Jerald G. Schutte

Everything You Always Wanted To Know About Elementary Statistics

(but were afraid to ask)

Jerald G. Schutte
Columbia University

Everything You Always Wanted To Know About Elementary Statistics

(but were afraid to ask)

Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging in Publication Data

SCHUTTE, JERALD G
Everything you always wanted to know about elementary statistics (but were afraid to ask).
Includes bibliographies and index.
1. Statistics. I. Title.
HA29.S425 1977 519.5 76-46638
ISBN 0-13-293506-6

PRENTICE-HALL METHODS OF SOCIAL SCIENCE SERIES

Herbert L. Costner and Neil Smelser, *Editors*

Printed in the United States of America

10 9 8 7 6 5 4 3

Prentice-Hall International, Inc., *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall of Canada, Ltd., *Toronto*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Prentice-Hall of Southeast Asia Pte. Ltd., *Singapore*
Whitehall Books Limited, Wellington, *New Zealand*

Prologue

As a social science researcher, but more important, a teacher, I suppose I should be horrified by the plague of panic and hostility that infects so many students having to study social science courses dealing with statistics. However, in a very real sense, I believe there are two fundamental reasons for this epidemic.

The first element reflects an increasing alienation from the use of numbers and statistics to describe the human condition. From the moment a baby is fitted with plastic wristband containing a number matching his mother's and identifying parentage, he is packaged, filed, and labeled in a hundred different numerical ways. Despite our impotent protests, society assigns us numerical identity on student cards, library identification, intelligence tests, driver's licenses, social security records, and even graveyard plots. We are further asked, almost daily, to internalize large doses of numerical information from all forms of mass media. We must understand weather reports, ozone counts, television ratings, political polls, and market pricing. We are forced to accept governmental decisions on the basis of statistical averages, not human emotions, and mathematical projections, not individual needs. Even our leisure activities seem bound up in numerical bureaucracy: fishing licenses, camping reservations, and even lotteries to determine who receives season tickets to athletic events. Is it any wonder many people tend to grow up with an increasing distaste for the use of numbers?

Although the reasons for this distaste are many, they seem to be compounded by a second element. This problem reflects several basic obstacles in learning to understand, accept, and use mathematics. The first obstacle is the public school system's apparent inability to pace the introduction of mathematical concepts with a child's developing capacity to comprehend. Students' introduction to the most abstract, and yet the most basic assumption of mathematics, the number system, comes at a time when they are least able to cope with abstractions. When we are taught to count, the use of ice cream sticks and tongue depressers as the unit, tens', and hundreds' place holders more likely conjures up a picture of popsicle orgies or dreaded trips to the doctor's office than integers in the real number system. As the education continues, the notion of pie charts seemingly evokes more pangs of hunger than visions of fractions. These valid but poorly timed techniques are continued over the years, interrupted at regular intervals by assignments not quite comprehended and tests on material not quite mastered. No wonder the unfortunate student despairs of ever understanding mathematics—he has become too bogged down in understanding the preliminary assumptions.

A second obstacle is that mathematics, like many other skills, involves certain essential chores of memorization. Even the advanced mathematician must consciously remember that $7 \times 8 = 56$. It is hardly possible to continue understanding higher-order abstractions unless we have the basic computations of addition, subtraction, multiplication, and division memorized. Yet we ask children to memorize these tables at a point in their life when it seems the attention span is at a minimum. It is small wonder that when later called upon to perform, our thought process creates anxiety over the inability to remember such seemingly simple things.

A third problem is generated by the weakness students often sense in the teacher's own grasp of mathematics. Sketchy explanations, vague instructions to "do the problem the best you can; you'll probably understand it as you do the work" deceives no one. Arbitrary conventions and lack of intuitive reasoning handicap the very authority whose function it is to make all these things clear. The message to the student becomes almost a self-fulfilling prophecy of doubt, rejection, and failure.

The fourth obstacle appears to be one brought about by fear of labeling within the peer group. If confusion and avoidance of math become the norm, words like "egghead," "brain," or "kissy" (to mention the least offensive) may be attached to those who have developed the reasoning abilities to cope with numbers. For females, the emphasis is often on social conformity and may entail the tacit awareness that doing well in math class risks *not* doing well with the men in that class.

For males, there is often a masculine mystique that emphasizes success with women and athletics; academic grinding is to be avoided at all costs.

One major consequence of these obstacles is that many college freshmen, having been caught up in this pattern, tend to draw away from the hard sciences into the social sciences, which have the reputation of being less quantitative. A great many of these students, convinced they lack the ability to deal with numbers, soon discover that practically every social science major requires a class in statistics; the ensuing encounter with the necessary statistics course becomes an almost universally dreaded and demoralizing walk down the path to pessimism and panic. The accompanying symptoms of ulcers, insomnia, and tears are often perpetuated by putting the class off until the senior year, where it becomes mandatory to pass in order to graduate.

This book is written with these students in mind. That is, it is a supplementary text for students in social science statistics classes. It is not meant to replace existing texts on the subject, although in some courses it could be relied upon to provide the main introduction to the field. Moreover, it is not intended to present a detailed mathematical treatment of the subject. Indeed, its purpose is to avoid the tortuous mathematical theorems and assumptions that account for most of the student's hesitations already. However, the book does not try to eliminate the mathematics (as is often done) in treating a statistical application. Rather, it periodically introduces the mathematical logic behind some formulae and derivations. By doing so, in a number of the simpler cases, it is hoped that the reader will generalize the comfort found in understanding these derivations to situations where the formulae appear more complex.

The technique used is a question and answer format. This is not meant to be a programmed approach. Rather, the manuscript actually derives from students over the past several years. These are their questions, and to a large extent, their answers. What I have tried to do is compile them in a coherent manner and provide supplementary information where I felt it was needed. The educational device employed is not the typical social science example used to illustrate certain statistical tools. Instead, I have tried wherever possible to introduce topics with the literary technique of analogy and metaphor. The reasoning is that a statistical concept (like any other concept) is rendered much more intuitive if removed from its world of mathematical symbols and anchored in the perceptual experience of the individual's everyday life. Thus, the student with a mental block for mathematics is more apt to remember the term *mean* by recalling a teeterboard or see-saw at the playground, or the concept of a *distribution* by imagining a roller-coaster, or the

rationale for *permutation* by thinking of horse races, than he ever will by drawing on his storehouse of partially remembered mathematical derivations.

The approach does not end here, however. After the concept is defined and an analogy is established, the symbolic formula and a more traditional example are elaborated. Where feasible, the mathematical rationale is explained. Lastly, reference to the ad hoc and historical nature of statistical terms and symbols is made in the Appendixes.

I believe this approach has three benefits. As a supplement, the student will want to refer to it for specific information. The question and answer format, broken down by chapter and following a logical progression of ideas, provides a more readable reference. Secondly, by providing a brief summary of the history of the methods and symbols, some of the mystique is taken from statistical analysis. Finally, by approaching the subject matter in a less complicated manner, the student is likely to internalize statistics with less distaste for the mathematics involved.

Part I serves as an introduction to the subject matter, focusing on the basic vocabulary of statistics (chap. 1) and the various levels of measurement (chap. 2). Part II deals with descriptive statistics. Having made the distinction between categorical and quantitative measurement, I treat them in chapters 3 and 4 respectively with discussions of the tables and graphs for each. Chapter 5 begins with a statement on central tendencies and chapter 6 relates these to measures of dispersion, including the concept of z-scores. Finally, chapters 7 and 8 introduce techniques useful for describing two quantitative variables, elaborating the ideas behind regression and correlation respectively.

Part III provides an introduction to hypothesis testing. In chapters 9 and 10 probability and probability distributions, such as the binomial, are related to the normal curve. Chapter 11 enumerates some of the ground rules for testing hypotheses when using sampling distributions. Chapters 12 to 14 treat specific cases of these tests. Chapter 12 illustrates the single sample case of testing and estimation. Chapter 13 explains the sampling distribution of the difference of means and introduces the F distribution. Chapter 14 extends the use of the F distribution and presents an example of simple one-way analysis of variance.

Part IV discusses special nonparametric techniques. In chapter 15 nonparametric tests of hypotheses are illustrated, and in chapter 16 traditional nonparametric relational measures are discussed. Finally, four appendixes are included: (I) historical notes, (II) a discussion of relevant statistical symbols, (III) a review of forgotten algebraic operations and tips on use of the slide rule and calculators, and (IV) a set of tables of statistical inference.

The book can be utilized in several different ways. As an introduction,

students may wish to read it first to introduce themselves to the subject matter in a main text. As a reference, the book can serve as further explanation for individual questions. As a supplement, students may utilize specific sections, such as the historical notes, the appendixes on algebra, slide rule, and statistical symbols, or the selected reference sections at the end of each chapter.

I have undoubtedly left out material that some will deem important. I may have included information that some will think is extraneous. However, in an effort to coordinate this subject matter with existing texts in the social sciences, I have tried to promote two ideas: never present a statistical technique until it is intuitively expressed; and always anchor this intuition in an analogy representing the everyday experiences of the student.

Acknowledgments

I wish to acknowledge the critical support of a number of people: Dr. Herbert Costner, for his copious reviews of several versions of this manuscript; Ed Stanford and his staff at Prentice-Hall, for their undying patience and renewed dedication to the concepts underlying this effort; James Keeshan, for his self-sacrificing service in producing the cartoon illustrations; John Light, for his continued support and feedback during the last stages of writing; and finally, my family, friends, and colleagues, without whose continued devotion and understanding I would have never mustered the motivation to put my ideas on paper.

JERALD SCHUTTE
Los Angeles, California

Contents

Appendixes

Introduction

I

Numbers, symbols, and words

A STUDENT'S LAMENT

1

I don't understand it! I don't understand it!

Why does a person need to study statistics?

Like it or not, you have been exposed to certain forms of statistics the better part of your life. Most of what we call current events are made known to us via such reports as opinion polls, stock averages, unemployment rates, and Gross National Products. Sports

and weather reports have familiarized us with concepts such as averages, percentages, and probabilities. And advertising has sensitized us to the differences between groups with fewer cavities, shinier hair, and softer hands.

Yet in each one of these political, social, and economic contexts, the availability of masses of information suggests a great deal of confusion and misuse. Therefore, it is essential for the serious student interested in contemporary society to gain some understanding of the concepts involved in the study of statistics, both to guard against inaccurate information and to make intelligent decisions.

Do I have to be good in math to understand statistics?

The language of statistics is powerful, yet the mathematical operations are as simple as those encountered in the fourth grade. At no time are we asked to deal with more than four basic operations: addition, subtraction, multiplication, and division. At first glance, these may appear to be combined in strange ways (e.g., a special case of dividing is finding the square root; a special case of multiplying is finding the power of a value). They may even be used to produce strange-sounding terms (e.g., standard error of the mean). But you will never be asked to compute anything that does not boil down to the four basic operations.

But I can't seem to work with numbers!

Many students take this attitude in social science statistics classes. The pattern appears to be the following: the first day of class there is a feeling of guarded optimism, "I can do it, I can do it." Tinges of panic set in after a lecture or two, "I must do it, I must do it." Then, the first homework assignment is encountered and the feeling is, "I don't understand it, I don't understand it." Finally, a missed lecture or a disappointing talk with the professor, and the student has confirmed what he suspected in the beginning, "I'm no good with numbers, I knew it; I'm just not good with mathematics at all." This problem is often intensified when the dreaded statistics requirement is ignored, only to be taken in the senior year. By this time the lag between the last high school math class and this sadistic requisite may have been four to six years.

The experience is particularly unfortunate because the problem is not one of having to feel comfortable with numbers specifically, or statistics in general. Rather it is to feel comfortable with the idea of using numbers and symbols, much the same as we do words, in creating

a statistical language which describes and predicts events in the "real" world.

Is statistics really a language or just masses of information?

Don't confuse *what* is being measured with *how* it is measured. Information such as unemployment or birth rates may communicate more or less depending upon the context and how it is put to use. This has very little to do with the means by which it is measured. The study of statistics provides us with a set of symbols and operations which become the means to accomplish this measurement. Just as words are labels which represent certain tangible or abstract concepts, statistical symbols and operations are also labels which can be used to eliminate some of the problems brought about by the ambiguity of using words to describe information and infer properties from it. When we combine numbers and symbols in certain specified ways, deriving more powerful concepts (e.g., mean and variance) and represent them with various Greek and English letters, we have an impressive language. And that is precisely what statistics is, a language. In everyday English, we describe situations by pointing out similarities and differences among events; statistics accomplishes the same goal. The beauty of a statistical vocabulary is that we can do this in a much more precise, less inconsistent manner than is possible through the use of everyday language.

As mentioned earlier, there are four basic mathematical symbols in the vocabulary of statistics ($+$, $-$, $\times$, and $\div$), one corresponding to each operation. However, to make a statement in the vocabulary we introduce a fifth symbol which acts as the verb in our language; the sign is equality ($=$). We can now combine symbols and numbers to make statements: $4 + 5 = 9$. This is a statement about the sum of 4 and 5. However, we could make this even more abstract by letting other symbols stand for the numbers: $x + y = z$. This is a general case in which we can substitute the particular values we happen to have. In the vocabulary of statistics, we represent combinations of general statements (composed of symbols and operations) by other symbols, e.g., $\bar{X} = \Sigma X_i / N$ and $s^2 = \Sigma(X_i - \bar{X})^2 / N - 1$. In this way we build each successive concept on those preceding it. Each term is given a name and it is these names which comprise the vocabulary of statistics.

However, just as we have basic symbols corresponding to the primary operations, so we have several basic definitions which define other aspects of the vocabulary of statistics. As in all logical systems, these definitions are givens and we must learn their meaning to use the language. We begin with the most general word statistics.

What does the word statistics *mean?*

Statistics (plural) is a branch of applied mathematics specializing in procedures for describing and reasoning from observations. It is generally acknowledged to be divided into two areas: descriptive and inferential statistics. The function of **descriptive statistics** is to describe observations collected in populations and samples. The function of **inferential statistics** is to infer something about populations, given descriptions of a sample.

How do I know whether I'm dealing with a population or a sample?

Any collection of objects, events, people, etc., which is defined because of some uniqueness, is called a **population**. The number of whales in the world can be a population as can the number of students in the class. The number of elements in a compound as well as the number of people at your next party qualify as populations. In short, a population is what you choose it to be. It is the unit whose characteristics you care to describe by observing them, directly or indirectly.

Unfortunately, some populations can be quite unwieldy. It would be difficult, at best, to measure something about all the men in the world—height, for example. Not only would it be expensive, but I dare say you may get lost trying to find all of them. Luckily, however, we can solve this problem by looking at a sample of them. A **sample** is a subset of a population. By looking at this subset and describing it, we can infer something about the larger population from which it came.

It is easy to see that at one time the students in your class may qualify as a sample, while at another they could be a population. The difference is whether you care to generalize the results to a larger group (as with a sample) or are content to describe only those you have selected (a population).

Why are different symbols used for populations and samples and what is the distinction?

A symbol used to indicate the measurement of a variable in a population is called a **parameter**. A symbol used to represent a given measure in a sample is called a **statistic** (singular). Greek letters are generally used for population parameters and English letters are used for sample statistics. In describing variables in a population and