Menu

Table Of Contents	CPV: pp. 1-23	CPV: pp. 24-25
CPV: pp. 26-30	CPV: pp. 31-33	CPV: pp. 34-36
CPV: pp. 37-52	CPV: pp. 53-61	CPV: pp. 62-74

Table of Contents

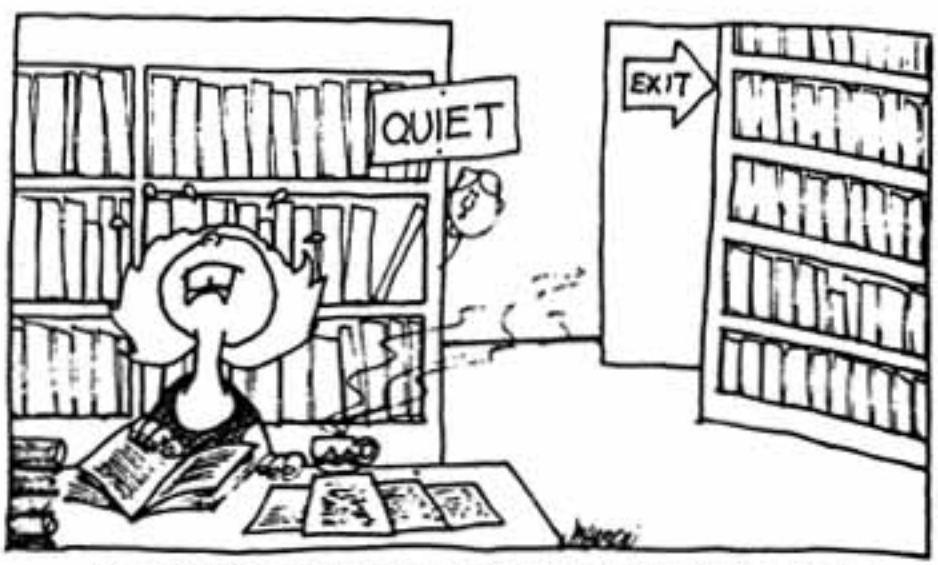
The Gambler's Fallacy

"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 while 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 Titanic or the explosion of the Hindenburg, were indeed very unlikely events. However, upon reflection you should be able to see that any specific 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."

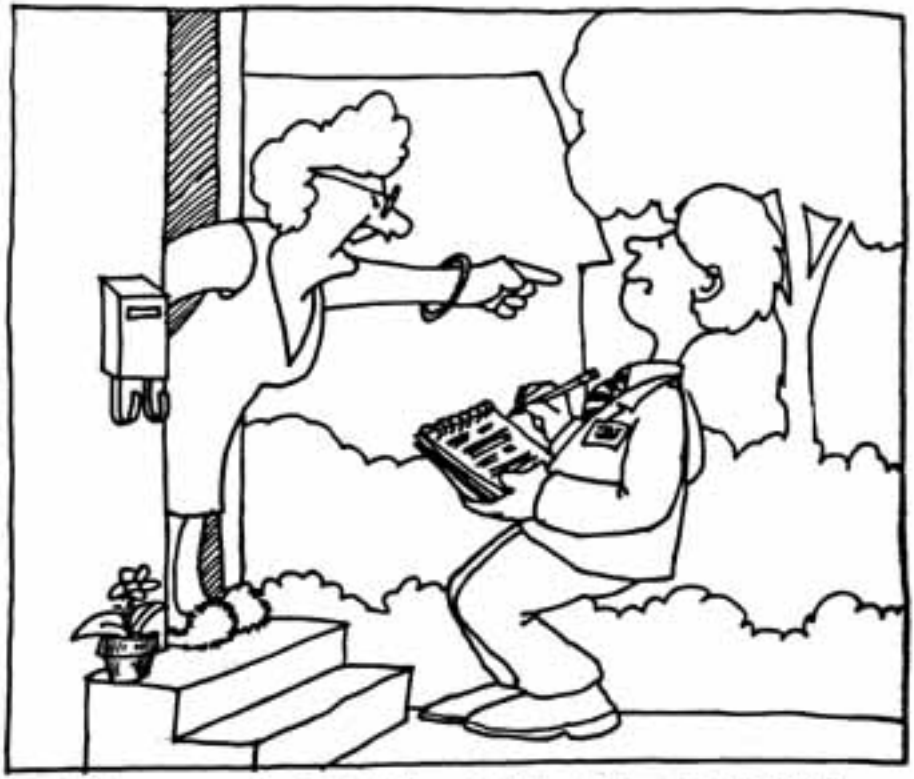
Improbable Probability. To explain why, Judge Raymond Sullivan attached a four-page appendix to his opinion that carried the necessary math far beyond the relatively simple formula of probability. Judge Sullivan was willing to assume it was unlikely that such a couple as the one described existed. But 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 <u>might</u> 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 <u>future</u> or <u>potential</u> 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 <u>make sure that in his experiment only random factors determine which unit gets which experimental treatment</u>. 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).

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.

A Scheme for Visualizing the Nature of Statistical Estimation

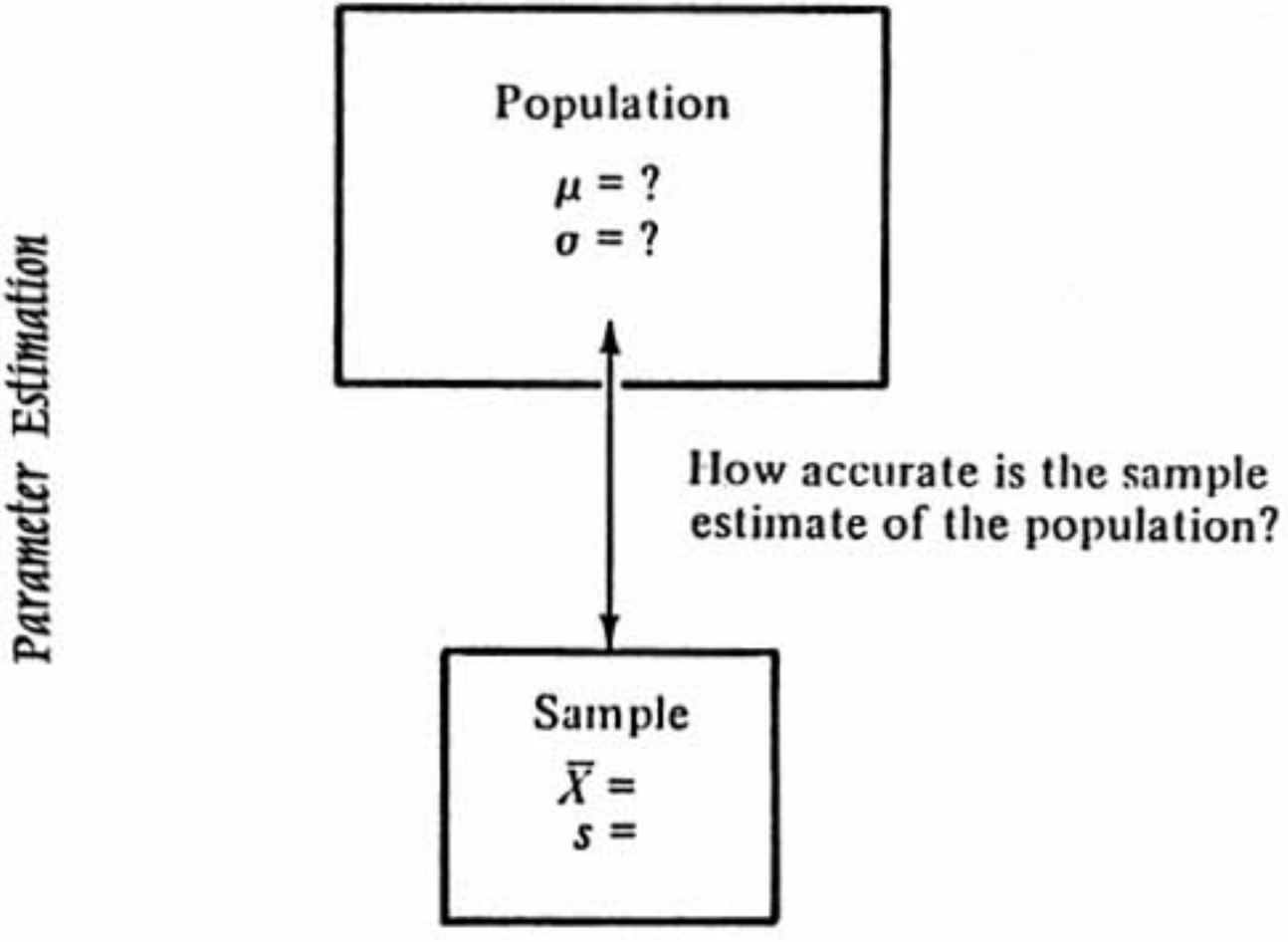


Photo by G. Paul Bishop, courtesy of Paul Bishop, Jr., Berkeley, Ca.

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), 19$_{87}$), 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.

The Urban Institute, Washington, D.C.

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.