Reference Guide on Statistics

DAVID H. KAYE AND DAVID A. FREEDMAN

David H. Kaye, M.A., J.D., is Regents' Professor, Arizona State University College of Law, and Fellow, Center for the Study of Law, Science, and Technology, Tempe, Arizona

David A. Freedman, Ph.D., is Professor of Statistics, University of California, Berkeley, California

CONTENTS

I. Introduction, 85
- A. Admissibility and Weight of Statistical Studies, 86
- B. Varieties and Limits of Statistical Expertise, 86
- C. Procedures that Enhance Statistical Testimony, 88
  1. Maintaining Professional Autonomy, 88
  2. Disclosing Other Analyses, 88
  3. Disclosing Data and Analytical Methods Before Trial, 89
  4. Presenting Expert Statistical Testimony, 89

II. How Have the Data Been Collected? 90
- A. Is the Study Properly Designed to Investigate Causation? 90
  1. Types of Studies, 90
  2. Randomized Controlled Experiments, 93
  3. Observational Studies, 94
  4. Can the Results Be Generalized? 96
- B. Descriptive Surveys and Censuses, 98
  1. What Method Is Used to Select the Units? 98
  2. Of the Units Selected, Which Are Measured? 101
- C. Individual Measurements, 102
  1. Is the Measurement Process Reliable? 102
  2. Is the Measurement Process Valid? 103
  3. Are the Measurements Recorded Correctly? 104

III. How Have the Data Been Presented? 104
- A. Are Rates or Percentages Properly Interpreted? 105
  1. Have Appropriate Benchmarks Been Provided? 105
  2. Have the Data-Collection Procedures Changed? 105
  3. Are the Categories Appropriate? 106
  4. How Big Is the Base of a Percentage? 107
  5. What Comparisons Are Made? 107
- B. Is an Appropriate Measure of Association Used? 108
- C. Does a Graph Portray Data Fairly? 110
  1. How Are Trends Displayed? 110
  2. How Are Distributions Displayed? 112
- D. Is an Appropriate Measure Used for the Center of a Distribution? 113

E. Is an Appropriate Measure of Variability Used? 114
IV. What Inferences Can Be Drawn from the Data? 115
A. Estimation, 117
1. What Estimator Should Be Used? 117
2. What Is the Standard Error? The Confidence Interval? 117
B. Significance Levels and Hypothesis Tests, 121
1. What Is the *p*-value? 121
2. Is a Difference Statistically Significant? 123
C. Evaluating Hypothesis Tests, 125
1. What Is the Power of the Test? 125
2. One- or Two-tailed Tests? 126
3. How Many Tests Have Been Performed? 127
4. Tests or Interval Estimates? 128
5. What Are the Rival Hypotheses? 129
D. Posterior Probabilities, 131
V. Correlation and Regression, 133
A. Scatter Diagrams, 134
B. Correlation Coefficients, 135
1. Is the Association Linear? 137
2. Do Outliers Influence the Correlation Coefficient? 137
3. Does a Confounding Variable Influence the Coefficient? 138
C. Regression Lines, 139
1. What Are the Slope and Intercept? 140
2. What Is the Unit of Analysis? 141
D. Statistical Models, 143
1. A Social Science Example, 145
2. Standard Errors, *t*-statistics, and Statistical Significance, 148
3. Summary, 148
Appendix, 151
A. Probability and Statistical Inference, 151
B. Technical Details on the Standard Error, the Normal Curve, and Significance Levels, 153
Glossary of Terms, 160
References on Statistics, 178

I. Introduction

Statistics, broadly defined, is the art and science 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.

Statistical assessments are prominent in many kinds of cases, ranging from antitrust to voting rights. Statistical reasoning can be crucial to the interpretation of psychological tests, toxicological and epidemiological studies, disparate treatment of employees, and DNA fingerprinting; this list could easily be extended.[1]

This reference guide describes the elements of statistical thinking. We hope that the explanations provided will permit judges and lawyers who deal with statistical evidence to understand the terminology, place the evidence in context, appreciate its strengths and weaknesses, and apply legal doctrine governing the use of such evidence. The reference guide is organized as follows:

- Section I provides an overview of the field, discusses the admissibility of statistical studies, and offers some suggestions about procedures that 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. The section reviews controlled experiments, observational studies, and surveys, indicating when these designs are likely to give useful data.
- Section III discusses the art of describing and summarizing data. The section considers the mean, median, and standard deviation. These are basic descriptive statistics, and most statistical analyses seen in court use them as building blocks. Section III also discusses trends and associations in data as summarized by graphs, percentages, and tables.
- Section IV describes the logic of statistical inference, emphasizing its foundations and limitations. In particular, this section explains statistical estimation, standard errors, confidence intervals, *p*-values, and hypothesis tests.
- Section V shows how relationships between two variables can be described by means of scatter diagrams, correlation coefficients, and regression lines. Statisticians often use regression techniques in an attempt to infer causation

1. *See generally* Statistics and the Law (Morris H. DeGroot et al. eds., 1986); Panel on Statistical Assessments as Evidence in the Courts, National Research Council, The Evolving Role of Statistical Assessments as Evidence in the Courts (Stephen E. Fienberg ed., 1989) [hereinafter The Evolving Role of Statistical Assessments as Evidence in the Courts]; Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers (1990); 1 & 2 Joseph L. Gastwirth, Statistical Reasoning in Law and Public Policy (1988); Hans Zeisel & David Kaye, Prove It with Figures: Empirical Methods in Law and Litigation (1997).

from association; section V briefly explains the techniques and some of their limitations.

- An appendix presents certain technical details, and the glossary defines many statistical terms that might be encountered in litigation.

A. Admissibility and Weight of Statistical Studies

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, since such studies may be offered to explain the basis for an expert's opinion or may be admissible under the learned treatise exception to the hearsay rule.[2] 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, such methods generally satisfy important aspects of the "scientific knowledge" requirement articulated in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*[3] Of course, a particular study may use a method that is entirely appropriate, but so poorly executed that it should be inadmissible under Federal Rules of Evidence 403 and 702.[4] Or, the method may be inappropriate for the problem at hand and thus lacks the "fit" spoken of in *Daubert*.[5] Or, the study may rest on data of the type not reasonably relied on by statisticians or substantive experts, and hence run afoul of Federal Rule of Evidence 703. Often, however, the battle over statistical evidence concerns weight or sufficiency rather than admissibility.

B. 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, such as error rates; probability theory plays a key role in this endeavor. Results may be used by

2. *See generally* 2 McCormick on Evidence §§ 321, 324.3 (John W. Strong ed., 5th ed. 1999). Studies published by government agencies also may be admissible as public records. *Id.* § 296. *See also* United States v. Esquivel, 88 F.3d 722, 727 (9th Cir. 1996) (taking judicial notice of 1990 census data showing the number of Hispanics eligible for jury service).

3. 509 U.S. 579, 589–90 (1993). For a discussion of the implications and scope of *Daubert* generally, see 1 Modern Scientific Evidence: The Law and Science of Expert Testimony § 1–3.0 (David L. Faigman et al. eds., 1997).

4. *See, e.g.,* Sheehan v. Daily Racing Form, Inc., 104 F.3d 940, 942 (7th Cir. 1997) ("failure to exercise the degree of care that a statistician would use in his scientific work, outside of the context of litigation" renders analysis inadmissible under *Daubert*).

5. 509 U.S. at 591; *cf.* People Who Care v. Rockford Bd. of Educ., 111 F.3d 528, 537–38 (7th Cir. 1997) ("a statistical study that fails to correct for salient explanatory variables, or even to make the most elementary comparisons, has no value as causal explanation and is therefore inadmissible in a federal court"); *Sheehan,* 104 F.3d at 942 (holding that expert's "failure to correct for any potential explanatory variables other than age" made the analyst's finding that "there was a significant correlation between age and retention" inadmissible).

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.

Individuals who specialize in using statistical methods—and whose professional careers demonstrate this orientation—are most likely to apply appropriate procedures and correctly interpret the results. On the other hand, forensic scientists and technicians often testify to probabilities or statistics derived from studies or databases compiled by others, even though some of these testifying experts lack the training or knowledge required to understand and apply the information. *State v. Garrison*[6] 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.'"[7]

At the same time, the choice of which data to examine, or how best to model a particular process, could require subject matter expertise that a statistician might lack. Statisticians often advise experts in substantive fields on the procedures for collecting data and often analyze data collected by others. As a result, cases involving statistical evidence often are (or should be) "two-expert" cases of interlocking testimony.[8] 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.[9]

6. 585 P.2d 563 (Ariz. 1978).

7. *Id.* at 566, 568.

8. 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.

9. In *Vuyanich v. Republic National Bank*, 505 F. Supp. 224, 319 (N.D. Tex. 1980), *vacated*, 723 F.2d 1195 (5th Cir. 1984), defendant's statistical expert criticized the plaintiffs' statistical model for an implicit, but restrictive, assumption about male and female salaries. The district court trying the case 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

C. Procedures that Enhance Statistical Testimony

1. Maintaining Professional Autonomy

Ideally, experts who conduct research in the context of litigation should proceed with the same objectivity that they would apply in other contexts. Thus, experts who testify (or who supply results that are used in testimony by others) should be free to do whatever analysis is required to address in a professionally responsible fashion the issues posed by the litigation.[10] Questions about the freedom of inquiry accorded to testifying experts, as well as the scope and depth of their investigations, may reveal some of the limitations to the analysis being presented.

2. Disclosing Other Analyses

Statisticians analyze data using a variety of statistical models and methods. There is 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 does settle on, however, the testifying expert may explain the history behind the development of the final statistical approach.[11] Indeed, some commentators have urged that counsel who know of other data sets 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.[12]

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.

10. *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).

11. *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).

12. 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.

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. Suggested procedures along these lines are available elsewhere.[13]

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.[14] 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.[15]

13. *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. *See also* David H. Kaye, *Improving Legal Statistics*, 24 L. & Soc'y Rev. 1255 (1990).

14. *See* Fed. R. Evid. 611.

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

II. How Have the Data Been Collected?

An analysis is only as good as the data on which it rests.[16] To a large extent, the design of a study determines the quality of the data. Therefore, the proper interpretation of data and their implications begins with an understanding of study design, and different designs help answer different questions. In many cases, statistics are introduced to show causation. Would additional information in a securities prospectus disclosure have caused potential investors to behave in some other way? Does capital punishment deter crime? Do food additives cause cancer? The design of studies intended to prove causation is the first and perhaps the most important topic of this section.

Another issue is the use of sample data to characterize a population: the population is the whole class of units that are of interest; the sample is a set of units chosen for detailed study. Inferences from the part to the whole are justified only when the sample is representative, and that is the second topic of this section.

Finally, it is important to verify the accuracy of the data collection. Errors can arise in the process of making and recording measurements on individual units. This aspect of data quality is the third topic in this section.

A. Is the Study Properly Designed to Investigate Causation?

1. Types of Studies

When causation is at issue, advocates have relied on three major types of information: anecdotal evidence, observational studies, and controlled experiments.[17] As we shall see, anecdotal reports can provide some information, but they are

16. For introductory treatments of data collection, see, e.g., David Freedman et al., Statistics (3d ed. 1998); Darrell Huff, How to Lie with Statistics (1954); David S. Moore, Statistics: Concepts and Controversies (3d ed. 1991); Hans Zeisel, Say It with Figures (6th ed. 1985); and Zeisel & Kaye, *supra* note 1.

17. When relevant studies exist before the commencement of the litigation, 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 and studies of how sex stereotyping affects perceptions of women in the workplace. *See, 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 accuracy); Price Waterhouse v. Hopkins, 490 U.S. 228, 235 (1989). Some psychologists have questioned the applicability of these experiments to litigation. *See, e.g.,* 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). For a rejoinder, see Susan T. Fiske et al., *What Constitutes a Scientific Review?: A Majority Retort to Barrett and Morris*, 17 Law & Hum. Behav. 217 (1993).

If no preexisting studies are available, a case-specific one may be devised. *E.g.*, United States v. Youritan Constr. Co., 370 F. Supp. 643, 647 (N.D. Cal. 1973) (investigating racial discrimination in the rental-housing market by using "testers"—who should differ only in their race—to rent a property),

more useful as a stimulus for further inquiry than as a basis for establishing association. Observational studies can establish that one factor is associated with another, but considerable analysis may be necessary to bridge the gap from association to causation.[18] Controlled experiments are ideal for ascertaining causation, but they can be difficult to undertake.

"Anecdotal evidence" means reports of one kind of event following another. Typically, the reports are obtained haphazardly or selectively, and the logic of "post hoc, ergo propter hoc" does not suffice to demonstrate that the first event causes the second. Consequently, while anecdotal evidence can be suggestive,[19] it can also be quite misleading.[20] For instance, some children who live near power lines develop leukemia; but does exposure to electrical and magnetic fields cause this disease? The anecdotal evidence is not compelling because leukemia also occurs among children who have minimal exposure to such fields.[21] It is necessary to compare disease rates among those who are exposed and those who are not. If exposure causes the disease, the rate should be higher among the exposed, lower among the unexposed. Of course, the two groups may differ in crucial ways other than the exposure. For example, children who live near power

aff'd in part, 509 F.2d 623 (9th Cir. 1975). For a critical review of studies using testers, see James J. Heckman & Peter Siegelman, *The Urban Institute Audit Studies: Their Methods and Findings*, *in* Clear and Convincing Evidence: Measurement of Discrimination in America 187 (Michael Fix & Raymond J. Struyk eds., 1993) (including commentary).

18. For example, smokers have higher rates of lung cancer than nonsmokers; thus smoking and lung cancer are associated.

19. In medicine, evidence from clinical practice is often the starting point for the demonstration of a causal effect. One famous example involves exposure of mothers to German measles during pregnancy, followed by blindness in their babies. N. McAlister Gregg, *Congenital Cataract Following German Measles in the Mother*, 3 Transactions Ophthalmological Soc'y Austl. 35 (1941), *reprinted in* The Challenge of Epidemiology 426 (Carol Buck et al. eds., 1988).

20. Indeed, some courts have suggested that attempts to infer causation from anecdotal reports are inadmissible as unsound methodology under *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993). *See, e.g.*, Haggerty v. Upjohn Co., 950 F. Supp. 1160, 1163–64 (S.D. Fla. 1996) (holding that reports to the Food and Drug Administration of "adverse medical events" involving the drug Halcion and "anecdotal case reports appearing in medical literature . . . can be used to generate hypotheses about causation, but not causation conclusions" because "scientifically valid cause and effect determinations depend on controlled clinical trials and epidemiological studies"); Cartwright v. Home Depot U.S.A., Inc., 936 F. Supp. 900, 905 (M.D. Fla. 1996) (excluding an expert's opinion that latex paint caused plaintiff's asthma, in part because "case reports . . . are no substitute for a scientifically designed and conducted inquiry").

21. *See* Committee on the Possible Effects of Electromagnetic Fields on Biologic Sys., National Research Council, Possible Health Effects of Exposure to Residential Electric and Magnetic Fields (1997); Zeisel & Kaye, *supra* note 1, at 66–67. There are serious problems in measuring exposure to electromagnetic fields, and results are somewhat inconsistent from one study to another. For such reasons, the epidemiologic evidence for an effect on health is quite inconclusive. *Id.*; Martha S. Linet et al., *Residential Exposure to Magnetic Fields and Acute Lymphoblastic Leukemia in Children*, 337 New Eng. J. Med. 1 (1997); Edward W. Campion, *Power Lines, Cancer, and Fear*, 337 New Eng. J. Med. 44 (1997) (editorial); Gary Taubes, *Magnetic Field-Cancer Link: Will It Rest in Peace?*, 277 Science 29 (1997) (quoting various epidemiologists).

lines could come from poorer families and be exposed to other environmental hazards. These differences could create the appearance of a cause-and-effect relationship, or they can mask a real relationship. Cause-and-effect relationships often are quite subtle, and carefully designed studies are needed to draw valid conclusions.[22]

Typically, a well-designed study will compare outcomes for subjects who are exposed to some factor—the treatment group—and other subjects who are not so exposed—the control group. A distinction must then be made between controlled experiments and observational studies. In a controlled experiment, the investigators decide which subjects are exposed to the factor of interest and which subjects go into the control group. In most observational studies, the subjects themselves choose their exposures. Because of this self-selection, the treatment and control groups are likely to differ with respect to important factors other than the one of primary interest.[23] (These other factors are called confounding variables or lurking variables.[24]) With studies on the health effects of power lines, family background is a possible confounder; so is exposure to other hazards.[25]

22. Here is a classic example from epidemiology. At one time, it was thought that lung cancer was caused by fumes from tarring the roads, because many lung cancer patients lived near roads that had recently been paved. This is anecdotal evidence. But the logic is quite incomplete, because many people without lung cancer were exposed to asphalt fumes. A comparison of rates is needed. Careful study showed that lung cancer patients had similar rates of exposure to tar fumes as other people; the real difference was in exposure to cigarette smoke. Richard Doll & A. Bradford Hill, *A Study of the Aetiology of Carcinoma of the Lung*, 2 Brit. Med. J. 1271 (1952).

23. 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. When 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; dependent variables also are called response variables. In contrast, the variables that represent the causes are called independent variables; independent variables also are called factors or explanatory variables.

24. A confounding variable is correlated with the independent variables and with the dependent variable. If the units being studied differ on the independent variables, they are also likely to differ on the confounder. Therefore, the confounder—not the independent variables—could be responsible for differences seen on the dependent variable.

25. Confounding is a problem even in careful epidemiologic studies. For example, women with herpes are more likely to develop cervical cancer than women who have not been exposed to the virus. It was concluded 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 Cervical Cancer and Human Papillomavirus (N. Muñoz et al. eds. 1992). For additional examples and discussion, see Freedman et al., *supra* note 16, at 12–27, 150–52; David Freedman, *From Association to Causation: Some Remarks on the History of Statistics,* 14 Stat. Sci. 243 (1999).

2. Randomized Controlled Experiments

In randomized controlled experiments, investigators assign subjects to treatment or control groups at random. The groups are therefore likely to be quite comparable—except for the treatment. Choosing at random tends to balance the groups with respect to possible confounders, and the effect of remaining imbalances can be assessed by statistical techniques.[26] Consequently, inferences based on well-executed randomized experiments are more secure than inferences based on observational studies.[27]

The following illustration brings together the points made thus far. Many doctors think that taking aspirin helps prevent heart attacks, but there is some controversy. Most people who take aspirin do not have heart attacks; this is anecdotal evidence for the protective effect, but proves very little. After all, most people do not suffer heart attacks—whether or not they take aspirin regularly. A careful study must compare heart attack rates for two groups: persons who take aspirin (the treatment group) and persons who do not (the controls). An observational study would be easy to do, but then the aspirin-takers are likely to be different from the controls. If, for instance, the controls are healthier to begin with, the study would be biased against the drug. Randomized experiments with aspirin are harder to do, but they provide much better evidence. It is the experiments that demonstrate a protective effect.

To summarize: First, outcome figures from a treatment group without a control group generally reveal very little and can be misleading. 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.[28] However, if the control group was created in any

26. *See infra* § IV.

27. Experiments, however, are often impractical, as in the power-line example. Even when randomized controlled experiments are feasible, true randomization can be difficult to achieve. *See, e.g.*, Kenneth F. Schulz, *Subverting Randomization in Controlled Trials*, 274 JAMA 1456 (1995); Rachel Nowak, *Problems in Clinical Trials Go Far Beyond Misconduct*, 264 Science 1538 (1994). For statistical purposes, randomization should be accomplished using some definite, objective method (like a random number generator on a computer); haphazard assignment may not be sufficient.

28. 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 of subjects to treatment or control groups puts statistical tests of significance on a secure footing. Freedman et al., *supra* note 16, at 503–24, 547–78.

Even more important, 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

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. Observational Studies

The bulk of the statistical studies seen in court are observational, not experimental. 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.[29]

Nevertheless, many studies of the deterrent effect of the death penalty have been conducted, all observational, and some have attracted judicial attention.[30] 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 punishment) and of controlling for potentially confounding variables (e.g., worsening economic conditions).[31] 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 randomized controlled experiment.[32]

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. See also* Colin Begg et al., *Improving the Quality of Reporting of Randomized Controlled Trials: The CONSORT Statement*, 276 JAMA 637 (1996).

29. *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).

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

31. A procedure often used to control for confounding in observational studies is regression analysis. The underlying logic is described *infra* § V.D and in Daniel L. Rubinfeld, Reference Guide on Multiple Regression, § II, in this manual. 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).

32. *See, e.g.*, Experimentation in the Law, *supra* note 29, 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

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.[33]
- 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.[34]

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)?

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.

33. 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. Furthermore, many different studies—of different types and on different populations—confirm the causal link. That is why most experts believe that smoking causes lung cancer and many other diseases. For a 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).

34. A. Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 Proc. Royal Soc'y Med. 295 (1965); Alfred S. Evans, Causation and Disease: A Chronological Journey 187 (1993).

- If the study 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?[35]

4. Can the Results Be Generalized?

Any study must be conducted on a certain group of subjects, at certain times and places, using certain treatments. With respect to these subjects, the study may be persuasive. There may be adequate control over confounding variables, and there may be an unequivocally large difference between the treatment and control groups. If so, the study's internal validity will not be disputed: for the subjects in the study, the treatment had an effect. But an issue of external validity remains. To extrapolate from the conditions of a study to more general circumstances always raises questions. For example, studies suggest that definitions of insanity given to jurors influence decisions in cases of incest;[36] would the definitions have a similar effect in cases of murder? Other studies indicate that recidivism rates for ex-convicts are not affected by temporary financial support after release.[37] Would the same results be obtained with different conditions in the labor market?

Confidence in the appropriateness of an extrapolation cannot come from the experiment itself.[38] It must come from knowledge about which outside factors

35. These questions are adapted from Freedman et al., *supra* note 16, at 28. For discussions of the admissibility or weight of studies that overlook obvious possible confounders, see *People Who Care v. Rockford Board of Education*, 111 F.3d 528, 537–38 (7th Cir. 1997) ("The social scientific literature on educational achievement identifies a number of other variables besides poverty and discrimination that explain differences in scholastic achievement, such as the educational attainments of the student's parents and the extent of their involvement in their children's schooling. . . . These variables cannot be assumed to be either randomly distributed across the different racial and ethnic groups in Rockford or perfectly correlated with poverty. . . ."); cases cited *supra* note 5 and *infra* note 230.

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

37. 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).

38. Suppose an epidemiologic study is conducted on the relationship between a toxic substance and a disease. The rate of occurrence of the disease in a group of persons exposed to the substance is compared to the rate in a control group, and the rate in the exposed group turns out to be more than double the rate in the control group. (More technically, the relative risk exceeds two.) Do these data imply that a plaintiff who was exposed to the toxic substance and contracted the disease probably would not have contracted the disease but for the exposure? If we assume that the substance causes the disease and all confounding has been properly accounted for (a judgment that might not be easy to defend),

would or would not affect the outcome.[39] Sometimes, several experiments or other studies, each having different limitations, all point in the same direction. 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.[40] Such convergent results strongly suggest the validity of the generalization.

then we can conclude that over half the cases of disease in the exposed group would not be there but for the exposure. Applying this arithmetic to a specific person, however, is problematic. For instance, the relative risk is an average over all the subjects included in the study. The exposures and susceptibilities almost certainly are not uniform, and the plaintiff's exposure and susceptibility cannot be known from the study. Nevertheless, several courts and commentators have stated that a relative risk of more than two demonstrates specific causation, or, conversely, that a relative risk of two or less precludes a finding of specific causation. *E.g.*, DeLuca v. Merrell Dow Pharms., Inc., 911 F.2d 941, 958–59 (3d Cir. 1990); Marder v. G.D. Searle & Co., 630 F. Supp. 1087, 1092 (D. Md. 1986) ("a two-fold increased risk is . . . the equivalent of the required legal burden of proof—a showing of causation by the preponderance of the evidence or, in other words, a probability of greater than 50%"), *aff'd sub nom.* Wheelahan v. G.D. Searle & Co., 814 F.2d 655 (4th Cir. 1987); Bert Black & David E. Lilienfeld, *Epidemiologic Proof in Toxic Tort Litigation*, 52 Fordham L. Rev. 732, 769 (1984); Michael D. Green et al., Reference Guide on Epidemiology, § VII, in this manual. A few commentators have sharply criticized this reasoning. Steven E. Fienberg et al., *Understanding and Evaluating Statistical Evidence in Litigation*, 36 Jurimetrics J. 1, 9 (1995); Diana B. Petitti, *Reference Guide on Epidemiology*, 36 Jurimetrics J. 159, 168 (1996) (review essay); D.A. Freedman & Philip B. Stark, *The Swine Flu Vaccine and Guillain-Barré Syndrome: A Case Study in Relative Risk and Specific Causation,* 23 Evaluation Rev. 619 (1999); James Robins & Sander Greenland, *The Probability of Causation Under a Stochastic Model for Individual Risk*, 45 Biometrics 1125, 1126 (1989); Melissa Moore Thompson, Comment, *Causal Inference in Epidemiology: Implications for Toxic Tort Litigation*, 71 N.C. L. Rev. 247 (1992).

39. Such judgments are easiest in the physical and life 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 often 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 by itself to establish causation. *See generally* Bruce N. Ames et al., *The Causes and Prevention of Cancer*, 92 Proc. Nat'l Acad. Sci. USA 5258 (1995); 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 Comparative Toxicity of Naturally Occurring Carcinogens, National Research Council, Carcinogens and Anticarcinogens in the Human Diet: A Comparison of Naturally Occurring and Synthetic Substances (1996); Committee on Risk Assessment of Hazardous Air Pollutants, National Research Council, Science and Judgment in Risk Assessment 59 (1994) ("There are reasons based on both biologic principles and empirical observations to support the hypothesis that many forms of biologic responses, including toxic responses, can be extrapolated across mammalian species, including *Homo sapiens*, but the scientific basis of such extrapolation is not established with sufficient rigor to allow broad and definitive generalizations to be made.").

40. 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.

B. Descriptive Surveys and Censuses

Having discussed the statistical logic of studies to investigate causation, we now turn to a second topic—sampling, that is, choosing units for study. A census tries to measure 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.[41]

1. What Method Is Used to Select the Units?

By definition, a census seeks to measure some characteristic of every unit in a whole 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.[42] Supplemental studies may enable statisticians to adjust for such omissions, but the adjustments may rest on uncertain assumptions.[43]

The methodological framework of a scientific survey is more complicated than that of a census. In surveys that use probability 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

41. For more extended treatment of these issues, see Shari Seidman Diamond, Reference Guide on Survey Research, § III, in this manual.

42. *See generally* Harvey M. Choldin, Looking for the Last Percent: The Controversy Over Census Undercounts 42–43 (1994).

43. For conflicting views on proposed adjustments to the 1990 census, see the exchanges of papers at 9 Stat. Sci. 458 (1994), 18 Surv. Methodology No. 1 (1992), 88 J. Am. Stat. Ass'n 1044 (1993), and 34 Jurimetrics J. 65 (1993). In *Wisconsin v. City of New York*, 517 U.S. 1 (1996), the Supreme Court resolved the conflict among the circuits over the legal standard governing claims that adjustment is compelled by statute or the Constitution. The Court unanimously determined that the exacting requirements of the equal protection clause, as explicated in congressional redistricting and state reapportionment cases, do not "translate into a requirement that the Federal Government conduct a census that is as accurate as possible" and do not provide any basis for "preferring numerical accuracy to distributive accuracy." *Id.* at 17, 18. The Court therefore applied a much less demanding standard to the Secretary's decision. Concluding that the government had shown "a reasonable relationship" between the decision not to make post hoc adjustments and "the accomplishment of an actual enumeration of the population, keeping in mind the constitutional purpose of the census . . . to determine the apportionment of the Representatives among the States," the Court held that the decision satisfied the Constitution. Indeed, having rejected the argument that the Constitution compelled statistical adjustment, the Court noted that the Constitution might prohibit such adjustment. *Id.* at 19 n.9, 20.

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.[44] In this case, the fit between the sampling frame and the population would be excellent.[45]

In other situations, the sampling frame may cover less of the population. In an obscenity case, for example, the defendant's opinion poll about community standards[46] should identify the population as all adults in the legally relevant community, but obtaining a full list 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.[47] The poll's measurement of community opinion will be biased, although the magnitude of this bias may not be great.

44. 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).

45. Likewise, in drug investigations the sampling frame for testing the contents of vials, bags, or packets seized by police easily can be devised to match the population of all the items seized in a single case. Because testing each and every item can be quite time-consuming and expensive, chemists often draw a probability sample, analyze the material that is sampled, and use the percentage of illicit drugs found in the sample to determine the total quantity of illicit drugs in all the items seized. *E.g.*, United States v. Shonubi, 895 F. Supp. 460, 470 (E.D.N.Y. 1995) (citing cases), *rev'd on other grounds*, 103 F.3d 1085 (2d Cir. 1997). For discussions of statistical estimation in such cases, see C.G.G. Aitken et al., *Estimation of Quantities of Drugs Handled and the Burden of Proof*, 160 J. Royal Stat. Soc'y 333 (1997); Dov Tzidony & Mark Ravreby, *A Statistical Approach to Drug Sampling: A Case Study*, 37 J. Forensic Sci. 1541 (1992); Johan Bring & Colin Aitken, *Burden of Proof and Estimation of Drug Quantities Under the Federal Sentencing Guidelines*, 18 Cardozo L. Rev. 1987 (1997).

46. On the admissibility of such polls, compare *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. 1990) ("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).

47. 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 16, at 334–35. 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 335, A-20 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 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.

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.[48] 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.[49] Selection bias is acute when constituents write their representatives, listeners call into radio talk shows, interest groups collect information from their members,[50] or attorneys choose cases for trial.[51] Selection bias also affects data from jury-reporting services that gather information from readily available sources.

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 too much discretion to the interviewers 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[52] leaves no room for selection bias. These procedures are used routinely to select individuals for jury

48. *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).

49. Nonresponse bias is discussed *infra* § II.B.2.

50. *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).

51. *See In re* Chevron U.S.A., Inc., 109 F.3d 1016 (5th Cir. 1997). In that case, the district court decided to try 30 cases to resolve common issues or to ascertain damages in 3,000 claims arising from Chevron's allegedly improper disposal of hazardous substances. The court asked the opposing parties to select 15 cases each. Selecting 30 extreme cases, however, is quite different from drawing a random sample of 30 cases. Thus, the court of appeals wrote that although random sampling would have been acceptable, the trial court could not use the results in the 30 extreme cases to resolve issues of fact or ascertain damages in the untried cases. *Id.* at 1020. Those cases, it warned, were "not cases calculated to represent the group of 3,000 claimants." *Id.*

52. In simple random sampling, units are drawn at random without replacement. In particular, each unit has the same probability of being chosen for the sample. 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.

duty;[53] they also have been used to choose "bellwether" cases for representative trials to resolve issues in all similar cases.[54]

2. Of the Units Selected, Which Are Measured?

Although probability sampling ensures that, within the limits of chance, the sample will be representative of the sampling frame, the question remains as to which units actually get measured. When objects like receipts are sampled for an audit, or vegetation is sampled for a study of the ecology of a region, all the selected units 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,[55] although supplemental study may establish that the nonrespondents do not differ systematically from the respondents with respect to the characteristics of interest[56] or may permit the missing data to

53. 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.

54. Hilao v. Estate of Marcos, 103 F.3d 767 (9th Cir. 1996); Cimino v. Raymark Indus., Inc., 751 F. Supp. 649 (E.D. Tex. 1990); *cf. In re* Chevron U.S.A., Inc., 109 F.3d 1016 (5th Cir. 1997) (discussed *supra* note 51). Although trials in a suitable random sample of cases can produce reasonable estimates of average damages, the propriety of precluding individual trials has been debated. *Compare* 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), *with Chevron*, 109 F.3d at 1021 (Jones, J., concurring); Robert G. Bone, *Statistical Adjudication: Rights, Justice, and Utility in a World of Process Scarcity*, 46 Vand. L. Rev. 561 (1993).

55. The 1936 *Literary Digest* election poll (see *supra* note 47) illustrates the danger. Only 24% of the 10 million people who received questionnaires returned them. Most of the respondents probably had strong views on the candidates, and most of them probably objected to President Roosevelt's economic program. 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 16, at 335–36.

In *United States v. Gometz*, 730 F.2d 475, 478 (7th Cir. 1984) (en banc), 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 follow up on persons who do not respond to a jury questionnaire," the court concluded that Congress "was not concerned with anything so esoteric as nonresponse bias." *Id.* at 479, 482.

56. 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

be imputed.[57]

In short, a good survey defines an appropriate population, uses an unbiased method for selecting the sample, has a high response rate, and gathers accurate information on the sample 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.[58] 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. Individual Measurements

1. Is the Measurement Process Reliable?

There are two main aspects to the accuracy of measurements—reliability and validity. In science, "reliability" refers to reproducibility of results.[59] 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

"outcry" of 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 nonrandom 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). For further criticism of this study, see David Streitfeld, *Shere Hite and the Trouble with Numbers*, 1 Chance 26 (1988).

57. Methods for "imputing" missing data are discussed in, e.g., Tanur, *supra* note 47, at 66 and Howard Wainer, *Eelworms, Bullet Holes, and Geraldine Ferraro: Some Problems with Statistical Adjustment and Some Solutions*, 14 J. Educ. Stat. 121 (1989) (with commentary). 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. Otherwise, imputation can be problematic.

58. *See infra* § IV.

59. 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.*, 509 U.S. 579, 590 n.9 (1993), for instance, distinguishes "evidentiary reliability" from reliability in the technical sense of giving consistent results. We use "reliability" to denote the latter.

same weight for the same object. Its errors, if any, are systematic; they always point in the same direction.

Reliability can be ascertained by measuring the same quantity several times. For instance, one method of DNA identification requires a laboratory to determine the lengths of fragments of DNA. By making duplicate measurements of DNA fragments, a laboratory can determine the likelihood that two measurements will differ by specified amounts.[60] Such results are needed when deciding whether an observed discrepancy between a crime sample and a suspect sample is sufficient to exclude the suspect.[61]

In many studies, descriptive information is obtained on the subjects. For statistical purposes, the information may have to be reduced to numbers, a process called "coding." The reliability of the coding process should be considered. 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.[62] 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. Another example involves the MMPI (Minnesota Multiphasic Personality Inventory), a pencil and paper test that, many psychologists agree, measures aspects of personality or psychological functioning. Its reliability can be quantified. But this does not make it a valid test of sexual deviancy.[63]

When an independent and reasonably accurate way of measuring the variable of interest is available, it may be used to validate the measuring system in ques-

60. *See* Committee on DNA Forensic Science: An Update, National Research Council, The Evaluation of Forensic DNA Evidence 139–41 (1996).

61. *Id.*; Committee on DNA Tech. in Forensic Science, National Research Council, DNA Technology in Forensic Science 61–62 (1992); David H. Kaye & George F. Sensabaugh, Jr., Reference Guide on DNA Evidence, § VII, in this manual.

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

63. *See* People v. John W., 229 Cal. Rptr. 783, 785 (Ct. App. 1986) (holding that because the use of the MMPI to diagnose sexual deviancy was not shown to be generally accepted as valid in the scientific community, a diagnosis based in part on the MMPI was inadmissible).

tion. 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).[64]

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 distort the results?[65]

III. How Have the Data Been Presented?

After data have been collected, they should be presented in a way that makes them intelligible. Data can be summarized with a few numbers or with graphical displays. However, the wrong summary can mislead.[66] Section III.A discusses rates or percentages, and gives some cautionary examples of misleading summaries, indicating the sorts of questions that might be considered when numerical summaries are presented in court. Percentages are often used to demonstrate statistical association, which is the topic of section III.B. Section III.C

64. *E.g.*, Washington v. Davis, 426 U.S. 229, 252 (1976); Albemarle Paper Co. v. Moody, 422 U.S. 405, 430–32 (1975). As the discussion of the correlation coefficient indicates, *infra* § V.B, the closer the coefficient is to 1, the greater the validity. Various statistics are used to characterize the reliability of laboratory instruments, psychological tests, or human judgments. These include the standard deviation as well as the correlation coefficient. *See infra* §§ III, V.

65. *See, e.g.*, McCleskey v. Kemp, 753 F.2d 877, 914–15 (11th Cir. 1985) (district court was unpersuaded by a statistical analysis of capital sentencing, in part because of various imperfections in the study, including discrepancies in the data and missing data; concurring and dissenting opinion concludes that the district court's findings on missing and misrecorded data were clearly erroneous because the possible errors were not large enough to affect the overall results; for an exposition of the study and response to such criticisms, see Baldus et al., *supra* note 62), *aff'd*, 481 U.S. 279 (1987); G. Heileman Brewing Co. v. Anheuser-Busch, Inc., 676 F. Supp. 1436, 1486 (E.D. Wis. 1987) ("many coding errors . . . affected the results of the survey"); EEOC v. Sears, Roebuck & Co., 628 F. Supp. 1264, 1304, 1305 (N.D. Ill. 1986) ("[E]rrors in EEOC's mechanical coding of information from applications in its hired and nonhired samples also make EEOC's statistical analysis based on this data less reliable." The EEOC "consistently coded prior experience in such a way that less experienced women are considered to have the same experience as more experienced men" and "has made so many general coding errors that its data base does not fairly reflect the characteristics of applicants for commission sales positions at Sears."), *aff'd*, 839 F.2d 302 (7th Cir. 1988); Dalley v. Michigan Blue Cross-Blue Shield, Inc., 612 F. Supp. 1444, 1456 (E.D. Mich. 1985) ("although plaintiffs show that there were some mistakes in coding, plaintiffs still fail to demonstrate that these errors were so generalized and so pervasive that the entire study is invalid").

66. *See generally* Freedman et al., *supra* note 16; Huff, *supra* note 16; Moore, *supra* note 16; Zeisel, *supra* note 16.

considers graphical summaries of data, while sections III.D and III.E discuss some of the basic descriptive statistics that are likely to be encountered in litigation, including the mean, median and standard deviation.

A. Are Rates or Percentages Properly Interpreted?

1. Have Appropriate Benchmarks Been Provided?

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. On the other hand, according to the *Wall Street Journal,* 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.[67] In this example, and in many other situations, it is helpful to look beyond a single number to some benchmark that places the isolated figure into perspective.

2. Have the Data-Collection Procedures Changed?

Changes in the process of collecting data can create problems of interpretation. Statistics on crime provide many examples. 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.[68] 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 Reports.[69]

Changes in data-collection procedures are by no means limited to crime statistics.[70] Indeed, almost all series of numbers that cover many years are affected by changes in definitions and collection methods. When a study includes such time series data, it is useful to inquire about changes and to look for any sudden jumps, which may signal such changes.[71]

67. Moore, *supra* note 16, at 161.

68. *Id.* at 162.

69. James P. Levine et al., Criminal Justice in America: Law in Action 99 (1986).

70. For example, improved survival rates for cancer patients may result from improvements in therapy. Or, the change may simply mean that cancers now are detected earlier, due to improvements in diagnostic techniques, so that patients with these cancers merely appear to live longer. *See* Richard Doll & Richard Peto, The Causes of Cancer: Quantitative Estimates of Avoidable Risks of Cancer in the United States Today app. C at 1278–79 (1981).

71. Moore, *supra* note 16, at 162.

3. Are the Categories Appropriate?

Misleading summaries also can be produced by choice of categories for comparison. In *Philip Morris, Inc. v. Loew's Theatres, Inc.*,[72] and *R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc.*,[73] 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 the defendant contained the data shown in Table 1.[74]

Table 1. Data used by defendant to refute plaintiffs' false advertising claim

	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%

Only 14% + 22% = 36% of the sample preferred Triumph to Merit, while 29% + 11% = 40% preferred Merit to Triumph.[75] 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."[76] The court correctly resisted this chicanery, finding that defendant's test results did not support the advertising claims.[77]

There was a similar distortion in claims for the accuracy of a home pregnancy test.[78] The manufacturer advertised the test as 99.5% accurate under laboratory conditions. The data underlying this claim are summarized in Table 2.

Table 2. Home pregnancy test results

	Actually pregnant	Actually not pregnant
Test says pregnant	197	0
Test says not pregnant	1	2
Total	198	2

72. 511 F. Supp. 855 (S.D.N.Y. 1980).
73. 511 F. Supp. 867 (S.D.N.Y. 1980).
74. 511 F. Supp. at 866.
75. *Id.* at 856.
76. *Id.* at 866.
77. *Id.* at 856–57. The statistical issues in these cases are discussed more fully in 2 Gastwirth, *supra* note 1, at 633–39.
78. This incident is reported in Arnold Barnett, *How Numbers Can Trick You*, Tech. Rev., Oct. 1994, at 38, 44–45.

Table 2 does indicate only one error in 200 assessments, or 99.5% overall accuracy. But the table also shows that the test can make two types of errors—it can tell a pregnant woman that she is not pregnant (a false negative), and it can tell a woman who is not pregnant that she is (a false positive). The reported 99.5% accuracy rate conceals a crucial fact—the company had virtually no data with which to measure the rate of false positives.[79]

4. How Big Is the Base of a Percentage?

Rates and percentages often provide effective summaries of data, but these statistics can be misinterpreted. A percentage makes a comparison between two numbers: one number is the base, and the other number is compared to that base. When the base is small, actual numbers may be more revealing than percentages. Media accounts in 1982 of a crime wave by the elderly give an example. 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.[80]

5. What Comparisons Are Made?

Finally, there is the issue of which numbers to compare. Researchers sometimes choose among alternative comparisons. It may be worthwhile to ask why they chose the one they did. Would another comparison give a different view? A government agency, for example, may want to compare the amount of service now being given 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 should be expected because of start-up problems.[81] If last year is used as the base, was it also part of the trend, or was it an unusually poor year? If the base year is not representative of other years, then the percentage may not portray the trend fairly.[82] No single question can be formulated to detect such distortions, but it may help to ask for the numbers from which the percentages were obtained;

79. Only two women in the sample were not pregnant; the test gave correct results for both of them. Although a false-positive rate of zero is ideal, an estimate based on a sample of only two women is not.

80. Mark H. Maier, The Data Game: Controversies in Social Science Statistics 83 (1991). *See also* Alfred Blumstein & Jacqueline Cohen, *Characterizing Criminal Careers*, 237 Science 985 (1987).

81. *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 . . . .").

82. Jeffrey Katzer et al., Evaluating Information: A Guide for Users of Social Science Research 106 (2d ed. 1982).

asking about the base can also be helpful. Ultimately, however, recognizing which numbers are related to which issues requires a species of clear thinking not easily reducible to a checklist.[83]

B. Is an Appropriate Measure of Association Used?

Many cases involve statistical association. Does a test for employee promotion have an exclusionary effect that depends on race or 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? This section discusses tables and percentage-based statistics that are frequently presented to answer such questions.[84]

Percentages often are used to describe the association between two variables. Suppose that a university alleged to discriminate against women in admitting students consists of only two colleges, engineering and business. The university admits 350 out of 800 male applicants; by comparison, it admits only 200 out of 600 female applicants. Such data commonly are displayed as in Table 3.[85]

Table 3. Admissions by gender

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

As Table 3 indicates, 350/800 = 44% of the males are admitted, compared with only 200/600 = 33% of the females. One way to express the disparity is to subtract the two percentages: 44% – 33% = 11 percentage points. Although such subtraction is commonly seen in jury discrimination cases,[86] the difference is inevitably small when the two percentages are both close to zero. If the selection rate for males is 5% and that for females is 1%, the difference is only 4 percentage points. Yet, females have only 1/5 the chance of males of being admitted, and that may be of real concern.[87]

83. For assistance in coping with percentages, see Zeisel, *supra* note 16, at 1–24.

84. Correlation and regression are discussed *infra* § V.

85. A table of this sort is called a "cross-tab" or a "contingency table." Table 3 is "two-by-two" because it has two rows and two columns, not counting rows or columns containing totals.

86. *See, e.g.*, D.H. Kaye, *Statistical Evidence of Discrimination in Jury Selection*, *in* Statistical Methods in Discrimination Litigation, *supra* note 11, at 13.

87. *Cf.* United States v. Jackman, 46 F.3d 1240, 1246–47 (2d Cir. 1995) (holding that the small percentage of minorities in the population makes it "inappropriate" to use an "absolute numbers" or "absolute impact" approach for measuring underrepresentation of these minorities in the list of potential jurors).

For Table 3, the selection ratio (used by the Equal Employment Opportunity Commission (EEOC) in its "80% rule")[88] is 33/44 = 75%, meaning that, on average, women have 75% the chance of admission that men have.[89] However, the selection ratio has its own problems. In the last example, if the selection rates are 5% and 1%, then the exclusion rates are 95% and 99%. The corresponding ratio is 99/95 = 104%, meaning that females have, on average, 104% the risk of males of being rejected. The underlying facts are the same, of course, but this formulation sounds much less disturbing.[90]

The odds ratio is more symmetric. If 5% of male applicants are admitted, the odds on a man being admitted are 5/95 = 1/19; the odds on a woman being admitted are 1/99. The odds ratio is (1/99)/(1/19) = 19/99. The odds ratio for rejection instead of acceptance is the same, except that the order is reversed.[91] Although the odds ratio has desirable mathematical properties, its meaning may be less clear than that of the selection ratio or the simple difference.

Data showing disparate impact are generally obtained by aggregating—putting together—statistics from a variety of sources. Unless the source material is fairly homogenous, aggregation can distort patterns in the data. We illustrate the problem with the hypothetical admission data in Table 3. Applicants can be classified not only by gender and admission but also by the college to which they applied, as in Table 4:

Table 4. 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 4 add up to the entries in Table 3; said more technically, Table 3 is obtained by aggregating the data in Table 4. Yet, there is no association between gender and admission in either college; men and women are ad-

88. 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). The rule is designed to help spot instances of substantially discriminatory practices, and the commission usually asks employers to justify any procedures that produce selection ratios of 80% or less.

89. The analogous statistic used in epidemiology is called the relative risk. *See supra* note 38; Michael D. Green et al., Reference Guide on Epidemiology, § III.A, in this manual. Relative risks are usually quoted as decimals rather than percentages; for instance, a selection ratio of 75% corresponds to a relative risk of 0.75. A variation on this idea is the relative difference in the proportions, which expresses the proportion by which the probability of selection is reduced. Kairys et al., *supra* note 44, at 776, 789–90; *cf.* David C. Baldus & James W.L. Cole, Statistical Proof of Discrimination § 5.1, at 153 (1980 & Supp. 1987) (listing various ratios that can be used to measure disparities).

90. The Illinois Department of Employment Security tried to exploit this feature of the selection

mitted at identical rates. 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, admits relatively few applicants; the engineering college, to which most of the men applied, is easier to get into. This example illustrates a common issue: association can result from combining heterogeneous statistical material.[92]

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.[93]

1. How Are Trends Displayed?

Graphs that plot values over time are useful for seeing trends. However, the scales on the axes matter. In Figure 1, the federal debt appears to skyrocket during the Reagan and Bush administrations; in Figure 2, the federal debt appears to grow slowly.[94] The moral is simple: Pay attention to the markings on the axes to determine whether the scale is appropriate.

ratio in *Council 31, Am. Fed'n of State, County and 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. *Id.* at 375. Recognizing that these layoffs ran afoul of the 80% rule (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, the retention rates showed no adverse impact under the 80% rule. *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 . . . shows conclusively that the layoffs did not have a disparate impact." *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.*

91. For women, the odds on rejection are 99 to 1; for men, 19 to 1. The ratio of these odds is 99/19. Likewise, the odds ratio for an admitted applicant being a man as opposed to a denied applicant being man is also 99/19.

92. Tables 3 and 4 are hypothetical, but closely patterned on a real example. *See* P.J. Bickel et al., *Sex Bias in Graduate Admissions: Data from Berkeley*, 187 Science 398 (1975). *See also* Freedman et al., *supra* note 16, at 17–20; Moore, *supra* note 16, at 246–47. The tables are 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). Another perspective on Table 3 may be helpful. The college to which a student applies is a confounder. *See supra* § II.A.1. In the present context, confounders often are called "omitted variables." For opinions discussing the legal implications of omitted variables, see cases cited *supra* note 5 and *infra* note 230.

93. *See generally* William S. Cleveland, The Elements of Graphing Data (1985); David S. Moore & George P. McCabe, Introduction to the Practice of Statistics 3–20 (2d ed. 1993). Graphs showing relationships among variables are discussed *infra* § V.

94. *See* Howard Wainer, *Graphs in the Presidential Campaign*, Chance, Winter 1993, at 48, 50.

Figure 1. The federal debt skyrockets under Reagan–Bush.

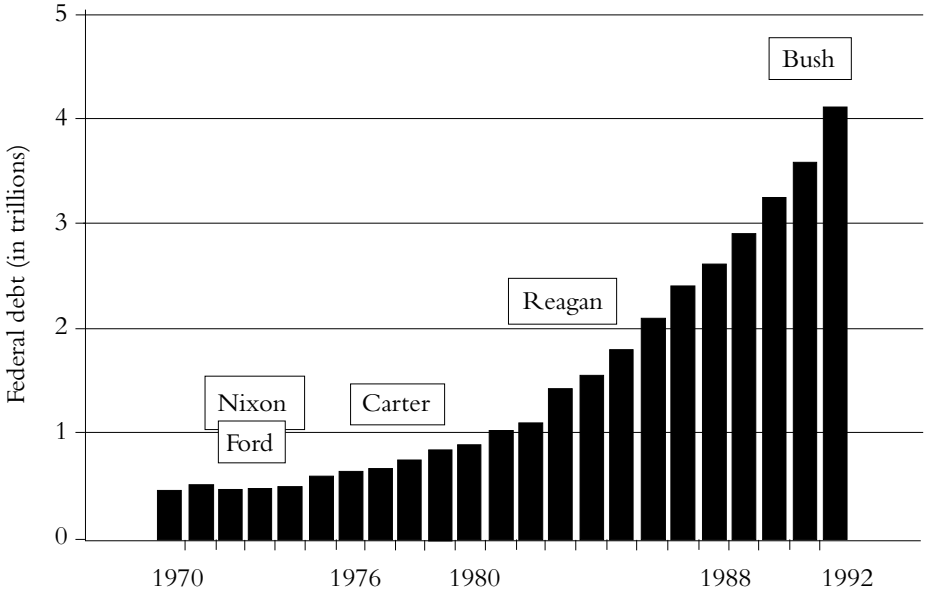
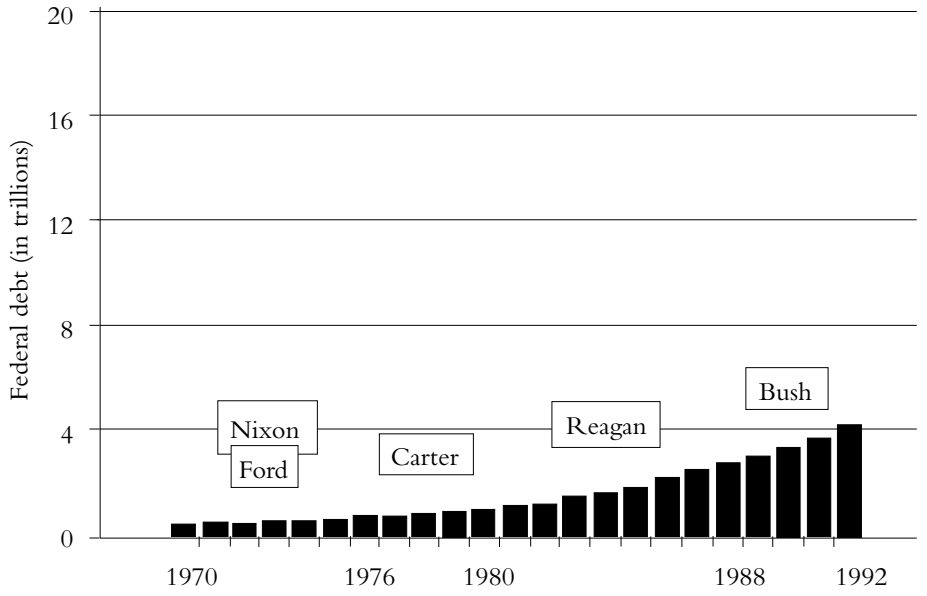


Figure 2. The federal debt grows steadily under Reagan–Bush.

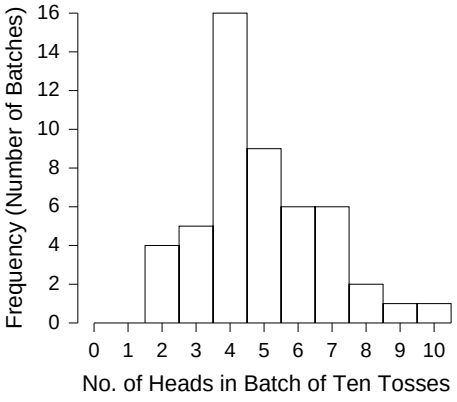


2. How Are Distributions Displayed?

A graph commonly used to display the distribution of data is the histogram.[95] One axis denotes the numbers, and the other indicates how often those fall within specified intervals (called "bins" or "class intervals"). For example, we flipped a quarter 10 times in a row and counted the number of heads in this "batch" of 10 tosses. With 50 batches, we obtained the following counts:[96]

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

Figure 3. Histogram showing how frequently various numbers of heads appeared in 50 batches of 10 tosses of a quarter.



The histogram is shown in Figure 3.[97] A histogram shows how the data are distributed over the range of possible values. The spread can be made to appear

95. For small batches of numbers, a "stem-and-leaf plot" may be more convenient. For instance, a stem-and-leaf plot for 11, 12, 23, 23, 23, 23, 33, 45, 69 is given below:

1	1 2
2	3 3 3 3
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, "2 | 3 3 3 3" stands for "23, 23, 23, 23."

96. The coin landed heads 7 times in the first 10 tosses; by coincidence, there were also 7 heads in the next 10 tosses; there were 5 heads in the third batch of 10 tosses; and so forth.

97. In Figure 3, the bin width is 1. There were no 0's or 1's in the data, so the bars over 0 and 1 disappear. There is a bin from 1.5 to 2.5; the four 2's in the data fall into this bin, so the bar over the

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.[98] It may be worth inquiring how the analyst chose the bin widths.

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

Perhaps the most familiar descriptive statistic is the mean (or "arithmetic mean"). The mean can be found by adding up all the numbers and dividing by how many there are. By comparison, the median is defined so that half the numbers are bigger than the median, and half are smaller.[99] Yet a third statistic is the mode, which is the most common number in the data set. These statistics are different, although they are not always clearly distinguished.[100] The mean takes account of all the data—it involves the total of all the numbers; however, particularly with small data sets, a few unusually large or small observations may have too much influence on the mean. The median is resistant to such outliers.

To illustrate the distinction between the mean and the median, consider a report that the "average" award in malpractice cases skyrocketed from $220,000

interval from 1.5 to 2.5 has height four. There is another bin from 2.5 to 3.5, which catches five 3's; the height of the corresponding bar is five. And so forth.

All the bins in Figure 3 have the same width, so this histogram is just like a bar graph. However, data are often published in tables with unequal intervals. The resulting histograms will have unequal bin widths; bar heights should be 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 16, at 31–41.

98. 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. The optimal bin width "depends on the subject matter and the goal of the analysis." Cleveland, *supra* note 93, at 125.

99. Technically, at least half the numbers are at the median or larger; at least half are at the median or smaller. When the distribution is symmetric, the mean equals the median. The values diverge, however, when the distribution is asymmetric, or skewed. The distinction between the mean and the median is critical to the interpretation of the Railroad Revitalization and Regulatory Reform Act, 49 U.S.C. § 11503 (1988), 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, whereas railroads prefer the median. The choice has important financial consequences, and much litigation has resulted. *See* David A. Freedman, *The Mean Versus the Median: A Case Study in 4-R Act Litigation*, 3 J. Bus. & Econ. Stat. 1 (1985).

100. In ordinary language, the arithmetic mean, the median, and the mode seem to be referred to interchangeably as "the average." In statistical parlance, the average is the arithmetic mean. The distinctions are brought out by the following question: How big an error would be made if every number in a batch were replaced by the "center" of the batch? The mode minimizes the number of errors; all errors count the same, no matter what their size. Similar distributions can have very different modes, and the mode is rarely used by statisticians. The median minimizes a different measure of error—the sum of all the differences between the center and the data points; signs are not taken into account when computing this sum, so positive and negative differences are treated the same way. The mean minimizes the sum of the squared differences.

in 1975 to more than $1 million in 1985.[101] The median award almost certainly was far less than $1 million,[102] and the apparently explosive growth may result from a few very large awards. Still, if the issue is whether insurers were experiencing more costs from jury verdicts, the mean is the more appropriate statistic: The total of the awards is directly related to the mean, not to the median.[103]

E. Is an Appropriate Measure of Variability Used?

The location of the center of a batch of numbers reveals nothing about the variations exhibited by these numbers.[104] Statistical measures of variability include the range, the interquartile range, and the standard deviation. The range is the difference between the largest number in the batch and the smallest. The range 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.[105] The interquartile range is the difference between the 25th and 75th percentiles.[106] The interquartile range contains 50% of the numbers and is resistant to changes in extreme values. The standard deviation is a sort of mean deviation from the mean.[107]

101. Kenneth Jost, *Still Warring Over Medical Malpractice: Time for Something Better*, A.B.A. J., May 1993, at 68, 70–71.

102. A study of cases in North Carolina reported an "average" (mean) award of about $368,000, and a median award of only $36,000. *Id.* at 71. In *TXO Production Corp. v. Alliance Resources Corp.*, 509 U.S. 443 (1993), briefs portraying the punitive damage system as 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 mean differs so dramatically from the median because the mean takes into account (indeed, is heavily influenced by) the magnitudes of the few very large awards; the median screens these out. Of course, representative data on verdicts and awards are hard to find. For a study using a probability sample of cases, see Carol J. DeFrances et al., *Civil Jury Cases and Verdicts in Large Counties*, Bureau Just. Stats. Special Rep., July 1995, at 1.

103. To get the total award, just multiply the mean by the number of awards; by contrast, the total cannot be computed from the median. (The more pertinent figure for the insurance industry is not the total of jury awards, but actual claims experience including settlements; of course, even the risk of large punitive damage awards may have considerable impact.) These and related statistical issues are pursued further in, e.g., Theodore Eisenberg & Thomas A. Henderson, Jr., *Inside the Quiet Revoluion in Products Liability*, 39 UCLA L. Rev. 731, 764–72 (1992); Scott Harrington & Robert E. Litan, *Causes of the Liability Insurance Crisis*, 239 Science 737, 740–41 (1988); Saks, *supra* note 81, at 1147, 1248–54.

104. 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.

105. Typically, the range increases with the size of the sample, i.e., the number of units chosen for the sample.

106. By definition, 25% of the data fall below the 25th percentile, 90% fall below the 90th percentile, and so on. The median is the 50th percentile.

107. As discussed in the Appendix, when the distribution follows the normal curve, about 68% of the data will be within one standard deviation of the mean, and about 95% will be within two standard deviations of the mean. For other distributions, the proportions of the data within specified numbers of standard deviations will be different.

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. Beyond this, any of the statistics can be supplemented with a figure that displays much of the data.[108]

IV. What Inferences Can Be Drawn from the Data?

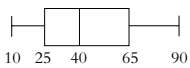
The inferences that may be drawn from a study depend on the quality of the data and the design of the study. As discussed in section II, the data might not address the issue of interest, might be systematically in error, or might be difficult to interpret due to confounding. We turn now to an additional concern—random error.[109] Are patterns in the data the result of chance? Would a pattern wash out if more data were collected?

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

Technically, the standard deviation is the square root of the variance; the variance is the mean square deviation from the mean. For instance, if the mean is 100, the datum 120 deviates from the mean by 20, and the square of 20 is $20^2 = 400$. If the variance (i.e., the mean of all the squared deviations) is 900, then the standard deviation is the square root of 900, that is, $\sqrt{900} = 30$. Among other things, taking the square root corrects for the fact that the variance is on a different scale than the measurements themselves. For example, if the measurements are of length in inches, the variance is in square inches; taking the square root changes back to inches.

To compare distributions on different scales, the coefficient of variation may be used: this statistic is the standard deviation, expressed as a percentage of the mean. For instance, 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 $2.8/5 = 56\%$.

108. For instance, the "five-number summary" lists the smallest value, the 25th percentile, the median, the 75th percentile, and the largest value. The five-number summary may be presented as a box plot. If the five numbers were 10, 25, 40, 65 and 90, the box plot 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 slightly different points in the distribution of numbers.

109. Random error is also called sampling error, chance error, or statistical error. Econometricians use the parallel concept of random disturbance terms.

Rights Act, if the proposed examination excludes a disproportionate number of women, the employer needs to show that the exam is job related.[110]

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 percentage points. The employer argues, however, that the disparity could just reflect random error. After all, only a small number of people took the test, and the sample could have included disproportionate numbers of high-scoring men and low-scoring women. Clearly, even if there were no overall difference in pass rates for male and female applicants, in some samples the men will outscore the women. More generally, a sample is unlikely to be a perfect microcosm of the population; statisticians call differences between the sample and the population, just due to the luck of the draw in choosing the sample, "random error" or "sampling error."

When assessing the impact of random error, a statistician might consider the following topics:

- *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, and there is no disparate impact: in the population of all 5,000 male and 5,000 female applicants, pass rates are equal. 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.[111]
- *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. For a subjectivist statistician, posterior probabilities may be com-

110. 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 11, 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).

111. "Hypothesis testing" is also called "significance testing." For details on the example, see *infra* Appendix, especially note 245.

puted using "Bayes' rule." Within the framework of classical statistical theory, however, such a posterior probability has no meaning.[112]

- *Applicability of statistical models.* Statistical inference—whether done with confidence intervals or significance probabilities, by objective methods or subjective—depends on the validity of statistical models for the data. If the data are collected on the basis of a probability sample or a randomized experiment, there will be statistical models that fit the situation very well, and inferences based on these models will be quite secure. Otherwise, calculations are generally based on analogy: this group of people is like a random sample, that observational study is like a randomized experiment. The fit between the statistical model and the data may then require examination: how good is the analogy?

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, we used the difference in pass rates for a sample of men and women 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 was estimated as 20 percentage points: 58% – 38% = 20 percentage points. 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,[113] leaving room for judgment in selecting an estimator.

2. What Is the Standard Error? The Confidence Interval?

An estimate based on a sample is likely to be off the mark, at least by a little, due to random error. The standard error gives the likely magnitude of this random error.[114] Whenever possible, an estimate should be accompanied by its standard

112. This classical framework is also called "objectivist" or "frequentist," by contrast with the "subjectivist" or "Bayesian" framework. In brief, objectivist statisticians view probabilities as objective properties of the system being studied. Subjectivists view probabilities as measuring subjective degrees of belief. Section IV.B.1 explains why posterior probabilities are excluded from the classical calculus, and section IV.C briefly discusses the subjectivist position. The procedure for computing posterior probabilities is presented *infra* Appendix. For more discussion, see David Freedman, *Some Issues in the Foundation of Statistics*, 1 Found. Sci. 19 (1995), *reprinted in* Topics in the Foundation of Statistics 19 (Bas C. van Fraasen ed., 1997).

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

114. "Standard errors" are also called "standard deviations," and courts seem to prefer the latter term, as do many authors.

error.[115] In our example, the standard error is about 10 percentage points: the estimate of 20 percentage points is likely to be off by something like 10 percentage points or so, in either direction.[116] 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 that population differences fall within plus-or-minus 1 standard error of the sample differences will 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.[117] 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.

The main point is that 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. 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. Confidence intervals are a technical refinement, and

115. The standard error can also be used to measure reproducibility of estimates from one random sample to another. *See infra* note 237.

116. The standard error depends on the pass rates of men and women in the sample, and the size of the sample. With larger samples, chance error will be smaller, so the standard error goes down as sample size goes up. ("Sample size" is the number of subjects in the sample.) 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.

117. Confidence levels are usually read off the normal curve (*see infra* Appendix). Technically, the area under the normal curve between –2 and +2 is closer to 95.4% than 95.0%; thus, statisticians often use ±1.96 SEs for a 95% confidence interval. However, the normal curve only gives an approximation to the relevant chances, and the error in that approximation will often be larger than the difference between 95.4% and 95.0%. For simplicity, we use ±2 SEs for 95% confidence. Likewise, we use ±1 SE for 68% confidence, although the area under the curve between –1 and +1 is closer to 68.3%. The normal curve gives good approximations when the sample size is reasonably large; for small samples, other techniques should be used.

"confidence" is a term of art.[118] 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,[119] but a high confidence level for a small interval is impressive,[120] indicating that the random error in the sample estimate is low.

Standard errors and confidence intervals are derived using statistical models of the process that generated the data.[121] If the data come from a probability

118. In the standard frequentist theory of statistics, one cannot make probability statements about population characteristics. *See, e.g.*, Freedman et al., *supra* note 16, at 383–86; *infra* § IV.B.1. Consequently, it is imprecise to suggest that "[a] 95% confidence interval means that there is a 95% probability that the 'true' relative risk falls within the interval." DeLuca v. Merrell Dow Pharms., Inc., 791 F. Supp. 1042, 1046 (D.N.J. 1992), *aff'd*, 6 F.3d 778 (3d Cir. 1993). 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).

Another misconception is that the confidence level gives the chance that repeated estimates fall into the confidence interval. *E.g.*, Turpin v. Merrell Dow Pharms., Inc., 959 F.2d 1349, 1353 (6th Cir. 1992) ("a confidence interval of '95 percent between 0.8 and 3.10' . . . means that random repetition of the study should produce, 95 percent of the time, a relative risk somewhere between 0.8 and 3.10"); United States *ex rel.* Free v. Peters, 806 F. Supp. 705, 713 n.6 (N.D. Ill. 1992) ("A 99% confidence interval, for instance, is an indication that if we repeated our measurement 100 times under identical conditions, 99 times out of 100 the point estimate derived from the repeated experimentation will fall within the initial interval estimate . . . ."), *rev'd in part*, 12 F.3d 700 (7th Cir. 1993). However, the confidence level does not give the percentage of the time that repeated estimates fall in the interval; instead, it gives the percentage of the time that intervals from repeated samples cover the true value.

119. Statements about the confidence in a sample without any mention of the interval estimate are practically meaningless. In *Hilao v. Estate of Marcos*, 103 F.3d 767 (9th Cir. 1996), for instance, "an expert on statistics . . . testified that . . . a random sample of 137 claims would achieve 'a 95% statistical probability that the same percentage determined to be valid among the examined claims would be applicable to the totality of [9,541 facially valid] claims filed.'" *Id.* at 782. Unfortunately, there is no 95% "statistical probability" that a percentage computed from a sample will be "applicable" to a population. One can compute a confidence interval from a random sample and be 95% confident that the interval covers some parameter. That can be done for a sample of virtually any size, with larger samples giving smaller intervals. What is missing from the opinion is a discussion of the widths of the relevant intervals.

120. Conversely, a broad interval signals that random error is substantial. In *Cimino v. Raymark Industries, 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 pending cases. The court then held a hearing to determine whether the samples were large enough to provide accurate estimates. 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.* at 664. 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. To be correct in a few instances with a 99% confidence interval is not very impressive—by definition, such intervals are broad enough to ensure coverage 99% of the time. *Cf.* Saks & Blanck, *supra* note 54.

121. Generally, statistical models enable the analyst to compute the chances of the various possible outcomes. For instance, the model may contain parameters, that is, numerical constants describing the population from which samples were drawn. *See infra* § V. That is the case for our example, where one

sample or a randomized controlled experiment,[122] the statistical model may be connected tightly to the actual data-collection process. In other situations, using the model may be tantamount to assuming that a sample of convenience is like a random sample, or that an observational study is like a randomized experiment.

Our example was based on a random sample, and that justified the statistical calculations.[123] In many contexts, the choice of an appropriate statistical model is not obvious.[124] When a model does not fit the data-collection process so well,

parameter is the pass rate of the 5,000 male applicants, and another parameter is the pass rate of the 5,000 female applicants. As explained in the Appendix, these parameters can be used to compute the chance of getting any particular sample difference. Using a model with known parameters to find the probability of an observed outcome (or one like it) is common in cases alleging discrimination in the selection of jurors. *E.g.*, Castaneda v. Partida, 430 U.S. 482, 496 (1977); Kaye, *supra* note 86, at 13; *cf.* Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 311 n.17 (1977) (computing probabilities of selecting black teachers). But when the values of the parameters are not known, the statistician must work backwards, using the sample data to estimate the unknown population parameters. That is the kind of statistical inference described in this section.

122. *See supra* § II.A–B.

123. As discussed in the Appendix, large random samples give rise to certain normally distributed statistics. Partly because the Supreme Court used such a model in *Hazelwood* and *Castaneda*, courts and attorneys sometimes are skeptical of 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 see* Branion v. Gramly, 855 F.2d 1256 (7th Cir. 1988) (questioning an apparently arbitrary assumption of normality), *discussed in* David H. Kaye, *Statistics for Lawyers and Law for Statistics*, 89 Mich. L. Rev. 1520 (1991) (defending the use of the normal approximation); Michael O. Finkelstein & Bruce Levin, *Reference Guide on Statistics: Non Lasciare Esperanza*, 36 Jurimetrics J. 201, 205 (1996) (review essay) ("The court was right to reject the normal distribution . . . ."). Whether a given variable is normally distributed is an empirical or statistical question, not a matter of law.

124. *See infra* § V. 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). As one statistician describes the issue:

> [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.

estimates and standard errors will be less probative.[125]

Standard errors and confidence intervals generally ignore systematic errors such as selection bias or non-response bias; in other words, these biases are assumed to be negligible.[126] 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.[127] 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; so does the confidence interval.[128] Likewise, the standard error does not address problems inherent in using convenience samples rather than random samples.[129]

B. Significance Levels 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 big as or bigger 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.

> 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 47, at 107, 117–18.

125. It still may be helpful to consider the standard error, perhaps as a minimal estimate for statistical uncertainty in the quantity being estimated.

126. For a discussion of such systematic errors, see *supra* § II.B.

127. Brock v. Merrell Dow Pharms., Inc., 874 F.2d 307, 311–12 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989).

128. In *Brock*, the court stated 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 situation, the standard error would vanish. Therefore, the standard error could disclose nothing about the impact of selective recall. The same conclusion holds even in the presence of sampling error.

129. *See supra* § II.B.1.

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:

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

In our example, $p = 5\%$. 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.[130] The *p*-value for the observed discrepancy is 5%, or .05.

In such cases, small *p*-values are evidence of disparate impact, while large *p*-values are evidence against disparate impact. Regrettably, multiple negatives are involved here. A statistical test is essentially an argument by contradiction. The "null hypothesis" asserts no difference in the population—that is, no disparate impact. Small *p*-values speak against the null hypothesis—there is disparate impact, because the observed difference is hard to explain by chance alone. Conversely, large *p*-values indicate that the data are compatible with the null hypothesis: the observed difference is easy to explain by chance. In this context, small *p*-values argue for the plaintiffs, while large *p*-values argue for the defense.[131]

Since *p* is calculated by assuming that the null hypothesis is correct (no real difference in pass rates), the *p*-value cannot give the chance that this hypothesis is true. The *p*-value merely gives the chance of getting evidence against the null hypothesis as strong or stronger than the evidence at hand—assuming the null hypothesis to be 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.[132]

130. *See infra* Appendix.

131. Of course, sample size must also be considered, among other factors. *See infra* § IV.C.

132. *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.*, Vasquez v. Hillery, 474 U.S. 254, 259 n.3 (1986) ("the District Court . . . ultimately accepted . . . a probability of 2 in 1,000 that the phenomenon was attributable to chance"); EEOC v. Olson's Dairy Queens, Inc., 989 F.2d 165, 167 (5th Cir. 1993) ("Dr. Straszheim concluded that the likelihood that [the] observed hiring patterns resulted from truly race-neutral hiring practices was less than one chance in ten thousand"); Capaci v. Katz & Besthoff, Inc., 711 F.2d 647, 652 (5th Cir. 1983) ("the highest probability of unbiased hiring was 5.367×10^{-20}"). Such statements 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 1 followed by 20 zeros, and 10^{-20} is the reciprocal of that number. The proverbial "one-in-a-million" is more dryly expressed as 1×10^{-6}.

Computing *p*-values requires statistical expertise. Many methods are available, but only some will fit the occasion. 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.[133]

Because the *p*-value is affected by sample size, it does not measure the extent or importance of a difference.[134] Suppose, for instance, that the 5,000 male and 5,000 female job applicants would differ in their pass rates, but only by a single 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.[135] In short, the *p*-value does not measure the strength or importance of an association.

2. Is a Difference Statistically Significant?

Statistical significance is determined by comparing a *p*-value to a preestablished value, the significance level.[136] 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. On the other hand, if the sample difference is far from the expected value—according to the null hypothesis—then 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, might be considered significant at

133. In this context, a parameter is an unknown numerical constant that is part of the statistical model. *See supra* note 121.

134. 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).

135. *Cf.* Frazier v. Garrison Indep. Sch. Dist., 980 F.2d 1514, 1526 (5th Cir. 1993) (rejecting claims of intentional discrimination in the use of a teacher competency examination that resulted in retention rates exceeding 95% for all groups).

136. Statisticians use the Greek letter alpha (α) to denote the significance level; α gives the chance of getting 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 or a "Type I error" (also called 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.

the .05 level. If the threshold were set lower, say at .01, the result would not be significant.[137]

In practice, statistical analysts often use certain preset significance levels—typically .05 or .01.[138] 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 figure.[139] An unexplained reference to "highly significant" results probably means that p is less than .01.[140]

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. Analysts may refer to a difference as "significant," meaning only that the p-value is below some threshold value. Significance depends not only on the magnitude of the effect, but also on the sample size (among other things). Thus, 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 practical significance is lacking—when the size of a disparity or correlation is negligible—there is no reason to worry about statistical significance.[141]

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.[142]

137. For another example of the relationship between a test statistic and significance, see *infra* § V.D.2.

138. The Supreme Court implicitly referred to this practice in *Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977), and *Hazelwood School District 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."

139. Some have suggested that data not "significant" at the .05 level should be disregarded. *E.g.*, Paul Meier et al., *What Happened in* Hazelwood: *Statistics, Employment Discrimination, and the 80% Rule*, 1984 Am. B. Found. Res. J. 139, 152, *reprinted in* Statistics and the Law, *supra* note 1, at 1, 13. This view is challenged in, e.g., Kaye, *supra* note 118, at 1344 & n.56, 1345.

140. 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 28, at 313, 316.

141. *E.g.*, Waisome v. Port Auth., 948 F.2d 1370, 1376 (2d Cir. 1991) ("though the disparity was found to be statistically significant, it was of limited magnitude"); *cf.* Thornburg v. Gingles, 478 U.S. 30, 53–54 (1986) (repeating the district court's explanation of why "the correlation between the race of the voter and the voter's choice of certain candidates was [not only] statistically significant," but also "so marked as to be substantively significant, in the sense that the results of the individual election would have been different depending upon whether it had been held among only the white voters or only the black voters").

142. *E.g.*, *Waisome*, 948 F.2d 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

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.[143]

C. Evaluating Hypothesis Tests

1. What Is the Power of the Test?

When a *p*-value is high, findings are not significant, and the null hypothesis is not rejected. This could happen for at least two reasons:

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 happened to be of the kind expected under the null hypothesis.

If the "power" of a statistical study is low, the second explanation may be plausible. Power is the chance that a statistical test will declare an effect when there is an effect to declare.[144] This chance depends on the size of the effect and

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. 1984); Sheehan v. Daily Racing Form, Inc., 104 F.3d 940, 941 (7th Cir. 1997) ("An affidavit by a statistician . . . states that the probability that the retentions . . . are uncorrelated with age is less than 5 percent.").

143. For more discussion, see Kaye, *supra* note 118; *cf. infra* note 167.

144. More precisely, power is the probability of rejecting the null hypothesis when the alternative hypothesis is right. (On the meaning of "alternative hypothesis," see *infra* § IV.C.5.) Typically, this probability will depend on the values of unknown parameters, as well as the pre-set significance level α. Therefore, no single number gives the power of the test. One can specify particular values for the parameters and significance level and compute the power of the test accordingly. *See infra* Appendix for an example. Power may be denoted by the Greek letter beta (β).

Accepting the null hypothesis when the alternative is true is known as a "false acceptance" of the null hypothesis or a "Type II error" (also called 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.

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 do not give the probabilities of the null and alternative hypotheses; *see supra* § IV.B.2; *infra* note 167. *See* D.H. Kaye, *Hypothesis Testing in the Courtroom*, *in* Contributions to the Theory and Application of Statistics: A Volume in Honor of Herbert Solomon 331, 341–43 (Alan E. Gelfand ed., 1987); *supra* § IV.B.1; *infra* note 165.

the size of the sample. Discerning subtle differences in the population requires large samples; even so, small samples may detect truly substantial differences.[145]

When a study with low power fails to show a significant effect, the results are more fairly described as inconclusive than as negative: the proof is weak because power is low.[146] On the other hand, when studies have a good chance of detecting a meaningful association, failure to obtain significance can be persuasive evidence that there is no effect to be found.[147]

2. One- or Two-tailed Tests?

In many cases, a statistical test can be done either one-tailed or two-tailed. The second method will produce a *p*-value twice as big as the first method. Since

145. For simplicity, the numerical examples of statistical inference in this reference guide presuppose large samples. Some courts have expressed uneasiness about estimates or analyses based on small samples; indeed, a few courts have refused even to consider such studies or formal statistical procedures for handling small samples. *See, e.g.*, Bunch v. Bullard, 795 F.2d 384, 395 n.12 (5th Cir. 1986) (that 12 of 15 whites and only 3 of 13 blacks passed a police promotion test created a prima facie case of disparate impact; however, "[t]he district court did not perform, nor do we attempt, the application of probability theories to a sample size as small as this" because "[a]dvanced statistical analysis may be of little help in determining the significance of such disparities"); United States v. Lansdowne Swim Club, 713 F. Supp. 785, 809–10 (E.D. Pa. 1989) (collecting cases). Other courts have been more venturesome. *E.g.*, Bazemore v. Friday, 751 F.2d 662, 673 & n.9 (4th Cir. 1984) (court of appeals applied its own *t*-test rather than the normal curve to quartile rankings in an attempt to account for a sample size of nine), *rev'd on other grounds*, 478 U.S. 385 (1986).

Analyzing data from small samples may require more stringent assumptions, but there is no fundamental difference in the meaning of confidence intervals and *p*-values. If the assumptions underlying the statistical analysis are justified—and this can be more difficult to demonstrate with small samples—then confidence intervals and test statistics are no less trustworthy than those for large samples. Aside from the problem of choosing the correct analytical technique, the concern with small samples is not that they are beyond the ken of statistical theory, but that (1) the statistical tests involving small samples might lack power, and (2) the underlying assumptions may be hard to validate.

146. In our 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* 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.

147. 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). 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., 1993); 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 Reseach Council, Combining Information: Statistical Issues and Opportunities for Research (1992); Symposium, *Meta-Analysis of Observational Studies*, 140 Am. J. Epidemiology 771 (1994). Unfortunately, the procedures have their own limitations. *E.g.*, Diana B. Petitti, Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis Methods for Quantitative Synthesis in Medicine (2d ed. 2000); Michael Oakes, Statistical Inference: A Commentary for the Social and Behavioural Sciences 157 (1986) ("a retrograde development"); John C. Bailar III, *The Promise and Problems of Meta-Analysis*, 337 New Eng. J. Med. 559 (1997) (editorial); Charles Mann, *Meta-Analysis in the Breech*, 249 Science 476 (1990).

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.[148]

Some courts have expressed a preference for two-tailed tests,[149] but a rigid rule is not required if *p*-values and significance levels are used as clues rather than as mechanical rules for statistical proof. 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.[150]

3. How Many Tests Have Been Performed?

Repeated testing complicates the interpretation of significance levels. If enough comparisons are made, random error almost guarantees that some will yield "significant" findings, even when there is no real effect. 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—can be repeated far too often.

148. 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 $p = .025$ (approximately). 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 $p = .05$. *See infra* Appendix (figure 13); Freedman et al., *supra* note 16, at 549–52.

According to formal statistical theory, the choice between one tail or two can sometimes be made by considering the exact form of the "alternative hypothesis." *See infra* § IV.C.5. In our example, the null hypothesis is that pass rates are 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. On the other hand, 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 16, at 551. Some experts think that the choice between one-tailed and two-tailed tests can often be made by considering the exact form of the null and alternative hypothesis.

149. *See, e.g.*, Baldus & Cole, *supra* note 89, § 9.1, at 308 n.35a; The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 38–40 (citing EEOC v. Federal Reserve 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 118, 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)). Arguments for one-tailed tests are discussed in Finkelstein & Levin, *supra* note 1, at 125–26; Richard Goldstein, *Two Types of Statistical Errors in Employment Discrimination Cases*, 26 Jurimetrics J. 32 (1985); Kaye, *supra* at 841.

150. One-tailed tests at the .05 level are viewed as weak evidence—no weaker standard is commonly used in the technical literature.

Such artifacts are commonplace. 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.[151] Even a single researcher may search for so many different relationships that a few will achieve statistical significance by mere happenstance. 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 somewhere along the way in a few thousand tosses of a coin means quite another.

There are statistical methods for coping with multiple looks at the data, which permit the calculation of meaningful *p*-values in certain cases.[152] However, no general solution is available, and the existing methods would be of little help in the typical case where analysts have tested and rejected a variety of regression models before arriving at the one considered the most satisfactory. In these situations, courts should not be overly impressed with claims that estimates are significant. Instead, they should be asking how analysts developed their models.[153]

4. Tests or Interval Estimates?

Statistical significance depends on the *p*-value, and *p*-values depend on sample size. Therefore, a "significant" effect could be small. Conversely, an effect that is "not significant" could be large.[154] 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—interval estimates may be valuable.[155] Seeing a plausible range of values for the quantity of interest helps describe 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

151. *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).

152. *See, e.g.*, Rupert G. Miller, Jr., Simultaneous Statistical Inference (2d ed. 1981).

153. *See, e.g.*, On Model Uncertainty and Its Statistical Implications: Lecture Notes in Econometric and Mathematical Systems (Theo K. Dijkstra ed., 1988); Frank T. Denton, *Data Mining As an Industry*, 67 Rev. Econ. & Stat. 124 (1985). 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.

154. *See supra* § IV.B.1.

155. 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 point estimate and standard error can be combined in a confidence interval.

estimate is that the pass rate for men is 20 percentage points higher than 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.[156] For instance, significance at the .05 level can be read off the 95% confidence interval.[157] In our example, zero is at the extreme edge of the 95% confidence interval, so we have "significant" evidence that the true difference in pass rates between male and female applicants is not zero. But there are values very close to zero inside the interval.

On the other hand, 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: say that 29 men and 20 women passed the test. The 95% confidence interval goes from -2 to 38 percentage points. Because a difference of zero falls within the 95% confidence interval, the null hypothesis—that the true difference is zero—cannot be rejected at the .05 level. But the interval extends to 38 percentage points, indicating that the population difference could be substantial. Lack of significance does not exclude this possibility.[158]

5. What Are the Rival Hypotheses?

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.[159] A small *p*-value means something is going on, besides random error; the alternative hypothesis should be viewed as one possible explanation—out of many—for the data.[160]

156. 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 118, at 1349 n.78; *cf.* Bailar & Mosteller, *supra* note 140, at 317.

157. Instead of referring to significance at the .05 level, some writers refer to "the 95 percent confidence level that is often used by scientists to reject the possibility that chance alone accounted for observed differences." Carnegie Comm'n on Science, Tech. & Gov't, Science and Technology in Judicial Decision Making: Creating Opportunities and Meeting Challenges 28 (1993).

158. We have used two-sided intervals, corresponding to two-tailed tests. One-sided intervals, corresponding to one-tailed tests, also are available.

159. Often, the null and alternative hypotheses are statements about possible ranges of values for parameters in a common statistical model. *See, e.g.*, *supra* note 148. 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.

160. *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 16, at 562–63; C.E.M. Hansel, ESP: A Scientific Evaluation (1966).

In *Mapes Casino, Inc. v. Maryland Casualty Co.*,[161] 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 18-month period, the win percentage at its craps tables was 6%, compared to 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 big to explain as the mere product of random chance. But it was not convinced by plaintiff's alternative hypothesis. The court pointed to other possible explanations (Runyonesque activities like "skimming," "scamming," and "crossroading") that might have accounted for the discrepancy without implicating the suspect employees.[162] In short, rejection of the null hypothesis does not leave the proffered alternative hypothesis as the only viable explanation for the data.[163]

In many studies, the validity of the model is secured by the procedures used to collect the data. There are formulas for standard errors and confidence intervals that hold when random samples are used. *See supra* §§ II.B, IV.A.2. There are statistical tests for comparing two random samples, or evaluating the results of a randomized experiment. *See supra* §§ II.A, IV.B.2. In such examples, the statistical procedures flow from the sampling method and the design of the study. On the other hand, if samples of convenience are used, or subjects are not randomized, the validity of the statistical procedures can be contested. *See* Freedman et al., *supra* note 16, at 387–88, 424, 557–65.

161. 290 F. Supp. 186 (D. Nev. 1968).

162. *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 out such possibilities.

163. *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., 885 F.2d 575 (9th Cir. 1989) (unsubstantiated rival hypothesis of "lack of interest" in "non-traditional" jobs insufficient to rebut prima facie case of gender discrimination); *cf. supra* § II.A (problem of confounding); *infra* note 230 (effect of omitting important variables from a regression model).

D. Posterior Probabilities

Standard errors, *p*-values, and significance tests are common techniques for assessing random error. These procedures rely on the sample data, and are justified in terms of the "operating characteristics" of the statistical procedures.[164] However, this frequentist approach does not permit the statistician to compute the probability that a particular hypothesis is correct, given the data.[165] For instance, a frequentist may postulate that a coin is fair: it has a 50-50 chance of landing heads, and successive tosses are independent; this is viewed as an empirical statement—potentially falsifiable—about the coin. On this basis, it is easy to calculate the chance that the coin will turn up heads in the next ten tosses:[166] the answer is 1/1,024. Therefore, observing ten heads in a row brings into serious question the initial hypothesis of fairness. Rejecting the hypothesis of fairness when there are ten heads in ten tosses gives the wrong result—when the coin is fair—only one time in 1,024. That is an example of an operating characteristic of a statistical procedure.

But what of the converse probability: if a coin lands heads ten times in a row, what is the chance that it is fair?[167] To compute such converse probabilities, it is necessary to postulate initial probabilities that the coin is fair, as well as probabilities of unfairness to various degrees.[168] And that is beyond the scope of frequentist statistics.[169]

164. "Operating characteristics" are the expected value and standard error of estimators, probabilities of error for statistical tests, and related quantities.

165. *See supra* § IV.B.1; *infra* Appendix. Consequently, quantities such as *p*-values or confidence levels cannot be compared directly to numbers like .95 or .50 that might be thought to quantify the burden of persuasion in criminal or civil cases. *See* Kaye, *supra* note 144; D.H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 Cornell L. Rev. 54 (1987).

166. Stated slightly more formally, if the coin is fair and each outcome is independent (the hypothesis), then the probability of observing ten heads (the data) is $\Pr(\text{data} \mid H_0) = (1/2)^{10} = 1/1{,}024$, where H_0 stands for the hypothesis that the coin is fair.

167. We call this a "converse probability" because it is of the form $\Pr(H_0 \mid \text{data})$ rather than $\Pr(\text{data} \mid H_0)$; an equivalent phrase, "inverse probability," also is used. The tendency to think of $\Pr(\text{data} \mid H_0)$ as if it were the converse probability $\Pr(H_0 \mid \text{data})$ is the "transposition fallacy." For instance, most United States senators are men, but very few men are senators. Consequently, there is a high probability that an individual who is a senator is a man, but the probability that an individual who is a man is a senator is practically zero. For examples of the transposition fallacy in court opinions, see cases cited *supra* note 142. *See also* Committee on DNA Forensic Science: An Update, *supra* note 60, at 133 (describing the fallacy in cases involving DNA identification evidence as the "prosecutor's fallacy"). The frequentist *p*-value, $\Pr(\text{data} \mid H_0)$, is generally not a good approximation to the Bayesian $\Pr(H_0 \mid \text{data})$; the latter includes considerations of power and base rates.

168. *See infra* Appendix.

169. In some situations, the probability of an event on which a case depends can be computed with objective methods. However, these events are measurable outcomes (like the number of heads in a series of tosses of a coin) rather than hypotheses about the process that generated the data (like the claim that the coin is fair). For example, in *United States v. Shonubi*, 895 F. Supp. 460 (E.D.N.Y. 1995), *rev'd*,

In the Bayesian or subjectivist approach, probabilities represent subjective degrees of belief rather than objective facts. The observer's confidence in the hypothesis that a coin is fair, for example, is expressed as a number between zero and one;[170] likewise, the observer must quantify beliefs about the chance that the coin is unfair to various degrees—all in advance of seeing the data.[171] These subjective probabilities, like the probabilities governing the tosses of the coin, are set up to obey the axioms of probability theory. The probabilities for the various hypotheses about the coin, specified before data collection, are called prior probabilities.

These prior probabilities can then be updated, using "Bayes' rule," given data on how the coin actually falls.[172] In short, Bayesian statisticians can compute posterior probabilities for various hypotheses about the coin, given the data.[173] Although such posterior probabilities can pertain directly to hypotheses of legal interest, they are necessarily subjective, for they reflect not just the data but also

103 F.3d 1085 (2d Cir. 1997), a government expert estimated for sentencing purposes the total quantity of heroin that a Nigerian defendant living in New Jersey had smuggled (by swallowing heroin-filled balloons) in the course of eight trips to and from Nigeria. He applied a method known as "resampling" or "bootstrapping." Specifically, he drew 100,000 independent simple random samples of size seven from a population of weights distributed as in customs data on 117 other balloon swallowers caught in the same airport during the same time period; he discovered that for 99% of these samples, the total weight was at least 2090.2 grams. 895 F. Supp. at 504. Thus, the researcher reported that "there is a 99% chance that Shonubi carried at least 2090.2 grams of heroin on the seven [prior] trips . . . ." *Id.* However, the Second Circuit reversed this finding for want of "specific evidence of what Shonubi had done." 103 F.3d at 1090. Although the logical basis for this "specific evidence" requirement is unclear, a difficulty with the expert's analysis is apparent. Statistical inference generally involves an extrapolation from the units sampled to the population of all units. Thus, the sample needs to be representative. In *Shonubi*, the government used a sample of weights, one for each courier on the trip at which that courier was caught. It sought to extrapolate from these data to many trips taken by a single courier—trips on which that other courier was not caught.

170. Here "confidence" has the meaning ordinarily ascribed to it rather than the technical interpretation applicable to a frequentist "confidence interval." Consequently, it can be related to the burden of persuasion. *See* Kaye, *supra* note 165.

171. For instance, let p be the unknown probability that coin lands heads: What is the chance that p exceeds .6? The Bayesian statistician must be prepared to answer all such questions. 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).

172. *See infra* Appendix.

173. *See generally* George E.P. Box & George C. Tiao, Bayesian Inference in Statistical Analysis (Wiley Classics Library ed., John Wiley & Sons, Inc. 1992) (1973). For applications to legal issues, see, e.g., Aitken et al., *supra* note 45, at 337–48; David H. Kaye, *DNA Evidence: Probability, Population Genetics, and the Courts*, 7 Harv. J.L. & Tech. 101 (1993).

the subjective prior probabilities—that is, the degrees of belief about the various hypotheses concerning the coin specified prior to obtaining the data.[174]

Such analyses have rarely been used in court,[175] and the question of their forensic value has been aired primarily in the academic literature.[176] Some statisticians favor Bayesian methods,[177] and some legal commentators have proposed their use in certain kinds of cases in certain circumstances.[178]

V. Correlation and Regression

Regression models are often used to infer causation from association; for example, such models are frequently introduced to prove disparate treatment in discrimination cases, or to estimate damages in antitrust actions. Section V.D explains the ideas and some of the pitfalls. Sections V.A–C cover some preliminary material, showing how scatter diagrams, correlation coefficients, and regression lines can be used to summarize relationships between variables.

174. In this framework, the question arises of whose beliefs to use—the statistician's or the factfinder's. *See*, *e.g.*, Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 Harv. L. Rev. 489 (1970) (proposing that experts give posterior probabilities for a wide range of prior probabilities, to allow jurors to use their own prior probabilities or just to judge the impact of the data on possible values of the prior probabilities). *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. The exception is paternity litigation; when genetic tests are indicative of paternity, testimony as to a posterior "probability of paternity" is common. *See*, *e.g.*, 1 Modern Scientific Evidence: The Law and Science of Expert Testimony, *supra* note 3, § 19–2.5.

176. *See*, *e.g.*, Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism (Peter Tillers & Eric D. Green eds., 1988); Symposium, *Decision and Inference in Litigation*, 13 Cardozo L. Rev. 253 (1991). The Bayesian framework probably has received more acceptance in explicating legal concepts such as the relevance of evidence, the nature of prejudicial evidence, probative value, and burdens of persuasion. *See*, *e.g.*, Richard D. Friedman, *Assessing Evidence*, 94 Mich. L. Rev. 1810 (1996) (book review); Richard O. Lempert, *Modeling Relevance*, 75 Mich. L. Rev. 1021 (1977); D.H. Kaye, *Clarifying the Burden of Persuasion: What Bayesian Decision Rules Do and Do Not Do*, 3 Int'l J. Evidence & Proof 1 (1999).

177. *E.g.*, Donald A. Berry, *Inferences Using DNA Profiling in Forensic Identification and Paternity Cases*, 6 Stat. Sci. 175, 180 (1991); 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). 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. *See*, *e.g.*, Freedman, *supra* note 112.

178. *E.g.,* Joseph C. Bright, Jr. et al., *Statistical Sampling in Tax Audits*, 13 L. & Soc. Inquiry 305 (1988); Ira Mark Ellman & David Kaye, *Probabilities and Proof: Can HLA and Blood Group Testing Prove Paternity?*, 54 N.Y.U. L. Rev. 1131 (1979); Finkelstein & Fairley, *supra* note 174; Kaye, *supra* note 173.

A. Scatter Diagrams

The relationship between two variables can be graphed in a scatter diagram.[179] Data on income and education for a sample of 350 men, ages 25 to 29, residing in Texas[180] provide an example. Each person in the sample corresponds to one dot in the diagram. As indicated in Figure 4, the horizontal axis shows the 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. 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.

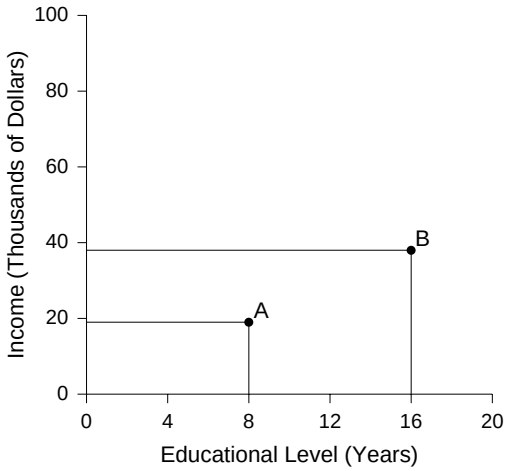
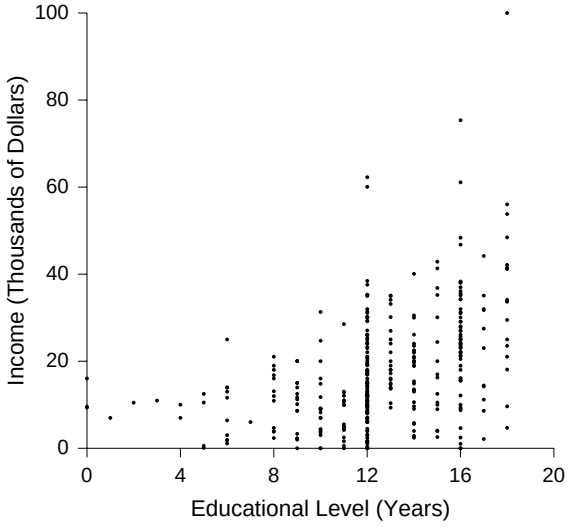


Figure 5 is the scatter diagram for the Texas data. The diagram confirms an obvious point. There is a "positive association" between income and education: in general, persons with a higher educational level have higher incomes. However, there are many exceptions to this rule, and the association is not as strong as one might expect.

179. These diagrams are also referred to as scatterplots or scattergrams.

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

Figure 5. Scatter diagram for income and education: men age 25 to 29 in Texas.[181]



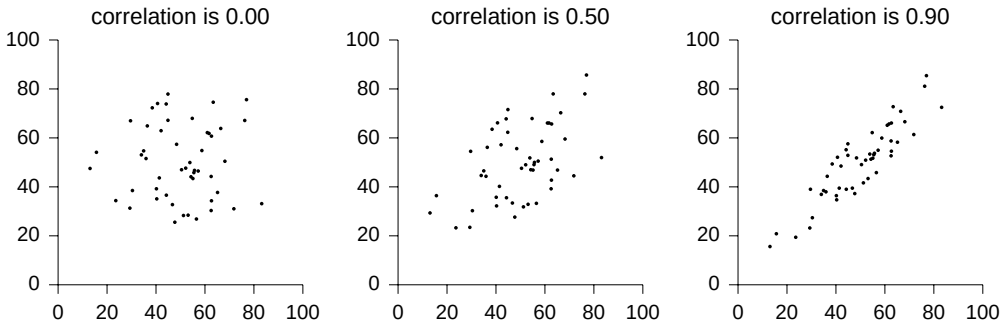
B. Correlation Coefficients

Two variables are positively correlated when their values tend to go up or down together.[182] Income and education in Figure 5 provides an example. The correlation coefficient (usually denoted by the letter *r*) is a single number that reflects the strength of an association. Figure 6 shows the values of *r* for three scatter diagrams.

181. 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.

182. Many statistics and displays are available to investigate association. The most common are the correlation coefficient and the scatter diagram.

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



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

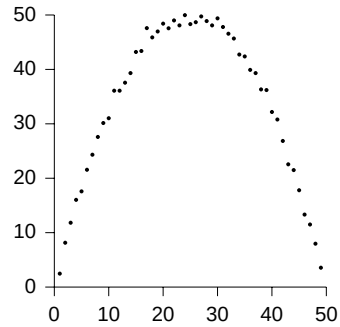
Moderate associations are the general rule in the social sciences; correlations larger than, say, 0.7 are quite unusual in many fields. For example, the correlation between college grades and first-year law school grades is under 0.3 at most law schools, while the correlation between LSAT scores and first-year law grades is generally about 0.4.[183] 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 was 0.43. The correlation coefficient cannot capture all the underlying information. Several issues may arise in this regard, and we consider them in turn.

183. Linda F. Wightman, Predictive Validity of the LSAT: A National Summary of the 1990–1992 Correlation Studies 10 (1993); *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.

1. 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 zero. When the scatter diagram reveals a strong nonlinear pattern, the correlation coefficient may not be a useful summary statistic.

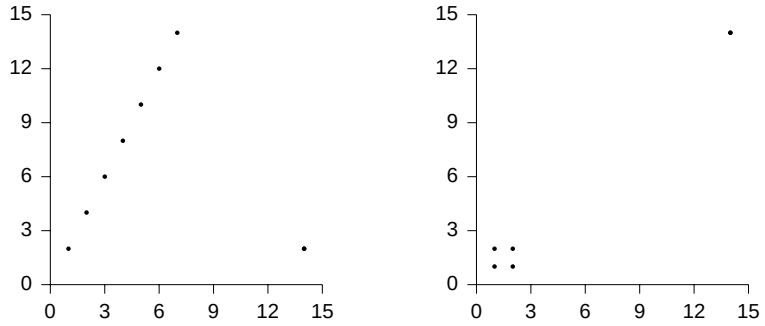
Figure 7. The correlation coefficient only measures linear association. The scatter diagram shows a strong nonlinear association with a correlation coefficient close to zero.



2. Do Outliers Influence the Correlation 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 nothing. Conversely, the right hand panel shows that one outlier (upper right hand corner) can raise a correlation of zero to nearly one.

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 zero to nearly one.



3. Does a Confounding Variable Influence the Coefficient?

The correlation coefficient measures the association between two variables. Investigators—and the courts—are usually more interested in causation. Association is not necessarily the same as causation. As noted in section II.A, the association between two variables may be driven largely by a "third variable" that has been omitted from the analysis. For an easy example, among school children, there is an association between shoe size and vocabulary. However, learning more words does not cause feet to get 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 harder to identify.

Technically, third variables are called confounders or confounding variables.[184] The basic methods of dealing with confounding variables involve controlled experiments[185] or the application, typically though a technique called "multiple regression,"[186] of "statistical controls."[187] In many examples, association really does reflect causation, but a large correlation coefficient is not enough to warrant such a conclusion. A large value of *r* only means that the dependent variable

184. *See supra* § II.A.1.

185. *See supra* § II.A.2.

186. Multiple regression analysis is discussed *infra* § V.D and again in Daniel L. Rubinfeld, Reference Guide on Multiple Regression, § II, in this manual.

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

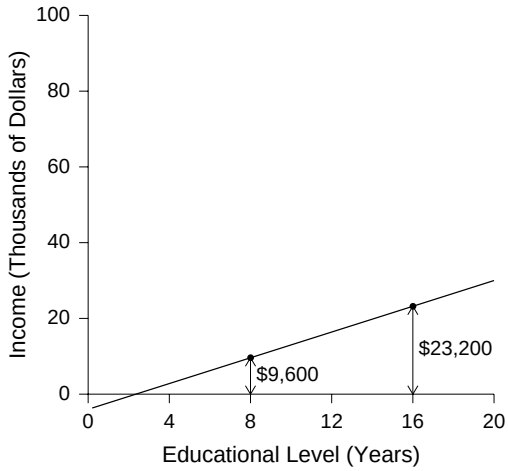
marches in step with the independent one—for any number of possible reasons, ranging from causation to confounding.[188]

C. Regression Lines

The regression line can be used to describe a linear trend in the data. The regression line for income on education in the Texas sample 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 eight years of education is estimated at $9,600, indicated by the height of the line at eight years; the average income for people with sixteen years of education is estimated at about $23,200.

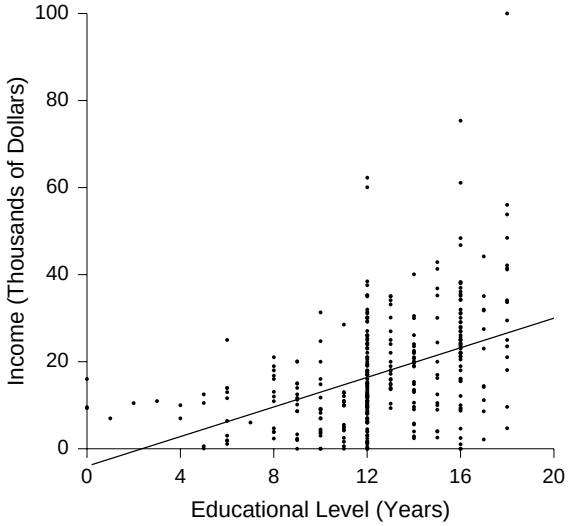
Figure 10 repeats the scatter diagram for income and education (see Figure 5); the regression line is plotted too. In a general way, the line shows the average trend of income as education increases. Thus, the regression line indicates the extent to which a change in one variable (income) is associated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

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



188. The square of the correlation coefficient, r^2, is sometimes called the proportion of variance "explained." However, "explained" is meant in a purely technical sense, and large values of r^2 need not point to a causal explanation.

Figure 10. Scatter diagram for income and education, with the regression line indicating the trend.



1. What Are the Slope and Intercept?

The regression line can be described in terms of its slope and intercept.[189] 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. The intercept is –$4,000. This is an estimate of the average income for persons with zero years of education. The estimate is not a good one, for such persons are far from the center of the diagram. In general, estimates based on the regression line become less trustworthy as we move away from the bulk of the data.

The slope has the same limitations as the correlation coefficient in measuring the degree of association:[190] (1) It only measures linear relationships; (2) it may

189. 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 zero. The slope of a line is akin to the grade of a road; the intercept gives the starting elevation. In 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 per year) × 16 years – $4,000 = $27,200 – $4,000 = $23,200

190. In fact, the correlation coefficient is the slope of the regression line if the variables are "standardized," that is, measured in terms of standard deviations away from the mean.

be influenced by outliers; and (3) it does not control for the effect of other variables. With respect to (1), the slope of $1,700 per year presents each additional year of education as having the same value, but some years of schooling surely are worth more and others less. With respect to (3), the association between education and income graphed in Figure 10 is partly causal, but there are other factors to consider, including the family backgrounds of the people in the sample. For instance, people with college degrees probably come from richer and better educated families than those who drop out after grade school. College graduates have other advantages besides the extra education. Factors like these must have some effect on income. That is why statisticians use the qualified language of "on average" and "associated with."[191]

2. What Is the Unit of Analysis?

If association between the characteristics of individuals is of interest, these characteristics should be measured on individuals. Sometimes the individual data are not available, but rates or averages are; correlations computed from rates or averages are termed "ecological." However, ecological correlations generally overstate the strength of an association. An example makes the point. The average income and average education can be determined for the men living in each state. The correlation coefficient for these 50 pairs of averages turns out to be 0.66. However, states do not go to school and do not earn incomes. People do. The correlation for income and education for all men in the United States is only about 0.44.[192] The correlation for state averages overstates the correlation for individuals—a common tendency for such ecological correlations.[193]

Ecological correlations are often used in cases claiming a dilution in the voting strength of a racial minority. In this type of voting rights case plaintiffs must prove three things: (1) the minority group constitutes a majority in at least one district of a proposed plan; (2) the minority group is politically cohesive, that is, votes fairly solidly for its preferred candidate; and (3) the majority group votes sufficiently as a bloc to defeat the minority-preferred candidate.[194] The first test is called compactness. The second and third tests deal with racially polarized voting.

191. 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 may run into problems. *See generally supra* § II.A, *infra* § V.D.

192. Correlations are computed from a public-use data tape, Bureau of the Census, Dep't of Commerce, for the March 1993 Current Population Survey.

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

194. *See* Thornburg v. Gingles, 478 U.S. 30, 50–51 (1986) ("First, the minority group must be able to demonstrate that it is sufficiently large and geographically compact to constitute a majority in a single-member district. . . . Second, the minority group must be able to show that it is politically

Of course, the secrecy of the ballot box means that racially polarized voting cannot be directly observed.[195] Instead, plaintiffs in these voting rights cases rely on scatter diagrams and regression lines to estimate voting behavior by racial or ethnic groups. The unit of analysis is typically the precinct; hence, the technique is called "ecological regression." For each precinct, public records may suffice to determine the percentage of registrants in each racial or ethnic group, as well as the percentage of the total vote for each candidate—by voters from all demographic groups combined. The statistical issue, then, is to estimate how each demographic subgroup voted.

Figure 11 provides an example. Each point in the scatter diagram 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.[196] The regression line is plotted too. 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.[197] However, the validity of such estimates is contested in statistical literature.[198]

cohesive. . . . Third, the minority must be able to demonstrate that the white majority votes sufficiently as a bloc to enable it . . . usually to defeat the minority's preferred candidate."). In subsequent cases, the Court has emphasized that these factors are not sufficient to make out a violation of section 2 of the Voting Rights Act. *E.g.*, Johnson v. De Grandy, 512 U.S. 997, 1011 (1994) ("*Gingles* . . . clearly declined to hold [these factors] sufficient in combination, either in the sense that a court's examination of relevant circumstances was complete once the three factors were found to exist, or in the sense that the three in combination necessarily and in all circumstances demonstrated dilution.").

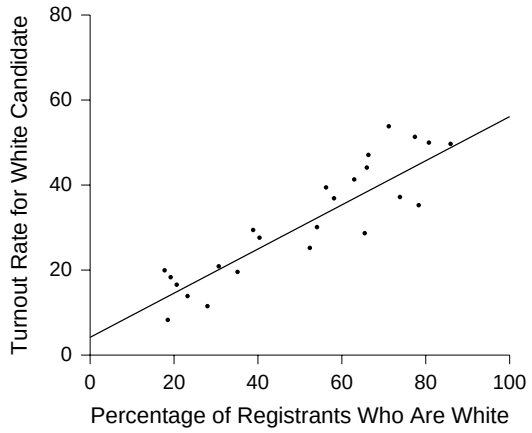
195. Some information could be obtained from exit polls. *E.g.*, Aldasoro v. Kennerson, 922 F. Supp. 339, 344 (S.D. Cal. 1995).

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

197. Figure 11 contemplates only one white candidate; more complicated techniques could be used if there were several candidates of each race. The intercept of the line is 4% and the slope is .52. Plaintiffs would conclude that only 4% of the black registrants voted for the white candidate, while 4% + 52% = 56% of the white registrants voted for the white candidate, which demonstrates polarization.

198. For further discussion of the problem of ecological regression in this context, see Stephen P. Klein & David A. Freedman, *Ecological Regression in Voting Rights Cases*, Chance, Summer 1993, at 38; Bernard Grofman & Chandler Davidson, Controversies in Minority Voting: The Voting Rights Act in Perspective (1992). The use of ecological regression increased considerably after the Supreme Court noted in *Thornburg v. Gingles*, 478 U.S. 30, 53 n.20 (1986), that "[t]he District Court found both methods [extreme case analysis and bivariate ecological regression analysis] standard in the literature for the analysis of racially polarized voting." *See*, *e.g.*, Teague v. Attala County, 92 F.3d 283, 285 (5th Cir. 1996) (one of "two standard methods for analyzing electoral data"); Houston v. Lafayette County, 56 F.3d 606, 612 (5th Cir. 1995) (holding that district court erred in ignoring ecological regression results). Nevertheless, courts have cautioned against "overreliance on bivariate ecological regression" in light of the inherent limitations of the technique (Lewis v. Alamance County, 99 F.3d 600, 604 n.3 (4th Cir. 1996)), and some courts have found ecological regressions unconvincing. *E.g.*, Aldasoro v. Kennerson,

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.[199]



D. Statistical Models

Statistical models are widely used in the social sciences and in litigation.[200] For example, the census suffers an undercount, more severe in certain places than others; if some statistical models are to be believed, the undercount can be corrected—moving seats in Congress and millions of dollars a year in entitlement funds.[201] Other models purport to lift the veil of secrecy from the ballot

922 F. Supp. 339 (S.D. Cal. 1995); Romero v. City of Pomona, 665 F. Supp. 853, 860 (C.D. Cal. 1987), *aff'd*, 883 F.2d 1418 (9th Cir. 1989); *cf.* Johnson v. Miller, 864 F. Supp. 1354, 1390 (S.D. Ga. 1994) ("mind-numbing and contradictory statistical data," including bivariate ecological regression, established "that some degree of vote polarization exists, but not in alarming quantities. Exact levels are unknowable."), *aff'd,* 515 U.S. 900 (1995).

Redistricting plans based predominantly on racial considerations are unconstitutional unless narrowly tailored to meet a compelling state interest. Shaw v. Reno, 509 U.S. 630 (1993). Whether compliance with the Voting Rights Act can be considered a compelling interest is an open question, but efforts to sustain racially motived redistricting on this basis have not fared well before the Supreme Court. *See* Abrams v. Johnson, 521 U.S. 74 (1997); Shaw v. Hunt, 517 U.S. 899 (1996); Bush v. Vera, 517 U.S. 952 (1996).

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

200. The frequency with which regression models are used is no guarantee that they are the best choice for a particular problem. *See*, *e.g.*, David W. Peterson, *Reference Guide on Multiple Regression*, 36 Jurimetrics J. 213, 214–15 (1996) (review essay). On the factors that might justify the choice of a particular model, see Moses, *supra* note 124.

201. *See supra* note 43.

box, enabling the experts to determine how racial or ethnic groups have voted—a crucial step in litigation to enforce minority voting rights.[202] This section discusses the statistical logic of regression models.[203]

A regression model attempts to combine the values of certain variables (the independent variables) in order to get expected values for another variable (the dependent variable). The model can be expressed in the form of a regression equation. A simple regression equation has only one independent variable; a multiple regression equation has several independent variables. Coefficients in the equation will often be interpreted as showing the effects of changing the corresponding variables. Sometimes, this interpretation can be justified. For instance, Hooke's law describes how a spring stretches in response to the load hung from it: strain is proportional to stress.[204] There will be a number of observations on a spring. For each observation, the physicist hangs a weight on the spring, and measures its length. A statistician could apply a regression model to these data: for quite a large range of weights,[205]

$$\text{length} = a + b \times \text{weight} + \varepsilon. \qquad (1)$$

The error term, denoted by the Greek letter epsilon (ε), is needed because measured length will not be exactly equal to $a + b \times$ weight. If nothing else, measurement error must be reckoned with. We model ε as a draw made at random with replacement from a box of tickets. Each ticket shows a potential error, which will be realized if that ticket is drawn. The average of all the potential errors in the box is assumed to be zero. In more standard statistical terminology, the εs for different observations are assumed to be "independent and identically distributed, with mean zero."[206]

In equation (1), a and b are parameters, unknown constants of nature that characterize the spring: a is the length of the spring under no load, and b is elasticity, the increase in length per unit increase in weight.[207] These parameters

202. *See supra* § V.C.2.

203. For a more detailed treatment, see Daniel L. Rubinfeld, Reference Guide on Multiple Regression at app., in this manual.

204. This law is named after Robert Hooke (England, 1653–1703).

205. The dependent or response variable in equation (1) is the length of the spring, on the left hand side of the equation. There is one independent or explanatory variable on the right hand side—weight. Since there is only one explanatory variable, equation (1) is a simple regression equation.

Hooke's law is only an approximation, although it is a very good one. With large enough weights, a quadratic term will be needed in equation (1). Moreover, beyond some point, the spring exceeds its elastic limit and snaps.

206. For some purposes, it is also necessary to assume that the errors follow the normal distribution.

207. *Cf. supra* note 121 (defining the term "parameter").

are not observable,[208] but they can be estimated by "the method of least squares."[209] In statistical notation, estimates are often denoted by hats; thus, $\hat{a}$ is the estimate for a, and $\hat{b}$ is the estimate for b.[210] Basically, the values of $\hat{a}$ and $\hat{b}$ are chosen to minimize the sum of the squared "prediction errors."[211] These errors are also called "residuals": they measure the difference between the actual length and the predicted length, the latter being $\hat{a} + \hat{b} \times$ weight.[212]

$$\text{residual} = \text{actual length} - \hat{a} - \hat{b} \times \text{weight} \tag{2}$$

Of course, no one really imagines there to be a box of tickets hidden in the spring. However, the variability of physical measurements (under many but by no means all circumstances) does seem to be remarkably like the variability in draws from a box.[213] In short, the statistical model corresponds rather closely to the empirical phenomenon.

1. A Social Science Example

We turn now to social science applications of the kind that might be seen in litigation. A case study would take us too far afield, but a stylized example of regression analysis used to demonstrate sex discrimination in salaries may give the idea.[214] We use a regression model to predict salaries (dollars per year) of employees in a firm using three explanatory variables: education (years of schooling completed), experience (years with the firm), and a dummy variable for

208. It might seem that a is observable; after all, one can measure the length of the spring with no load. However, the measurement is subject to error, so one observes not a but $a + \varepsilon$. *See* equation (1). The parameters a and b can be estimated, even estimated very well, but they cannot be observed directly.

209. The method was developed by Adrien-Marie Legendre (France, 1752–1833) and Carl Friedrich Gauss (Germany, 1777–1855) to fit astronomical orbits.

210. Another convention is use Greek letters for the parameters and English letters for the estimates.

211. Given trial values for a and b, one computes residuals as in equation (2), and then the sum of the squares of these residuals. The "least squares" estimates $\hat{a}$ and $\hat{b}$ are the values of a and b that minimize this sum of squares. These least squares values can be computed from the data by a mathematical formula. They are the intercept and slope of the regression line. *See supra* § V.C.1; Freedman et al., *supra* note 16, at 208–10.

212. The residual is observable, but because the estimates $\hat{a}$ and $\hat{b}$ are only approximations to the parameters a and b, the residual is only an approximation to the error term in equation (1). The term "predicted value" is used in a specialized sense, because the actual values are available too; statisticians often refer to "fitted value" rather than "predicted value," to avoid possible misinterpretations.

213. This is Gauss's model for measurement error. *See* Freedman et al., *supra* note 16, at 450–52.

214. For a more extended treatment of the concepts, see Daniel L. Rubinfeld, Reference Guide on Multiple Regression, at app., in this manual.

gender, taking the value 1 for men and 0 for women.[215] The equation is[216]

$$\text{salary} = a + b \times \text{education} + c \times \text{experience} + d \times \text{gender} + \varepsilon \qquad (3)$$

Equation (3) is a statistical model for the data, with unknown parameters *a, b, c,* and *d*; here, *a* is the intercept and the others are regression coefficients; ε is an unobservable error term. This is a formal analog of Hooke's law, shown as equation (1); the same assumptions are made about the errors. In other words, an employee's salary is determined as if by computing

$$a + b \times \text{education} + c \times \text{experience} + d \times \text{gender} \qquad (4)$$

then adding an error drawn at random from a box of tickets. The expression (4) is the expected value for salary given the explanatory variables (education, experience, gender); the error term in equation (3) represents deviations from the expected.

The parameters in equation (3) are estimated from the data using least squares. If the estimated coefficient for the dummy variable turns out to be positive and statistically significant (by a *t*-test[217]), 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 are entered into equation (3) as statistical controls, precisely in order to claim that adjustment has been made for differences in backgrounds.

Suppose the estimated equation turns out as follows:

$$\text{predicted salary} = \$7{,}100 + \$1{,}300 \times \text{education} + \$2{,}200 \times \text{experience} + \$700 \times \text{gender} \qquad (5)$$

That is, $\hat{a} = \$7{,}100$, $\hat{b} = \$1{,}300$, and so forth. According to equation (5), 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, the company gives men a salary premium of $700 over women with the same education and expe-

215. A dummy variable takes only two values (e.g., 0 and 1) and serves to identify two mutually exclusive and exhaustive categories.

216. In equation (3), the variable on the left hand side, salary, is the response variable. On the right hand side are the explanatory variables—education, experience, and the dummy variable for gender. Because there are several explanatory variables, this is a multiple regression equation rather than a simple regression equation; *cf. supra* note 205.

Equations like (3) are suggested, somewhat loosely, by "human capital theory." However, there remains considerable uncertainty about which variables to put into the equation, what functional form to assume, and how error terms are supposed to behave. Adding more variables is no panacea. *See* Peterson, *supra* note 200, at 214–15.

217. *See infra* § V.D.2.

rience, on average. For example, a male employee with 12 years of education (high school) and 10 years of experience would have a predicted salary of

$$\$7{,}100 + \$1{,}300 \times 12 + \$2{,}200 \times 10 + \$700 \times 1 = \$7{,}100 + \$15{,}600 + \$22{,}000 + \$700 = \$45{,}400 \quad (6)$$

A similarly situated female employee has a predicted salary of only

$$\$7{,}100 + \$1{,}300 \times 12 + \$2{,}200 \times 10 + \$700 \times 0 = \$7{,}100 + \$15{,}600 + \$22{,}000 + \$0 = \$44{,}700 \quad (7)$$

Notice the impact of the dummy variable: $700 is added to equation (6), but not to equation (7).

A major step in proving discrimination is establishing that the estimated coefficient of the dummy variable—$700 in our numerical illustration—is statistically significant. This 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, both for men and women. Similarly, each extra year of experience is worth the same across all levels of education, both for men and women. Furthermore, the premium paid to men does not depend systematically on education or experience. Ability, quality of education, or quality of experience are assumed not to make any systematic difference to the predictions of the model.[218]

The assumptions about the error term—that the errors are independent and identically distributed from person to person in the data set—turn out to be critical for computing *p*-values and demonstrating statistical significance. Regression modeling that does not produce statistically significant coefficients is unlikely to establish discrimination, and statistical significance cannot be established unless stylized assumptions are made about unobservable error terms.[219]

The typical regression model is based on a host of such assumptions; without them, inferences cannot be drawn from the model. With Hooke's law—equation (1)—the model rests on assumptions that are relatively easy to validate experimentally. For the salary discrimination model—equation (3)—validation seems more difficult.[220] Court or counsel may well inquire: What are the assumptions behind the model, and why do they apply to the case at bar? In this regard, it is important to distinguish between situations where (1) the nature of the relationship between the variables is known and regression is being used to make quantitative estimates, and (2) where the nature of the relationship is largely unknown and regression is being used to determine the nature of the relation-

218. Technically, these omitted variables are assumed to be uncorrelated with the error term in the equation.

219. *See supra* note 124.

220. Some of the material in this section is taken from Freedman, *supra* note 112, at 29–35.

ship—or indeed whether any relationship exists at all. The statistical basis for regression theory was developed to handle situations of the first type, with Hooke's law being an example. The basis for the second type of application is analogical, and the tightness of the analogy is a critical issue.

2. Standard Errors, t-statistics, and Statistical Significance

Statistical proof of discrimination depends on the significance of $\hat{d}$ (the estimated coefficient for gender); significance is determined by the *t*-test, using the standard error of $\hat{d}$. The standard error of $\hat{d}$ measures the likely difference between $\hat{d}$ and d, the difference being due to the action of the error term in equation (3). The *t*-statistic is $\hat{d}$ divided by its standard error. For example, in equation (5), $\hat{d}$ = \$700. If the standard error of $\hat{d}$ is \$325, then t = \$700/\$325 = 2.15. This is significant, that is, hard to explain as the mere product of random chance. Under the null hypothesis that $d = 0$, there is only about a 5% chance that the absolute value of t (denoted $|t|$) is greater than 2. A value of t greater than 2 would therefore demonstrate statistical significance.[221] On the other hand, if the standard error is \$1,400, then t = \$700/\$1,400 = 0.5, and the discrepancy could easily result from chance. Of course, the parameter d is only a construct in a model. If the model is wrong, the standard error, *t*-statistic, and significance level are rather difficult to interpret.

Even if the model is granted, there is a further issue: the 5% is a probability for the data given the model, namely, $P(|t| > 2 \,|\, d = 0)$. However, the 5% is often misinterpreted as $P(d = 0 \,|\, \text{data})$. This misinterpretation is commonplace in the social science literature, and it appears in some opinions describing expert testimony.[222] For an objectivist statistician, $P(d = 0 \,|\, \text{data})$ makes no sense: parameters do not exhibit chance variation. For a subjectivist statistician, $P(d = 0 \,|\, \text{data})$ makes good sense, but its computation via the *t*-test could be seriously in error, because the prior probability that $d = 0$ has not been taken into account.[223]

3. Summary

The main ideas of regression modeling can be captured in a hypothetical exchange between a plaintiff seeking to prove salary discrimination and a company denying that allegation. Such a dialog might proceed as follows:

1. Plaintiff argues that the defendant company pays male employees more than females, which establishes prima facie case of discrimination.[224]

221. The cutoff at 2 applies to large samples. Small samples require higher thresholds.

222. *See supra* § IV.B and notes 142, 167.

223. For an objectivist, the vertical bar " | " in $P(|t| > 2 \,|\, d = 0)$ means "computed on the assumption that." For a subjectivist, the bar would signify a conditional probability. *See supra* § IV.B.1, C; *infra* Appendix.

224. The conditions under which a simple disparity between two groups amounts to a prima facie case that shifts the burden of proof to the defendant in Title VII and other discrimination cases have yet

2. The company responds that the men are paid more because they are better educated and have more experience.
3. Plaintiff tries to refute the company's theory by fitting a regression equation like equation (5). Even after adjusting for differences in education and experience, men earn $700 a year more than women, on average. This remaining difference in pay shows discrimination.
4. The company argues that a small difference like $700 could be the result of chance, not discrimination.
5. Plaintiff replies that the coefficient of "gender" in equation (5) is statistically significant, so chance is not a good explanation for the data.

Statistical significance is determined by reference to the observed significance level, which is usually abbreviated to *p*.[225] The *p*-value depends not only on the $700 difference in salary levels, but also on the sample size, among other things.[226] The bigger the sample, other things being equal, the smaller is *p*—and the tighter is plaintiff's argument that the disparity cannot be explained by chance. Often, a cutoff at 5% is used; if *p* is less than 5%, the difference is "statistically significant."[227]

In some cases, the *p*-value has been interpreted as the probability that defendants are innocent of discrimination. However, such an interpretation is wrong: *p* merely represents the probability of getting a large test statistic, given that the model is correct and the true coefficient of "gender" is zero.[228] Therefore, even if the model is undisputed, a *p*-value less than 50% does not necessarily demonstrate a "preponderance of the evidence" against the null hypothesis. Indeed, a *p*-value less than 5% or 1% might not meet the preponderance standard.

In employment discrimination cases, and other contexts too, a wide variety of models are used. This is perhaps not surprising, for specific equations are not dictated by the science. Thus, in a strongly contested case, our dialog would be likely to continue with an exchange about which model is better. Although

to be articulated clearly and comprehensively. *Compare* EEOC v. Olson's Dairy Queens, Inc., 989 F.2d 165, 168 (5th Cir. 1993) (reversing district court for failing to find a prima facie case from the EEOC's statistics on the proportion of African-Americans in defendant's workforce as compared to the proportion of food preparation and service workers in the Houston Standard Metropolitan Statistical Area), *with* Wilkins v. University of Houston, 654 F.2d 388 (5th Cir. 1981) (holding that the district court correctly found that plaintiffs' proof of simple disparities in faculty salaries of men and women did not constitute a prima facie case), *vacated and remanded on other grounds*, 459 U.S. 809 (1982), *aff'd on remand*, 695 F.2d 134 (5th Cir. 1983). *See generally*, D.H. Kaye, *Statistical Evidence: How to Avoid the "Diderot Effect" of Getting Stumped*, Inside Litig., Apr. 1988, at 21. Richard Lempert, *Befuddled Judges: Statistical Evidence in Title VII Cases*, *in* Controversies in Civil Rights (Bernard Grofman ed., forthcoming 2000).

225. *See supra* § IV.B.1.

226. The *p*-value depends on the estimated value of the coefficient and its standard error. These quantities can be computed from (1) the sample size, (2) the means and SDs of the variables, and (3) the correlations between pairs of variables. The computation is rather intricate.

227. *See supra* § IV.B.2.

228. *See supra* §§ IV.B, V.D.2.

statistical assumptions[229] are challenged in court from time to time, arguments more commonly revolve around the choice of variables. One model may be questioned because it omits variables that should be included—for instance, skill levels or prior evaluations;[230] another model may be challenged because it includes "tainted" variables reflecting past discriminatory behavior by the firm.[231] Frequently, each side will have its own equations and its own team of experts; the court then must decide which model—if either—fits the occasion.[232]

229. *See generally supra* § V.D.1 (discussion following equation (7)); Finkelstein & Levin, *supra* note 1, at 397–403; Daniel L. Rubinfeld, Reference Guide on Multiple Regression, in this manual. One example of a statistical assumption is the independence from subject to subject of the error term in equation (3); another example is that the errors have mean zero and constant variance.

230. *E.g.*, Smith v. Virginia Commonwealth Univ., 84 F.3d 672 (4th Cir. 1996) (dispute over omitted variables precludes summary judgment). *Compare* Bazemore v. Friday, 478 U.S. 385 (1986), *on remand*, 848 F.2d 476 (4th Cir. 1988) *and* Sobel v. Yeshiva Univ., 839 F.2d 18, 34 (2d Cir. 1988) (failure to include variables for scholarly productivity did not vitiate plaintiffs' regression study of salary differences because "Yeshiva's experts . . . [offered] no reason, in evidence or analysis, for concluding that they correlated with sex"), *with* Penk v. Oregon State Bd. of Higher Educ., 816 F.2d 458, 465 (9th Cir. 1987) ("Missing parts of the plaintiffs' interpretation of the board's decision-making equation included such highly determinative quality and productivity factors as teaching quality, community and institutional service, and quality of research and scholarship . . . that . . . must have had a significant influence on salary and advancement decisions.") *and* Chang v. University of R.I., 606 F. Supp. 1161, 1207 (D.R.I. 1985) (plaintiff's regression not entitled to substantial weight because the analyst "excluded salient variables even though he knew of their importance").

The same issue arises, of course, with simpler statistical models, such as those used to assess the difference between two proportions. *See*, *e.g.*, Sheehan v. Daily Racing Form, Inc., 104 F.3d 940, 942 (7th Cir. 1997) ("Completely ignored was the more than remote possibility that age was correlated with a legitimate job-related qualification, such as familiarity with computers. Everyone knows that younger people are on average more comfortable with computers than older people are, just as older people are on average more comfortable with manual-shift cars than younger people are.").

231. Michael O. Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases*, 80 Colum. L. Rev. 737 (1980).

232. *E.g.*, *Chang*, 606 F. Supp. at 1207 ("it is plain to the court that [defendant's] model comprises a better, more useful, more reliable tool than [plaintiff's] counterpart"); Presseisen v. Swarthmore College, 442 F. Supp. 593, 619 (E.D. Pa. 1977) ("[E]ach side has done a superior job in challenging the other's regression analysis, but only a mediocre job in supporting their own . . . and the Court is . . . left with nothing."), *aff'd*, 582 F.2d 1275 (3d Cir. 1978).

Appendix

A. Probability and Statistical Inference

The mathematical theory of probability consists of theorems derived from axioms and definitions. The mathematical reasoning is not controversial, but there is some disagreement as to how the theory should be applied; that is, statisticians may differ on the proper interpretation of probabilities in specific applications. There are two main interpretations. For a subjectivist statistician, probabilities represent degrees of belief, on a scale between 0 and 1. An impossible event has probability 0, an event that is sure to happen has probability 1. For an objectivist statistician, probabilities are not beliefs; rather, they are inherent properties of an experiment. If the experiment can be repeated, then in the long run, the relative frequency of an event tends to its probability. For instance, if a fair coin is tossed, the probability of heads is 1/2; if the experiment is repeated, the coin will land heads about one-half the time. If a fair die is rolled, the probability of getting an ace (one spot) is 1/6; if the die is rolled many times, an ace will turn up about one-sixth of the time.[233] (Objectivist statisticians are also called frequentists, while subjectivists are Bayesians, after the Reverend Thomas Bayes, England, c.1701–1761.)

Statisticians also use conditional probability, that is, the probability of one event given that another has occurred. For instance, suppose a coin is tossed twice. One event is that the coin will land HH. Another event is that at least one H will be seen. Before the coin is tossed, there are four possible, equally likely, outcomes: HH, HT, TH, TT. So the probability of HH is 1/4. However, if we know that at least one head has been obtained, then we can rule out two tails TT. In other words, given that at least one H has been obtained, the conditional probability of TT is 0, and the first three outcomes have conditional probability 1/3 each. In particular, the conditional probability of HH is 1/3. This is usually written as P(HH | at least one H) = 1/3. More generally, the probability of any event B is denoted as P(B); the conditional probability of B given A is written as P($B \mid A$).

Two events A and B are independent if the conditional probability of B given that A occurs is equal to the conditional probability of B given that A does not occur. Statisticians often use "~A" to denote the event that A does not occur, so A and B are independent if P($B \mid A$) = P($B \mid \sim A$). If A and B are inde-

233. Probabilities may be estimated from relative frequencies, but probability itself is a subtler idea. For instance, suppose a computer prints out a sequence of ten letters H and T (for heads and tails), which alternate between the two possibilities H and T as follows: H T H T H T H T H T. The relative frequency of heads is 5/10 or 50%, but it is not at all obvious that the chance of an H at the next position is 50%.

pendent, then the probability that both occur is equal to the product of the probabilities:

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B) \tag{1}$$

This is the multiplication rule (or product rule) for independent events. If events are dependent, then conditional probabilities must be used:

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B \,|\, A) \tag{2}$$

This is the multiplication rule for dependent events.

Assessing probabilities, conditional probabilities, and independence is not entirely straightforward. Inquiry into the basis for expert judgment may be useful, and casual assumptions about independence should be questioned.[234]

Bayesian statisticians assign probabilities to hypotheses as well as to events; indeed, for them, the distinction between hypotheses and events may not be a sharp one. If H_0 and H_1 are two hypotheses[235] which govern the probability of an event A, a Bayesian statistician might use the multiplication rule (2) to find that

$$\mathrm{P}(A \text{ and } \mathrm{H}_0) = \mathrm{P}(A \,|\, \mathrm{H}_0)\ \mathrm{P}(\mathrm{H}_0) \tag{3a}$$

and

$$\mathrm{P}(A \text{ and } \mathrm{H}_1) = \mathrm{P}(A \,|\, \mathrm{H}_1)\ \mathrm{P}(\mathrm{H}_1) \tag{3b}$$

Reasoning further that $\mathrm{P}(A) = \mathrm{P}(A \text{ and } \mathrm{H}_0) + \mathrm{P}(A \text{ and } \mathrm{H}_1)$, the statistician would conclude that

$$\mathrm{P}(\mathrm{H}_0 \,|\, A) = \frac{\mathrm{P}(A \,|\, \mathrm{H}_0)\mathrm{P}(\mathrm{H}_0)}{\mathrm{P}(A \,|\, \mathrm{H}_0)\mathrm{P}(\mathrm{H}_0) + \mathrm{P}(A \,|\, \mathrm{H}_1)\mathrm{P}(\mathrm{H}_1)} \tag{4}$$

This is a special case of Bayes' rule, which yields the conditional probability of hypothesis H_0 given that event A has occurred. For example, H_0 might be the hypothesis that blood found at the scene of a crime came from a person unrelated to the defendant; H_1 might deny H_0 and assert that the blood came from the defendant; and A could be the event that blood from both the crime scene and the defendant is type A. Then $\mathrm{P}(\mathrm{H}_0)$ is the prior probability of H_0, based on subjective judgment, while $\mathrm{P}(\mathrm{H}_0 \,|\, A)$ is the posterior probability—the prior probability updated using the data. Here, we have observed a match in type A blood,

234. For problematic assumptions of independence in litigation, see, e.g., *Branion v. Gramly*, 855 F.2d 1256 (7th Cir. 1988); *People v. Collins*, 438 P.2d 33 (Cal. 1968); D.H. Kaye, *The Admissibility of "Probability Evidence" in Criminal Trials* (pts. 1 & 2), 26 Jurimetrics J. 343 (1986), 27 Jurimetrics J. 160 (1987).

235. H_0 is read "H-sub-zero," while H_1 is "H-sub-one."

which occurs in about 42% of the population, so $P(A \mid H_0) = 0.42$.[236] Because the defendant has type A blood, the match probability given that the blood came from him is $P(A \mid H_1) = 1$. If the prior probabilities were, say, $P(H_0) = P(H_1) = 0.5$, then according to (4), the posterior probability would be

$$P(H_0 \mid A) = \frac{0.42 \times 0.5}{0.42 \times 0.5 + 1 \times 0.5} = 0.30 \quad (5)$$

Conversely, the posterior probability that the blood is from the defendant would be

$$P(H_1 \mid A) = 1 - P(H_0 \mid A) = 0.70 \quad (6)$$

Thus, the data make it more probable that the blood is the defendant's: the probability rises from the prior value of $P(H_1) = 0.50$ to the posterior value of $P(H_1 \mid A) = 0.70$.

A frequentist statistician would be hesitant to quantify the probability of hypotheses like H_0 and H_1. Such a statistician would merely report that if H_0 is true, then the probability of type A blood is 42%, whereas if H_1 is true, the probability is 100%.

More generally, H_0 could refer to parameters in a statistical model. For example, H_0 might specify equal selection rates for a population of male and female applicants; H_1 might deny H_0 and assert that the selection rates are not equal; and A could be the event that a test statistic exceeds 2 in absolute value. In such situations, the frequentist statistician would compute $P(A \mid H_0)$ and reject H_0 if this probability fell below a figure such as 0.05.

B. *Technical Details on the Standard Error, the Normal Curve, and Significance Levels*

This section of the Appendix describes several calculations for the pass rate example of section IV. In that example, the population consisted of all 5,000 men and 5,000 women in the applicant pool. Suppose by way of illustration 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 at random from the population, and 50 women. In our sample, the pass rate for the men was 58% and the pass rate for the women was 38%, so 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.

236. Not all statisticians would accept the identification of a population frequency with $P(A \mid H_0)$; indeed, H_0 has been translated into a hypothesis that the true donor has been randomly selected from the population, which is a major step needing justification.

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 estimate 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[237]

$$\sqrt{\frac{5{,}000 - 50}{5{,}000 - 1}} \times \sqrt{\frac{P_{men}\,(1 - P_{men})}{50} + \frac{P_{women}\,(1 - P_{women})}{50}} \qquad (7)$$

In expression (7), 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. With the 60% and 35% figures we have postulated, the standard deviation of the sample differences would be 9.6 percentage points:

$$\sqrt{\frac{5{,}000 - 50}{5{,}000 - 1}} \times \sqrt{\frac{.60\,(1 - .60)}{50} + \frac{.35\,(1 - .35)}{50}} = .096 \qquad (8)$$

Figure 12 shows the histogram for the sample differences.[238] The graph is drawn so the area between two values gives the relative frequency of sample

237. *See, e.g.*, Freedman et al., *supra* note 16, at 414, 503–04; Moore & McCabe, *supra* note 93, 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. On the standard deviation, see *supra* § III.E; Freedman et al., *supra* note 16, at 67–72.

238. The "probability histogram" in Figure 12 shows the "distribution" of the sample differences, indicating the relative likelihoods of the various ranges of possible values; likelihood is represented by

differences 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 for a sample of 50 men and 50 women chosen at random, the difference between the pass rates for the sample men and women will be in the range from 20 to 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.[239] 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 into expression (7) yields

$$\sqrt{\frac{5{,}000 - 50}{5{,}000 - 1}} \times \sqrt{\frac{.58\ (1 - .58)}{50} + \frac{.38\ (1 - .38)}{50}} = .097 \qquad (9)$$

That is about 10 percentage points—the standard error reported in section IV.A.2.[240]

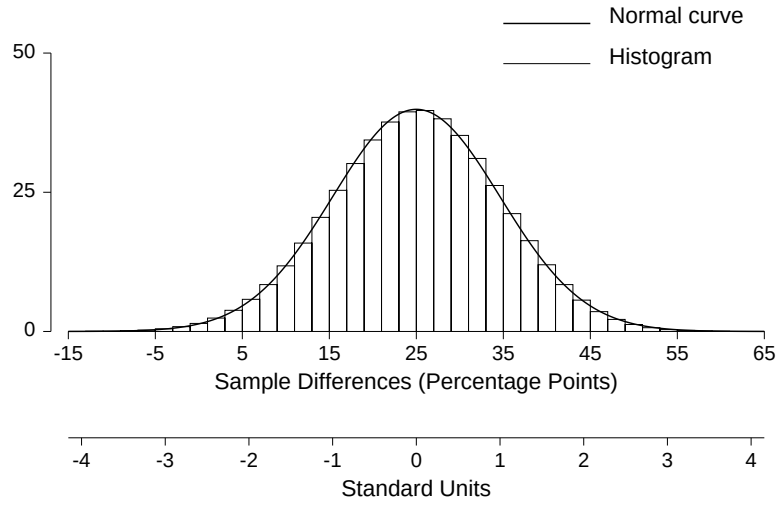
area. The lower horizontal scale shows "standard units," that is, deviations from the expected value relative to the standard error. In our example, the expected value is 25 percentage points and the standard error is 9.6 percentage points. Thus, 35 percentage points would be expressed as $(35 - 25)/9.6 = 1.04$ standard units. The vertical scale in the figure shows probability per standard unit. Probability is measured on a percentage scale, with 100% representing certainty; the maximum shown on the vertical scale in the figure is 50, i.e., 50% per standard unit. *See* Freedman et al., *supra* note 16, at 80, 315.

239. The normal curve is the famous bell-shaped curve of statistics, whose equation is

$$y = \frac{100\%}{\sqrt{2\pi}} e^{-x^2/2}$$

240. There is little difference between (8) and (9)—the standard error does not depend very strongly on the pass rates.

Figure 12. The distribution of the sample difference in pass rates when $P_{men} = 60\%$ and $P_{women} = 35\%$



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, while 95% is between –2 and 2, and 99.7% is between –3 and 3, approximately.

We turn to *p*-values.[241] 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 zero, because $P_{men} - P_{women} = 0$. Since the overall pass rate in the sample is 48%, we use this value to estimate both P_{men} and P_{women} in expression (7):

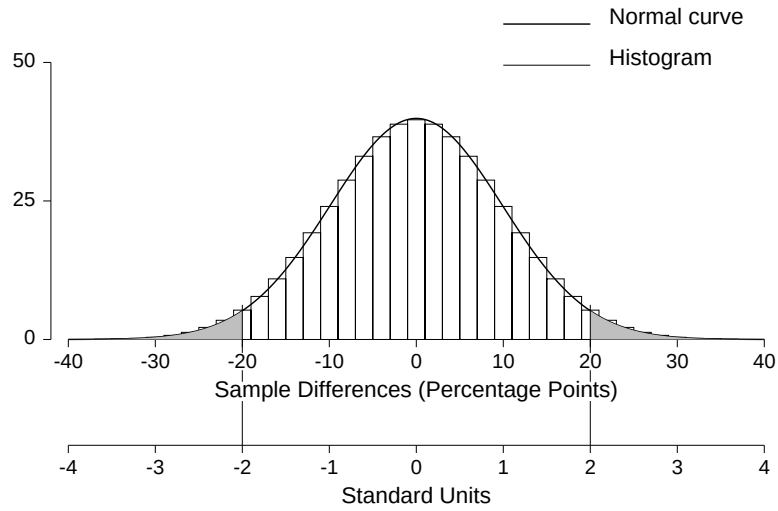
$$\sqrt{\frac{5{,}000 - 50}{5{,}000 - 1}} \times \sqrt{\frac{.48\ (1 - .48)}{50} + \frac{.48\ (1 - .48)}{50}} = .099 \qquad (10)$$

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:

241. *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%.[242]

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%.



Finally, we calculate power.[243] 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

242. 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 may in turn be approximated by the area under the normal curve beyond ±1.9, which is 5.7%. *See, e.g.*, Freedman et al., *supra* note 16, at 318. Keeping track of the edges of the rectangles is called the "continuity correction." *Id.* 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 245.

243. *See supra* note 144.

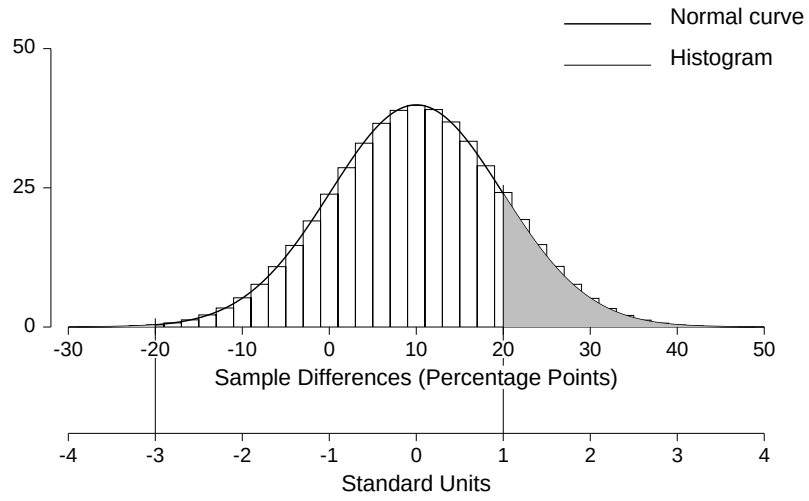
percentage points by equation (1), and the SE estimated from the sample will be about the same. On that basis, only sample differences larger than 20 percentage points or smaller than –20 points will be declared significant.[244] About 1/6 of the area under the normal curve in Figure 14 lies in this region.[245] 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 zero, because it is drawn according to the requirements of the null hypothesis. Figure 14 is centered at 10, because the alternative hypothesis is used to determine the center, rather than the null hypothesis.

244. The null hypothesis asserts a difference of zero. 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 zero, so 20 is 1 SE to the right of the expected value, while –20 is 3 SEs to the left.

245. Let t = sample difference/SE, where the SE is estimated from the data, as in expression (10). 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 expression (7); 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 242.

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 –3. That shaded area is about 1/6. Power is about 1/6, or 17%.



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 A. Freedman et al., Statistics (3d ed. 1998).

adjust for. See control for.

alpha (α). A symbol often used to denote the probability of a Type I error. See Type I error; size. Compare beta.

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. That is, the researchers make a list of areas, choose some at random, and interview people in the selected areas. This is a cost-effective way to draw a sample of people. See probability sample; sampling frame.

arithmetic mean. See mean.

average. See mean.

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

beta (β). A symbol sometimes used to denote power, and sometimes to denote the probability of a Type II error. See Type II error; power. Compare alpha.

bias. A systematic tendency for an estimate to be too high or too low. An estimate is "unbiased" if the bias is zero. (Does not mean prejudice, partiality, or discriminatory intent.) See non-sampling error. Compare sampling error.

bin. A class interval in a histogram. See class interval; histogram.

binary variable. A variable that has only two possible values (e.g., gender). Also called a "dummy variable."

binomial distribution. A distribution for the number of occurrences in repeated, independent "trials" where the probabilities are fixed. For example, the number of heads in 100 tosses of a coin follows a binomial distribution. When the probability is not too close to zero or one and the number of trials is large, the binomial distribution has about the same shape as the normal distribution. See normal distribution; Poisson distribution.

blind. See double-blind experiment.

bootstrap. Also called resampling; Monte Carlo method. A procedure for estimating sampling error by constructing a simulated population on the basis of the sample, then repeatedly drawing samples from this simulated population.

categorical data; categorical variable. See qualitative variable. Compare quantitative 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). The chi-squared statistic 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, one might take households at random, then interview all 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. See multistage cluster sample.

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 mean: SD/mean, or SE/expected value. See expected value; mean; standard deviation; standard error.

collinearity. See multicollinearity.

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 about 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. Compare bias.

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. An association between the dependent and independent variables in an observational study may not be causal, but may instead be due to confounding. 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 gets large, are seldom 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. Compare discrete variable.

control for. Statisticians may "control for" the effects of confounding variables in nonexperimental data by making comparisons for smaller and more homogeneous groups of subjects, or by entering the confounders as explanatory variables in a regression model. To "adjust for" is perhaps a better phrase in the regression context, because in an observational study the confounding factors are not under experimental control; statistical adjustments are an imperfect substitute. See regression model.

control group. See controlled experiment.

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, 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 pre-existing differences. Of course, in many situations, a randomized controlled experiment is impractical, and investigators must then rely on observational studies. Compare observational study.

convenience sample. A non-random sample of units, also called a "grab sample." Such samples are easy to take, but may suffer from serious bias. Mall samples are convenience samples.

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. Compare correlation coefficient; standard error; variance.

covariate. A variable that is related to other variables of primary interest in a study; a measured confounder; a statistical control in a regression equation.

criterion. The variable against which an examination 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.

descriptive statistics. Like the mean or standard deviation, used to summarize data.

differential validity. Differences in the correlation 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. Compare continuous variable.

distribution. See frequency distribution; probability distribution; sampling distribution.

disturbance term. A synonym for 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 patients in the control group. In a *single-blind experiment*, the patients do not know whether they are in treatment or control; however, the diagnosticians have this information.

dummy variable. Generally, a dummy variable takes only the values 0 or 1, and distinguishes one group of interest from another. See binary variable; regression model.

econometrics. 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 "disturbance term."

estimator. A sample statistic used to estimate the value of a population parameter. For instance, the sample mean commonly is used to estimate the popu-

lation mean. The term "estimator" connotes a statistical procedure, while an "estimate" connotes a particular numerical result.

expected value. See random variable.

experiment. See controlled experiment; randomized controlled experiment. Compare observational study.

explanatory variable. See independent variable, regression model.

factors. See independent variable.

Fisher's exact test. When comparing two sample proportions, for instance the proportions of whites and blacks getting a promotion, an investigator may wish 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. For more details, see Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers 156–59 (1990). See hypergeometric distribution; *p*-value; significance test; statistical hypothesis.

fitted value. See residual.

fixed significance level. Also alpha; size. A pre-set 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. Compare observed significance level; *p*-value.

frequency distribution. Shows how often specified values occur in a data set.

Gaussian distribution. A synonym for the normal 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; variance. Compare regression model.

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. With 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. How many times will items of a certain type come into the sample? The hypergeometric distribution gives the probabilities. For more details, see 1 William Feller, An Introduction to Prob-

ability Theory and its Applications 41–42 (2d ed. 1957). Compare Fisher's exact test.

hypothesis. See alternative hypothesis; null hypothesis; one-sided hypothesis; significance test; statistical hypothesis; two-sided hypothesis.

hypothesis test. See significance test.

independence. Events are independent when the probability of one is unaffected by the occurrence or non-occurrence of the other. Compare conditional probability.

independent variable. Independent variables (also called explanatory variables or factors) are used in a regression model to predict the dependent variable. For instance, the unemployment rate has been used as the independent variable in a model for predicting the crime rate; the unemployment rate is the independent variable in this model, and the crime rate is the dependent variable. See regression model. Compare dependent variable.

indicator variable. See dummy variable.

interquartile range. Difference between 25th and 75th percentile. See percentile.

interval estimate. A "confidence interval," or an estimate coupled with a standard error. See confidence interval; standard error. Compare point estimate.

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.

level. The level of a significance test is denoted alpha (α). See alpha; fixed significance level; observed significance level; *p*-value; significance test.

linear combination. To obtain a linear combination of two variables, multiply the first variable by some constant, multiply the second variable by another constant, and add the two products. For instance, $2u + 3v$ 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. Compare risk.

lurking variable. See confounding variable.

mean. Also, the average; the expected value of a random variable. 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, as in "weighted mean" or "weighted average." See random variable. Compare median; mode.

median. The median is another way to find the center of a batch of numbers. The median is the 50th 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.) Compare mean; mode; percentile.

meta-analysis. Attempts to combine information from all studies on a certain topic. For example, in the epidemiologic context, a meta-analysis may attempt to provide a summary odds ratio and confidence interval for the effect of a certain exposure on a certain disease.

mode. The most commonly observed value. Compare mean; median.

model. See probability model; regression model; statistical model.

multicollinearity. Also, collinearity. The existence of correlations among the "independent variables" in a regression model. See independent variable; regression model.

multiple comparison. Making several statistical tests on the same data set. Multiple comparisons complicate the interpretation of a *p*-value. For example, if 20 divisions of a company are examined, 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. Compare *p*-value; significance test; statistical 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*-squared; regression model. Compare correlation coefficient.

multiple regression. A regression equation that includes two or more independent variables. See regression model. Compare simple regression.

multivariate methods. Methods for fitting models with multiple variables, especially, multiple response variables; occasionally, multiple explanatory variables. See regression model.

multi-stage 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. Compare controlled experiment.

nonresponse bias. Systematic error created by differences between respondents and nonrespondents. If the nonresponse rate is high, this bias may be severe.

non-sampling error. A catch-all term for sources of error in a survey, other than sampling error. Non-sampling 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 non-response 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 need be nothing wrong with a distribution that differs from the normal.

null hypothesis. For example, a hypothesis that there is no difference between two groups from which samples are drawn. See significance test; statistical hypothesis. Compare alternative 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 to the death rate for non-smokers. 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.) Compare confounding variable; controlled experiment.

observed significance level. A synonym for *p*-value. See significance test. Compare fixed significance level.

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; the odds against rain are 1 to 2.

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 to 5% of people who are not exposed, then the odds of the disease in the exposed group are 10/90 = 1/9, compared to 5/95 = 1/19 in the unexposed group. The odds ratio is 19/9 = 2.1. An odds ratio of 1 indicates no association. Compare relative risk.

one-sided hypothesis. Excludes the possibility that a parameter could be, e.g., less than the value asserted in the null hypothesis. A one-sided hypothesis leads to a one-tailed test. See significance test; statistical hypothesis; compare two-sided hypothesis.

one-tailed test. See significance test.

outlier. An observation that is far removed from the bulk of the data. Outliers may indicate faulty measurements and they may exert undue influence on summary statistics, 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 significance test; statistical hypothesis.

parameter. A numerical characteristic of a population or a model. See probability model.

percentile. To get the percentiles of a data set, array the data from the smallest value to the largest. Take the 90th percentile by way of example: 90% of the values fall below the 90th percentile, and 10% are 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: 50% of the values fall below the median, and 50% are above. 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). Compare mean; median; quartile.

placebo. See double-blind experiment.

point estimate. An estimate of the value of a quantity expressed as a single number. See estimator. Compare confidence interval; interval estimate.

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 is 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. Compare sample; sampling frame.

posterior probability. See Bayes' rule.

power. The probability that a statistical test will reject the null hypothesis. To compute power, one 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 beta; significance test. Compare alpha; size; *p*-value.

practical significance. Substantive importance. Statistical significance does not necessarily establish practical significance. With large samples, small differences can be statistically significant. See significance test.

predicted value. See residual.

predictive validity. A 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 percent; 100% corresponds to 1, while 5% corresponds to .05, and so forth.

probability density. Describes the probability distribution of a random variable. The chance that the random variable falls in an interval equals the area below the density and above the interval. (However, not all random variables have densities.) See probability distribution; random variable.

probability distribution. Gives probabilities for possible values or ranges of values of a random variable. Often, the distribution is described in terms of a 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 minimize selection bias, but can be expensive to draw.

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. Compare continuous variable; discrete variable; dummy variable. See variable.

quartile. The 25th or 75th percentile. See percentile. Compare median.

***R*-squared (R^2).** Measures how well a regression equation fits the data. R^2 varies between zero (no fit) and one (perfect fit). R^2 does not measure the validity of underlying assumptions. See regression model. Compare multiple correlation coefficient; standard error of regression.

random error. Sources of error that are haphazard in their effect. These are reflected 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 a pair of dice are thrown, 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 7, with chance 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 will be off (in either direction) by something like a standard error (SE) or so. If the random variable has a more or less normal distribution, there is about a 68% chance for it to fall in the range "expected value − SE" to "expected value + SE." See normal curve; standard error.

randomization. See controlled experiment; randomized 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, that is, by randomization. See controlled experiment. Compare observational study.

range. The difference between the biggest and the 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 (simple) regression equation.

regression model. A "regression model" attempts to combine the values of certain variables (the "independent" or "explanatory" variables) in order to get expected values for another variable (the "dependent" variable). Sometimes, "regression model" refers to a probability model for the data; if no qualifications are made, the model will generally be linear, and errors will be assumed independent across observations, with common variance; the coefficients in the linear combination are called "regression coefficients"; these are parameters. At times, "regression model" refers to an equation (the "regression equation") estimated from data, typically by least squares.

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, as well as "statistical controls" like 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. Salary would be the dependent variable; education, experience, and the dummy would be the independent variables. See least squares; multiple regression; random error; variance. Compare general linear model.

relative risk. A measure of association used in epidemiology. For instance, if 10% of all people exposed to a chemical develop a disease, compared to 5% of people who are not exposed, then 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. For more details, see Abraham M. Lilienfeld & David E. Lilienfeld, Foundations of Epidemiology 209 (2d ed. 1980). Compare odds ratio.

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

resampling. See bootstrap.

residual. The difference between an actual and a "predicted" value. The predicted value comes typically from a regression equation, and is also called the "fitted value." See regression model; independent variable.

response variable. See 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 modified slightly.

sample. A set of units collected for study. Compare population.

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,

$$\text{estimate} = \text{true value} + \text{bias} + \text{sampling error}.$$

Sampling error is also called "chance error" or "random error." See standard error. Compare bias; non-sampling 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; scatter diagram. 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."

selection bias. Systematic error due to non-random selection of subjects for study.

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. Compare specificity.

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. A significance test involves formulating a statistical hypothesis and a test statistic, computing a *p*-value, and comparing *p* to some pre-established value ("alpha") 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 were 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. The "one-tailed" test would reject the null hypothesis if the sample 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 hypothesis could be tested against the "two-sided alternative" 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 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, that is, the one-tailed test would appear to give stronger evidence against the null hypothesis. See *p*-value; statistical hypothesis; *t*-statistic.

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. One takes a unit at random (as if by lottery), sets it aside, takes another at random from what is left, and so forth.

simple regression. A regression equation that includes only one independent variable. Compare multiple regression.

size. A synonym for alpha (α).

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. Compare sensitivity.

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 will not make the feet grow. Such non-causal correlations are said to be "spuri-

ous." (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.) Compare confounding variable.

standard deviation (SD). The SD indicates how far a typical element deviates from the average. For instance, in round numbers, the average height of women age 18 and over in the United States is 5 feet 4 inches. However, few women are exactly average; most will deviate from average, at least by a little. The SD is sort of an average deviation from average. For the height distribution, the SD is 3 inches. The height of a typical woman is around 5 feet 4 inches, but is off that average value by something like 3 inches.

For distributions that follow the normal curve, about 68% of the elements are in the range "mean – SD" to "mean + SD." Thus, about 68% of women have heights in the range 5 feet 1 inch to 5 feet 7 inches. Deviations from the average that exceed three or four SDs are extremely unusual. Many authors use "standard deviation" to also mean standard error. See 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. Compare expected value; standard deviation.

standard error of regression. Indicates how actual values differ (in some average sense) from the fitted values in a regression model. See regression model; residual. Compare *R*-squared.

standardization. See standardized variable.

standardized variable. Transformed to have mean zero and variance one. This involves two steps: (1) subtract the mean, (2) divide by the standard deviation.

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

statistical controls. Procedures that try to filter out the effects of confounding variables on non-experimental data, for instance, by "adjusting" through statistical procedures (like multiple regression). Variables in a multiple regression equation. See multiple regression; confounding variable; observational study. Compare controlled experiment.

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 "tested" against the data

with 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 test. See significance test.

statistical significance. See *p*-value.

stratified random sample. A type of probability sample. One divides the population up into relatively homogeneous groups called "strata," and draws a random sample separately from each stratum.

systematic sampling. The elements of the population are numbered consecutively as 1, 2, 3 . . . . Then, every *k*th element is chosen. If $k = 10$, for instance, the sample would consist of items 1, 11, 21 . . . . Sometimes the starting point is chosen at random from 1 to *k*.

***t*-statistic.** A test statistic, used to make the "*t*-test." The *t*-statistic indicates how far away an estimate is from its expected value, relative to the standard error. The expected value is computed using the null hypothesis that is being tested. Some authors refer to the *t*-statistic, others to the "*z*-statistic," especially when the sample is large. In such cases, 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 bigger than 2, or smaller than –2, then 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% = 13/100 of the area under the normal curve is outside the range from –1.5 to 1.5. On the other hand, $t = 3$ is remarkable: only 3/1,000 of the area lies outside the range from –3 to 3. This discussion is predicated 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." A *t*-test is not appropriate for small samples drawn

from a population that is not normal. See *p*-value; significance test; statistical hypothesis.

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." Such 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 *p*-value; statistical hypothesis; *t*-statistic.

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

treatment group. See controlled experiment.

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. A two-sided alternative hypothesis suggests a two-tailed test. See statistical hypothesis; significance test. Compare one-sided hypothesis.

two-tailed test. See significance test.

Type I error. A statistical test makes a "Type I error" when (1) the null hypothesis is true and (2) the test rejects the null hypothesis, i.e., there is a false positive. 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 significance test; statistical hypothesis. Compare alpha; Type II error.

Type II error. A statistical test makes a "Type II error" when (1) the null hypothesis is false and (2) 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 significance test; statistical hypothesis. Compare beta; Type I error.

unbiased estimator. An estimator that is correct on average, over the possible data sets. The estimates have no systematic tendency to be high or low. Compare bias.

uniform distribution. For example, a whole number picked at random from 1 to 100 has the uniform distribution: all values are equally likely. Similarly, a uniform distribution is obtained 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.

validity. The extent to which an instrument measures what it is supposed to, rather than something else. The validity of a standardized test is often 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. For example, in a study of households, household income; in a study of people, employment status (employed, unemployed, not in labor force).

variance. The square of the standard deviation. Compare standard error; covariance.

***z*-statistic.** See *t*-statistic.

***z*-test.** See *t*-test.

References on Statistics

General Surveys

David Freedman et al., Statistics (3d ed. 1998).

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 (3d ed. 1991).

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 G. 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 & Supp. 1987).

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

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

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

1 & 2 Modern Scientific Evidence: The Law and Science of Expert Testimony (David L. Faigman et al. eds., 1997)

Ramona Paetzold & Steven L. Willborn, The Statistics of Discrimination: Using Statistical Evidence in Discrimination Cases (1994)

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

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

Hans Zeisel & David Kaye, Prove It with Figures: Empirical Methods in Law and Litigation (1997)

General Reference

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