APPLICATIONS, USES, AND ABUSES OF STATISTICS

Edited by
Wilbert Marcellus Leonard II
Professor of Sociology

FOREWORD

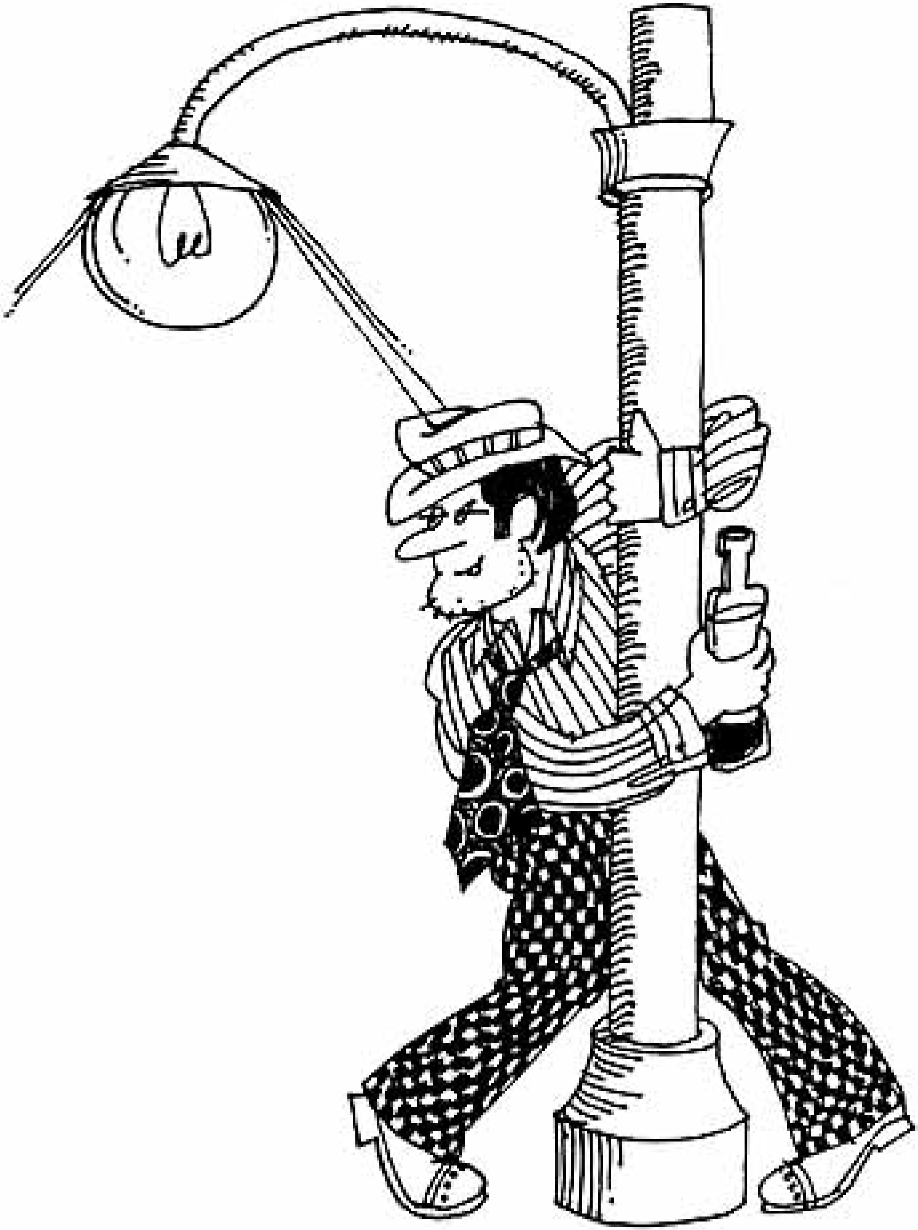
The purpose of this Reader is to supplement and complement the standard statistics text. Having taught statistics for two and one-half decades, I have been somewhat disquieted by the amount of time necessary to convey the principles of computation (even with the assistance of computer technology) and interpretation of statistical formulae at the expense of practical application and usage. To fill this void, I have edited a wide range of statistical materials and selected those that would facilitate students' appreciation of the actual use, and sometimes abuse, of quantitative tools. My hope is that these Vignettes stimulate the "statistical imagination" and cultivate "mathematical literacy."

I have organized the "Readings" into three major sections. First, materials are arranged by their use in various fields, e.g., journalism/social policy, business, health, politics, biology and energy. These essays lucidly illustrate some contemporary uses of statistical analysis in several fields. Second, materials are organized by the subtopic within statistics that they most clearly fall into, e.g., Frequency Distributions and Graphic Representation, Measures of Central Tendency, Measures of Dispersion, the Normal Curve and z-Scores, and Measures of Association and Regression Analysis. These topics fall under the rubric of Descriptive Statistics. Examples of Hypothesis Testing and Parameter Estimation are subcomponents of Inferential Statistics and are organized as such. Third, several miscellaneous statistics selections are also contained herein.

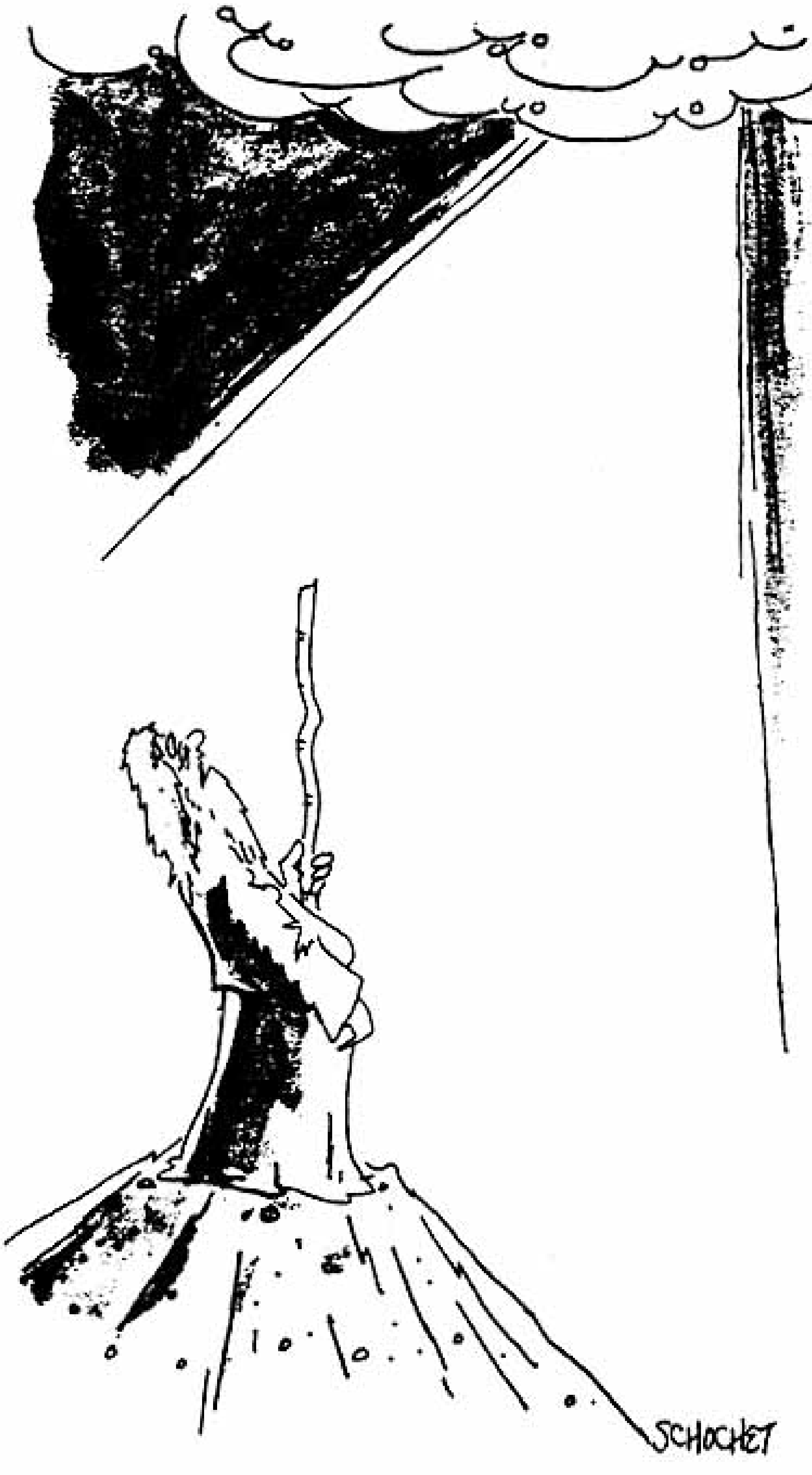
This supplement was supported by the Faculty and Instructional Development Program at Illinois State University and the first edition was completed during the summer of 1981. The second edition was produced during the summer of 1983 and the third edition during the fall of 1987. I would like to express my thanks to the authors of the original writings. The essays have been adapted or used in total for inclusion in this booklet. My appreciation to the staff in the Department of Sociology - Pat Walsh, Mickey McCombs, Christine Gruber - and Penny Williams for their assistance on this, and other, projects.

Wilbert Marcellus Leonard, II
Professor of Sociology
Winter 1996

Table of Contents



It has been said that many people use statistics in much the same way that an inebriate uses a lamp pole: more for support than for illumination.



"Seventy-three percent are in favor of one through five, forty-one percent find six unfair, thirteen percent are opposed to seven, sixty-two percent applauded eight, thirty-seven percent . . ."

IF NUMBERS MAKE YOU NUMB...

...You're not alone. You can count yourself among a multiplying percentage of people who find themselves taxed by the array of figures upon which the smooth functioning of this increasingly complex society is calculated.

Innumeracy, the inability to understand and deal with numbers, is the mathematical equivalent of illiteracy, the inability to read. It afflicts consumers who can't balance their checkbooks or who don't understand credit-card interest, doctors who unwittingly misinform patients about the risks of operations, gamblers who wager money on near-impossible odds. Also included are investors duped by stock-market scams, and fast-food franchises that, for the sake of undereducated, underpaid employees, must place pictures of menu items on cash registers that automatically add appropriate taxes.

What would one megaton of TNT do? (The world's nuclear-weapons stockpiles contain 25,000 megatons of explosive power for every man, woman and child on earth.) What does it mean when a country spends a million dollars, or a billion? (It takes about 11½ days for a million seconds to pass—almost 32 years for a billion seconds to tick off.)

What are your odds of being killed by terrorists? (In a recent year, 17 Americans were killed by terrorists, out of 28 million who traveled abroad. That's one in 1.6 million.)

"I'm distressed by a society which depends so completely on mathematics and science and yet seems so indifferent to the innumeracy and scientific illiteracy of so many of its citizens," writes mathematician John Allen Paulos (*Innumeracy: Mathematical Illiteracy and Its Consequences*, New York: Hill and Wang, 1988, page 134).

Those citizens, says Dr. Paulos, "seem to have no mathematical frame of reference and no basic understandings on which to build. They're afraid. They've been intimidated by officious and sometimes sexist teachers and others who may themselves suffer from math anxiety" (page 88).

Society's failure to effectively educate people about numbers contributes to disparities in employment, among other problems. For example, women often are steered away from heavy mathematics and science curriculums. This denies them better, more lucrative jobs, though they're just as capable as men of understanding math.

"Women, in particular, may end up in lower-paying fields because they do everything in their power to avoid a chemistry or an economics course with mathematics or statistics prerequisites," writes Dr. Paulos. "I've seen too many bright women go into sociology and too many dull men go into business, the only difference between them being that the men managed to scrape through a couple of college math courses" (pages 78-79).

What dosage of medicine should you take? Which investment will produce a better return for you? What, exactly, is a calorie and how many of them do you need? When the media scream about "epidemics" such as AIDS, how scared should you get?

Everyone of us needs to be able to use numbers to reason inductively, to analyze, to make proper judgments, to estimate. Acquiring basic numeracy means overcoming the fear of figures.

Norman L. Shoaf

what do statisticians do?

For many people the word "statistics" brings visions of zig-zag graphs, batting averages, passes completed—all the tabulations on the sports pages or columns of figures in the business section of newspapers. This is only one kind of statistics; the statistician today does far more than construct and examine graphs and tables. He or she must develop a greater understanding of the origins of data, their possible meanings, and especially their accuracy. Knowing that Council Bluffs, Iowa had 60,348 inhabitants in 1970 is not enough. How was the count made? Did it include students temporarily away at school? Or those students temporarily living there during the school year? How about the newborn babies in the hospital? Did the enumerators do their job well? Important decisions, such as who gets what amount of Federal funds in revenue sharing, depend upon the answers to these questions.

U.S. Government statisticians conduct Censuses of Population, Housing, Manufactures, and Agriculture. Compilations are made of sales, production, inventories, payrolls, and other internal industrial and business data. These statistics tell the alert business manager how his industry is growing, how his company is growing in relation to the industry, and what plans he should make for future expansion. Government officials could not function without this information; planning would come to a standstill.

The statistician also helps the manager make wise decisions. When a food chain, for example, considers opening a new branch of a supermarket, it turns to the statistician to help find the best location. Sample surveys help determine the prospects for success, by measuring such items as population concentrations, income levels of potential customers, the availability of transportation, the needs of the community and existing competition. The expert interpretation of these data puts the statistician right at the heart of the final decision.

Central to the work of many statisticians is the use of computers. Statistical calculations that once took weeks of labor can now be done on a high-speed electronic computer in a few seconds. Some statisticians use the computer to analyze data. Others use it to help solve thorny statistical problems whose mathematical complexity might otherwise be overwhelming. Almost all agree that what the test tube is to the chemist, the modern computer is to the statistician.

Statistics is a changing field with new methods being generated constantly. The user of statistical techniques regularly has more and better tools at hand to help solve his problems.

There seems to be no limit to the areas of challenge to the statistician. Let's look at some specific examples of statistical problems, in widely ranging fields.

<u>Careers in Statistics</u>, American Statistical Association, p. 5.

SIR RONALD A. FISHER

The ideas and methods that we study as "statistics" were mostly invented in the nineteenth and twentieth centuries by people working on problems that required analysis of data. Astronomy, biology, social science, and even surveying can claim a role in the birth of statistics. But if anyone can claim to be "the father of statistics," that honor belongs to Sir Ronald A. Fisher (1890–1962).

Fisher's writings helped organize statistics as a distinct field of study whose methods apply to practical problems across many disciplines. He systematized the mathematical theory of statistics and invented many new techniques. But the randomized comparative experiment is perhaps Fisher's greatest contribution.

Like many statistical pioneers, Fisher was driven by the demands of practical problems. Beginning in 1919, he worked on agricultural field experiments at Rothamsted in England. How should we arrange the planting of different crop varieties or the application of different fertilizers to get a fair comparison among them? Because fertility and other variables change as we move across a field, experimenters used elaborate checkerboard planting arrangements to obtain fair comparisons. Fisher had a better idea: "arrange the plots deliberately at random."

This chapter explores statistical designs for producing data to answer specific questions like "Which crop variety has the highest mean yield?" Fisher's innovation, the deliberate use of chance in producing data, is the central theme of the chapter and one of the most important ideas in statistics.

Garbage In, Garbage Out

Several years ago *Time* magazine took a hard look at the American fascination with statistics -- and the doubtful quality of some of those statistics. In an editorial essay entitled "The Science and Snares of Statistics," *Time* suggested:

> This dedication to numbers has created its own pitfalls for the innocent -- and opportunities for the purveyors. There is an air of certainty about the decimal point or the fractionalized percentage -- even in areas where the measurement is statistically absurd or the data basically unknowable. A classic example is a survey made some years ago, which solemnly reported that 33-1/3% of all the coeds at Johns Hopkins University had married faculty members. True enough, Johns Hopkins had only three women students at the time, and one of them married a faculty member. The American Medical Association announces not that very few people dream in color, but that "only 5% of Americans" dream in color. New York City has 8,000,000 rats. How does anybody know? Statisticians have a phrase for this, borrowed from the computer industry on which they now rely. The phrase is "garbage in, garbage out" — meaning that the result that comes out is only as good as the material that is fed in.
>
> For the sake of drama or publicity, numbers are slapped on nearly everything -- and the bigger the number the better . . . . Newsmen during the Detroit race riots pressed a harried fire chief for damage estimates. His guess: $500 million. So far, in the cooling aftermath of riot, insurance companies are processing only $84 million worth of damage claims, and the overall loss is now put at $144 million. For newsmen, the National Safety Council issues forecasts of expected highway deaths over holiday weekends usually with a prediction tacked on of "record fatalities". What the forecast never says is that the record is due to population increases and wider use of automobiles, and that the fatality rate is usually just as high proportionately on other weekends -- holiday weekends are just a bit longer . . . .
>
> Since it is obviously impractical to poll the nation on anything less important than the selection of a President, one cherished statistical tool is the sample. Not even statisticians can agree on how big or good a sample can be relied upon as representing the whole. Dr. Alfred C. Kinsey's celebrated reports were criticized by statisticians not so much for their moral implications but because they made sweeping presumptions on the basis of too small a sample (in the male study, only 5,300 men provided data). The Nielsen ratings, by which television programs live or die, have been justly attacked because Nielsen recorders are necessarily hooked to the sets of those viewers willing to have a recorder -- a special class by definition, whose tastes may or may not correspond with those of the unpolled millions of the total TV audience.
>
> Still the state of the art of statistics has come a long way since 1661, when its founding father, London Haberdasher John Graunt, began a careful count and found that more boy babies died in infancy than girls, and concluded that therefore, there must be more women than men in Britain. Today's scientists, who no longer believe that anything is absolutely certain, also believe that many things are predictably probable. And it is the computer, fed with vast amounts

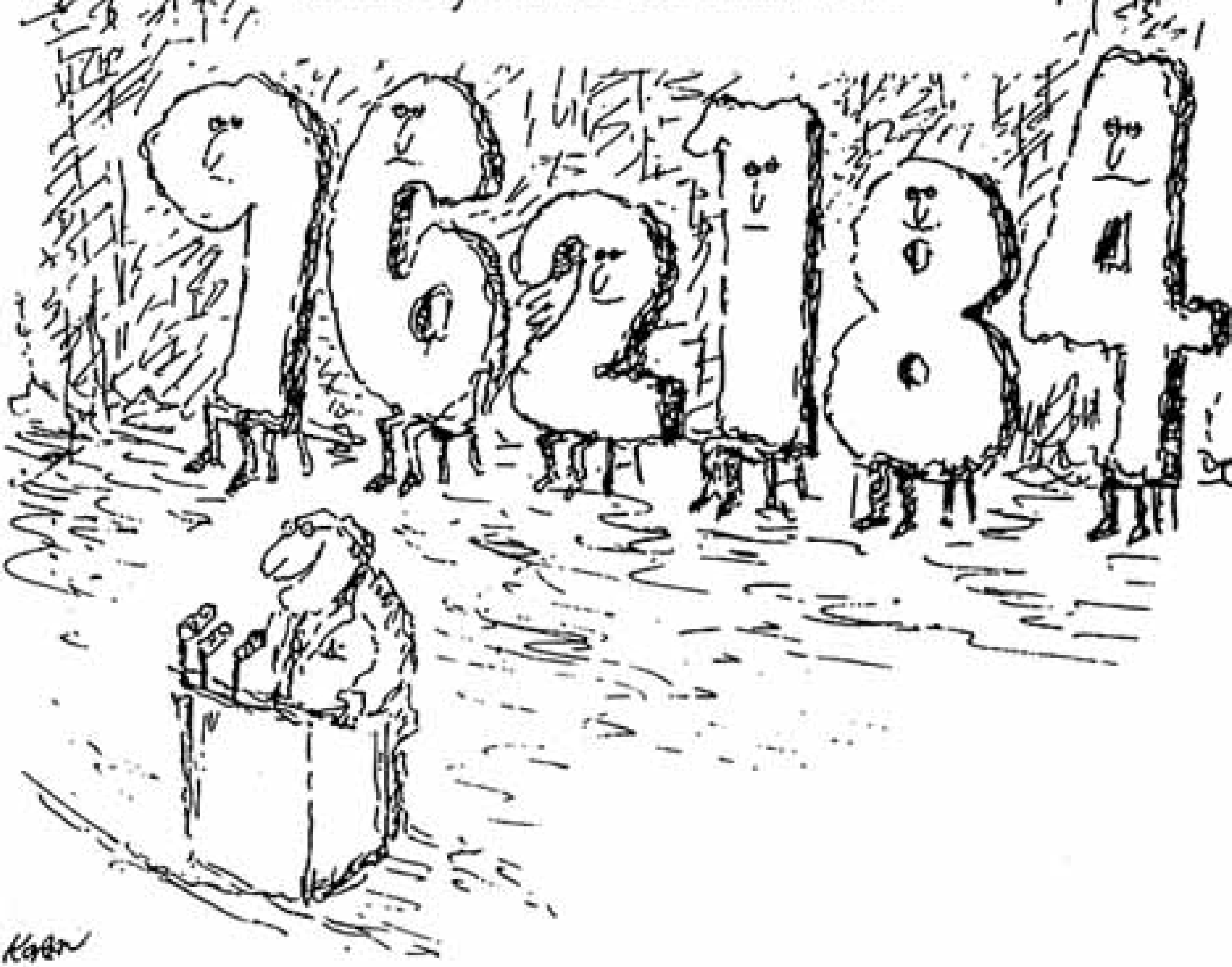
of past data, that can project or at least outline the alternatives of several possible futures. "The computer has enshrined statistics," says M.I.T.'s Professor Harold A. Freeman. "Without it, statistics would still be a grubby business." Where once all they had to do was count, and perhaps draw graphs, statisticians are now "programmers," with a mystique all their own. Unquestionable, for the moment, numbers are king. But perhaps the time has come for society to be less numerically conscious and therefore less willing to be ruled by statistics.[1]

Since it doesn't seem likely that society will become "less numerically conscious," at least in the near future, a better suggestion might be that we try to become more statistically sophisticated, more familiar with proper usage of statistics and more aware of their limitations. We might move in this direction by including statistics in the standard high school -- or even elementary school — curriculum. Maybe the time has come for Reading, 'Riting, and Regression analysis.

[1]Condensed from "The Science and Snares of Statistics," Time (September 8, 1967): 29.

Understanding and Using Statistics-Basic Concepts, pp. 37-38, Marty J. Schmidt (Second edition), 1979.

The Uses of Statistics in Selected Fields



"Tonight, we're going to let the statistics speak for themselves."

From the Literature

Sociology is a rather new discipline. In fact, so new that many modern-day sociologists regard Emile Durkheim (1858–1917) as one of its founders. Durkheim set the tone for sociology as a research-oriented science of society that uses various statistical techniques for obtaining, summarizing, and interpreting data in order to acquire knowledge.

Emile Durkheim is especially noted for several major works, including *Le Suicide*, which was written in French and published in 1897. It was translated into English in 1951.

Durkheim's *Suicide* is clearly a classic. Many of the ideas contained in it are presented in almost every basic sociology text. We often remember Durkheim for his typology of suicides: altruistic, anomic, and egoistic. He should also be remembered for publishing the first empirical study in which a sociologist used a variety of simple yet appropriate statistical tools (such as rates, percentages, and proportions) to summarize and interpret data.

Among the data described and summarized in *Suicide* are the following:

suicides per 100,000 inhabitants

deaths per 100,000 inhabitants

percentage differences in suicide rates between two consecutive years

suicide rate per million inhabitants in 11 European countries

number of insane persons per 1000 inhabitants of each religious faith by country

Based on these and other research findings, Durkheim highlighted the relationship between group conditions and suicide. Specifically, he showed that individuals are integrated into social groups and that the risk of suicide depends, in part, on the extent of that integration. In addition, he noted that suicide depends on the quality of group life, which can range from stable and supportive to unstable and nonsupportive. In this context his data suggest that suicide becomes a viable option for persons weakly integrated into the group. It also becomes a viable option among highly integrated individuals when the group itself supports self-destruction. Finally, it is a viable option when individuals are caught up in a disorganized environment, an environment that does not provide clear-cut guidelines for meaningful actions.

Lyman Ott, Richard F. Larson, & William Mendenhall, <u>Statistics</u> (Boston: Duxbury Press), 1987: 20.

When we laugh at monkeys in the zoo, we may be laughing at ourselves. Statistics have shown a correlation between animal and human behavior and between animal behavior and natural phenomena.

Because Man is a primate, scientists are carefully observing monkeys, chimpanzees, gorillas and other primates to learn more about human behavior. Jane Goodall, famous for studying chimpanzees in the African wilds, gathers data on such chimp behavior as lone play, group play, mating, grooming, restlessness, aggressiveness, and closeness to others.

At the Stanford Outdoor Primate Facility (SOPF) at Stanford University, California, the nature, frequency, and duration of these behaviors have been sampled on a statistically random basis and the data fed into a computer. Correlational statistics are being compiled on development behavior and hormonal changes during puberty. Scientists want to learn if social behavior in adolescent chimps is related to body biology.

Because they had the data available, Stanford researchers were able to study another interesting correlation. Reports from China after the catastrophic earthquake on February 4, 1975, indicated that striking changes in animal behavior had been observed before the quake. The researchers decided to go back to look at their data to compare chimp behavior at SOPF before California quakes.

The researchers correlated dates of shocks registered on nearby U.S. Geological Survey equipment with chimp behavior. They discovered that the chimps had been significantly more restless and group oriented (statistically speaking) for 12 hours preceding seismic shocks in June and July of 1975.

Since the behavior data were collected and stored in the computer before the previously unpublished seismic activity was made available, it is assumed no observer bias colored the data. The researchers, however, caution that such a small animal group (15) and limited events (2) make too small a sample for generalizing from their statistical analysis. Nonetheless, speculation on the possibilities of using animals to predict earthquakes is fun. And certainly this study demonstrates the value of quantitative recording of even commonplace events.

Source: Ukena, Ann Seymour, Statistics Today, Harper & Row, Publishers, 1978, p. xiii-xix.

STATISTICS TODAY IN BUSINESS

When you are in business, you are always trying to outguess your customers. You have to. Unless you can judge ahead of time how many of any item customers will buy, you cannot know how many to order if you are a retailer, or how many to manufacture if you are a manufacturer. It's a tough problem.

Statistics help a lot. People in business gather statistical data to learn customer preferences and attitudes. This is called market research and helps companies to judge what turns consumers on and off.

The Coco-Cola Co., for instance, confirmed through market research that the famous "Coke" bottle does indeed induce customers to "reach for a Coke." That contour shape is known throughout the world. But Coca-Cola executives were not sure the bottle shape would play any significant role in sales of a new 16-ounce size the company planned to introduce several years ago.

A straight-walled bottle instead of the traditional contour one would save glass and speed the production line when the bottles were being filled. But executives want to be sure no sales would be lost.

So, a market test was run to check the hypothesis that potential customers identify with the "Coke" bottle shape sufficiently strongly to be lured into purchase more often.

Researchers interviewed about 800 people in shopping centers in some five cities. Shoppers were shown one bottle or another, then asked questions on intention to buy "Coca-Cola" that day. Later, answers were cross-tabulated and checked with known buying patterns from previous years at each shopping center.

When the data were analyzed, the contour bottle indeed proved to have extra selling power in the 16-ounce size. The traditional shape was retained.

Market research takes many forms. Sometimes businesses want to evaluate the potential effects of promotional programs before spending large sums of money. Then computer studies may be used. The computer compares past history with expected changes in customer habits. From these projected results, marketing executives decide whether the contemplated campaign is likely to produce statistically significant improvements in company sales.

You can be sure that whatever you intend to merchandise as a business person, you will look at statistics before making a major decision.

Source: Ukena, Ann Seymour. Statistics Today, Harper & Row, Publishers, 1978, pp. 350-351.

Energy. Where is it going to come from, and how are we going to use it in the future? Researchers are busy gathering statistics trying to forecast answers to both those questions.

Many persons suppose the sun is likely to be a leading energy source by the year 2000, while others believe wind will provide us considerable power. Researchers, however, predict no more than 8 percent of the United State's total energy needs are apt to be supplied by solar power by the year 2000. Wind is rated even lower.

These figures come from research being conducted on solar and wind energy as potential future power sources in Arizona, California, and New York, among other states. Similar research is going on for other energy forms: nuclear, geothermal, hydro(water), oil, gas, and coal.

Accurate forecasting depends on reliable statistics. Utility companies use statistical data to forecast what our main energy sources will be so that they can decide now how to invest in equipment for future power production. Manufacturers of autos, appliances, steel, and other products also must know what fuels will be available. And the general public needs to know in order to make intelligent policy decisions about environment that may change with varying fuel choices.

With statistics we can forecast our future energy habits now so we can tabulate statistically our future needs against possible energy sources. Determining need is a complicated process. Social customs, technology, climate, economics, and environment must all be considered.

Researchers must find answers to questions like: How many people are in a household? What new appliances have caught consumer fancy?

Will technical breakthroughs make solar power more widespread than predicted? How often are furnaces or air conditioners turned on? What's the price of energy? What pollutants might various energy sources create?

Researchers reduce all these environmental, social, and economic variables to statistical data. These data are converted to computer models that play-off statistics, one against another, in changing patterns to simulate actual behavior. Sometimes these computer-projected patterns are checked out in live situations.

For example, 8-hour workdays determine when most residential and industrial customers make peak demand on utilities. Were utilities to price power based on time of day--cheaper when demand is low, more expensive when demand is high--would consumers change their habits? Computer models say maybe. A utility company in Connecticut is conducting real-life experiments to find out.

Interpreting the statistical data from the experiments being conducted now will help shape how we use energy in the future.

Source: Okeun, Ann Seymour, Statistics Today, Harper & Row, Publishers, 1978, pp. 68-69.

Statistics help health professionals provide quality care to patients everyday. As nurses make their hospital rounds, they are constantly on the lookout for patients' vital signs: temperature, blood pressure, respiratory and pulse rates. They have to understand the standard deviation for each of these vital signs so they can take action if they spot a statistically significant variation.

Health professionals who work in special-care areas, such as the cardiac unit, rely on statistics even more. In these units, patients who have had heart attacks are hooked to monitoring machines that record heart rhythms, pressure within heart chambers, blood pressure, and other important body functions, which give clues to the likelihood of repeat heart attacks. Nurses, medical technicians, and physicians keep tabs on the machines that display precise information. These health professionals recognize when statistically significant changes in information from any body function is displayed.

Nurses, for example, who observed abnormal heart rhythms might administer medicine or help the patient relax, as measures to restore normal rhythms. Arrhythmias, as abnormal heart rhythms are called, are the number one cause of death after the most common type of heart attack, myocardial infarction. In the 15 years since arrhythmia monitoring was first introduced, nurses have reduced arrhythmia deaths within hospitals by about 20 percent.

The same statistical approach to caring for all patients is taking place in many hospitals around the country, though monitoring machines are not used. In a new auditing technique, diseases are reduced to a standard list of nursing-care criteria. These criteria are derived from statistical analysis of patient charts. Stroke patients, as one example, might require help in bathing, turning, eating, physical therapy; they might have to be watched carefully for speech difficulties, pupil changes, reduced bodily movements.

Once the criteria are established, a minimum desirable compliance percent is set; perhaps 98 percent for one disease, 90 percent for another. A compliance record compiled mainly from nurses' notes is audited regularly. Nursing quality control committees meet regularly, too, to analyze compliance deficiencies. The committees then make recommendations on what needs to be done to bring the quality of care up to the statistically desired rate.

Thus valid statistical procedures provide objective, consistent decisions in situations for which the chief guideline formerly was intuition. Health professionals now combine quantitative data with qualitative concern to bring top care to sick people.

Source: Ukena, Ann Seymour, Statistics Today, Harper & Row, Publishers, 1978, pp. 286-287.

Journalists would be surprised to be told they are walking statisticians since most hate math. They have, however, been trained to gather observations, which they process and reduce to readable articles.

Observations analyzed and reduced to quantitative terms are called data. Data organized in summary forms are called statistics. Thus, in a sense, a journalist is a statistician who summarizes data in a way the general public can understand.

Journalists also consciously use statistics in developing meaningful stories. They interpret masses of public data gathered everyday by government and other organizations. From this data they write exposes of illegal campaign contributions, for example, or why suicide rates are climbing among suburban youth.

Typical of modern journalism was an important statistical analysis made several years ago by two reporters on the Philadelphia Inquirer. They decided to scientifically measure the administration of justice in the local court system.

The reporters were interested in answers to such questions as: Are blacks and whites treated alike in the judicial system? How similar are sentences for the same crime among judges in Philadelphia? Do certain types of crimes result in harsher sentences even though legally they are of the same seriousness as other crimes?

The reporters studied 1034 persons indicted during 1971 for one of four major crimes: murder, rape, aggravated robbery, and aggravated assault and battery. These 1034 defendants represented a statistica--, random 39 percent sample of all persons indicted for those crimes during 1971.

More than 10,000 court documents and some 20,000 pages of notes of testimony concerning trials and other legal proceedings were used to feed data into the computer for analysis.

The Inquirer story of the study disclosed fascinating information, interesting to courts around the country as well as to Philadelphia residents. For example, only 30 percent of all persons convicted under one judge actually served prison sentences of any sort. What is the implication of such a statistic? Does it mean, as many claim, that criminals are being released back into society?

Noticeable statistical differences among blacks and whites also showed up. Black criminals were sent to jail for 7 months or longer 20 percent more often than white criminals, for example, while black judges imposed longer sentences 6 percent more often than white judges.

These are just a few of many facts the journalists wrote about after their statistical analysis. But these facts indicate the useful sociological information buried in dry statistics, information journalists and other social-problem researchers can present to society at large. Then it is up to society to decide what use to make of this knowledge.

Source: Ukena, Ann Seymour. Statistics Today, Harper & Row, Publishers, 1978, pp. 14-15.

Political opinion polling before elections is an excellent example of predicting results from a small sample. Pollsters need check only a small fraction of the voting population to come close to what actually happens on any given election night.

Often the forecasts are within five percent above or below the voted preference. That may appear to be a wide margin of error. But when you are assessing independent choices made by millions of individuals, that tolerance is sufficiently realistic to prove useful.

Pollsters achieve reasonable predictions by careful statistical probability principles. The goal is for evey eligible voter to stand an equal chance to be included in a sample. If this would be achieved, a truly random sample would result.

However, true random samples seldom work out. There are too many difficulties: finding the right people at home to give an accurate cross section of socioeconomic status, sex, age, race, and other variables, including honest answers to pollster questions.

Instead of a true random sample, pollsters try for artificial randomization that gives a representative sampling of the voting population. Various techniques are used, among them stratification.

For example, California has 10 million registered voters, out of 15 million eligible adults, of who some 8 or 9 million vote in any election. Poll sampling in that state stratifies, or divides the state into population concentrations. Then a sample of about 1000 persons is distributed throughout the state. That is, since Los Angeles county represents 38 percent of the state's voting population, 38 percent, or 380, will come from that county. And so it goes.

Within each stratification, census tract information provides the breakdown on the variables necessary to reflect an accurate picture of that voting community.

Usually those to be polled are put into distributive groups based on information collected from census data. Pollsters begin at some random start then pick, say, every one-thousandth individual until they have sufficient people. Extra persons are included to allow for those who cannot be reached and for balancing variables.

The above explanation is overly simplified and does not explain correction procedures for biases that inevitably creep in: if interviews are during the day, the sample is biased against many employed people; if in the evening, movie-goers are neglected. It's difficult to overcome them all.

But reputable pollsters know how to compensate for many of these biases. Nowadays we have fewer major blunders, such as Truman's unexpected win over Dewey, that made laughingstocks of opinion polls in years past.

Source: Ukena, Ann Seymour. Statistics Today, Harper & Row, Publishers, 1978, pp. 212-213.

From the Literature

Public policy is often based on social research findings. For example, one of the major policy interpretations set by the U.S. Supreme Court was based, in fact, on a series of research findings. That policy interpretation put an end to legally sanctioned segregation in the public schools, and it came about as a direct result of the Supreme Court desegregation decision (*Brown v. Board of Education*) of 1954. For many experts in the field of race relations, the 1954 Supreme Court ruling was the beginning of the militant civil rights activities.

Prior to the Court's decision, the seventeen Southern states and the District of Columbia had two complete sets of school systems (at the elementary and secondary levels). One set was for whites, the other for blacks. Previous Court rulings had upheld the doctrine of "equal but separate" school systems. Social research data, however, systematically revealed that the two systems, while separate, were not equal.

Sample sizes for the social research studies were rather large in several instances and required considerable summarization. One, for example, involved data collected on nearly 15,000 Southern blacks. Some of the data summarized and made available to the Court during its deliberations included the following:

> A U.S. Commission on Civil Rights study reported the median number of school years completed by blacks who were 25 years of age and over was 8.1, while the median for whites was 11.4.
>
> Performance tests showed that blacks in Northern schools (Illinois, New York, Ohio, and Pennsylvania) scored higher, on the average, than whites in Southern schools (Arkansas, Georgia, Kentucky, and Mississippi).
>
> These same performance tests showed white Southerners scored, on the average, higher than black Southerners.

Lyman Ott, Richard F. Larson, & William Mendenhall, Statistics (Boston: Duxbury Press), 1987: 104.

FLORENCE NIGHTINGALE

Florence Nightingale (1820–1910) won fame as a founder of the nursing profession and as a reformer of health care. As chief nurse for the British army during the Crimean War, from 1854 to 1856, she found that lack of sanitation and disease killed large numbers of soldiers hospitalized by wounds. Her reforms reduced the death rate at her military hospital from 42.7% to 2.2%, and she returned from the war famous. She at once began a fight to reform the entire military health care system, with considerable success.

One of the chief weapons Florence Nightingale used in her efforts was data. She had the facts, because she reformed record keeping as well as medical care. She was a pioneer in using graphs to present data in a vivid form that even generals and members of Parliament could understand. Her inventive graphs are a landmark in the growth of the new science of statistics. She considered statistics essential to understanding any social issue and tried to introduce the study of statistics into higher education.

In beginning our study of statistics, we will follow Florence Nightingale's lead. This chapter and the next will stress the analysis of data as a path to understanding. Like her, we will start with graphs to see what data can teach us. Along with the graphs we will present numerical summaries, just as Florence Nightingale calculated detailed death rates and other summaries. Data for Florence Nightingale were not dry or abstract, because they showed her, and helped her show others, how to save lives. That remains true today.

Tables and graphs are often confusing even though they are intended to present information concisely and unambiguously. Because of an inability to read tables and graphs, many people either misinterpret them or rely on an author's summary of what the data mean. However, another person's interpretation of a table or graph may be deliberately biased, accidentally misleading, or incomplete.

Tables and graphs have a lot of information packed into them, but if they have been properly organized, you can easily understand them by following certain steps (Wallis and Roberts, 1962:195–207). The steps outlined below are keyed to Table 3.1 and Figure 3.2.

1. **Begin by reading the title of the table or graph carefully; it will tell you what information is being presented.** Table 3.1 shows median annual incomes in the United States.

2. **Find out the source of the information.** You will want to know whether the source is reliable, whether its techniques for gathering and presenting data are sound. The figures originated from the U.S. Bureau of the Census, a highly trusted source. If you know the source of data, you can investigate further on your own.

3. **Read any notes accompanying the table or graph.** Not all tables and graphs have notes, but if they do, the notes should be read for further information about the nature of the data. The notes in Table 3.1 and in Figure 3.2 explain that all the data refer to the total money income of full-time and part-time workers, ages 25 and over, in a March 1988 survey.

4. **Examine any footnotes.** Footnotes in Table 3.1 and Figure 3.2 indicate that the data are categorized by the highest grade actually completed. Although you may have assumed this correctly, years of schooling could have referred to the total number of years in school, regardless of the grade level attained.

Table 3.1 Median Annual Income By Sex, Race, and Education

Demographic Group	Overall Median Income	Years of Schooling[1] Less Than 8	8	9–11	12	13–15	16 or More
White males	$22,189	$8,983	$11,178	$14,957	$21,016	$25,361	$34,889
Black males	$13,193	$6,655	$ 9,101	$10,604	$13,966	$19,597	$25,621
White females	$ 9,411	$4,989	$ 5,674	$ 6,451	$ 8,916	$12,331	$18,777
Black females	$ 7,899	$4,432	$ 4,562	$ 5,270	$ 9,284	$13,681	$20,658

Note: These figures include the total money income of full-time and part-time workers, ages 25 and over, as of March 1988.
[1]In terms of highest grade completed.
SOURCE: U.S. Bureau of the Census, *Money Income of Households, Families, and Persons in the United States: 1987*, Current Population Reports, Series P-60, No. 162 (Washington, DC: U.S. Government Printing Office, 1989a), pp. 140, 144.

5. **Look at the headings across the top and down the left-hand side of the table or graph.** To observe any pattern in the data, it is usually necessary to keep both types of headings in mind. Table 3.1 and Figure 3.2 show the median annual income of black and white males and females for several levels of education.

6. **Find out what units are being used.** Data can be expressed in percentages, hundreds, thousands, millions, billions, means, and so forth. In Table 3.1 and Figure 3.2, the units are dollars and years of schooling.

7. **Check for trends in the data.** For tables, look down the columns (vertically) and across the rows (horizontally) for the highest figures, lowest figures, trends, irregularities, and sudden shifts. If you read Table 3.1 vertically, you would be able to see how income varies by race and sex within each level of education. If you read the table horizontally, you could see how income varies with educational attainment for white males, black males, white females, and black females. A major advantage of graphs is that the sudden shifts, trends, irregularities, and extremes are easier to spot than they are in tables.

8. **Draw conclusions from your own observations.** Table 3.1 and Figure 3.2 show that although income tends to rise with educational level for both blacks and whites, it increases much less for black men and for women of both races than for white men. At each level of schooling, black men earn less than white men. In fact, white male high school dropouts have incomes almost as high as black high school graduates; white high school graduates earn only some $4,000 less each year than black males with sixteen or more years of education. Black women appear to improve their earning power through college education to a greater extent than do white women.

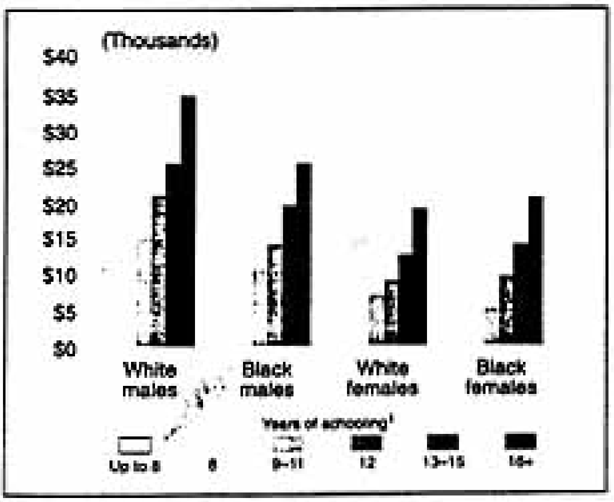
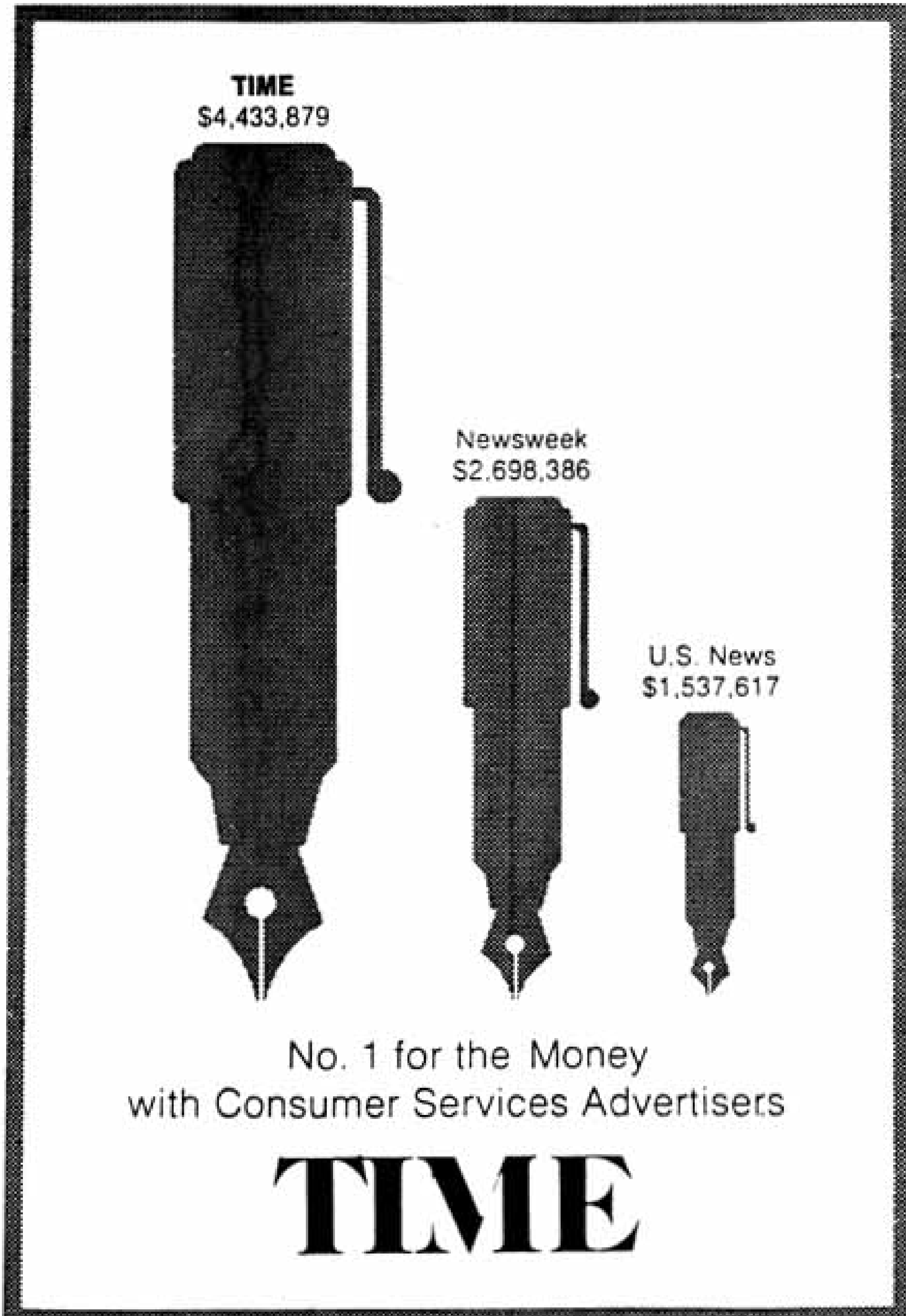
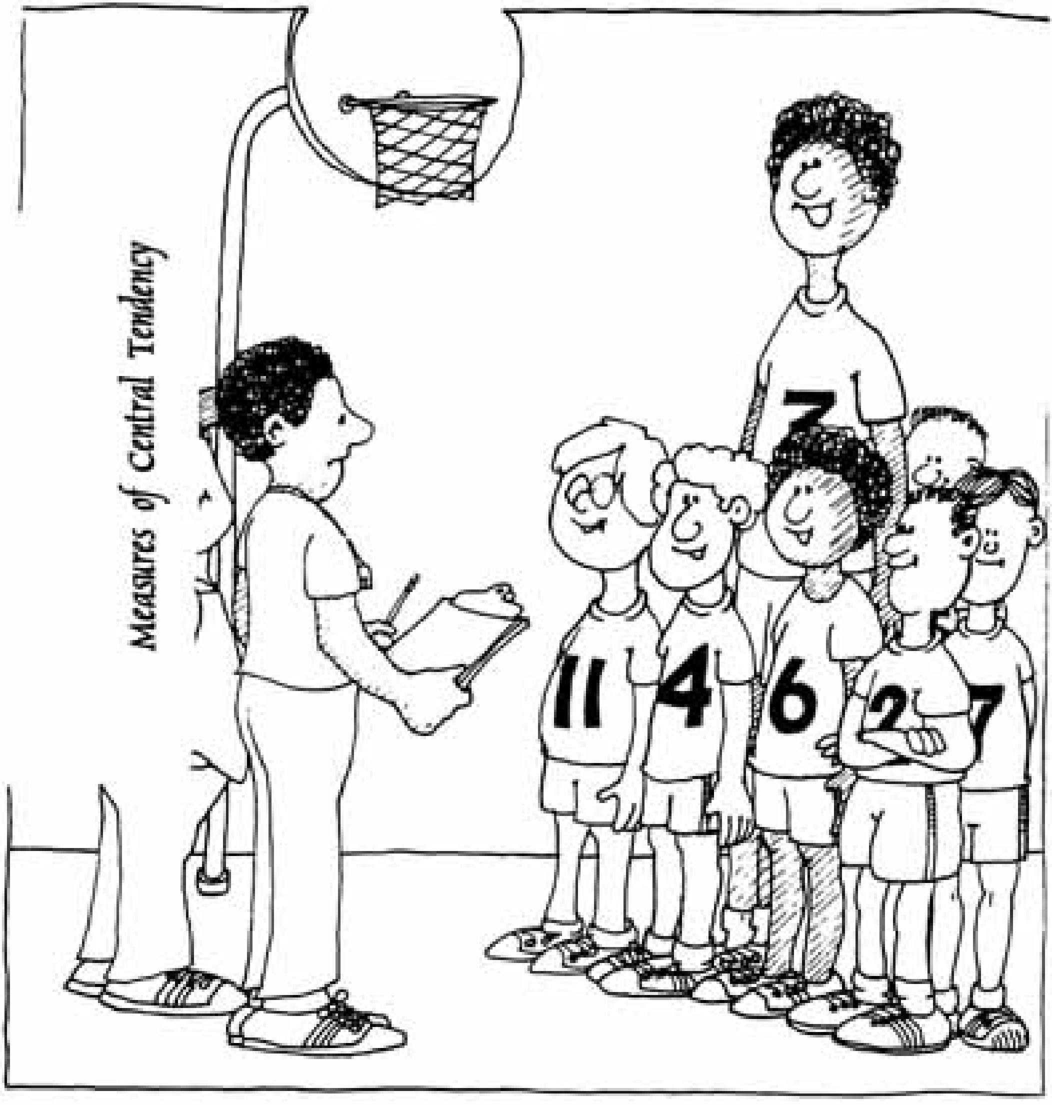


Figure 3.2 Median Annual Income by Sex, Race, and Education

Note: These figures include the total money income of full-time and part-time workers, ages 25 and over, as of March 1988.
[1]In terms of highest grade completed.
SOURCE: U.S. Bureau of the Census, *Money Income of Households, Families, and Persons in the U.S.: 1987*, Current Population Reports, Series P-60, No. 162 (Washington, DC: U.S. Government Printing Office, 1989a), pp. 140, 144.



An attractive but misleading bar graph. [Copyright © 1971 by Time, Inc. Reproduced by permission.]



"Should we scare the opposition by announcing our mean height or lull them by announcing our median height?"

On the Average[1]

The figure 2.2 children per adult female was felt to be in some respects absurd, and a Royal Commission suggested that the middle classes be paid money to increase the average to a rounder and more convenient number.

----PUNCH

In former times, when the hazards of sea voyages were much more serious than they are today, when ships buffeted by storms threw a portion of their cargo overboard, it was recognized that those whose goods were sacrificed had a claim in equity to indemnification at the expense of those whose goods were safely delivered. The value of the lost goods was paid for by agreement between all those whose merchandise had been in the same ship. This sea damage to cargo in transit was known as havaria and the word came naturally to be applied to the compensation money which each individual was called on to pay. From this Latin word derives our modern word "average". Thus the idea of an average has its roots in primitive insurance. Quite naturally with the growth of shipping, insurance was put on a firmer footing whereby the risk was shared, not simply by those whose goods were at risk on a particular voyage, but by large groups of traders. Eventually the carrying of such risks developed into a separate skilled and profit-making profession. This entailed the payment to the underwriter of a sum of money which bore a recognizable relation to the risk involved.

The idea of an average is common property. However scanty our knowledge of arithmetic, we are all at home with the idea of goal averages, batting and bowling averages and the like. We realize that the purpose of an average is to represent a group of individual values in a simple and concise manner so that the mind can get a quick understanding of the general size of the individuals in the group, undistracted by fortuitous and irrelevant variations. It is of the utmost importance to appreciate this fact that the average is to act as a representative. It follows that it is the acme of nonsense to go through all the rigmarole of the arithmetic to calculate the average of a set of figures which do not in some real sense constitute a single family. Suppose a prosperous medical man earning £3,000 a year had a wife and two children none of whom were gainfully employed and that the doctor had in his household a maid to whom he paid £150 a year and that there was a jobbing gardner who received £40 a year. We can go through all the processes of calculating the average income for this little group. Six people between them earn £3,190 in the year. Dividing the total earnings by the number of people, we may determine the average earnings of the group to be £531 13s.4d. But this figure is no more than an imposter in the robes of an average. It represents not a single person in the group. It gives the reader a totally meaningless figure, because he cannot make one single reliable deduction from it. This is an extreme example, but mock averages are calculated with great abandon. Few people ask themselves: What conclusions will be drawn from this average that I am about to calculate? Will it create a false impression?

The idea of an average is so handy that it is not surprising that several kinds of averages have been invented so that as wide a field as possible may be covered without misrepresentation. We have a choice of averages; and we pick out the one which is appropriate both to our data and our purpose. We should not let ourselves fall into the error that because the idea of an average is easy to grasp there is no more to be said on the subject.

[1] From M.J. Moroney, Facts from Figures 2d ed. (London: Penguin Books, 1953), pp. 34-35.

THE FAR SIDE

"Bob and Ruth! Come on in .... Have you met Russell and Bill, our 1.5 children?"

MAJOR CHARACTERISTICS OF EACH MEASURE OF CENTRAL TENDENCY

Mode

1. The mode is the most frequent or probable measurement or score in a distribution.
2. There can be more than one mode per distribution of scores.
3. The mode is not influenced by extreme scores.
4. Modes of subsets cannot be combined to determine the mode of the whole set.
5. The mode can be calculated when the ends of the distribution are open, provided that it does not fall in an open-ended interval.
6. The mode's value can change by organizing the data into different categories.
7. The mode is applicable to both qualitative and quantitative data.

Median

1. The median is the central value with 50% of the scores larger than it and 50% smaller.
2. There is only one median per distribution.
3. The median is not influenced by extreme values.
4. Medians of subsets cannot be combined to determine the median of the whole set.
5. The median can be calculated when the ends of the distribution are open, provided that it does not fall in an open-ended interval.
6. The median's value is rather stable even when data are organized into different categories.
7. The median is applicable to quantitative data only.

Mean

1. The mean is the sum of all scores divided by the number of scores.
2. There is only one mean per distribution.
3. The mean is influenced by extreme scores, and thus it may not be very representative of the distribution.
4. Means of subsets, when weighted, can be combined to determine the mean of the whole set.
5. The mean cannot be calculated when the ends of the distribution are open.
6. The mean is applicable to interval and ratio data only.

Lyman Ott, Richard F. Larson, & William Mendenhall, <u>Statistics</u> (Boston: Duxbury Press), 1987: 118.


DON'T WORRY, THAT ROPE IS ONE INCH THICK ON THE AVERAGE

It is not uncommon for Chambers of Commerce to report the mean annual temperature without reporting variability. A city with an "ideal" annual temperature of, say, 74° may go below zero in winter and above 100° in summer. The term "ideal" becomes a misleading statistical abstraction.

Sir Francis Galton Discovers Variability

There are two important uses for measures of variability: (1) describing the spread of distributions, and (2) describing an individual's standing within a group independent of the units of measurement. z-scores could tell us that a person's height is 1 standard deviation above the mean and his weight 0.5 standard deviations above the mean without mentioning inches or pounds. These ideas may seem obvious and indispensable now, but this was not always so.

While the concept of average originated at some uncertain period in antiquity, the idea of a variability measure has a much shorter history. During the first part of the nineteenth century, mathematicians described dispersion by probable error and the interquartile deviation.[1] The science of statistics was well underway at the time, but statisticians were slow to appreciate the utility of these mathematical curiosities. The idea of using variability indices, as we use them here, originated primarily with the English biologist, anthropologist, and psychologist Sir Francis Galton (1822-1911).

Galton's overriding interest was in uncovering the laws of inheritance; to this end he spent an active lifetime collecting and analyzing massive numbers of measurements on all sorts of phenomena. Sweet peas, race horses, dogs, and all classes of Englishmen were brought to his laboratory and measured in dozens of ways. When available statistical methods of his day would not suffice, he invented his own. The concepts of correlation and linear regression were developed by Galton as means of describing the relation between characteristics of parents and their offspring. He also pioneered the use of the normal curve in analyzing frequency distributions of biological and psychological characteristics. All these contributions, in turn, would not have been possible had he not brought variability into the laboratory.

> It is hard to believe now, but the idea of describing variability in groups was quite foreign to Galton's contemporaries. In 1890 he wrote: It seems to be a great loss of opportunity when, after observations have been laboriously collected, and been subsequently discussed in order to obtain mean values from them, that the small amount of extra trouble is not taken, which would determine other values whereby to express the variety of all the individuals in those groups.. . . . There are numerous problems of special interest to anthropologists that deal solely with variety.
>
> There can be little doubt that most persons fail to have adequate conceptions of the orderliness of variability, and think it is useless to pay scientific attention to variety, as being in their view, a subject wholly beyond the powers of definition. They forget that what is confessedly undefined in the individual may be definite in the group, and that uncertainty as regards the one is in no way incompatible with statistical assurance as regards the other . . . .
>
> Greater interest is attached to individuals who occupy positions towards the middle of a marshalled series than to those who stand about its middle. An average man is morally and intellectually an uninteresting being.[2]

Just as important, moreover, was his realization that distributions of different variables can be compared to one another if the observations are specified in units of variability. Galton's term for measures of variability was "statistical units," whose "office", he said,

> is to make the variabilities of totally different classes, such as horses, men, mice, plants, proficiency in classics, etc., comparable on equal terms. The statistical unit of each series is derived from the series itself.[3]

He no doubt recognized that his discovery of a use for measures of variability was of major consequence, and he described the moment of insight with a scene that brings to mind Newton's apple tree:

> As these lines are being written, the circumstances under which I first clearly grasped the important generalization that the laws of Heredity were solely concerned with deviations expressed in statistical units, are vividly recalled to my memory. It was in the grounds of Naworth Castle, where an invitation had been given to ramble freely. A temporary shower drove me to seek refuge in a reddish recess in the rock by the side of the pathway. There the idea flashed across me, and I forgot everything else for a moment in my great delight.[4]

Applications of Galton's "important generalization" go far beyond z-scores and t-scores. Without the ability to measure variability, and to use that index to specify an individual's standing in a group, we would have to end statistics at this point. Correlation, regression, and most inferential statistics require a means of specifying an individual's place in a group independently of the original units of measurement.

[1]The probable error is very similar to our modern standard deviation; the standard deviation was brought into general use by Karl Pearson in the 1890's.
[2]From Galton's Anthropometric Laboratory, Notes and Memoirs, No. 1, quoted in Karl Pearson, The Life, Letters and Labours of Francis Galton (Cambridge: Cambridge University Press, 1924), II, 384-385.
[3]Sir Francis Galton, Memories of My Life, 3d ed. (London: Methuen, 1909), p. 298.
[4]Ibid., p. 300.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 116-117.

The Normal Curve

The Normal Distribution

The normal curve that is so familiar in modern statistics was known among early-nineteenth-century mathematicians as the Law of Error--a term that describes its first practical applications in science.[1]

Astronomers of the 1830s were greatly concerned with errors. Mathematical theorems had been developed that predicted the orbits of various heavenly bodies very precisely, but theory could not always be advanced or verified through the telescope because human error entered into every observation. Differences in observers' reaction times and minute variations in spatial judgments from observation to observation added a personal equation to every measurement, one that had to be circumvented if possible. Thus, astronomers turned to mathematicians like Pierre La Place and Karl Gauss for detailed descriptions of errors. These men, perhaps more than any others, laid the foundation for modern statistics by pointing out that errors are distributed as a normal curve (although the term normal curve did not come into use until several decades later) and that the probabilities associated with different error sized can be predicted from a knowledge of the curve. This made it possible to predict the "most probable orbit" of heavenly bodies from a number of observations.

Sir Francis Galton, however, should receive primary credit for pointing out that the Law of Error also describes a multitude of biological and psychological distributions. Making the bridge between mathematics and biology, Galton wrote in Natural Inheritance:

> I need hardly remind the reader that the Law of Error upon which these Normal Values are based, was excogitated for the use of astronomers and others who are concerned with extreme accuracy of measurement, and without the slightest idea until the time of Quetelet that they might be applicable to human measures. But Errors, Differences, Deviations, Divergencies, Dispersions, and Individual Variations, all spring from the same kind causes . . . . The Law of Error finds a footing wherever the individual peculiarities are wholly due to the combined influence of a multitude of "accidents". . . . All persons conversant with statistics are aware that this supposition brings Variability within the grasp of the laws of Chance, with the result that the relative frequency of Deviations of different amounts admist of being calculated, when these amounts are measured in terms of any self-contained unit of variability.[2]

Galton was fully aware that using the mathematical abstraction to describe real distributions required an ultimately unprovable assumption, namely, that the underlying characteristic was normally distributed; he was also aware that many naturally occurring distributions are not normal in shape. He made it clear, however, that he believed the assumption of normality justified, in some cases, even if not provable:

> It has been objected to some of my former work, especially in Hereditary Genius, that I pushed the applications of the Law of Frequency of Error somewhat too far. I may have done so, rather by incautious phrases than in reality; but I am sure that with the evidence now before me, the applicability of that law is more than justified. . . . I am satisfied to claim that the Normal Law is a fair average representation of the Observed Curves during nine tenths of their course . . . the agreement of the Curve of Stature with the Normal Curve is very fair, and forms the mainstay of my inquiry into the laws of Natural Inheritance. It has already been said that mathematicians laboured at the Law of Error for one set of purposes, and we are entering into the fruits of their labours for another.[3]

[1]Abraham DeMoivre published the equation for the curve in 1733, but it saw little practical application for nearly 100 years. Those interested in a more complete history of the normal curve -- and of many other methods treated in statistics -- should see Helen M. Walker's classic, Studies in the History of the Statistical Method (Baltimore: Williams and Wilkins Co., 1929).

[2]Sir Francis Galton, Natural Inheritance (London: Macmillan and Co., 1894), pp. 54,55.

[3]Ibid., pp. 56, 57.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp.140-141.

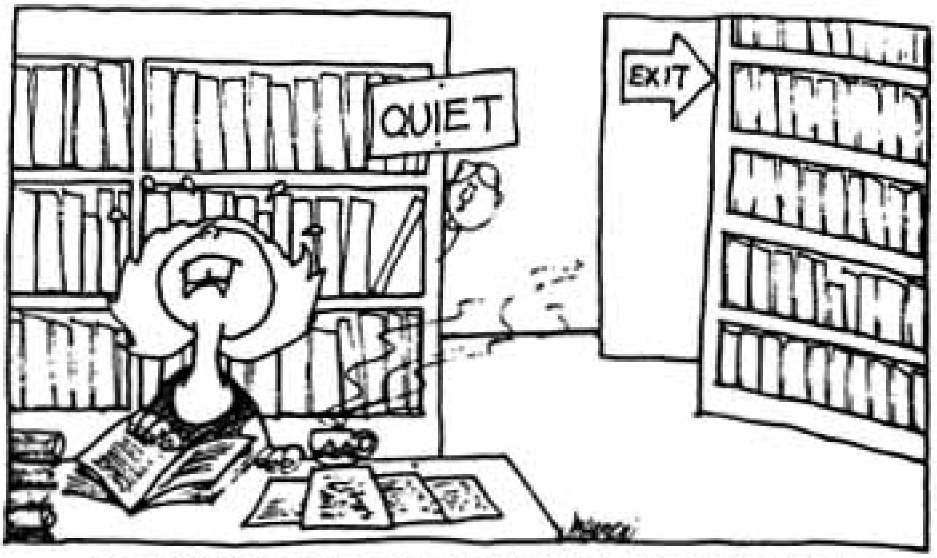
The Gambler's Fallacy

"When you're hot you're hot" is more than just a song title. It is an expression that summarizes some irrational feelings all of us--including nongamblers--occasionally fall prey to, known as the "gambler's fallacy."

Suppose you flip a coin six times and it turns up heads every time. What would you bet is the outcome of the next flip? Most people would probably choose heads; it is hard to fight the belief that there is a "run" or "streak" in progress. A few might choose tails for just the opposite reason, thinking that the laws of probability demand an evening up of the score. Both lines of reasoning are faulty, however, the probability that heads will appear on the next slip is exactly 0.50, just as it has been on every preceding flip. "But," someone might ask, "isn't it extremely unlikely that seven heads in a row will turn up?" It is. The probability of obtaining seven consecutive heads is only 0.0078. However, once six consecutive heads have already occurred (the probability of that happening is 0.0156), the probability that the next flip will be heads is 0.50. Each flip is statistically independent of all preceding flips. When events are independent, the probability that one occurs is in no way altered by the occurrence or nonoccurrence of the other.

If you still have trouble discounting the fallacy, consider the story of a man who was caught bringing a bomb on board an airplane. When questioned, he replied there was nothing to worry about--he was just a professor of statistics deathly afraid of being bombed on an airplane. He calculations showed that while the probability was very low that someone carrying a bomb would board any given airplane, it is much lower that two such people would board the same airplane. So he attempted to lower the probability of getting on board with a madman by bringing his own bomb.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, p. 244-245.



"I've had it! Simulated wood, simulated leather, simulated coffee, and now simulated probabilities!"

Courtesy of Dr. David Blackwell, University of California, Berkeley

DAVID BLACKWELL

Statistical practice rests in part on statistical theory. Statistics has been advanced not only by people concerned with practical problems, from Florence Nightingale to R. A. Fisher and John Tukey, but also by people whose first love is mathematics for its own sake. David Blackwell (1919–) is one of the major contemporary contributors to the mathematical study of statistics.

Blackwell grew up in Illinois, earned a doctorate in mathematics at the age of 22, and in 1944 joined the faculty of Howard University in Washington, D.C. "It was the ambition of every black scholar in those days to get a job at Howard University," he says. "That was the best job you could hope for." Society changed, and in 1954 Blackwell became professor of statistics at the University of California at Berkeley.

Washington, D.C., had an active statistical community, and the young mathematician Blackwell soon began to work on mathematical aspects of statistics. He explored the behavior of statistical procedures which, rather than working with a fixed sample, keep taking observations until there is enough information to reach a firm conclusion. He found insights into statistical inference by thinking of inference as a game in which nature plays against the statistician. Blackwell's work uses probability theory, the mathematics that describes chance behavior. We must travel the same route, though only a short distance. This chapter presents, in a rather informal fashion, the probabilistic ideas needed to understand the reasoning of inference.

Leicester, England, June 22 - The congregation of more than 300 was singing "inflame, we pray, our inmost hearts, with fire from heaven above," when lightning struck the Church of St. James the Greater.

"The whole place was suddenly bathed in light," said the vicar, the Rev. Lawrence Jackson. Some dust fell on him, but no one was hurt, damage was negligible, and the service went on.

The event described above actually occurred in 1960. What do you suppose the probability is that lightning would strike that particular church <u>while</u> the congregation was singing that particular line? It is no doubt practically 0.0.

It is quite common after disasters or other momentous events for news commentators or other journalists to dwell on the unlikely string of events that preceded and led to the event. Many such events, such as the sinking of the <u>Titanic</u> or the explosion of the <u>Hindenburg</u>, were indeed very unlikely events. However, upon reflection you should be able to see that <u>any specific</u> event is really almost "impossible" in a probabilistic sense. With this in mind, imagine how you would have responded had you been on the jury in the 1968 trial described below:[1]

> Trial by Mathematics. After an elderly woman was mugged in an alley in San Pedro, Calif., a witness saw a blonde girl with a ponytail run from the alley and jump into a yellow car driven by a bearded Negro. Eventually tried for the crime, Janet and Malcolm Collins were faced with the circumstantial evidence that she was white, blonde and wore a ponytail while her Negro husband owned a yellow car and wore a beard. The prosecution, impressed by the unusual nature and number of matching details, sought to persuade the jury by invoking a law rarely used in a courtroom--the mathematical law of statistical probability.
>
> The jury was indeed persuaded, and ultimately convicted the Collinses (TIME, Jan. 8, 1965). Small wonder. With the help of an expert witness from the mathematics department of a nearby college, the prosecutor explained that the probability of a set of events actually occurring is determined by multiplying together the probabilities of each of the events. Using what he considered "conservative" estimates (for example, the chances of a car's being yellow were 1 to 10, the chances of a couple in a car being interracial 1 in 1,000), the prosecutor multiplied all the factors together and concluded that the odds were 1 in 12 million that any other couple shared the characteristics of the defendants.
>
> Only One Couple. The logic of it all seemed overwhelming, and few disciplines pay as much homage to logic as do the law and math. But neither works right with the wrong premises. Hearing an appeal of Malcolm Collins' conviction, the California Supreme Court recently turned up some serious defects, including the fact that not even the odds were all they seemed.
>
> To begin with, the prosecution failed to supply evidence that "any of the individual probability factors listed were even roughly accurate." Moreover, the factors were not shown to be fully independent of one another as they must be to satisfy the mathematical law; the factor of a Negro with a beard, for instance, overlaps the possibility that the bearded Negro may be part of an interracial couple. The 12 million to 1 figure, therefore, was just "wild conjecture". In addition, there was not complete agreement among the witnesses about the characteristics in question. "No mathematical equation," added the court, "can prove

beyond a reasonable doubt (1) that the guilty couple in fact possessed the characteristics described by the witnesses, or even (2) that only one couple possessing those distinctive characteristics could be found in the entire Los Angeles area."

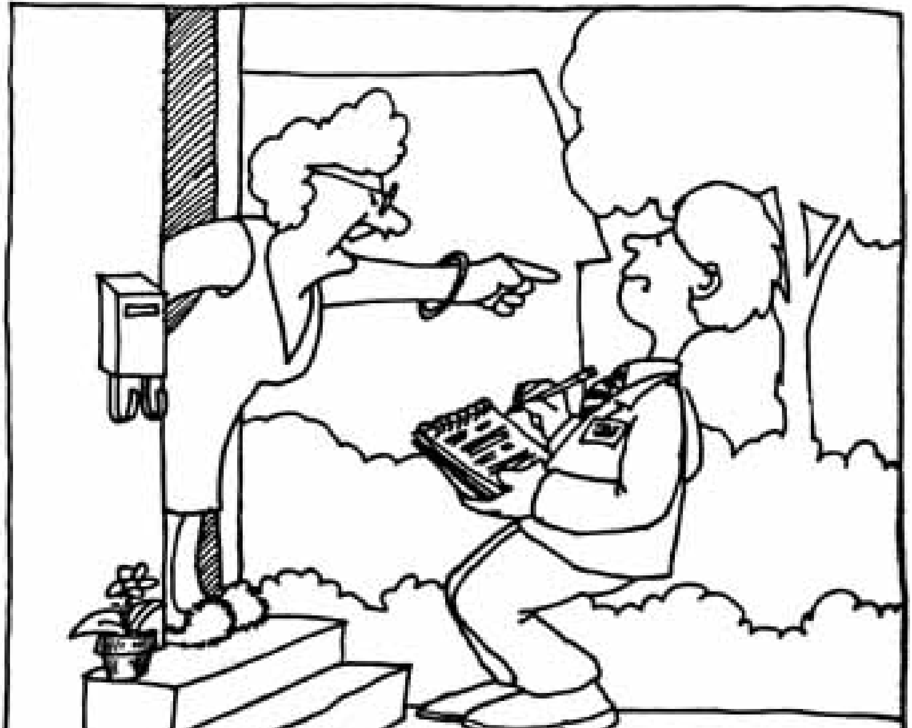
Improbable Probability. To explain why, Judge Raymond Sullivan attached a four-page appendix to his opinion that carried the necessary math far beyond the relatively simple formula of probability. Judge Sullivan was willing to assume it was unlikely that such a couple as the one described existed. But since such a couple did exist -- and the Collinses demonstrably did exist -- there was a perfectly acceptable mathematical formula for determining the probability that another such couple existed. Using the formula and the prosecution's figure of 12 million, the judge demonstrated to his own satisfaction and that of five concurring justices that there was a 41% chance that at least one other couple in the area might satisfy the requirements.[2]

"Undoubtedly," said Sullivan, "the jurors were unduly impressed by the mystique of the mathematical demonstration but were unable to assess its relevancy or value." Neither could the defense attorney have been expected to know of the sophisticated rebuttal available to them. Janet Collins is already out of jail, has broken parole and lit out for parts unknown. But Judge Sullivan concluded that Malcolm Collins, who is still in prison at the California Conservation Center, had been subjected to "trial by mathematics" and was entitled to a reversal of his conviction. He could be tried again, but the odds are against it.

[1]"Trial by Mathematics," Time (April 26, 1968).
[2]The proof involved is essentially the same as that behind the common parlor trick of betting that in a group of 30 people, at least two will have the same birthday; in that case, the probability is 70%.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 245-246.

Sampling



"Have you ever thought of adding an indicator of how people feel about having their opinions asked every other day?"

Whom Does the Sample Represent?

"Everyone knows that a representative sample of lemon pie, for instance, must include the meringue on the top, the lemon in the middle, and the crust on the bottom." --George Gallup and Saul Rae[1]

As you read reports of experiments in the behavioral sciences you may notice that in some instances researchers selecting people for samples take the greatest pains to ensure that the sample is representative of the population. The Gallup Poll, for instance, is based on a highly complex stratified random sample, which takes into account "considerations of geography, occupation, age, sex, political affiliation, race, religion, and general cultural background."[2] On the other hand, a psychologist studying the ability of human beings to understand speech in the presence of distracting noise may use only female college students between the ages of 18 and 20 as subjects; other psychologists interested in general principles of learning may base their conclusions entirely on data obtained from laboratory-raised albino rats. When are representative sampling procedures needed, and when aren't they? To answer that question, one needs to think very carefully about (1) what is being studied, and (2) to what population will sample results be generalized. W.L. Hays, in his book, Statistics for Psychologists, offers some guidelines on the matter in a section entitled "To What Populations Do Our Inferences Refer?"

> Most psychologists who use inferential statistics in research rely on the model of simple random sampling. Yet how does one go about getting such a "truly" random sample? It is not easy to do, unless, as in all probability sampling, each and every potential member of the population may somehow be listed. Then, by means of a device such as random number tables, individuals may be assigned to the sample with approximately equal probabilities.
>
> However, in behavioral sciences such as psychology, interest often lies in experimental effects that, presumably, should apply to a very large population of men or other living organisms. Such a listing procedure is simply not possible. Still other experiments may refer to all possible measurements that *might* be made of some phenomenon under various experimental conditions, where estimated true values may be sought from the experimental observation of a few instances. Here, the population is not only infinite, it is hypothetical, since it includes all *future* or *potential* observations of that phenomenon under the different conditions. In sampling from such experimental populations, where there is no possibility of listing the elements for random assignment to the sample, the only recourse of the experimenter is to draw his basic experimental units in some more or less random, "haphazard," way, and then *make sure that in his experiment only random factors determine which unit gets which experimental treatment*. In other words, there are two ways in which randomness is important in an experiment: the first is in the selection of the sample as a whole, and the second is in the allotment of individuals to experimental treatments. Each kind of randomness is important for the "generalizability" of the experimental results, so that when one does an experiment he usually takes pains to see that both kinds of randomness are present. However, even given that individual cases are assigned to experimental manipulations at random, the possible inferences are still limited by the fundamental population from which the total sample is drawn.

How does one know the population to which the statistical inferences drawn from a sample apply? If random sampling is to be assumed, the population is defined by the sample and the manner in which it is drawn. The only population to which the inferences strictly apply is that in which individuals have equal likelihood of appearing in the sample. It should be obvious that simple random samples from one population may not be random samples of another population. For example, suppose that some one wishes to sample American college students. He obtains a directory of college students from a midwestern university and, using a random number table, takes a sample of these students. He is not, however, justified in calling this a random sample of the population of American college students, although he may be justified in calling this a random sample of students at that university. The population is defined not by what he said, but rather by what he did to get the sample. For any sample, one should always ask the question, "What is the set of potential cases that could have appeared in my sample with equal probability?" If there is some well-defined set of cases which fits this qualification, then inferences may be made to that population. However, if there is some population whose members could not have been represented in the sample with equal probability, then inferences do not necessarily apply to that population when methods based on simple random sampling are used. Any generalization beyond the population actually sampled at random must rest on extrastatistical, scientific, considerations.[3]

[1]G. Gallup and S. Rae, The Pulse of Democracy (New York: Glenwood Press, 1968), p. 64.
[2]Ibid., p. 60.
[3]From Statistics by William L. Hays, (Holt, Rinehart and Winston, Inc., 1963).

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 268-269.

HOW THE NIELSEN TV RATINGS ARE ARRIVED AT

Since 1950 the A.C. Nielsen Company has been an integral part of the television industry. Its major purpose is to measure television audiences, document its growth and characteristics for advertisers, agencies, broadcasters and others involved in the medium. Since the number of televised sporting events has continued to grow it is of interest to know how the actual ratings are arrived at. The first thing we must emphasize is that the Nielsen ratings are not intended to nor do they measure program quality. Instead, they provide reliable and quantitative estimates of tv audience size and characteristics.

The rating techniques are based upon sampling theory and are scientifically valid. A sample (part of the total phenomenon of interest) is selected because to study the entire population of 71 million tv homes would be financially and practically prohibitive. Sampling is not a last resort but a highly efficient procedure for estimating characteristics of a larger population from which it is selected. Nielsen selects a probability sample--technically it is an area probability sample--of about 1200 households. These households are randomly selected and households cannot volunteer to be part of the sample. Each sample household's television set is installed with a device samller than a cigar box--technically known as the Storage Instantaneous Audimeter (SIA)--which monitors and records in its computer-like memory whether or not the tv is on and what channel it is tuned to. When Nielsen's Central Office Computer in Dunedin, Florida wishes to retrieve this information, it is sent along special telephone lines to the Office. This procedure enables Nielsen to determine what shows are turned on. To determine who—not what--is tuned in a separate sample of National Audience Composition (NAC) households keep a diary of their viewing habits.

Through these procedures, the Nielsen organization produces a tv rating—a statistical estimate of the number of homes tuned to a program. Hence if a program receives a rating of 44.4 as did a recent Super Bowl, it means that nearly 44½ percent of U.S. tv homes were tuned in to that program. Since over 71 million households (98 percent of the total number of households have tv sets) a rating of 44.4 means that an estimated 31½ million tv households tuned in. In other words:

Rating X 71 million = Number of Households Tuned In

It must be emphasized, however, that the figures are estimates--but accurate ones. If the Nielsen sample constructed a rating of 44.4 percent, the true rating lies somewhere between 43.1 and 45.7 sixty-seven percent of the time. One-third of the time the "error" may be larger but when repeated ratings are taken the range of error correspondingly reduces.

In summary, ratings appear to benefit the television audience--because they provide a barometer of peoples' likes and dislikes. It also provides the tv industry--advertisers, agencies, networks, tv stations, program producers--with vital information regarding the public's viewing habits and preferences.

SOURCE: "Nielsen Television 78", A.C. Nielsen Co. (Chicago, IL: Media Research Services Group, 1978).

Sample Size and Error of Estimate

To emphasize a point, statistical methods themselves do not automatically reject bad data. Any group of observations, for instance, can be used to compute a confidence interval for the mean at the 95 percent confidence level and, in general, the larger the sample size, the narrower the interval. So, if a sociological researcher knows that 40 interviews provide a good estimate, why shouldn't he obtain 400 interviews to get an even more precise estimate? The problem is that data quality is sometimes affected by the number of observations that must be made. W.A. Wallis and H.V. Roberts addressed the problem of sample size and error in The Nature of Statistics:

> A large number of measurements made hurriedly or superficially may not represent as much true information as a small number made carefully. In extreme cases, poor data can be so misleading as to be worse than no information at all. A rather paradoxical example of the effective use of samples is the Bureau of the Census' use of them to check on the accuracy of the census. Although sampling error is almost absent from the census, the nonsampling error is considerable--that is, such errors as those arising from failure to make questions clearly understood, from misrecording replies, from faulty tabulation, from omitting people who should have been interviewed. In the sample census, however, these nonsampling errors may be reduced enough to offset the sampling error, for it is cheaper and easier to select, train, and supervise a few hundred well-qualified interviewers to conduct a few thousand careful interviews than it is to select, train, and supervise 150,000 interviewers to conduct a complete census of the population. Similarly, in measuring the useful life of the equipment in a telephone plant, the practical choice is not between measurements for a sample of the equipment and equally accurate measurements for all the equipment, but between fairly precise measurements of a sample made carefully by competent engineers, and crude measurements of the whole plant made hastily by less skilled people. Even in laboratory experiments in the sciences, the difficulties of precise measurement are often so great that it is better to reduce the number of items measured in order to take more care with the individual measurements. . . .
>
> The reader may wonder why, in view of the advantages of sampling, the entire population of the United States is enumerated completely every ten years. Aside from the overriding fact that the Constitution requires this, perhaps the most important reason is that information is required for very small groups of the population--such as small towns, individual neighborhoods in cities, etc.--as well as for the country as a whole. Even so, however, about half the questions on the 1950 census were asked only of a sample--for some questions a 20 percent sample, and for some questions a 3-1/3 percent sample (namely, a 16-2/3 percent subsample of the 20 percent sample).[1]

[1]W.A. Wallis and H.V. Roberts, The Nature of Statistics (New York: The Free Press, 1962), p. 138.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 333-334.

A Scheme for Visualizing the Nature of Statistical Estimation

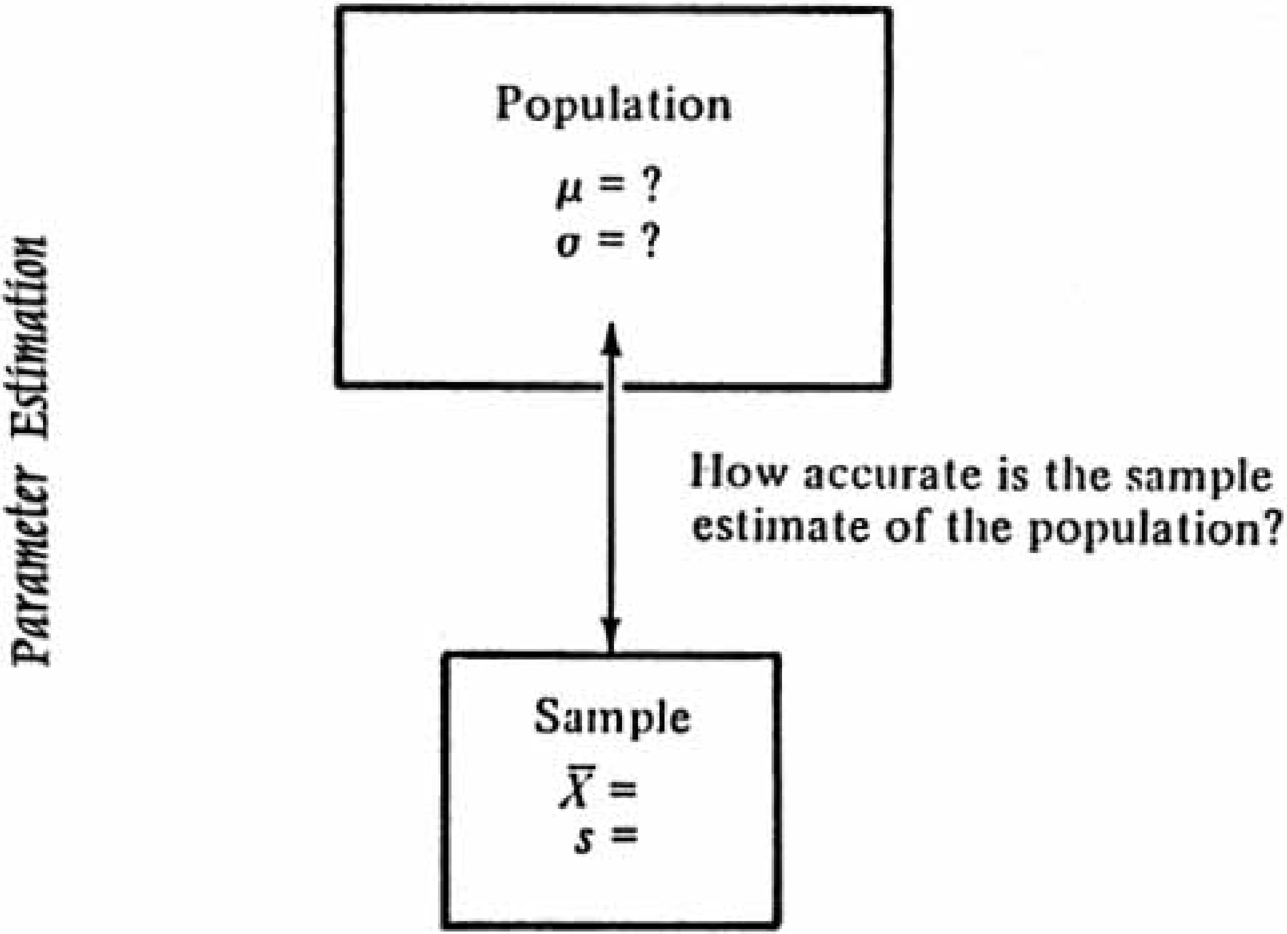


Photo by G. Paul Bishop, courtesy of Paul Bishop, Jr., Berkeley, Ca.

JERZY NEYMAN

The most-used methods of statistical inference are confidence intervals and tests of significance. Both are products of the twentieth century. From complex and sometimes confusing origins, statistical tests took their current form in the writings of R. A. Fisher, whom we met at the beginning of Chapter 3. Confidence intervals appeared in 1934, the brainchild of Jerzy Neyman (1894–1981).

Neyman was trained in Poland and, like Fisher, worked at an agricultural research institute. He moved to London in 1934 and in 1938 joined the University of California at Berkeley. He founded Berkeley's Statistical Laboratory and remained its head even after his official retirement as a professor in 1961. Retirement did not slow Neyman's work—he remained active until the end of his long life and almost doubled his list of publications after "retiring." Statistical problems arising from astronomy, biology, and attempts to modify the weather attracted his attention.

Neyman ranks with Fisher as a founder of modern statistical practice. In addition to introducing confidence intervals, he helped systematize the theory of sample surveys and reworked significance tests from a new point of view. Fisher, who was very argumentative, disliked Neyman's approach to tests and said so. Neyman, who wasn't shy, replied vigorously.

Tests and confidence intervals are our topic in this chapter. Like most users of statistics, we will stay close to Fisher's approach to tests. You can find some of Neyman's ideas in the optional final section.

HOW RELIABLE IS THAT POLL?

CASE STUDY

It is almost impossible to read a daily newspaper or listen to the radio or view telecasts without hearing about some opinion poll or economic survey. For many of us comes the inevitable question: How reliable are the percentages derived from these samples of public opinion? Do the national polls conducted by the Gallup and Harris organizations, the news media, and so on really provide accurate estimates of the percentages of people in the United States who favor various propositions?

A report of the results of a poll conducted by the *New York Times*/WCBS-TV provides a clue to these answers (*New York Times*, May 14, 1985). The object of the poll was to determine the opinions of New York residents concerning race relations in the city. Nested in the middle of the report is a box titled "How Poll Was Conducted," which explains, among other things, that the poll consisted of telephone interviews with 1557 adults in all parts of New York City. Describing the reliability of the poll results, the box states that "in theory, in 19 cases out of 20 the results based on such samples will differ by no more than 3 percentage points in either direction from what would have been obtained by interviewing all adult New Yorkers." In addition, the margin of error for smaller groups, racial or ethnic, which would represent only fractions of the total of 1557 adults in the sample, would be larger than 3 percent.

The size of the sample of the *New York Times*/WCBS-TV poll is typical of the sample size chosen for most major national polls. And, it is quite common to read that the margin of error for these polls is plus or minus 3 percent. Is this correct and how did the pollsters arrive at this figure?

In this chapter you will learn how sample statistics are used to estimate the values of population parameters, such as population means and proportions, and you will learn how to evaluate the reliability of these estimates. Then you will use what you have learned to reexamine the reliability of the *New York Times*/WCBS-TV poll.

William Mendenhall, Introduction to Probability and Statistics (Boston, Duxbury), 1987), pp. 297, 336-337.

Do these calculations confirm the *New York Times* statement that the sample percentages will vary less than 3 percent from the actual population percentages? The answer is "yes but." It is yes if we assume that the sample is a simple random sample. Based on simple random sampling, our calculations show that the bound on the error of estimation would be less than 3 percent. The "but" is necessary because it is often difficult, if not impossible, to draw a simple random sample. This is because it is often impossible to acquire a complete list of all adults in the population from which the sample should be selected. For example, a telephone listing would omit the impoverished adults in the city who do not have telephones and who may have very different views about racial relations than those included in the *New York Times*/WCBS-TV samples. Was this group sampled and, if not, is the group large enough to affect the poll results? Most pollsters employ sampling procedures that attempt to sample all segments of a population and to achieve something close to random sampling. To the extent that they are successful, the bound on the error of estimation of a population percentage for a sample of approximately 1600 respondents is probably less than 3 percent.

Table 11.2
Approximate 95% Confidence Intervals for an Observed Sample Frequency of .50 in Samples of Various Sizes

Sample Size	*Confidence Interval*	*Margin of Error*
10	.20 to .80	+/− .30
25	.28 to .72	+/− .22
50	.36 to .64	+/− .14
100	.40 to .60	+/− .10
250	.44 to .56	+/− .06
500	.46 to .54	+/− .04
1000	.47 to .53	+/− .03
1500	.48 to .52	+/− .02

$\sqrt{\frac{PQ}{N}} = \sqrt{\frac{(.5)(.5)}{10}}$

Ronald N. Giere, Understanding Scientific Reasoning (New York: Holt, 1984, p. 236.

JANET NORWOOD

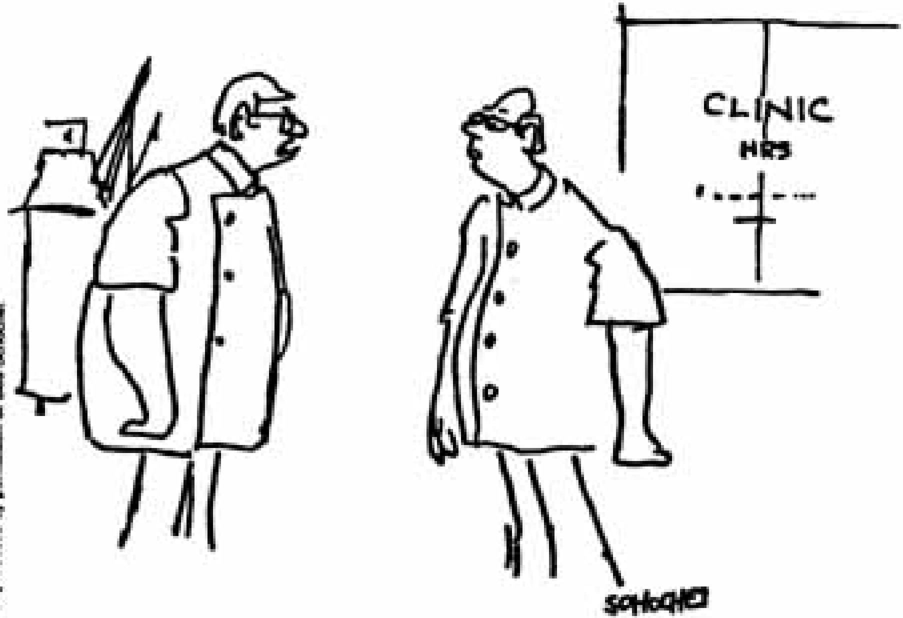
The commissioner of labor statistics is one of the nation's most influential statisticians. As head of the Bureau of Labor Statistics, the commissioner supervises the collection and interpretation of data on employment, earnings, and many other economic and social trends.

The data collected by the Bureau of Labor Statistics are often politically sensitive, as when a report released just before an election shows rising unemployment. For this reason, the bureau must remain objective and independent of political influence. To safeguard the bureau's independence, the commissioner is appointed by the president and confirmed by the Senate for a fixed term of four years. The commissioner must have statistical skill, administrative ability, and a facility for working with both Congress and the president.

Janet Norwood served three terms as commissioner, from 1979 to 1991, under three presidents. When she retired, the *New York Times* said (December 31, 1991) that she left with "a near-legendary reputation for nonpartisanship and plaudits that include one senator's designation of her as a 'national treasure.'" Norwood says, "There have been times in the past when commissioners have been in open disagreement with the Secretary of Labor or, in some cases, with the President. We have guarded our professionalism with great care."

Some of the most important statistics produced by the Bureau of Labor Statistics are proportions. The monthly unemployment rate, for example, is the proportion of the labor force that is unemployed this month. Methods for inference about proportions are the topic of this chapter.

Measures of Significance: Hypothesis Testing



"Sure your patients have 50% fewer cavities. That's because they have 50% fewer teeth!"

Where Are All of Those Tests?

Hypothesis testing is a widely used approach to actual research problems. And yet, this excerpt may represent the first contact with formal statistical decision making for many readers. If statistical decisions in government, industry, and science have indeed had such a profound impact on our lives, why haven't we heard of hypothesis testing and statistical decision making before undertaking a systematic study of statistics?

Dr. Irwin D.J. Bross answered this question in the introductory chapter of his delightful book, Design for Decision. He briefly discussed the revolution in many fields of science that occurred with the application of sophisticated statistical methods in the period from 1920 to 1940. Then, turning to the matter of public awareness of these methods, he commented.

> I cannot blame you if, at this point, you scratch your head and murmur, "All this looks suspiciously like the old ballyhoo. If Statistical Decision is such a world-shaking affair why haven't I felt some of the tremors?" You may not have heard of the statistical "revolution" that I mentioned earlier, and, to digress a bit, let me explain why you may not have heard of these matters. The main reason is that publications on the subject are written only for fellow specialists (and even these worthies have trouble understanding them). It may take twenty years before these ideas reach other scientists in a comprehensible form and even longer before they are taught to students. Specific techniques (in cookbook form) may be transmitted more rapidly, but the ideas diffuse very slowly.
>
> A few scientists, it is true, have tried to write for the public. But while the public has eagerly accepted the television sets, wonder drugs, and bigger strawberries that scientific research has produced, they have been profoundly uninterested in the fundamental ideas, the Scientific Method, that have made this research fruitful. People must have the very latest electronic gadget, but they cling tenaciously to ideas and methods of thinking that were obsolete three hundred years ago.
>
> This delay in the transmission of ideas is, I believe, one of the factors which has led our civilization to its present crisis. Moreover, the already dangerous situation is steadily getting worse because it is increasingly difficult to translate the language of science -- a symbolic one -- into everyday English.[1]

Take another look at the way scientific results are reported in the magazines and newspapers. Occasionally you may find reference to probability, but rarely, if ever, will you see a reference to a hypothesis test or to a statistical decision. Rejected null hypotheses in the scientific journals have a way of turning into "proof" in the papers.

[1]From I.D.J. Bross, Design for Decision (New York: Macmillan, 1953), pp.3-4.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 294-295.

KARL PEARSON

Karl Pearson (1857–1936), a professor at University College in London, had already published nine books before he turned his abundant energy to statistics in 1893. Of course, Pearson didn't really take up statistics, which was not yet a separate field of study. He took up problems of heredity and evolution, which led him into statistics.

Pearson developed a family of curves—we would call them density curves—for describing biological data that don't follow a normal distribution. He then asked how he could test whether one of these curves actually fit a set of data well. In 1900 he invented a method, the chi-square test. Pearson's chi-square test has the honor of being the oldest inference procedure still in use. It is now most often used for problems somewhat different from the one that motivated Pearson, as we will see in this chapter.

After Pearson, statistics was a field of study. Fisher and Neyman in the 1920s and 1930s would provide much of its present form, but here is what the leading historian of statistics says about the origins:

Before 1900 we see many scientists of different fields developing and using techniques we now recognize as belonging to modern statistics. After 1900 we begin to see identifiable statisticians developing such techniques into a unified logic of empirical science that goes far beyond its component parts. There was no sharp moment of birth; but with Pearson and Yule and the growing numbers of students in Pearson's laboratory, the infant discipline may be said to have arrived.[1]

Chi-Square in Research

The chi-square statistic is very popular in behavioral science research; a glance through any of the psychological or sociological journals, for instance, suggests that a large proportion of the "facts" in these disciplines are based on chi-square evidence. One reason, no doubt, for the popularity of chi-square is that it can be used in many nonparametric tests. This means that the variable under study need not come from a normally distributed population or represent the higher levels of measurement. Following is a short summary of a typical chi-square application from the area of social psychology.

Do Americans support the Bill of Rights? And does it matter who asks them? A number of studies suggest that a majority of adult Americans will not endorse the Bill of Rights (first ten amendments to the U.S. Constitution) when the original text is placed before them but not identified as the Bill of Rights. Dr. William Samuel conducted a study to see if this was true in one area of Sacramento, California, and also whether it made a difference if a "hip" or "straight" canvasser asked for the endorsement.[1]

Thirteen college-age researchers solicited signatures at a number of middle-class homes in a Sacramento suburb. Seven were dressed in "straight" attire and six in "hip" costume. Each researcher carried three different statements: One was a paraphrased version of the real Bill of Rights (which guarantees a number of basic freedoms), one was a negative paraphrased version that urged restriction of these rights, and a third was a "wishy washy" paraphrase that attempted to take a middle ground between the other two versions. At each house a researcher introduced himself or herself as a representative of a student group called Youth for America; the resident was then asked to read <u>one</u> of the paraphrases and to sign it if he agreed with it.

There was thus two independent variables: attire of the canvasser ("hip" or "straight") and version of the Bill of Rights (real, negative, and "wishy washy"). All respondents were exposed to one level of each independent variable, and all were measured on the same dependent variable, "signature or no signature".

As Samuel expected, people approached on three days of testing[2] were apparently more ready to sign the negative version than the real or "wishy washy" version: 62 percent of those approached endorsed the negative version, while only 46 percent signed the "wishy washy" text, and 44 percent the real paraphrase. But are these differences significant? Is the difference between 62 percent and 44 percent, for instance, due to the effect of the independent variable "version read," or to chance variability. When the chi-square was computed for the appropriate frequencies, the difference appeared to be real ($X^2 = 7.52$, $df = 2$, $p < 0.025$). Thus, the chi-square test supports the conclusion that most of these people would not endorse the Bill of Rights.

However, some other interesting results appeared on closer examination of the data. The above-mentioned preference for the negative version seemed to hold only when the canvasser was dressed as a "straight"; with "hip" canvassers, the frequencies of signatures for different versions was non-significant ($X^2 = 1.46$, $df = 2$). Furthermore, "straights" were more likely to obtain signature than "hips" (for "hip" and "straight" signature rates, $X^2 = 4.46$, $df = 1$, $p < 0.05$).

[1]W. Samuel, "Response to Bill of Rights Paraphrases as Influenced by the Hip or Straight Attire of the Opinion Solicitor," <u>Journal of Applied Social Psychology</u>, 2(1972), 47-62.

[2]The canvassing was actually conducted on four days of testing, but the results were complicated by the fact that one day's canvassing occurred on the Sunday following the Kent State and Jackson State shooting incidents. See Samuel's article for a description of what happened on this day, and the author's theoretical interpretation of these results.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 367-368.

From the Literature

One of the most famous community studies, *Elmtown's Youth: The Impact of Social Classes on Adolescents*, was published by August B. Hollingshead in 1949. Hollingshead focused his research on the relationship between the social stratification system and the social behavior of adolescents (aged 13 to 19) in Elmtown. Elmtown was studied because it was thought to be a "typical Middle Western community" with a clearcut class structure: upper, upper middle, middle, lower middle, and lower.

Hollingshead was particularly interested in showing the dependence of a series of social behaviors on social class position (the independent variable). He used chi-square, a simple, effective statistical tool, to test the dependence of particular social behaviors on class position.

Because so few of the adolescents were from upper-class families, he combined the upper and upper-middle classes into a single category. Thus the independent variable, social class, was divided into four categories: upper and upper-middle, middle, lower-middle, and lower. The number of categories for the dependent variables ranged from two (working mothers versus nonworking mothers, for example) to six (religious affiliation: Federated, Methodist, Lutheran, Catholic, Baptist, and no affiliation).

Some of the major findings from this study, including the chi-square test results and p-values, follow:

> The higher the class, the more likely parents were counseled on the schoolwork of the child; the lower the class, the more likely parents were counseled on the discipline of the child ($\chi^2 = 19.41, p < .01$).
>
> The higher the class, the greater the number of athletic events, high school dances, and evening plays and parties attended ($\chi^2 = 152.91, p < .01$, for athletic events; $\chi^2 = 95.41$, $p < .01$, for dances; and $\chi^2 = 131.24$, $p < .01$, for plays and parties).
>
> Church affiliation is related to class position: Higher classes are affiliated with the Federated and Methodist churches, the lower classes with the Lutheran, Catholic, and Baptist churches ($\chi^2 = 300.00, p < .01$).

Lyman Ott, Richard F. Larson, & William Mendenhall, Statistics (Boston: Duxbury Press), 1987: 377.

WILLIAM S. GOSSET

What would cause the head brewer of the famous Guinness brewery in Dublin, Ireland, not only to use statistics but to invent new statistical methods? The search for better beer, of course.

William S. Gosset (1876–1937), fresh from Oxford University, joined Guinness as a brewer in 1899. He soon became involved in experiments and in statistics to understand the data from these experiments. What are the best varieties of barley and hops for brewing? How should they be grown, dried, and stored? The results of the field experiments, as you can guess, varied. Statistical inference can uncover the pattern behind the variation. The statistical methods available at the turn of the century ended with a version of the z test for means—even confidence intervals were not yet available.

Gosset faced in his job the problem we noted in using the z test to introduce the reasoning of statistical tests: he didn't know the population standard deviation σ. What is more, field experiments give only small numbers of observations. Just replacing σ by s in the z statistic and calling the result roughly normal wasn't accurate enough. So Gosset asked the key question, What is the exact sampling distribution of the statistic $(\bar{x} - \mu)/s$?

By 1907 Gosset was brewer-in-charge of Guiness's experimental brewery. He also had the answer to his question and had calculated a table of critical values for his new distribution. We call it the t distribution. The new t test identified the best barley variety, and Guinness promptly bought up all the available seed. Guinness allowed Gosset to publish his discoveries, but not under his own name. He used the name "Student," and the t test is sometimes called "Student's t" in his honor. Gosset's statistical work helped him become head brewer, a more interesting title than professor of statistics.

t-Tests in Research

The t-test for the significance of the difference between two means has been applied to questions in all areas of science, from agronomy to zymurgy. Here is one example from psychology.

Do noisy environments affect human performance? Some people are required to spend their working hours amidst high noise levels. Those who live near airports, railroad tracks, and busy highways may be similarly condemned throughout their leisure hours. It is fairly well established that long exposure to even moderate noise levels leads to some permanent hearing loss. Researchers are less in agreement, however, about the effect of noise on human performance in different situations.

Finkelman and Glass reasoned that predictability of noise might be an important factor in determining whether or not performance is impaired by noise when people are working at the limits of their mental ability. To test this idea, they had volunteer subjects carry on mental tasks while being subjected to noise. In the repeated-measures experiment, subjects receiving the "predictable noise" condition were treated to blasts of 80-db noise through earphones, presented for 9 sec. each time, with blasts spaced at regular intervals. Under the "unpredictable noise" condition, subjects received 80-db noise at irregular intervals, in blasts of varying duration. All subjects were scored on a dependent variable "number of errors on a digit recall task."

The mean number of errors on the task was 4.0 for the "predictable noise" group and 8.0 for the "unpredictable noise" group. The obtained t for this difference was 2.37, significant at the 0.05 level (degrees of freedom were unspecified). The obtained significant difference suggests that some kinds of noise conditions do affect performance.

[1] J.M. Finkelman and D.C. Glass, "Reappraisal of the Relationship Between Noise and Human Performance by Means of a Subsidiary Task Measure," Journal of Applied Psychology 54(1970): 211-213.

Analysis of Variance in Research

Without question, the analysis of variance is the most popular statistical technique applied to controlled experiments in the behavioral sciences. A conceptual understanding of ANOVA principles is necessary if one hopes to comprehend and evaluate much of the current research literature in psychology, sociology, and education. An introduction to ANOVA, therefore, is an important part of a first course in applied statistics, even though very few students actually go on to perform experiments themselves.

A quick glance through publications such as the Journal of Experimental Psychology, Sociometry, the Journal of Applied Psychology, or any of the other current behavioral science journals will reveal that ANOVA results are not always presented in the same form. Frequently, you will find ANOVA tables that are abbreviated forms of the tables illustrated in the course. For instance, in reporting the results of an experiment designed to examine the effects of "Speaker Credibility" on "Persuasion," the experimenter might tell you only that "The effect associated with credibility was significant ($F = 34.7$, $df = 3.76$, $p < 0.001$)." Consider for a moment the information contained between the parentheses, and you will recognize that it contains all of the important information displayed in larger tables with SS, MS, df, and "Source of Variation" columns. Because you know that there are three degrees of freedom associated with the F-obtained numerator variance, you will infer there were four experimental groups, each of which received different levels of an independent variable called "Credibility." The error variance has 76 degrees of freedom associated with it. Without even reading the "Methods" section of the article, you should be able to calculate that each group had twenty subjects. (df error $= k(n-1)$). If df error $= 76$ and $k = 4$, then n must equal 20. What about the individual SS and MS values? It is true that you cannot recover them from the parenthetical information, but then it is unlikely that you would ever want to. It is the ratio of the two MS values as expressed by F-observed that helps you evaluate the experiment, not the separate SS and MS values. In the last decade, authors and editors have begun to realize that printing complete ANOVA summary tables is a waste of space, and current reports are likely to omit all but the essential information.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 393-394.

Does Party-Going Increase Smoking Behavior?

Many cigarette smokers report that they smoke more at parties or other social gatherings than they do otherwise. Is this true? If so, why?

A recent pair of studies by Brett Silverstein, Lynn Kozlowski and Stanley Schachter was conducted to address some of these questions.[1] In an earlier study, these researchers had demonstrated that manipulation of urinary pH (acidity) can influence the number of cigarettes people choose to smoke. In the studies discussed below, they investigated the possibility that some aspect of social situations raises urinary acidity, thereby producing more smoking.

In one study, eighteen smokers simply recorded the number of cigarettes smoked during the day; from these data, the mean number of cigarettes consumed each waking hour was determined for each person. Subjects also kept track of their social activities, and each smoker's days were individually categorized as "social" or "nonsocial", according to several criteria. For these subjects, the mean number of cigarettes smoked during the social day was 31.23, and the mean number during the nonsocial day was 27.85, a significant difference ($t = 2.71$, $df = 17$, $p < 0.02$). Because it was arguable that this difference arises from the simple fact that social days were generally longer than nonsocial days, the mean number of cigarettes smoked per hour on these days was also determined. On social days, subjects averaged 1.85 cigarettes an hour, and nonsocial days they averaged 1.73. This difference between means was almost, but not quite, significant at the 0.05 level ($t = 2.01$, $df = 17$, $p < 0.06$). From this result, the experimenters concluded that these people did smoke more when engaged in social activity. The next step was to examine the role of urinary pH in this situation.

For smokers, it was determined that urinary pH readings were lower at the end of social days (mean = 5.86) than at the end of nonsocial days (mean = 6.30); this difference was statistically significant ($t = 3.49$, $df = 15$, $p < 0.01$). But does this drop cause more cigarette smoking? In further tests with other subjects, pH reading dropped from a mean of 6.43 before a two-hour party to 6.00 after the party ($t = 3.23$, $df = 14$, $p < 0.01$). Moreover, this drop appeared for both smokers and nonsmokers, refuting the possible argument that smoking causes pH changes rather than the reverse. These results, along with the earlier evidence that pH manipulation alters smoking behavior, suggest a partial explanation for increased smoking at parties.

[1] B. Silverstein, L. Kozlowski, and S. Schachter, "Social Life, Cigarette Smoking, and Urinary pH," Journal of Experimental Psychology: General, 106 (1977): 20-23.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 320-321.

Measures of Association and Regression Analysis

"How did I get into this business? Well, I couldn't understand regression and correlation in college, so I settled for this instead."

Lyman Ott, Richard F. Larson, & William Mendenhall, Statistics (Boston: Duxbury Press), 1987: 430.

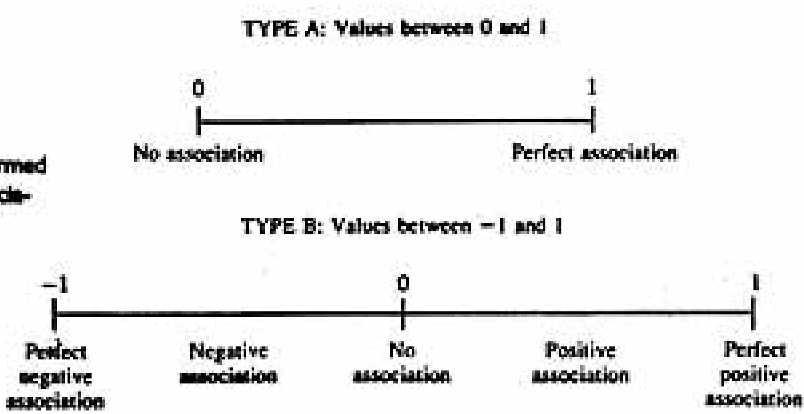
From the Literature

Howard Schuman and his colleagues reported on their investigation of the relationship between effort and grades in college (*Social Forces*, June 1985). The researchers recognized two widely held assumptions: (1) hard work along socially prescribed lines produces rewards and (2) the relationship between hard work and rewards is valid even for most researchers who emphasize the importance of other variables, such as natural ability and sheer luck.

Schuman and colleagues had as the overall focus of their study the test of the popular maxim "Genius is one percent inspiration and ninety-nine percent perspiration." To test this maxim, as well as a number of specific hypotheses, they drew upon the responses of a random sample of 522 students attending a major midwestern university. The interview questionnaire tapped such variables as hours studied, grade point average (GPA), and SAT scores.

GPA and SAT scores were validated later by obtaining access to official student records. The variable "hours studied" was divided into two components: hours studied during the week and hours studied on weekends. These two components were moderately related (r = .47). Small positive associations were found individually between GPA and hours studied (r = .11), as well as between GPA and the percentage of classes attended (r = .28). Taken as a group, classes attended, hours studied, and total SAT scores were found to provide a meaningful prediction of GPA.

The investigators also reported that performance on early tests was highly correlated with performance on later tests, and that later tests were even more highly correlated with the final grades earned in a class (r varied between .81 and .88). Other patterns were also reported. They included the finding that the total number of hours typically spent studying was related to the number of hours studied yesterday (r = .74), scores on the first test (r = .82), scores on the second test (r = .84), and scores on the third test (r = .86). Finally, SAT scores, by themselves, were related to GPA (r = .44).



Two types of normed measures of association

Surprising Correlations

Sometimes correlation coefficients only tell us what we already know by confirming the obvious. Just as often, however, unsuspected relations emerge from the data.

Take the relation between the effectiveness of teachers and the subjective ratings students give them. It is only reasonable to suspect that college students know a good teacher when they see one and, when asked to grade the performances of their instructors, assign the best grades to the best teachers. One recent study suggests, however, that just the opposite may be true. Roden and Rodin approached this issue by developing a highly objective measure of the amount of material learned from the instructor in an undergraduate calculus course.[1] They calculated the mean amount learned by each of twelve classes and then determined the mean rating each class assigned its teacher. The scatter plot in Figure 6.9 shows the surprising results: A negative correlation between subjective ratings of teachers and amount learned is clearly indicated by the slope of scatter-plot points, and the computed correlation coefficient for these data was $r = -0.75$. Rodin and Rodin appropriately point out that any explanation for the negative correlation offered at this point would be speculative, but they do interpret the results to indicate that "students are less than perfect judges of teaching effectiveness if the latter is measured by how much they have learned."[2]

Or, consider the relation that surely must exist between students' backgrounds in mathematics and the grades they receive in statistics courses. Undoubtedly, the student entering a college sophomore-level statistics course having already taken calculus and advanced algebra is better off than a student who has had only elementary algebra. But are the highest statistics grades received by those with the more extensive math backgrounds? The answer is, "That's an empirical question." This means that, strong as our intuitions on the matter may be, the only way to find out for sure is to examine the evidence.

Dr. Leonard Giambra addressed this question by examining the relation among mathematics backgrounds, overall grade-point averages, and grades received by 201 students in his statistics classes.[3] Surprisingly, he found no apparent relation between statistics grades and math backgrounds—students who had taken calculus did not, for example, receive a disproportionate percentage of the A's and B's. There was, however, some correlation (0.25) between statistics grades and overall grade-point averages, suggesting to Dr. Giambra that a student's grade in statistics depends more on his overall ability than on his math background.

The results from these two studies may or may not apply to situations beyond the immediate ones in which they were obtained. But they effectively illustrate the utility of correlational analysis in distinguishing between objective reality and cherished suppositions.

[1] M. Rodin and B. Rodin, "Student Evaluation of Teachers," Science 177 (1972): 1164-1166.
[2] Ibid., p. 1166.
[3] L. Giambra, "Mathematical Background and Grade-Point Average as Predictors of Course Grade in an Undergraduate Behavioral Statistics Course," American Psychologist 25 (1970): 366-367.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 175-177.

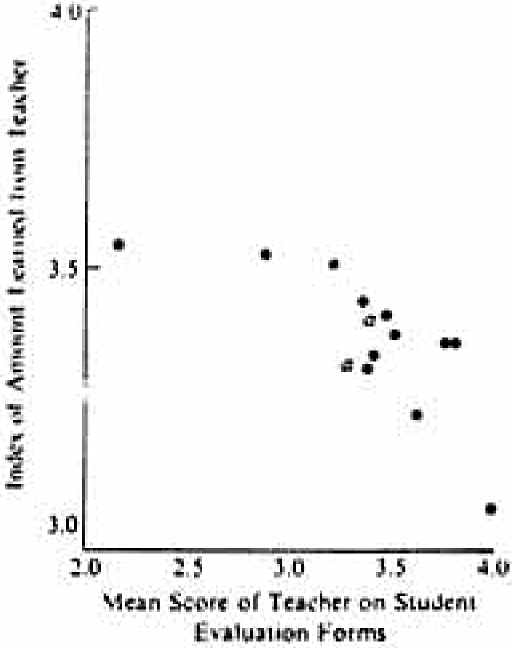


Figure 6.9 Relation between objective and subjective criteria of good teaching ($r = -0.75$). The points labeled *a* are for two sections taught by the same instructor.

Source: From M. J. Rodin and B. Rodin, "Student Evaluation of Teachers," *Science*, 177 (September 1972): 1164–1166. Reprinted by permission from Dr. Miriam J. Rodin and Dr. Burton Rodin and the American Association for the Advancement of Science. Copyright 1972 by the American Association for the Advancement of Science.

Measures of Association

There are many ways to measure association and each one may be interpreted differently. These comments amplify the discussion in the text and should help in interpreting these measures.

The measures most often used in sociology are *gamma, tau, rho, lambda*, and *r* (Pearson correlation coefficient). Most measures of association, and all of the ones just mentioned, have values between 0 and 1. Values close to 0 indicate a low relationship, those near 1 indicate a high relationship. In other words, the higher the value the greater the relationship between any two variables.

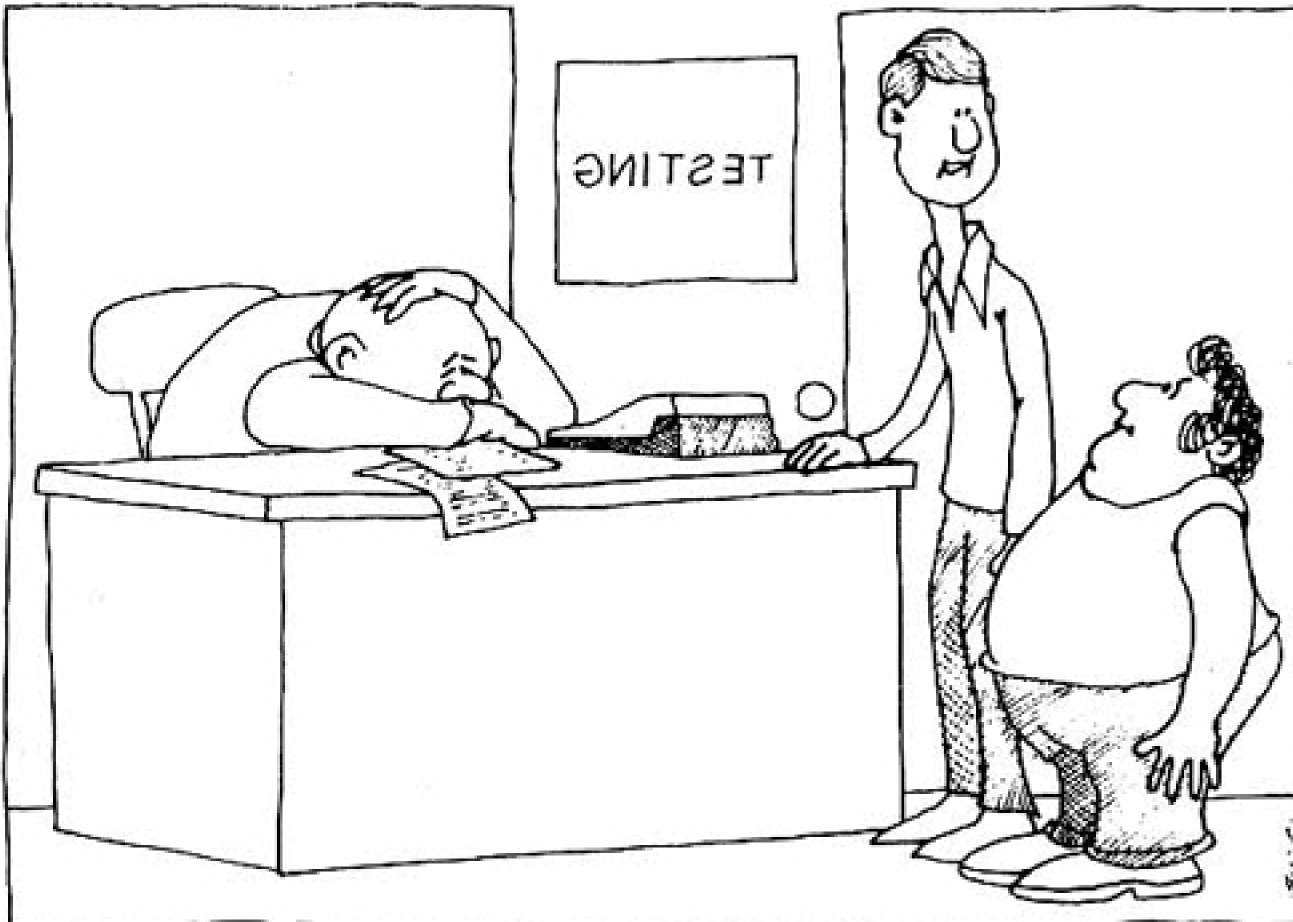
The direction of the relationship is also included in these statistics. Direction can be positive (direct) or negative (inverse). A positive relation obtains when one variable increases as the other variable increases, or when one decreases as the other decreases. The two variables move or change in the same direction. A perfect positive relationship equals +1.00. On the other hand, a negative relation obtains when one variable increases as the other decreases. A perfect negative relationship equals −1.00.

There is no one simple interpretation for the various statistics. However, for the most commonly used measure, *r*, the simplest interpretation is in terms of r^2, which is the variance explained. If our $r = .70$, $r^2 = .49$, which is the proportion of the variance explained by one variable on the other. With other measures, it is necessary to rely on the magnitude of the relationship in order to interpret it; the closer to 1 the particular measure is, the greater the relationship between them.

In order to make easier a decision about the correct interpretation of these measures (percentages, rates, and measures of association), sociologists frequently report a "*P* value." Generally at the bottom of a table you will see $P = .05$ (or some other value), or a statement like, "correlation significant at 0.05 level." This statistic is telling you that the correlation is not due to chance. Put another way, if the study were done over again, you would find approximately the same correlation.

In the social sciences, if a relation could occur at the .05 level, or .01 or .001, it is considered statistically significant. To say that a finding is statistically significant is only to say that the finding is very unlikely to be due to chance. To say that this is not a chance finding, however, is not to say what it is. And it should not be confused with being theoretically or practically signifitant. However, it does mean that whatever was found is likely what is there, that is, if you do the study again, you will likely find the same thing.

Robert Hagedorn, (General Editor), Sociology, Dubuque, Iowa: William C. Brown, 1983: 57.



"He says we've ruined his positive association between height and weight."

Darrell Huff, in How to Lie with Statistics,[1] reports there is a positive correlation between the ages of women and the angles of their feet in walking. Younger women tend to point their toes straight ahead; older women tend to walk with toes pointed out. Is it possible that the aging process causes women's feet to turn out? When one assumes, because two variables are correlated, that changes in one cause changes in the other, one has a post hoc error. The above relationship no more proves that aging causes feet to turn out than it proves the opposite--turning the feet out causes an increase in age.[2]

There are times, of course, when high correlations appear and no one is tempted to make the post hoc error. Reportedly, there are high correlations between (1) the sizes of schoolboys' feet and the quality of their handwriting;[3] (2) the number of storks' nests and the number of human births in northwest Europe;[4] and (3) the salaries of Presbyterian ministers in Massachusetts and the price of rum in Havana.[5] In some cases we can suggest a possible third factor that might account for the relation, but tend to be older than children with smaller feet, and handwriting probably improves with age. The number of storks' nests in Europe is related to the number of chimneys, which is related to the number of houses, which is related to the number of human births. The reason for relation 3 is anybody's guess.

There are other times, however, when we suspect that correlated variables might indeed represent a cause-and-effect relationship, but we cannot jump to that conclusion strictly on the basis of our correlational information. If, for instance, a medical study found that people who drink more coffee tend to have more coronary heart disease problems than people who drink less coffee, could we conclude that coffee drinking causes heart problems? It is possible that people who drink more coffee have, on the average, more sedentary jobs and thus get less daily exercise. Not until we can rule out all other possible causes can we begin to make cause-and-effect arguments based on correlational data. This is the problem faced by the research specialist in epidemiology (a science that attempts to understand and control disease in populations). Dr. Paul Milvey, a biophysicist who specializes in epidemiological problems, discusses some of the problems involved in demonstrating a causal link between physical exercise and risk of coronary heart disease (CHD):

> Most of the several hundred studies that deal with the relationship of physical activity to CHD or mortality from all diseases show the general picture of those people who are physically active tending to have a lower incidence of disease at any particular age than those who are not.
>
> The scientific and medical field pursuing these studies is called epidemiology. Epidemiologists seek to discover through statistical studies of man and his environment the cause or causes of pathology and disease. Whether the problem is cholera in 19th century England (solved), the Legionnaires' disease in Philadelphia last year (solved), the primary cause of lung cancer (solved), bladder cancer (unsolved) or CHD (unsolved), epidemiologists seek to demonstrate a cause-and-effect relationship between two entities.
>
> The criteria for demonstrating this relationship are quite specific and are difficult to satisfy. Sophisticated mathematical techniques often are employed so the epidemiologist can achieve valid conclusions. It's a difficult discipline because, while it's very easy to prove "association," it is infinitely more difficult to demonstrate "cause and effect."

For example, in the 1960s (Michael) Yudkin studied the relationship between dietary sugar and CHD and its final manifestation, the heart attack. He took very careful dietary histories of three groups of hospital patients: (1) patients hospitalized for any one of a large variety of reasons (broken legs to appendectomies); (2) patients suffering from CHD; and (3) patients who had had myocardial infarctions (heart attacks).

The amount of sugar each group consumed over the months and years prior to hospitalization was shown to be least for the "control" patients with the variety of diseases unrelated to their hearts and arteries, intermediate for the CHD patients and highest for the heart attack patients.

This study seemed strongly to indicate (nothing is ever quite proven in science) that sugar consumption caused or perhaps was one of several causing factors in the development of CHD and its final manifestation, the heart attack. But several investigators questioned this.

In the 1970s, two groups, working independently, showed that there was a better association or correlation between the development of this disease and the amount of the patients' smoking. And there was an even better correlation--and excellent correlation--between smoking and sugar consumption. (Those who smoked more also consumed more sugar.)

From a scientific and medical point of view, there is no obvious reason why sugar and CHD should have a causal relationship (fats, especially saturated fats, are quite probably related to CHD, but not sugar). But one can suggest any of a number of good, medically sound explanations or mechanisms by which smoking might cause CHD. And there is statistical evidence that also shows this. So we were fooled for a number of years by the earlier study which showed a correlation or association between CHD and sugar consumption, when in fact no causal relationship existed.

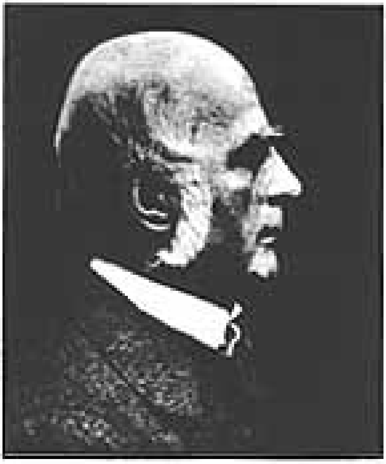
In this case, the epidemiologist calls sugar a "confounding variable." It goes along with the real causal variable. It's absent when the real cause is absent, it's there when the real cause is there, and it's a damned confusing problem to handle in all epidemiological studies.

Why are sugar and smoking consumption correlated? We can only speculate. We do know that the lower socio-economic groups in our country eat more sugar (junk food) than the higher socio-economic classes. Although I haven't checked it out, they almost certainly smoke more as well. So both smoking and sugar consumption may, in this light, be seen to be "caused" by one's socio-economic status in life.

We also know the CHD is higher in the less affluent classes. So perhaps the cause of CHD should be looked for in some common aspect of the environment or life style of the less-advantaged socio-economic classes: Do they go for checkups less frequently, do they smoke and consume fats or salt in larger amounts?

It's very difficult for the epidemiologist to pinpoint the environmental causes. . . .[6]

[1]Darrell Huff, How to Lie with Statistics (New York: W.W. Norton, 1954), Chapter 8.
[2]Huff suggests that older women were raised during a time when toeing out was encouraged; young women today are encouraged to walk with a different posture.
[3]W.A. Wallis and H.V. Roberts, The Nature of Statistics (New York: Free Press, 1962), p. 108.
[4]Ibid., p. 108.
[5]Huff., How to Lie with Statistics, p. 90.
[6]From P. Milvey, "Getting to the Heart," Runner's World, 12 (April 1977): 27-31.



SIR FRANCIS GALTON

The least-squares method will happily fit a straight line to any two-variable data. It is an old method, going back to the French mathematician Legendre in about 1805. Legendre invented least squares for use on data from astronomy and surveying. It was Sir Francis Galton (1822–1911), however, who turned "regression" into a general method for understanding relationships. He even invented the word.

Galton was one of the last gentleman scientists, an upper-class Englishman who studied medicine at Cambridge and explored Africa before turning to the study of heredity. He was well connected here also: Charles Darwin, who published *The Origin of Species* in 1859, was his cousin.

Galton was full of ideas but was no mathematician. He didn't even use least squares, preferring to avoid unpleasant computations. But Galton's ideas led eventually to the machinery for inference about regression that we will meet in this chapter. He asked: If people's heights are distributed normally in every generation, and height is inherited, what is the relationship between generations? He discovered a straight-line relationship between the heights of parent and child and found that tall parents tended to have children who were taller than average but less tall than their parents. He called this "regression toward mediocrity." Galton went further: he described inheritance by a straight-line relationship with responses y that have a normal distribution about the line for every fixed input x. This is the model for regression we use in this chapter.

It Tastes Good, But Is It Good for You?

Sometimes regression lines are useful for purposes other than for predicting future events from past events. The coefficients that define the line--slope and intercept--are sometimes valuable in their own right for analyzing characteristics of a linear relation between two variables.

Consider, for instance, a problem faced by two General Food Corporation researchers, Elisabeth Street and Mavis Carroll. They were involved in an attempt to develop an "easy-to-prepare, nutritious, on-the-run meal," code named H.[1] In the laboratory it is quite easy to ensure that such concoctions contain certain nutrients, but there is little a prior insurance that humans will metabolize those nutrients efficiently; there is even less assurance that they will find them palatable. Such factors must be determined empirically by testing the finished product on living subjects.

A rigorously controlled test using human subjects indicated that, indeed, H was as tasty as another product C (so called because it has a casein base), which was already on the market. The researchers were also interested, however, in finding whether or not the protein content of H would be metabolized efficiently under conditions of actual use. They approached this question in another carefully controlled study using rats as subjects. Rats--rather than humans--were used in this study because the metabolic processes involved in protein utilization are fairly similar in rats and humans and because the relative efficiency of that utilization is more clearly and quickly reflected in the animals' body weights.

Thirty rats were randomly divided into three groups of ten each; each group contained animals of comparable weights, that is, all groups had equal-age light, medium, and heavy rats.[2] All animals were individually weighed before the experiment, and then each group was given a different diet for 28 days. One group was fed only on a liquid version of H, another group received only a solid version of H, and the third (control) group was fed on casein. Each of the three diets contained about 9 percent protein by weight. During this experimental period the animals were allowed to eat as much as they wished of the designated food, and the food intake of each animal was carefully monitored.

At the end of the 28-day feeding period, the researchers had two measures on each animal: body weight gain (in grams) over the 28 days, and the 28-day protein intake. These data constitute paired observations on two variables for each rat and are displayed in the scatter-plot diagram shown in Figure 7.8. Each dot or cross thus represents data from one animal.

Figure 7.8 shows clearly that animals on liquid and solid H took in more protein and gained more weight than animals on the casein diet. The researchers reasoned, however, that more analysis of these data was needed in order to answer the question of whether or not H protein was used more efficiently than casein protein. It is possible, after all, that H was simply more tasty than casein. Casual inspection of Figure 7.8 does not indicate whether or not a gram of H protein can be expected to produce a greater weight gain than a gram of casein protein. So, a separate regression line was computed for each group's paired observations. These are shown (without data points) in Figure 7.9.

Notice that the H diets produced steeper regression lines than did the casein diet. The calculated slopes, b, of the liquid and solid H groups were respectively, 3.72 and 3.66, while b for the casein group was 2.91. So, the researchers concluded, differences in weight gain between H and casein groups were not due entirely to differences in protein intake. The differences in b values indicated that "For a given increase in protein intake, the H diets resulted in a greater increase in weight gain than did the casein diet."[3]

[1]E. Street and M.B. Carroll, "Preliminary Evaluation of a New Food Product," In J. Tanur et al. (eds.), Statistics: A Güide to the Unknown (San Francisco: Holden-Day, 1972), pp. 220-228.
[2]The procedure used here was similar to the matched pairs experimental design. Here, matched trios of rats were arranged before the experiment; one member of each trio was randomly assigned to each group.
[3]Street and Carroll, op. cit., p. 224.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 198-201.

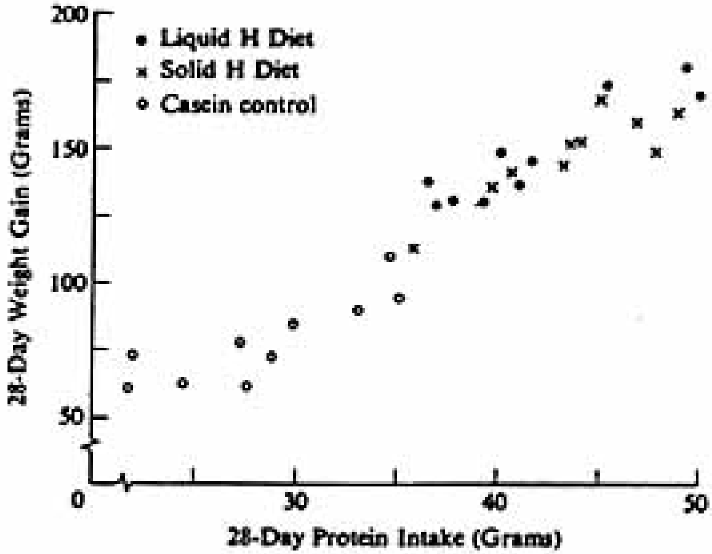


Figure 7.8 Relationship of 28-day protein intake and weight gain in young male rats.

Source: Figures 7.8 and 7.9 are taken from E. Street and M. B. Carroll, "Preliminary Evaluation of a New Food Product," in J. Tanur et al. (eds.), *Statistics: A Guide to the Unknown* (San Francisco: Holden-Day, 1972), p. 223. Reprinted by permission.

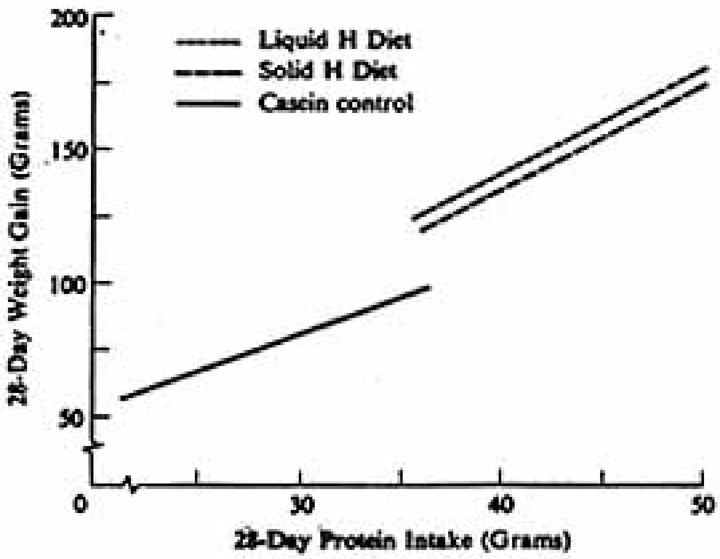


Figure 7.9 Estimated regression of 28-day weight gain on protein intake for young male rats.

"It was a numbers explosion."

SCRUTINIZING POPULAR REPORTS OF SOCIAL SCIENCE RESEARCH

We are being bombarded daily with such a mass of new information that it is difficult to process it adequately. It is increasingly important to become a critical, selective, and informed consumer of information. Discussed below are several means for better evaluating reports on social science research that you may encounter in the media.

Maintain a Skeptical Attitude
Be skeptical, because the media have a tendency to sensationalize and distort. For example, the media may report that a university researcher spent $500,000 to find out that love keeps families together when, in fact, this was only one small aspect of the larger research project. Moreover, chances are the media have oversimplified even this part of the researcher's conclusions.

Consider the Source of Information
It is important to know, for example, whether a study on the relationship between cancer and smoking has been sponsored by the tobacco industry or by the American Cancer Society. On a 1985 "60 Minutes" segment, a representative of a tobacco company denied the existence of any research linking throat and mouth cancer with dipping snuff. A medical researcher contended that putting a "pinch between your cheek and gum" has, in the long run, led to cancer in humans. Whom do you believe? At the very least you want to know the background of the source of information before making a judgment about scientific conclusions.

Determine Whether a Control Group Has Been Used
Knowing whether a control group has been used in the research may be important. For instance, increases in self-esteem and physical energy may be reported in a study of participants in a meditation program. Was this because of the respect and attention they were given during the training period or because of the meditation techniques themselves? Or a study may report that the productivity of a group of workers in an office increased dramatically because the workers were allowed to participate in work-related decisions. Was the productivity increase due to the employees' being involved in something new and exciting or because of the participation in decision making itself? Without one or more control groups, you cannot be certain of what caused the changes in the meditation participants or in the office workers.

Do Not Mistake Correlation for Causation
A correlation between two variables does not necessarily mean that one caused the other. For example, at one time the percentage of Americans who smoked was increasing at the same time life expectancy was increasing. Did this mean that smoking caused people to live longer? Actually, a third factor—improved health care—accounts for the increased life expectancy. Do not assume that two events are related just because they occur together.

STATISTICAL DESIGN AND THE EQUINE QUADRUPED

The genesis of modern-day statistics seems to be accounted for, in great part, by the efforts of an English gentleman named R.A. Fisher. Much of Fisher's research was done during the twenties and before, with the beginning of his scholarly publications in the early 1930's. Fisher's work was with agricultural experimentation, which ultimately led to blocking theory and split-plot statistical design. It is his agricultural research that was most interesting to the author, and the remainder of the paper will be devoted to a systematic and logical analysis of the genesis of statistical design as it relates to the equine quadruped.

Fisher's early work was done with garden plots and was initially as much an interest in flowers and truck products as it was with statistics, but being a mathematician he became interested in statistical design as a means of recording, controlling, and analyzing the results of his many and varied experiments.

In addition to experimental design as a means of controlling variables, Fisher exercised other controls with regard to plant growth. One of the major variables was the use of fertilizer, about which the author will take the liberty of making some assumptions.

Much of Fisher's experimentation took place during the twenties, perhaps some before then, and on in to the thirties. Since this period of time was prior to the widespread use of synthetic and chemically contrived fertilizers, the author will assume that natural fertilizers were still in wide use and that, further, Mr. Fisher was an ardent user of said natural fertilizers as a means of achieving the "big" effect on the experimental variable.

Further, that one of the major animals of England has, and will always be, the equine quadruped. Said quadruped had long been a mainstay in English society and, yea, even government when we consider the effect of the horse on extending the power and stature of the medieval knight in defense of whatever ruler. Since the ubiquitous quadruped was so prevalent, approached only in numbers perhaps by the lowly bovine quadruped and the chicken, the author will assume that the majority of natural fertilizer was produced by the equine quadruped. That being the case, Mr. Fisher must have been an ardent user of the fecal matter of the equine quadruped as an experimental variable; there seems to be little current question on this point.

What is truly amazing, in the opinion of the author, is that we have come so far in the field of chemical fertilizers only to find the field of statistics still grossly encumbered and shot through and through with horseshit.

SOURCE: Harper's Magazine, January, 1974, p. 5.