

1

.....

Basic Concepts

1.1. Introduction

A brief glance through almost any recently published medical journal will show that statistical methods are playing an increasingly visible role in modern medical research. At the very least, most research papers quote (at least) one 'p-value' to underscore the 'significance' of the results which the authors wish to communicate. At the same time, a growing number of papers are now presenting the results of relatively sophisticated, 'multi-factor' statistical analyses of complex sets of medical data. This proliferation in the use of statistical methods has also been paralleled by the increased involvement of professionally trained statisticians in medical research as consultants to and collaborators with the medical researchers themselves.

The primary purpose of this book is to provide medical researchers with sufficient understanding to enable them to read, intelligently, statistical methods and discussion appearing in medical journals. At the same time, we have tried to provide the means for researchers to undertake the simpler analyses on their own, if this is their wish. And by presenting statistics from this perspective, we hope to extend and improve the common base of understanding which is necessary whenever medical researchers and statisticians interact.

It seems obvious to us that statisticians involved in medical research need to have some understanding of the related medical knowledge. We also believe that in order to benefit from statistical advice, medical researchers require some understanding of the subject of statistics. This first chapter provides a brief introduction to some of the terms and symbols which recur throughout the book. It also establishes what statisticians talk about (random variables,

probability distributions) and how they talk about these concepts (standard notation). We are very aware that this material is difficult to motivate; it seems so distant from the core and purpose of medical statistics. Nevertheless, ‘these dry bones’ provide a skeleton which allows the rest of the book to be more precise about statistics and medical research than would otherwise be possible. Therefore, we urge the reader to forbear with these beginnings, and read beyond the end of chapter 1 to see whether we do not put flesh onto these dry bones.

1.2. Random Variables, Probability Distributions and Some Standard Notation

Most statistical work is based on the concept of a random variable. This is a quantity that, theoretically, may assume a wide variety of actual values, although in any particular realization we only observe a single value. Measurements are common examples of random variables; take the weights of individuals belonging to a well-defined group of patients, for example. Regardless of the characteristic that determines membership in the group, the actual weight of each individual patient is almost certain to differ from that of other group members. Thus, a statistician might refer to the random variable representing the weight of individual patients in the group, or population of interest. Another example of a random variable might be a person’s systolic blood pressure; the variation in this measurement from individual to individual is frequently quite substantial.

To represent a particular random variable, statisticians generally use an upper case Roman letter, say X or Y . The particular value which this random variable represents in a specific case is often denoted by the corresponding lower case Roman letter, say x or y . The probability distribution (usually shortened to the distribution) of any random variable can be thought of as a specification of all possible numerical values of the random variable, together with an indication of the frequency with which each numerical value occurs in the population of interest.

It is common statistical shorthand to use subscripted letters – x_1, x_2, \dots, x_n , for example – to specify a set of observed values of the random variable X . The corresponding notation for the set of random variables is $X_i, i = 1, 2, \dots, n$, where X_i indicates that the random variable of interest is labelled X and the symbols $i = 1, 2, \dots, n$ specify the possible values of the subscripts on X . Similarly, using n as the final subscript in the set simply indicates that the size of the set may vary from one instance to another, but in each particular instance it will be a fixed number.

Subscripted letters constitute extremely useful notation for the statistician, who must specify precise formulae which will subsequently be applied in particular situations which vary enormously. At this point it is also convenient to introduce the use of Σ , the upper case Greek letter sigma. In mathematics, Σ represents summation. To specify the sum $X_1 + X_2 + X_3$ we would simply write $\sum_{i=1}^3 X_i$. This expression specifies that the subscript i should take the values 1, 2 and 3 in turn, and we should sum the resulting variables. For a fixed but unspecified number of variables, say n , the sum $X_1 + X_2 + \dots + X_n$ would be represented by $\sum_{i=1}^n X_i$.

A set of values x_1, x_2, \dots, x_n is called a sample from the population of all possible occurrences of X . In general, statistical procedures which use such a sample assume that it is a random sample from the population. The random sample assumption is imposed to ensure that the characteristics of the sample reflect those of the entire population, of which the sample is often only a small part.

There are two types of random variables. If we ignore certain technicalities, a discrete random variable is commonly defined as one for which we can write down all its possible values and their corresponding frequencies of occurrence. In contrast, continuous random variables are measured on an interval scale, and the variable can assume any value on the scale. Of course, the instruments which we use to measure experimental quantities (e.g., blood pressure, acid concentration, weight, height, etc.) have a finite resolution, but it is convenient to suppose, in such situations, that this limitation does not prevent us from observing any plausible measurement. Furthermore, the notation which statisticians have adopted to represent all possible values belonging to a given interval is to enclose the end-points of the interval in parentheses. Thus, (a, b) specifies the set of all possible values between a and b , and the symbolic statement ' $a < X < b$ ' means that the random variable X takes a value in the interval specified by (a, b) .

The probability distribution of a random variable is often illustrated by means of a histogram or bar graph. This is a picture which indicates how frequently each value of the random variable occurs, either in a sample or in the corresponding population. If the random variable is discrete, the picture is generally a simple one to draw and to understand. Figure 1.1a shows a histogram for the random variable, S , which represents the sum of the showing faces of two fair dice. Notice that there are exactly 11 possible values for S . In contrast to this situation, the histogram for a continuous random variable, say systolic blood pressure, X , is somewhat more difficult to draw and to understand. One such histogram is presented in figure 1.1b. Since the picture is intended to show both the possible values of X and also the frequency with which they arise, each rectangular block in the graph has an area equal to the propor-

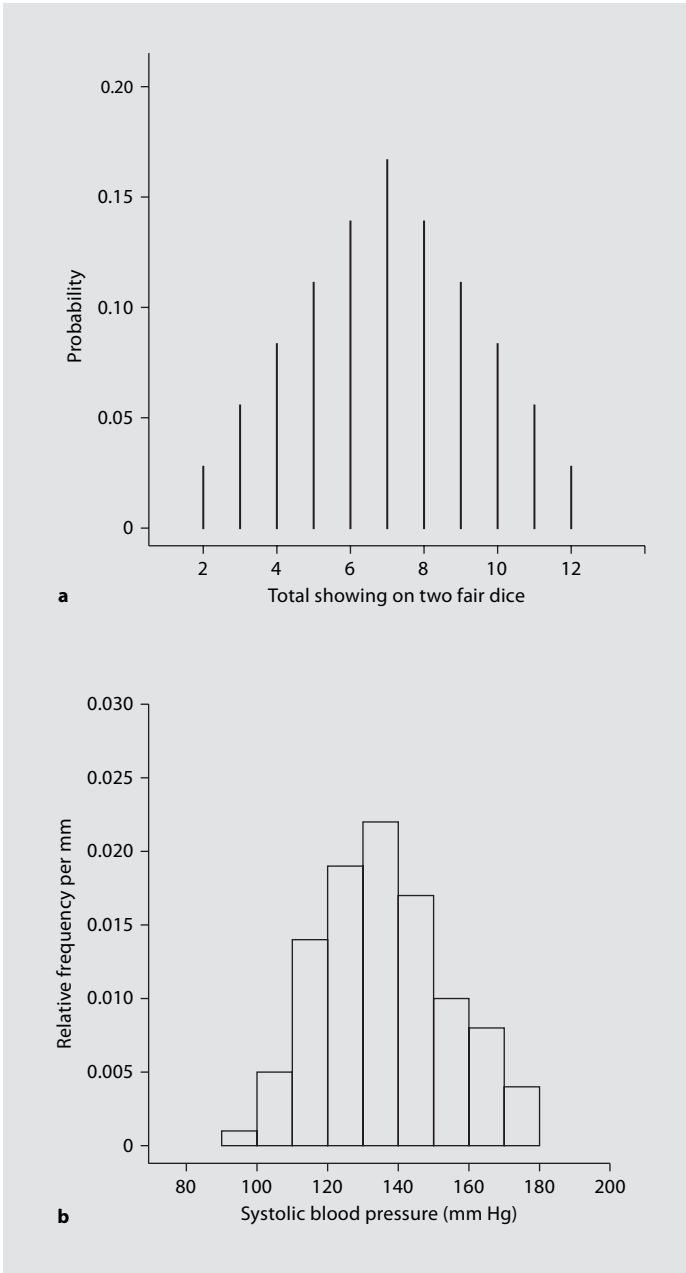


Fig. 1.1. Histograms of random variables. **a** The discrete random variable, S , representing the sum of the showing faces for two fair dice. **b** One hundred observations on the continuous random variable, X , representing systolic blood pressure.

tion of the sample represented by all outcomes belonging to the interval on the base of the block. This has the effect of equating frequency, or probability of occurrence, with area and is known as the ‘area = probability’ equation for continuous random variables.

To a statistician, histograms are simply an approximate picture of the mathematical way of describing the distribution of a continuous random variable. A more accurate representation of the distribution is obtained by using the equation of a curve which can best be thought of as a ‘smooth histogram’; such a curve is called a probability density function. A more convenient term, and one which we intend to use, is probability curve.

Figure 1.2a shows the probability curve, or smooth histogram, for the continuous random variable, X , which we used above to represent systolic blood pressure. This curve is, in fact, the probability curve which has the characteristic shape and equation known as a ‘normal distribution’. Random variables that have a normal distribution will recur in subsequent chapters, and we intend to explain their properties and uses in more detail at that time. For the present, however, we want to concentrate on the concept of the area = probability equation. Figure 1.2b shows two shaded areas. One is the area below the curve and above the interval $(110, 130)$. Recall that the symbol $(110, 130)$ represents all blood pressure measurements between 110 and 130 mm Hg. Because of the area = probability equation for the continuous random variable X , the shaded area above $(110, 130)$ corresponds, pictorially, to the probability that systolic blood pressure in the population is between 110 and 130 mm Hg. This area can be calculated mathematically, and in this particular example the value is 0.323. To represent this calculation in a symbolic statement we would write $\Pr(110 < X < 130) = 0.323$; the equation states that the probability that X , a systolic blood pressure measurement in the population, is between 110 and 130 mm Hg is equal to 0.323.

The second shaded area in figure 1.2b is the area below the probability curve corresponding to values of X in the interval $(165, \infty)$, i.e., the probability that a systolic blood pressure measurement in the population exceeds 165 mm Hg. By means of certain calculations we can determine that, for this specific example, the probability that systolic blood pressure exceeds 165 mm Hg is 0.023; the concise mathematical description of this calculation is simply $\Pr(X > 165) = 0.023$.

Although the probability curve makes it easy to picture the equality of area and probability, it is of little direct use for actually calculating probabilities since areas cannot be read directly from a picture or sketch. Instead, we need a related function called the cumulative probability curve. Figure 1.3 presents the cumulative probability curve for the normal distribution shown in figure 1.2a. The horizontal axis represents the possible values of the random