Reference Guide on Statistics

David H. Kaye
David A. Freedman

David H. Kaye, A.M., J.D., is Professor of Law, Arizona State University College of Law, Tempe, Arizona. **David A. Freedman,** Ph.D., is Professor of Statistics, University of California, Berkeley, California.

This page left blank intentionally for proper pagination when printing two-sided

Contents

I. Introduction

Statistics, broadly defined, is the science and art of gaining information from data. For statistical purposes, *data* mean observations or measurements, expressed as numbers. A *statistic* may refer to a particular numerical value, derived from the data. Baseball statistics, for example, is the study of data about the game; a player's batting average is a statistic.

The field of statistics includes methods for (1) collecting data, (2) analyzing data, and (3) drawing inferences from data. This reference guide describes the underlying ideas of statistics as they relate to legal proceedings. Statistical assessments figure prominently in antitrust, discrimination, fraud, homicide, sexual assault, trademark, toxic tort, and many other kinds of cases.[1] Typically, the most difficult arguments about such studies concern their probative value.[2]

This reference guide focuses on the nature of statistical thinking rather than on the rules of evidence or substantive legal doctrine. We hope that the explanations provided, although summary and nonmathematical in form, will permit judges who are confronted with statistical testimony to understand more of the

[1] *See generally* David C. Baldus & James W. L. Cole, Statistical Proof of Discrimination (1980); Statistics and the Law (Morris H. DeGroot et al. eds., 1986); The Evolving Role of Statistical Assessments as Evidence in the Courts (Stephen E. Fienberg ed., 1989); Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers (1990); 1 & 2 Joseph L. Gastwirth, Statistical Reasoning in Law and Public Policy (1988).

[2] Statistical studies suitably designed to address a material issue generally will be admissible under the Federal Rules of Evidence. The hearsay rule rarely is a serious barrier to the presentation of statistical studies. *See* Linda A. Bailey et al., Reference Guide on Epidemiology n.1, in this manual. Likewise, since most statistical methods relied on in court are described in textbooks and journal articles and are capable of producing useful results when carefully and appropriately applied, the methodology generally satisfies the "scientific knowledge" requirement articulated in Daubert v. Merrell Dow Pharmaceuticals, Inc., 113 S. Ct. 2786, 2795 (1993). For a discussion of the implications and scope of *Daubert* generally, see Margaret A. Berger, Evidentiary Framework §§ I, III, in this manual; Bert Black et al., *Science and the Law in the Wake of Daubert: A New Search for Scientific Knowledge,* 72 Tex. L. Rev. 715 (1994); Richard D. Friedman, *The Death and Transfiguration of Frye* , 34 Jurimetrics J. 133 (1994); Susan R. Poulter, *Daubert and Scientific Evidence: Assessing Evidentiary Reliability in Toxic Tort Cases,* 1993 Utah L. Rev. 1307; Symposium, *Scientific Evidence After the Death of Frye,* 15 Cardozo L. Rev. 1745 (1994). Of course, a particular study may use a method that is entirely appropriate for some problems, but that is so poorly executed that it should be inadmissible under Fed. R. Evid. 403 and 702. Or, the method may be inappropriate for the problem at hand and thus lack the "fit" spoken of in *Daubert*. 113 S. Ct. at 2796. Or, the study may rest on data of the type not reasonably relied on by statisticians or substantive experts and hence run afoul of Fed. R. Evid. 703. *See, e.g*., Faust F. Rossi, Expert Witnesses 43–98 (1991); Ronald L. Carlson, *Policing the Bases of Modern Expert Testimony* , 39 Vand. L. Rev. 577 (1986); Paul R. Rice, *Inadmissible Evidence as a Basis for Expert Opinion Testimony: A Response to Professor Carlson* , 40 Vand. L. Rev. 583 (1987); Michael C. McCarthy, Note, *"Helpful" or "Reasonably Reliable"?: Analyzing the Expert Witness's Methodology Under Federal Rules of Evidence 702 and 703* , 77 Cornell L. Rev. 350 (1992). More often, however, the battle over statistical evidence concerns weight or sufficiency rather than admissibility.

terminology, to place the evidence in context, to appreciate its strengths and weaknesses, and to develop and apply legal doctrine governing the use of statistical evidence.

The reference guide is organized as follows:

- Section I provides an overview of the field and offers some suggestions about procedures to encourage the best use of statistical expertise in litigation.
- Section II addresses data collection. The design of a study is the most important determinant of its quality. Section II describes the design of surveys, controlled experiments, and observational studies. It indicates when these procedures are likely to produce useful data for various purposes.
- Section III discusses methods for extracting and summarizing the most important features of data. *Descriptive statistics* is the art of describing and summarizing data, and section III considers the meaning, useful ness, and limitations of such descriptive statistics as the mean, median, standard deviation, correlation coefficient, and slope of a regression line. These are the basic descriptive statistics, and most statistical analyses seen in court use them as building blocks.
- Section IV describes the logic of statistical inference, emphasizing its foundations and limitations. In particular, it explains statistical estimation, standard errors, confidence intervals, *p*-values, and hypothesis tests.

A. Varieties and Limits of Statistical Expertise

For convenience, the field of statistics may be divided into three subfields: probability, theoretical statistics, and applied statistics. Theoretical statistics is the study of the mathematical properties of statistical procedures; probability theory plays a key role in this endeavor. Results may be used by applied statisticians who specialize in particular types of data collection, such as survey research, or in particular types of analysis, such as multivariate methods.

Statistical expertise is not confined to those with degrees in statistics. Because statistical reasoning underlies all empirical research, researchers in many fields are exposed to statistical ideas. Experts with advanced degrees in the physical, medical, and social sciences and some of the humanities may receive formal training in statistics. Such specializations as biostatistics, epidemiology, econometrics, and psychometrics are primarily statistical, with an emphasis on methods and problems most important to the related substantive discipline.

Experience with applied statistics is the best indication of the type of statistical expertise needed in court. By and large, individuals who think of themselves as specialists in using statistical methods—and whose professional careers demonstrate this orientation—are most likely to apply appropriate procedures and correctly interpret the results.[3] At the same time, the choice of which data to examine or how best to model a particular process may require subject matter expertise that a statistician may lack. Statisticians typically advise experts in substantive fields on the procedures for collecting data and usually analyze data collected by others. As a result, cases involving statistical evidence often are (or should be) "two-expert" cases of interlocking testimony.[4] A labor economist, for example, may supply a definition of the relevant labor market from which an employer draws its employees, and the statistical expert may contrast the racial makeup of those hired to the racial composition of the labor market. Naturally, the value of the statistical analysis depends on the substantive economic knowledge that informs it.[5]

B. Procedures That Enhance Statistical Testimony

1. Maintaining professional autonomy

Ideally, experts who conduct research for litigants should proceed with the same objectivity that they would apply in other contexts. Thus, if experts testify or if their results are used in testimony by others, they should be free to do whatever analysis and have access to whatever data are required to address the problems

[3] Forensic scientists and technicians often testify to probabilities or statistics derived from studies or databases compiled by others, even though some of these experts lack the training or knowledge required to understand and apply the information. *See* Andre A. Moenssens, *Foreword: Novel Scientific Evidence in Criminal Cases: Some Words of Caution*, 84 J. Crim. L. & Criminology 1, 19 (1993) ("Most forensic experts who use . . . [probability and] statistics have no idea of how the calculations were made, and are not statisticians themselves."). We believe that courts should be more discerning in assessing the qualifications of these experts to opine on matters that they cannot explain adequately. *See* Paul C. Giannelli, *Expert Testimony and the Confrontation Clause*, 22 Cap. U. L. Rev. 45 (1993). State v. Garrison, 585 P.2d 563 (Ariz. 1978), illustrates the problem. In a murder prosecution involving bite mark evidence, a dentist was allowed to testify that "the probability factor of two sets of teeth being identical in a case similar to this is, approximately, eight in one million," even though "he was unaware of the formula utilized to arrive at that figure other than that it was 'computerized.'" *Id*. at 566, 568.

[4] Sometimes a single witness presents both the substantive underpinnings and the statistical analysis. Ideally, such a witness has extensive expertise in both fields, although less may suffice to qualify the witness under Fed. R. Evid. 702. In deciding whether a witness who clearly is qualified in one field may testify in a related area, courts should recognize that qualifications in one field do not necessarily imply qualifications in the other. *See, e.g.*, United States *ex rel.* DiGiacomo v. Franzen, 680 F.2d 515, 516 (7th Cir. 1982) (state criminalist testified not only to her finding matching hairs but also to a study that she vaguely recalled on the probability of coincidental matches); Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224, 286 (N.D. Tex. 1980) (plaintiffs' expert "is an impressive expert on statistics, but not on compensation or other personnel practices"), *modified in part*, 521 F. Supp. 656 (N.D. Tex. 1981), *vacated,* 723 F.2d 1195 (5th Cir.), *cert. denied*, 469 U.S. 1073 (1984).

[5] In *Vuyanich,* 505 F. Supp. at 319, defendant's statistical expert criticized the plaintiffs' statistical model for an implicit, but restrictive, assumption about male and female salaries. The district court accepted the model because the plaintiffs' expert had a "very strong guess" about the assumption, and her expertise included labor economics as well as statistics. *Id*. It is doubtful, however, that economic knowledge sheds much light on the assumption, and it would have been simple to perform a less restrictive analysis. In this case, the court may have been overly impressed with a single expert who combined substantive and statistical expertise. Once the issue is defined by legal and substantive knowledge, some aspects of the statistical analysis will turn on statistical considerations alone, and expertise in another subject will not be pertinent.

the litigation poses in a professionally responsible fashion.[6] Questions about the freedom of inquiry accorded to testifying experts and the scope and depth of experts' investigations may reveal the experts' approach to acquiring and extracting relevant information.

2. Disclosing other analyses

Statisticians may analyze data using a variety of statistical models and methods. There is nothing underhanded in, and much to be said for, looking at the data in a variety of ways. To permit a fair evaluation of the analysis that the statistician may settle on, however, the testifying expert should explain the history behind the development of the final statistical approach.[7]

3. Disclosing data and analytical methods before trial

The collection of data often is expensive, and data sets typically contain at least some minor errors or omissions. Careful exploration of alternative modes of analysis also can be expensive and time-consuming. To minimize the occurrence of distracting debates at trial over the accuracy of data and the choice of analytical techniques, and to permit informed expert discussions of method, pretrial procedures should be used, particularly with respect to the accuracy and scope of the data, and to discover the methods of analysis.[8] Suggested procedures along these lines are available elsewhere.[9]

[6] *See* The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 164 (recommending that the expert be free to consult with colleagues who have not been retained by any party to the litigation and that the expert receive a letter of engagement providing for these and other safeguards).

[7] *See, e.g*., Mikel Aickin, *Issues and Methods in Discrimination Statistics, in* Statistical Methods in Discrimination Litigation 159 (David H. Kaye & Mikel Aickin eds., 1986). Some commentators have urged that counsel who know of other data samples or analyses that do not support the client's position should reveal this fact to the court rather than attempt to mislead the court by presenting only favorable results. The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 167; *cf*. William W Schwarzer, *In Defense of "Automatic Disclosure in Discovery,"* 27 Ga. L. Rev. 655, 658–59 (1993) ("[T]he lawyer owes a duty to the court to make disclosure of core information."). The Panel on Statistical Assessments as Evidence in the Courts also recommends that "if a party gives statistical data to different experts for competing analyses, that fact be disclosed to the testifying expert, if any." The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 167. Whether and under what circumstances a particular statistical analysis might be so imbued with counsel's thoughts and theories of the case that it should receive protection as the attorney's work product is an issue beyond the scope of this reference guide.

[8] *See* Fed. R. Civ. P. 16(c), 26(a)(2)(B) (Supp. 1993); Black et al., *supra* note 2, at 791. We also think that a pretrial procedure used in England deserves consideration. In most cases, Order 38, Rule 37, like Fed. R. Civ. P. 26(a)(2)(B), demands that an expert produce a written report before trial. Evidence Rules, S.I. 1989, No. 2427, *reprinted in* The Supreme Court Practice 83 (6th Cum. Supp. 1988). But Order 38, Rule 38 goes beyond the Federal Rules in explicitly authorizing the judge to require the experts to participate in the pretrial identification of disputed issues. Evidence Rules, S.I. 1987, No. 1423, *reprinted in* The Supreme Court Practice, *supra* at 83–84. This rule allows the court to

> direct that there be a meeting "without prejudice" of such experts . . . for the purpose of identifying those parts of their evidence which are in issue. Where such a meeting takes place the experts may prepare a joint statement indicating those parts of their evidence on which they are, and those on which they are not, in agreement. *Id.*

[9] *See* The Special Comm. on Empirical Data in Legal Decision Making, Recommendations on Pretrial Proceedings in Cases with Voluminous Data, *reprinted in* The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, app. F. When the parties are alerted before trial to the criticisms of their

4. Presenting expert statistical testimony

The most common format for the presentation of evidence at trial is sequential. The plaintiff's witnesses are called first, one by one, without interruption except for cross-examination, and testimony is in response to specific questions rather than by an extended narration. Although traditional, this structure is not compelled by the Federal Rules of Evidence.[10] Some alternatives have been proposed that might be more effective in cases involving substantial statistical testimony. For example, when the reports of witnesses go together, the judge might allow their presentations to be combined and the witnesses to be questioned as a panel rather than sequentially. More narrative testimony might be allowed, and the expert might be permitted to give a brief tutorial on statistics as a preliminary to some testimony. Instead of allowing the parties to present their experts in the midst of all the other evidence, the judge might call for the experts for opposing sides to testify at about the same time. Some courts, particularly in bench trials, may have both experts placed under oath and, in effect, permit them to engage in a dialogue. In such a format, experts are able to say whether they agree or disagree on specific issues. The judge and counsel can interject questions. Such practices may improve the judge's understanding and reduce the tensions associated with the experts' adversarial role.[11]

data or analyses, it may be possible to determine whether the putative problems have much effect on the results. *E.g.*, David H. Kaye, *Improving Legal Statistics*, 24 Law & Soc'y Rev. 1255 (1990).

[10] *See* Fed. R. Evid. 611.

[11] The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 174.

This page left blank intentionally for proper pagination when printing two-sided

II. How Have the Data Been Collected?

An analysis is only as good as the data on which it rests. Along with an examination of the statistical analysis, therefore, it is important to verify the quality of the data collection[12] and to identify its limitations.

A. Individual Measurements

1. Is the measurement process reliable?

In science, *reliability* refers to reproducibility of results.[13] A reliable measuring instrument returns consistent measurements of the same quantity. A scale, for example, is reliable if it reports the same weight for the same object time and again. It may not be accurate—it may always report a weight that is too high or one that is too low—but the perfectly reliable scale always reports the same weight for the same object. Its errors, if any, are systematic; they always point in the same direction.

Reliability can be ascertained by repeatedly measuring the same quantity. The predominant method of DNA identification, for instance, requires laboratories to determine the molecular weight of fragments of DNA. By making duplicate measurements of the same fragments, laboratories can determine the likelihood that two measurements of the same fragment will differ by a specified amount.[14] Ascertaining the usual range of such random error is essential in deciding whether an observed discrepancy between a crime sample and a suspect's

[12] For introductory treatments of data collection, see, *e.g.*, Stephen K. Campbell, Flaws and Fallacies in Statistical Thinking (1974); David Freedman et al., Statistics (2d ed. 1991); Darrell Huff, How to Lie with Statistics (1954); Jeffrey Katzer et al., Evaluating Information: A Guide for Users of Social Science Research (2d ed. 1982); David S. Moore, Statistics: Concepts and Controversies (2d ed. 1985); Robert S. Reichard, The Figure Finaglers (1974); Richard P. Runyon, Winning with Statistics: A Painless First Look at Numbers, Ratios, Percentages, Means, and Inference (1977); Hans Zeisel, Say It with Figures (6th ed. 1985).

[13] Courts often use *reliable* to mean "that which can be relied on" for some purpose, such as establishing probable cause or crediting a hearsay statement when the declarant is not produced for confrontation. Daubert v. Merrell Dow Pharmaceuticals, Inc., 113 S. Ct. 2786, 2795 n.9 (1993), for instance, distinguishes "evidentiary reliability," or "trustworthiness," from "scientific reliability," or "consistent results." Here, we use the term in the latter sense to clarify the many components of ultimate trustworthiness.

[14] *See, e.g.*, B. Budowle et al., *Fixed-Bin Analysis for Statistical Evaluation of Continuous Distributions of Allelic Data from VNTR Loci, for Use in Forensic Comparisons,* 48 Am. J. Hum. Genet. 841 (1991); B. S. Weir & B. S. Gaut, *Matching and Binning DNA Fragments in Forensic Science*, 34 Jurimetrics J. 9 (1993).

sample is sufficient to exclude the suspect as a possible source of the crime sample.[15]

In some social science studies, researchers examine recorded information and characterize it. For instance, in a study of death sentencing in Georgia, legally trained evaluators examined short summaries of cases and ranked them according to the defendant's culpability.[16] Two different aspects of reliability are worth considering. First, the "within-observer" variability of judgments should be small—the same evaluator should rate essentially identical cases the same way. Second, the "between-observer" variability should be small—different evaluators should rate the same cases the same way.

2. Is the measurement process valid?

Reliability is necessary, but not sufficient, to ensure accuracy. In addition to reliability, *validity* is needed. A valid measuring instrument measures what it is supposed to. Thus, a polygraph measures certain physiological responses to stimuli. It may accomplish this task reliably. Nevertheless, it is not valid as a lie detector unless increases in pulse rate, blood pressure, and the like are well correlated with conscious deception.

When an independent and highly accurate way of measuring the *variable*[17] of interest is available, it may be used to validate the measuring system in question. Breathalyzer readings may be validated against alcohol levels found in blood samples. Employment test scores may be validated against job performance. A common measure of validity is the *correlation coefficient* between the *criterion* (job performance) and the predictor (the test score).[18]

3. Are the measurements recorded correctly?

Judging the adequacy of data collection may involve examining the process by which measurements are recorded and preserved. Are responses to interviews coded and logged correctly? Are all the responses to a survey included? If gaps or mistakes are present, do they appear randomly so they do not distort the results?

Once it is shown that measurements are reliable, valid, and properly recorded, inferences can be made. The purpose of collecting and analyzing the data may be to describe something, such as the prevalence of a blood type, or it

[15] Committee on DNA Technology in Forensic Science, National Research Council, DNA Technology in Forensic Science 61–62 (1992).

[16] David C. Baldus et al., Equal Justice and the Death Penalty: A Legal and Empirical Analysis 49–50 (1990).

[17] For present purposes, a variable is a numerical characteristic of units in a study. For instance, in a survey of people, the unit of analysis is the person, and variables might include income (in dollars per year) and educational level (years of schooling completed). In a study of school districts, the unit of analysis is the district, and variables might include average family income of residents and average test scores of students.

[18] *E.g.*, Washington v. Davis, 426 U.S. 229, 252 (1976); Albemarle Paper Co. v. Moody, 422 U.S. 405, 430–32 (1975). *See* discussion of the correlation coefficient *infra* § III.F.2. Various statistics are used to characterize the reliability of laboratory instruments, psychological tests, or human judgments. These include the *standard deviation* (SD) as well as the correlation coefficient. *See infra* § III.E.

may be to investigate a question of cause and effect, such as the deterrent effect of capital punishment.

B. Descriptive Surveys and Censuses

A census measures some characteristic of every unit in a *population* of individuals or objects. A survey, alternatively, measures characteristics only in part of a population. The accuracy of the information collected in a census or survey depends on how the units are selected, which units are actually measured, and how the measurements are made.

1. What method is used to select the units to be measured?

By definition, a census seeks to measure every unit in the population. It may fall short of this goal, in which case the question must be asked whether the missing data are likely to differ in some systematic way from the data that are collected. The U.S. Bureau of the Census estimates that the past six censuses failed to count everyone, and there is evidence that the undercount is greater in certain subgroups of the population. Supplemental studies may enable statisticians to adjust for such omissions, but the adjustments may rest on uncertain assumptions.[19]

The methodological framework of a scientific survey is more complicated than that of a census. In surveys that use random sampling methods, a *sampling frame*, that is, an explicit list of units in the population, is created. Individual units then are selected by a kind of lottery procedure, and measurements are made on these sampled units. For example, a defendant charged with a notorious crime who seeks a change of venue may commission an opinion poll to show that popular opinion is so adverse and deep-rooted that it will be difficult to impanel an unbiased jury. The population consists of all persons in the jurisdiction who might be called for jury duty. A sampling frame here could be the list of these persons as maintained by appropriate officials.[20] In this case, the fit between the sampling frame and the population would be excellent.

[19] For conflicting views on proposed adjustments to the 1990 census, see Stephen E. Fienberg, *The New York City Census Adjustment Trial: Witness for the Plaintiffs*, 34 Jurimetrics J. 65 (1993); David A. Freedman, *Adjusting the Census of 1990,* 34 Jurimetrics J. 99 (1993); John E. Rolph, *The Census Adjustment Trial: Reflections of a Witness for the Plaintiffs* , 34 Jurimetrics J. 85 (1993); Kenneth W. Wachter, *The Census Adjustment Trial: An Exchange,* 34 Jurimetrics J. 107 (1993). *See also* Symposium , *Undercount in the 1990 Census* , 88 J. Am. Stat. Ass'n 1044 (1993). Similarly, the courts are divided over the legal standard governing claims that adjustment is statutorily or constitutionally compelled. *Compare* New York City v. United States Dep't of Commerce, 63 U.S.L.W. 2128 (2d Cir. 1994) (equal protection clause requires government to show compelling interest that could justify Secretary of Commerce's refusal to adjust 1990 census) *with* City of Detroit v. Franklin, 4 F.3d 1367 (6th Cir. 1993) (neither statutes nor constitution requires adjustment), *cert. denied,* 114 S. Ct. 1212 (1994); Tucker v. Dep't of Commerce, 958 F.2d 1411 (7th Cir.) (issue is not justiciable), *cert. denied,* 113 S. Ct. 407 (1992).

[20] If the jury list is not compiled properly from appropriate sources, it might be subject to challenge. *See* David Kairys et al., *Jury Representativeness: A Mandate for Multiple Source Lists* , 65 Cal. L. Rev. 776 (1977).

In other situations, the sampling frame may cover less of the population. In an obscenity case, for example, the defendant's poll of opinion about community standards[21] should identify all adults in the legally relevant community as the population, but obtaining the names of all such people may not be possible. If names from a telephone directory are used, people with unlisted numbers are excluded from the sampling frame. If these people, as a group, hold different opinions from those included in the sampling frame, the poll will not reflect this difference, no matter how many individuals are polled and no matter how well their opinions are elicited.[22] The poll's measurement of community opinion will be biased, although the magnitude of this bias may not be great.

Not all surveys use random selection. In some commercial disputes involving trademarks or advertising, the population of all potential purchasers of the products is difficult to identify. Some surveyors may resort to an easily accessible subgroup of the population, such as shoppers in a mall.[23] Such *convenience samples* may be biased by the interviewer's discretion in deciding whom to interview—a form of selection bias—and the refusal of some of those approached to participate—nonresponse bias.[24] Selection bias is acute when constituents write their representatives, listeners call into radio talk shows, or interest groups collect information from their members.[25] Selection bias also affects data from jury-reporting services that gather information from readily available sources.[26]

[21] On the admissibility of such polls, compare, *e.g.*, Saliba v. State, 475 N.E.2d 1181, 1187 (Ind. Ct. App. 1985) ("Although the poll did not . . . [ask] the interviewees . . . whether the particular film was obscene, the poll was relevant to an application of community standards") with United States v. Pryba, 900 F.2d 748, 757 (4th Cir.) ("Asking a person in a telephone interview as to whether one is offended by nudity, is a far cry from showing the materials . . . and then asking if they are offensive," so exclusion of the survey results was proper), *cert. denied*, 498 U.S. 924 (1990).

[22] A classic example of selection bias is the 1936 *Literary Digest* poll. After successfully predicting the winner of every U.S. presidential election since 1916, the *Digest* used the replies from 2.4 million respondents to predict that Alf Landon would win 57% to 43%. In fact, Franklin Roosevelt won by a landslide vote of 62% to 38%. *See* Freedman et al., *supra* note 12, at 306. The *Digest* was so far off, in part, because it chose names from telephone books, rosters of clubs and associations, city directories, lists of registered voters, and mail order listings. *Id.* at 306–08, A-13 n.6. In 1936, when only one household in four had a telephone, the people whose names appeared on such lists tended to be more affluent. Lists that overrepresented the affluent had worked well for sampling in earlier elections, when rich and poor voted along similar lines, but the bias in the sampling frame proved fatal when the Great Depression made economics a salient consideration for voters. *See* Judith M. Tanur, *Samples and Surveys*, *in* Perspectives on Contemporary Statistics 55, 57 (David C. Hoaglin & David S. Moore eds., 1992). Today, survey organizations conduct polls by telephone, but most voters have telephones, and these organizations select the numbers to call at random rather than sampling names from telephone books.

[23] *E.g.*, R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc., 511 F. Supp. 867, 876 (S.D.N.Y. 1980) (questioning the propriety of basing a "nationally projectable statistical percentage" on a suburban mall intercept study).

[24] Nonresponse bias is discussed *infra* § II.B.2.

[25] *E.g.*, Pittsburgh Press Club v. United States, 579 F.2d 751, 759 (3d Cir. 1978) (tax-exempt club's mail survey of its members to show little sponsorship of income-producing uses of facilities was held to be inadmissible hearsay because it "was neither objective, scientific nor impartial"), *rev'd on other grounds*, 615 F.2d 600 (3d Cir. 1980). So, too, veterans groups collected instances of multiple myeloma (a form of cancer) among veterans of the Hiroshima and Nagasaki occupation forces. They claimed that the number of cases was unusual and called for government study and compensation. Such anecdotal evidence, based on a few cases without systematic comparison or data collection, may be an incentive for more careful investigation but may also reflect rumor and speculation rather than fact. In this instance, a committee of the National Research

Various procedures are available to cope with selection bias. In quota sampling, the interviewer is instructed to interview so many women, so many older men, so many ethnic minorities, or the like. But quotas alone still leave vast discretion in selecting among the members of each category and therefore do not solve the problem of selection bias.

Probability sampling methods, in contrast, ideally are suited to avoid selection bias. Once the conceptual population is reduced to a tangible sampling frame, the units to be measured are selected by some kind of lottery that gives each unit in the sampling frame a known, nonzero probability of being chosen. Selection according to a table of random digits or the like[27] leaves no room for selection bias.[28]

2. Of the units selected, which are measured?

Although probability sampling ensures that, within the limits of chance, the sample of units selected will be representative of the sampling frame, the question remains as to which units actually get measured. When objects like receipts (for an audit) or vegetation (for a study of the ecology of a region) are sampled, all can be examined. Human beings are more troublesome. Some may refuse to respond, and the survey should report the nonresponse rate. A large nonresponse rate warns of bias,[29] but it does not necessarily demonstrate bias. Supplemental

Council found no evidence that the rate of multiple myeloma for the "atomic veterans" was higher than that in similar populations. Moore, *supra* note 12, at 124.

[26] For example, a study from the mid-1980s found that the average award in medical malpractice cases was $962,258. The figure comes from a jury-reporting service that relies on newspaper accounts and other sources that are likely to report predominantly large awards. Kenneth Jost, *Still Warring Over Medical Malpractice*, A.B.A. J., May 1993, at 68, 71 (citing an interview with Neil Vidmar). On the limitations of jury verdict service and other reports of jury awards, see, *e.g.*, Theodore Eisenberg & Thomas A. Henderson, Jr., *Inside the Quiet Revolution in Products Liability*, 39 UCLA L. Rev. 731, 765 n.100 (1992).

[27] In simple random sampling, each unit has the same probability of being chosen. More complicated methods, such as stratified sampling and cluster sampling, have advantages in certain applications. In systematic sampling, every fifth, tenth, or hundredth (in mathematical jargon, every *n*th) unit in the sampling frame is selected. If the starting point is selected at random and the units are not in any special order, then this procedure is comparable to simple random sampling.

[28] Before 1968, most federal districts used the "key man" system for compiling lists of eligible jurors. Individuals believed to have extensive contacts in the community would suggest names of prospective jurors, and the qualified jury wheel would be made up from those names. To reduce the risk of discrimination associated with this system, the Jury Selection and Service Act of 1968, 28 U.S.C. §§ 1861–1878 (1988), substituted the principle of "random selection of juror names from the voter lists of the district or division in which court is held." S. Rep. No. 891, 90th Cong., 1st Sess. 10 (1967), *reprinted in* 1968 U.S.C.C.A.N. 1792, 1793.

[29] The 1936 *Literary Digest* election poll illustrates the danger. Only 24% of the 10 million people who received questionnaires returned them. Most of these respondents probably had strong views on the candidates, and most of them probably objected to President Roosevelt's innovative economic programs. This self-selection is likely to have biased the poll. Maurice C. Bryson, *The Literary Digest Poll: Making of a Statistical Myth,* 30 Am. Statistician 184 (1976); Freedman et al., *supra* note 12, at 307–08.

In United States v. Gometz, 730 F.2d 475, 478 (7th Cir.) (en banc), *cert. denied*, 469 U.S. 845 (1984), the Seventh Circuit recognized that "a low rate of response to juror questionnaires could lead to the underrepresentation of a group that is entitled to be represented on the qualified jury wheel." Nevertheless, the court held that under the Jury Selection and Service Act of 1968, 28 U.S.C. §§ 1861–1878 (1988), the clerk did not abuse his discretion by failing to take steps to increase a response rate of 30%. According to the court, "Congress wanted to make it possible for all qualified persons to serve on juries, which is different from forcing all qualified persons to be available for jury service." *Gometz*, 730 F.2d at 480. Although it might "be a good thing to study may establish that the nonrespondents do not differ systematically from the respondents with respect to the characteristics of interest[30] or may permit the missing data to be imputed.[31]

In short, a convincing survey defines an appropriate population to study, uses an unbiased method for selecting units to measure, with a reliable and valid procedure for gathering information on the units selected for study, and succeeds in gathering information on all or a fair cross section of these units. When these goals are met, the sample tends to be representative of the population: The measurements within the sample describe fairly the characteristics in the population. It remains possible, however, that despite every precaution, the sample, being less than exhaustive, is not representative; proper statistical analysis helps address the magnitude of this risk, at least for probability samples.[32] Of course, surveys may be useful even if they fail to meet all of the criteria given above; but then, additional arguments are needed to justify the inferences.

C. Experiments

In many cases, the court needs more than a description of a population. It seeks an answer to a question of causation. Would additional information in a securities prospectus disclosure have caused potential investors to behave any differently? Does the similarity in the names of two products lead consumers to buy one brand because of their familiarity with the other brand? Does capital punishment deter crime? Do food additives cause cancer?

follow up on persons who do not respond to a jury questionnaire," the court concluded that Congress merely "wanted to make it possible for all qualified persons to serve on juries" and "was not concerned with anything so esoteric as nonresponse bias." *Id.* at 479, 480, 482.

[30] Even when demographic characteristics of the sample match those of the population, however, caution still is indicated. In the 1980s, a behavioral researcher sent out 100,000 questionnaires to explore how women viewed their relationships with men. Shere Hite, Women and Love: A Cultural Revolution in Progress (1987). She amassed a huge collection of anonymous letters from thousands of women disillusioned with love and marriage, and she wrote that these responses established that the "outcry" of some feminists "against the many injustices of marriage—exploitation of women financially, physically, sexually, and emotionally" is "just and accurate." *Id*. at 344. The outcry may indeed be justified, but this research does little to prove the point. About 95% of the 100,000 inquiries did not produce responses. The nonrespondents may have had less distressing experiences with men and therefore did not see the need to write autobiographical letters. Furthermore, this systematic difference would be expected within every demographic and occupational class. Therefore, the argument that the sample responses are representative because "those participating according to age, occupation, religion, and other variables known for the U.S. population at large in most cases quite closely mirrors that of the U.S. female population" is far from convincing. *Id.* at 777. In fact, the results of this non-random sample differ dramatically from those of polls with better response rates. *See* Chamont Wang, Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety 174–76 (1993). *See also* David Streitfeld, *Shere Hite and the Trouble with Numbers,* 1 Chance 26 (1988).

[31] Methods for "imputing" missing data are sketched in, *e.g.*, Tanur, *supra* note 22, at 55. For more technical references, see, *e.g*., Donald B. Rubin, Multiple Imputation for Nonresponse in Surveys (1987); Imputation and Editing of Faulty or Missing Survey Data (Faye Aziz & Fritz Scheuren eds., 1978). Efforts to fill in missing data can be problematic. The "easy case" is one in which the response rate is so high that even if all nonrespondents had responded in a way adverse to the proponent of the survey, the substantive conclusion would be unaltered.

[32] *See infra* § IV.

Controlled experiments are, far and away, the best vehicle for establishing a causal relationship. Such experiments may exist before the commencement of the litigation. If so, it becomes the task of the lawyer and appropriate experts to explain this research to the court. Examples of such "off-the-shelf" research are experiments pinpointing conditions under which eyewitnesses tend to err in identifying criminals[33] or studies of how sex stereotyping affects perceptions of women in the workplace.[34] Even if no preexisting studies are available, a case-specific one may be devised:[35] A psychologist may simulate the conditions of a particular eyewitness's identification to see whether comparable identifications tend to be correct;[36] an organization investigating racial discrimination in the rental-housing market may send several "testers" (who, it is hoped, differ only in their race) to rent a property.[37]

A well-designed experiment shows how one variable responds to changes in variables under the control of the experimenter. Variables not directly controlled should be subject only to random fluctuations. For example, to verify that a fertilizer improves crop yields, it is insufficient only to report that the yield is high in a fertilized field. It may be that the yield would have been higher without the fertilizer. To compare the outcome with fertilizer to the outcome without fertilizer, two essentially identical fields can be planted, and fertilizer can be applied only to one field. If the conditions in the fields are nearly identical, any large difference in the yields must be the result of fertilizer. By definition, other possible causes have been eliminated.

To the extent that the two fields are not truly identical, but differ in a myriad of ways that are hard to specify but could affect the yield, the experiment may be replicated on many fields randomly assigned to be fertilized or not.[38] These strategies of control and randomization are the earmarks of good experiments.

[33] *E.g.*, State v. Chapple, 660 P.2d 1208, 1223–24 (Ariz. 1983) (reversing a conviction for excluding expert testimony about scientific research on eyewitness testimony). For citations to the case law and scientific literature, see, *e.g.*, 1 McCormick on Evidence § 206(A) (John William Strong ed., 4th ed. 1992).

[34] The testimony of a social psychologist about stereotyping played a limited—and controversial—role in Price Waterhouse v. Hopkins, 490 U.S. 228, 235 (1989). *Compare* Gerald V. Barrett & Scott B. Morris, *The American Psychological Association's Amicus Curiae Brief in Price Waterhouse v. Hopkins: The Values of Science Versus the Values of the Law*, 17 Law & Hum. Behav. 201 (1993) *with* Susan T. Fiske et al., *What Constitutes a Scientific Review?: A Majority Retort to Barrett and Morris*, 17 Law & Hum. Behav. 217 (1993).

[35] For a review of the law on such pretrial experiments and a proposal that the parties be encouraged to cooperate in the design of such experiments, see 1 McCormick on Evidence, *supra* note 33, § 202.

[36] Willem A. Wagenaar, *The Proper Seat: A Bayesian Discussion of the Position of the Expert Witness*, 12 Law & Hum. Behav. 499 (1988) (describing the difficulty of presenting the results of such an experiment to a court in the Netherlands).

[37] *E.g.*, United States v. Youritan Constr. Co., 370 F. Supp. 643, 647 (N.D. Cal. 1973), *aff'd in part*, 509 F.2d 623 (9th Cir. 1975).

[38] *Statistically significant* differences are those that are so large that they rarely would occur with an ineffectual fertilizer just because the fields randomly selected for the fertilizer treatment happen to be the best for growth. The techniques for establishing statistical significance are considered *infra* § IV.

1. What are the independent and dependent variables?

In investigating a possible cause-and-effect relationship, the variable that characterizes the effect is called the *dependent variable,* since it may depend on the causes.[39] In contrast, the variables that represent the causes are called *independent variables*.[40] In the fertilizer experiment, crop yield is the dependent variable. It depends on such independent variables as the density of planting, the level of irrigation or rainfall, the nature of the soil, and the extent of insect infestation. Listing such variables is a useful exercise because it focuses attention on which factors are under control (and can be excluded as causes of the observed differences) and which are not (and may mask a causal relation or give a false appearance of one).

2. What are the confounding variables?

A *confounding variable* is correlated with the independent variables and with the dependent variable. Since a confounding variable changes with one or more independent variables, it is generally not possible to determine whether changes in the independent variables caused changes in the dependent variable or whether changes in the confounding variable did—especially if the investigator did not collect data on the *confounder* . For example, many studies have been conducted to determine whether physical exercise increases life span. In one such study, the physical fitness of a large number of men was measured. Over the next sixteen years, about twice as many men in the lowest fitness quartile died as did men in the highest quartile.[41] One interpretation is that maintaining a high level of physical activity protects against death. However, both physical fitness and mortality are correlated with general health at the beginning of the study; thus, it is possible that the highly fit men lived longer, not because they exercised, but simply because they were healthier to begin with.[42] A disproportionate number of healthier men in the high fitness group biases the study in favor of finding improved survival in that group.

Randomly assigning subjects to a treatment and a control group eliminates this problem.[43] In experiments on human beings, it is especially difficult to ensure that the treatment and control groups are identical, but with random selection the many factors not under the experimenter's control tend to balance out

[39] Dependent variables also may be called response variables.

[40] Independent variables also may be called factors or explanatory variables.

[41] Leiv Sandvik et al., *Physical Fitness as a Predictor of Mortality Among Healthy, Middle-Aged Norwegian Men* , 328 New Eng. J. Med. 533 (1993). The lower mortality among the fittest men was attributable chiefly to a lower risk of dying of cardiovascular causes.

[42] Gregory D. Curfman, *The Health Benefits of Exercise: A Critical Reappraisal* , 328 New Eng. J. Med. 574, 575 (1993).

[43] "A randomized, controlled trial of physical activity for the primary prevention of cardiovascular disease . . . has never been performed and is probably not feasible because of problems related to compliance and cost." *Id*. *But see* M. A. Fiatarone et al., *Exercise Training and Nutritional Supplementation for Physical Frailty in Very Elderly People,* 330 New Eng. J. Med. 1769 (1994).

in both groups.[44] A handful of well-designed studies is far more convincing than any number of biased ones.

Two insights are important. First, outcome figures from a treatment group without a control group reveal very little and can be misleading.[45] Comparisons are essential. Second, if the control group was obtained through random assignment before treatment, a difference in the outcomes between treatment and control groups may be accepted, within the limits of statistical error, as the true measure of the treatment effect.[46] However, if the control group was created in any other way, differences in the groups that existed before treatment may contribute to differences in the outcomes or mask differences that otherwise would be observed. Thus, observational studies succeed to the extent that their treatment and control groups are comparable, apart from the treatment.

3. Can the results be generalized?

All experiments are conducted with a sample of a certain population, at a certain place, at a certain time, and with a limited number of treatments. With respect to the sample studied, the experiment may be persuasive. It may have succeeded in controlling all confounding variables and in finding an unequivocally large difference between the treatment and control groups. If so, its "internal validity" will not be disputed; in the sample studied, the treatment has an effect.

But an issue of "external validity" remains. To extrapolate from the limiting conditions of an experiment always raises questions. If juries react differently to competing instructions on the law of insanity in cases of housebreaking and of incest,[47] would the difference persist if the charge were rape or murder? Would the failure of ex-convicts to react to transitory payments after release hold if conditions in the employment market were to change radically?[48]

Confidence in the appropriateness of an extrapolation cannot come from the experiment itself. It must come from knowledge about which outside factors

[44] Of course, the possibility that the two groups will not be comparable in some unrecognized way can never be eliminated. Random assignment, however, allows the researcher to compute the probability of seeing a large difference in the outcomes when the treatment actually has no effect. When this probability is small, the difference in the response is said to be "statistically significant." *See infra* § IV.B.2.

Randomization also ensures that the assignment of subjects to treatment and control groups is free from conscious or unconscious manipulation by investigators or subjects. Randomization may not be the only way to ensure such protection, but "it is the simplest and best understood way to certify that one has done so." Philip W. Lavori et al., *Designs for Experiments—Parallel Comparisons of Treatment, in* Medical Uses of Statistics 61, 66 (John C. Bailar III & Frederick Mosteller eds., 2d ed. 1992). To avoid ambiguity, the researcher should be explicit "about how the randomization was done (e.g., table of random numbers) and executed (e.g., by sealed envelopes prepared in advance)." *Id*.

[45] For an effort to identify circumstances in which such studies may be informative, see John C. Bailar III et al., *Studies Without Internal Controls, in* Medical Uses of Statistics, *supra* note 44, at 105.

[46] The problem of statistical error is treated *infra* § IV.

[47] *See* Rita James Simon, The Jury and the Defense of Insanity 58–59 (1967).

[48] For an experiment on income support and recidivism, see Peter H. Rossi et al., Money, Work, and Crime: Experimental Evidence (1980). The interpretation of the data has proved controversial. *See* Hans Zeisel, *Disagreement over the Evaluation of a Controlled Experiment,* 88 Am. J. Soc. 378 (1982) (with commentary).

would or would not affect the outcome.[49] Sometimes, several experiments or other studies, each having different limitations, all point in the same direction.[50] Such convergent results strongly suggest the validity of the generalization.[51]

D. Observational Studies of Causation

The bulk of the statistical studies seen in court are observational, not experimental. In an experiment the investigators select certain units for treatment. In an *observational study* the investigators have no control over who or what receives the treatment. Take the question of whether capital punishment deters murder. To do a randomized controlled experiment, people would have to be assigned randomly to a control group and a treatment group. The controls would know that they could not receive the death penalty for murder, while those in the treatment group would know they could be executed. The rate of subsequent murders by the subjects in these groups would be observed. Such an experiment is unacceptable—politically, ethically, and legally.[52]

Nevertheless, many studies of the deterrent effect of the death penalty have been conducted, all observational, and some have attracted judicial attention.[53] Researchers have catalogued differences in the incidence of murder in states with and without the death penalty, and they have analyzed changes in homicide rates and execution rates over the years. In such observational studies, investigators may speak of control groups (such as the states without capital punish-

[49] Such judgments are easiest in the natural sciences, but even here, there are problems. For example, it may be difficult to infer human reactions to substances that affect animals. First, there are inconsistencies across test species: A chemical may be carcinogenic in mice but not in rats. Extrapolation from rodents to humans is even more problematic. Second, to get measurable effects in animal experiments, chemicals are administered at very high doses. Results are extrapolated—using mathematical models—to the very low doses of concern in humans. However, there are many dose-response models to use and few grounds for choosing among them. Generally, different models produce radically different estimates of the "virtually safe dose" in humans. David A. Freedman & Hans Zeisel, *From Mouse to Man: The Quantitative Assessment of Cancer Risks*, 3 Stat. Sci. 3 (1988). For these reasons, many experts—and some courts in toxic tort cases—have concluded that evidence from animal experiments is generally insufficient to establish causation. *See* Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and Bendectin Litigation* , 86 Nw. U. L. Rev. 643 (1992); Lois S. Gold et al., *Rodent Carcinogens: Setting Priorities*, 258 Science 261 (1992); D. Krewski et al., *A Model-Free Approach to Low-Dose Extrapolation*, 90 Envtl. Health Persp. 279 (1991); Susan R. Poulter, *Science and Toxic Torts: Is There a Rational Solution to the Problem of Causation?,* 7 High Tech. L.J. 189 (1993) (epidemiological evidence on humans is needed). *See also* Committee on Risk Assessment Methodology, National Research Council, Issues in Risk Assessment (1993).

[50] This is the case, for example, with eight studies indicating that jurors who approve of the death penalty are more likely to convict in a capital case. Phoebe C. Ellsworth, *Some Steps Between Attitudes and Verdicts, in* Inside the Juror 42, 46 (Reid Hastie ed., 1993). Nevertheless, in Lockhart v. McCree, 476 U.S. 162 (1986), the Supreme Court held that the exclusion of opponents of the death penalty in the guilt phase of a capital trial does not violate the constitutional requirement of an impartial jury.

[51] *See* Zeisel, *supra* note 12, at 252–62.

[52] *Cf*. Experimentation in the Law: Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law (Federal Judicial Center 1981) [hereinafter Experimentation in the Law] (study of ethical issues raised by controlled experimentation in the evaluation of innovations in the justice system).

[53] *See generally* Hans Zeisel, *The Deterrent Effect of the Death Penalty: Facts v. Faith*, 1976 Sup. Ct. Rev. 317.

ment) and of controlling for potentially confounding variables (e.g., worsening economic conditions).[54] However, association is not causation, and the causal inferences that can be drawn from such analyses rest on a less secure foundation than that provided by a controlled randomized experiment.[55]

Of course, observational studies can be very useful. The evidence that smoking causes lung cancer in humans, although largely observational, is compelling. In general, observational studies provide powerful evidence in the following circumstances.

- The association is seen in studies of different types among different groups. This reduces the chance that the observed association is due to a defect in one type of study or a peculiarity in one group of subjects.
- The association holds when the effects of plausible confounding variables are taken into account by appropriate statistical techniques, such as comparing smaller groups that are relatively homogeneous with respect to the factor.[56]

[54] A procedure often used to control for confounding in observational studies is regression analysis. The underlying logic is described *infra* § III.F.3. The early enthusiasm for using multiple regression analysis to study the death penalty was not shared by reviewers. *Compare* Isaac Ehrlich, *The Deterrent Effect of Capital Punishment: A Question of Life and Death*, 65 Am. Econ. Rev. 397 (1975) *with, e.g*., Lawrence R. Klein et al., *The Deterrent Effect of Capital Punishment: An Assessment of the Estimates, in* Panel on Research on Deterrent and Incapacitative Effects, National Research Council, Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates 336 (Alfred Blumstein et al. eds., 1978); Edward Leamer, *Let's Take the Con Out of Econometrics* , 73 Am. Econ. Rev. 31 (1983); Richard O. Lempert, *Desert and Deterrence: An Assessment of the Moral Bases of the Case for Capital Punishment,* 79 Mich. L. Rev. 1177 (1981); Richard O. Lempert, *The Effect of Executions on Homicides: A New Look in an Old Light* , 29 Crime & Delinq. 88 (1983).

[55] *See, e.g* ., Experimentation in the Law, *supra* note 52, at 18:

> [G]roups selected without randomization will [almost] always differ in some systematic way other than exposure to the experimental program. Statistical techniques can eliminate chance as a feasible explanation for the differences, . . . [b]ut without randomization there are no certain methods for determining that observed differences between groups are not related to the preexisting, systematic difference . . . [C]omparison between systematically different groups will yield ambiguous implications whenever the systematic difference affords a plausible explanation for apparent effects of the experimental program.

[56] The idea is to control for the influence of a confounder by making comparisons separately within groups for which the confounding variable is nearly constant and therefore has little influence over the variables of primary interest. For example, smokers are more likely to get lung cancer than nonsmokers. Age, gender, social class, and region of residence are all confounders, but controlling for such variables does not really change the relationship between smoking and cancer rates. On the basis of observational studies, most experts believe that smoking does cause lung cancer (and many other diseases). For a recent review of the literature, see 38 International Agency for Research on Cancer (IARC), World Health Org., IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans: Tobacco Smoking (1986). However, the associations seen in observational studies, even good ones, can be misleading. For example, women with herpes are more likely to develop cervical cancer than women who have not been exposed to the virus. For a time, it was believed that herpes caused cancer. In other words, the association was thought to be causal. Later research suggests that herpes is only a marker of sexual activity. Women who have had multiple sexual partners are more likely to be exposed not only to herpes but also to human papilloma virus. Certain strains of papilloma virus seem to cause cervical cancer, while herpes does not. Apparently, the association between herpes and cervical cancer is not causal but is due to the effect of other variables. *See* Viral Etiology of Cervical Cancer (Richard Peto & Harald zur Hausen eds., 1986); The Epidemiology of Human Papillomavirus and Cervical Cancer (N. Muñoz et al. eds., 1992). For additional examples and discussion, see Freedman et al., *supra* note 12, at 11–25, 133–48.

- There is a plausible explanation for the effect of the independent variables; thus, the causal link does not depend on the observed association alone. Other explanations linking the response to confounding variables should be less plausible.[57]

When these criteria are not fulfilled, observational studies may produce legitimate disagreement among experts, and there is no mechanical procedure for ascertaining who is correct. In the end, deciding whether associations are causal is not a matter of statistics, but a matter of good scientific judgment, and the questions that should be asked with respect to data offered on the question of causation can be summarized as follows:

- Was there a control group? If not, the study has little to say about causation.
- If there was a control group, how were subjects assigned to treatment or control: through a process under the control of the investigator (a controlled experiment) or a process outside the control of the investigator (an observational study)?
- If it was a controlled experiment, was the assignment made using a chance mechanism (randomization), or did it depend on the judgment of the investigator?
- If the data came from an observational study or a nonrandomized controlled experiment, how did the subjects come to be in treatment or in control groups? Are the groups comparable? What factors are confounded with treatment? What adjustments were made to take care of confounding? Were they sensible?[58]

[57] David S. Moore & George P. McCabe, Introduction to the Practice of Statistics 202 (2d ed. 1993).

[58] These questions are adapted from Freedman et al., *supra* note 12, at 25. As with controlled experiments, chance variation sometimes produces an apparent association between variables when none really exists (*see infra* § IV).

III. How Have the Data Been Presented?

After data have been collected, they should be presented in a way that makes them intelligible and revealing. Huge quantities of data can be summarized with a few numbers or with graphical displays. However, the wrong summary or a distorted graph can mislead.[59]

A. Is the Data Display Sufficiently Complete?

Selective presentation of numerical information is like quoting someone out of context. A television commercial for the Investment Company Institute (the mutual fund trade association) said that a $10,000 investment made in 1950 in an average common stock mutual fund would have increased to $113,500 by the end of 1972. The *Wall Street Journal* indicated that the same investment spread over all the stocks making up the New York Stock Exchange Composite Index would have grown to $151,427. Mutual funds performed worse than the stock market as a whole.[60] In this example, and in many other situations, it is helpful to look beyond a single number to some comparison or benchmark that places the isolated figure into perspective.

Even complete and accurate data can mislead if changes in the process of collecting the data are not reported. For example, the number of petty larcenies reported in Chicago more than doubled between 1959 and 1960—not because of an abrupt crime wave—but because a new police commissioner introduced an improved reporting system.[61] For many years, researchers ignored New York City crime statistics because it was common practice for the precincts to underreport crime to protect the reputations of their neighborhoods. When New York City shifted to a centralized reporting system, burglary reports increased more than fourteenfold in three years.[62] During the 1970s, police officials in Washington, D.C., "demonstrated" the success of President Nixon's law-and-order campaign by valuing stolen goods at $49, just below the $50 threshold for inclusion in the Federal Bureau of Investigation's (FBI) Uniform Crime

[59] *See generally* Campbell, *supra* note 12; Freedman et al., *supra* note 12; Huff, *supra* note 12; Katzer et al., *supra* note 12; Moore, *supra* note 12; Runyon, *supra* note 12; Zeisel, *supra* note 12.

[60] Moore, *supra* note 12, at 128.

[61] *Id*. at 129.

[62] Mark H. Maier, The Data Game: Controversies in Social Science Statistics 80–81 (1991).

Reports.[63] Likewise, in the mid-1970s, the Indianapolis police department tripled the number of crime reports deemed "without merit," which hence went uncounted in the Uniform Crime Reports.[64]

Changes in the collection of data over the years are by no means limited to crime statistics. In 1971, President Nixon signed the National Cancer Act, calling for a war on cancer of the "same kind of concentrated effort that split the atom and took man to the moon."[65] Two decades and hundreds of billions of dollars later, advocates of the war on cancer recognize that no general cure is close at hand. They are encouraged, however, by the development of cures for some cancers, and they cite improved survival rates for other cancers. Some epidemiologists question the inference that the changes in survival rates reflect a successful assault on the disease. Because some kinds of cancers now are detected earlier, patients with these cancers merely appear to live longer.[66]

Almost all series of numbers that cover many years are affected by changes in definitions and collection methods. When considering time series data, it is worth looking for any sudden jumps, which may signal a change in definitions or data collection procedures.[67]

[63] James P. Levine et al., Criminal Justice in America: Law in Action 99 (1986); Maier, *supra* note 62, at 81.

[64] Maier, *supra* note 62, at 81; Harold E. Pepinsky & Paul Jesilow, Myths That Cause Crime 28 (1985).

[65] As quoted in Ralph W. Moss, The Cancer Syndrome 16 (1980). *See also* Richard M. Nixon, *Acting Against Cancer*, Sat. Evening Post, July/Aug. 1986, at 67.

[66] *See* Maier, *supra* note 62, at 55; James E. Enstrom & Donald F. Austin, *Interpreting Cancer Survival Rates*, 195 Science 847 (1977); *Cancer: Illusory Progress?*, Sci. Am., June 1987, at 29. For a more recent discussion of the difficulties in interpreting trends in incidence and death rates for cancers, see Tim Beardsley, *A War Not Won*, Sci. Am., Jan. 1994, at 130; National Cancer Inst., Evaluating the National Cancer Program: An Ongoing Process (1994) (transcript of the President's Cancer Panel meeting, Sept. 23, 1993, on file with the National Cancer Institute).

[67] Moore, *supra* note 12, at 129. Another problem can arise from collapsing categories in a table. In Philip Morris, Inc. v. Loew's Theatres, Inc., 511 F. Supp. 855 (S.D.N.Y. 1980), and R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc., 511 F. Supp. 867 (S.D.N.Y. 1980), Philip Morris and R.J. Reynolds sought an injunction to stop the maker of Triumph low-tar cigarettes from running advertisements claiming that participants in a national taste test preferred Triumph to other brands. Plaintiffs alleged that claims that Triumph was a "national taste test winner" or Triumph "beats" other brands were false and misleading. An exhibit introduced by defendant contained the following data:

	Triumph much better than Merit	Triumph somewhat better than Merit	Triumph about the same as Merit	Triumph somewhat worse than Merit	Triumph much worse than Merit
Number	45	73	77	93	36
Percentage	14%	22%	24%	29%	11%

511 F. Supp. at 866. Only 14% + 22% = 36% of the sample preferred Triumph to Merit, while 29% + 11% = 40% preferred Merit to Triumph. *Id.* at 856. By selectively combining categories, however, defendant attempted to create a different impression. Since 24% found the brands about the same, and 36% preferred Triumph, defendant claimed that a clear majority (36% + 24% = 60%) found Triumph "as good or better than Merit." *Id.* at 866. The court correctly resisted this chicanery, finding that defendant's test results did not support the advertising claims. *Id.* at 856–57. The statistical issues in these cases are discussed more fully in 2 Gastwirth, *supra* note 1, at 633–39. For a hypothetical, but strikingly similar example of selective collapsing of categories, see Richard P. Runyon, How Numbers Lie: A Consumer's Guide to the Fine Art of Numerical Deception 67–70 (1981).

Few summaries of data are intended to mislead; most try to bring out broad features of the data. All descriptive statistics, however, are simplifications, and there are times when the details they omit are important. The statistical analyst should be able to explain why the summary statistics used are sufficient to capture the relevant aspects of the data. For instance, the proportion of applicants who pass an entrance examination for a police academy is sufficient to indicate how significant a barrier the test is for that group of tested individuals. For this purpose, it is not necessary to know how each individual scored.[68] In other situations, a graph may reveal a pattern not evident from the summary statistic.[69]

B. Are Rates or Percentages Properly Interpreted?

Rates and percentages effectively summarize data, but these statistics can be misinterpreted. A percentage is a summary that makes a comparison between two numbers. One number is the base, and the other number is compared with that base. When the base is small, actual numbers may be more revealing than percentages. For example, there were media accounts in 1982 of a crime wave by the elderly. The annual Uniform Crime Reports showed a near tripling of the crime rate by older people since 1964, while crimes by younger people only doubled. But people over 65 years of age account for less than 1% of all arrests. In 1980, for instance, there were only 151 arrests of the elderly for robbery out of 139,476 total robbery arrests.[70]

Usually, the small-base problem is obvious if the presentation is reasonably complete. An expert who says that 50% of the people interviewed had a certain opinion also should reveal how many individuals were contacted and how many expressed an opinion.[71] Then we know whether the 50% is 2 out of 4 or 500 out of 1,000.

Finally, there is the issue of which numbers to compare.[72] Researchers sometimes choose among alternative comparisons. It may be worthwhile to ask why they chose the one they did. Does it give a fair picture, or would another comparison give a different view? A government agency, for example, may want to compare the amount of service being given this year with that of earlier years—but what earlier year ought to be the baseline? If the first year of operation is used, a large percentage increase due to start-up problems for a new

[68] If the analyst wants to examine the effect of the test on different subgroups, the proportions in each relevant subgroup must be considered. *See, e.g.*, Bouman v. Block, 940 F.2d 1211 (9th Cir.), *cert. denied*, 112 S. Ct. 640 (1991); 1 Gastwirth, *supra* note 1, at 254–55.

[69] *See infra* § III.C.3.

[70] Maier, *supra* note 62, at 83. *See also* Alfred Blumstein & Jacqueline Cohen, *Characterizing Criminal Careers*, 237 Science 985 (1987).

[71] For a poll suggesting that male and female trial attorneys have different impressions of the behavior of male and female litigators but omitting the number of respondents by category, see Stephanie B. Goldberg, *"Good Girl" Litigators*, A.B.A. J., June 1993, at 33.

[72] *See, e.g.*, Runyon, *supra* note 67, at 75–79.

agency should be expected.[73] If last year is used as the baseline, was last year also part of an increasing service trend, or was it an unusually poor year? If the base year is not representative of the other years, the percentage may not portray the trend fairly.[74] No single question can be formulated to detect such distortions. The judge can ask for the numbers from which the percentages were obtained, and asking about the base can expose distortions. Ultimately, however, the judge must recognize which numbers relate to which issues—a species of clear thinking that is not reducible to a checklist.[75]

C. Does a Graph Portray Data Fairly?

Graphs are useful for revealing key characteristics of a batch of numbers, trends over time, and the relationships among variables.[76]

1. Displaying distributions: histograms

A graph commonly used to display the distribution of a batch of numbers is the *histogram*.[77] One axis shows the numbers, and the other indicates how often those fall within specified intervals (called a *bin* or a *class interval*). For example, we flipped a quarter ten times in a row and counted the number of heads in this "batch" of ten tosses. For 50 batches, we got the following data:

7 7 5 6 8   4 2 3 6 5   4 3 4 7 4   6 8 4 7 4   7 4 5 4 3
4 4 2 5 3   5 4 2 4 4   5 7 2 3 5   4 6 4 9 10   5 5 6 6 4

The data are shown in Figure 1 below (with a bin width of 1).

[73] *Cf*. Michael J. Saks, *Do We Really Know Anything About the Behavior of the Tort Litigation System—And Why Not?*, 140 U. Pa. L. Rev. 1147, 1203 (1992) (using 1974 as the base year for computing the growth of federal product liability filings exaggerates growth because "1974 was the first year that product liability cases had their own separate listing on the cover sheets. . . . The count for 1974 is almost certainly an understatement . . . .").

[74] Katzer et al., *supra* note 12, at 106.

[75] For some assistance in the task of coping with percentages, see Zeisel, *supra* note 12, at 1–24.

[76] *See generally* William S. Cleveland, The Elements of Graphing Data (1985); Moore & McCabe, *supra* note 57, at 3–20.

[77] For small batches of numbers, stem-and-leaf plots show all the values and how they are distributed. A stem-and-leaf plot for 11, 12, 23, 23, 23, 23, 33, 45, 69 is shown below:

1	12
2	3333
3	3
4	5
5	
6	9

The numbers to the left of the line are the first digits; those to the right are the second digits. Thus, the entry "2 | 3333" stands for "23, 23, 23, 23."

Figure 1

Histogram showing how frequently various numbers of heads appeared in 50 batches of 10 tosses of a quarter. The bin width is 1.

insert figure 1 here

Figure 1 shows how the number of heads per batch of ten tosses is distributed over the full range of possible values. The spread can be made to appear larger or smaller, however, by changing the scale of the horizontal axis. Likewise, the shape can be altered somewhat by changing the size of the bins.[78] It may be worth inquiring how the analyst chose the bin width.[79]

2. Displaying trends

Graphs that plot many values of a variable over time are useful for seeing trends. However, the scales on the axes matter. Figures 2 and 3 show how the scale of an axis can be changed to give a different appearance to the same data.[80] In Figure 2, the federal debt appears to skyrocket during the Reagan and Bush administrations, whereas in Figure 3, the federal debt grows steadily during the same years. The moral is simple: Pay attention to the markings on the axes to determine whether the scale is appropriate.

[78] In Figure 1, all the bins have equal widths. The histogram is just like a bar graph. However, government agencies often publish economic and social data in tables with unequal intervals. The resulting histograms have unequal bin widths; bar heights are calculated so that the areas (height × width) are proportional to the frequencies. In general, a histogram differs from a bar graph in that it represents frequencies by area, not height. *See* Freedman et al., *supra* note 12, at 29–40.

[79] As the width of the bins decreases, the graph becomes more detailed. But the appearance becomes more ragged until finally the graph is effectively a plot of each datum. No general rule can be stated as to what bin width is optimal: "[T]he tolerable loss depends on the subject matter and the goal of the analysis." Cleveland, *supra* note 76, at 125.

[80] The data are taken from figures in Howard Wainer, *Graphical Answers to Scientific Questions*, Chance, Fall 1993, at 48, 50. This flexibility in presentation applies to other types of graphs as well. *See* Runyon, *supra* note 67, at 37–39.

Figure 2
The federal debt skyrockets under Reagan-Bush.

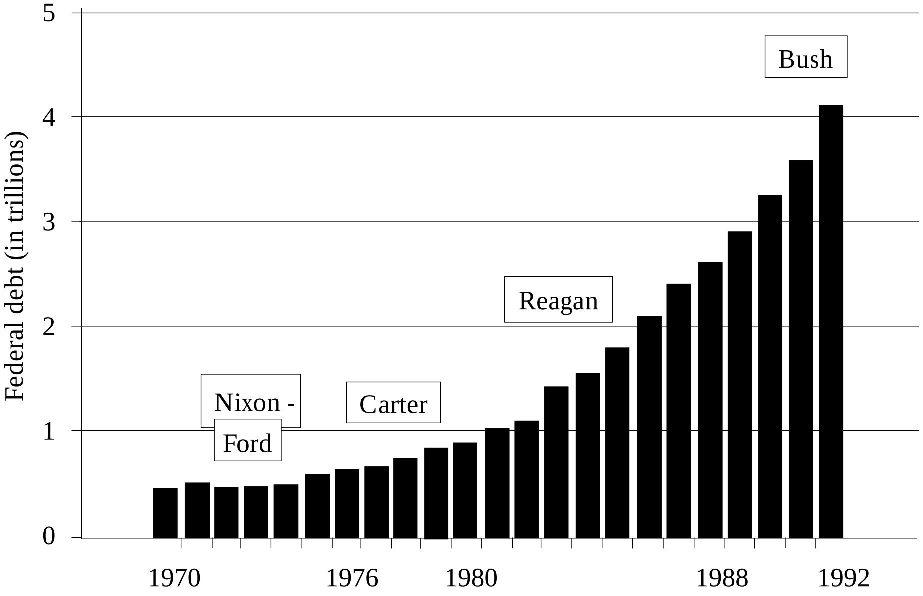
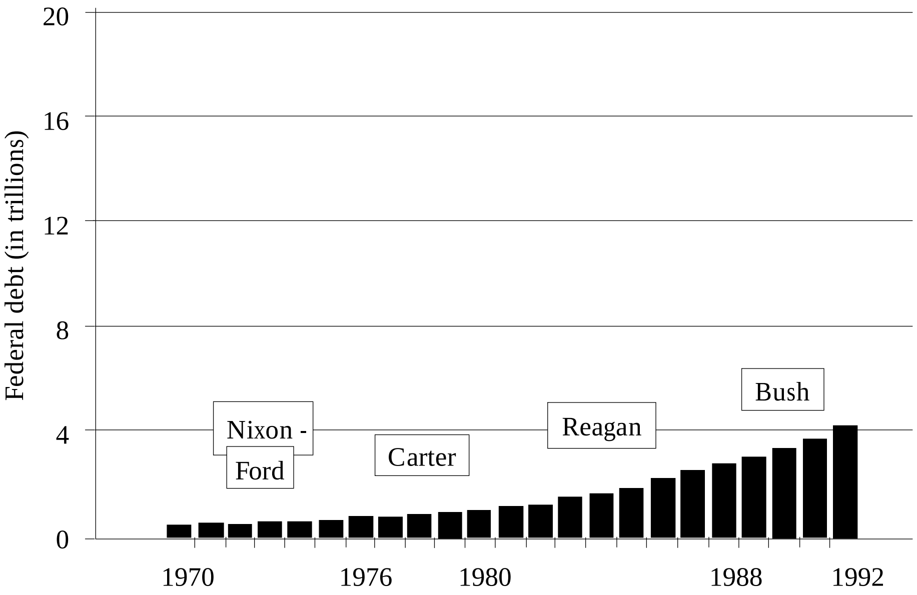


Figure 3
The federal debt grows steadily under Reagan-Bush.



3. Displaying association: scatter diagrams

The relationship between two variables can be shown in a *scatter diagram* (also known as a scatterplot or scattergram). Data on income and education for a sample of 350 men aged 25 to 29 in Texas[81] provide an illustration. Each person in the sample corresponds to one dot in the diagram. As indicated in Figure 4, the horizontal axis shows this person's education, and the vertical axis shows his income. Person A completed 8 years of schooling (grade school) and had an income of $19,000 dollars. Person B completed 16 years of schooling (college) and had an income of $38,000.

Figure 4

Plotting a scatter diagram. The horizontal axis shows educational level, and the vertical axis shows income.

insert figure 4 here

Figure 5 (next page) is the scatter diagram for all the Texas data. This scatter diagram confirms an obvious point. There is a positive association between income and education. In general, people with higher educational levels have higher incomes. However, there are many exceptions to this rule, and the association is not as strong as one might expect. The correlation coefficient is a numerical measure of the strength of the association.[82]

[81] These data are from a public-use data tape, Bureau of the Census, U.S. Dep't of Commerce, for the Current Population Survey of March 1988. Income and education are self-reported. Income is truncated at $100,000 and education (years of schooling completed) at 18 years.

[82] For a discussion of correlation coefficients, see i*nfra* § III.F.2.

Figure 5

Scatter diagram for income and education; men aged 25 to 29 in Texas.[83]

insert figure 5 here

D. Is an Appropriate Measure Used for the Center of a Distribution?

Perhaps the most familiar descriptive statistic is the arithmetic mean, or average. The *mean* of a batch of numbers lies somewhere in the middle of the data. The mean can be found by adding up all the numbers and dividing by how many there are. The *median* has a different definition. Half the numbers are bigger than the median, and half are smaller.[84] Yet a third statistic is the *mode* —the most common number in the data set. These measures have different properties.[85] The mean takes account of all the data—it involves the total of all the numbers—but, particularly with small data sets, a few unusually large or small

[83] Education may be compulsory, but the Current Population Survey generally finds a small percentage of respondents who report very little schooling. Such respondents will be found at the lower left corner of the scatter diagram.

[84] Technically, at least half the numbers are at least as large as the median, and at least half are as small as the median. When the distribution is symmetric, the mean equals the median. The values diverge, however, when the distribution is asymmetric, or skewed.

[85] How big an error do you make in replacing every number by the “center” of the batch? (1) The mode minimizes the number of errors; for the mode, all “errors” count the same, no matter what their sizes are. Consequently, similar distributions can have very different modes, and the mode is rarely useful. (2) The median minimizes a different measure of error—the sum of all the differences (treating positive and negative differences the same) between the center and the data points. (3) The mean minimizes the sum of the squared differences.

observations can cause it to shift substantially. The median, in contrast, is more resistant to such *outliers*.

Which statistic is most useful depends on the purpose of the analysis. For example, what should be made of a report that the average award in malpractice cases skyrocketed from $220,000 in 1975 to more than $1 million in 1985?[86] It might be noted that the median award almost certainly was far less than $1 million[87] and that the apparently explosive growth may be nothing more than the addition of a tiny fraction of very large awards. Still, if the issue is whether insurers were experiencing more costs from jury verdicts, then the mean is the more appropriate statistic. The total of the awards is related directly to the mean,[88] but this figure cannot be recovered from the median.[89]

E. Is an Appropriate Measure of Variability Used?

The location of the center of a batch of numbers reveals nothing about the variations that these numbers exhibit.[90] Statistical measures of variability include the *range*, the interquartile range, the mean absolute deviation, and the *standard deviation*. The range is the difference between the high and the low. It seems natural, and it indicates the maximum spread in the numbers, but it is generally the most unstable because it depends entirely on the most extreme values. The interquartile range is the difference between the 25th and 75th percentiles.[91] It contains 50% of the numbers and is more resistant to changes in the extreme values. The mean absolute deviation depends on all the numbers. It is calculated by averaging the differences between each number and the mean. The

[86] Jost, *supra* note 26, at 68, 70–71.

[87] A study of cases in North Carolina reported an "average" (mean) award of $367,737 and a median award of only $36,500. *Id*. at 71. In TXO Prod. Corp. v. Alliance Resources Corp., 113 S. Ct. 2711 (1993), briefs portraying punitive damages awards as being out of control reported mean punitive awards some ten times larger than the median awards described in briefs defending the current system of punitive damages. *See* Michael Rustad & Thomas Koenig, *The Supreme Court and Junk Social Science: Selective Distortion in Amicus Briefs*, 72 N.C. L. Rev. 91, 145–47 (1993). The two measures differ so dramatically because the mean allows a few huge awards to overwhelm the effects of many smaller ones.

Another dispute over the choice of the mean or the median involves the Railroad Revitalization and Regulatory Reform Act, 49 U.S.C. § 11503, which forbids the taxation of railroad property at a higher rate than other commercial and industrial property. To compare the rates, tax authorities often use the mean, but railroads prefer the median. *See* David A. Freedman, *The Mean Versus the Median: A Case Study in 4-R Act Litigation,* 3 J. Bus. & Econ. Stat. 1 (1985).

[88] To get the total, just multiply the mean by the number of awards. The more pertinent figure is not the total of jury awards, but actual claims experience, including settlements.

[89] These and related statistical issues are pursued further in, *e.g.*, Eisenberg & Henderson, *supra* note 26, at 731, 764–72; Scott Harrington & Robert E. Litan, *Causes of the Liability Insurance Crisis*, 239 Science 737, 740–41 (1988); Saks, *supra* note 73, at 1147, 1248–54.

[90] The numbers 1, 2, 5, 8, 9 have 5 as their mean and median. So do the numbers 5, 5, 5, 5, 5. In the first batch, the numbers vary considerably about their mean; in the second, the numbers do not vary at all.

[91] By definition, 25% of the data fall below the 25th percentile. The median is the 50th percentile.

standard deviation is like the mean absolute deviation except that the squared differences[92] from the mean are averaged, and the square root is extracted.[93]

There are no hard-and-fast rules as to which statistic is the best. In general, the bigger these measures of spread are, the more the numbers are dispersed. Particularly in small data sets, the standard deviation can be influenced heavily by a few outlying values. To remove this influence, the mean and the standard deviation can be recomputed with the outliers discarded.[94] Beyond this, any of the statistics can be supplemented with a figure that displays much of the data.[95]

F. Is an Appropriate Measure of Association Used?

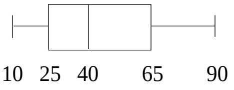
Many cases involve statistical association. Does an employer's requirement of passing a test for promotion have an exclusionary effect that depends on race? Does the salary of workers depend on gender? Does the incidence of murder vary with the rate of executions for convicted murderers? Do consumer purchases of a product depend on the presence or absence of a product warning?

Statistics, such as the mean and the standard deviation, describe each variable in isolation. They do not describe the extent to which two variables are associated. This section will discuss statistics—percentages, proportions, ratios, correlation coefficients, and slopes of regression lines—that can be used to describe the association between two variables.[96]

[92] If a difference is 10, the squared difference is $10 \times 10 = 100$. The mean of the squared differences is known as the *variance*.

[93] The square root of 100 is 10. Taking the square root corrects for the fact that the variance is on a different scale than the measurements themselves. If the measurements are of length in inches, the variance is in square inches. Taking the square root changes back to inches.

[94] Alternatively, a five-number summary, which lists the smallest value, the 25th percentile, the median, the 75th percentile, and the largest value, may be given. The five-number summary may be presented as a boxplot. If the five numbers were 10, 25, 40, 65, and 90, the boxplot would look like the following:



There are many variations on this idea in which the boundaries of the box or the whiskers extending from it represent different points in the distribution.

[95] The measures of variability discussed above depend on the units of measurement. To facilitate comparisons of the variability of different distributions, another statistic known as the *coefficient of variation* often is used. It is the standard deviation expressed as a percentage of the mean. Consider the batch of numbers 1, 4, 4, 7, 9. The mean is 25/5 = 5, the variance is (16 + 1 + 1 + 4 + 16)/5 = 7.6, and the standard deviation is $\sqrt{7.6} = 2.8$. The coefficient of variation is 2.8/5 = 56%.

[96] Even if there is an association, however, there will often be a second issue: Is the association causal? For instance, women may be paid less than men because of gender discrimination; or, the difference may be due to the influence of other *covariates*, such as education or experience. On the question of causation, see *supra* §§ II.C–D, which explains why controlled experiments are the best way to eliminate other variables as possible causes of an observed association.

1. Percentage-related statistics

Percentages often are used to describe the association between two variables. Suppose that a university consisting of only two colleges, engineering and business, admits 550 out of 1,400 students: 350 out of 800 male applicants are admitted, but only 200 out of 600 female applicants are admitted. Such data commonly are displayed in the form of a table:[97]

Table 1
Admissions by Gender

Decision	Male	Female	Total
Admit	350	200	550
Deny	450	400	850
Total	800	600	1,400

The entries in Table 1 indicate that 350/800 = 44% of the men are admitted, compared with only 200/600 = 33% of the women. The resulting selection ratio (used by the Equal Employment Opportunity Commission (EEOC) in its "80% rule")[98] is 33/44 = 75%, meaning that, on average, women have 75% the chance of admission that men have.[99] Another way to express the disparity is to subtract the two percentages: 44 percentage points – 33 percentage points = 11 percentage points.

One difficulty with the simple difference, however, is that it is inevitably small when the two percentages are both close to zero. If the selection rate for men is 5% and that for women is 1%, the difference is only 4 percentage points; yet, on average, women have only 1/5 the chance of men to be selected—and that may be of real concern.

The ratio of the selection rates also has its problems. In the last example, if the selection rates are 5% and 1%, the exclusion rates are 95% and 99%, respectively. The corresponding ratio is 99/95 = 104%, meaning that women have, on average, 104% the chance of men to be rejected. The underlying facts are the same, of course, but this formulation sounds much less disturbing.[100]

[97] A table of this sort also is called a cross-tabulation, or a contingency table. Table 1 is "two-by-two" because it has two rows and two columns, not counting rows or columns containing the totals.

[98] The EEOC generally regards any procedure that selects candidates from the least successful group at a rate less than 80% of the rate for the most successful group as having an adverse impact. EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (1993).

[99] The analogous statistic used in epidemiology is called the relative risk. A variation on this idea is the relative difference in the proportions, which expresses the proportion by which the probability of selection is reduced. Baldus & Cole, *supra* note 1, § 5.1; Kairys et al., *supra* note 20, at 776, 789–90.

[100] The Illinois Department of Employment Security tried to exploit this feature of the ratio in Council 31, Am. Fed'n of State, County & Mun. Employees v. Ward, 978 F.2d 373 (7th Cir. 1992). In January 1985, the department laid off 8.6% of the blacks on its staff in comparison with 3.0% of the whites on its staff. *Id.* at 375. Recognizing that these layoffs ran afoul of the 80% rule if analyzed in terms of those selected to be laid

Another statistic, the *odds ratio,* avoids this asymmetry. If 5% of male applicants are admitted, the *odds* of a man being admitted are 5%/95% = 1/19; the odds of a woman being admitted are 1%/99% = 1/99. The ratio of these quantities is (1/99)/(1/19) = 19/99. The odds ratio for rejection instead of acceptance is the same, except that the order is reversed.[101] Likewise, when the odds of an admitted applicant being a man as opposed to the odds of a denied applicant being a man is considered, the odds ratio also becomes 99/19.

Although the odds ratio has desirable mathematical properties,[102] its meaning may be less clear than that of the selection ratio or the simple difference. To gauge the magnitude of the association implicit in a two-by-two table, any of the statistics presented here may be considered.

Finally, to illustrate the point that association does not necessarily imply causation, consider again the hypothetical admission data in Table 1. Applicants can be classified not only by gender and admission but also by the college to which they applied, as in Table 2:

Table 2
Admissions by Gender and College

	Engineering		Business	
Decision	Male	Female	Male	Female
Admit	300	100	50	100
Deny	300	100	150	300

The entries in Table 2 add up to the entries in Table 1. Yet, there is no association between gender and admission in either college; men and women are admitted in identical percentages.

Combining two colleges with no association produces a university in which gender is associated strongly with admission. The explanation for this paradox: the business college, to which most of the women applied, is hard to get into; the engineering college, to which most of the men applied, is easier to get into. This example illustrates a common issue in discrimination cases: the effect of other, often unreported, variables on an observed association.[103] When a study is

off—since 3.0%/8.6% = 35%, which is far less than 80%—the department instead presented the selection ratio for retention. *Id.* at 375–76. Since black employees were retained at 91.4%/97.0% = 94% of the white rate, use of a retention rate analysis showed no adverse impact. *Id.* at 376. When a subsequent wave of layoffs was challenged as discriminatory, the department argued "that its retention rate analysis is the right approach to this case and that . . [it] shows conclusively that the layoffs did not have a disparate impact," because they comported with the 80% rule. *Id*. at 379. The Seventh Circuit disagreed and, in reversing an order granting summary judgment to defendants on other grounds, left it to the district court on remand "to decide what method of proof is most appropriate." *Id.*

[101] For women, the odds of rejection are 99 to 1; for men, 19 to 1. The ratio of these odds is 99/19.

[102] *See, e.g.*, Finkelstein & Levin, *supra* note 1, at 2–4; Joseph L. Fleiss, Statistical Methods for Rates and Proportions 56–99 (2d ed. 1981); Steve Selvin, Statistical Analysis of Epidemiologic Data app. C (1991).

[103] The example is taken from Moore, *supra* note 12, at 205–06, and inspired by more complex data on graduate admissions in 1973 at the University of California at Berkeley analyzed in P. J. Bickel et al., *Sex Bias*

said to have omitted important variables, some experts find it helpful to consider how large a value of the omitted variable would be needed to explain away the reported results.[104]

2. Correlation coefficients

Two variables are positively correlated when their values tend to go up or down together.[105] Consider the scatter diagram for income and education in Figure 5. As a rule, people with below-average educational levels also have below-average incomes, while people with higher educational levels generally have higher incomes. The association is positive. The correlation coefficient (usually denoted by *r*) is a single number that measures the strength of a linear association. Figure 6 shows the values of *r* for several scatter diagrams.

Figure 6
The correlation coefficient measures the strength of linear association.

$r = 0.0$ $r = 0.5$ $r = 0.9$

insert figure 6

A correlation coefficient of 0 indicates no linear association between the variables, while a coefficient of +1 indicates a perfect linear relationship: All the dots in the scatter diagram fall on a straight line that slopes up. The maximum value for *r* is +1. Sometimes, there is a negative association between two variables. Large values of one variable tend to go with small values of the other. The age of a car and its fuel economy in miles per gallon provide an example. Negative association is indicated by negative values for *r*. The extreme case is an

in Graduate Admissions: Data from Berkeley, 187 Science 398 (1975). *See also* Freedman et al., *supra* note 12, at 16–19. Table 2 is an instance of Simpson's Paradox. *See generally* Myra L. Samuels, *Simpson's Paradox and Related Phenomena* , 88 J. Am. Stat. Ass'n 81 (1993).

[104] *See, e.g.,* Joseph L. Gastwirth, *Methods for Assessing the Sensitivity of Statistical Comparisons Used in Title VII Cases to Omitted Variables*, 33 Jurimetrics J. 19 (1992); Joseph L. Gastwirth, *Employment Discrimination: A Statistician's Look at Analysis of Disparate Impact Claims* , 11 Law & Ineq. J. 151 (1992).

[105] Many statistics and displays are available to investigate correlation. The most common are the correlation coefficient and the scatter diagram.

r of –1, indicating that all the points in the scatter diagram lie on a straight line that slopes down.

Moderate associations are the general rule in the social sciences. Correlations larger than about 0.7 are unusual. For example, the correlation between college grades and first-year law school grades is under 0.3 at most law schools, while the correlation between LSAT scores and first-year law school grades is generally about 0.4.[106] The correlation between heights of fraternal twins is about 0.5, while the correlation between heights of identical twins is about 0.95. In Figure 5, the correlation between income and education is 0.43. The correlation coefficient cannot capture all the underlying information. Several questions may arise in this regard, and we consider them in turn.

a. Is the association linear?

The correlation coefficient is designed to measure linear association. Figure 7 shows a strong nonlinear pattern with a correlation close to 0.

Figure 7

The correlation coefficient only measures linear association. The scatter diagram shows a strong nonlinear association with a correlation coefficient of nearly 0.

insert figure 7 here

b. Do outliers influence the coefficient?

The correlation coefficient can be distorted by outliers—a few points that are far removed from the bulk of the data. The left-hand panel in Figure 8 shows that one outlier (lower right-hand corner) can reduce a perfect correlation to nearly

[106] Linda F. Wightman, Predictive Validity of the LSAT: A National Summary of the 1990–1992 Correlation Studies 10 (1993) (draft final report on data from 167 law schools); *cf.* Linda F. Wightman & David G. Muller, An Analysis of Differential Validity and Differential Prediction for Black, Mexican American, Hispanic, and White Law School Students 11–13 (1990). A combination of LSAT and undergraduate grade point average has a higher correlation with first-year law school grades than either item alone. The multiple correlation coefficient is typically about 0.5. Wightman, *supra* at 10.

nothing. Conversely, the right-hand panel shows that one outlier (upper right-hand corner) can raise a correlation from 0 to nearly 1.

Figure 8

The correlation coefficient can be distorted by outliers. The left-hand panel shows an outlier (in the lower right-hand corner) that destroys a nearly perfect correlation. The right-hand panel shows an outlier (in the upper right-hand corner) that changes the correlation from 0 to nearly 1.

insert figure 8 here

c. Does a third variable influence the coefficient?

The correlation coefficient measures the association between two variables. Investigators—and the courts—may be more interested in causation. However, association is not necessarily the same as causation. Indeed, the association between two variables may be driven largely by a third variable that has been omitted from the analysis. For instance, among schoolchildren, there is an association between shoe size and vocabulary. However, learning more words does not cause feet to grow bigger, and swollen feet do not make children more articulate. In this case, the third variable is easy to spot—age. In more realistic examples, the driving variable may be more difficult to identify.

Of course, in many other examples the association really does reflect causation, but a large correlation coefficient is not enough to warrant this conclusion. Technically, third variables are called "confounders," or "confounding variables." The basic methods for dealing with a confounding variable involve controlled experiments[107] or the application, typically through a technique called *multiple regression,*[108] of *statistical controls*.[109]

[107] *See supra* § II.C.2.

[108] Multiple regression analysis is discussed in Daniel L. Rubinfeld, Reference Guide on Multiple Regression, in this manual.

[109] For the reasons stated *supra* § II.D, efforts to control confounding in observational studies are generally less convincing than randomized controlled experiments.

3. Regression lines

The *regression line* can be used to describe a linear trend in the data. The regression line for income on education is shown in Figure 9. The height of the line estimates the average income for a given educational level. For example, the average income for people with 8 years of education is estimated at $9,600, indicated by the height of the line at 8 years; the average income for people with 16 years of education is estimated at about $23,200.

Figure 9
The regression line for income and education, and its estimates.

insert figure 9

Figure 10 repeats the scatter diagram for income and education (see Figure 5); the regression line is plotted too. In a general way, the line shows the average trend of income as education increases. Thus, the regression line indicates the extent to which a change in one variable (income) is associated with a change in another variable (education).

a. What are the slope and intercept?

The regression line can be described in terms of its slope and intercept.[110] In Figure 10, the slope is $1,700 per year. On average, each additional year of education is associated with an additional $1,700 of income. Next, the intercept is –$4,000. This is an estimate of the average income for people with 0 years of education. The estimate is not a good one, for such people are far from the center

[110] The regression line, like any straight line, has an equation of the form $y = mx + b$. Here, m is the slope, that is, the change in y per unit change in x. The slope is the same anywhere along the line. Mathematically, that is what distinguishes straight lines from curves. The intercept b is the value of y when x is 0. The slope of a line is akin to the grade of a road; the intercept tells you the starting elevation. For example (Figure 9), the regression line estimates an average income of $23,200 for people with 16 years of education. This may be computed from the slope and intercept as follows:

$$(\$1{,}700 \text{ per year}) \times 16 \text{ years} - \$4{,}000 = \$27{,}200 - \$4{,}000 = \$23{,}200$$

of the diagram. In general, estimates based on the regression line become less trustworthy as you move away from the bulk of the data.

Figure 10

Scatter diagram for income and education; the regression line indicates the trend.

insert figure 10

b. What does the slope ignore?

The slope has the same limitations as the correlation coefficient in measuring the degree of association.[111] It only measures linear relationships, it may be influenced by outliers, and it does not control for the effect of other variables. Although the slope of $1,700 per year of education presents each additional year of education as having the same value, some years of schooling surely are worth more and others less. Likewise, the association between education and income graphed in Figure 10 is partly causal, but there are other factors to consider as well, including family backgrounds. People with college degrees probably come from more affluent and better educated families than people who drop out after grade school. They have other advantages besides extra education. Such factors

[111] In fact, the correlation coefficient is the slope of a regression line with the variables in standardized form, that is, measured in terms of standard deviations away from the mean.

must have some effect on income. This is why statisticians use the guarded language of "on average" and "associated with."[112]

c. What is the unit of analysis?

If the association between the characteristics of individuals is of interest, those characteristics should be measured on individuals. Sometimes, however, the individual data are not available, but rates or averages are. "Ecological" correlations are computed from such rates or averages; however, ecological correlations generally overstate the strength of an association. An example makes the point. The Bureau of the Census divides the United States into nine geographic areas. The average income and average education can be determined for the men living in each region. The correlation coefficient for these nine pairs of averages turns out to be 0.7.[113] However, geographic regions do not attend school and do not earn incomes. People do. The correlation for income and education for men in the United States is only about 0.4.[114] The correlation for regional averages overstates the correlation for individuals—a common tendency for such ecological correlations.[115]

Scatter diagrams and regression lines are used often in voting rights cases, where the unit of analysis is the voting precinct. Each point in Figure 11 shows data for a precinct in the 1982 Democratic primary election for auditor in Lee County, South Carolina. The horizontal axis shows the percentage of registrants who are white. The vertical axis shows the turnout rate for the white candidate.[116] The regression line is plotted too.

[112] Many investigators would use multiple regression to isolate the effects of one variable on another—for instance, the independent effect of education on income. Such efforts, like all attempts to infer causation from observational data (*see supra* § II), may run into problems. *See* David A. Freedman, *As Others See Us: A Case Study in Path Analysis*, 12 J. Educ. Stat. 101 (1987).

[113] *See* Freedman et al., *supra* note 12, at 140–41 (using 1988 Current Population Survey).

[114] *Id*. at 140 (using 1988 Current Population Survey).

[115] The ecological correlation uses only the average figures, but within each region there is a lot of spread about the average. The ecological correlation overlooks this individual variation.

[116] By definition, this turnout rate equals the number of votes for the candidate, divided by the number of registrants; the rate is computed separately for each precinct.

Figure 11

Turnout rate for the white candidate plotted against the percentage of registrants who are white. Precinct-level data, 1982 Democratic primary for auditor, Lee County, South Carolina.

insert figure 11

Source: Data are from James W. Loewen & Bernard Grofman, *Recent Developments in Methods Used in Vote Dilution Litigation*, 21 Urb. Law. 589, tbl. 1, at 591 (1989).

In this sort of diagram, the slope is often interpreted as the difference between the white turnout rate and the black turnout rate for the white candidate; the intercept would be interpreted as the black turnout rate for the white candidate. However, the validity of such estimates is contested in the statistical literature. The problem comes from the ecological nature of the regression, that is, making the voting precinct the unit of analysis rather than the individual voter.[117]

[117] The secrecy of the ballot box prevents one from obtaining voting data on individuals, although exit polls may provide some information. For further discussion of the problem of ecological regression in this context, see Symposium, *Statistical and Demographic Issues Underlying Voting Rights Cases,* 15 Evaluation Rev. 659 (1991); James W. Loewen & Bernard Grofman, *Recent Developments in Methods Used in Vote Dilution Litigation,* 21 Urb. Law. 589, tbl. 1, at 591 (1989); Stephen P. Klein & David A. Freedman, *Ecological Regression in Voting Rights Cases*, Chance, Summer 1993, at 38; Stephen P. Klein et al., *Ecological Regression Versus the Secret Ballot,* 31 Jurimetrics J. 393 (1991); Arthur Lupia & Kenneth McCue, *Why the 1980s Measures of Racially Polarized Voting Are Inadequate for the 1990s*, 12 Law & Pol'y 353 (1990).

This page left blank intentionally for proper pagination when printing two-sided

IV. What Inferences Can Be Drawn from the Data?

The inferences that reasonably may be drawn from a study depend on the quality of the data. As discussed in section II, the data may not address the issue of interest, or may be systematically in error, or may be difficult to interpret due to confounding. We turn now to an additional concern—*random error.*[118] Are patterns in the data the result of chance? Would a pattern wash out if more data were collected? If measurements on individual units are unreliable,[119] the errors may combine to produce a false pattern. Even if the measurements on individual units are free from error, the sample may not be representative of the population.

The laws of probability are central to analyzing random error. By applying these laws, the statistician can assess the likely impact of chance error, using *standard errors, confidence intervals, significance probabilities, hypothesis tests,* or *posterior probability distributions*. The following example illustrates the ideas. An employer plans to use a standardized examination to select trainees from a pool of 5,000 male and 5,000 female applicants. This total pool of 10,000 applicants is the statistical population. Under Title VII of the Civil Rights Act, if the proposed examination excludes a disproportionate number of women, the employer must show that the exam is job related.[120]

To see whether there is disparate impact, the employer administers the exam to a sample of 50 men and 50 women drawn at random from the population of job applicants. In the sample, 29 of the men but only 19 of the women pass; the sample pass rates are therefore 29/50 = 58% and 19/50 = 38%. The employer announces that it will use the exam anyway, and several applicants bring an action under Title VII.

Disparate impact seems clear. The difference in sample pass rates is 20 percentage points: 58% – 38% = 20%. The employer argues, however, that the disparity could just reflect random error. After all, only a small number of people

[118] Random error is also called *sampling error*, *chance error,* or statistical error. Econometricians use the parallel concept of *random disturbance term.*

[119] *See supra* § II.A.1.

[120] The seminal case is Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971). The requirements and procedures for the validation of tests can go beyond a simple showing of job-relatedness. *See, e.g.*, Richard R. Reilly, *Validating Employee Selection Procedures*, *in* Statistical Methods in Discrimination Litigation, *supra* note 7, at 133; Michael Rothschild & Gregory J. Werden, *Title VII and the Use of Employment Tests: An Illustration of the Limits of the Judicial Process*, 11 J. Legal Stud. 261 (1982).

took the test, and the sample just may have happened to include disproportionate numbers of high-scoring men and low-scoring women. Clearly, even if there was no overall difference in pass rates for male and female applicants, in some samples men will outscore women. A statistician then might be asked to address such topics as the following:

- *Estimation*. Plaintiffs use the difference of 20 percentage points between the sample men and women to estimate the disparity between all male and female applicants. How good is this estimate? Precision can be expressed using the standard error or a confidence interval.
- *Statistical Significance*. Suppose the defendant is right—in the population of all 5,000 male and 5,000 female applicants, the pass rates are equal; there is no disparate impact. How likely is it that a random sample of 50 men and 50 women will produce a disparity of 20 percentage points or more? This chance is known as a *p*-value. Statistical significance is determined by reference to the *p*-value, and hypothesis testing is the technique for computing *p*-values or determining statistical significance.[121]
- *Posterior probability*. Given the observed disparity of 20 percentage points in the sample, what is the probability that—in the population as a whole—men and women have equal pass rates? This question is of direct interest to the courts. However, within the framework of classical statistical theory, such a posterior probability has no meaning.[122] For a subjectivist statistician, posterior probabilities may be computed using *Bayes' rule*.

A. Estimation

1. What estimator should be used?

An *estimator* is a statistic computed from sample data and used to estimate a numerical characteristic of the population. For example, the difference in pass rates for a sample of men and women is used to estimate the corresponding disparity in the population of all applicants. In our sample, the pass rates were 58% and 38%; the difference in pass rates for the whole population is estimated to be 20 percentage points: 58% – 38% = 20%. In more complex problems, statisticians may have to choose among several estimators. Generally, estimators that tend to make smaller errors are preferred. However, this idea can be made precise in more than one way,[123] leaving room for judgment in selecting an estimator.

[121] Hypothesis testing is also called significance testing.

[122] This classical framework is also called "objective" or "frequentist." Contrast with the subjectivist approach; *see infra* § IV.C.

[123] Furthermore, reducing error in one context may increase error in other contexts; there may also be a trade-off between accuracy and simplicity.

2. What is the standard error?

The estimate of 20 percentage points is likely to be off, at least by a little, due to random error. The standard error gives the likely magnitude of this random error.[124] Whenever possible, an estimate should be accompanied by its standard error.[125] In our example, the standard error is about 10 percentage points: The estimate of 20 percentage points is likely to be off by about 10 percentage points or so, in either direction.[126] Since the pass rates for all 5,000 men and 5,000 women are unknown, we cannot say exactly how far off the estimate is going to be, but 10 percentage points gauges the likely magnitude of the error.

Confidence intervals make the idea more precise. Statisticians who say the population difference falls within plus-or-minus 1 standard error of the sample difference would be correct about 68% of the time. To write this more compactly, we can abbreviate standard error as SE. A 68% confidence interval is the range

estimate – 1 SE   to   estimate + 1 SE

In our example, the 68% confidence interval goes from 10 to 30 percentage points. If a higher confidence level is wanted, the interval must be widened. The 95% confidence interval is about

estimate – 2 SE   to   estimate + 2 SE

This runs from 0 to 40 percentage points. Although 95% confidence intervals are used commonly, there is nothing special about 95%. For example, a 99.7% confidence interval is about

estimate – 3 SE   to   estimate + 3 SE

This stretches from –10 to 50 percentage points.[127]

[124] Standard errors are also called standard deviations, and courts seem to prefer the latter term, as do many authors. *See infra* notes 145, 149.

[125] The standard error can also be used to measure reproducibility of estimates from one random sample to another. *See infra* the Appendix.

[126] The standard error depends on the pass rates of men and women in the sample and on the size of the sample. Chance error is smaller for larger samples, so the standard error goes down as sample size goes up. The Appendix gives the formula for computing the standard error of a difference in rates based on random samples. Generally, the formula for the standard error must take into account the method used to draw the sample and the nature of the estimator. Statistical expertise is needed to choose the right formula.

[127] A negative value, such as -10%, indicates that for the whole population, more women than men are estimated to pass the test. The 68%, 95%, and 99.7% come from the normal curve. *See infra* the Appendix. When there are samples of reasonable size, an estimator like the pass rate difference will follow the normal curve fairly well. Statisticians call this the *central limit theorem*. The probability that our estimator will be within 2 standard errors of the true population figure is approximately equal to the area under the normal curve between -2 and +2. This area is about 95%. For a more complete description of the normal curve and its use in large samples, see, *e.g.*, Freedman et al., *supra* note 12, at 73–89, 282–302. Of course, many estimators do not follow the normal curve, and other procedures then must be used to obtain confidence intervals.

A confidence interval is based on the standard error. If the standard error is small, the estimate probably is close to the truth. If the standard error is large, the estimate may be seriously wrong.

3. What do standard errors and confidence intervals mean?

An estimate based on a sample will differ from the exact population value due to random error; the standard error measures the likely size of the random error. Confidence intervals are a technical refinement, and confidence is a term of art.[128] For a given confidence level, a narrower interval indicates a more precise estimate. For a given sample size, increased confidence can be attained only by widening the interval. A high confidence level alone means very little, but a high confidence level resulting in a small interval is impressive.[129] It indicates that the random error in the sample estimate is low.

Both the standard error and the confidence interval are derived using a particular model of statistical error. A statistical model expresses the way random error works and generally contains parameters that characterize the population from which the samples were drawn.[130] The data in our example came from a random sample, and that guaranteed the validity of the statistical calculations.[131]

[128] In the standard frequentist theory of statistics, one cannot make probability statements about population characteristics. *See, e.g.*, Freedman et al., *supra* note 12, at 351–53; *infra* § IV.B.1. Because of the limited technical meaning of confidence, it has been argued that the term is misleading and should be replaced by a more neutral one, such as frequency coefficient, in courtroom presentations. David H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 Wash. L. Rev. 1333, 1354 (1986).

[129] Conversely, a broad interval signals that random error is substantial. In Cimino v. Raymark Indus., Inc., 751 F. Supp. 649 (E.D. Tex. 1990), the district court drew certain random samples from more than 6,000 pending asbestos cases, tried these cases, and used the results to estimate the total award to be given to all plaintiffs in the 6,000 cases. The court then held a hearing to determine whether the samples were large enough to provide accurate estimates. *Id.* at 664. The court's expert, an educational psychologist, testified that the estimates were accurate because the samples matched the population on such characteristics as race and the percentage of plaintiffs still alive. *Id*. However, the matches occurred only in the sense that population characteristics fell within very broad 99% confidence intervals computed from the samples. The court thought that matches within the 99% confidence intervals proved more than matches within 95% intervals. *Id*. Unfortunately, this is backwards. It is not very impressive to be correct in a few instances with a 99% confidence interval, because, by definition, such intervals are broad enough to ensure coverage 99% of the time. *Cf.* Michael J. Saks & Peter David Blanck, *Justice Improved: The Unrecognized Benefits of Aggregation and Sampling in the Trial of Mass Torts*, 44 Stan. L. Rev. 815 (1992).

[130] In our example, one parameter is the pass rate of the 5,000 male applicants; another parameter is the pass rate of the 5,000 female applicants. These two parameters determine the probabilities of observing the various possible values for the sample difference, according to a set of mathematical equations. The statistical problem consists of working backwards from the sample data to the population parameters.

When the parameters are known, the analyst may use the model to find the probability of an observed outcome (or one like it). This approach is common in cases alleging discrimination in the selection of jurors. *E.g.*, Castaneda v. Partida, 430 U.S. 482, 496 (1977); David H. Kaye, *Statistical Evidence of Discrimination in Jury Selection, in* Statistical Methods in Discrimination Litigation, *supra* note 7, at 13. *Cf.* Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 311 n.17 (1977) (computing probabilities of selecting black teachers). Although such problems in applied probability theory are not explicitly treated in this reference guide, the Appendix presents some relevant calculations.

[131] Partly because the Supreme Court used models giving rise to variables that are approximately normal in *Hazelwood* and *Castaneda*, courts and attorneys sometimes are skeptical of models and analyses that produce other types of random variables. *See, e.g.*, EEOC v. Western Elec. Co., 713 F.2d 1011 (4th Cir. 1983), *discussed in* David H. Kaye, *Ruminations on Jurimetrics: Hypergeometric Confusion in the Fourth Circuit,* 26 Jurimetrics J. 215 (1986). *But cf*. Branion v. Gramly, 855 F.2d 1256 (7th Cir. 1988) (questioning an apparently

The choice of an appropriate model in other situations may be less obvious.[132] When a model does not describe well the process giving rise to the data, the estimate and its standard error are less probative.[133]

Furthermore, the standard error and the confidence interval generally ignore systematic errors, such as selection bias or nonresponse bias.[134] For example, one court—reviewing studies of whether a particular drug causes birth defects—observed that mothers of children with birth defects may be more likely to remember taking a drug during pregnancy than women with normal children.[135] This selective recall would bias comparisons between samples from the two groups of women. The standard error for the estimated difference in drug usage between the two groups ignores this bias, as does the confidence interval.[136]

arbitrary assumption of normality), *cert. denied*, 490 U.S. 1008 (1989), *discussed in* David H. Kaye, *Statistics for Lawyers and Law for Statistics,* 89 Mich. L. Rev. 1520 (1991). Whether a given variable is normally distributed is an empirical or statistical question, not a matter of law. That a particular model has been used in a previous case may be of limited value in deciding whether it is appropriate in the case at bar. *See generally* Statistical Methods in Discrimination Litigation, *supra* note 7, at iii; Laurens Walker & John Monahan, *Social Facts: Scientific Methodology as Legal Precedent,* 76 Cal. L. Rev. 877 (1988).

[132] For examples of legal interest, see, *e.g.*, Mary W. Gray, *Can Statistics Tell Us What We Do Not Want to Hear?: The Case of Complex Salary Structures*, 8 Stat. Sci. 144 (1993); Arthur P. Dempster, *Employment Discrimination and Statistical Science*, 3 Stat. Sci. 149 (1988). One statistician describes the issue as follows:

> [A] given data set can be viewed from more than one perspective, can be represented by a model in more than one way. Quite commonly no unique model stands out as "true" or correct; justifying so strong a conclusion might require a depth of knowledge that is simply lacking. So it is not unusual for a given data set to be analyzed in several apparently reasonable ways. If conclusions are qualitatively concordant, that is regarded as grounds for placing additional trust in them. But more often, only a single model is applied, and the data are analyzed in accordance with it . . . .
>
> Desirable features in a model include (i) tractability, (ii) parsimony, and (iii) realism. That there is some tension among these is not surprising.
>
> *Tractability*. A model that is easy to understand and to explain is tractable in one sense. Computational tractability can also be an advantage, though with cheap computing available not too much weight can be given to it.
>
> *Parsimony*. Simplicity, like tractability, has a direct appeal, not wisely ignored—but not wisely over-valued either. If several models are plausible and more than one of them fits adequately with the data, then in choosing among them, *one* criterion is to prefer a model that is simpler than the other models.
>
> *Realism* . . . . First, does the model reflect well the actual . . . [process that generated the data]? This question is really a host of questions, some about the distributions of the random errors, others about the mathematical relations among the [variables and] parameters. The second aspect of realism is sometimes called robustness: If the model is *false* in certain respects, how badly does that affect estimates, significance test results, etc., that are based on the flawed model?

Lincoln E. Moses, *The Reasoning of Statistical Inference,* *in* Perspectives on Contemporary Statistics, *supra* note 22, at 107, 117–18.

[133] It still may be helpful to consider the standard error, perhaps as a minimal estimate for statistical uncertainty.

[134] For a discussion of such systematic errors, see *supra* § II.B.

[135] Brock v. Merrell Dow Pharmaceuticals, Inc., 874 F.2d 307, 311–12 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989), *cert. denied,* 494 U.S. 1046 (1990).

[136] In *Brock*, the court held that the confidence interval took account of bias (in the form of selective recall) as well as random error. 874 F.2d at 311–12. With respect, we disagree. Even if sampling error were nonexistent, which would be the case if one could interview every woman who had a child in the period that the drug was available, selective recall would produce a difference in the percentages of reported drug exposure between mothers of children with birth defects and those with normal children. In this hypothetical situa-

Likewise, the standard error does not address problems inherent in using convenience samples rather than random samples.[137]

B. *p*-values and Hypothesis Tests

1. What is the *p*-value?

In our example, 50 men and 50 women were drawn at random from 5,000 male and 5,000 female applicants. An exam was administered to this sample, and in the sample, the pass rates for the men and women were 58% and 38%, respectively; the sample difference in pass rates was 58 – 38 = 20 percentage points. The *p*-value answers the following question: If the pass rates among all 5,000 male applicants and 5,000 female applicants were identical, how probable would it be to find a discrepancy as large as or larger than the 20 percentage point difference observed in our sample? The question is delicate, because the pass rates in the population are unknown—that is why a sample was taken in the first place.

The assertion that the pass rates in the population are the same is called the *null hypothesis*. The null hypothesis asserts that there is no difference between men and women in the whole population—differences in the sample are due to the luck of the draw. The *p*-value is the probability of getting data as extreme as, or more extreme than, the actual data, given that the null hypothesis is true:

$$p = \Pr(\text{extreme data} \mid \text{null hypothesis in model})$$

If the null hypothesis is true, there is only a 5% chance of getting a difference in the pass rates of 20 percentage points or more.[138] The *p*-value for the observed discrepancy is 5%, or .05.

In such examples, small *p*-values are evidence of disparate impact, while large *p*-values are evidence against disparate impact. Regrettably, multiple negatives are involved here. The null hypothesis asserts no difference in the population—that is, no disparate impact. Small *p*-values argue against the null hypothesis; that is, small *p*-values argue there is disparate impact. Generally, by indicating that the magnitude of the observed difference is improbable if the null hypothesis is true, small *p*-values undermine the null hypothesis. The smaller the *p*-value for a given study, the more surprising it would be to see such differences under the null hypothesis. Conversely, large *p*-values indicate that the data are compatible with the null hypothesis.

However, since *p* is calculated by assuming the null hypothesis, the *p*-value cannot give the chance that this hypothesis is true. The *p*-value merely gives the

tion, the standard error would vanish. Therefore, the standard error can disclose nothing about the impact of selective recall.

[137] *See supra* § II.B.1.

[138] *See infra* the Appendix.

chance of getting evidence against the null hypothesis as strong or stronger than the evidence at hand—assuming the null hypothesis is correct. No matter how many samples are obtained, the null hypothesis is either always right or always wrong. Chance affects the data, not the hypothesis. With the frequency interpretation of chance, there is no meaningful way to assign a numerical probability to the null hypothesis, or to any alternative hypothesis, for that matter.[139]

Computing *p*-values requires statistical expertise. Many methods are available, but only some will fit the occasion.[140] Sometimes standard errors will be part of the analysis, while other times they will not be. Sometimes a difference of 2 standard errors will imply a *p*-value of about .05, other times it will not. In general, the *p*-value depends on the model and its parameters, the size of the sample, and the sample statistics.

Because the *p*-value is affected by sample size, it does not measure the extent or importance of a difference.[141] Suppose, for instance, that the 5,000 male and 5,000 female job applicants would differ in their pass rates, but only by 1 percentage point. This difference might not be enough to make a case of disparate impact, but by including enough men and women in the sample, the data could be made to have an impressively small *p*-value. This *p*-value would confirm that the 5,000 men and 5,000 women have different pass rates, but it would not show the difference is substantial.

Likewise, in considering whether two quantities are correlated[142] in a population from which a random sample has been drawn, the *p*-value depends on the correlation in the sample as well as on the number of data points. Statistical significance may result from a small correlation and a large number of points. In short, the *p*-value does not measure the strength or importance of an association.[143]

[139] *See, e.g.,* The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 196–98; David H. Kaye, *Statistical Significance and the Burden of Persuasion,* Law & Contemp. Probs., Autumn 1983, at 13. Some opinions suggest a contrary view. *E.g.,* Fudge v. Providence Fire Dep't, 766 F.2d 650, 658 (1st Cir. 1985) ("Widely accepted statistical techniques have been developed to determine the likelihood an observed disparity resulted from mere chance."); Capaci v. Katz & Besthoff, Inc., 711 F.2d 647, 652 (5th Cir. 1983) ("the highest probability of unbiased hiring was 5.367×10^{-20}"), *cert. denied,* 466 U.S. 927 (1984). Such statements appear to confuse the probability of the kind of outcome observed, which is computed under some model of chance, with the probability that chance is the explanation for the outcome. (In scientific notation, 10^{20} is one followed by twenty zeros, and 10^{-20} is the reciprocal of that number. The proverbial "one in a million" is more dryly expressed as 1×10^{-6}.)

[140] *See, e.g.*, Thomas J. Sugrue & William B. Fairley, *A Case of Unexamined Assumptions: The Use and Misuse of the Statistical Analysis of Castaneda/Hazelwood in Discrimination Litigation,* 24 B.C. L. Rev. 925 (1983).

[141] Some opinions seem to equate small *p*-values with gross or substantial disparities. *E.g.*, Craik v. Minnesota St. Univ. Bd., 731 F.2d 465, 479 (8th Cir. 1984). Other courts have emphasized the need to decide whether the underlying sample statistics reveal that a disparity is large. *E.g.*, McCleskey v. Kemp, 753 F.2d 877, 892–94 (11th Cir. 1985), *aff'd*, 481 U.S. 279 (1987).

[142] *See supra* § III.F.2.

[143] The conventional procedures used to compute a *p*-value for a correlation depend on the normality of the underlying process for generating the data. The scatter diagram itself gives some useful clues as to whether this assumption is satisfied. Basically, the scatter diagram should be roughly circular or oval in shape. The diagrams in Figure 6 confirm the assumption of normality. The diagram in Figure 5 is incompatible with the assumption, because the cloud of points widens as one moves from left to right along the horizontal axis. In the

2. Is a difference statistically significant?

Statistical significance is determined by comparing a *p*-value to a preestablished value, the *significance level*.[144] If an observed difference is in the middle of the distribution that would be expected under the null hypothesis, there is no surprise. The sample data are of the type that often would be seen when the null hypothesis is true: The difference is not significant, and the null hypothesis cannot be rejected. Conversely, if the sample difference is far from the expected value—according to the null hypothesis—the sample is unusual: The difference is significant, and the null hypothesis is rejected. In our example, the 20 percentage point difference in pass rates for the men and women in the sample, whose *p*-value was about .05, would be significant at the .05 level. If the threshold were set lower, for instance at .01, the result would not be significant.

In practice, statistical analysts use certain preset significance levels—typically .05 or .01.[145] The .05 level is the most common in social science, and an analyst who speaks of "significant" results without specifying the threshold probably is using this level.[146] An unexplained reference to "highly significant" results probably means that *p* is less than .01.[147]

Since the term "significant" is merely a label for certain kinds of *p*-values, it is subject to the same limitations as are *p*-values themselves. Significant differences are evidence that something besides random error is at work, but they are not evidence that this "something" is legally or practically important. Statisticians distinguish between statistical and practical significance to make the point. When

jargon of the field, Figure 5 shows *heteroscedasticity,* while the diagrams in Figure 6 are *homoscedastic.* Figures 7 and 8 are also incompatible with the normality assumption; Figure 7 shows a strong nonlinear pattern, while Figure 8 has outliers. Under other circumstances, too, procedures may be available to test the adequacy of a model's assumptions. *See, e.g.*, David C. Hoaglin, *Diagnostics*, *in* Perspectives on Contemporary Statistics, *supra* note 22, at 123; David A. Belsley et al., Regression Diagnostics: Identifying Influential Data and Sources of Collinearity (1980). Sometimes, however, assumptions are tested only by introducing other assumptions that are even more obscure.

[144] Statisticians use the Greek letter α (*alpha*) to denote the significance level; α gives the chance of obtaining a significant result, assuming that the null hypothesis is true. Thus, α represents the chance of what is variously termed a false rejection of the null hypothesis, a type I error, a false positive, or a false alarm. For example, suppose $\alpha = 5\%$. If investigators do many studies, and the null hypothesis happens to be true in each case, then about 5% of the time they would obtain significant results—and falsely reject the null hypothesis.

[145] The Supreme Court implicitly referred to this practice in Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977), and Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 311 n.17 (1977). In these footnotes, the Court described the null hypothesis as "suspect to a social scientist" when a statistic from "large samples" falls more than "two or three standard deviations" from its expected value under the null hypothesis. Although the Court did not say so, these differences produce *p*-values of about .05 and .01 when the statistic is normally distributed. The Court's standard deviation is our standard error.

[146] Some have intimated that data not significant at the .05 level should be disregarded. This view is challenged in, *e.g.*, Kaye, *supra* note 128, at 1344 & n.56, 1345. *But see* Paul Meier et al., *What Happened in Hazelwood: Statistics, Employment Discrimination, and the 80% Rule,* 1984 Am. B. Found. Res. J. 139, 152.

[147] Merely labeling results as significant or not significant without providing the underlying information that goes into this conclusion is of limited value. *See, e.g.*, John C. Bailar, III & Frederick Mosteller, *Guidelines for Statistical Reporting in Articles for Medical Journals: Amplifications and Explanations, in* Medical Uses of Statistics, *supra* note 44, at 313, 316.

practical significance is lacking—when the size of a disparity or correlation is negligible—there is no reason to worry about statistical significance.[148]

As noted above, it is easy to mistake the *p*-value for the probability that there is no difference. Likewise, if results are significant at the .05 level, it is tempting to conclude that the null hypothesis has only a 5% chance of being correct.[149] This temptation should be resisted. From the frequentist perspective, statistical hypotheses are either true or false—probabilities govern the samples, not the models and hypotheses. The significance level tells us what is likely to happen when the null hypothesis is correct; it cannot tell us the probability that the hypothesis is true. Significance comes no closer to expressing the probability that the null hypothesis is true than does the underlying *p*-value.[150]

3. Questions about hypothesis tests

a. What is the power of the test?

When a *p*-value is high, findings are not significant, and the null hypothesis is not rejected. There are at least two possible explanations:

1. There is no difference in the population—the null hypothesis is true; or
2. There is some difference in the population—the null hypothesis is false—but, by chance, the data are of the kind expected under the null hypothesis.[151]

If the *power* of a statistical study is low, the second is a reasonable explanation for the data. Power is the chance that a statistical test will declare an effect when there is an effect to declare.[152] This chance depends on the size of the effect and

[148] *E.g.*, Waisome v. Port Auth., 948 F.2d 1370, 1376 (2d Cir. 1991) ("[T]hough the disparity was found to be statistically significant, it was of limited magnitude . . . .") (citations omitted).

[149] *E.g., id.* at 1376 ("Social scientists consider a finding of two standard deviations significant, meaning there is about one chance in 20 that the explanation for a deviation could be random . . . ."); Rivera v. City of Wichita Falls, 665 F.2d 531, 545 n.22 (5th Cir. 1982) ("A variation of two standard deviations would indicate that the probability of the observed outcome occurring purely by chance would be approximately five out of 100; that is, it could be said with a 95% certainty that the outcome was not merely a fluke."); Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224, 272 (N.D. Tex. 1980) ("[I]f a 5% level of significance is used, a sufficiently large t-statistic for the coefficient indicates that the chances are less than one in 20 that the true coefficient is actually zero."), *vacated*, 723 F.2d 1195 (5th Cir.), *cert. denied*, 469 U.S. 1073 (1984).

[150] For more discussion, see Kaye, *supra* note 139.

[151] Tests also may reject—or fail to reject—because the statistical model does not fit the situation. *See infra* § IV.B.3.e.

[152] More precisely, power is the probability of rejecting the null hypothesis when the alternative hypothesis is right. Typically, this depends on the values of unknown parameters, as well as on the preset significance level (α). *See supra* notes 130, 144. Therefore, no single number gives the power of the test. The expert can specify particular values for the parameters and significance level and compute the power of the test accordingly. *See infra* the Appendix for an example. Power may be denoted by the Greek letter β (*beta*).

Accepting the null hypothesis when the alternative is true is known as a false acceptance of the null hypothesis, a type II error, a false negative, or a missed signal. The chance of a false negative may be computed from the power, as $1 - \beta$. Frequentist hypothesis testing keeps the risk of a false positive to a specified level (such as $\alpha = .05$) and then tries to minimize the chance of a false negative ($1 - \beta$) for that value of α. Regrettably, the notation is in some degree of flux; many authors use β to denote the chance of a false negative; then, it is β that should be minimized.

the size of the sample. Discerning subtle differences in the population requires large samples.

When a study with low power fails to show a significant effect, one should not treat the negative result as strong proof that there is no effect. The study is described more fairly as inconclusive than as negative.[153] In contrast, when studies have a good chance of detecting a meaningful association, failure to obtain significant findings can be persuasive evidence that there is no effect to be found.[154]

b. One-tailed versus two-tailed tests

In many cases, a statistical test can be either *one-tailed* or *two-tailed*. The second method will generally produce a *p*-value twice as big as the first method. Since small *p*-values are evidence against the null hypothesis, a one-tailed test seems to produce stronger evidence than a two-tailed test. However, this difference is largely illusory.[155]

Some courts have expressed a preference for two-tailed tests,[156] but a rigid rule is not required if *p*-values and significance levels are used as clues rather than as

Some commentators have claimed that the cutoff for significance should be chosen to equalize the chance of a false positive and a false negative, on the ground that this criterion corresponds to the "more-probable-than-not" burden of proof. Unfortunately, the argument is fallacious because α and β apply to data, not hypotheses. *See supra* § IV.B.1.

[153] In our pass rate example, with $\alpha = .05$, power to detect a difference of 10 percentage points between the male and female job applicants is only about 1/6. *See infra* the Appendix. Not seeing a "significant" difference therefore provides only weak proof that the difference between men and women is smaller than 10 percentage points. We prefer estimates accompanied by standard errors to tests, because the former seem to make the state of the statistical evidence clearer: The estimated difference is 20 ± 10 percentage points, indicating that a difference of 10 percentage points is quite compatible with the data.

[154] Some formal procedures are available to aggregate results across studies. *See In re* Paoli R.R. Yard PCB Litig., 916 F.2d 829 (3d Cir. 1990), *cert. denied sub nom.* General Elec. Co. v. Knight, 499 U.S. 961 (1991). In principle, the power of the collective results will be greater than the power of each study. *See, e.g.*, The Handbook of Research Synthesis 226–27 (Harris Cooper & Larry V. Hedges eds., 1994); Larry V. Hedges & Ingram Olkin, Statistical Methods for Meta-Analysis (1985); Jerome P. Kassirer, *Clinical Trials and Meta-Analysis: What Do They Do for Us?*, 327 New Eng. J. Med. 273, 274 (1992) ("[C]umulative meta-analysis represents one promising approach."); National Research Council, Combining Information: Statistical Issues and Opportunities for Research (1992). Unfortunately, these procedures have their own limitations. *E.g.*, Diana Petitti, Meta-Analysis, Decision Analysis, Cost-Effectiveness Analysis in Medicine: Methods for Quantitative Synthesis of Information (1994); Michael Oakes, Statistical Inference: A Commentary for the Social and Behavioral Sciences 157 (1986) ("a retrograde development"); Charles Mann, *Meta-Analysis in the Breech*, 249 Science 476 (1990).

[155] In our pass rate example, the *p*-value of the test is approximated by a certain area under the normal curve. The one-tailed procedure uses the tail area under the curve to the right of 2, giving $p = .025$. The two-tailed procedure contemplates the area to the left of -2, as well as the area to the right of 2. Now there are two tails, and $p = .05$. According to formal statistical theory, the choice between one tail and two sometimes can be made by considering the exact form of the alternative hypothesis. The null hypothesis held that pass rates were equal for men and women in the whole population of applicants. The alternative hypothesis may exclude a priori the possibility that women have a higher pass rate and hold that more men will pass than women. This asymmetric alternative suggests a one-tailed test. Conversely, the alternative hypothesis may simply be that pass rates for men and women in the whole population are unequal. This symmetric alternative admits the possibility that women may score higher than men and points to a two-tailed test. *See, e.g.*, Freedman et al., *supra* note 12, at 495–98 .

[156] *See, e.g.*, Baldus & Cole, *supra* note 1, at 308 n.35a (1980 & Supp. 1987); The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 38–40 (citing EEOC v. Federal Reserve

mechanical devices for deferring to or dismissing statistical proofs. One-tailed tests make it easier to reach a threshold like .05, but if .05 is not used as a magic line, then the choice between one tail and two is less important—as long as the choice and its effect on the *p*-value are made explicit.[157]

c. How many tests have been performed?

Repeated applications of significance testing complicate the interpretation of a significance level. If enough studies are conducted, random error almost guarantees that some will yield significant findings, even when there is no real effect.[158] Consider the problem of deciding whether a coin is biased. The probability that a fair coin will produce ten heads when tossed ten times is $(1/2)^{10} = 1/1{,}024$. Observing ten heads in the first ten tosses, therefore, would be strong evidence that the coin is biased. Nevertheless, if a fair coin is tossed a few thousand times, it is likely that at least one string of ten consecutive heads will appear. The test—looking for a run of ten heads—has been repeated far too often.[159]

The problem of multiple testing can affect statistical models with many possible equations and parameters. Almost any large data set—even pages from a table of random digits—will contain some unusual pattern that can be uncovered by a diligent search. Having detected the pattern, the analyst can perform a statistical test for it, blandly ignoring the search effort. Statistical significance is bound to follow. Ten heads in the first ten tosses means one thing; a run of ten heads in a few thousand tosses of a coin means another.

There are statistical methods for coping with multiple looks at the data, which permit the calculation of meaningful *p*-values in certain cases.[160] However, no general solution is available, and the existing methods would be of little help in the typical case where analysts have run through a variety of regression models to arrive at the one considered the most satisfactory. In these situations, courts

Bank, 698 F.2d 633 (4th Cir. 1983), *rev'd on other grounds sub nom.* Cooper v. Federal Reserve Bank, 467 U.S. 867 (1984)); Kaye, *supra* note 128, at 1358 n.113; David H. Kaye, *The Numbers Game: Statistical Inference in Discrimination Cases*, 80 Mich. L. Rev. 833 (1982) (citing Hazelwood Sch. Dist. v. United States, 433 U.S. 299 (1977)). An argument for one-tailed tests is made by Richard Goldstein, *Two Types of Statistical Errors in Employment Discrimination Cases*, 26 Jurimetrics J. 32 (1985).

[157] One-tailed tests at the .05 level are viewed as weak evidence—no weaker standard is commonly used in the technical literature. *But see* Richard Lempert, *Statistics in the Courtroom: Building on Rubinfeld*, 85 Colum. L. Rev. 1098, 1099 (1985) ("[T]he values of social science are not the values of law.").

[158] Since research that fails to uncover significance is not usually published, reviews of the literature may produce an unduly large number of studies finding statistical significance. *E.g.*, Stuart J. Pocock et al., *Statistical Problems in the Reporting of Clinical Trials: A Survey of Three Medical Journals*, 317 New Eng. J. Med. 426 (1987).

[159] For advice on spotting comparable abuses in biomedical studies, see James L. Mills, *Data Torturing*, 329 New Eng. J. Med. 1196 (1993).

[160] *See, e.g.*, Yosef Hochberg & Ajit C. Tamhane, Multiple Comparison Procedures (1987); Rupert G. Miller, Jr., Simultaneous Statistical Inference (2d ed. 1981); Peter H. Westfall & S. Stanley Young, Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment (1993); Joseph L. Gastwirth & Samuel W. Greenhouse, *Estimating a Common Relative Risk: Application in Equal Employment,* 82 J. Am. Stat. Ass'n 38 (1987); Robert Follett & Finis Welch, *Testing for Discrimination in Employment Practices,* Law & Contemp. Probs., Autumn 1983, at 171; Kaye, *supra* note 130, at 13.

should not be overly impressed with claims that estimates are significant. Instead, they should be asking how analysts developed their models.[161]

d. What are the interval estimates?

Statistical significance depends on the *p*-value, and the *p*-value depends on sample size. Therefore, a significant effect may be small. Conversely, an effect that is not significant may be large.[162] By inquiring into the magnitude of an effect, courts can avoid being misled by *p*-values. To focus attention where it belongs—on the actual size of an effect and the reliability of the statistical analysis—the court may ask for an *interval estimate.*[163] Seeing a plausible range of values for the quantity of interest enables the court to decide whether this quantity is large or small and to consider the statistical uncertainty in the estimate.

In our example, the 95% confidence interval for the difference in the pass rates of men and women ranged from 0 to 40 percentage points. Our best estimate is that the pass rate for men is 20 percentage points higher than that for women; and the difference may plausibly be as little as 0 or as much as 40 percentage points. The *p*-value does not yield this information. The confidence interval contains the information provided by a significance test—and more.[164] For instance, significance at the .05 level can be read off the 95% confidence interval. In our example, 0 is at the extreme edge of the 95% confidence interval; thus, we have significant evidence that the true difference in pass rates between male and female applicants is not 0. But there are values very close to 0 inside the interval. This may help us consider whether the difference is practically significant.

In contrast, suppose a significance test fails to reject the null hypothesis. The confidence interval may prevent the mistake of thinking there is positive proof for the null hypothesis. To illustrate, let us change our example slightly: 29 men and 20 women passed the test. The 95% confidence interval goes from –2 to 38 percentage points. Because a difference of 0 falls within the 95% confidence interval, the null hypothesis—that the true difference is 0—cannot be rejected at the .05 level. However, the interval extends to 38 percentage points, indicating

[161] *See, e.g*., Persi Diaconis, *Theories of Data Analysis: From Magical Thinking Through Classical Statistics, in* Exploring Data Tables, Trends, and Shapes 1, 8–9 (David C. Hoaglin et al. eds., 1985); Frank T. Denton, *Data Mining As an Industry,* 67 Rev. Econ. & Stat. 124 (1985); David A. Freedman, *A Note on Screening Regression Equations,* 37 Am. Statistician 152 (1983). Intuition may suggest that the more variables included in the model, the better. However, this idea often seems to be wrong. Complex models may reflect only accidental features of the data. Standard statistical tests offer little protection against this possibility when the analyst has tried a variety of models before settling on the final specification.

[162] *See supra* § IV.B.1.

[163] An interval estimate may be composed of a *point estimate* —like the sample mean used to estimate the population mean—together with its standard error, or the two can be combined into a confidence interval. The first alternative may be more informative.

[164] Accordingly, it has been argued that courts should demand confidence intervals (whenever they can be computed) to the exclusion of explicit significance tests and *p*-values. Kaye, *supra* note 128, at 1349 n.78; *cf*. Bailar & Mosteller, *supra* note 147, at 317.

that the population difference could be substantial. Lack of significance does not exclude this possibility.[165]

e. What are the other explanations for the findings?

The *p*-value of a statistical test is computed on the basis of a model for the data—the null hypothesis. Usually, the test is made in order to argue for the *alternative hypothesis*—another model. However, on closer examination, both models may prove to be unreasonable.[166] A small *p*-value indicates the occurrence of something besides random error; the alternative hypothesis should be viewed as one possible explanation out of many for the data.[167]

In *Mapes Casino, Inc. v. Maryland Casualty Co*.,[168] for example, the court recognized the importance of explanations that the proponent of the statistical evidence had failed to consider. In this action to collect on an insurance policy, Mapes Casino sought to quantify the amount of its loss due to employee defalcation. The casino argued that certain employees were using an intermediary to cash in chips at other casinos. It established that over an eighteen-month period the win percentage at its craps tables was 6%, compared with an expected value of 20%. The court recognized that the statistics were probative of the fact that something was wrong at the craps tables—the discrepancy was too large to explain as the mere product of random chance. However, the court was not convinced by the plaintiff's alternative hypothesis. The court pointed to other possible explanations (such Runyonesque activities as skimming, scamming, and crossroading) that might have accounted for the discrepancy without implicating the suspect employees.[169] In short, rejection of the null hypothesis does not leave the proffered alternative hypothesis as the only viable explanation for the data.[170]

[165] We have used two-sided intervals corresponding to two-tailed tests. One-sided intervals corresponding to one-tailed tests also are available.

[166] Often, the null and alternative hypotheses are statements about possible ranges of values for parameters in a common statistical model. Computations of standard errors, *p*-values, and power all take place within the confines of this basic model. The statistical analysis looks at the relative plausibility for competing values of the parameters but makes no global assessment of the reasonableness of the basic model. Inquiry by the court may be advisable.

[167] *See, e.g*., Paul Meier & Sandy Zabell, *Benjamin Peirce and the Howland Will*, 75 J. Am. Stat. Ass'n 497 (1980) (competing explanations in a forgery case). Outside the legal realm there are many intriguing examples of the tendency to think that a small *p*-value is definitive proof of an alternative hypothesis, even though there are other plausible explanations for the data. *See, e.g.*, Freedman et al., *supra* note 12, at 503–04; C.E.M. Hansel, ESP: A Scientific Evaluation (1966).

[168] 290 F. Supp. 186 (D. Nev. 1968).

[169] *Id.* at 193. Skimming consists of taking off the top before counting the drop; scamming is cheating by collusion between dealer and player; and crossroading involves professional cheaters among the players. *Id*. In plainer language, the court seems to have ruled that the casino itself might be cheating, or there could have been cheaters other than the particular employees identified in the case. At the least, plaintiff's statistical evidence did not rule such possibilities out of bounds.

[170] *Compare* EEOC v. Sears, Roebuck & Co., 839 F.2d 302, 312 & n.9, 313 (7th Cir. 1988) (EEOC's regression studies showing significant differences did not establish liability because surveys and testimony supported the rival hypothesis that women generally had less interest in commission sales positions) *with* EEOC v. General Tel. Co. of N.W., Inc., 885 F.2d 575 (9th Cir. 1989) (unsubstantiated rival hypothesis of lack of interest in nontraditional jobs insufficient to rebut prima facie case of gender discrimination), *cert. denied*, 498 U.S. 950 (1990); *cf. supra* § II.C (problem of confounding).

C. Posterior Probabilities

Standard errors, *p*-values, and significance tests are often used to assess random error. These assessments rely on the sample data and are justified in terms of the operating characteristics of the statistical procedures.[171] However, this frequentist approach does not permit the statistician to compute the probability that a particular hypothesis is correct, given the data.[172]

In the Bayesian approach, probabilities represent subjective degrees of belief rather than objective facts. This approach allows the calculation of posterior probabilities for various hypotheses given the data.[173] However, such probabilities must be "personal," for they reflect not just the data, but also the statistician's, or perhaps the fact finder's,[174] subjective *prior probabilities*—that is, degree of belief about the hypotheses, prior to obtaining the data.[175]

[171] Operating characteristics are the expected value and standard error of estimators, probabilities of error for statistical tests, and so forth.

[172] *See supra* § IV.B.1. Consequently, quantities such as *p*-values or confidence levels cannot be compared directly with numbers like .95 or .50 that might be thought to quantify the burden of persuasion in civil or criminal cases. *See* David H. Kaye, *Hypothesis Testing in the Courtroom*, *in* Contributions to the Theory and Application of Statistics 331 (Alan E. Gelfand ed., 1987); David H. Kaye, *Apples and Oranges*: *Confidence Coefficients and the Burden of Persuasion*, 73 Cornell L. Rev. 54 (1987).

[173] *See, e.g.,* Joseph B. Kadane, *A Statistical Analysis of Adverse Impact of Employer Decisions,* 85 J. Am. Stat. Ass'n 925 (1990) (analysis of data in an age discrimination case); David H. Kaye, *Statistical Evidence of Discrimination*, 77 J. Am. Stat. Ass'n 773, 780 (1982) (Bayesian analysis of the data in Swain v. Alabama, 380 U.S. 202 (1965)); Kaye, *supra* note 156, at 848–52 (analysis of data from Hazelwood Sch. Dist. v. United States, 433 U.S. 299 (1977)).

[174] *E.g.*, Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence,* 83 Harv. L. Rev. 489 (1970); Kadane, *supra* note 173. *But see* Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 Harv. L. Rev. 1329 (1971) (arguing that efforts to describe the impact of evidence on a juror's subjective probabilities would unduly impress jurors and undermine the presumption of innocence and other legal values).

[175] *See generally* David H. Kaye, *Introduction: What is Bayesianism*?, *in* Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism 1 (Peter Tillers & Eric D. Green eds., 1988); Brian Skyrms, Choice and Chance: An Introduction to Inductive Logic (3d ed. 1986).

To date, such analyses rarely have been used in court,[176] and the question of their forensic value has been aired primarily in academic literature.[177] Some statisticians favor Bayesian methods,[178] and some legal commentators have proposed their use in certain kinds of cases.[179]

[176] *See* The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 193. The one area where Bayesian techniques are often used is parentage testing in civil cases. *Compare* State v. Spann, 617 A.2d 247, 257 (N.J. 1993) *with* Plemel v. Walter, 735 P.2d 1209, 1215 (Or. 1987).

[177] *See, e.g.,* Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism, *supra* note 175; Symposium, *Decision and Inference in Litigation*, 13 Cardozo L. Rev. 253 (1991).

[178] Donald A. Berry, *Inferences Using DNA Profiling in Forensic Identification and Paternity Cases,* 6 Stat. Sci. 175, 180 (1991); Stephen E. Fienberg & Joseph B. Kadane, *The Presentation of Bayesian Statistical Analyses in Legal Proceedings*, 32 Statistician 88 (1983); Stephen E. Fienberg & Mark J. Schervish, *The Relevance of Bayesian Inference for the Presentation of Statistical Evidence and for Legal Decisionmaking*, 66 B.U. L. Rev. 771 (1986); Kadane, *supra* note 173; Kathryn Roeder, *DNA Fingerprinting: A Review of the Controversy,* 9 Stat. Sci. 222 (1994); *cf.* I. W. Evett et al., *An Illustration of the Advantages of Efficient Statistical Methods for RFLP Analysis in Forensic Science*, 52 Am. J. Hum. Genet. 498, 499 (1993) (favoring presentation of the likelihood ratio for expressing the weight of DNA evidence). Nevertheless, many statisticians question the general applicability of Bayesian techniques: The results of the analysis may be substantially influenced by the prior probabilities, which in turn may be quite arbitrary.

[179] *E.g.*, Ira Mark Ellman & David Kaye, *Probabilities and Proof: Can HLA and Blood Group Testing Prove Paternity?,* 54 N.Y.U. L. Rev. 1131 (1979); David H. Kaye, *DNA Evidence: Probability, Population Genetics, and the Courts*, 7 Harv. J.L. & Tech. 101 (1993); Joseph C. Bright et al., *Statistical Sampling in Tax Audits*, 13 Law & Soc. Inquiry 305 (1988); authorities cited *supra* note 174.

Bayesian procedures are sometimes defended on the ground that the beliefs of any rational observer must conform to the Bayesian rules. However, the definition of “rational” is purely formal. *See* Peter C. Fishburn, *The Axioms of Subjective Probability*, 1 Stat. Sci. 335 (1986); David Kaye, *The Laws of Probability and the Law of the Land,* 47 U. Chi. L. Rev. 34 (1979).

This page left blank intentionally for proper pagination when printing two-sided

Appendix: Technical Details on the Standard Error, the Normal Curve, and the *P*-Value

This appendix describes several calculations for our pass rate example. The population consisted of all 5,000 men and 5,000 women in the applicant pool. By way of illustration, suppose that the pass rates for these men and women were 60% and 35%, respectively; so the population difference is $60 - 35 = 25$ percentage points. We chose 50 men and 50 women at random from the population. In our sample, the pass rate for the men was 58%, and the pass rate for the women was 38%; thus, the sample difference was $58 - 38 = 20$ percentage points. Another sample might have pass rates of 62% and 36%, for a sample difference of $62 - 36 = 26$ percentage points. And so forth.

In principle, we can consider the set of all possible samples from the population and make a list of the corresponding differences. This is a long list. Indeed, the number of distinct samples of 50 men and 50 women that can be formed is immense—nearly 5×10^{240}, or 5 followed by 240 zeros. Our sample difference was chosen at random from this list. Statistical theory enables us to make some precise statements about the list and hence about the chances in the sampling procedure.

- The average of the list—that is, the average of the differences over the 5×10^{240} possible samples—equals the difference between the pass rates of all 5,000 men and 5,000 women. In more technical language, the expected value of the sample difference equals the population difference. Even more tersely, the sample difference is an unbiased estimator of the population difference.
- The standard deviation (SD) of the list—that is, the standard deviation of the differences over the 5×10^{240} possible samples—is equal to:[180]

[180] *See, e.g.*, Freedman et al., *supra* note 12, at 337; Moore & McCabe, *supra* note 57, at 590–91. The standard error for the sample difference equals the standard deviation of the list of all possible sample differences, making the connection between standard error and standard deviation. If we drew two samples at random, the difference between them would be on the order of $\sqrt{2} \approx 1.4$ times this standard deviation. The standard error can therefore be used to measure reproducibility of sample data. *See supra* notes 125–26. On the standard deviation, see *supra* § III.E; *see also* Freedman et al., *supra* note 12, at 67.

$$\sqrt{\frac{5{,}000 - 50}{5{,}000 - 1}} \times \sqrt{\frac{P_{\text{men}}\left(1 - P_{\text{men}}\right)}{50} + \frac{P_{\text{women}}\left(1 - P_{\text{women}}\right)}{50}} \qquad (1)$$

In equation (1), P_{men} stands for the proportion of the 5,000 male applicants who would pass the exam, and P_{women} stands for the corresponding proportion of women. When $P_{\text{men}} = 60\%$ and $P_{\text{women}} = 35\%$, the standard deviation of the sample differences would be 9.6 percentage points:

$$\sqrt{\frac{5{,}000 - 50}{5{,}000 - 1}} \times \sqrt{\frac{.60\left(1 - .60\right)}{50} + \frac{.35\left(1 - .35\right)}{50}} = .096 \qquad (2)$$

Figure 12

The distribution of the sample difference in pass rates when $P_{\text{men}} = 60\%$ and $P_{\text{women}} = 35\%$.

insert figure 12 here

Figure 12 shows the histogram for the sample differences.[181] The graph is drawn so the area between two values gives the relative frequency of sample dif-

[181] The *probability histogram* in Figure 12 shows the distribution of the sample differences, indicating the relative likelihood of the various ranges of possible values; likelihood is represented by area. The lower horizontal scale shows standard units, that is, deviations from the expected value relative to the standard error. In our example, the expected value is 25 percentage points and the standard error is 9.6 percentage points. Thus, 35 percentage points would be expressed as (35 – 25)/9.6 = 1.04 standard units. The vertical scale shows probability per standard unit. *See* Freedman et al., *supra* note 12, at 75, 289.

ferences falling in that range, among all 5×10^{240} possible samples. For instance, take the range from 20 to 30 percentage points. About half the area under the histogram falls into this range. Therefore, given our assumptions, there is about a 50% chance that a sample chosen at random will have a male-female pass rate difference between 20 and 30 percentage points. The “central limit theorem” establishes that the histogram for the sample differences follows the normal curve, at least to a good approximation. Figure 12 shows this curve for comparison. The main point is that chances for the sample difference can be approximated by areas under the normal curve.

Generally, we do not know the pass rates P_{men} and P_{women} in the population. We chose 60% and 35% just by way of illustration. Statisticians would use the pass rates in the sample—58% and 38%—to estimate the pass rates in the population. Substituting the sample pass rates in equation (1) yields:

$$\sqrt{\frac{5{,}000-50}{5{,}000-1}} \times \sqrt{\frac{.58\left(1-.58\right)}{50}+\frac{.38\left(1-.38\right)}{50}} = .097 \tag{3}$$

That is about 10 percentage points—the standard error reported in section IV.A.2.[182]

To sum up, the histogram for the sample differences follows the normal curve, centered at the population difference. The spread is given by the standard error. That is why confidence levels can be based on the standard error, with confidence levels read off the normal curve: 68% of the area under the curve is between –1 and 1, 95% is between –2 and 2, and 99.7% is between –3 and 3, approximately.

We turn to *p*-values.[183] Consider the null hypothesis that the men and women in the population have the same overall pass rates. In that case, the sample differences are centered at 0, because $P_{\text{men}} - P_{\text{women}} = 0$. Since the overall pass rate in the sample is 48%, we use this value to estimate both P_{men} and P_{women} in equation (1):

$$\sqrt{\frac{5{,}000-50}{5{,}000-1}} \times \sqrt{\frac{.48\left(1-.48\right)}{50}+\frac{.48\left(1-.48\right)}{50}} = .099 \tag{4}$$

Again, the standard error (SE) is about 10 percentage points. The observed difference of 20 percentage points is 20/10 = 2.0 SEs. As shown in Figure 13, differences of that magnitude or larger have about a 5% chance of occurring:

[182] There is little difference between equations (2) and (3)—the standard error does not depend strongly on the pass rates.

[183] *See supra* § IV.B.1.

About 5% of the area under the normal curve lies beyond ±2. (In Figure 13, this tail area is shaded.) The *p*-value is about 5%.[184]

Figure 13

P-value for observed difference of 20 percentage points, computed using the null hypothesis. The chance of getting a sample difference of 20 points in magnitude (or more) is about equal to the area under the normal curve beyond ±2. That shaded area is about 5%.

insert figure 13 here

Finally, we calculate power.[185] We are making a two-tailed test at the .05 level. Instead of the null hypothesis, we assume an alternative: In the applicant pool, 55% of the men would pass, and 45% of the women. So there is a difference of 10 percentage points between the pass rates. The distribution of sample differences would now be centered at 10 percentage points (see Figure 14). Again, the sample differences follow the normal curve. The true SE is about 10 percentage points by equation (1), and the SE estimated from the sample will be about the same. On that basis, only sample differences larger than 20 percentage

[184] Technically, the *p*-value is the chance of getting data as extreme as, or more extreme than, the data at hand. *See supra* § IV.B.1 . That is the chance of getting a difference of 20 percentage points or more on the right, together with the chance of getting -20 or less on the left. This chance equals the area under the histogram to the right of 19, together with the area to the left of -19. (The rectangle whose area represents the chance of getting a difference of 20 is included, and likewise for the rectangle above -20.) The area under the histogram in turn may be approximated by the area under the normal curve beyond ±1.9, which is 5.7%. *See, e.g.*, Freedman et al., *supra* note 12, at 291. Keeping track of the edges of the rectangles is called the "continuity correction." As a technical matter, the histogram is computed assuming pass rates of 48% for the men and the women. Other values could be dealt with in a similar way. *See infra* note 187.

[185] *See supra* note 152.

points or smaller than –20 points will be declared significant.[186] About 1/6 of the area under the normal curve in Figure 14 lies in this region.[187] Therefore, the power of the test against the specified alternative is only about 1/6. In the figure, it is the shaded area that corresponds to power.

Figures 12, 13, and 14 have the same shape: The central limit theorem is at work. However, the histograms are centered differently, because the values of P_{men} and P_{women} are different in all three figures. Figure 12 is centered at 25 percentage points, reflecting our illustrative values of 60% and 35% for the pass rates. Figure 13 is centered at 0, because it is drawn according to the requirements of the null hypothesis. Figure 14 is centered at 10 percentage points, because the alternative hypothesis is used to determine the center, rather than the null hypothesis.

Figure 14

Power when P_{men} = 55% and P_{women} = 45%. The chance of getting a significant difference (at the 5% level, two-tailed) is about equal to the area under the normal curve, to the right of +1 or to the left of –2. That shaded area is about 1/6. Power is about 1/6, or 17%.

[186] The null hypothesis asserts a difference of 0: In Figure 13, 20 percentage points is 2 SEs to the right of the value expected under the null hypothesis; likewise, -20 is 2 SEs to the left. However, Figure 14 takes the alternative hypothesis to be true; on that basis, the expected value is 10 instead of 0, so 20 is 1 SE to the right of the expected value, while -20 is 3 SEs to the left.

[187] Let t = sample difference/SE, where the SE is estimated from the data, as in equation (4). One formal version of our test rejects the null hypothesis if $|t| \geq 2$. To find the power, we replace the estimated SE by the true SE, computed as in equation (1), and we replace the probability histogram by the normal curve. These approximations are quite good. The size can be approximated in a similar way, given a common value for the two population pass rates. Of course, more exact calculations are possible. *See supra* note 184.

This page left blank intentionally for proper pagination when printing two-sided

Glossary of Terms

The following terms and definitions are adapted from a variety of sources, including: Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers (1990), and David Freedman et al., Statistics (2d ed. 1991).

Alpha (α). Also, size. A symbol often used to denote the probability of a Type I error. See Type I Error.

Alternative Hypothesis. A statistical hypothesis that is contrasted with the null hypothesis in a significance test. See Statistical Hypothesis; Significance Test.

Area Sample. An area sample is a probability sample in which the sampling frame is a list of geographical areas (i.e., one makes a list of areas, chooses some at random, and interviews people in the selected areas). This is a cost-effective way to draw a sample of people. See Probability Sample; Sampling Frame.

Bayes' Rule. An investigator may start with a subjective probability (the "prior") that expresses degrees of belief about a parameter or a hypothesis. Then data are collected according to some statistical model. Bayes' rule gives a procedure for combining the prior with the data to compute the "posterior" probability, which expresses the investigator's beliefs about the parameter or hypothesis given the data.

Beta (β). A symbol used sometimes to denote power and sometimes to denote the probability of a type II error. See Type II Error; Power.

Bias. A systematic tendency for an estimate to be too high or too low. An estimate is unbiased if the bias is 0. See Nonsampling Error.

Bin. A class interval in a histogram. See Class Interval; Histogram.

Binary Variable. A variable that has only two possible values (e.g., the gender of an employee).

Binomial Distribution. A distribution for the number of occurrences in repeated, independent trials where the probabilities are fixed. For example, the number of heads out of 100 tosses of a coin follows a binomial distribution. (The probability of heads is 1/2 on each toss.) When the probability is not too close to 0 or 1 and the number of trials is large, the binomial distribution has about the same shape as the normal distribution. See Normal Distribution; Poisson Distribution.

Bootstrapping. A procedure for estimating sampling error by generating a simulated population from the sample, then repeatedly drawing samples from this population.

Categorical Data; Categorical Variable. See Qualitative Variable.

Central Limit Theorem. Shows that under suitable conditions, the probability histogram for a sum (or average, or rate) will follow the normal curve.

Chance Error. See Random Error; Sampling Error.

Chi-Squared (χ^2). A statistic that measures the distance between the data and expected values computed from a statistical model. If χ^2 is too large to explain by chance, the data contradict the model. The definition of large depends on the context. See Statistical Hypothesis; Significance Test.

Class Interval. Also, bin. The base of a rectangle in a histogram; the area of the rectangle shows the percentage of observations in the class interval. See Histogram.

Cluster Sample. A type of random sample. For example, a statistician might take households at random, then interview all the people in the selected households. This is a cluster sample of people: A cluster consists of all the people in a selected household. Generally, clustering reduces the cost of interviewing.

Coefficient of Determination. A statistic (more commonly known as R^2) that describes how well a regression equation fits the data. See R-Squared.

Coefficient of Variation. A statistic that measures spread relative to the center of the distribution: SD/average, or SE/expected value.

Conditional Probability. The probability that one event will occur given that another has occurred.

Confidence Coefficient. See Confidence Interval.

Confidence Interval. An estimate, expressed as a range, for a quantity in a population. If an estimate from a large sample is unbiased, a 95% confidence interval is the range from two standard errors below to two standard errors above the estimate. Intervals obtained this way cover the true value about 95% of the time, and 95% is the confidence level, or the confidence coefficient. See Unbiased Estimator; Standard Error.

Confidence Level. See Confidence Interval.

Confounding. See Confounding Variable; Observational Study.

Confounding Variable; Confounder. A variable that is correlated with the independent variables and the dependent variable. When confounding is suspected, an association between the dependent and independent variable may not be causal. See Controlled Experiment; Observational Study.

Consistency; Consistent. See Consistent Estimator.

Consistent Estimator. An estimator that tends to become more and more accurate as the sample size grows. (Inconsistent estimators, which do not become more accurate as the sample size grows, are generally not used by statisticians.)

Content Validity. The extent to which a skills test is appropriate to its intended purpose, as evidenced by a set of questions that adequately reflect the domain being tested.

Continuous Variable. A variable that has arbitrarily fine gradations, such as a person's height.

Control Group. See Controlled Experiment.

Control for. Statisticians "control for" the effects of confounding variables in nonexperimental data by making comparisons for smaller and more homogeneous groups of subjects or by using regression models. See Regression Model.

Controlled Experiment. An experiment where the investigators determine which subjects are put into the treatment group and which are put into the control group. Subjects in the treatment group are exposed by the investigators to some influence—the treatment; those in the control group are not so exposed. For instance, in an experiment to evaluate a new drug, subjects in the treatment group are given the drug, while subjects in the control group are given some other therapy. The outcomes in the two groups are compared to see whether the new drug works. Randomization—that is, randomly assigning subjects to each group—is usually the best way to assure that any observed difference between the two groups comes from the treatment rather than preexisting differences. Of course, in many situations, a randomized controlled experiment is impractical, and investigators must then rely on observational studies.

Convenience Sample. Also, grab sample. A nonrandom sample of units; for instance, for a "mall sample," the interviewer picks respondents from the crowd in a shopping mall.

Correlation Coefficient. A number between –1 and 1 that indicates the extent of the linear association between two variables. Often, the correlation coefficient is abbreviated as *r* .

Covariance. A quantity that describes the statistical interrelationship of two variables.

Covariate. A variable that is related to other variables of primary interest in a study.

Criterion. The variable against which a skills test or other selection procedure is validated. See Predictive Validity.

Data. Observations or measurements, usually of units in a sample taken from a larger population.

Dependent Variable. See Independent Variable; Regression Model.

Descriptive Statistic. A statistic, such as the mean or the standard deviation, used to summarize data.

Differential Validity. Differences in the relationship between skills test scores and outcome measures across different subgroups of test takers.

Discrete Variable. A variable that has only a finite number of possible values, such as the number of automobiles owned by a household.

Random Disturbance Term. See Error Term.

Double -Blind Experiment . An experiment with human subjects in which neither the diagnosticians nor the subjects know who is in the treatment group or the control group. This is accomplished by giving a placebo treatment to subjects in the control group.

Dummy Variable. Generally, a dummy variable takes only the values 0 or 1 and distinguishes one group of interest from another. For example, in a regression study of salary differences between men and women in a firm, the analyst may include a dummy variable for gender and statistical controls such as education and experience to adjust for productivity differences between men and women. The dummy variable would be defined as 1 for the men, 0 for the women. See Regression Model.

Econometrics. The statistical study of economic issues.

Epidemiology. Statistical study of disease or injury in human populations.

Error Term. The part of a statistical model that describes random error (i.e., the impact of chance factors unrelated to variables in the model). In econometric models, the error term is called a random disturbance term.

Estimator. A sample statistic used to estimate a population parameter. For instance, the sample mean commonly is used to estimate the population mean. The term “estimator” connotes a statistical procedure, while an “estimate” connotes a particular numerical result.

Expected Value. See Random Variable.

Fisher's Exact Test. When comparing two sample proportions (e.g., the proportions of whites and blacks getting a promotion), an investigator may want to test the null hypothesis that promotion does not depend on race. Fisher's exact test is one way to arrive at a *p*-value. The calculation is based on the hypergeometric distribution. See Hypergeometric Distribution; Statistical Hypothesis; Significance Test; *p*-Value.

Fixed Significance Level. Also, alpha, size. A preset level, such as 0.05 or 0.01. If the *p*-value of a test falls below this level, the result is deemed statistically significant. See Significance Test.

Frequency Distribution. Shows how often specified values occur in a data set.

Gaussian Distribution. See Normal Distribution.

General Linear Model. Expresses the dependent variable as a linear combination of the independent variables plus an error term whose components may be dependent and have differing variances. See Error Term; Linear Combination; Regression Model; Variance.

Grab Sample. See Convenience Sample.

Heteroscedastic. See Scatter Diagram.

Histogram. A plot showing how observed values fall within specified intervals, called bins or class intervals. Generally, matters are arranged so the area under the histogram, but over a class interval, gives the frequency or relative frequency of data in that interval. In a probability histogram, the area gives the chance of observing a value that falls in the corresponding interval.

Homoscedastic. See Scatter Diagram.

Hypergeometric Distribution. Suppose a sample is drawn at random without replacement from a finite population. The number of times that items of a certain type come into the sample is given by the hypergeometric distribution.

Hypothesis Test. See Significance Test.

Independence. Events are independent when the probability of one is unaffected by the occurrence or nonoccurrence of the other.

Independent Variable. The independent variable is used in a regression model to predict values of the dependent variable. For instance, the unemployment rate has been used as the independent variable in a model for predicting the crime rate; the latter is the dependent variable in this application. See Regression Model.

Indicator Variable. See Dummy Variable.

Interval Estimate. A confidence interval; or, a point estimate coupled with a standard error. See Confidence Interval; Standard Error.

Least Squares. See Least-Squares Estimator; Regression Model.

Least-Squares Estimator. An estimator that is computed by minimizing the sum of the squared residuals. See Residual.

Linear Combination. To obtain a linear combination of two variables, the first variable is multiplied by some constant, the second variable is multiplied by another constant, and the two products are added (e.g., $2u + 3v$ is a linear combination of u and v).

Loss Function. Statisticians may evaluate estimators according to a mathematical formula involving the errors (i.e., differences between actual values and estimated values). The loss may be the total of the squared errors or the total of the absolute errors, etc. Loss functions seldom quantify real losses but may be useful summary statistics and may prompt the construction of useful statistical procedures.

Mean. The mean is one way to find the center of a batch of numbers: Add up the numbers, and divide by how many there are. Weights may be employed, too, as in weighted mean or weighted average. Also, the expected value of a random variable; average. See Random Variable.

Median. The median is another way to find the center of a batch of numbers. The median is the fiftieth percentile. Half the numbers are larger, and half are smaller. (To be very precise, at least half the numbers are greater than or equal to the median; at least half the numbers are less than or equal to the median; for small data sets, the median may not be uniquely defined.)

Meta-Analysis. Attempts to combine information from all studies in a certain collection.

Mode. The most commonly observed value.

Multicollinearity. Also, collinearity. The existence of correlations among the independent variables in a regression model. See Independent Variable; Regression Model.

Multiple Comparison. An examination of more than one test statistic relating to the same data set. Multiple comparisons complicate the interpretation of a *p*-value. For example, if twenty divisions of a company are examined for disparities, and one division is found to have a disparity significant at the 0.05 level, the result is not surprising; indeed, it should be expected under the null hypothesis.

Multiple Correlation Coefficient. A number that indicates the extent to which one variable can be predicted as a linear combination of other variables. Its magnitude is the square root of R^2. See Linear Combination; R-*S*quared; Regression Model.

Multiple Regression. A regression equation that includes two or more independent variables. See Regression Model.

Multistage Cluster Sample. A probability sample drawn in stages, usually after stratification; the last stage will involve drawing a cluster. See Cluster Sample; Probability Sample; Stratified Random Sample.

Natural Experiment. An observational study in which treatment and control groups have been formed by some natural development; however, the assignment of subjects to groups is judged akin to randomization. See Observational Study.

Nonsampling Error. A catch-all term for sources of error in a survey, other than sampling error. Nonsampling errors cause bias. One example is selection bias: The sample is drawn in a way that tends to exclude certain subgroups in the population. A second example is nonresponse bias: People who do not respond to a survey are usually different from respondents. A final example: Response bias arises, for instance, if the interviewer uses a loaded question.

Normal Distribution. Also, Gaussian distribution. The density for this distribution is the famous bell-shaped curve. Statistical terminology notwithstanding, there is nothing wrong with a distribution that differs from the normal.

Null Hypothesis. A hypothesis that there is no difference between two groups from which samples are drawn. See Statistical Hypothesis.

Observational Study. A study in which subjects select themselves into groups; investigators then compare the outcomes for the different groups. For example, studies of smoking are generally observational. Subjects decide whether or not to smoke; the investigators compare the death rate for smokers with the death rate for nonsmokers. In an observational study, the groups may differ in important ways that the investigators do not notice; controlled experiments minimize this problem. The critical distinction is that in a controlled experiment, the investigators intervene to manipulate the circumstances of the subjects; in an observational study, the investigators are passive observers. (Of course, running a good observational study is hard work and may be quite useful.)

Observed Significance Level. See p-Value.

Odds. The probability that an event will occur divided by the probability that it will not. For example, if the chance of rain tomorrow is 2/3, then the odds on rain are (2/3)/(1/3) = 2/1, or 2 to 1.

Odds Ratio. A measure of association, often used in epidemiology. For instance, if 10% of all people exposed to a chemical develop a disease, compared with 5% of people who are not exposed, the odds of the disease in the exposed group are 10/90 = 1/9, compared with 5/95 = 1/19 in the unexposed group. The odds ratio is 19/9 = 2.1. An odds ratio of 1 indicates no association.

One-Sided Hypothesis. Excludes the possibility that a parameter could be, for example, less than the value asserted in the null hypothesis. A one-sided hypothesis leads to a one-tailed test. See Statistical Hypothesis; Significance Test.

One-Tailed Test. See Significance Test.

Outlier. An observation that is far removed from the bulk of the data. Outliers may indicate a faulty measurement; they may exert undue influence on a summary statistic, such as the mean or the correlation coefficient.

p-Value. The output of a statistical test. The probability of getting, just by chance, a test statistic as large as or larger than the observed value. Large *p*-values are consistent with the null hypothesis; small *p*-values undermine this hypothesis. However, *p* itself does not give the probability that the null hypothesis is true. If *p* is smaller than 5%, the result is said to be statistically significant. If *p* is smaller than 1%, the result is highly significant. The *p*-value is also called the observed significance level. See Statistical Hypothesis; Significance Test.

Parameter. A numerical characteristic of a population or of a model. See Probability Model.

Percentile. To get the 90th percentile, for instance, of a data set, the data are arrayed from the smallest value to the largest. Then 90% of the values fall below the 90th percentile, and 10% fall above. (To be very precise, at least 90% of the data are at the 90th percentile or below; at least 10% of the data are at the 90th percentile or above.) The 50th percentile is the median. When the LSAT first was scored on a 10–50 scale in 1982, a score of 32 placed a test taker at the 50th percentile; a score of 40 was at the 90th percentile (approximately).

Point Estimate. An estimate of the value of a quantity expressed as a single number.

Poisson Distribution. The Poisson distribution is a limiting case of the binomial distribution, when the number of trials is large and the common probability is small. The parameter of the approximating Poisson distribution is the number of trials times the common probability, which gives the "expected" number of events. When this number is large, the Poisson distribution may be approximated by a normal distribution.

Population. Also, universe. All the units of interest to the researcher.

Posterior Probability. See Bayes' Rule.

Power. The probability that a statistical test will reject the null hypothesis. To compute power, the analyst has to fix the size of the test and specify parameter values outside the range given in the null hypothesis. A powerful test has a good chance of detecting an effect, when there is an effect to be detected. See Significance Test; Beta.

Practical Significance. Substantive importance. Statistical significance does not necessarily establish practical significance. Small differences can be statistically significant in large samples.

Predictive Validity. A psychological or skills test has predictive validity to the extent that test scores are well correlated with later performance or, more generally, with outcomes that the test is intended to predict.

Prior Probability. See Bayes' Rule.

Probability. Chance, on a scale from 0 to 1. Impossibility is represented by 0, certainty by 1. Equivalently, chances may be quoted in percentages; 100% corresponds to 1; 5% to .05; and so forth.

Probability Density. Describes the probability distribution for a random variable. The chance that the random variable falls in an interval equals the area below the density and above the interval. See Probability Distribution; Variable.

Probability Distribution. Gives probabilities for possible values of a random variable. Often, the distribution is described in terms of the density. See Probability Density.

Probability Histogram. See Histogram.

Probability Model. Relates probabilities of outcomes to parameters; also, Statistical Model. The latter connotes unknown parameters.

Probability Sample. A sample drawn from a sampling frame by some objective chance mechanism; each unit has a known probability of being sampled. Such samples are expensive to draw but minimize selection bias.

Psychometrics. The study of psychological measurement and testing.

Qualitative Variable; Quantitative Variable. A qualitative or categorical variable describes qualitative features of subjects in a study (e.g., marital status—never married, married, widowed, divorced, separated). A quantitative variable describes numerical features of the subjects (e.g., height, weight, income). This is not a hard-and-fast distinction, because qualitative features may be given numerical codes, as in a dummy variable. Quantitative variables may be classified as discrete or continuous. Concepts like the mean and the standard deviation apply only to quantitative variables. See Discrete Variable; Continuous Variable.

Quartile. The 25th or 75th percentile. See Percentile.

R-*Squared* (R ^{2})*.* Measures how well a regression equation fits the data. R^2 varies between 0 (no association) and 1 (perfect fit). Generally, R^2 does not measure the validity of underlying assumptions. See Regression Model.

Random Error. Sources of error that are haphazard in their effect. These are re - flected in the error term of a statistical model. Some authors refer to random error as chance error or sampling error. See Regression Model.

Random Variable. A variable whose possible values occur according to some probability mechanism. For example, if you throw a pair of dice, the total number of spots is a random variable. The chance of two spots is 1/36, the chance of three spots is 2/36, and so forth; the most likely number is seven, with a chance of 6/36. The expected value of a random variable is the weighted average of the possible values; the weights are the probabilities. In our example, the expected value is

$$\frac{1}{36}\times 2+\frac{2}{36}\times 3+\frac{3}{36}\times 4+\frac{4}{36}\times 5+\frac{5}{36}\times 6+\frac{6}{36}\times 7$$
$$+\frac{5}{36}\times 8+\frac{4}{36}\times 9+\frac{3}{36}\times 10+\frac{2}{36}\times 11+\frac{1}{36}\times 12=7$$

In many problems, the weighted average is computed with respect to the density; then sums must be replaced by integrals. The expected value need not be a possible value for the random variable. Generally, a random variable will be somewhere around its expected value, but it will be off (in either direction) by something like 1 standard error or so. See Standard Error.

Randomization. See Controlled Experiment.

Randomized Controlled Experiment. A controlled experiment in which subjects are placed into the treatment and control groups at random—as if by lot. See Controlled Experiment.

Range. The difference between the biggest and smallest values in a batch of numbers.

Regression Coefficient. A constant in a regression equation. See Regression Model.

Regression Diagnostics. Procedures intended to check whether the assumptions of a regression model are appropriate.

Regression Equation. See Regression Model.

Regression Line. The graph of a regression equation with only one dependent variable and one independent variable.

Regression Model. A regression model attempts to combine the values of certain variables (the independent variables) to obtain expected values for another variable (the dependent variable). A hypothetical example illustrates the idea. An analyst might try to predict salaries of employees in a firm using education, experience—and a dummy variable for gender, taking the value 1 for men and 0 for women. Here, salary is the dependent variable (the variable being predicted), while education, experience, and the dummy are the independent variables (the variables entered into the equation to make the predictions).

Sometimes, "regression model" refers to a statistical model for the data; if no qualifications are made, the model will generally be linear, and errors will be assumed independent, with common variance. At other times, "regression model" refers to an equation estimated from data.

In our example, salary (dollars per year) is predicted from education (years of schooling completed) and experience (years with the company)—along with the dummy variable man, taking the value 1 for male employees and 0 for female employees. The model is

$$\text{salary} = a + b \times \text{education} + c \times \text{experience} + d \times \text{man} + u \tag{1}$$

Equation (1) is a statistical model for the data, with unknown parameters a, b, c, d; these parameters are regression coefficients; a often is called the intercept, and u is an error term, with a component for each employee.

The parameters in equation (1) are estimated from the data using least squares. If the estimated coefficient d for the dummy variable turns out to be positive and statistically significant (by a t-test), that would be taken as evidence of disparate impact: Men earn more than women, even after adjusting for differences in background factors that might affect productivity. Education and experience would be entered into equation (1) as statistical controls, precisely in order to claim that adjustment had been made for differences in background.

Suppose the estimated equation turns out as follows:

$$\begin{aligned}\text{predicted salary} = \$7{,}100 + \$1{,}300 \times \text{education} &+ \\ \$2{,}200 \times \text{experience} + \$700 \times \text{man}&\end{aligned} \tag{2}$$

According to equation (2), every extra year of education is worth on average \$1,300; similarly, every extra year of experience is worth on average \$2,200; and most important, men receive a premium of \$700 over women with the same education and experience, on average.

Some numerical examples will illustrate equation (2). A male employee with 12 years of education (high school) and 10 years of experience would have a predicted salary of

$$\begin{aligned}&\$7{,}100 + \$1{,}300 \times 12 + \$2{,}200 \times 10 + \$700 \times 1 = \\ &\qquad \$7{,}100 + \$15{,}600 + \$22{,}000 + \$700 = \$45{,}400\end{aligned} \tag{3}$$

A similarly situated female employee has a predicted salary of only

$$\begin{aligned}&\$7{,}100 + \$1{,}300 \times 12 + \$2{,}200 \times 10 + \$700 \times 0 = \\ &\qquad \$7{,}100 + \$15{,}600 + \$22{,}000 + \$0 = \$44{,}700\end{aligned} \tag{4}$$

Notice the impact of the dummy variable: $700 is added to equation (3) but not to equation (4).

A male employee with 16 years of education (college) and 6 years of experience would have a predicted salary of

$$\begin{aligned}&\$7{,}100 + \$1{,}300 \times 16 + \$2{,}200 \times 6 + \$700 \times 1 = \\ &\qquad \$7{,}100 + \$20{,}800 + \$13{,}200 + \$700 = \$41{,}800\end{aligned} \tag{5}$$

A similarly situated female employee has a predicted salary of only

$$\begin{aligned}&\$7{,}100 + \$1{,}300 \times 16 + \$2{,}200 \times 6 + \$700 \times 0 = \\ &\qquad \$7{,}100 + \$20{,}800 + \$13{,}200 + \$0 = \$41{,}100\end{aligned} \tag{6}$$

In equation (1), u is an error term, with one component for each employee; these components are random errors. Equation (2) has corresponding residuals. For each employee, there is a difference (or residual) between the salary predicted from the equation and the actual salary:

$$\text{actual} = \text{predicted} + \text{residual} \tag{7}$$

The residuals are approximations to the random errors in equation (1).

A critical step in the argument is stablishing that the coefficient d of the dummy variable in equation (1) is "statistically significant." This step depends on the statistical assumptions built into the model. For instance, each extra year of education is assumed to be worth the same (on average) across all levels of experience, for both men and women; similarly, each extra year of experience is worth the same across all levels of education, for both men and women; furthermore, the premium paid to men does not depend systematically on education or experience. Ability, quality of education, and quality of experience are assumed not to make any systematic difference to the predictions of the model.

Moreover, there are technical assumptions that must be made about the error term u: for instance, that its components—the random errors—are independent from person to person in the data set but have the same variance. Some assumptions of this general nature will be found to underlie typical applications of regression techniques; such assumptions should be identified and their reasonableness assessed.

The term "predicted" in equation (2) has a specialized meaning, since the analyst has available the data being predicted. For that reason, statisticians often refer to "fitted values" rather than to "predicted values." See Random Error; Independence; Least Squares; Regression Model; Multiple Regression; *t*-Test; Dummy Variable; Random Variable; Variance.

Relative Risk. A measure of association used in epidemiology. For instance, if 10% of all people exposed to a chemical develop a disease, compared with 5% of people who are not exposed, the disease occurs twice as frequently among the exposed people: The relative risk is 10%/5% = 2. A relative risk of 1 indicates no association.

Reliability. The extent to which a measuring instrument gives the same results on repeated measurement of the same thing.

Residual. The difference between an actual and a predicted value. The predicted value comes typically from a regression equation and also is called the "fitted value." See Regression Model; Independent Variable.

Risk. Expected loss. "Expected" means on average, over the various data sets that could be generated by the statistical model under examination. Usually, risk cannot be computed exactly but has to be estimated, because the parameters in the statistical model are unknown and must be estimated. See Loss Function; Random Variable.

Robust. A statistic or procedure that does not change much when data or assumptions are slightly modified.

Sample. A set of units collected for study.

Sample Size. The number of units in a sample.

Sampling Distribution. The distribution of the values of a statistic, over all possible samples from a population. For example, suppose a random sample is drawn. Some values of the sample mean are more likely, others are less likely. The sampling distribution specifies the chance that the sample mean will fall in one interval rather than another.

Sampling Error. A sample is part of a population. When a sample is used to estimate a numerical characteristic of the population, the estimate is likely to differ from the population value, because the sample is not a perfect microcosm of the whole. If the estimate is unbiased, the difference between the estimate and the exact value is sampling error. More generally, estimate = true value + bias + sampling error

Sampling error is also called chance error or random error. See Standard Error.

Sampling Frame. A list of units designed to represent the entire population as completely as possible. The sample is drawn from the frame.

Scatter Diagram. Also, scatterplot, scattergram. A graph showing the relationship between two variables in a study; each dot represents one subject. One variable is plotted along the horizontal axis, the other variable is plotted along the vertical axis. A scatter diagram is homoscedastic when the spread is more or less the same inside any vertical strip. If the spread changes from one strip to another, the diagram is heteroscedastic.

Sensitivity. In clinical medicine, the probability that a test for a disease will give a positive result given that the patient has the disease. Sensitivity is analogous to the power of a statistical test.

Sensitivity Analysis. Analyzing data in different ways to see how results depend on methods or assumptions.

Significance Level. See Fixed Significance Level; *p*-Value.

Significance Test. Also, statistical test, hypothesis test, test of significance; statistical hypothesis; *p*-value; *t*-test. A significance test involves formulating a statistical hypothesis and a test statistic, computing a *p*-value, and comparing *p* with some preestablished value to decide if the test statistic is significant. The idea is to see whether the data conform to the predictions of the null hypothesis. Generally, a large test statistic goes with a small *p*-value, and small *p*-values would undermine the null hypothesis.

For instance, suppose that a random sample of male and female employees was given a skills test, and the mean scores of the men and women were different in the sample. To judge whether the difference is due to sampling error, a statistician might consider the implications of competing hypotheses about the difference in the population. The null hypothesis would say that on average, in the population, men and women have the same scores: The difference observed in the data is then just due to sampling error. A one-sided alternative hypothesis would be that on average, in the population, men score higher than women. A one-tailed test would reject the null hypothesis if the sample of men score substantially higher than the women—so much so that the difference is hard to explain on the basis of sampling error.

In contrast, the null could be tested against the two-sided alternative hypothesis that on average, in the population, men score differently than women—higher or lower. The corresponding two-tailed test would reject the null hypothesis if the sample of men score substantially higher—or substantially lower—than the women.

The one-tailed and two-tailed tests would both be based on the same data and use the same *t*-statistic. However, if the men in the sample score higher than the women, the one-tailed test would give a *p*-value only half as large as the two-tailed test (i.e., the one-tailed test would appear to give stronger evidence against the null hypothesis).

Significant. See *p*-Value; Practical Significance; Significance Test.

Simple Random Sample. A random sample in which each unit in the sampling frame has the same chance of being sampled. For example, the statistician takes a unit at random (as if by lottery), sets it aside, takes another at random from what is left, and so forth.

Size . The size of a statistical test is a synonym for alpha (α). See Alpha.

Specificity. In clinical medicine, the probability that a test for a disease will give a negative result given that the patient does not have the disease. Specificity is analogous to $1 - \alpha$, where α is the significance level of a statistical test.

Spurious Correlation. When two variables are correlated, one is not necessarily the cause of the other. The vocabulary and shoe size of children in elementary school, for instance, are correlated—but learning more words does not make their feet grow. Such noncausal correlations are said to be spurious. (Originally, the term seems to have been applied to the correlation between two rates with the same denominator: Even if the numerators are unrelated, the common denominator will create some association.)

Standard Deviation (SD). Indicates how far a typical element deviates from the average. For instance, in round numbers, the average height of women aged eighteen and over in the United States is 5 feet 4 inches, and the SD is 3 inches. Typical woman are about 5 feet 4 inches in height; they are off this something like 3 inches.

For distributions that follow the normal curve, about 95% of the elements are in the range "mean –2 SD" to "mean +2 SD". Deviations from the average that exceed 3 or 4 SDs are extremely unusual. Many authors use standard deviation also to mean standard error.

Standard Error (SE). Indicates the likely size of the sampling error in an estimate. Many authors use the term "standard deviation" instead of standard error.

Standard Error of Regression. Indicates how actual values differ (in some average sense) from the fitted values in a regression model. See Regression Model.

Standardization. See Standardized Variable.

Standardized Variable. Transformed to have a mean of 0 and a variance of 1. This involves two steps: (1) subtract the mean, and (2) divide by the standard deviation.

Statistic. A number that characterizes or summarizes data. A statistic refers to a sample; a parameter or a true value refers to a population or a probability model.

Statistical Control. See Control for.

Statistical Hypothesis. Data may be governed by a probability model; parameters are numerical characteristics describing features of the model. Generally, a statistical hypothesis is a statement about the parameters in a probability model. The null hypothesis may assert that certain parameters have specified values or fall in specified ranges; the alternative hypothesis would specify other values or ranges. The null hypothesis is compared with the data using a test statistic; the null hypothesis may be rejected if there is a statistically significant difference between the data and the predictions of the null hypothesis.

Typically, the investigator seeks to demonstrate the alternative hypothesis; the null hypothesis would explain the findings as a result of mere chance, and the investigator uses a significance test to rule out this explanation. See Significance Test.

Statistical Model. See Probability Model.

Statistical Significance. See *p*-Value; Significance Test.

Statistically Significant. See *p*-Value.

Stratified Random Sample. A type of probability sample. The analyst divides the population up into relatively homogeneous groups called strata and draws a random sample separately from each stratum.

Stratum; Strata. See Stratified Random Sample.

t-Statistic. A test statistic used to make the *t*-test. The *t*-statistic tells you how far away an estimate is from its expected value, relative to the standard error. The expected value is computed using the null hypothesis. Some authors refer to the *t*-statistic, others to the *z*-statistic, especially when the sample is large. A *t*-statistic larger than 2 or 3 in absolute value makes the null hypothesis rather unlikely—the estimate is too many standard errors away from its expected value. See Statistical Hypothesis; Significance Test; *t*-Test.

t-Test. A statistical test based on the *t*-statistic. Large *t*-statistics are beyond the usual range of sampling error. For example, if *t* is larger than 2 or smaller than -2, the estimate is statistically significant at the 5% level: Such values of *t* are hard to explain on the basis of sampling error. The scale for *t*-statistics is tied to areas under the normal curve. For instance, a *t*-statistic of 1.5 is not very striking, because 13%, or 13/100, of the area under the normal curve is outside the range from –1.5 to 1.5. Conversely, $t = 3$ is remarkable: Only 3/1,000 of the area lies outside the range from –3 to 3. This discussion is based on having a reasonably large sample; in that context, many authors refer to the *z*-test rather than the *t*-test.

For small samples drawn at random from a population known to be normal, the *t*-statistic follows "Student's *t*-distribution" (when the null hypothesis holds) rather than the normal curve; larger values of *t* are required to achieve significance. See Statistical Hypothesis; Significance Test; *p*-Value.

Test Statistic. A statistic used to judge whether data conform to the null hypothesis. The parameters of a probability model determine expected values for the data; differences between expected values and observed values are measured by a test statistic. Test statistics include the chi-squared statistic (χ^2) and the *t*-statistic. Generally, small values of the test statistic are consistent with the null hypothesis; large values lead to rejection. See Statistical Hypothesis; *p*-Value; *t*-Statistic; Chi-Squared.

Time Series. A series of data collected over time—for instance, the Gross National Product of the United States from 1940 to 1990.

Two -Sided Hypothesis. An alternative hypothesis asserting that the values of a parameter are different from—either greater than or less than—the value asserted in the null hypothesis. See Statistical Hypothesis; Significance Test.

Two-Tailed Test. See Significance Test.

Type I Error. A statistical test makes a type I error when (a) the null hypothesis is in fact true, and (b) the test rejects the null hypothesis (i.e., there is a false alarm). For instance, a study of two groups may show some difference between samples from each group, even when there is no difference in the population. When a statistical test deems the difference to be significant in this situation, it makes a type I error. See Statistical Hypothesis; Significance Test.

Type II Error. A statistical test makes a type II error when (a) the null hypothesis is in fact not true, and (b) the test fails to reject the null hypothesis (i.e., there is a false negative). For instance, there may not be a significant difference between samples from two groups when, in fact, the groups are different. See Statistical Hypothesis; Significance Test.

Unbiased Estimator. An estimator that is correct on average, over the possible data sets. The estimates have no systematic tendency to fall high or low.

Uniform Distribution. For example, if an investigator picks a whole number at random from 1 to 100, it has the uniform distribution: All values are equally likely. Similarly, one gets a uniform distribution by picking a real number at random between 0.75 and 3.25: The chance of landing in an interval is proportional to the length of the interval. The uniform distribution, without further qualification, is presumably on the unit interval (which goes from 0 to 1).

Validity. The extent to which a test instrument measures what it is supposed to, rather than something else. The validity of a standardized test often is indicated, in part, by the correlation coefficient between the test scores and some outcome measure.

Variable. A property of units in a study, which varies from one unit to another (e.g., incomes of households) in a study of households; employment status of individuals (employed, unemployed, not in labor force) in a study of people.

Variance. The square of the standard deviation. See Standard Deviation.

z-Statistic. See *t*-Statistic.

z-Test. See *t*-Test.

References on Statistics

General Surveys

David Freedman et al., Statistics (2d ed. 1991).

Darrell Huff, How to Lie with Statistics (1954).

Gregory A. Kimble, How to Use (and Misuse) Statistics (1978).

David S. Moore, Statistics: Concepts and Controversies (2d ed. 1985).

David S. Moore & George P. McCabe, Introduction to the Practice of Statistics (2d ed. 1993).

Michael Oakes, Statistical Inference: A Commentary for the Social and Behavioral Sciences (1986).

Perspectives on Contemporary Statistics (David C. Hoaglin & David S. Moore eds., 1992).

Statistics: A Guide to the Unknown (Judith M. Tanur et al. eds., 2d ed. 1978).

Hans Zeisel, Say It with Figures (6th ed. 1985).

Reference Works for Lawyers and Judges

David C. Baldus & James W. L. Cole, Statistical Proof of Discrimination (1980).

David W. Barnes & John M. Conley, Statistical Evidence in Litigation: Methodology, Procedure, and Practice (1986).

James Brook, A Lawyer's Guide to Probability and Statistics (1990).

The Evolving Role of Statistical Assessments as Evidence in the Courts (Stephen E. Fienberg ed., 1989).

Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers (1990).

Statistical Methods in Discrimination Litigation (David H. Kaye & Mikel Aickin eds., 1986).

General Reference

International Encyclopedia of Statistics (William H. Kruskal & Judith M. Tanur eds., 1978).