Introduction to Statistics

A Fresh Approach

Gottfried Noether

Gottfried E. Noether
University of Connecticut

to Statistics

A Fresh Approach

Houghton Mifflin Company
Boston

New York
Atlanta
Geneva, Illinois
Dallas
Palo Alto

Printed in the U. S. A.

Library of Congress Catalog Card Number: 77-135750

ISBN: 0-395-05001-4

Editor's Introduction

The introductory course in statistics has always represented a serious pedagogic problem. While it will be terminal for many students, for others it will provide a basis for studying specialized methods within their major fields. The main function of such a course is to introduce them to variability, uncertainty, and some common statistical methods of drawing inferences from observed data. Such a course should be intellectually stimulating; the ability to reason statistically should be more crucial to the success of a student than his mathematical ability.

Traditional textbooks do not generally serve the needs adequately. In most cases the basic statistical ideas do not appear until very late. Valuable time is first devoted to the intellectually dull task of organizing and graphing data, and to interminable computations of means, modes, medians, and variances. Sometimes effort is still devoted to rather pointless discussions of skewness, kurtosis, and geometric means. Students with little interest in mathematics find a week or more of combinatorics frustrating if not mystifying. Finally one-quarter or one-semester courses which attempt to cover most of the important methods in statistics run out of time or become dull compendia of formulae. It is preferable to concentrate on basic ideas and a few methods and to let the applied departments introduce specialized techniques when needed at a small cost of time.

What are called for are innovative approaches which avoid these pitfalls and are efficient in exhibiting many basic ideas of inference in contexts which are intellectually stimulating. Professor Noether has presented us with such a book. The main technique is the exploitation of nonparametric methods. This book is not a study of nonparametric methods in statistics. Rather, it is the effective use of these methods to illustrate statistical ideas.

What are the advantages of this approach? The basic ideas of inference appear at the beginning. The methods developed are easy to apply and require a minimal amount of computation. The methods are simple in principle; the common sense logic behind them is easy to perceive and to explain. The probability background used in most of the illustrations is minimal, requiring often no more than easy applications of the binomial distribution and its tables. Although no attempt is made at complete coverage, this vehicle is remarkably effective in covering many basic ideas with a few methods and illustrations. Finally, there are the advantages of the robustness and of the wide applicability of the nonparametric methods discussed.

This is a well conceived, carefully thought out, and well written book. I fully anticipate that Professor Noether's success with a preliminary version will be shared by many teachers who use this book.

Herman Chernoff

Preface

In writing this book I have tried to avoid two dangers that seem ever present when teaching a first course in statistics to nonmajors. At one extreme students are overwhelmed by probability theory. At the other extreme they are bored by what are to them endless and mostly meaningless computations. In the present book, topics in the theory of probability and in descriptive statistics are held to bare essentials.

The aim of this book is to allow students to concentrate on basic ideas without becoming involved in technical and computational detail. Accordingly, such concepts as estimation and hypothesis testing are discussed in terms of the binomial model, which is much simpler than the normal model. One-, two-, and k-sample problems are solved nonparametrically before the student is introduced to t-tests and the analysis of variance; rank correlation precedes least square regression and product moment correlation. This arrangement is based on my conviction and experience that nonparametric methods are not only safer—particularly in the hands of beginners—but are also conceptually and computationally simpler than the corresponding normal theory methods.

A reasonably good high school background in mathematics is sufficient for the book.

The book is intended chiefly for a one-semester or two-quarter course; but it provides sufficient subject matter for a year course. A flexible organization permits the instructor of a one-semester course to choose either a purely nonparametric or a combination nonparametric-normal theory approach. I have covered the contents of Chapters 1 through 16 in a one-semester course, meeting three times a week, by omitting all or most of Chapter 4 on random variables.

A possible compromise between complete omission and complete inclusion of Chapter 4 in a one-semester course is to take up, rather briefly, the ideas of Sections 21 through 24 before going on to Chapter 5.

Before attempting a careful study of Chapters 18 through 21, I find it desirable to cover all of Chapter 4 so that the students will have a fuller understanding of the concepts of mean and variance of random variables and sums of random variables.

An instructor, who feels that students who will spend only one semester on the study of statistics should have some acquaintance with t-tests, can replace Chapter 15 (nonparametric k-sample procedures) and Chapter 16 (rank correlation) with parts of Chapters 18 and 19 (one- and two-sample normal theory procedures, respectively).

In Chapters 12 through 16 the following sections may be omitted without loss of continuity:

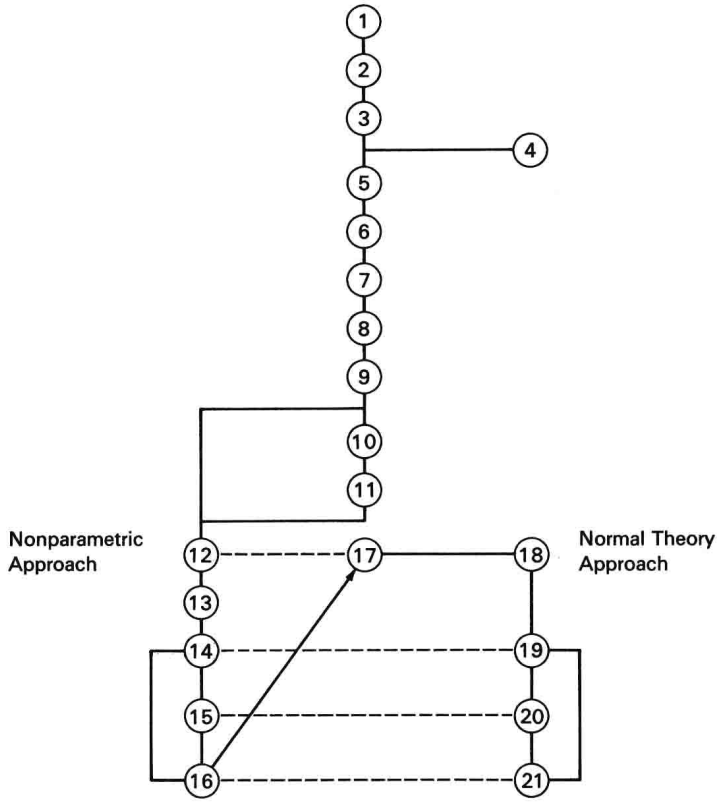
Chapter 12: Sections 87–90
Chapter 14: Sections 102 and 109 (Section 105 is needed only for an understanding of the Kruskal-Wallis test in Chapter 15)
Chapter 15: Sections 114, 116–119
Chapter 16: Section 125.

In a two-quarter course it should be possible to cover most of the material through Chapter 19. But again, an instructor may want to replace Chapter 14 with sections of Chapter 20 on the analysis of variance and/or Chapter 16 with sections of Chapter 21 on linear regression.

A few problems at the end of individual chapters are starred. These problems are somewhat theoretical in nature and in some cases require knowledge of Chapter 4.

I am grateful for permission to reproduce and/or use certain published tables. Acknowledgement is made at the appropriate place. I express my deep appreciation to Herman Chernoff, Ralph D'Agostino, and John Pratt who have read the entire manuscript and offered many helpful suggestions. I take full responsibility for any shortcomings. I am thankful to graduate students Michael Barthel and Teng-Shan Weng for checking computational details and to Miss Sandra Tamborello and my daughter Monica for typing the manuscript. Lastly, I am indebted to Houghton Mifflin Company for their assistance in bringing out this book.

Gottfried E. Noether


1
2
3
4
5
6
7
8
9
10
11
Nonparametric Approach
12
17
18
Normal Theory Approach
13
14
19
15
20
16
21

Contents

1

What Is Statistics?

"If experimentation is the Queen of the Sciences, surely statistical methods must be regarded as the Guardian of the Royal Virtue."
(From a letter to *Science* by Myron Tribus)

1 To most people the word "statistics" brings visions of endless columns of numbers, mysterious graphs, and frightening charts that show how the government is spending our tax money. At one time the word referred exclusively to numerical information required by governments for the conduct of state. Statisticians were people who collected masses of numerical information. Some statisticians still do, but many do not. They help in the conduct and interpretation of scientific experiments and professional investigations. Along with changes in the work of statisticians has gone a change in the meaning of the word "statistics" itself. "Statistics" may refer to numerical information, like football statistics or financial statistics. However the word also refers to a subject, a subject like mathematics or economics. When we use the word "statistics" in our discussions, we shall usually have this second meaning in mind. The few times when the word refers to numerical information as such, the fact will be clear from the context.

How, then, can we describe the field of statistics? A U.S. Civil Service Commission document says that "statistics is the science of the collection, classification, and measured evaluation of facts as a basis for inference. It is a body of techniques for acquiring accurate knowledge from incomplete information; a scientific system for the collection, organization, analysis, interpretation and presentation of information which can be stated in numerical form." I just hope that this definition does not keep anyone from wanting to study statistics!

Actually, the following statement taken from the book *The Nature of Statistics* by Wallis and Roberts is much more appropriate for our purposes: "Statistics is a body of methods for making wise decisions in the face of uncertainty." While this may not reflect as precise and all-inclusive a view of statistics as the earlier statement, it describes very succinctly that aspect of statistics that will be of greatest interest to us in this book, namely, how to use incomplete information to make decisions and draw valid conclusions.

The kind of decisions (or conclusions) that a statistician has in mind are decisions based on numerical information of one kind or another. Most of us are quite accustomed to making decisions of this sort. For example, after having noted on several occasions how much time it takes to get to the station or airport, we decide on how much time to allow in the future for such a trip. In general, experience coupled with common sense is quite enough to produce satisfactory results.

But there are many problems of a statistical nature for which common sense alone is unable to provide adequate answers. What are needed are more formal methods—methods of *statistical inference* as they are usually called—that have been developed and analyzed by mathematical statisticians using the calculus of probability. Starting with Chapter 6 we study some of the simpler such methods and the rationale behind them. In Chapters 2–5 we learn some basic probability. And during the remainder of this chapter we illustrate certain ideas that are basic for all of statistical inference. Of necessity this preliminary discussion will rely heavily on intuition.

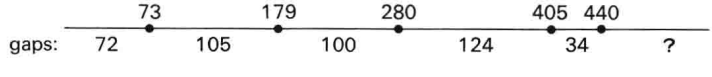
The Taxi Problem

2 Suppose that you are waiting at a street corner for a taxi. Several drive by, but they are all occupied. You begin to wonder how many taxis there are in this city. Clearly there are not enough to go around. But how many? You start watching the numbers on the shields of the taxis going by:

405, 280, 73, 440, 179.

The next taxi stops and picks you up. You could simply ask your taxi driver how many taxis are registered in the city. But before you do so, you wonder whether on the basis of your observations you cannot make an educated guess. Indeed you can, as we shall see.

We represent our information graphically by marking the 5 observed numbers on a line as follows:



The endpoint on the left corresponds, of course, to taxi #1. We should like to know the value of the endpoint on the right. How can we make a guess? It is certainly reasonable to expect that the gap to the right of the largest observed number, 440 in our case, is similar in size to the observed gaps between the known numbers. So let us compute the average gap size and add it to the largest number. We then find

$$\begin{aligned} 440 + \frac{72 + 105 + 100 + 124 + 34}{5} &= 440 + \frac{435}{5} \\ &= 440 + 87 \\ &= 527. \end{aligned}$$

Using statistical terminology, we would say that our *estimate* of the total number of cabs is 527. Having arrived at an estimate, we are, of course, curious how good our estimate is. Our taxi driver informs us that there are 550 taxis in the city. Thus our estimate is in error by 23, less than 5 per cent, which is quite good considering that we observed only 5 out of 550 cabs.

By now, some students may be thinking that we are playing a rather silly game. Since we actually want to pursue the example further, it may be helpful to point out that this game, which goes by the name of *serial number analysis*, was most useful during World War II. There, German tanks took the place of taxi cabs. It was found at the end of the war that statistical estimates of German tank production were much more accurate than estimates based on more orthodox intelligence sources.

A little earlier we mentioned that we were going to illustrate certain basic ideas of statistical inference. Statistical inference is concerned with drawing useful conclusions from available data, usually called the *sample*, about a larger aggregate called the *population*. In our taxi example the sample consists of the taxis numbered 73, 179, 280, 405, and 440 that we happened to observe. The population consists of all taxis from 1 to 550. In this example the quantity that has the greatest interest is the number 550. Before the taxi driver told us, this number was unknown to us. But by looking at the 5 numbers in our sample, we were able to make a rather good guess. In other words, we found an *estimate* of the unknown quantity.

A different kind of problem arises if we have some preconceived idea of what the total number of taxis might be. Somebody might have assured us that there are at least 1000 taxis in the city. After waiting unsuccessfully for a taxi to pick us up and noting that the largest number in our sample is only 440, we may have some doubts about the claim that there are that many taxis around. We have a *statistical hypothesis* that we want to test. The hypothesis states that there are at least 1000 taxis in the city. But our sample information throws such great doubt on the correctness of the hypothesis that we prefer to believe otherwise. We may argue something like this. If it is really true that there are at least1000 taxis around, in a sample of 5 there should ordinarily

be one with a number greater than, say, 600. The fact that the largest number in our sample is only 440 makes the original hypothesis rather untenable. Presumably the true number of taxis in the city is considerably less than 1000. In Chapter 3 we get more definite evidence to this effect.

3 Estimation and tests of hypotheses are the two main problems of statistical inference that we are going to discuss in our course. There are many different aspects to testing and estimation. Another look at the taxi estimation problem suggests some of these.

It has perhaps occurred to some of you that there are other estimates of the total number of taxis than the one we have used. Another look at our graph suggests that the middle number in the sample, i.e., 280, is approximately halfway between the two endpoints. Since the upper endpoint represents the total number of cabs and the lower endpoint is 1, we can use this fact to find another estimate of the total number of cabs. However, before we look at details, let us introduce some symbols and define an important concept.

We denote the total number of cabs by the symbol T. Next, suppose that we have a set of numbers, like our taxi numbers. When we arrange these numbers according to size, say, from the smallest to the largest, the number in the middle is called the *median* of our set of numbers. In our group of 5 taxi numbers, the median is 280. In general, we denote the median of a set of numbers by the symbol M.

Our earlier remark can now be expressed as follows. The median M of our observed numbers is approximately halfway between 1 and T. Now, the exact value of the number halfway between 1 and T is $(1 + T)/2$, and, therefore, M is an estimate of $(1 + T)/2$. Accordingly, $2M - 1$ is an estimate of T. In our case,

$$2M - 1 = 2 \times 280 - 1 = 559,$$

so our second estimate of T is 559.

The fact that there is a second estimate raises the following problem. In practice, which estimate should we use, the first one or the second one, or possibly still a third one? Probably with a little thought, you can come up with still different ways of estimating the total number of taxis.

In our example, the second estimate is clearly better than the first estimate, since 559 is closer to the true value 550 than is 527. But of course, in general, there is no taxi driver around to tell us the true value after we have computed our different estimates. The Germans certainly did not tell the Allies how many tanks they had built during a given period of time. The Allies only found out for sure after the end of the war, when the information was purely academic. In most statistical problems, we never find out the true value.

How can we decide then which estimate to use, the first or the second? Before we give an answer to this question, let us look at some additional data, representing three more sets of 5 observations of taxi numbers. These numbers were taken from tables of *random numbers* rather than observing actual taxi-cabs. We will hear a great deal more about random numbers later on. At this moment all we have to know is that the results are comparable to an experiment in which three statisticians stand at three busy street corners in a city with 550 taxis numbered from 1 to 550, each writing down the numbers of the first five taxis that happen to pass by. The numbers in our original sample as well as those in the additional three samples are listed in Table 1.1. Easy com-

TABLE 1.1

Sample 1	405	280	73	440	179
Sample 2	72	132	189	314	290
Sample 3	485	65	108	382	298
Sample 4	450	485	56	383	399

putations produce the estimates in Table 1.2, where the first

TABLE 1.2

	gap estimate		*median estimate*	
Sample 1	527	(23)	559	(9)
Sample 2	376	(174)	377	(173)
Sample 3	581	(31)	595	(45)
Sample 4	581	(31)	797	(247)

estimate is called the gap estimate and the second estimate is called the median estimate, and where the numbers in parentheses indicate by how much each estimate is in error.

One thing is clear from these computations. Neither estimate is consistently better than the other. In two samples the gap estimate happens to be closer to the true value; in one case the median estimate is closer. And in one case, the two estimates practically coincide. On the basis of the available information a clear-cut decision as to which estimate is better is not possible. However, by continuing the experiment we would eventually find that on the average, the gap estimate is somewhat closer to the true value than the median estimate. Some indication of this can be found in our samples. On the average the first estimate deviates less from the true value than the second estimate. While the average error of the median estimate is 118.5, that of the gap estimate is only 64.8. This is where the mathematical statistician comes in. With the help of the theory of probability a mathematical statistician can not only determine which of two estimates is better, but can also determine how much better one estimate is than the other. However such investigations are beyond the scope of our course. In general we have to take the word of a mathematical statistician that a recommended procedure has desirable properties.

There is a more convenient way for computing the gap estimate.