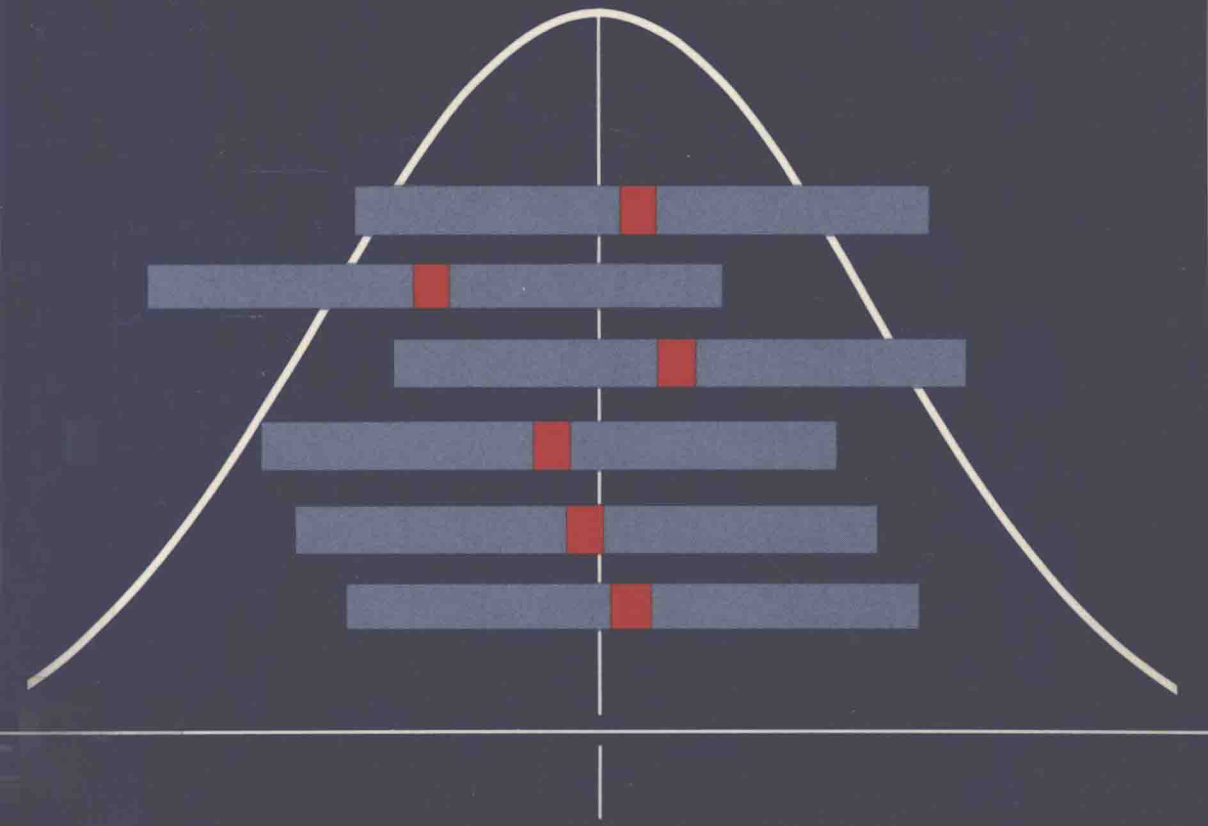
# Statistics
## Concepts and Controversies



David S. Moore

# Statistics

## Concepts and Controversies

**David S. Moore**
Purdue University

**W. H. Freeman and Company**
New York

*Sponsoring Editor:* Peter Renz
*Project Editor:* Dick Johnson
*Production Coordinator:* M. Y. Mim
*Illustration Coordinator:* Cheryl Nufer
*Line Art:* Tim Keenan
*Original Cartoons:* John Johnson
*Compositor:* Holmes Composition Service
*Printer and Binder:* The Maple-Vail Book Manufacturing Group.

**Library of Congress Cataloging in Publication Data**

Moore, David S.
Statistics: concepts and controversies.

Includes bibliographical references and index.
1. Statistics. I. Title.
QA276.12.M66 001.4'22 78–12740
ISBN 0–7167–1022–6
ISBN 0–7167–1021–8 pbk.

AMS (MOS) Subject Classification 6201; Statistics, elementary exposition of.



Printed in the United States of America

9 MP 1 0 8 9 8 7 6 5

# Statistics

## Concepts and Controversies

Does not He see my ways,
and number all my steps?
Job

But even the hairs of your head
are all numbered.
Jesus

Hell is inaccurate.
Charles Williams

# To the Teacher

In Japan, October 18 is National Statistics Day. The Japanese, with their usual thoroughness, have made public and official what is quietly recognized elsewhere: that statistics, no longer satisfied to be the assistant of researchers and government planners, now forces herself into the consciousness of students of every discipline and citizens of every occupation. This book is written for those students and citizens. There are texts on statistical theory and texts on statistical methods. This is neither. It is a book on statistical ideas and their relevance in public policy and in the human sciences from medicine to sociology.

I have developed this material during six years of teaching students (usually freshmen and sophomores) from Purdue's School of Humanities, Social Science, and Education. The students come from many disciplines and are fulfilling a dreaded mathematical sciences requirement. Future psychology and sociology majors often choose the course as preparation for later study of statistical methods; the other students will probably never again encounter statistics as a discipline. My intention is to make statistics accessible by teaching verbally rather than symbolically, and to bring statistics out of the technician's closet by discussing applications and issues of broad public concern. I have also used much of this material as supplementary reading in more traditional statistics courses, which usually neglect many of the concepts and issues discussed here.

So the book is popular, written for readers interested in ideas rather than technique. Yet this appellation requires several qualifications. First, readers will find genuine intellectual content, probably more than in a technique-laden methods course. I have even included several simple techniques (use of random digits for sampling and for simulation, computation of simple descriptive statistics and of index numbers) on the grounds that talking about a median without ever computing one may be empty. Second, I have been positive in my approach to statistics. I am annoyed that so many popular presentations suggest that statistics

is a subcategory of lying. Third, this is a text, organized for study and provided with abundant exercises at the end of each chapter. I hope that this organization will not deter those admirable individuals who seek pleasure and learning in uncompelled reading.

I am grateful to many colleagues for comments and suggestions. Professors William Erickson of the University of Michigan and Paul Speckman of the University of Oregon used the first draft in their classes and graciously provided both their own reactions and those of their students. They will forgive me if I proved at times hard to sway. The mathematicians and statisticians who have taught from the first draft noted the challenge of teaching nonmathematical material. I have tried to provide detailed guidance for teachers in the Instructor's Manual. Here, only a few suggestions are in order. First, try to establish a "humanities course" atmosphere, with much discussion in class. Many of the exercises involve discussion and can be modified to ask "Come prepared to discuss" rather than seek written answers. There are of course techniques to be learned, but classes should not be primarily problem-oriented. Second, use the collection of readings *Statistics: A Guide to the Unknown,* by J. M. Tanur et al. (eds.), as supplementary material. It is referred to often in the text, and complements this book well.

Statistics is a subject of growing importance to general audiences, and statisticians are increasingly aware of the need to introduce their subject to a wider public. There is as yet no consensus on how this should be done. It is my opinion that words remain as effective as computer terminals. I enjoy teaching this material as much as any. The orientation toward discussion brings students and teachers closer than in a more technical course. I hope that you also enjoy it.

# Introduction

Most of us associate "statistics" with the words of the play-by-play announcer at the end of the sports broadcast, "And we thank our statistician, Alan Roth. . . ." We meet the statistician as the person who compiles the batting averages or yards gained. Statisticians do indeed work with numerical facts (which we call *data*), but usually for more serious purposes. Statistics originated as *state*-istics, an accessory to governments wanting to know how many taxable farms or military-age men their realms contained. The systematic study of data has now infiltrated most areas of academic or practical endeavor. Here are some examples of statistical questions.

1. The Bureau of Labor Statistics reports that the unemployment rate last month was 6.5%. What exactly does that figure mean? How did the government obtain this information? (Neither you nor I were asked if we were employed last month.) How accurate is the unemployment rate given?
2. The Gallup Poll reports that 42% of the American public currently approve the President's performance in office. Where did that information come from? How accurate is it?
3. What kind of evidence links smoking to increased incidence of lung cancer and other health problems? You may have heard that much of this evidence is "statistical."
4. A medical researcher claims that vitamin C is not effective in reducing the incidence of colds and flu. How can an experiment be designed to prove or disprove this claim?
5. Do gun control laws reduce violent crime? Both proponents and opponents of stricter gun legislation offer numerical arguments in favor of their position. Which of these arguments are sense and which are nonsense?

*The aim of statistics is to provide insight by means of numbers.* In pursuit of this aim, statistics divides the study of data into three parts:

I. Collecting data
II. Describing and presenting data
III. Drawing conclusions from data

This book is organized into three parts following this same pattern. The second of these divisions is often called *descriptive statistics*; the third is often called *statistical inference*. I hasten to add that we will not leave the interesting business of drawing conclusions to the end of the book. Collecting and organizing data usually suggest conclusions (not always correct conclusions), and we will have much to say about these informal inferences in the first two parts of our study.

Your goals in reading this book should be threefold. First, reach an understanding of statistical ideas in themselves. The basic concepts and modes of reasoning of statistics are major intellectual accomplishments (almost all developed within this century) worthy of your attention. Second, acquire the ability to deal critically with numerical arguments. Many persons are unduly credulous when numerical arguments are used; they are impressed by the solid appearance of a few numbers and do not attempt to penetrate the substance of the argument. Others are unduly cynical; they think numbers are liars by nature and never trust them. Numerical arguments are like any others. Some are good, some are bad, and some are irrelevant. A bit of quantitative sophistication will enable you to hold your own against the number-slinger. Third, gain an understanding of the impact of statistical ideas on public policy and in your primary area of academic study. The list of statistical questions given above hints at the considerable impact of statistics in areas of public policy. I will only add that the impact is sometimes aimed at your pocketbook. For example, each 1% rise in the Consumer Price Index automatically triggers a billion dollar increase in government spending on such things as social security payments. You pay a share of that billion dollars, and you should know something about the Consumer Price Index and other creatures in the statistical zoo. The invasion of many academic areas by statistics is even more dramatic. For example, two political scientists recently compiled the percentage of articles appearing in a leading political science journal that made use of numerical data. Here are their results.[1]

| | | |
|---|---|---|
| 1946–1948 | 12% | of all articles used numerical data |
| 1950–1952 | 16% | " " " " " " |
| 1963–1965 | 40% | " " " " " " |
| 1968–1970 | 65% | " " " " " " |

It is clear that a political scientist must now be prepared to deal with statistics.

Economists, psychologists, sociologists, and educators have long considered statistics a basic part of their tool kit. Not even historians and literary scholars can ignore statistical methods. It is now common, for example, to attempt to decide the authorship of a disputed historical or literary document by analyzing quan-

titative characteristics of writing style (sentence length, vocabulary counts, frequency of certain grammatical constructions, etc.). Comparing these characteristics of the disputed document with documents by known authors often leads to a decision about authorship of the disputed document. An outstanding investigation of this type concerned the authorship of 12 of the papers originally published in *The Federalist.* These papers were published anonymously in 1787–1788 to persuade the citizens of New York State to ratify the Constitution. There is general agreement as to the authors of most of those papers: John Jay wrote 5, James Madison wrote 14, and Alexander Hamilton wrote 51. The disputed 12 may belong to either Madison or Hamilton. A statistical study of the style of these papers gave good reason to think that all were written by Madison.*

I hasten to state that one need not entirely approve of the infiltration of fields such as political science and history by quantitative methods. One might well agree with the remarks of Lewis A. Coser, president of the American Sociological Association in 1975, who warned the Association's annual meeting that "if concepts and theoretical notions are weak, no measurement, however precise, will advance an explanatory science."[2] In other words, it may be misleading to attempt to measure what you don't understand. But whether we like or dislike the increasing use of statistical arguments, we must be prepared to deal with them. Even if you wish only to rebut your local statistician, this book aims to give you the conceptual tools to do so.

## NOTES

1. James L. Hutter, "Statistics and Political Science," *Journal of the American Statistical Association,* Volume 67 (1972), p. 735.
2. Reported in *The New York Times,* August 30, 1975.

*See Frederick Mosteller and David L. Wallace, "Deciding Authorship," in J. M. Tanur et al. (eds.), *Statistics: A Guide to the Unknown* (San Francisco: Holden-Day, 1972). This book of readings contains many outstanding examples of the uses of statistics, and we shall refer to it often.

# Statistics

## Concepts and Controversies

# Contents

Part I

# Collecting Data

Before numbers can be used for good or evil, we must collect them. Of course we could make up data, a common enough practice. Leaving invention aside, many statistical studies are based on *available data,* that is, data not gathered specifically for the study at hand, but lying about in files or records kept for other reasons. Available data must be used with caution. Here is an example.

> The American Cancer Society, in a booklet called "The Hopeful Side of Cancer," claims that about one in three cancer patients is now cured, while in 1930 only one in five patients was cured. That's encouraging. But where does this encouraging estimate come from? From the state of Connecticut. Why Connecticut? Because it is the only state that kept records of cancer patients in 1930. It is a matter of available data. But Connecticut is not typical of the entire nation. It has no large cities, and few blacks. Cancer death rates are higher in large cities than in rural locations, and higher among blacks than among whites. We are left without clear knowledge of the national trend in cancer cures.*

Historians must rely on available data. The rest of us can make an effort to obtain data that bear directly on the questions we wish to ask. Such data are obtained by either *observation* or *experiment.* Observation is passive: The observer wishes to record data without interfering with the process being observed. Experimentation is active: The experimenter attempts to completely control the

*This example is taken from a *Newsday* dispatch that appeared in the *Lafayette Journal and Courier* of January 29, 1977.