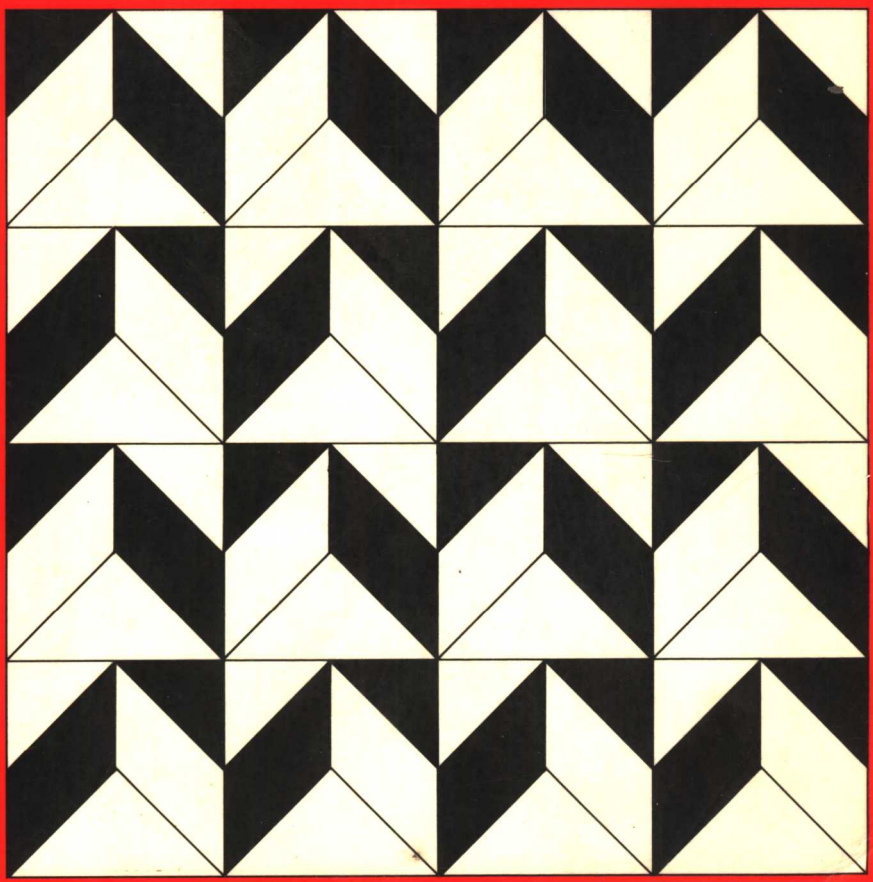
David S. Phillips

Basic Statistics for Health Science Students



Basic Statistics for Health Science Students

David S. Phillips

University of Oregon Medical School

W. H. Freeman and Company
San Francisco

Preface

Individuals in the health sciences who are involved in research, directly or indirectly, have come to realize that it is necessary for them to have some familiarity with statistics. Of course, researchers who are actively collecting data must be able to apply statistics, but those who attempt to keep up with the current scientific literature in their fields need also to understand the uses and misuses of statistical techniques. It is for these individuals that this book has been written. It may be used as a text by the individuals who wish to approach the topic of statistics by themselves but is intended more for use as a course textbook or as a quick reference book for students who have previously taken a course in statistics.

Having served as a statistical consultant to the staff and students of a health sciences center and having taught statistics to dental, medical, nursing, and basic medical science Ph.D. students for several years, I have found a need for a simply written introductory book that covers a wide variety of statistical problems. While there are a number of very good introductory textbooks

available, most of these do not deal with the types of problems confronted by the health scientist. Furthermore, the need is for a nontheoretical, cookbook approach. Most of the physicians and nurses who have come to me for statistical help have no more interest in the theory underlying a statistical test than they do in the theory underlying the building of an X-ray machine. Both the X ray and the statistical test are viewed as tools that may be useful in answering a particular question. Typically I am asked, "What formula should I use with these data to answer this question and how do I interpret the results?" Hence, this book attempts to cover, in cookbook fashion, a wide variety of statistical techniques that I have found being used by health professionals. It is assumed only that the reader has some familiarity with high-school level algebra.

The topic of statistics may be divided into two broad areas—descriptive statistics and inferential statistics. The first half of this book deals with descriptive statistics commonly encountered in the health sciences, and the second half covers some inferential statistics.

The goal of descriptive statistics is just that—to describe data. In 1975 there were 3728 physicians licensed to practice medicine in Oregon. This statement tells us the population under study but really does not describe it. Depending upon our interests we could describe this population in a variety of ways: the number of physicians in each specialty, their average age, the percent who are females, the patient/physician ratio, etc. When we have data that we wish to describe we may do this either pictorially (tables and graphs) or numerically (ratio, percents, rates, correlation). Most data can be described in more than one way so that the choice of method is up to the investigator. In making this choice the investigator should choose the method that will most clearly present the point that he or she wishes to make with a particular audience. This is to say that the same data may be presented in different forms for different audiences. A student who has done a thesis project on congenital heart disease would present his or her data one way in a talk for the PTA but in a different way in a paper given at the American Heart Association meeting and perhaps in a third way for the defense of his or her thesis before a group of professors.

Just as the person who is presenting that data must be careful in selecting the form to be used in the presentation, so must the consumer be careful. The speaker or writer is going to use the best means available to make his or her point, and the listener or reader must consider that if the data were presented and/or organized in another fashion they might lend themselves to another interpretation. We are used to being skeptical when we read the used-car ads

in the newspaper and see cold remedies advertised on TV, but we need that same skepticism when we are confronted with "scientific" data in lectures and journals.

Most chapters have exercises so the student can test his or her understanding of the material. Answers to the exercises are at the end of the book.

I am grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Longman Group Ltd., London, for permission to reprint Tables III, IV, and VII from their book *Statistical Tables for Biological, Agricultural and Medical Research* (6th edition, 1970).

I also wish to express my appreciation to the Biometrika Trustees for permission to reprint Tables 1 and 18 from E. S. Pearson and H. O. Hartley's *Biometrika Tables for Statisticians, Vol. I*, 3rd edition; to the Institute of Educational Research at Indiana University for permission to reproduce the materials in Appendix E; to the Institute for Mathematical Statistics for permission to reprint the material in Appendix F; to Lederle Laboratories, a division of American Cyanamid Company, for permission to reproduce the material found in Appendix G; and to the American Statistical Association for permission to use the material composing Appendix I and Appendix J.

September 1977
Portland, Oregon

David S. Phillips

Contents

Chapter 1
Tables and Graphs

In this chapter we will consider ways to present data in a tabular or graphic fashion. All kinds of data lend themselves to pictorial presentation, and consequently the variety of forms is limited only by one's imagination. Because of the variety of forms, it is impossible to construct a general set of "rules" to be followed in generating tables and graphs. However, authors should keep in mind that the aim is to summarize data; so the material should be presented in a clear and concise manner.

The best way to evaluate a table or a graph is to ask the question, "Can this table or graph stand alone? Is it clear enough to be self-explanatory?" If you have to refer to the text of a journal article to understand a graph, it is not self-sufficient. The best way to make a table self-sufficient is to label it properly. This labeling begins with the title and is carried on throughout the table. The title should contain information about what kind of data are in the table, when the data was collected, where it was collected and who was involved. "Immature births by county of residence, Washington, 1977" tells us exactly

what we will find in the accompanying table. The headings in a table and the axes on a graph should be appropriately labeled. In any case where the units of measurement are not obvious, they should be indicated.

Tables

Tables have the general form indicated in Table 1.1. Each table is given a number so that it can be referenced, and this number and the title appear above the table. Headings and subheadings should be clearly labeled. If the table is being used in a talk or an article to quickly make a particular point, it should be kept brief. A table composed of eight columns of numbers covering an 8×11 page does not lend itself to a quick conclusion. Large tables of this type are appropriate for reference works or thesis appendixes where the reader may want access to detailed information for further study.

Table 1.1. *Title: typical table*

Stub	*Heading* *Subheading*	*Heading* *Subheading*

Graphs

What we have said about tables also applies to graphs: keep them concise, label them appropriately. Graphs are given numbers also and the number of the figure and its title usually appear below the figure. The variable under study is customarily plotted on the horizontal axis of the graph, the **abscissa;** the vertical axis, the **ordinate,** contains enumerative data, such as the number of cases, rate per 10,000, percentage of patients, etc.

Frequency Polygons

Several forms of graphs occur with such frequency that they have been given specific names. Suppose that we are studying a group of pediatric patients with the following ages: 1, 1, 2, 2, 2, 2, 3, 3, 4, 5, 6. We can plot these data

by listing the ages on the abscissa and the number of patients on the ordinate. For each age group we count the number of patients and then put a dot on the graph above that age and corresponding to the number of patients in that group. After we have done this for each age group, we join the points with a line and we extend this line down to the abscissa at each end to keep the graph from hanging in mid-air. This type of graph is called a **frequency polygon.**

Some individuals prefer to use relative frequencies or percentages on the ordinate when plotting data of this nature. As can be seen from comparing the three graphs in Figure 1.1, the overall shape of the figure remains the same regardless of which of these is used on the ordinate. Relative frequencies or percentages may be especially useful when the intent is to compare two or more sets of scores. However, caution should be used when interpreting such data if the numbers of observations are small.

Histograms

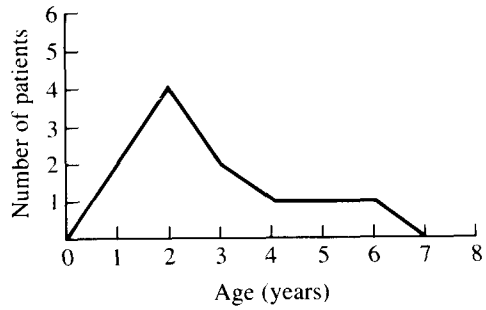
If we use the same graph but erect a column proportional to the number of cases at that age at each point on the abscissa, we will produce Figure 1.2, which is a **histogram.** Frequency polygons and histograms can be used to present the same data; so the choice is up to the personal preference of the author. If several sets of data are to be presented simultaneously, this may be done with a frequency polygon. Histograms are usually limited to two sets of data on one figure. As with frequency polygons, histograms may have percentages or relative frequencies plotted on their ordinates.

Ogives

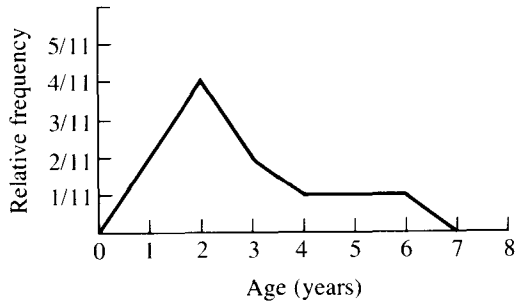
An **ogive** or cumulative percentage curve is generated if we plot the cumulated percent on the ordinate and the age on the abscissa. By picking an age on the abscissa, reading up to the curve and over to the ordinate, we can determine what percent of our patients are below a particular age. For example, from Figure 1.3 we see that 90.9% of our pediatric patients are 5 years of age or younger.

Bar Graphs

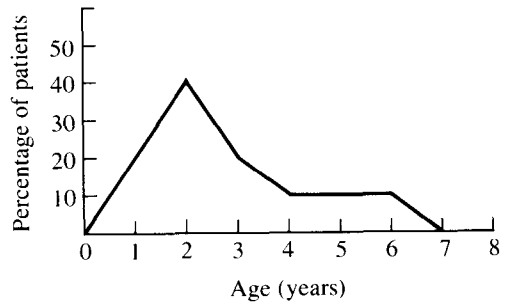
Suppose that we are studying the types of accidental deaths occurring in Lincoln County during 1972 and we find the following data: home 4, auto 6, public 5, and occupational 3. We can graph these data by listing the cause of



Frequency polygon



Relative frequency polygon



Percentage frequency polygon

Figure 1.1. *Age distribution of pediatric patients in study X, 1976*

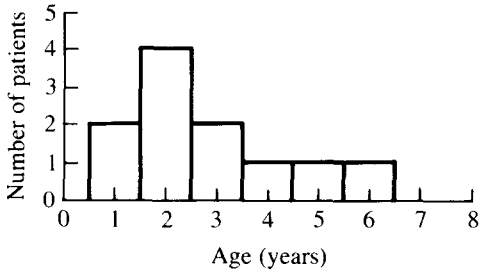


Figure 1.2. *Histogram of age of pediatric patients in study X, 1976*

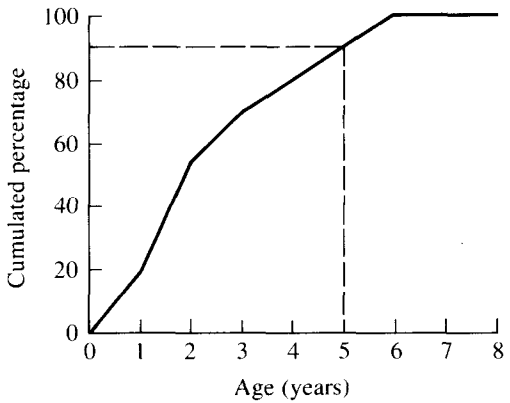


Figure 1.3. *Ogive of age of pediatric patients in study X, 1976*

death on the abscissa, the number of deaths on the ordinate, and erecting a column proportional to the number of cases of each. This is a **bar graph.** The bar graph (Figure 1.4) and the histogram (Figure 1.2) look very similar, but they are for different types of data. The data plotted on the abscissa of a histogram are assumed to have an underlying continuum and thus can be ordered, that is, age 3 comes before age 4 but after age 2. The data on the abscissa of a bar graph do not have an underlying continuum, and thus order is not important. Deaths from autos could have been plotted first rather than deaths in the home.

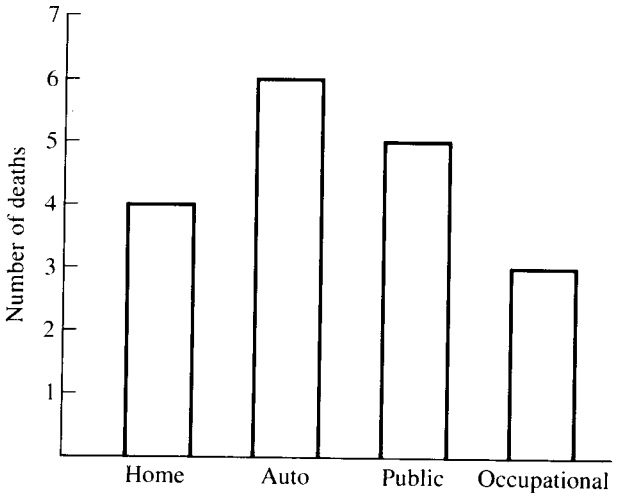


Figure 1.4. *Bar graph of causes of accidental deaths, Lincoln County, 1972*

Pie Graphs

Another way to present the data in Figure 1.4 pictorially would be to use a **pie diagram.** Here the number of deaths from each cause is converted to a percentage of the total deaths and plotted as a proportional part of a circle or pie as in Figure 1.5.

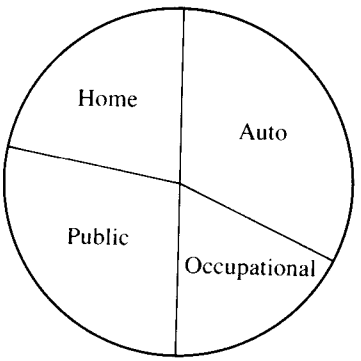


Figure 1.5. *Pie graph of accidental deaths, Lincoln County, 1972*

Shapes of Distributions

Certain terms are commonly used to describe the shapes of distributions as they deviate from "normal." We will define a "normal distribution" more rigorously in a later chapter, but for now this term is used to refer to a symmetrical, bell-shaped distribution such as Figure 1.6.

Figure 1.6. *Normal distribution*

Skewness

If most of the cases pile up at one end of the distribution so that it is no longer symmetrical, we say that the distribution is **skewed.** The tail of the distribution (the end with the smaller number of cases) determines the type of skewness. If the tail is on the left-hand side of the figure, the distribution is negatively skewed (Figure 1.7); if the tail is on the right, it is positively skewed (Figure 1.8).

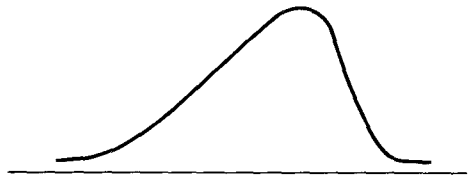


Figure 1.7. *Negatively skewed distribution*

Figure 1.8. *Positively skewed distribution*

Kurtosis

Skewness tells us something about the symmetry of the distribution. A distribution can be symmetrical and still not be normal. If we grabbed the peak of a normal distribution and either pulled it straight up or pushed it straight down, it would still be symmetrical but it would not be normal. The term **kurtosis** is used to refer to the peakedness of a distribution. A distribution that is flatter than normal is referred to as **platykurtic** (Figure 1.9). A distribution that is more peaked than normal is called **leptokurtic,** while a normal distribution is referred to as **mesokurtic.**

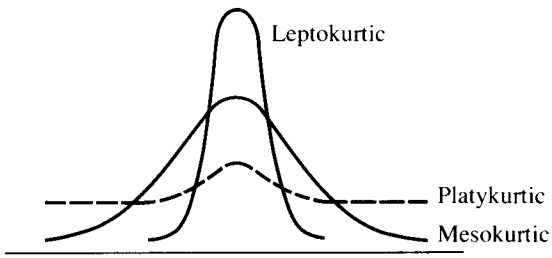


Figure 1.9. *Types of curves*

Warning

When you are confronted with a table or graph you should ask yourself, "If these data were presented in another way, would they imply another conclusion?" When we see a commercial on TV or an ad in a newspaper we know that someone is trying to sell us something; so we tend to view the material somewhat skeptically. Unfortunately we too often accept articles in scientific journals as proven fact when we should view them skeptically also. The author of a journal article is trying to sell us something too, for instance, that treatment A is superior to treatment B. Study the tables and graphs carefully and do not settle for a first impression.

If we make a frequency polygon of the marriage rate in Oregon for the years 1968–1974, we may vary the overall impression conveyed by the figure by varying the ordinate. Figure 1.10 shows a frequency polygon of these data ranging from a rate of 7.9/1000 in 1968 to 8.8/1000 in 1974. Figure 1.11 is