GREGORY A. KIMBLE

# How to Use (and misuse) Statistics


PRENTICE-HALL, INC., *Englewood Cliffs, N.J. 07632*

*Library of Congress Cataloging in Publication Data*

KIMBLE, GREGORY A.
How to use (and misuse) statistics.

(A Spectrum Book)
Includes bibliographical references and index.
1. Statistics. I. Title.
HA29.K4875 519 5 77–27101
ISBN 0–13–436204–7
ISBN 0–13–436196–2 pbk.



Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

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*

# How to Use (and misuse) Statistics

**Gregory A. Kimble** is chairman of the Psychology Department at Duke University. He has published widely in the professional psychological literature and has contributed to the *Encyclopedia Americana* and the *Encyclopaedia Britannica*.

*To LLK and HP-65*

# Contents

# Preface

When I was an undergraduate psychology major at Carleton College, longer ago than most of you are old, I took the path of least resistance through the curriculum. As there always will be, there were loopholes in the requirements and I managed to locate most of them. This meant that I graduated almost untouched by hard science, mathematics, and statistics.

Later I was to suffer for my sloth. As a graduate student, they made me take remedial courses to repair my deficiencies. And that experience dramatically changed my outlook on quantitative materials. Taking those courses, particularly statistics, working with the subject matter, and eventually teaching a course in advanced statistics on my first job at Brown University showed me the importance of what I had rejected as an undergraduate. My opinion now is that a course in statistics can be the most liberalizing course a student can find in the curriculum and perhaps should be required of everyone. The world we live in is a world of uncertainty. An acquaintance with the way a statistician thinks is as useful as anything I know in the struggle that we all must carry on in order to cope with what we cannot predict.

This last statement may provide a hint that I will not be much concerned with calculations in this book. What I want you to come away with is an appreciation of a style of thought and a respectable level of statistical literacy. I see no necessity, with these as my objectives, to dwell on formulas and computations. For those of you who find security in arithmetic, a final section of the book presents some of the technical tools. But what I want you to take away from your reading does not require mastery of that section.

I have tried very hard to keep the materials in this book lively and interesting, but I hope that I have not been *too* successful. By that I mean that it would be unfortunate if my routine flipness and irreverence and my occasional brushes with obscenity were to distract you and make you miss the serious points I have to offer. For I am really very serious about these materials. Understood, they will enrich your lives just as they have mine. So please: Read for understanding. If sometimes understanding is also fun, so much the better.

GREGORY A. KIMBLE

*I wish to express my thanks to everyone*
*who helped me with this book, especially Norma Karlin,*
*Maxine Kluck, Tamara Easton, and Dan Thomann.*

# The Nature of Statistics

What is *statistics*? Ask an ordinary person and you are apt to get a response that reflects Disraeli's opinion that "there are three kinds of lies: lies, damned lies and statistics." Ask a statistician and the answer will be different: Statistics is a branch of science dealing with the collection, analysis, and interpretation of data. As the title of this book tells you, both answers suggest topics worth writing about. Although the popular notion that you can "prove anything with statistics" is far too extreme, it is possible to misuse statistics in order to mislead people. But statistics can also aid your understanding of many of the problems that face the world today. The pages to follow will present examples of both of these uses of statistics. It is useful to think of statistics as serving two general functions, a descriptive one and an interpretive one. This chapter presents an overview of both functions.

## DESCRIPTIVE STATISTICS

*Descriptive statistics* are exactly what their name implies, numbers that describe some situation of interest. Batting averages, rates of unemployment, rates of mental illness, sales of automobiles in November, number of children in the average family, average income, frequencies of sexual activity, and the well-known tables of heights, weights, and lengths of life are all examples.

The value of descriptive statistics is that they give an efficient summary of some type of information. A professional baseball player may come to bat 500 or even 600 times during a season. A list of the player's performance on every such occasion would be so long and complicated as to make it impossible to obtain a very clear overall impression of the record. In 1976 Carew of the Minnesota Twins came to bat 605 times and got 200 hits. A complete record of all Carew's at-bats, together with an indication of what happened on each occasion, would take pages. The important information can be conveyed briefly, however, by dividing 200 by 605 and reporting that Carew's batting average was .331.

Batting averages are straightforward and uncomplicated because the operations involved in their computation are completely clear. We know what it means to be at bat and to get a hit. We also know what it means to divide one of these terms by the other. Sometimes, however, things are not so clear.

### Sizes of Cities and How They Grow

When you look up a city's population in *The World Almanac*, you will find a comfortably definite number. For Boulder, Colorado, where I live, the number is 66,870. But what exactly does this mean? For example, does it include students at the university? Probably not, unless they have established residence in Boulder. How many additional people does this add to the population? There is no real way to answer the question because a student at the university turns out to be a statistical abstraction. The university operates under a legislated limit of 20,000 students per year. But this does not mean 20,000 human beings; it means 300,000 student credit hours. For purposes of limiting the size of the university, a "student" is 15 credit hours of work. Since the average student takes fewer than 15 credit hours of courses, the number of people is some indefinite number greater than 20,000. Some unknown fraction of them (those who are not Boulder residents) should be added to 66,870 to produce a more accurate figure for the population of the city.

The estimate of Boulder's population is complicated further by the fact that whether a person lives in Boulder is an arbitrary matter, because city limits are set arbitrarily. In the case of our city this fact is important, because a very large number of people who are Boulder residents for most purposes live beyond the city limits. Various ways of estimating the number of such people suggest that there are 20,000–25,000 of them. These figures, together with a guess that perhaps 15,000 students should be added to the population, indicate that the true population of Boulder is probably closer to 100,000 than to the almanac figure of about 67,000.

Similar ambiguities exist in reports of the rates of growth for cities. To make this point, I shall use some statistics for Raleigh and Durham, North Carolina, where I lived before I moved to Colorado. Populations of the two cities for 1960 and 1970, along with some statistics that can be derived from them, appear in Table 1–1.

Obviously, Raleigh is bigger than Durham. It also appears that Raleigh is growing faster. But again there is more to the situation than meets the eye. Both cities are university cities and exact population figures depend upon how students are counted. Once more there is also a problem relating to the setting of city limits, which accounts for some of the apparent difference in rates of growth. If my memory is correct, Raleigh extended its

**TABLE 1–1**
**Populations of Raleigh and Durham, North Carolina, 1960 and 1970**

| City | 1960 Population | 1970 Population | Increase in Population | Percentage Increase in Population |
|---|---|---|---|---|
| Durham | 78,302 | 95,438 | 17,136 | 22 |
| Raleigh | 93,931 | 123,793 | 29,862 | 32 |

limits sometime in the 1960s but Durham did not. If Raleigh's expansion brought an additional 9,218 people into the city, this would account for the difference between a 22% and a 32% increase in the sizes of the two communities.

It may be important to point out that all of this is not just an idle exercise designed to get across statistical ideas. Such information has great practical significance. A merchant considering Boulder as a place to open a branch store might make quite different decisions depending upon whether he believes that Boulder's population is 67,000 or 100,000. Recognizing such concerns, the U.S. Office of Management and Budget (OMB) has defined Standard Metropolitan Statistical Areas (SMSAs) on the basis of census data and reports populations for 266 urban areas. From time to time it also combines areas when they seem to represent a single economic unit. The results can be spectacular. In 1970 the Raleigh SMSA had a population of 228,453 and a rank in the nation of 135. By 1975 the OMB had combined regions and created a Raleigh/Durham SMSA. It has a population of 462,300 and is 78th in the nation.

### Unemployment

For a good many years, the following question has been high in the minds of everyone: "Are we in a recession or not?" Whether we are or not depends upon the definition of a recession. Several economic indices, including production in heavy industry, the sizes of inventories in retail stores, and rate of unemployment provide possible bases for such a definition. The definition in terms of unemployment seems particularly relevant to human welfare. As a result, most of us pay some attention to the statistics on unemployment and take these figures as an important index of the economic health of the nation, possibly saying to ourselves that if the unemployment rate exceeds 9%, the nation's economy is in a bad way—we are in a recession. To take this position is to define a time of recession as a time when more than 9% of the population is out of work.

Thus the answer to the question, "Are we in a recession?" depends upon the answer to the more specific question, "Is the rate of unemployment more than 9%?"

In November 1975 official U.S. government figures indicated that the jobless rate was 8.3%. Albert Sindlinger, a market research analyst, criticized this figure, claiming that it should have been over 9.2%, perhaps as high as 10.6%. In short, by the definition of a recession as 9% unemployment, we were in one according to Sindlinger but not according to the official figures. How could such a disagreement come about?

To answer this question, it is important to understand how the percentage of unemployment is determined. The calculation is almost the same as the computation of a batting average. It is the number of people without jobs divided by the total number of people in the labor force. In November 1975, the official computation was: 7,717,257 unemployed divided by 92,979,000 in the labor force equals .083, or 8.3%.

The difference between this calculation and that of a batting average is that both of the basic terms in the calculation are a little vague. Of the two, the concept of labor force is the more definite, and Sindlinger agreed with the official figure in the dispute mentioned above. The labor force is everyone over 16 who might be expected to work. During the school year it does not include students unless they hold a job, but it does include them during summer vacations. Thus the size of the labor force increases by a few million each June and a "seasonal adjustment" enters the calculation of the rate of unemployment. A second major group omitted from the labor force are housewives. They enter the force only if they start looking for a job. Whatever one makes of this treatment of certain groups, the definition of labor force is precise enough to create no great problem.

The definition of an unemployed person is much more difficult. One definition might be everyone who is looking for a job. This definition would then include everyone who has a job but is trying to find a better one. Defined that way, the unemployment rate may typically be as high as 11 or 12%. Or one could go to the other extreme and say that a person is not really unemployed unless he has been out of work for a while—15 weeks being a period for which 1975 data are available. Defined in those terms, the percentage of unemployed was only about 3.3%.

Obviously, the jobless rate is a more elusive figure than one might have thought. Looking further into the definition of unemployment only complicates matters. Does part-time work count as employment? If so, is an hour's work a week on a college work–study program enough to qualify a person as employed, or must it be more? How about people on the Public Service Employment Program who have jobs created by the government specifically for the unemployed? Should they be counted or not? The list

could go on. I mention these two items because they appear to have been responsible for the dispute between Sindlinger and the Feds. If the official count of the number of employed people included 700,000 work–study students and 315,000 Public Service Employees (reasonable estimates) in November, and if they should not have been counted as employed, the rate of unemployment was more like 9.4% than the reported 8.3%. It is the elusiveness of such figures that allows them so frequently to come up for argument in politics, for example in the Carter–Ford debates of 1976.

### Rates of Mental Illness

Those who believe that the world is going to hell in a hand-basket will tell you that this unhappy trend includes an increase in the rate of mental illness. The greater complexity of life and the resulting greater stress are to blame, according to this view. Deciding the truth of this assertion poses very special problems. For one thing, in order to detect the alleged change it would be desirable to make a comparison between rates of mental illness for widely separated points in time—ideally 75 or 100 years. Such a comparison might, for example, look at admission rates (proportions of the population admitted to mental hospitals) as revealed by older hospital records and newer ones. Such studies have been done and superficially the data suggest that a trend toward increased frequency of mental illness actually exists.

There are several things wrong with these[1] data, however:

1. Mental illness is diagnosed more frequently now than it was 100 years ago. Complaints now seen as mental disorders once were not. This means that the comparison must be for such disturbances as schizophrenia, which have been identified by the same symptoms for about as long as psychiatric diagnosis has been around.
2. Whatever the disorder, the increased availability of mental hospitals means that more people are undergoing treatment now.
3. The population now contains a greater proportion of old people and they have a higher incidence of mental disorder.

Correction for these features of the data reveals that the incidence of mental disorder is probably no different now from what it was in the middle of the last century.[2]

[1]Along with Edwin B. Newman and others, I am fighting what appears at the moment to be a losing battle against a variety of grammatical atrocities. *Data* is a plural word. It refers to more than one item of information. The singular is *datum*. Thus you should speak of "these data" and "this datum."

[2]W. A. Wallis and H. V. Roberts, *The Nature of Statistics* (New York: Free Press, 1962).

### Crime Rates

Increases and decreases in crime rates are also harder to prove than you might think, for several reasons:

1. One basis for determining the number of crimes is the number reported, but if people change in their willingness to report crimes, a change in rate could be detected where none existed.
2. Another basis for defining crime rate is the number of individuals caught and booked for criminal acts. If the police forces improve and more criminals are apprehended, the crime rate may erroneously appear to increase on this basis.
3. In a similar way, better systems of keeping criminal records will produce an increased crime rate that is apparent but not real.
4. The definition of "crime" will influence the rate at which crimes occur. In some times and in some places, homosexuality and the possession of small amounts of marijuana are major crimes, felonies. If the laws change and these infractions become misdemeanors rather than felonies, the felony rate will decline.

## OPERATIONISM AND SOME RELATED IDEAS

If you are willing to put up with a brief discussion that is slightly philosophical, the materials just presented can be made to illustrate some very important points about straight thinking on many topics. My plan for this section is to develop the essential ideas in a position called *operationism*. I hope to convince you that the straight thinking just referred to is operational thinking. In these antiintellectual times, such thinking has been subjected to a great deal of bad mouthing. If I am successful in my effort, you will come away from your reading of this section inoculated against the effects of such mindless criticism.

Operationism is a general view of science which holds that scientific statements are meaningful only if they lead to observations that can be made on the real world. Otherwise, these statements are meaningless. A key concept in the operationistic position is that of operational definitions, which are definitions of concepts in terms of physical procedures.

In the most general sense a definition is a statement that gives the meaning of a word or group of words. The last sentence is an example. It is a definition of the word *definition*. Definitions of this type pose a small problem. They make no sense if you happen not to understand the terms employed, for example "gives the meaning of." To make this point more realistically, suppose I tell you (correctly) that "a $Z$ score is the deviation