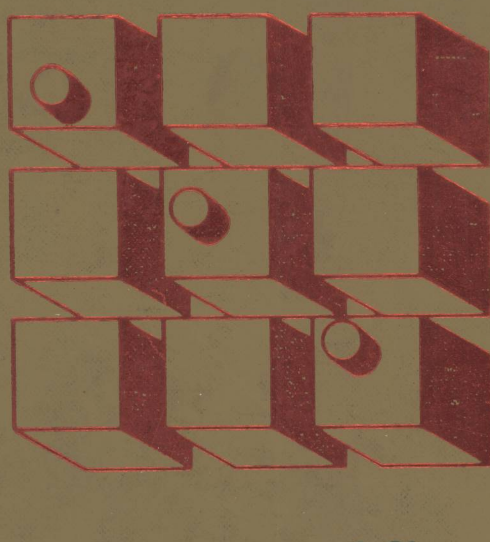
# elementary statistical procedures



Clinton I. Chase

# elementary statistical procedures

SECOND EDITION



Clinton I. Chase
Indiana University



McGraw-Hill Book Company

New York St. Louis San Francisco Auckland Düsseldorf Johannesburg
Kuala Lumpur London Mexico Montreal New Delhi Panama
Paris São Paulo Singapore Sydney Tokyo Toronto

**Library of Congress Cataloging in Publication Data**

Chase, Clinton I
Elementary statistical procedures.

Includes index.
1. Statistics. I. Title.
HA29.C5435 1976 519.5 75-14144
ISBN 0-07-010681-9

ELEMENTARY STATISTICAL PROCEDURES



1 2 3 4 5 6 7 8 9 0 D O D O 7 9 8 7 6

This book was set in Times Roman.
The editors were Stephen D. Dragin and J. W. Maisel;
the designer was Anne Canevari Green;
the production supervisor was Leroy A. Young.
The drawings were done by Danmark & Michaels, Inc.
R. R. Donnelley & Sons Company was printer and binder.

# elementary statistical procedures

# Preface

The second edition of *Elementary Statistical Procedures* is tied to the same objectives as the first edition; however, a number of ways in which these objectives are to be pursued have been altered. The intent remains to present the basic methodology of social statistics in a direct, uncluttered manner.

But two major points arose from the experience students and professors had with the first edition. First, in some parts of the book attempts to reduce a concept to its simplest components stripped it too bare. Consequently, some efforts to simplify content in the first edition were not as helpful as intended. Second, opportunities to apply skills, although presented contiguous to the appropriate concepts, were insufficient.

Therefore, in the second edition the intent remains to keep the content at a basic level, but various concepts have been expanded in an effort to reflect more faithfully the essential ideas. For example, the section on probability has largely been redone, introduction to correlation has been expanded to capitalize more on the intuitive grasp of the concepts, the *t*-test —especially the section dealing with large and small sample problems—has been extensively rewritten, and the introduction to analysis of variance has been completely revised. A number of other, less pervasive changes have also been made, but the intent in all cases is the same—to present only the essential content but with greater detail and clarity than before.

Students and their instructors often asked for more opportunities to apply the concepts developed in the text. Two or three exercises, even though placed directly at the point in the text where a process was developed, appeared insufficient. As a result the total number of exercises has been nearly doubled. The intent here is to allow the students to become more confident of their skill with a process before moving on to a new point. The additional exercises appear necessary to accomplish this.

A tangential interest in the second edition is to promote research designs more than was done in the first edition. To this end some of the research vocabulary (e.g., experimental and control groups) has been introduced, simple research designs have been used as examples, and problems have been set in the context of research activities. The intent here is not only to promote research designs but also to suggest that statistical tests are intended to be applied in real situations. Also, the students should realize that they are acquiring skills that they, themselves, can apply in the research problems they are encountering.

An introduction to computer usage has also been added to the second edition as an additional aid. Many instructors are getting their students to the computer as soon as possible because rapid data processing plays so large a role in all modern science. To facilitate student use of the computer, a brief introduction is presented in Appendix D. This place in the text was chosen because different instructors will choose different points in the course to introduce the computer. Also, the topic is not actually statistical in content, and hence it rightly should not be in the main body of the text. The introduction is admittedly brief, but sufficient enough so that a persistent student with a little guidance can get onto the computer with simple, canned programs.

During the years of the first edition many students and their instructors submitted interesting and helpful comments to me. My appreciation is extended to all these people. The changes in the second edition are in a large part a result of responses from its readers. They have been most helpful. It is hoped that readers will continue to advise me with the second edition of this book.

And finally, I am especially grateful for the help of Ms. Pat Chase, whose assistance has added immeasurably to the development of this book.

*Clinton I. Chase*

# elementary statistical procedures

# Contents

# 1

# definitions and concepts

As we read the evening newspaper, one characteristic that quickly strikes us is that feature writers load their articles with numbers. What is the probability that another baseball player will break Henry Aaron's home run record in this century? What is the average age of members of the Green Bay Packers team? What percentage of marriages ends in divorce? What is the average salary made by women faculty members at Excelsior University as compared with the average salary for men at similar ranks?

The news media pour voluminous amounts of numbers onto a public that drinks them in insatiably. Some compilations of data are merely expository in that they are intended to support no argument, while others are clearly devised (and often cleverly contrived) to support a particular point of view. How much of this information can we trust? Do we need to know more about how data are legitimately manipulated before we can answer this question?

This question leads us to the topic of this book. The popular idea of statistics as large compilations of data is *not* the topic we shall study.

Clearly all sections of society need data, but data are useful only to the extent that we can put order into the masses of numbers that social and scientific agencies collect. We therefore need methods of managing these masses of data to answer the questions we have in mind. It is these methods that are the topics of the chapters ahead. Thus, this book is not on statistics but on statistical methods.

The techniques described in successive chapters will do one of only two things. They will describe a set of data, or they will provide a basis for making a generalization about a large group of individuals when only a selected portion of such a group has been observed. Certain procedures, then, are called *descriptive methods* because they point up a characteristic of the group being observed. Other techniques are called *inferential procedures*, because they allow us to make inferences about large numbers of individuals when only a small sample from the larger group has been observed.

Let us look at some examples of these two methods. I have in my files IQs of all children in my sixth-grade classes in Lewiston, Idaho. What do these data tell me? Just thumbing through the files I cannot really say. But statistical methods can tell me what the typical IQ is, or between what two IQs the bulk of the students fall. These are descriptive methods in that they summarize a mass of data into a few simple ideas.

On the other hand, we may want to say something about a large group of people—all ten-year-old middle-class boys, for example—but we can reasonably study only a very small sample of this population. We shall look at methods which will indeed allow us to speculate about the nature of a population, once we know something about a sample from that population. These are inferential techniques, because they allow us to make reasonable projections about a large group of individuals when we have studied only a small portion of the group.

At this point we need to identify the terms *statistic* and *parameter*, but to do this we first must talk about samples and populations.

A population is any group of individuals all of whom have at least one characteristic in common. We typically think of populations in terms of census data, such as the population of a nation, a state, or a city. Certainly these concepts fit our definition, since they each describe a group of people who have a common characteristic—locality of residence. But populations can be defined on bases other than residence. For example, all women with naturally red hair or all men with beards or all people with annual incomes above $100,000 are populations, and the defining characteristic has nothing to do with residence. Populations may be large—such as all persons within the city limits of New York on a given day—or they may be small—such as all persons who serve in forest lookout stations in Clearwater County, Idaho, or all persons over one hundred years of age in California.

In research in the social sciences, as well as in other disciplines, it is often impossible to study entire populations. Instead we select smaller portions of the population and from these make inferences as to what the population is like. These smaller portions of a population we call *samples.*

In choosing samples our hope is to get a segment of the population which looks just like—is representative of—the entire population. There are several techniques for selecting representative samples, but the one on which the inferential techniques in successive chapters rely is randomization. For a sample to be truly random, all individuals in the population must have an equal chance to be selected each time a selection is made. In this way the selection of one individual has nothing to do with determining who else will be selected in the sample.

In actual practice the individuals we observe are often not truly random samples from the parent population. Therefore, we go to some length to compare the characteristics of the sample with known characteristics of the population so that we may decide whether our sample looks sufficiently like the population to venture on with further assessments. For example, we cannot observe all the six-year-old children in Detroit, and so we go to five randomly chosen public schools dispersed across the city and observe children from one first-grade class in each school. This is not a method for selecting a random sample. However, we know something about the expected varieties of IQs for all six-year-olds, we know something about the numbers of people in various socio-economic levels, and we know something about the numbers in various races in the city. With this type of information we could look at our sample to see how closely it fits the known traits of the population. If the sample seems to fit the population on several traits, we often hypothesize that it also will look like the population in other ways. This assumption may be in error; but when populations are large, truly random selection techniques are next to impossible to implement, and the above alternative is often the next resort.

When we have selected a sample from a population, we usually compute certain characteristics of that sample, e.g., an average for some trait. Characteristics of samples are called *statistics.* However, when we are dealing with populations, characteristics, such as averages, are then called *parameters.* Thus, if we define our population as all people taking elementary statistics at Eks University, the average height of these people would be a parameter. But if we put all the names of these students into a barrel, stir them well, and, blindfolded, draw out 10 names (replacing each name once it was drawn out—why?), the average height of this group of 10 people would be a statistic. Statistics tell us something about samples taken from given populations; parameters tell us something about populations.

Statistical procedures are derived out of mathematical models of what is presumed to be "reality." For example, most of us have experienced being graded "on the curve." This method of grading is based on the belief that human talents are distributed in the population in such a way that many people cluster around average, a few are above average and a few below, and a very few are very much above average and a similar-sized group very much below. If we believe that human traits are indeed distributed in this manner, mathematicians can fit to this model of the world a curve showing how proportions of the population will vary with talent level. Statisticians then exploit this curve in describing groups of people and in making inferences about populations from samples from those populations.

But the well-known bell-shaped curve is not the only model we shall exploit. Statisticians utilize a variety of mathematical models of the world's phenomena as a foundation for developing their methods. It will often help students to understand a procedure if they attend first to the model, then to the procedure.

Mathematical models not only tell us how given characteristics are expected to be distributed in the world but also provide us with a concept of the scale with which we might measure varying amounts of the condition. However, in the real world we typically measure less precisely than the mathematical model would prescribe.

Stevens (1946) has dealt with this problem in describing four classes of measurement scales. The *nominal* scale is merely the assigning of numbers to classes to identify them. Serial numbers on bicycles might be an example. The number does not tell us about varying degrees of quality; it merely distinguishes one style or model from another. Numbers worn by athletes similarly identify one player from another but tell us nothing about the size or quality of the player.

The *ordinal scale* is more sophisticated. Here larger scores reflect more of the quantity or quality, but units along the scale are unequal in size. The order of winning a race is an example. The difference in speed between the first- and second-place runners is not necessarily equal to the difference between the second and third place. A teacher-made arithmetic test is also an ordinal scale. The difficulty level of one problem is not presumed to be equal to the difficulty of other problems.

*Interval* scales have equal units and are therefore more precise than ordinal scales. However, we do not know where true zero is on the scale. The Fahrenheit thermometer is an interval scale. We have equal units along the scale, but zero on the thermometer does not indicate an absence of all heat. Standard scores on an intelligence test are also on an interval scale.

The *ratio* scale also has equal units, but zero on the scale means an absence of the quality being assessed. Measures of length and of weight are ratio scales.

Stevens (1959) has argued that because of the nature of various scales, not all statistical procedures are appropriate for all kinds of measures. For example, suppose we have collected data using an ordinal scale. Stevens' view has been that any technique which requires arithmetic manipulations of these scores, e.g., averaging, is inappropriate. How can we add units that are unequal in length?

On the other hand, Burke (1963) has argued that if our statistical methods correspond with our mathematical model of the population with which we are dealing, we need not worry much about the nature of the measurement scale. The fact that we measure clumsily does not change the nature of the population and should not deter us from using the statistical procedures indicated by our concept of the population.

Labovitz (1967, 1970) has supported Burke's position, and has indeed presented empirical evidence to bear out his argument. He concluded that researchers should use the most powerful statistical procedures available, even with less precise scales such as ordinal measures.

The controversy continues to flare up periodically. The sample-oriented researcher often prefers to select statistical procedures based on the nature of his measuring scale. The conclusion-oriented researcher who is seeking data to test hypotheses about the nature of populations may prefer statistical procedures that correspond with a mathematical model of the population.

The data we apply to our mathematical models are often referred to as variables. A *variable* is a specified condition that can take any of a set of values. For example, height is a variable, and so are mental age, the number of children in various classrooms in the elementary schools of Pocatello, Idaho, and the strength of right-hand grip for a sample of ten-year-olds.

Statistical workers deal with many kinds of variables, but all these kinds can be sorted into about two categories. In the first category we have data obtained by counting indivisible units, such as children in a family or errors on a true-false test. We call a succession of these units a *discrete* series. The data are always collections of whole numbers, since at no time do we have a part of a unit. If we are going from house to house tabulating the number of children in the family, each report is of a whole number of children. There is no such thing as a fractional part of a child. Families are made up of two, three, four, etc., children, never three and a fraction or two and a fraction. If there are two children in a family and another child is born, the number of children changes from two to three without passing through a series of fractional parts between those two numbers. The series is a discrete one.

However, many kinds of data come in units which are divisible into an infinite number of fractional parts. This is the second category of data. To expand an observation from one unit to the next, we must pass through a large number of fractional parts of that next unit. For example, when a child grows from 43 to 44 inches in height, she passes through all fractions of that forty-fourth inch before she actually acquires the fine height of a full 44 inches. Numbers which are obtained from a succession of units, each of which can be divided an infinitely large number of times, are called a *continuous* series. If we have driven 2 miles, to go 3 we must pass through all fractional parts of that next mile; if we walk a block, we must walk through all fractional parts of that segment of the street. All such measurement kinds of data represent a continuous series.

We may wish at times to visualize discrete and continuous data which progress together. For example, a mile race may mean that the runners must go around the race track four times. The actual distance of a mile represents a continuous variable; however, if we count the runner's progress by laps around the track, we are in the realm of discrete data. Either the runner has completed one, two, three laps, or he has not.

But what if a runner has completed 2.67 laps? Shall we count him as having completed only two laps? But he is really nearer to completing three laps than two. When we are counting events in a discrete series which is paralleled by a continuous one, we typically mark off units at the halfway point between two successive numbers. For example, if we are counting inches, the unit labeled 4 inches would run from 3.5 to 4.5. The unit labeled 2 would run from 1.5 to 2.5, and so on, as shown in Fig. 1-1. In other words we shall think of any given number in a continuous series as the midpoint of a unit. The unit may be small ($\frac{1}{8}$ inch, an ounce) or it may be large (a mile, a decade), size being a relative matter. In any case, the numbers we use to label the units will represent the midpoint of the unit. Thus, if our unit is $\frac{1}{2}$ inch, the number 3 would represent the third half inch in a series of $\frac{1}{2}$-inch units, and it would be the midpoint of that segment of length that runs from $\frac{1}{4}$ inch below the third half inch to $\frac{1}{4}$ inch above it. If our unit is the mile, 3 would represent the distance from 2.5 to 3.5 miles, etc.

Thus, in statistical methods we often have occasions to deal with a discrete series of units which is paralleled by a continuous series. When this occurs, the units in the discrete series label the midpoints of units in the continuous series, and the units in the continuous series are thought of as beginning a half unit below the labeled point and extending a half unit beyond it.

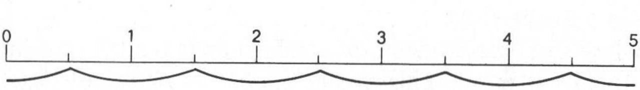



**Figure 1-1**

## SUMMARY

Not everyone today is a statistician, but everyone, almost without choice, must be a consumer of statistics. In either case, knowledge of basic methods of manipulating data is salient to intelligent behavior in a quantitative world.

This is a book about methods of dealing with data—statistical methods. Some of these methods tell us how to describe a body of data. These are called descriptive statistics. Other techniques allow us to make judgments about a large group of individuals when we have observed only a carefully chosen segment of the total group. These techniques are categorized under the term statistical inference.

All individuals who have one or more characteristics in common can be defined as a population. Characteristics of a population are called parameters. A collection of individuals who represent a portion of a given population is called a sample. Characteristics of samples are called statistics.

Conditions observed by social scientists provide numbers. These numbers represent a discrete series, in which each unit is an indivisible whole, or a continuous series, in which case a given unit may be divided an infinite number of times. A discrete series is reported only in whole numbers, such as the number of students in a class, whereas a continuous series may involve combinations of whole numbers and fractions.

## REFERENCES

Burke, Cletus J.: Measurement Scales and Statistical Models, in M. H. Marx, "Theories of Contemporary Psychology," pp. 147–159, The Macmillan Company, New York, 1963.

Labovitz, Sanford: Some Observations on Measurement and Statistics, *Social Forces*, **46**:151–160 (1967).

———: The Assignment of Numbers to Rank Order Categories, *American Social Review*, **35**:515–524 (1970).

Stevens, S. S.: Measurement, in C. West Churchman, "Measurement: Definitions and Theory," pp. 18–36, John Wiley & Sons, Inc., New York, 1959.

———: On the Theory of Scales and Measurement, *Science*, **103**:667–680 (1946).

---

## PROBLEMS*

1 List four populations other than those tied to geographical boundaries such as city limits or state boundary lines. For each of these populations describe how you might draw a random sample.

* Problem answers are found in Appendix E.

2 Label each of the following conditions as a parameter or a statistic, and state why you have so labeled them.
- **a** The average age of all red-haired children in the sixth grade in Centerville Public School
- **b** The proportion of children who are left-handed among a sample of 40 children stopped at random on the playground of Rogers Elementary School
- **c** The average weight of 10 girls whose names have been drawn from a hat containing names of all girls in the senior class at Centerville High School
- **d** The range in height (the difference between the shortest and the tallest) of all the males who were convicted of a felony in the State of Oregon in 1973

3 Decide whether the following variables represent discrete or continuous series.
- **a** The number of dogs caught by the dogcatcher in a given city on December 12, 1960
- **b** The time it takes a rat to run through a T maze
- **c** The number of children whose fathers are college professors at Eks University
- **d** Amount of change in height for a given child in a 6-month period
- **e** The number of birthdays a sample of 10 elementary school children have had

# 2

# frequency distributions and graphic representation of data

The collection of research data results in an accumulation of numbers which stand for various amounts of some observed condition. If we wish to know if drug X has an influence on the number of errors rats make in running a maze, the numbers that we collect in our research could represent errors made by rats. If we want to know the effect of instructional method A, compared with method B, on the learning of arithmetic in grade 5, the numbers we collect here might represent the numbers of correct answers on an arithmetic achievement test given to students who were taught A and students who were taught B. In any case, researchers are collectors of numbers, and they seek out ways to put order into the masses of numbers they collect.

## FREQUENCY DISTRIBUTIONS

One of the most fundamental techniques for putting order into a disarray of data is the frequency distribution. Basically it is a systematic procedure for arranging individuals from least to most in relation to some quantifiable