1

Introduction

This text is primarily concerned with the interrelationships between multiple variables that are collected on study subjects. For example, we may be interested in how age, blood pressure, serum cholesterol, body mass index, and gender affect a patient's risk of coronary heart disease. The methods that we shall discuss involve descriptive and inferential statistics. In descriptive statistics, our goal is to understand and summarize the data that we have actually collected. This can be a major task in a large database with many variables. In inferential statistics, we seek to draw conclusions about patients in the population at large from the information collected on the specific patients in our database. This requires first choosing an appropriate model that can explain the variation in our collected data and then using this model to estimate the accuracy of our results. The purpose of this chapter is to review some elementary statistical concepts that we shall need in subsequent chapters. Although this text is self-contained, I recommend that readers who have not had an introductory course in biostatistics start off by reading one of the many excellent texts on biostatistics that are available (see, for example, Katz, 2006).

1.1. Algebraic notation

This text assumes that the reader is familiar with high school algebra. In this section we review notation that may be unfamiliar to some readers.

- We use parentheses to indicate the order of multiplication and addition; brackets are used to indicate the arguments of functions. Thus, $a(b + c)$ equals the product of a and $b + c$, while $a[b + c]$ equals the value of the function a evaluated at $b + c$.
- The function $\log[x]$ denotes the natural logarithm of x. You may have seen this function referred to as either $\ln[x]$ or $\log_e[x]$ elsewhere.
- The constant $e = 2.718\ldots$ is the base of the natural logarithm.
- The function $\exp[x] = e^x$ is the constant e raised to the power x.

1. Introduction

- The function

$$\text{sign}[x] = \begin{cases} 1 : \text{if } x > 0 \\ 0 : \text{if } x = 0 \\ -1 : \text{if } x < 0. \end{cases} \tag{1.1}$$

- The absolute value of x is written $|x|$ and equals

$$|x| = x\ \text{sign}[x] = \begin{cases} x : \text{if } x \geq 0 \\ -x : \text{if } x < 0. \end{cases}$$

- The expression $\int_a^b f[x]dx$ denotes the area under the curve $f[x]$ between a and b. That is, it is the region bounded by the function $f[x]$ and the x-axis and by vertical lines drawn between $f[x]$ and the x-axis at $x = a$ and $x = b$. For example, if $f[x]$ denotes the curve drawn in Figure 1.11 then the area of the shaded region in this figure is $\int_a^b f[x]dx$. With the exception of the occasional use of this notation, no calculus is used in this text.

Suppose that we have measured the weights of three patients. Let $x_1 = 70$, $x_2 = 60$, and $x_3 = 80$ denote the weight of the first, second, and third patient, respectively.

- We use the Greek letter Σ to denote summation. For example,

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 70 + 60 + 80 = 210.$$

  When the summation index is unambiguous we will drop the subscript and superscript on the summation sign. Thus, $\sum x_i$ also equals $x_1 + x_2 + x_3$.
- We use the Greek letter Π to denote multiplication. For example,

$$\prod_{i=1}^{3} x_i = \prod x_i = x_1\, x_2\, x_3 = 70 \times 60 \times 80 = 336\,000.$$

- We use braces to denote sets of values; $\{i :\ x_i > 65\}$ is the set of integers for which the inequality to the right of the colon is true. Since $x_i > 65$ for the first and third patient, $\{i :\ x_i > 65\} = \{1,\ 3\}$ = the integers one and three. The summation

$$\sum_{\{i:\ x_i > 65\}} x_i = x_1 + x_3 = 70 + 80 = 150.$$

  The product

$$\prod_{\{i:\ x_i > 65\}} x_i = 70 \times 80 = 5600.$$

1.2. Descriptive statistics

1.2.1. Dot plot

Suppose that we have a sample of n observations of some variable. A **dot plot** is a graph in which each observation is represented by a dot on the y-axis. Dot plots are often subdivided by some grouping variable to permit a comparison of the observations between the two groups. For example, Bernard et al. (1997) performed a randomized clinical trial to assess the effect of intravenous ibuprofen on mortality in patients with sepsis. People with sepsis have severe systemic bacterial infections that may be due to a wide number of causes. Sepsis is a life-threatening condition. However, the mortal risk varies considerably from patient to patient. One measure of a patient's mortal risk is the Acute Physiology and Chronic Health Evaluation (APACHE) score (Bernard et al., 1997). This score is a composite measure of the patient's degree of morbidity that was collected just before recruitment into the study. Since this score is highly correlated with survival, it was important that the treatment and control groups be comparable with respect to baseline APACHE score. Figure 1.1 shows a dot plot of the baseline APACHE scores for study subjects subdivided by treatment group. This plot indicates that the treatment and placebo groups are comparable with respect to baseline APACHE score.

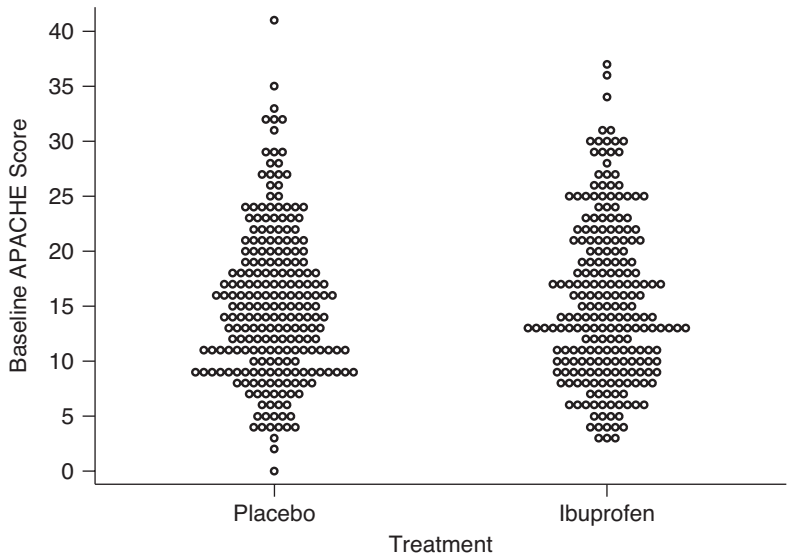

Figure 1.1 Dot plot of baseline APACHE score subdivided by treatment (Bernard et al., 1997).

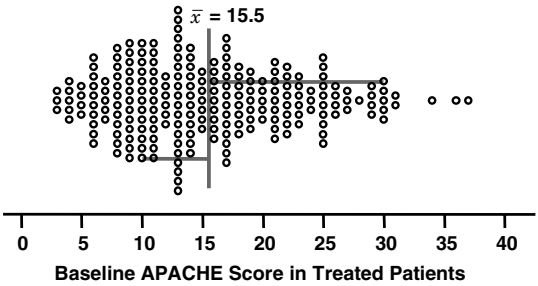

Figure 1.2 Dot plot for treated patients in the Ibuprofen in Sepsis Study. The vertical line marks the sample mean, while the length of the horizontal lines indicates the residuals for patients with APACHE scores of 10 and 30.

1.2.2. Sample mean

The **sample mean** $\bar{x}$ for a variable is its average value for all patients in the sample. Let x_i denote the value of a variable for the i^{th} study subject $(i = 1, 2, \ldots, n)$. Then the sample mean is

$$\bar{x} = \sum_{i=1}^{n} x_i / n = (x_1 + x_2 + \cdots + x_n)/n, \tag{1.2}$$

where n is the number of patients in the sample. In Figure 1.2 the vertical line marks the mean baseline APACHE score for treated patients. This mean equals 15.5. The mean is a measure of central tendency of the x_is in the sample.

1.2.3. Residual

The **residual** for the i^{th} study subject is the difference $x_i - \bar{x}$. In Figure 1.2 the length of the horizontal lines show the residuals for patients with APACHE scores of 10 and 30. These residuals equal $10 - 15.5 = -5.5$ and $30 - 15.5 = 14.5$, respectively.

1.2.4. Sample variance

We need to be able to measure the variability of values in a sample. If there is little variability, then all of the values will be near the mean and the residuals will be small. If there is great variability, then many of the residuals will be large. An obvious measure of sample variability is the average absolute value of the residuals, $\sum |x_i - \bar{x}|/n$. This statistic is not commonly used because it is difficult to work with mathematically. A more mathematician-friendly

measure of variability is the **sample variance**, which is

$$s^2 = \sum (x_i - \bar{x})^2/(n-1). \tag{1.3}$$

You can think of s^2 as being the average squared residual. (We divide the sum of the squared residuals by $n-1$ rather than n for mathematical reasons that are not worth explaining at this point.) Note that the greater the variability of the sample, the greater the average squared residual and, hence, the greater the sample variance.

1.2.5. Sample standard deviation

The **sample standard deviation** s is the square root of the sample variance. Note that s has the same units as x_i. For patients receiving ibuprofen in Figure 1.1 the variance and standard deviation of the APACHE score are 52.7 and 7.26, respectively.

1.2.6. Percentile and median

Percentiles are most easily defined by an example; the 75th **percentile** is that value that is greater or equal to 75% of the observations in the sample. The **median** is the 50th percentile, which is another measure of central tendency.

1.2.7. Box plot

Dot plots provide all of the information in a sample on a given variable. They are ineffective, however, if the sample is too large and may require more space than is desirable. The mean and standard deviation give a terse description of the central tendency and variability of the sample, but omit details of the data structure that may be important. A useful way of summarizing the data that provides a sense of the data structure is the **box plot** (also called the **box-and-whiskers** plot). Figure 1.3 shows such plots for the APACHE data in each treatment group. In each plot, the sides of the box mark the 25th

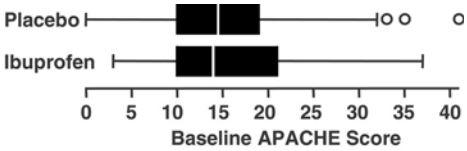

Figure 1.3 Box plots of APACHE scores of patients receiving placebo and ibuprofen in the Ibuprofen in Sepsis Study.

and 75th percentiles, which are also called the **quartiles**. The vertical line in the middle of the box marks the median. The width of the box is called the **interquartile range**. The middle 50% of the observations lie within this range. Whiskers extend on either side of the box. The vertical bars at the end of each whisker mark the most extreme observations that are not more than 1.5 times the interquartile range from their adjacent quartiles. Any values beyond these bars are plotted separately as in the dot plot. They are called **outliers** and merit special consideration because they may have undue influence on some of our analyses. Figure 1.3 captures much of the information in Figure 1.1 in less space.

For both treated and control patients the largest APACHE scores are farther from the median than are the smallest scores. For treated subjects the upper quartile is farther from the median than is the lower quartile. Data sets in which the observations are more stretched out on one side of the median than the other are called **skewed**. They are **skewed to the right** if values above the median are more dispersed than are values below. They are **skewed to the left** when the converse is true. Box plots are particularly valuable when we wish to compare the distributions of a variable in different groups of patients, as in Figure 1.3. Although the median APACHE values are very similar in treated and control patients, the treated patients have a slightly more skewed distribution. (It should be noted that some authors use slightly different definitions for the outer bars of a box plot. The definition given here is that of Cleveland, 1993.)

1.2.8. Histogram

This is a graphic method of displaying the distribution of a variable. The range of observations is divided into equal intervals; a bar is drawn above each interval that indicates the proportion of the data in the interval. Figure 1.4 shows a histogram of APACHE scores in control patients. This graph also shows that the data are skewed to the right.

1.2.9. Scatter plot

It is often useful to understand the relationship between two variables that are measured on a group of patients. A **scatter plot** displays these values as points in a two-dimensional graph: the x-axis shows the values of one variable and the y-axis shows the other. For example, Brent et al. (1999) measured baseline plasma glycolate and arterial pH on 18 patients admitted for ethylene glycol poisoning. A scatter plot of plasma glycolate versus arterial pH for these patients is plotted in Figure 1.5. Each circle on this graph shows

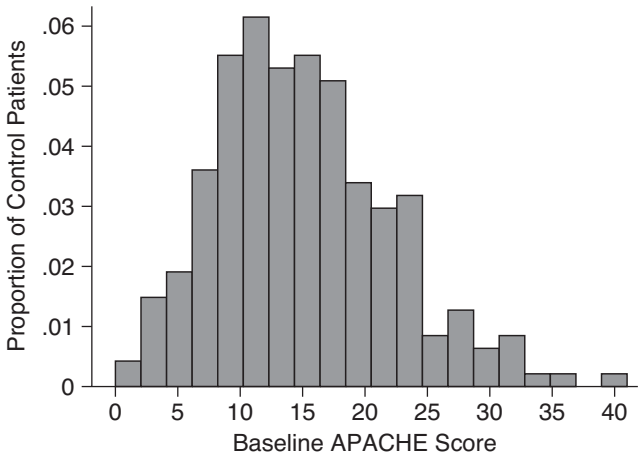

Figure 1.4 Histogram of APACHE scores among control patients in the Ibuprofen in Sepsis Study.

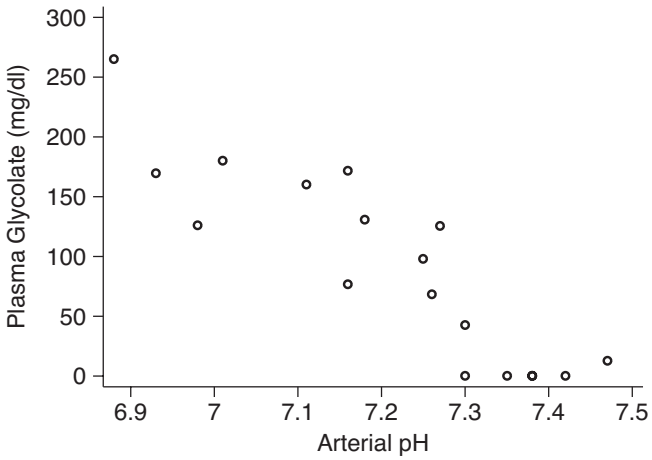

Figure 1.5 Scatter plot of baseline plasma glycolate vs. arterial pH in 18 patients with ethylene glycol poisoning (Brent et al., 1999).

the plasma glycolate and arterial pH for a study subject. Note that patients with high glycolate levels tended to have low pHs, and that glycolate levels tended to decline with increasing pH.

1.3. The Stata Statistical Software Package

The worked examples in this text are performed using Stata version 10 (StataCorp, 2007). Excellent documentation is available for this software. At a minimum, I suggest you read their *Getting Started* manual. This text is not intended to replicate the Stata documentation, although it does explain

the commands that it uses. Appendix B provides a list of these commands and the page number where each command is first explained.

To follow the examples in this book you will need to purchase a license for *Stata/IC 10* for your computer and install it following the directions in the *Getting Started* manual. When you launch the Stata program you will see a screen with four windows. These are the Command window where you will type your commands, the Results window where output is written, the Review window where previous commands are stored, and the Variables window where variable names will be listed. A Stata command is executed when you press the Enter key at the end of a line in the command window. Each command is echoed back in the Results window followed by the resulting output or error message. Graphic output appears in a separate Graph window. In the examples given in this text, I have adopted the following conventions. All Stata commands and output are written in monospaced fonts. Commands that are entered by the user are written in bold face; variable names, labels and other text chosen by the user are italicized, while command names and options that must be entered as is are printed in an upright font. Output from Stata is written in a smaller point size. Highlighted output is discussed in the comments following each example. Numbers in boxes on the right margin refer to comments that are given at the end of each example.

1.3.1. Downloading data from my website

An important feature of this text is the use of real data sets to illustrate methods in biostatistics. These data sets are located at `biostat.mc.vanderbilt.edu/dupontwd/wddtext/`. In the examples, I assume that you are using a Microsoft Windows computer and have downloaded the data into a folder on your C drive called *WDDtext*. I suggest that you create such a folder now. (Of course the location and name of the folder is up to you but if you use a different name you will have to modify the file address in my examples. If you are using Stata on a Macintosh, Linux, or Unix computer you will need to use the appropriate disk and folder naming conventions for these computers.) Next, use your web browser to go to `biostat.mc.vanderbilt.edu/dupontwd/wddtext/` and click on the blue underlined text that says Data Sets. A page of data sets will appear. Click on 1.3.2.Sepsis. A dialog box will ask where you wish to download the sepsis data set. Enter *C:/WDDtext* and click the download button. A Stata data set called *1.3.2.Sepsis.dta* will be copied to your *WDDtext* folder. You are now ready to run the example in the next section.

1.3.2. Creating histograms with Stata

The following example shows the contents of the Results window after entering a series of commands in the Command window. Before replicating this example on your computer, you must first download *1.3.2.Sepsis.dta* as described in the preceding section.

```
. *  Examine the Stata data set 1.3.2.Sepsis.dta.  Create                    1
. *  histograms of baseline APACHE scores in treated and
. *  control patients.
. use  C:\WDDtext\1.3.2.Sepsis.dta                                           2

. describe                                                                   3

Contains data from c:\WDDtext\1.3.2.Sepsis.dta
  obs:           455
 vars:             2                          16 Apr 2002 15:36
 size:         5,460 (99.5% of memory free)

------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
treat           float  %9.0g       treatmnt   Treatment
apache          float  %9.0g                  Baseline APACHE Score
------------------------------------------------------------------------------

Sorted by:

. list  treat apache in 1/3                                                  4

     +-------------------+
     |    treat   apache |
     |-------------------|
  1. |  Placebo       27 |
  2. | Ibuprofen      14 |
  3. |  Placebo       33 |
     +-------------------+                                                   5

. browse                                                                     6

. histogram apache, by(treat) bin(20) percent                                7
```

Comments

1 Command lines that start with an asterisk (*) are treated as comments and are ignored by Stata.

2 The *use* command specifies the name of a Stata data set that is to be used in subsequent Stata commands. This data set is loaded into memory where it may be analyzed or modified. Data sets may also be opened by clicking on the folder icon on the Stata toolbar. In Section 4.21 we will illustrate how to create a new data set using Stata.

3 The *describe* command provides some basic information about the current data set. The *1.3.2.Sepsis* data set contains 455 observations. There are two variables called *treat* and *apache*. The labels assigned to these variables are *Treatment* and *Baseline APACHE Score*.

4 The *list* command gives the values of the specified variables; *in 1/3* restricts this listing to the first three observations in the file.

5 At this point the Review, Variables, Results, and Command windows should look like those in Figure 1.6. (The size of these windows has been changed to fit in this figure.) Note that if you click on any command in the Review window it will appear in the Command window where you can edit and re-execute it. This is particularly useful for fixing command errors. When entering variables in a command you may either type them directly or click on the desired variable from the Variables window. The latter method avoids spelling mistakes.

6 Typing *browse* opens the Stata Data Browser window (there is a button on the toolbar that does this as well). This command permits you to review but not modify the current data set. Figure 1.7 shows this window, which presents the data in a spreadsheet format with one row per patient and one column per variable. (Stata also has an *edit* command that allows you to both view and change the current data. I recommend that you use the *browse* command when you only wish to view your data in order to avoid accidental changes.)

I will refer to rows and columns of the Stata Data Browser as *observations* and *variables*, respectively. An observation consists of all of the variable values in a single row, which will usually consist of all variable values collected on a single patient. A row of variable values is also called a *record*.

7 This command produces Figure 1.8. This figure appears in its own Graph window. The *by(treat)* option causes separate histograms for each treatment to be drawn side by side in the same graph; *bin(20)* groups the data into 20 bins with a separate bar for each bin in each panel. The *percent* option causes the *y*-axis to be the percentage of patients on each