre la cantidad de variabilidad presente en un conjunto de datos. Si todos los valores son iguales, no hay dispersión; si no son todos iguales, hay dispersión presente en los datos. La cantidad de dispersión puede ser pequeña cuando los valores, aunque diferentes, están cerca uno del otro. La Figura 2.5.1 muestra los polígonos de frecuencia para dos poblaciones que tienen medias iguales pero diferentes cantidades de variabilidad. La población B, que es más variable que la población A, está más dispersa. Si los valores están muy dispersos, la dispersión es mayor. Otros términos utilizados como sinónimos de dispersión incluyen variación, dispersión y dispersión.

**The Range** One way to measure the variation in a set of values is to compute the *range*. The range is the difference between the largest and smallest value in a set of observations. If we denote the range by *R*, the largest value by *xL*, and the smallest value by *xs*, we compute the range as follows:

*R* = *xL* - *xS*



**EXAMPLE 2.5.1**

We wish to compute the range of the ages of the sample subjects discussed in Example 2.4.2.

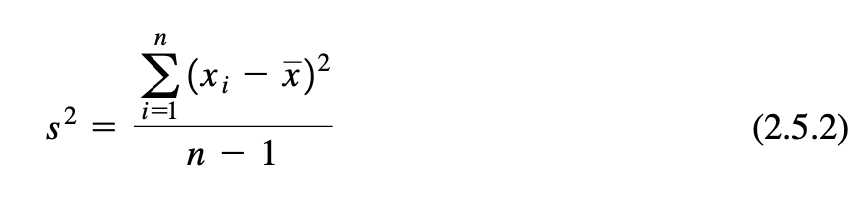
**Solution:** Since the youngest subject in the sample is 30 years old and the oldest is 82, we compute the range to be

*R* = 82 - 30 = 52 s

The usefulness of the range is limited. The fact that it takes into account only two val- ues causes it to be a poor measure of dispersion. The main advantage in using the range is the simplicity of its computation. Since the range, expressed as a single measure, imparts minimal information about a data set and therefore, is of limited use, it is often preferable to express the range as a number pair, [*x*S, *x*L], in which *x*S and *x*L are the smallest and largest values in the data set, respectively. For the data in Example 2.5.1, we may express the range as the number pair [30, 82]. Although this is not the tradi- tional expression for the range, it is intuitive to imagine that knowledge of the minimum and maximum values in this data set would convey more information than knowing only that the range is equal to 52. An infinite number of distributions, each with quite differ- ent minimum and maximum values, may have a range of 52.

**The Variance** When the values of a set of observations lie close to their mean, the dispersion is less than when they are scattered over a wide range. Since this is true, it would be intuitively appealing if we could measure dispersion relative to the scatter of the values about their mean. Such a measure is realized in what is known as the *vari- ance*. In computing the variance of a sample of values, for example, we subtract the mean from each of the values, square the resulting differences, and then add up the squared differences. This sum of the squared deviations of the values from their mean is divided by the sample size, minus 1, to obtain the sample variance. Letting *s*2 stand for the sample variance, the procedure may be written in notational form as follows:

La varianza. Cuando los valores de un conjunto de observaciones se encuentran cerca de su media, la dispersión es menor que cuando están dispersos en un amplio rango. Dado que esto es cierto, sería intuitivamente atractivo si pudiéramos medir la dispersión relativa a la dispersión de los valores alrededor de su media. Dicha medida se realiza en lo que se conoce como la varianza. Al calcular la varianza de una muestra de valores, por ejemplo, restamos la media de cada uno de los valores, elevamos al cuadrado las diferencias resultantes y luego sumamos las diferencias al cuadrado. Esta suma de las desviaciones al cuadrado de los valores con respecto a su media se divide por el tamaño de la muestra, menos 1, para obtener la varianza muestral. Si s² representa la varianza muestral, el procedimiento puede escribirse en forma de notación de la siguiente manera:



**EXAMPLE 2.5.2** Let us illustrate by computing the variance of the ages of the subjects discussed in Example 2.4.2.

**Solution:**

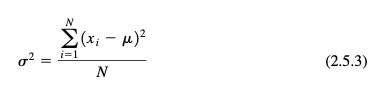
**Degrees of Freedom** The reason for dividing by *n* - 1 rather than *n*, as we might have expected, is the theoretical consideration referred to as *degrees of freedom.*

Grados de libertad La razón para dividir por n - 1 en lugar de n, como podríamos haber esperado, es la consideración teórica conocida como grados de libertad.

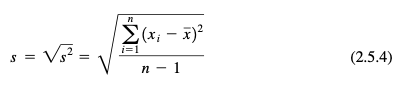
In computing the variance, we say that we have *n* - 1 *degrees of freedom.* We reason as follows. The sum of the deviations of the values from their mean is equal to zero, as can be shown. If, then, we know the values of *n* - 1 of the deviations from the mean, we know the *n*th one, since it is automatically determined because of the necessity for all *n* values to add to zero. From a practical point of view, dividing the squared differ- ences by *n* - 1 rather than *n* is necessary in order to use the sample variance in the inference procedures discussed later. The concept of degrees of freedom will be revis- ited in a later Chapter. Students interested in pursuing the matter further at this time should refer to the article by Walker (2).

Al calcular la varianza, decimos que tenemos n - 1 grados de libertad. Razonamos de la siguiente manera. La suma de las desviaciones de los valores con respecto a su media es igual a cero, como se puede demostrar. Si, entonces, conocemos los valores de n - 1 de las desviaciones con respecto a la media, conocemos la n-ésima, ya que se determina automáticamente debido a la necesidad de que todos los valores n sumen cero. Desde un punto de vista práctico, dividir las diferencias al cuadrado por n - 1 en lugar de n es necesario para utilizar la varianza muestral en los procedimientos de inferencia que se analizan más adelante. El concepto de grados de libertad se revisará en un capítulo posterior. Los estudiantes interesados ​​en profundizar en el tema en este momento deben consultar el artículo de Walker (2).

When we compute the variance from a finite population of *N* values, the proce- duresoutlined above are followed except that we subtract m from each *x* and divide by *N* rather than *N* - 1. If we let s2 stand for the finite population variance, the formula is as follows:



**Standard Deviation** The variance represents squared units and, therefore, is not an appropriate measure of dispersion when we wish to express this concept in terms of the original units. To obtain a measure of dispersion in original units, we merely take the square root of the variance. The result is called the *standard deviation.* In general, the standard deviation of a sample is given by



The standard deviation of a finite population is obtained by taking the square root of the quantity obtained by Equation 2.5.3.

**The Coefficient of Variation** The standard deviation is useful as a measure of variation within a given set of data. When one desires to compare the dispersion in two sets of data, however, comparing the two standard deviations may lead to fallacious results. It may be that the two variables involved are measured in different units. For example, we may wish to know, for a certain population, whether serum cholesterol levels, measured in milligrams per 100 ml, are more variable than body weight, measured in pounds.

El coeficiente de variación. La desviación estándar es útil como medida de variación dentro de un conjunto de datos. Sin embargo, al comparar la dispersión en dos conjuntos de datos, comparar ambas desviaciones estándar puede generar resultados erróneos. Es posible que las dos variables involucradas se midan en unidades diferentes. Por ejemplo, podríamos querer saber, para una población determinada, si los niveles de colesterol sérico, medidos en miligramos por 100 ml, son más variables que el peso corporal, medido en libras.

Furthermore, although the same unit of measurement is used, the two means may be quite different. If we compare the standard deviation of weights of first-grade chil- dren with the standard deviation of weights of high school freshmen, we may find that the latter standard deviation is numerically larger than the former, because the weights themselves are larger, not because the dispersion is greater.

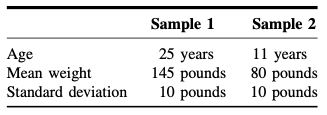
What is needed in situations like these is a measure of relative variation rather than absolute variation. Such a measure is found in the *coefficient of variation,* which expresses the standard deviation as a percentage of the mean. The formula is given by



We see that, since the mean and standard deviations are expressed in the same unit of measurement, the unit of measurement cancels out in computing the coefficient of vari- ation. What we have, then, is a measure that is independent of the unit of measurement.

**EXAMPLE 2.5.3**

Suppose two samples of human males yield the following results:



We wish to know which is more variable, the weights of the 25-year-olds or the weights of the 11-year-olds.

**Solution:**

A comparison of the standard deviations might lead one to conclude that the two samples possess equal variability. If we compute the coefficients of variation, however, we have for the 25-year-olds

C.V. = 145 11002 = 6.9%

and for the 11-year-olds

C.V. = 80 11002 = 12.5%

If we compare these results, we get quite a different impression. It is clear from this example that variation is much higher in the sample of 11-year- olds than in the sample of 25-year-olds.

The coefficient of variation is also useful in comparing the results obtained by different persons who are conducting investigations involving the same variable. Since the coefficient of variation is independent of the scale of measurement, it is a useful statistic for comparing the variability of two or more variables measured on different scales. We could, for example, use the coefficient of variation to compare the variabil- ity in weights of one sample of subjects whose weights are expressed in pounds with the variability in weights of another sample of subjects whose weights are expressed in kilograms.

**Computer Analysis** Computer software packages provide a variety of possibil- ities in the calculation of descriptive measures. Figure 2.5.2 shows a printout of the descriptive measures available from the MINITAB package. The data consist of the ages from Example 2.4.2.

In the printout *Q*1 and *Q*3 are the first and third quartiles, respectively. These meas- ures are described later in this chapter. N stands for the number of data observations, and N\* stands for the number of missing values. The term SEMEAN stands for *standard error of the mean.* This measure will be discussed in detail in a later chapter. Figure 2.5.3 shows, for the same data, the SAS® printout obtained by using the PROC MEANS statement.

**Percentiles and Quartiles** The mean and median are special cases of a fam- ily of parameters known as *location parameters.* These descriptive measures are called location parameters because they can be used to designate certain positions on the hori- zontal axis when the distribution of a variable is graphed. In that sense the so-called loca- tion parameters “locate” the distribution on the horizontal axis. For example, a distribution with a median of 100 is located to the right of a distribution with a median of 50 when the two distributions are graphed. Other location parameters include percentiles and quar- tiles. We may define a percentile as follows:

Percentiles y cuartiles. La media y la mediana son casos especiales de una familia de parámetros conocidos como parámetros de ubicación. Estas medidas descriptivas se llaman parámetros de ubicación porque pueden usarse para designar ciertas posiciones en el eje horizontal cuando se grafica la distribución de una variable. En ese sentido, los llamados parámetros de ubicación “ubican” la distribución en el eje horizontal. Por ejemplo, una distribución con una mediana de 100 se ubica a la derecha de una distribución con una mediana de 50 cuando se grafican las dos distribuciones. Otros parámetros de ubicación incluyen percentiles y cuartiles. Podemos definir un percentil de la siguiente manera:

**DEFINITION**

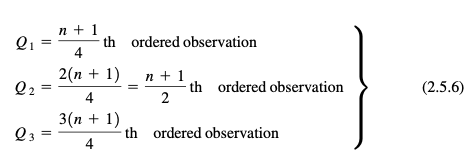
**Given a set of *n* observations *x*1, *x*2, . . . *xn*, the *p*th percentile *P* is the value of *X* such that *p* percent or less of the observations are less than *P* and** (**100** 􏰌 ***p***) **percent or less of the observations are greater than *P*.**

Dado un conjunto de n observaciones x1, x2, . . . xn, el percentil p P es el valor de X tal que el p por ciento o menos de las observaciones son menores que P y el (100 ≤ p) por ciento o menos de las observaciones son mayores que P.

Subscripts on *P* serve to distinguish one percentile from another. The 10th per- centile, for example, is designated *P*10, the 70th is designated *P*70, and so on. The 50th percentile is the median and is designated *P*50. The 25th percentile is often referred to as the *first quartile* and denoted *Q*1. The 50th percentile (the median) is referred to as the second or *middle quartile* and written *Q* 2, and the 75th percentile is referred to as the *third quartile*, *Q* 3.

Los subíndices de P sirven para distinguir un percentil de otro. El percentil 10, por ejemplo, se designa como P10, el 70, como P70, y así sucesivamente. El percentil 50 es la mediana y se designa como P50. El percentil 25 se suele denominar primer cuartil y se denota como Q1. El percentil 50 (la mediana) se denomina segundo cuartil o cuartil medio y se escribe como Q2, y el percentil 75, tercer cuartil, como Q3.

When we wish to find the quartiles for a set of data, the following formulas are used:



**Interquartile Range** As we have seen, the range provides a crude measure of the variability present in a set of data. A disadvantage of the range is the fact that it is computed from only two values, the largest and the smallest. A similar measure that reflects the variability among the middle 50 percent of the observations in a data set is the *interquartile range.*

**DEFINITION**

**The interquartile range (IQR) is the difference between the third and first quartiles: that is,**

IQR 􏰍 ***Q***3 􏰌 ***Q***1 **(2.5.7)**

A large IQR indicates a large amount of variability among the middle 50 percent of the relevant observations, and a small IQR indicates a small amount of variability among the relevant observations. Since such statements are rather vague, it is more informative to compare the interquartile range with the range for the entire data set. A comparison may be made by forming the ratio of the IQR to the range (*R*) and multiplying by 100. That is, 100(IQR/*R*) tells us what percent the IQR is of the overall range.

**Kurtosis** Just as we may describe a distribution in terms of skewness, we may describe a distribution in terms of kurtosis.

Curtosis Así como podemos describir una distribución en términos de asimetría, podemos describir una distribución en términos de curtosis.

**DEFINITION**

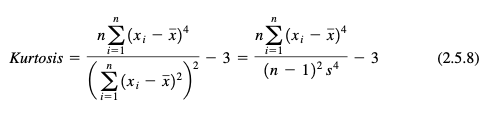
***Kurtosis* is a measure of the degree to which a distribution is “peaked” or flat in comparison to a normal distribution whose graph is charac- terized by a bell-shaped appearance.**

La curtosis es una medida del grado en el cual una distribución es “picuda” o plana en comparación con una distribución normal cuyo gráfico se caracteriza por una apariencia en forma de campana.

A distribution, in comparison to a normal distribution, may possess an excessive propor- tion of observations in its tails, so that its graph exhibits a flattened appearance. Such a distribution is said to be *platykurtic*. Conversely, a distribution, in comparison to a nor- mal distribution, may possess a smaller proportion of observations in its tails, so that its graph exhibits a more peaked appearance. Such a distribution is said to be *leptokurtic*. A normal, or bell-shaped distribution, is said to be *mesokurtic*.

Una distribución, en comparación con una distribución normal, puede presentar una proporción excesiva de observaciones en sus colas, de modo que su gráfico presenta una apariencia aplanada. Dicha distribución se denomina platicúrtica. Por el contrario, una distribución, en comparación con una distribución normal, puede presentar una proporción menor de observaciones en sus colas, de modo que su gráfico presenta una apariencia más picuda. ​​Dicha distribución se denomina leptocúrtica. Una distribución normal, o con forma de campana, se denomina mesocúrtica.

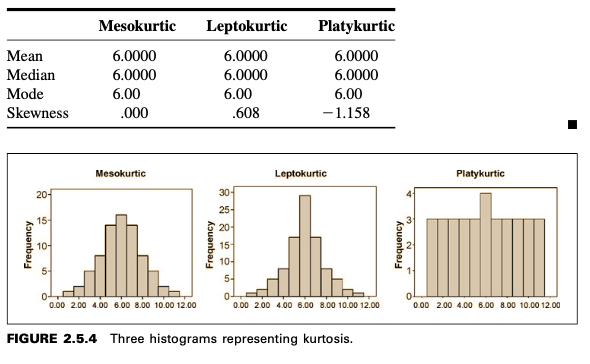
Kurtosis can be expressed as



Manual calculation using Equation 2.5.8 is usually not necessary, since most statisti- cal packages calculate and report information regarding kurtosis as part of the descrip- tive statistics for a data set. Note that each of the two parts of Equation 2.5.8 has been reduced by 3. A perfectly mesokurtic distribution has a kurtosis measure of 3 based on the equation. Most computer algorithms reduce the measure by 3, as is done in Equation 2.5.8, so that the kurtosis measure of a mesokurtic distribution will be equal to 0. A leptokurtic distribution, then, will have a kurtosis measure 􏰆 0, and a platykur- tic distribution will have a kurtosis measure 􏰇 0. Be aware that not all computer pack- ages make this adjustment. In such cases, comparisons with a mesokurtic distribution are made against 3 instead of against 0. Graphs of distributions representing the three types of kurtosis are shown in Figure 2.5.4.

**EXAMPLE 2.5.4**

Consider the three distributions shown in Figure 2.5.4. Given that the histograms rep- resent frequency counts, the data can be easily re-created and entered into a statistical package. For example, observation of the “mesokurtic” distribution would yield the fol- lowing data: 1, 2, 2, 3, 3, 3, 3, 3, ... , 9, 9, 9, 9, 9, 10, 10, 11. Values can be obtained from the other distributions in a similar fashion. Using SPSS software, the following descriptive statistics were obtained for these three distributions:



**Box-and-Whisker Plots** A useful visual device for communicating the infor- mation contained in a data set is the *box-and-whisker plot.* The construction of a box- and-whisker plot (sometimes called, simply, a *boxplot*) makes use of the quartiles of a data set and may be accomplished by following these five steps:

1. **Represent the variable of interest on the horizontal axis.**
2. **Draw a box in the space above the horizontal axis in such a way that the left end of the box aligns with the first quartile *Q* 1 and the right end of the box aligns with the third quartile *Q* 3.**
3. **Divide the box into two parts by a vertical line that aligns with the median *Q* 2.**
4. **Draw a horizontal line called a *whisker* from the left end of the box to a point that aligns with the smallest measurement in the data set.**
5. **Draw another horizontal line, or whisker, from the right end of the box to a point that aligns with the largest measurement in the data set.**

Examination of a box-and-whisker plot for a set of data reveals information regarding the amount of spread, location of concentration, and symmetry of the data.

The following example illustrates the construction of a box-and-whisker plot.

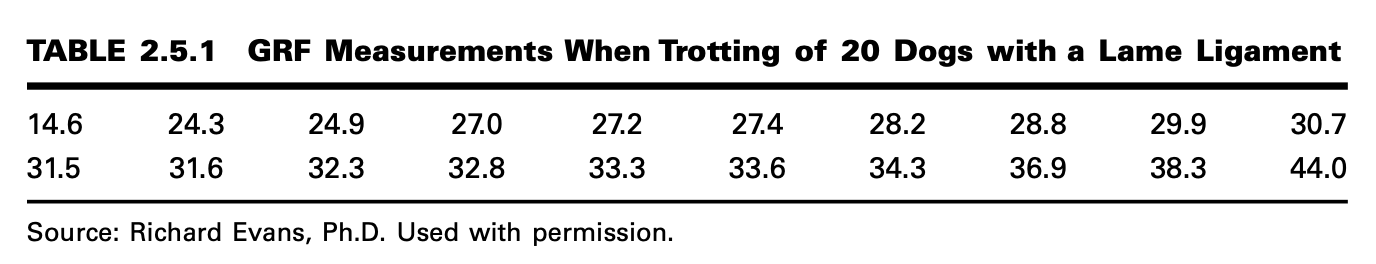
**EXAMPLE 2.5.5**

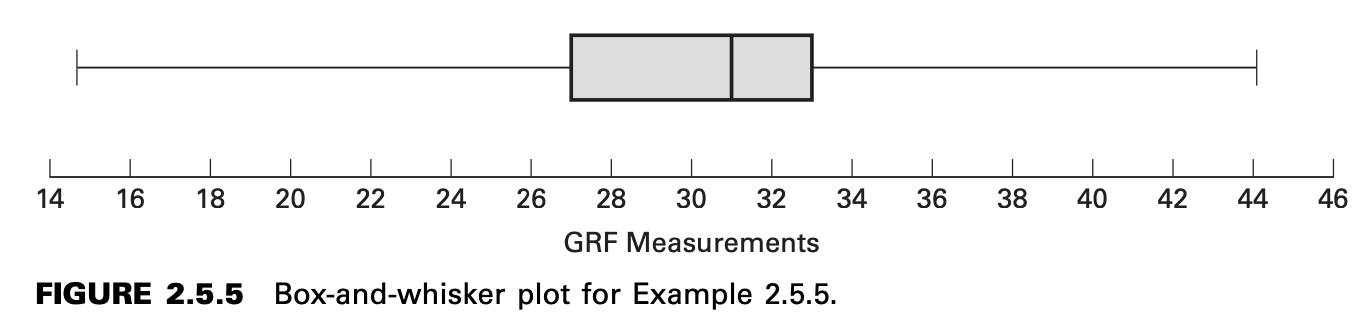
Evans et al. (A-7) examined the effect of velocity on ground reaction forces (GRF) in dogs with lameness from a torn cranial cruciate ligament. The dogs were walked and trotted over a force platform, and the GRF was recorded during a certain phase of their performance. Table 2.5.1 contains 20 measurements of force where each value shown is the mean of five force measurements per dog when trotting.

**Solution:**

The smallest and largest measurements are 14.6 and 44, respectively. The first quartile is the *Q*1 = 120 + 12>4 = 5.25th measurement, which is 27.2 + 1.252127.4 - 27.22 = 27.25. The median is the *Q* 2 + 120 + 12 >2 = 10.5th measurement or 30.7 + 1.52131.5 - 30.72 = 31.1; and the third quartile is the *Q*3 + 3120 + 12>4 = 15.75th measurement, which is equal to 33.3 + 1.752133.6 - 33.32 = 33.525. The interquartile range is IQR = 33.525 - 27.25 = 6.275. The range is 29.4, and the IQR is 10016.275>29.42 = 21 percent of the range. The resulting box-and-whisker plot is shown in Figure 2.5.5. s

Examination of Figure 2.5.5 reveals that 50 percent of the measurements are between about 27 and 33, the approximate values of the first and third quartiles, respec- tively. The vertical bar inside the box shows that the median is about 31.





Many statistical software packages have the capability of constructing box-and- whisker plots. Figure 2.5.6 shows one constructed by MINITAB and one constructed by NCSS from the data of Table 2.5.1. The procedure to produce the MINITAB plot is shown in Figure 2.5.7. The asterisks in Figure 2.5.6 alert us to the fact that the data set contains one unusually large and one unusually small value, called *outliers.* The outliers are the dogs that generated forces of 14.6 and 44. Figure 2.5.6 illustrates the fact that box-and- whisker plots may be displayed vertically as well as horizontally.

An outlier, or a typical observation, may be defined as follows.

**DEFINITION**

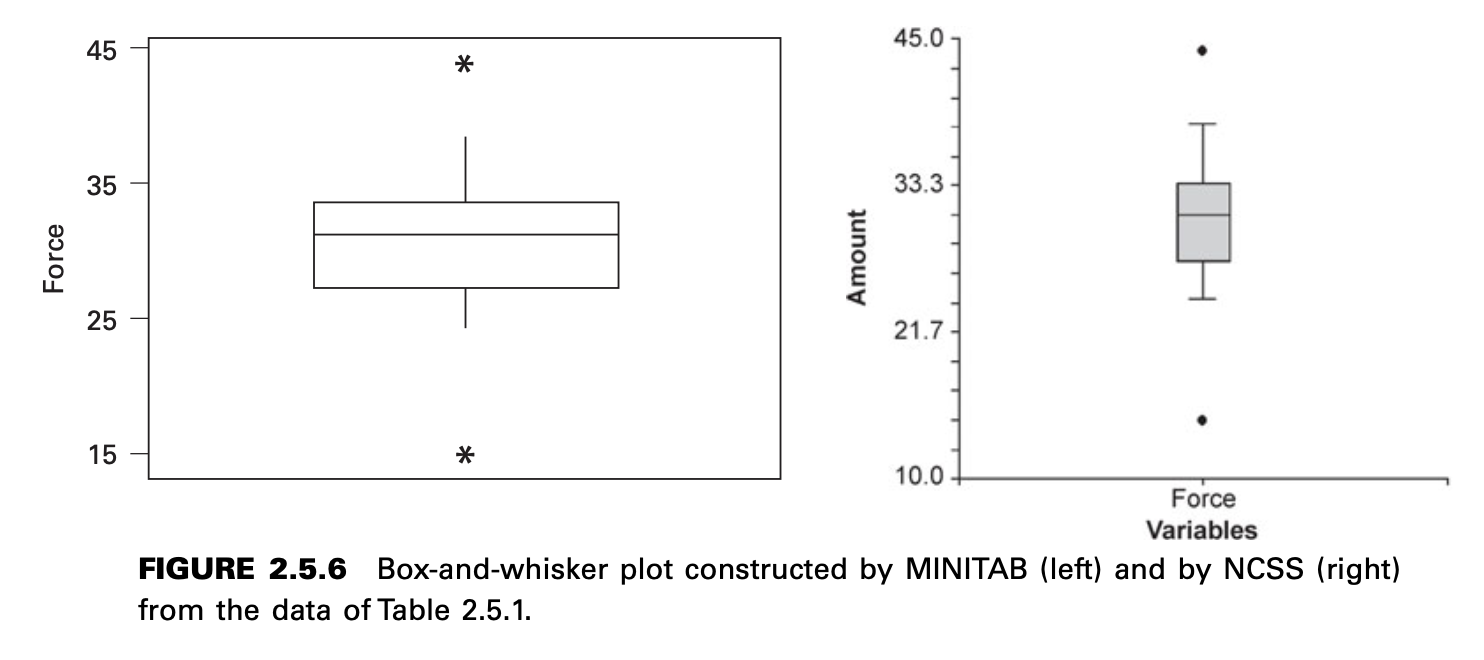
**An *outlier* is an observation whose value, *x*, either exceeds the value of the third quartile by a magnitude greater than 1.5(IQR) or is less than the value of the first quartile by a magnitude greater than 1.5(IQR). That is, an observation of *x***>***Q* 3** 􏰎 **1.5(IQR) or an observation of *x***<***Q* 1** 􏰌 **1.5(IQR) is called an outlier.**

Un valor atípico es una observación cuyo valor, x, supera el valor del tercer cuartil en una magnitud mayor que 1,5 (RIC) o es menor que el valor del primer cuartil en una magnitud mayor que 1,5 (RIC). Es decir, una observación de x > Q 3 ≤ 1,5 (RIC) o una observación de x < Q 1 ≤ 1,5 (RIC) se denomina valor atípico.

For the data in Table 2.5.1 we may use the previously computed values of *Q*1, *Q*3, and IQR to determine how large or how small a value would have to be in order to be con- sidered an outlier. The calculations are as follows:

*x* > 27.25 - 1.516.2752 = 17.8375 and *x* 7 33.525 + 1.516.2752 = 42.9375

For the data in Table 2.5.1, then, an observed value smaller than 17.8375 or larger than 42.9375 would be considered an outlier.



he SAS® statement PROC UNIVARIATE may be used to obtain a box-and- whisker plot. The statement also produces other descriptive measures and displays, including stem-and-leaf plots, means, variances, and quartiles.

**Exploratory Data Analysis** Box-and-whisker plots and stem-and-leaf dis- plays are examples of what are known as *exploratory data analysis* techniques. These techniques, made popular as a result of the work of Tukey (3), allow the investigator to examine data in ways that reveal trends and relationships, identify unique features of data sets, and facilitate their description and summarization.

**EXERCISES**

For each of the data sets in the following exercises compute (a) the mean, (b) the median, (c) the mode, (d) the range, (e) the variance, (f) the standard deviation, (g) the coefficient of variation, and (h) the interquartile range. Treat each data set as a sample. For those exercises for which you think it would be appropriate, construct a box-and-whisker plot and discuss the usefulness in under- standing the nature of the data that this device provides. For each exercise select the measure of central tendency that you think would be most appropriate for describing the data. Give reasons to justify your choice.

**2.5.1** Porcellini et al. (A-8) studied 13 HIV-positive patients who were treated with highly active antiretro- viral therapy (HAART) for at least 6 months. The CD4 T cell counts 1\* 106>*L*2 at baseline for the 13 subjects are listed below.

**2.5.2** Shair and Jasper (A-9) investigated whether decreasing the venous return in young rats would affect ultrasonic vocalizations (USVs). Their research showed no significant change in the number of ultrasonic vocalizations when blood was removed from either the superior vena cava or the carotid artery. Another important variable measured was the heart rate (bmp) during the withdrawal of blood. The table below presents the heart rate of seven rat pups from the experiment involving the carotid artery.

**2.5.3** Butz et al. (A-10) evaluated the duration of benefit derived from the use of noninvasive positive- pressure ventilation by patients with amyotrophic lateral sclerosis on symptoms, quality of life, and survival. One of the variables of interest is partial pressure of arterial carbon dioxide (PaCO2). The values below (mm Hg) reflect the result of baseline testing on 30 subjects as established by arterial blood gas analyses.

**2.5.4** According to Starch et al. (A-11), hamstring tendon grafts have been the “weak link” in anterior cruciate ligament reconstruction. In a controlled laboratory study, they compared two techniques for reconstruction: either an interference screw or a central sleeve and screw on the tibial side. For eight cadaveric knees, the measurements below represent the required force (in newtons) at which initial failure of graft strands occurred for the central sleeve and screw technique.   
**2.5.5** Cardosi et al. (A-12) performed a 4-year retrospective review of 102 women undergoing radical hysterectomy for cervical or endometrial cancer. Catheter-associated urinary tract infection was observed in 12 of the subjects. Below are the numbers of postoperative days until diagnosis of the infection for each subject experiencing an infection.  
   
   
**2.5.6** The purpose of a study by Nozawa et al. (A-13) was to evaluate the outcome of surgical repair of a pars interarticularis defect by segmental wire fixation in young adults with lumbar spondy- lolysis. The authors found that segmental wire fixation historically has been successful in the treatment of nonathletes with spondylolysis, but no information existed on the results of this type of surgery in athletes. In a retrospective study, the authors found 20 subjects who had the sur- gery between 1993 and 2000. For these subjects, the data below represent the duration in months of follow-up care after the operation.

2.5.7 See Exercise 2.3.1.

2.5.8 See Exercise 2.3.2.

2.5.9 See Exercise 2.3.3.

2.5.10 See Exercise 2.3.4.

2.5.11 See Exercise 2.3.5.

2.5.12 See Exercise 2.3.6.

2.5.13 See Exercise 2.3.7.

2.5.14 In a pilot study, Huizinga et al. (A-14) wanted to gain more insight into the psychosocial conse- quences for children of a parent with cancer. For the study, 14 families participated in semistruc- tured interviews and completed standardized questionnaires. Below is the age of the sick parent with cancer (in years) for the 14 families.

**2.6 SUMMARY**

In this chapter various descriptive statistical procedures are explained. These include the organization of data by means of the ordered array, the frequency distribution, the rela- tive frequency distribution, the histogram, and the frequency polygon. The concepts of central tendency and variation are described, along with methods for computing their more common measures: the mean, median, mode, range, variance, and standard devia- tion. The reader is also introduced to the concepts of skewness and kurtosis, and to exploratory data analysis through a description of stem-and-leaf displays and box-and- whisker plots.

We emphasize the use of the computer as a tool for calculating descriptive meas- ures and constructing various distributions from large data sets.

**REVIEW QUESTIONS AND EXERCISES**

1. Define:  
    (a) Stem-and-leaf display

**(b)** Box-and-whisker plot

(c) Percentile

**d)** Quartile

(e) Location parameter

**(f)** Exploratory data analysis

(g) Ordered array

(h) Frequency distribution

(i) Relative frequency distribution

(k) Parameter

**(l)** Frequency polygon

(m) True class limits

1. Define and compare the characteristics of the mean, the median, and the mode.
2. What are the advantages and limitations of the range as a measure of **dispersion?**
3. **Explain the rationale for using *n* - 1 to compute the sample variance.**
4. **What is the purpose of the coefficient of variation?**

**6.** What is the purpose of Sturges’s rule?

1. **What is another name for the 50th percentile (second or middle quartile)?**
2. **Describe from your field of study a population of data where knowledge of the central tendency and dispersion would be useful. Obtain real or realistic synthetic values from this population and compute the mean, median, mode, variance, and standard deviation.**
3. **Collect a set of real, or realistic, data from your field of study and construct a frequency distribu- tion, a relative frequency distribution, a histogram, and a frequency polygon.**
4. **Compute the mean, median, mode, variance, and standard deviation for the data in Exercise 9.**
5. **Find an article in a journal from your field of study in which some measure of central tendency and dispersion have been computed.**
6. **The purpose of a study by Tam et al. (A-15) was to investigate the wheelchair maneuvering in individuals with lower-level spinal cord injury (SCI) and healthy controls. Subjects used a modi- fied wheelchair to incorporate a rigid seat surface to facilitate the specified experimental measure- ments. Interface pressure measurement was recorded by using a high-resolution pressure-sensitive mat with a spatial resolution of 4 sensors per square centimeter taped on the rigid seat support. During static sitting conditions, average pressures were recorded under the ischial tuberosities. The data for measurements of the left ischial tuberosity (in mm Hg) for the SCI and control groups are shown below.  
      
   Control 131 115 124 131 122 117 88 114 150 169**

**SCI 60 150 130 180 163 130 121 119 130 148  
   
Source: Eric W. Tam, Arthur F. Mak, Wai Nga Lam, John H. Evans, and York  
 Y. Chow, “Pelvic Movement and Interface Pressure Distribution During Manual Wheel- chair Propulsion,” *Archives of Physical Medicine and Rehabilitation, 84* (2003),  
 1466 –1472.  
   
(a) Find the mean, median, variance, and standard deviation for the controls.**

**(b) Find the mean, median variance, and standard deviation for the SCI group.  
   
(c) Construct a box-and-whisker plot for the controls.**

**(d) Construct a box-and-whisker plot for the SCI group.  
   
(e) Do you believe there is a difference in pressure readings for controls and SCI subjects in this study?**

1. **Johnson et al. (A-16) performed a retrospective review of 50 fetuses that underwent open fetal myelomeningocele closure. The data below show the gestational age in weeks of the 50 fetuses undergoing the procedure.**

**(a)** Construct a stem-and-leaf plot for these gestational ages.

**(b)** Based on the stem-and-leaf plot, what one word would you use to describe the nature of the data?

**(c)** Why do you think the stem-and-leaf plot looks the way it does?

**(d)** Compute the mean, median, variance, and standard deviation.

14. The following table gives the age distribution for the number of deaths in New York State due to accidents for residents age 25 and older.

For these data construct a cumulative frequency distribution, a relative frequency distribution, and a cumulative relative frequency distribution.

15. Krieser et al. (A-17) examined glomerular filtration rate (GFR) in pediatric renal transplant recip- ients. GFR is an important parameter of renal function assessed in renal transplant recipients. The following are measurements from 19 subjects of GFR measured with diethylenetriamine penta- acetic acid. (Note: some subjects were measured more than once.)

**(a)** Compute mean, median, variance, standard deviation, and coefficient of variation.

**(b)** Construct a stem-and-leaf display.

**(c)** Construct a box-and-whisker plot.

**(d)** What percentage of the measurements is within one standard deviation of the mean? Two standard deviations? Three standard deviations?

16. The following are the cystatin C levels (mg/L) for the patients described in Exercise 15 (A-17). Cystatin C is a cationic basic protein that was investigated for its relationship to GFR levels. In addition, creatinine levels are also given. (Note: Some subjects were measured more than once.)

**(a)** For each variable, compute the mean, median, variance, standard deviation, and coefficient of variation.

**(b)** For each variable, construct a stem-and-leaf display and a box-and-whisker plot.

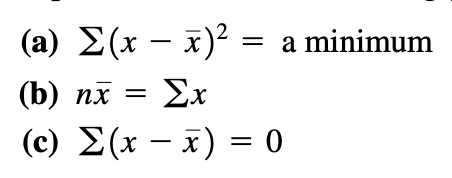
**(c)** Which set of measurements is more variable, cystatin C or creatinine? On what do you base your answer?

1. **Give three synonyms for variation (variability).**
2. **The following table shows the age distribution of live births in Albany County, New York, for 2000.**

For these data construct a cumulative frequency distribution, a relative frequency distribution, and a cumulative relative frequency distribution.

19. Spivack (A-18) investigated the severity of disease associated with *C. difficile* in pediatric inpa- tients. One of the variables they examined was number of days patients experienced diarrhea. The data for the 22 subjects in the study appear below. Compute the mean, median, variance, and stan- dard deviation.

20. Express in words the following properties of the sample mean:



21. Your statistics instructor tells you on the first day of class that there will be five tests during the term. From the scores on these tests for each student, the instructor will compute a measure of central tendency that will serve as the student’s final course grade. Before taking the first test, you must choose whether you want your final grade to be the mean or the median of the five test scores. Which would you choose? Why?

22. Consider the following possible class intervals for use in constructing a frequency distribution of serum cholesterol levels of subjects who participated in a mass screening:

Which set of class intervals do you think is most appropriate for the purpose? State specifically for each one why you think the other two are less desirable.

23. On a statistics test students were asked to construct a frequency distribution of the blood creatine levels (units/liter) for a sample of 300 healthy subjects. The mean was 95, and the standard devi- ation was 40. The following class interval widths were used by the students:

**a)** 1 **(b)** 5 **(c)** 10

**(d)** 15 **(e)** 20 **(f)** 25

Comment on the appropriateness of these choices of widths.

24. Give a health sciences–related example of a population of measurements for which the mean would be a better measure of central tendency than the median.

1. **Give a health sciences–related example of a population of measurements for which the median would be a better measure of central tendency than the mean.**
2. **Indicate for the following variables which you think would be a better measure of central ten- dency, the mean, the median, or mode, and justify your choice:  
      
   (a) Annual incomes of licensed practical nurses in the Southeast.  
    (b) Diagnoses of patients seen in the emergency department of a large city hospital.**

**(c) Weights of high-school male basketball players.**

1. **Refer to Exercise 2.3.11. Compute the mean, median, variance, standard deviation, first quartile, third quartile, and interquartile range. Construct a boxplot of the data. Are the mode, median, and mean equal? If not, explain why. Discuss the data in terms of variability. Compare the IQR with the range. What does the comparison tell you about the variability of the observations?**
2. **Refer to Exercise 2.3.12. Compute the mean, median, variance, standard deviation, first quartile, third quartile, and interquartile range. Construct a boxplot of the data. Are the mode, median, and mean equal? If not, explain why. Discuss the data in terms of variability. Compare the IQR with the range. What does the comparison tell you about the variability of the observations?**
3. **Thilothammal et al. (A-19) designed a study to determine the efficacy of BCG (bacillus Calmette- Guérin) vaccine in preventing tuberculous meningitis. Among the data collected on each subject was a measure of nutritional status (actual weight expressed as a percentage of expected weight for actual height). The following table shows the nutritional status values of the 107 cases studied.**

**(a)** For these data compute the following descriptive measures: mean, median, mode, variance, standard deviation, range, first quartile, third quartile, and IQR.

**(b)** Construct the following graphs for the data: histogram, frequency polygon, stem-and-leaf plot, and boxplot.

**(c)** Discuss the data in terms of variability. Compare the IQR with the range. What does the com- parison tell you about the variability of the observations?

**(d)** What proportion of the measurements are within one standard deviation of the mean? Two standard deviations of the mean? Three standard deviations of the mean?

**(e)** What proportion of the measurements are less than 100?

**(f)** What proportion of the measurements are less than 50?

**Exercises for Use with Large Data Sets Available on the Following Website: www.wiley.com/college/daniel**

Refer to the dataset NCBIRTH800. The North Carolina State Center for Health Statistics and Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (A-20) make publicly available birth and infant death data for all children born in the state of North Carolina. These data can be accessed at www.irss.unc.edu/ncvital/bfd1down.html. Records on birth data go back to 1968. This comprehensive data set for the births in 2001 con- tains 120,300 records. The data represents a random sample of 800 of those births and selected variables. The variables are as follows:

**Variable Label**

**PLURALITY SEX**

**MAGE WEEKS MARITAL RACEMOM**

**HISPMOM**

**GAINED SMOKE**

**DRINK**

**TOUNCES TGRAMS LOW**

**PREMIE**

**Description**

Number of children born of the pregnancy

Sex of child (1 = male, 2 = female)

Age of mother (years)

Completed weeks of gestation (weeks)

Marital status (1 = married, 2 = not married)

Race of mother (0 = other non-White, 1 = White, 2 = Black 3 = American Indian, 4 = Chinese, 5 = Japanese, 6 = Hawaiian, 7 = Filipino, 8 = Other Asian or Pacific Islander)

Mother of Hispanic origin (C = Cuban, M = Mexican, N = Non-Hispanic, O = other and unknown Hispanic, P = Puerto Rican, S = Central/South American, U = not classifiable)

Weight gained during pregnancy (pounds)

0 = mother did not smoke during pregnancy

1 = mother did smoke during pregnancy

0 = mother did not consume alcohol during pregnancy 1 = mother did consume alcohol during pregnancy Weight of child (ounces)

Weight of child (grams)

0 = infant was not low birth weight

1 = infant was low birth weight

0 = infant was not premature

1 = infant was premature

Premature defined at 36 weeks or sooner

**1. 2. 3. 4. 5.**

**6. 7.**

For the variables of MAGE, WEEKS, GAINED, TOUNCES, and TGRAMS: Calculate the mean, median, standard deviation, IQR, and range.

For each, construct a histogram and comment on the shape of the distribution. Do the histograms for TOUNCES and TGRAMS look strikingly similar? Why? Construct box-and-whisker plots for all four variables.

Construct side-by-side box-and-whisker plots for the variable of TOUNCES for women who admit- ted to smoking and women who did not admit to smoking. Do you see a difference in birth weight in the two groups? Which group has more variability?

Construct side-by-side box-and-whisker plots for the variable of MAGE for women who are and are not married. Do you see a difference in ages in the two groups? Which group has more vari- ability? Are the results surprising?

Calculate the skewness and kurtosis of the data set. What do they indicate?

**REFERENCES**

**Methodology References**

1. H. A. STURGES, “The Choice of a Class Interval,” *Journal of the American Statistical Association, 21* (1926), 65–66.
2. HELEN M. WALKER, “Degrees of Freedom,” *Journal of Educational Psychology, 31* (1940), 253–269.
3. JOHN W. TUKEY, *Exploratory Data Analysis,* Addison-Wesley, Reading, MA, 1977.  
      
   **Applications References**

* A-1. FARHAD ATASSI, “Oral Home Care and the Reasons for Seeking Dental Care by Individuals on Renal Dialy- sis,” *Journal of Contemporary Dental Practice, 3* (2002), 31–41.
* A-2. VALLABH JANARDHAN, ROBERT FRIEDLANDER, HOWARD RIINA, and PHILIP EDWIN STIEG, “Identifying Patients at Risk for Postprocedural Morbidity after Treatment of Incidental Intracranial Aneurysms: The Role of Aneurysm Size and Location,” *Neurosurgical Focus, 13* (2002), 1–8.
* A-3. A. HOEKEMA, B. HOVINGA, B. STEGENGA, and L. G. M. De BONT, “Craniofacial Morphology and Obstruc- tive Sleep Apnoea: A Cephalometric Analysis,” *Journal of Oral Rehabilitation, 30* (2003), 690–696.
* A-4. DAVID H. HOLBEN, “Selenium Content of Venison, Squirrel, and Beef Purchased or Produced in Ohio, a Low  
     
  Selenium Region of the United States,” *Journal of Food Science, 67* (2002), 431–433.
* A-5. ERIK SKJELBO, THEONEST K. MUTABINGWA, IB BYGBJERG, KARIN K. NIELSEN, LARS F. GRAM, and KIM BRØSEN, “Chloroguanide Metabolism in Relation to the Efficacy in Malaria Prophylaxis and the *S*-Mephenytoin Oxida-  
     
  tion in Tanzanians,” *Clinical Pharmacology & Therapeutics, 59* (1996), 304–311.
* A-6. HENRIK SCHMIDT, POUL ERIK MORTENSEN, SÀREN LARS FÀLSGAARD, and ESTHER A. JENSEN, “Autotransfu-  
     
  sion After Coronary Artery Bypass Grafting Halves the Number of Patients Needing Blood Transfusion,”  
     
  *Annals of Thoracic Surgery, 61* (1996), 1178–1181.
* A-7. RICHARD EVANS, WANDA GORDON, and MIKE CONZEMIUS, “Effect of Velocity on Ground Reaction Forces in  
     
  Dogs with Lameness Attributable to Tearing of the Cranial Cruciate Ligament,” *American Journal of Veteri-*    
  *nary Research, 64* (2003), 1479–1481.
* A-8. SIMONA PORCELLINI, GUILIANA VALLANTI, SILVIA NOZZA, GUIDO POLI, ADRIANO LAZZARIN, GUISEPPE  
     
  TAMBUSSI, and ANTONIO GRASSIA, “Improved Thymopoietic Potential in Aviremic HIV Infected Individu-  
     
  als with HAART by Intermittent IL-2 Administration,” *AIDS, 17* (2003), 1621–1630.
* A-9. HARRY N. SHAIR and ANNA JASPER, “Decreased Venous Return Is Neither Sufficient nor Necessary to Elicit  
     
  Ultrasonic Vocalization of Infant Rat Pups,” *Behavioral Neuroscience*, *117* (2003), 840–853.
* A-10. M. BUTZ, K. H. WOLLINSKY, U. WIDEMUTH-CATRINESCU, A. SPERFELD, S. WINTER, H. H. MEHRKENS, A. C. LUDOLPH, and H. SCHREIBER, “Longitudinal Effects of Noninvasive Positive-Pressure Ventilation in Patients  
     
  with Amyotophic Lateral Sclerosis,” *American Journal of Medical Rehabilitation, 82* (2003), 597–604.
* A-11. DAVID W. STARCH, JERRY W. ALEXANDER, PHILIP C. NOBLE, SURAJ REDDY, and DAVID M. LINTNER, “Multi- stranded Hamstring Tendon Graft Fixation with a Central Four-Quadrant or a Standard Tibial Interference Screw  
     
  for Anterior Cruciate Ligament Reconstruction,” *American Journal of Sports Medicine, 31* (2003), 338–344.
* A-12. RICHARD J. CARDOSI, ROSEMARY CARDOSI, EDWARD C. GRENDYS Jr., JAMES V. FIORICA, and MITCHEL S. HOFFMAN, “Infectious Urinary Tract Morbidity with Prolonged Bladder Catheterization After Radical Hys-  
     
  terectomy,” *American Journal of Obstetrics and Gynecology, 189* (2003), 380–384.
* A-13. SATOSHI NOZAWA, KATSUJI SHIMIZU, KEI MIYAMOTO, and MIZUO TANAKA, “Repair of Pars Interarticularis Defect by Segmental Wire Fixation in Young Athletes with Spondylolysis,” *American Journal of Sports Medicine, 31*    
  (2003), 359–364.
* A-14. GEA A. HUIZINGA, WINETTE T. A. van der GRAAF, ANNEMIKE VISSER, JOS S. DIJKSTRA, and JOSETTE E. H.  
     
  M. HOEKSTRA-WEEBERS, “Psychosocial Consequences for Children of a Parent with Cancer,” *Cancer Nursing,*    
  *26* (2003), 195–202.
* A-15. ERIC W. TAM, ARTHUR F. MAK, WAI NGA LAM, JOHN H. EVANS, and YORK Y. CHOW, “Pelvic Movement and  
     
  Interface Pressure Distribution During Manual Wheelchair Propulsion,” *Archives of Physical Medicine and Reha-*    
  *bilitation, 84* (2003), 1466–1472.
* A-16. MARK P. JOHNSON, LESLIE N. SUTTON, NATALIE RINTOUL, TIMOTHY M. CROMBLEHOLME, ALAN W. FLAKE,  
     
  LORI J. HOWELL, HOLLY L. HEDRICK, R. DOUGLAS WILSON, and N. SCOTT ADZICK, “Fetal Myelomeningocele Repair: Short-Term Clinical Outcomes,” *American Journal of Obstetrics and Gynecology, 189* (2003), 482–487.

A-17.

A-18. A-19. A-20.

**DESCRIPTIVE STATISTICS**

D. M. Z. KRIESER, A. R. ROSENBERG, G. KAINER, and D. NAIDOO, “The Relationship between Serum Crea- tinine, Serum Cystatin C, and Glomerular Filtration Rate in Pediatric Renal Transplant Recipients: A Pilot Study,” *Pediatric Transplantation, 6* (2002), 392–395.

JORDAN G. SPIVACK, STEPHEN C. EPPES, and JOEL D. KLIEN, “*Clostridium Difficile*–Associated Diarrhea in a Pediatric Hospital,” *Clinical Pediatrics, 42* (2003), 347–352.

N. THILOTHAMMAL, P. V. KRISHNAMURTHY, DESMOND K. RUNYAN, and K. BANU, “Does BCG Vaccine Pre- vent Tuberculous Meningitis?” *Archives of Disease in Childhood, 74* (1996), 144–147.

North Carolina State Center for Health Statistics and Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill. Birth data set for 2001 found at www.irss.unc.edu/ncvital/ bfd1down.html. All calculations were performed by John Holcomb and do not represent the findings of the Center or Institute.