Chapter 1

one

Introduction

1.1 What is statistics

Statistics is a branch of science that deals with the art of collecting, classifying, displaying, analyzing and interpreting results from research. This definition implies that the statistician is more concerned with the scientific observations (i.e. data) than with the actual biological material involved in the study. In fact, many of the statistical techniques described in this book had their beginnings in "nonbiological" settings, such as agriculture, business and engineering. Statistical procedures which have been developed for use in one field of science have almost invariably found applications in a number of other fields. There are, however, statistical procedures which are more frequently used in the biomedical field. The application of statistical methods to the biological and life sciences is typically called biostatistics (or, sometimes, biometry). This book will concentrate on statistical techniques that are most widely used.

The definition given to statistics above conveniently breaks into 2 subcategories: descriptive statistics and inferential statistics.

1.1.1 Descriptive statistics

Descriptive statistics is the part of the subject that most layman think of when they hear the word "statistics". This area involves the collection, presentation and description of numerical data through the use of graphical, tabular or numerical devices. Many of the concepts involved in the area of descriptive statistics are part of people's every day observations. For example, the concept of average is used in referring to a person's height, to the weather forecast, or to a ballplayer's performance. Dispersion, or variability, is another familiar concept. While the average gives us a notion of a typical or usual value, dispersion tells us that individual observations differ from person to person or day to day. When a teacher says that the better a student reads, the better he (or she) will perform at arithmetic, then the statistical concept of correlation, or association, is being referenced. The manner in which these concepts of average, dispersion

and correlation are implemented within the analysis of observations from biomedical experiments will be developed more fully in the ensuing chapters.

1.1.2 Inferential statistics

The area of inferential statistics is concerned with the generalities that can be inferred from specific observations. For example, in testing the safety and efficacy of a new drug, an experimenter will typically have only a limited number of patients available. The researcher then is interested in the extent to which generalizations to all patients can be made from the experimental results. Thus, inferential statistics attempts to reach conclusions concerning a large group (or population) on the basis of studying only a small subset (or sample).

Notice that the conclusions reached in the area of inferential statistics will depend strongly upon the selection of appropriate descriptive statistics. Indeed, the failure to choose appropriate descriptive statistics has often been responsible for faulty scientific conclusions.

1.2 Study design and data collection

Much of the material in this book relates to methods that are used in the analysis of data. Obtaining relevant data requires a carefully drawn plan that identifies the population of interest, the procedure used to randomly select units for study, and the process used in the observation/measurement of the attributes of interest. Two standard methods of data collection are sample surveys and experiments (that may involve sampling). Sample survey design deals with ways to select a random sample that is representative of the population of interest and from which a valid inference can be made. Unfortunately, it is very easy to select nonrepresentative samples that lead to misleading conclusions about the population, emphasizing the need for the careful design of sample surveys. Experimental design involves the creation of a plan for determining whether or not there are differences between groups. The design attempts to control extraneous factors so that the only reason for any observed differences between groups is the factor under study. Since it is very difficult to take all extraneous factors into account in a design, we also use the random allocation of subjects to the groups. We hope that through the use of the random assignment, we can control for factors that have not been included in the design itself. Experimental design is also concerned with determining the appropriate sample size for the study. Sometimes we also analyze data that were already collected. In this case, we need to understand how the data were collected in order to determine the appropriate of analysis.

1.3 The purpose of this textbook

The preceding section has discussed the importance of proper planning prior to experimentation. It will be assumed that proper design and sampling techniques have been followed for any given set of observations presented in the examples and exercises. This text will discuss and explain appropriate statistical procedures, along with limitations and assumptions, to be used in varying biological situations. It should not, however, be assumed that the analysis provided in an example is the only appropriate analysis to be considered. It should be remembered that statistics is both an art and a science and that there are no formulas whereby one can decide whether the data are amenable to certain statistical techniques. Only time, study, thought and experience can provide this. In fact, it is even incorrect to assume that 2 experienced statisticians would come to the same conclusions concerning adequacy or suitability of a specific statistical analysis. This is not to imply that there is confusion over the techniques involved, but rather that there is more than one way to appropriately examine a situation statistically. Thus, it is hoped that this textbook will provide an understanding of the proper application of statistical methods in scientific research. The application of statistical methods requires more than the ability to use statistical software. In this text, we give priority to understanding the context for the use of statistical procedures. This context includes the study's goal, the data, and how the data are collected and measured. We do not focus on the derivation of formulas. Instead, we present the rationale for the different statistical procedures and when and why they use. We would like the reader to think instead of simply memorizing formulas.

1.4 Introduction to some basic terms and notation

To obtain an understanding of statistics, it is necessary to be able to speak its language. The following definitions are intended to give a general understanding and are not necessarily mathematically complete.

A **population** is a collection, or set, of individuals, objects, or measurements about whom we wish to make inferences. Notice that this definition is not limited only to a collection of people. We are interested in making inferences or generalizations about the population. Thus, the population of concern must be carefully and well defined. The set of "all diabetic males over age 50" is all example of a well-defined population. A **sample** is a subset of the population. The result of the sample will be used to infer generalizations of the population. A **variable** is a characteristic of interest about each individual element of population or sample. Example of variables might be a diabetics

weight, age, or blood pressure. Frequently variables are represented by a letter of the alphabet, usually x. A **data set** (or set of observations) is the set of values collected for the variables of interest from each of the elements belonging to the sample. We will use the symbol n to represent the number of observations collected in a sample. To distinguish between the different observations that make up the sample we will use subscripts. Thus, if we were interested in the weight of diabetics, X_1 would be the weight of the first diabetic, X_2 the weight of the second diabetic, . . . , X_n the weight of the nth (or last) diabetic. A **parameter** is a summary descriptive characteristic of a population of observations. The average weight of all diabetics over age 50 is an illustration of a parameter. A parameter is typically symbolized by a Greek letter. The most commonly used parameters in the ensuing chapters are given below.

μ ("mu"): The mean (or average) of a population.

σ^2 ("sigma square"): The variance of a population.

σ ("sigma"): The standard deviation of a population.

θ ("theta"): Binomial proportion.

ρ ("rho"): Population correlation coefficient.

A **statistic** is a summary descriptive characteristic of a sample of observations. The average value of the data set would be an example of a statistic. A sample statistic is typically used to estimate (or infer information about) a corresponding population parameter. Most of the sample statistics that we will study will be assigned symbolic names that are letters of the English alphabet as follows:

$\bar{x}$ ("x bar") : The mean (or average) of a sample.

s^2 ("s square"): The variance of a sample.

s: The standard deviation of a sample.

p: Sample proportion.

r: Sample correlation coefficient.

(Lu Wenli and Wang Peishan)

Chapter 2

two

Descriptive Statistics

In this chapter, we will introduce statistical method to describe the data at hand in some concise way. In smaller studies this step can be accomplished by listing each data point. In general, however, this procedure is tedious or impossible and, even if it were possible, would not give an overall picture of what the data look like.

2.1 Types of data

Before discussing how data can be summarised and displayed, it is first necessary to distinguish between different types of data.

2.1.1 Nominal data

Nominal data are data that one can name. They are not measured but simply counted. They often consist of unordered "either-or" type observations, for example: Dead or Alive; Male or Female; Cured or Not Cured; Pregnant or Not Pregnant. However, they often can have more than two categories, for example: blood group O, A, B, AB, country of origin, racial group or social class. The methods of presentation of nominal data are limited in scope.

2.1.2 Ordered categorical or ranked data

If there are more than two categories of classification it may be possible to order them in some way. For example, after treatment a patient may be either improved, the same or worse; a woman may never have conceived, conceived but spontaneously aborted, or given birth to a live infant. In some studies it may be appropriate to assign ranks. For example, patients with rheumatoid arthritis may be asked to order their preference for four dressing aids. Here although numerical values are assigned to each aid one cannot treat them as numerical values. They are in fact only codes for best, second best, third choice and worst.

2.1.3 Numerical discrete data

Such data consist of counts. The number of previous babies born to the 1000 women who had just given birth is an example of numerical discrete data. Another example might be the number of deaths in a hospital per year.

2.1.4 Numerical continuous data

Such data are measurements that can, in theory at least, take any value within a given range, for example, temperature in degrees Celsius or height in centimeters.

For simplicity it is often the case in medicine that continuous data are dichotomised to make nominal data. Thus diastolic blood pressure, which is continuous, is converted into hypertension (>90mmHg) and normotension ($\leqslant 90$mmHg). This clearly leads to a loss of information, but often makes the data easier to summarize. Another example, age can be measured in years (numerical continuous), placed into young, middle-aged, and elderly age groups (ordinal), or classified as economically productive (ages 16 to 64) and dependent (under 16 and over 64) age groups (nominal). It is possible to convert a higher-level scale (numerical discrete or continuous) into a lower-level scale (ordinal and nominal) but not to convert from a lower level to a higher level.

2.2 Measures of central tendency

One type of measure useful for summarizing data defines the center, or middle, of the sample. This type of measure is a measure of location.

2.2.1 The arithmetic mean

The simplest and probably most familiar measure of central tendency is the average of observations in the data set. However, due to the varying interpretations that are given to the term "average", statisticians have uniformly decided to say "arithmetic mean" or more simply, just "mean" or "sample mean". The arithmetic mean is the sum of all the observations divided by the number of observations. It is written in statistical terms as

$$\bar{x} = \frac{\sum x}{n}$$

The symbol $\sum$ means summation. If we have the data for the entire population,

not for just a sample of observations from the population, the mean is denoted by the Greek letter μ (pronounced "mu").

Example 2.1 A group of 16 subjects are measured for concentration of prothrombin in plasma. The data, where prothrombin is expressed in milligrams/100ml(mg/100ml), is shown as follows

22	17	23	18	25	17	23	17
15	18	20	17	23	21	22	24

What is the mean concentration of prothrombin?

$$\bar{x} = \frac{22 + 17 + 23 + \cdots + 21 + 22 + 24}{16} = \frac{322}{16} = 20.1\,\text{mg/100ml}$$

2.2.2 The median

Another frequently used measure of central tendency is the median, defined as the middle value of a data set which has been ranked in order according to size. The median is calculated as follows

1. Rank the observations in order of size (smallest to largest).

2. (a) If n is odd, then the median is determined as the $[(n+1)/2]^{\text{th}}$ largest observation.

(b) If n is even, then the median is determined as the value halfway between the $(n/2)^{\text{th}}$ and the $[(n/2)+1]^{\text{th}}$ observations.

For example, the median for a sample of size 33 is thus the 17^{th} largest value. The value 17 comes from (33 + 1)/2. When sample size is even, as in the case of the data on systolic blood pressure readings presented in Example 2.2, there is no observed sample value such as one-half of the sample falls below it and one-half falls above it.

By convention, we use the average of the two middle sample values as the median—that is, the average of the $(16/2)^{\text{th}}$ and $[(16/2)+1]^{\text{th}}$ largest values.

Example 2.2 A group of 16 subjects are measured for systolic blood pressure(mmHg), is shown as follows

128	134	152	120	144	130	132	122
128	142	138	136	118	126	116	126

Step 1. The data are sorted in ascending order:

116	118	120	122	126	126	128	128
130	132	134	136	138	142	144	152

Step 2. Since $n = 16$ is even, we find that

$$\frac{n}{2} = \frac{16}{2} = 8$$

Steps. Median = Value halfway between the 8^{th} and 9^{th} position

$$\text{Median} = \frac{128 + 130}{2} = 129\text{mmHg}$$

2.2.3 The geometric mean

Another measure that is used in these situations is the geometric mean. The sample geometric mean for n observations is the n^{th} root of the product of the values — that is,

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$$

Note that since the nth root is used in its calculation, the geometric mean cannot be used when a value is negative or zero.

We use the geometric mean to measure central tendency when the numbers reflect population counts that are extremely variable. For example, in a laboratory setting, the growth in the number of bacteria per area is examined over time. The number of microbes per area does not change by the same amount from one period to the next, but the change is proportional to the number of microbes that were present during the previous time period. Another way of saying this is that the growth is multiplicative, not additive.

There is another way we can find the geometric mean. We can transform the observations to a logarithmic scale. Use of the logarithmic scale provides for accurate calculation of the geometric mean. After finding the logarithm of the geometric mean, we will transform the value back to the original scale and have the value of the geometric mean. In this section we shall use logarithms to the base 10, although other bases could be used. The logarithm of the product of n values is

$$\lg(x_1 \cdot x_2 \cdot \cdots \cdot x_n) = \sum_{i=1}^{n} \lg x_i$$

In addition, taking the n^{th} root of a product on the arithmetic scale becomes division by n on the logarithmic scale — that is, finding the mean logarithm. In symbols, this is

$$\sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \text{antilg} \frac{\sum_{i=1}^{n} \lg x_i}{n}$$

We now have the logarithm of the geometric mean, and, to obtain the geometric mean, we must take the antilogarithm of the mean logarithm — that is,

$$\bar{x}_g = \text{antilg} \frac{\sum_{i=1}^{n} \lg x_i}{n}$$

Example 2.3 Suppose that the number of microbes observed from six different areas are the following: 100, 100, 1000, 1000, 10,000, and 1,000,000. What is the geometric mean?

$$\begin{aligned}\bar{x}_g &= \text{antilg}\frac{\sum_{i=1}^{n}\lg x_i}{n}\\ &= \text{antilg}\frac{\lg 100+\lg 100+\lg 1000+\lg 1000+\lg 10,000+\lg 1,000,000}{6}\\ &= \text{antilg}\frac{2+2+3+3+4+6}{6}\\ &= \text{antilg}\,3.33\\ &= 2154.43\end{aligned}$$

The arithmetic mean of these observations is 168,700, a much larger value than the geometric mean and also much larger than five of the six values. The usual mean does not provide a good measure of central tendency in this case.

2.2.4 The mode

The mode is the most frequently occurring value. When all the values occur the same number of times, we usually say that there is no unique mode. When two values occur the same number of times and more than any other values, the distribution is said to be bimodal. If there are three values that occur the same number of times and more than any other value, the distribution could be called trimodal. Usually one would not go beyond trimodal in labeling a distribution. It is not unexpected to have no unique mode when dealing with continuous data, since it is unlikely that two units have exactly the same values of a continuous variable. However, in our data set of prothrombin present in Example 2.1, the value of 17 occurs four times, more frequently than any other reading, and is thus the mode. Although prothrombin is a continuous variable, the measurer often has a preference for values ending in zero, resulting in multiple observations of some values.

2.2.5 Use of the measures of central tendency

Now that we understand how these three measures of central tendency are defined and found, when are they used? Note that in calculating the mean, we summed the observations. Hence, we can only calculate a mean when we can perform arithmetic operations on the data. We cannot perform meaningful arithmetic operations on nominal data. Therefore, the mean should only be used when we are working with continuous data, although sometimes we find it being used with ordinal data as well. The median does not require us to sum observations, and thus it can be used with continuous and

ordinal data, but it also cannot be used with nominal data. The mode can be used with all types of data because it simply says which level of the variable occurs most frequently.

The mean is affected by extreme values, whereas the median is not. Hence, if we are studying a variable such as income that has some extremely large values, that is positively skewed, the mean will reflect these large values and move away from the center of the data. The median is unaffected, and it remains at the center of the data. For data that are symmetrically distributed or approximately so, the mean and median will be the same or very close to each other. As was just mentioned, the SBP readings ranged from 116 to 152mmHg for the 16 observations. The sample mean was 130.75mmHg, and the sample median was 129mmHg. These two values do not differ very much, since the data set contains observations that are relatively extreme on both the low and high end. However, the two values 2mmHg have caused the mean of 130. 75mmHg to be slightly larger than the median of 129mmHg.

The geometric mean has also been used in the estimation of population counts — for example, of mosquitos — through the use of capture procedures over several time points or areas. These counts can be quite variable by time or area, and hence, the geometric mean is the preferred measure of central tendency in this situation.

2.3 Measures of dispersion

It is not sufficient to characterized a data set solely by an indication of its central location. To illustrate, let's consider the following 2 cases.

Data Set 1: 4, 5, 5, 5, 5, 6

Data Set 2: 1, 2, 5, 5, 8, 9

These data sets have the same mean, median and mode (all equal to 5), however, the second set is more variable than the first. We will now consider several frequently used sample measures of variability, or dispersion.

2.3.1 The range

The simplest measure of dispersion is the range. It is defined as the difference between the largest and smallest observation.

Example 2.4 The ranges for the above 2 data sets are

(1) Range = 6 − 4 = 2

(2) Range = 9 − 1 = 8

Although the range has the advantage of being easy to compute and to understand, it has the disadvantage of being sensitive to extreme values and of ignoring the sample size. We expect large samples to include occasional extreme values and hence,