





Reference Manual on Scientific Evidence: Third Edition


ISBN
978-0-309-21421-6

1038 pages
6 x 9
PAPERBACK (2011)

Committee on the Development of the Third Edition of the Reference Manual on Scientific Evidence; Federal Judicial Center; National Research Council

 Add book to cart

 Find similar titles

 Share this PDF



Visit the National Academies Press online and register for...

- ✓ Instant access to free PDF downloads of titles from the
 - NATIONAL ACADEMY OF SCIENCES
 - NATIONAL ACADEMY OF ENGINEERING
 - INSTITUTE OF MEDICINE
 - NATIONAL RESEARCH COUNCIL
- ✓ 10% off print titles
- ✓ Custom notification of new releases in your field of interest
- ✓ Special offers and discounts

Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press. Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences. Request reprint permission for this book

Reference Guide on Statistics

DAVID H. KAYE AND DAVID A. FREEDMAN

David H. Kaye, M.A., J.D., is Distinguished Professor of Law and Weiss Family Scholar, The Pennsylvania State University, University Park, and Regents' Professor Emeritus, Arizona State University Sandra Day O'Connor College of Law and School of Life Sciences, Tempe.

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

[Editor's Note: Sadly, Professor Freedman passed away during the production of this manual.]

CONTENTS

- I. Introduction, 213
 - A. Admissibility and Weight of Statistical Studies, 214
 - B. Varieties and Limits of Statistical Expertise, 214
 - C. Procedures That Enhance Statistical Testimony, 215
 - 1. Maintaining professional autonomy, 215
 - 2. Disclosing other analyses, 216
 - 3. Disclosing data and analytical methods before trial, 216
- II. How Have the Data Been Collected? 216
 - A. Is the Study Designed to Investigate Causation? 217
 - 1. Types of studies, 217
 - 2. Randomized controlled experiments, 220
 - 3. Observational studies, 220
 - 4. Can the results be generalized? 222
 - B. Descriptive Surveys and Censuses, 223
 - 1. What method is used to select the units? 223
 - 2. Of the units selected, which are measured? 226
 - C. Individual Measurements, 227
 - 1. Is the measurement process reliable? 227
 - 2. Is the measurement process valid? 228
 - 3. Are the measurements recorded correctly? 229
 - D. What Is Random? 230
- III. How Have the Data Been Presented? 230
 - A. Are Rates or Percentages Properly Interpreted? 230
 - 1. Have appropriate benchmarks been provided? 230
 - 2. Have the data collection procedures changed? 231
 - 3. Are the categories appropriate? 231
 - 4. How big is the base of a percentage? 233
 - 5. What comparisons are made? 233
 - B. Is an Appropriate Measure of Association Used? 233

- C. Does a Graph Portray Data Fairly? 236
 - 1. How are trends displayed? 236
 - 2. How are distributions displayed? 236
- D. Is an Appropriate Measure Used for the Center of a Distribution? 238
- E. Is an Appropriate Measure of Variability Used? 239
- IV. What Inferences Can Be Drawn from the Data? 240
 - A. Estimation, 242
 - 1. What estimator should be used? 242
 - 2. What is the standard error? The confidence interval? 243
 - 3. How big should the sample be? 246
 - 4. What are the technical difficulties? 247
 - B. Significance Levels and Hypothesis Tests, 249
 - 1. What is the p -value? 249
 - 2. Is a difference statistically significant? 251
 - 3. Tests of interval estimates? 252
 - 4. Is the sample statistically significant? 253
 - C. Evaluating Hypothesis Tests, 253
 - 1. What is the power of the test? 253
 - 2. What about small samples? 254
 - 3. One tail or two? 255
 - 4. How many tests have been done? 256
 - 5. What are the rival hypotheses? 257
 - D. Posterior Probabilities, 258
- V. Correlation and Regression, 260
 - A. Scatter Diagrams, 260
 - B. Correlation Coefficients, 261
 - 1. Is the association linear? 262
 - 2. Do outliers influence the correlation coefficient? 262
 - 3. Does a confounding variable influence the coefficient? 262
 - C. Regression Lines, 264
 - 1. What are the slope and intercept? 265
 - 2. What is the unit of analysis? 266
 - D. Statistical Models, 268
- Appendix, 273
 - A. Frequentists and Bayesians, 273
 - B. The Spock Jury: Technical Details, 275
 - C. The Nixon Papers: Technical Details, 278
 - D. A Social Science Example of Regression: Gender Discrimination in Salaries, 279
 - 1. The regression model, 279
 - 2. Standard errors, t -statistics, and statistical significance, 281
- Glossary of Terms, 283
- References on Statistics, 302

I. Introduction

Statistical assessments are prominent in many kinds of legal cases, including antitrust, employment discrimination, toxic torts, and voting rights cases.¹ This reference guide describes the elements of statistical reasoning. We hope the explanations will help judges and lawyers to understand statistical terminology, to see the strengths and weaknesses of statistical arguments, and to apply relevant legal doctrine. The 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 evidence.
- Section II addresses data collection and explains why the design of a study is the most important determinant of its quality. This section compares experiments with observational studies and surveys with censuses, indicating when the various kinds of study are likely to provide useful results.
- Section III discusses the art of summarizing data. This section considers the mean, median, and standard deviation. These are basic descriptive statistics, and most statistical analyses use them as building blocks. This section also discusses patterns in data that are brought out by graphs, percentages, and tables.
- Section IV describes the logic of statistical inference, emphasizing foundations and disclosing limitations. This section covers estimation, standard errors and confidence intervals, *p*-values, and hypothesis tests.
- Section V shows how associations can be described by scatter diagrams, correlation coefficients, and regression lines. Regression is often used to infer causation from association. This section explains the technique, indicating the circumstances under which it and other statistical models are likely to succeed—or fail.
- An appendix provides some technical details.
- The glossary defines statistical terms that may be encountered in litigation.

1. See generally *Statistical Science in the Courtroom* (Joseph L. Gastwirth ed., 2000); *Statistics and the Law* (Morris H. DeGroot et al. eds., 1986); 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* (2d ed. 2001); 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).

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, because 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 Because most statistical methods relied on in court are described in textbooks or journal articles and are capable of producing useful results when properly applied, these methods generally satisfy important aspects of the "scientific knowledge" requirement in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*³ Of course, a particular study may use a method that is entirely appropriate but that is 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 lack the "fit" spoken of in *Daubert*.⁵ Or the study might 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 theory, theoretical statistics, and applied statistics. Probability theory is the mathematical study of outcomes that are governed, at least in part, by chance. Theoretical statistics is about the properties of statistical procedures, including error rates; probability theory plays a key role in this endeavor. Applied statistics draws on both of these fields to develop techniques for collecting or analyzing particular types of data.

2. See generally 2 McCormick on Evidence §§ 321, 324.3 (Kenneth S. Broun ed., 6th ed. 2006). Studies published by government agencies also may be admissible as public records. *Id.* § 296.

3. 509 U.S. 579, 589–90 (1993).

4. See *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 152 (1999) (suggesting that the trial court should "make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."); *Malletier v. Dooney & Bourke, Inc.*, 525 F. Supp. 2d 558, 562–63 (S.D.N.Y. 2007) ("While errors in a survey's methodology usually go to the weight accorded to the conclusions rather than its admissibility, . . . 'there will be occasions when the proffered survey is so flawed as to be completely unhelpful to the trier of fact.'") (quoting *AHP Subsidiary Holding Co. v. Stuart Hale Co.*, 1 F.3d 611, 618 (7th Cir.1993)).

5. *Daubert*, 509 U.S. at 591; *Anderson v. Westinghouse Savannah River Co.*, 406 F.3d 248 (4th Cir. 2005) (motion to exclude statistical analysis that compared black and white employees without adequately taking into account differences in their job titles or positions was properly granted under *Daubert*); *Malletier*, 525 F. Supp. 2d at 569 (excluding a consumer survey for "a lack of fit between the survey's questions and the law of dilution" and errors in the execution of the survey).

Statistical expertise is not confined to those with degrees in statistics. Because statistical reasoning underlies many kinds of empirical research, scholars in a variety of fields—including biology, economics, epidemiology, political science, and psychology—are exposed to statistical ideas, with an emphasis on the methods most important to the discipline.

Experts who specialize in using statistical methods, and whose professional careers demonstrate this orientation, are most likely to use appropriate procedures and correctly interpret the results. By contrast, forensic scientists often lack basic information about the studies underlying their testimony. *State v. Garrison*⁶ illustrates the problem. In this 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 lacks. As a result, cases involving statistical evidence frequently are (or should be) “two expert” cases of interlocking testimony. A labor economist, for example, may supply a definition of the relevant labor market from which an employer draws its employees; the statistical expert may then compare the race of new hires to the racial composition of the labor market. Naturally, the value of the statistical analysis depends on the substantive knowledge that informs it.⁸

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 would be required in other contexts. Thus, experts who testify (or who supply results used in testimony) should conduct the analysis required to address in a professionally responsible fashion the issues posed by the litigation.⁹ Questions about the freedom of inquiry accorded to testifying experts,

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

7. *Id.* at 566, 568. For other examples, see David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* § 12.2 (2d ed. 2011).

8. In *Vuyovich 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.

9. 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

as well as the scope and depth of their investigations, may reveal some of the limitations to the testimony.

2. *Disclosing other analyses*

Statisticians analyze data using a variety of methods. There is much to be said for looking at the data in several ways. To permit a fair evaluation of the analysis that is eventually settled on, however, the testifying expert can be asked to explain how that approach was developed. According to some commentators, counsel who know of analyses that do not support the client's position should reveal them, rather than presenting only favorable results.¹⁰

3. *Disclosing data and analytical methods before trial*

The collection of data often is expensive and subject to errors and omissions. Moreover, careful exploration of the data can be time-consuming. To minimize debates at trial over the accuracy of data and the choice of analytical techniques, pretrial discovery procedures should be used, particularly with respect to the quality of the data and the method of analysis.¹¹

II. How Have the Data Been Collected?

The interpretation of data often depends on understanding “study design”—the plan for a statistical study and its implementation.¹² Different designs are suited to answering different questions. Also, flaws in the data can undermine any statistical analysis, and data quality is often determined by study design.

In many cases, statistical studies are used to show causation. Do food additives cause cancer? Does capital punishment deter crime? Would additional disclosures

by any party to the litigation and that the expert receive a letter of engagement providing for these and other safeguards).

10. *Id.* 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 National Research Council 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.

11. 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 Law & Soc’y Rev. 1255 (1990).

12. For introductory treatments of data collection, see, for example, David Freedman et al., *Statistics* (4th ed. 2007); Darrell Huff, *How to Lie with Statistics* (1993); David S. Moore & William I. Notz, *Statistics: Concepts and Controversies* (6th ed. 2005); Hans Zeisel, *Say It with Figures* (6th ed. 1985); Zeisel & Kaye, *supra* note 1.

in a securities prospectus cause investors to behave differently? The design of studies to investigate causation is the first topic of this section.¹³

Sample data can be used to describe a population. The population is the whole class of units that are of interest; the sample is the set of units chosen for detailed study. Inferences from the part to the whole are justified when the sample is representative. Sampling is the second topic of this section.

Finally, the accuracy of the data will be considered. Because making and recording measurements is an error-prone activity, error rates should be assessed and the likely impact of errors considered. Data quality is the third topic of this section.

A. Is the Study Designed to Investigate Causation?

1. Types of studies

When causation is the issue, anecdotal evidence can be brought to bear. So can observational studies or controlled experiments. Anecdotal reports may be of value, but they are ordinarily more helpful in generating lines of inquiry than in proving causation.¹⁴ Observational studies can establish that one factor is associ-

13. See also Michael D. Green et al., Reference Guide on Epidemiology, Section V, in this manual; Joseph Rodricks, Reference Guide on Exposure Science, Section E, in this manual.

14. In medicine, evidence from clinical practice can be the starting point for discovery of cause-and-effect relationships. For examples, see David A. Freedman, *On Types of Scientific Enquiry*, in *The Oxford Handbook of Political Methodology* 300 (Janet M. Box-Steffensmeier et al. eds., 2008). Anecdotal evidence is rarely definitive, and 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., *McClain v. Metabolife Int'l, Inc.*, 401 F.3d 1233, 1244 (11th Cir. 2005) (“simply because a person takes drugs and then suffers an injury does not show causation. Drawing such a conclusion from temporal relationships leads to the blunder of the *post hoc ergo propter hoc* fallacy.”); *In re Baycol Prods. Litig.*, 532 F. Supp. 2d 1029, 1039–40 (D. Minn. 2007) (excluding a meta-analysis based on reports to the Food and Drug Administration of adverse events); *Leblanc v. Chevron USA Inc.*, 513 F. Supp. 2d 641, 650 (E.D. La. 2007) (excluding plaintiffs’ experts’ opinions that benzene causes myelofibrosis because the causal hypothesis “that has been generated by case reports . . . has not been confirmed by the vast majority of epidemiologic studies of workers being exposed to benzene and more generally, petroleum products.”), *vacated*, 275 Fed. App’x. 319 (5th Cir. 2008) (remanding for consideration of newer government report on health effects of benzene); cf. *Matrixx Initiatives, Inc. v. Siracusano*, 131 S. Ct. 1309, 1321 (2011) (concluding that adverse event reports combined with other information could be of concern to a reasonable investor and therefore subject to a requirement of disclosure under SEC Rule 10b-5, but stating that “the mere existence of reports of adverse events . . . says nothing in and of itself about whether the drug is causing the adverse events”). Other courts are more open to “differential diagnoses” based primarily on timing. E.g., *Best v. Lowe’s Home Ctrs., Inc.*, 563 F.3d 171 (6th Cir. 2009) (reversing the exclusion of a physician’s opinion that exposure to propenyl chloride caused a man to lose his sense of smell because of the timing in this one case and the physician’s inability to attribute the change to anything else); *Kaye et al.*, *supra* note 7, §§ 8.7.2 & 12.5.1. See also *Matrixx Initiatives, supra*, at 1322 (listing “a temporal relationship” in a single patient as one indication of “a reliable causal link”).

ated with another, but work is needed to bridge the gap between association and causation. Randomized controlled experiments are ideally suited for demonstrating causation.

Anecdotal evidence usually amounts to reports that events of one kind are followed by events of another kind. Typically, the reports are not even sufficient to show association, because there is no comparison group. For example, some children who live near power lines develop leukemia. Does exposure to electrical and magnetic fields cause this disease? The anecdotal evidence is not compelling because leukemia also occurs among children without exposure.¹⁵ 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 and lower among the unexposed. That would be association.

The next issue is crucial: Exposed and unexposed people may differ in ways other than the exposure they have experienced. For example, children who live near power lines could come from poorer families and be more at risk from other environmental hazards. Such differences can create the appearance of a cause-and-effect relationship. Other differences can mask a real relationship. Cause-and-effect relationships often are quite subtle, and carefully designed studies are needed to draw valid conclusions.

An epidemiological classic makes the point. 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 recently had been tarred. This is anecdotal evidence. But the argument is incomplete. For one thing, most people—whether exposed to asphalt fumes or unexposed—did not develop lung cancer. A comparison of rates was needed. The epidemiologists found that exposed persons and unexposed persons suffered from lung cancer at similar rates: Tar was probably not the causal agent. Exposure to cigarette smoke, however, turned out to be strongly associated with lung cancer. This study, in combination with later ones, made a compelling case that smoking cigarettes is the main cause of lung cancer.¹⁶

A good study design compares outcomes for subjects who are exposed to some factor (the treatment group) with outcomes for other subjects who are

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

16. Richard Doll & A. Bradford Hill, *A Study of the Aetiology of Carcinoma of the Lung*, 2 *Brit. Med. J.* 1271 (1952). This was a matched case-control study. Cohort studies soon followed. See Green et al., *supra* note 13. For a review of the evidence on causation, 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).

not exposed (the control group). Now there is another important distinction to be made—that between controlled experiments and observational studies. In a controlled experiment, the investigators decide which subjects will be exposed and which subjects will go into the control group. In observational studies, by contrast, the subjects themselves choose their exposures. Because of self-selection, the treatment and control groups are likely to differ with respect to influential factors other than the one of primary interest. (These other factors are called lurking variables or confounding variables.)¹⁷ With the health effects of power lines, family background is a possible confounder; so is exposure to other hazards. Many confounders have been proposed to explain the association between smoking and lung cancer, but careful epidemiological studies have ruled them out, one after the other.

Confounding remains a problem to reckon with, even for the best observational research. For example, women with herpes are more likely to develop cervical cancer than other women. Some investigators concluded that herpes caused cancer: In other words, they thought the association was causal. Later research showed that the primary cause of cervical cancer was human papilloma virus (HPV). Herpes was a marker of sexual activity. Women who had multiple sexual partners were more likely to be exposed not only to herpes but also to HPV. The association between herpes and cervical cancer was due to other variables.¹⁸

What are “variables?” In statistics, a variable is a characteristic of units in a study. With a study of people, the unit of analysis is the person. Typical variables include income (dollars per year) and educational level (years of schooling completed): These variables describe people. With a study of school districts, the unit of analysis is the district. Typical variables include average family income of district residents and average test scores of students in the district: These variables describe school districts.

When investigating a cause-and-effect relationship, the variable that represents the effect is called the dependent variable, because it depends on the causes. The variables that represent the causes are called independent variables. With a study of smoking and lung cancer, the independent variable would be smoking (e.g., number of cigarettes per day), and the dependent variable would mark the presence or absence of lung cancer. Dependent variables also are called outcome variables or response variables. Synonyms for independent variables are risk factors, predictors, and explanatory variables.

17. For example, a confounding variable may be correlated with the independent variable and act causally on the dependent variable. If the units being studied differ on the independent variable, they are also likely to differ on the confounder. The confounder—not the independent variable—could therefore be responsible for differences seen on the dependent variable.

18. For additional examples and further discussion, see Freedman et al., *supra* note 12, at 12–28, 150–52; David A. Freedman, *From Association to Causation: Some Remarks on the History of Statistics*, 14 Stat. Sci. 243 (1999). Some studies find that herpes is a “cofactor,” which increases risk among women who are also exposed to HPV. Only certain strains of HPV are carcinogenic.

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 comparable, except for the treatment. This minimizes the role of confounding. Minor imbalances will remain, due to the play of random chance; the likely effect on study results can be assessed by statistical techniques.¹⁹ The bottom line is that causal inferences based on well-executed randomized experiments are generally more secure than inferences based on well-executed observational studies.

The following example should help bring the discussion together. Today, we know that taking aspirin helps prevent heart attacks. But initially, there was some controversy. People who take aspirin rarely have heart attacks. This is anecdotal evidence for a protective effect, but it proves almost nothing. After all, few people have frequent heart attacks, whether or not they take aspirin regularly. A good study compares heart attack rates for two groups: people who take aspirin (the treatment group) and people who do not (the controls). An observational study would be easy to do, but in such a study the aspirin-takers are likely to be different from the controls. Indeed, they are likely to be sicker—that is why they are taking aspirin. The study would be biased against finding a protective effect. Randomized experiments are harder to do, but they provide better evidence. It is the experiments that demonstrate a protective effect.²⁰

In summary, data from a treatment group without a control group generally reveal very little and can be misleading. Comparisons are essential. If subjects are assigned to treatment and control groups at random, a difference in the outcomes between the two groups can usually be accepted, within the limits of statistical error (*infra* Section IV), as a good measure of the treatment effect. However, if the groups are created in any other way, differences that existed before treatment may contribute to differences in the outcomes or mask differences that otherwise would become manifest. Observational studies succeed to the extent that the 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 conduct a randomized controlled experiment, people would need to be assigned randomly to a treatment group or a control group. People in the treatment group would know they were subject to the death penalty for murder; the

19. Randomization of subjects to treatment or control groups puts statistical tests of significance on a secure footing. Freedman et al., *supra* note 12, at 503–22, 545–63; see *infra* Section IV.

20. In other instances, experiments have banished strongly held beliefs. *E.g.*, Scott M. Lippman et al., Effect of Selenium and Vitamin E on Risk of Prostate Cancer and Other Cancers: The Selenium and Vitamin E Cancer Prevention Trial (SELECT), 301 JAMA 39 (2009).

controls would know that they were exempt. Conducting such an experiment is not possible.

Many studies of the deterrent effect of the death penalty have been conducted, all observational, and some have attracted judicial attention. Researchers have catalogued differences in the incidence of murder in states with and without the death penalty and have analyzed changes in homicide rates and execution rates over the years. When reporting on such observational studies, investigators may speak of “control groups” (e.g., the states without capital punishment) or claim they are “controlling for” confounding variables by statistical methods.²¹ However, association is not causation. The causal inferences that can be drawn from analysis of observational data—no matter how complex the statistical technique—usually rest on a foundation that is less secure than that provided by randomized controlled experiments.

That said, observational studies can be very useful. For example, there is strong observational evidence that smoking causes lung cancer (*supra* Section II.A.1). Generally, observational studies provide good evidence in the following circumstances:

- The association is seen in studies with different designs, on different kinds of subjects, and done by different research groups.²² That reduces the chance that the association is due to a defect in one type of study, a peculiarity in one group of subjects, or the idiosyncrasies of one research group.
- The association holds when effects of confounding variables are taken into account by appropriate methods, for example, comparing smaller groups that are relatively homogeneous with respect to the confounders.²³
- There is a plausible explanation for the effect of the independent variable; alternative explanations in terms of confounding should be less plausible than the proposed causal link.²⁴

21. A procedure often used to control for confounding in observational studies is regression analysis. The underlying logic is described *infra* Section V.D and in Daniel L. Rubinfield, Reference Guide on Multiple Regression, Section II, in this manual. *But see* Richard A. Berk, Regression Analysis: A Constructive Critique (2004); Rethinking Social Inquiry: Diverse Tools, Shared Standards (Henry E. Brady & David Collier eds., 2004); David A. Freedman, Statistical Models: Theory and Practice (2005); David A. Freedman, *Oasis or Mirage*, Chance, Spring 2008, at 59.

22. For example, case-control studies are designed one way and cohort studies another, with many variations. *See, e.g.*, Leon Gordis, Epidemiology (4th ed. 2008); *supra* note 16.

23. The idea is to control for the influence of a confounder by stratification—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 materially 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 International Agency for Research on Cancer, *supra* note 16.

24. 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). Plausibility, however, is a function of time and circumstances.

Thus, evidence for the causal link does not depend on observed associations alone.

Observational studies can produce legitimate disagreement among experts, and there is no mechanical procedure for resolving such differences of opinion. In the end, deciding whether associations are causal typically is not a matter of statistics alone, but also rests on scientific judgment. There are, however, some basic questions to ask when appraising causal inferences based on empirical studies:

- Was there a control group? Unless comparisons can be made, 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 through a process outside the control of the investigator (an observational study)?
- 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 treatment and control groups comparable?
- If not, what adjustments were made to address confounding?
- Were the adjustments sensible and sufficient?²⁵

4. *Can the results be generalized?*

Internal validity is about the specifics of a particular study: Threats to internal validity include confounding and chance differences between treatment and control groups. *External validity* is about using a particular study or set of studies to reach more general conclusions. A careful randomized controlled experiment on a large but unrepresentative group of subjects will have high internal validity but low external validity.

Any study must be conducted on certain subjects, at certain times and places, and using certain treatments. To extrapolate from the conditions of a study to more general conditions raises questions of external validity. For example, studies suggest that definitions of insanity given to jurors influence decisions in cases of incest. Would the definitions have a similar effect in cases of murder? Other studies indicate that recidivism rates for ex-convicts are not affected by provid-

25. Many courts have noted the importance of confounding variables. *E.g.*, *People Who Care v. Rockford Bd. of Educ.*, 111 F.3d 528, 537–38 (7th Cir. 1997) (educational achievement); *Hollander v. Sandoz Pharms. Corp.*, 289 F.3d 1193, 1213 (10th Cir. 2002) (stroke); *In re Proportionality Review Project (II)*, 757 A.2d 168 (N.J. 2000) (capital sentences).

ing them with temporary financial support after release. Would similar results be obtained if conditions in the labor market were different?

Confidence in the appropriateness of an extrapolation cannot come from the experiment itself. It comes from knowledge about outside factors that would or would not affect the outcome.²⁶ Sometimes, several studies, each having different limitations, all point in the same direction. This is the case, for example, with studies indicating that jurors who approve of the death penalty are more likely to convict in a capital case.²⁷ Convergent results support the validity of generalizations.

B. Descriptive Surveys and Censuses

We now turn to a second topic—choosing units for study. A census tries to measure some characteristic of every unit in a population. This is often impractical. Then investigators use sample surveys, which measure characteristics for only part of a population. The accuracy of the information collected in a census or survey depends on how the units are selected for study and how the measurements are made.²⁸

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 one must ask

26. 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 responses 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, e.g., Bruce N. Ames et al., *The Causes and Prevention of Cancer*, 92 Proc. Nat’l Acad. Sci. USA 5258 (1995); 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.”).

27. Phoebe C. Ellsworth, *Some Steps Between Attitudes and Verdicts*, in *Inside the Juror* 42, 46 (Reid Hastie ed., 1993). Nonetheless, 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.

28. See Shari Seidman Diamond, *Reference Guide on Survey Research*, Sections III, IV, in this manual.

whether the missing data are likely to differ in some systematic way from the data that are collected.²⁹ The methodological framework of a scientific survey is different. With probability methods, a sampling frame (i.e., an explicit list of units in the population) must be created. Individual units then are selected by an objective, well-defined chance procedure, and measurements are made on the sampled units.

To illustrate the idea of a sampling frame, suppose that a defendant in a criminal case seeks a change of venue: According to him, popular opinion is so adverse that it would be difficult to impanel an unbiased jury. To prove the state of popular opinion, the defendant commissions a survey. The relevant population consists of all persons in the jurisdiction who might be called for jury duty. The sampling frame is the list of all potential jurors, which is maintained by court officials and is made available to the defendant. In this hypothetical case, the fit between the sampling frame and the population would be excellent.

In other situations, the sampling frame is more problematic. In an obscenity case, for example, the defendant can offer a survey of community standards.³⁰ The population comprises all adults in the legally relevant district, but obtaining a full list of such people may not be possible. Suppose the survey is done by telephone, but cell phones are excluded from the sampling frame. (This is usual practice.) Suppose too that cell phone users, as a group, hold different opinions from landline users. In this second hypothetical, the poll is unlikely to reflect the opinions of the cell phone users, no matter how many individuals are sampled and no matter how carefully the interviewing is done.

Many surveys do not use probability methods. In commercial disputes involving trademarks or advertising, the population of all potential purchasers of a product is hard to identify. Pollsters may resort to an easily accessible subgroup of the population, for example, shoppers in a mall.³¹ Such convenience samples may be biased by the interviewer's discretion in deciding whom to approach—a form of

29. The U.S. Decennial Census generally does not count everyone that it should, and it counts some people who should not be counted. There is evidence that net undercount is greater in some demographic groups than others. Supplemental studies may enable statisticians to adjust for errors and omissions, but the adjustments rest on uncertain assumptions. See Lawrence D. Brown et al., *Statistical Controversies in Census 2000*, 39 *Jurimetrics J.* 347 (2007); David A. Freedman & Kenneth W. Wachter, *Methods for Census 2000 and Statistical Adjustments*, in *Social Science Methodology* 232 (Steven Turner & William Outhwaite eds., 2007) (reviewing technical issues and litigation surrounding census adjustment in 1990 and 2000); 9 *Stat. Sci.* 458 (1994) (symposium presenting arguments for and against adjusting the 1990 census).

30. On the admissibility of such polls, see *State v. Midwest Pride IV, Inc.*, 721 N.E.2d 458 (Ohio Ct. App. 1998) (holding one such poll to have been properly excluded and collecting cases from other jurisdictions).

31. *E.g.*, *Smith v. Wal-Mart Stores, Inc.*, 537 F. Supp. 2d 1302, 1333 (N.D. Ga. 2008) (treating a small mall-intercept survey as entitled to much less weight than a survey based on a probability sample); *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).

selection bias—and the refusal of some of those approached to participate—non-response bias (*infra* Section II.B.2). Selection bias is acute when constituents write their representatives, listeners call into radio talk shows, interest groups collect information from their members, or attorneys choose cases for trial.³²

There are procedures that attempt to correct for selection bias. In quota sampling, for example, the interviewer is instructed to interview so many women, so many older people, so many ethnic minorities, and the like. But quotas still leave discretion to the interviewers in selecting members of each demographic group and therefore do not solve the problem of selection bias.³³

Probability methods are designed to avoid selection bias. Once the population is reduced to a sampling frame, the units to be measured are selected by a lottery that gives each unit in the sampling frame a known, nonzero probability of being chosen. Random numbers leave no room for selection bias.³⁴ Such procedures are used to select individuals for jury duty. They also have been used to choose “bellwether” cases for representative trials to resolve issues in a large group of similar cases.³⁵

32. *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). *Cf. 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 3000 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 3000 claimants.” *Id.* See *infra* note 35.

A well-known 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 the popular vote, 57% to 43%. In fact, Franklin Roosevelt won by a landslide vote of 62% to 38%. See Freedman et al., *supra* note 12, 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.

33. See Freedman et al., *supra* note 12, at 337–39.

34. 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. *Id.* at 339–41. 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 units are not in any special order, then systematic sampling is often comparable to simple random sampling.

35. *E.g.*, *In re Simon II Litig.*, 211 F.R.D. 86 (E.D.N.Y. 2002), *vacated*, 407 F.3d 125 (2d Cir. 2005), *dismissed*, 233 F.R.D. 123 (E.D.N.Y. 2006); *In re Estate of Marcus Human Rights Litig.*, 910

2. Of the units selected, which are measured?

Probability sampling ensures that within the limits of chance (*infra* Section IV), the sample will be representative of the sampling frame. The question remains regarding which units actually get measured. When documents are sampled for audit, all the selected ones can be examined, at least in principle. Human beings are less easily managed, and some will refuse to cooperate. Surveys should therefore report nonresponse rates. A large nonresponse rate warns of bias, although supplemental studies may establish that nonrespondents are similar to respondents with respect to characteristics of interest.³⁶

In short, a good survey defines an appropriate population, uses a probability 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. Data from the sample can be extrapolated

F. Supp. 1460 (D. Haw. 1995), *aff'd sub nom.* 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), *rev'd*, 151 F.3d 297 (5th Cir. 1998); *cf. In re Chevron U.S.A., Inc.*, 109 F.3d 1016 (5th Cir. 1997) (discussed *supra* note 32). Although trials in a suitable random sample of cases can produce reasonable estimates of average damages, the propriety of precluding individual trials raises questions of due process and the right to trial by jury. See Thomas E. Willging, Mass Torts Problems and Proposals: A Report to the Mass Torts Working Group (Fed. Judicial Ctr. 1999); *cf. Wal-Mart Stores, Inc. v. Dukes*, 131 S. Ct. 2541, 2560–61 (2011). The cases and the views of commentators are described more fully in David H. Kaye & David A. Freedman, *Statistical Proof*, in 1 *Modern Scientific Evidence: The Law and Science of Expert Testimony* § 6:16 (David L. Faigman et al. eds., 2009–2010).

36. For discussions of nonresponse rates and admissibility of surveys conducted for litigation, see *Johnson v. Big Lots Stores, Inc.*, 561 F. Supp. 2d 567 (E.D. La. 2008) (fair labor standards); *United States v. Dentsply Int'l, Inc.*, 277 F. Supp. 2d 387, 437 (D. Del. 2003), *rev'd on other grounds*, 399 F.3d 181 (3d Cir. 2005) (antitrust).

The 1936 *Literary Digest* election poll (*supra* note 32) illustrates the dangers in nonresponse. Only 24% of the 10 million people who received questionnaires returned them. Most of the respondents probably had strong views on the candidates and 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 12, at 335–36. Even when demographic characteristics of the sample match those of the population, caution is indicated. See David Streitfeld, *Shere Hite and the Trouble with Numbers*, 1 *Chance* 26 (1988); Chamont Wang, *Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety* 174–76 (1993).

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.” Nonetheless, 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; *cf. In re United States*, 426 F.3d 1 (1st Cir. 2005) (reaching the same result with respect to underrepresentation of African Americans resulting in part from nonresponse bias).

to describe the characteristics of the population. Of course, surveys may be useful even if they fail to meet these criteria. But then, additional arguments are needed to justify the inferences.

C. Individual Measurements

1. Is the measurement process reliable?

Reliability and validity are two aspects of accuracy in measurement. In statistics, reliability refers to reproducibility of results.³⁷ A reliable measuring instrument returns consistent measurements. A scale, for example, is perfectly reliable if it reports the same weight for the same object time and again. It may not be accurate—it may always report a weight that is too high or one that is too low—but the perfectly reliable scale always reports the same weight for the same object. Its errors, if any, are systematic: They always point in the same direction.

Reliability can be ascertained by measuring the same quantity several times; the measurements must be made independently to avoid bias. Given independence, the correlation coefficient (*infra* Section V.B) between repeated measurements can be used as a measure of reliability. This is sometimes called a test-retest correlation or a reliability coefficient.

A courtroom example is DNA identification. An early method of identification required laboratories to determine the lengths of fragments of DNA. By making independent replicate measurements of the fragments, laboratories determined the likelihood that two measurements differed by specified amounts.³⁸ Such results were needed to decide whether a discrepancy between a crime sample and a suspect sample was sufficient to exclude the suspect.³⁹

Coding provides another example. In many studies, descriptive information is obtained on the subjects. For statistical purposes, the information usually has to be reduced to numbers. The process of reducing information to numbers is called “coding,” and the reliability of the process should be evaluated. For example, 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.⁴⁰

37. 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 Pharms., Inc.*, 509 U.S. 579, 590 n.9 (1993), for example, distinguishes “evidentiary reliability” from reliability in the technical sense of giving consistent results. We use “reliability” to denote the latter.

38. See National Research Council, *The Evaluation of Forensic DNA Evidence* 139–41 (1996).

39. *Id.*; National Research Council, *DNA Technology in Forensic Science* 61–62 (1992). Current methods are discussed in David H. Kaye & George Sensabaugh, *Reference Guide on DNA Identification Evidence*, Section II, in this manual.

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

Two different aspects of reliability should be considered. First, the “within-observer variability” of judgments should be small—the same evaluator should rate essentially identical cases in similar ways. Second, the “between-observer variability” should be small—different evaluators should rate the same cases in essentially 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, for example, in pulse rate or blood pressure. The measurements may be reliable. Nonetheless, the polygraph is not valid as a lie detector unless the measurements it makes are well correlated with lying.⁴¹

When there is an established way of measuring a variable, a new measurement process can be validated by comparison with the established one. Breathalyzer readings can be validated against alcohol levels found in blood samples. LSAT scores used for law school admissions can be validated against grades earned in law school. A common measure of validity is the correlation coefficient between the predictor and the criterion (e.g., test scores and later performance).⁴²

Employment discrimination cases illustrate some of the difficulties. Thus, plaintiffs suing under Title VII of the Civil Rights Act may challenge an employment test that has a disparate impact on a protected group, and defendants may try to justify the use of a test as valid, reliable, and a business necessity.⁴³ For validation, the most appropriate criterion variable is clear enough: job performance. However, plaintiffs may then turn around and challenge the validity of performance ratings. For reliability, administering the test twice to the same group of people may be impractical. Even if repeated testing is practical, it may be statistically inadvisable, because subjects may learn something from the first round of testing that affects their scores on the second round. Such “practice effects” are likely to compromise the independence of the two measurements, and independence is needed to estimate reliability. Statisticians therefore use internal evidence

41. See *United States v. Henderson*, 409 F.3d 1293, 1303 (11th Cir. 2005) (“while the physical responses recorded by a polygraph machine may be tested, ‘there is no available data to prove that those specific responses are attributable to lying.’”); National Research Council, *The Polygraph and Lie Detection* (2003) (reviewing the scientific literature).

42. As the discussion of the correlation coefficient indicates, *infra* Section V.B, the closer the coefficient is to 1, the greater the validity. For a review of data on test reliability and validity, see Paul R. Sackett et al., *High-Stakes Testing in Higher Education and Employment: Appraising the Evidence for Validity and Fairness*, 63 *Am. Psychologist* 215 (2008).

43. See, e.g., *Washington v. Davis*, 426 U.S. 229, 252 (1976); *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 430–32 (1975); *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971); *Lanning v. S.E. Penn. Transp. Auth.*, 308 F.3d 286 (3d Cir. 2002).

from the test itself. For example, if scores on the first half of the test correlate well with scores from the second half, then that is evidence of reliability.

A further problem is that test-takers are likely to be a select group. The ones who get the jobs are even more highly selected. Generally, selection attenuates (weakens) the correlations. There are methods for using internal measures of reliability to estimate test-retest correlations; there are other methods that correct for attenuation. However, such methods depend on assumptions about the nature of the test and the procedures used to select the test-takers and are therefore open to challenge.⁴⁴

3. *Are the measurements recorded correctly?*

Judging the adequacy of data collection involves an examination of the process by which measurements are taken. Are responses to interviews coded correctly? Do mistakes distort the results? How much data are missing? What was done to compensate for gaps in the data? These days, data are stored in computer files. Cross-checking the files against the original sources (e.g., paper records), at least on a sample basis, can be informative.

Data quality is a pervasive issue in litigation and in applied statistics more generally. A programmer moves a file from one computer to another, and half the data disappear. The definitions of crucial variables are lost in the sands of time. Values get corrupted: Social security numbers come to have eight digits instead of nine, and vehicle identification numbers fail the most elementary consistency checks. Everybody in the company, from the CEO to the rawest mailroom trainee, turns out to have been hired on the same day. Many of the residential customers have last names that indicate commercial activity (“Happy Valley Farriers”). These problems seem humdrum by comparison with those of reliability and validity, but—unless caught in time—they can be fatal to statistical arguments.⁴⁵

44. See Thad Dunning & David A. Freedman, *Modeling Selection Effects*, in *Social Science Methodology* 225 (Steven Turner & William Outhwaite eds., 2007); Howard Wainer & David Thissen, *True Score Theory: The Traditional Method*, in *Test Scoring* 23 (David Thissen & Howard Wainer eds., 2001).

45. See, e.g., *Malletier v. Dooney & Bourke, Inc.*, 525 F. Supp. 2d 558, 630 (S.D.N.Y. 2007) (coding errors contributed “to the cumulative effect of the methodological errors” that warranted exclusion of a consumer confusion 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). *But see Dalley v. Mich. 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.”).

D. What Is Random?

In the law, a selection process sometimes is called “random,” provided that it does not exclude identifiable segments of the population. Statisticians use the term in a far more technical sense. For example, if we were to choose one person at random from a population, in the strict statistical sense, we would have to ensure that everybody in the population is chosen with exactly the same probability. With a randomized controlled experiment, subjects are assigned to treatment or control at random in the strict sense—by tossing coins, throwing dice, looking at tables of random numbers, or more commonly these days, by using a random number generator on a computer. The same rigorous definition applies to random sampling. It is randomness in the technical sense that provides assurance of unbiased estimates from a randomized controlled experiment or a probability sample. Randomness in the technical sense also justifies calculations of standard errors, confidence intervals, and *p*-values (*infra* Sections IV–V). Looser definitions of randomness are inadequate for statistical purposes.

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.⁴⁶ Section III.A discusses rates or percentages and provides some cautionary examples of misleading summaries, indicating the kinds of questions that might be considered when summaries are presented in court. Percentages are often used to demonstrate statistical association, which is the topic of Section III.B. Section III.C 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?

The selective presentation of numerical information is like quoting someone out of context. Is a fact that “over the past three years,” a particular index fund of large-cap stocks “gained a paltry 1.9% a year” indicative of poor management? Considering that “the average large-cap value fund has returned just 1.3% a year,”

46. See generally Freedman et al., *supra* note 12; Huff, *supra* note 12; Moore & Notz, *supra* note 12; Zeisel, *supra* note 12.

a growth rate of 1.9% is hardly an indictment.⁴⁷ In this example and many others, it is helpful to find a benchmark that puts the figures 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 one year—not because of an abrupt crime wave, but because a new police commissioner introduced an improved reporting system.⁴⁸ For a time, police officials in Washington, D.C., “demonstrated” the success of a law-and-order campaign by valuing stolen goods at \$49, just below the \$50 threshold then used for inclusion in the Federal Bureau of Investigation’s Uniform Crime Reports.⁴⁹ Allegations of manipulation in the reporting of crime from one time period to another are legion.⁵⁰

Changes in data collection procedures are by no means limited to crime statistics. 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.

3. Are the categories appropriate?

Misleading summaries also can be produced by the choice of categories to be used for comparison. In *Philip Morris, Inc. v. Loew’s Theatres, Inc.*,⁵¹ and *R.J. Reynolds Tobacco Co. v. Loew’s Theatres, Inc.*,⁵² 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 1.⁵³ Only $14\% + 22\% = 36\%$ of the sample preferred Triumph to Merit, whereas

47. Paul J. Lim, *In a Downturn, Buy and Hold or Quit and Fold?*, N.Y. Times, July 27, 2008.

48. James P. Levine et al., *Criminal Justice in America: Law in Action* 99 (1986) (referring to a change from 1959 to 1960).

49. D. Seidman & M. Couzens, *Getting the Crime Rate Down: Political Pressure and Crime Reporting*, 8 Law & Soc’y Rev. 457 (1974).

50. Michael D. Maltz, *Missing UCR Data and Divergence of the NCVS and UCR Trends*, in *Understanding Crime Statistics: Revisiting the Divergence of the NCVS and UCR* 269, 280 (James P. Lynch & Lynn A. Addington eds., 2007) (citing newspaper reports in Boca Raton, Atlanta, New York, Philadelphia, Broward County (Florida), and Saint Louis); Michael Vasquez, *Miami Police: FBI: Crime Stats Accurate*, Miami Herald, May 1, 2008.

51. 511 F. Supp. 855 (S.D.N.Y. 1980).

52. 511 F. Supp. 867 (S.D.N.Y. 1980).

53. *Philip Morris*, 511 F. Supp. at 866.

29% + 11% = 40% preferred Merit to Triumph. By selectively combining categories, however, the defendant attempted to create a different impression. Because 24% found the brands to be about the same, and 36% preferred Triumph, the defendant claimed that a clear majority (36% + 24% = 60%) found Triumph “as good [as] or better than Merit.”⁵⁴ The court resisted this chicanery, finding that defendant’s test results did not support the advertising claims.⁵⁵

Table 1. Data Used by a 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

There was a similar distortion in claims for the accuracy of a home pregnancy test. 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

Table 2 does indicate that only one error occurred 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.⁵⁶

54. *Id.*

55. *Id.* at 856–57.

56. Only two women in the sample were not pregnant; the test gave correct results for both of them. Although a false-positive rate of 0 is ideal, an estimate based on a sample of only two women is not. These data are reported in Arnold Barnett, *How Numbers Can Trick You*, Tech. Rev., Oct. 1994, at 38, 44–45.

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. Putting them on the same base (100) makes it easy to compare them.

When the base is small, however, a small change in absolute terms can generate a large percentage gain or loss. This could lead to newspaper headlines such as “Increase in Thefts Alarming,” even when the total number of thefts is small.⁵⁷ Conversely, a large base will make for small percentage increases. In these situations, actual numbers may be more revealing than percentages.

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 should be the baseline? If the first year of operation is used, a large percentage increase should be expected because of startup problems. 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, the percentage may not portray the trend fairly. 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; asking about the base can also be helpful.⁵⁸

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.⁵⁹

Percentages often are used to describe the association between two variables. Suppose that a university alleged to discriminate against women in admitting

57. Lyda Longa, *Increase in Thefts Alarming*, Daytona News-J. June 8, 2008 (reporting a 35% increase in armed robberies in Daytona Beach, Florida, in a 5-month period, but not indicating whether the number had gone up by 6 (from 17 to 23), by 300 (from 850 to 1150), or by some other amount).

58. For assistance in coping with percentages, see Zeisel, *supra* note 12, at 1–24.

59. Correlation and regression are discussed *infra* Section V.

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 3.⁶⁰

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,⁶¹ 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 one-fifth the chance of males of being admitted, and that may be of real concern.

Table 3. Admissions by Gender

Decision	Male	Female	Total
Admit	350	200	550
Deny	450	400	850
Total	800	600	1400

For Table 3, the selection ratio (used by the Equal Employment Opportunity Commission in its “80% rule”) is $33/44 = 75\%$, meaning that, on average, women have 75% the chance of admission that men have.⁶² 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 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.

60. 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.

61. See, e.g., *State v. Gibbs*, 758 A.2d 327, 337 (Conn. 2000); *Primeaux v. Dooley*, 747 N.W.2d 137, 141 (S.D. 2008); D.H. Kaye, *Statistical Evidence of Discrimination in Jury Selection*, in *Statistical Methods in Discrimination Litigation 13* (David H. Kaye & Mikel Aickin eds., 1986).

62. A procedure that selects candidates from the least successful group at a rate less than 80% of the rate for the most successful group “will generally be regarded by the Federal enforcement agencies as evidence of adverse impact.” EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (2008). 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.

The analogous statistic used in epidemiology is called the relative risk. See Green et al., *supra* note 13, Section III.A. Relative risks are usually quoted as decimals; for example, a selection ratio of 75% corresponds to a relative risk of 0.75.

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.⁶³ 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 homogeneous, 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

Decision	Engineering		Business	
	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. Expressed in a more technical manner, 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 admitted 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 is that the business college, to which most of the women applied, admits relatively few applicants. It is easier to be accepted at the engineering college, the college to which most of the men applied. This example illustrates a common issue: Association can result from combining heterogeneous statistical material.⁶⁴

63. 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 a man is also 99/19.

64. 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). The tables are an instance of Simpson's Paradox.

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.

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 rate of all crimes of domestic violence in Florida (per 100,000 people) appears to decline rapidly over the 10 years from 1998 through 2007; in Figure 2, the same rate appears to drop slowly.⁶⁵ The moral is simple: Pay attention to the markings on the axes to determine whether the scale is appropriate.

Figure 1

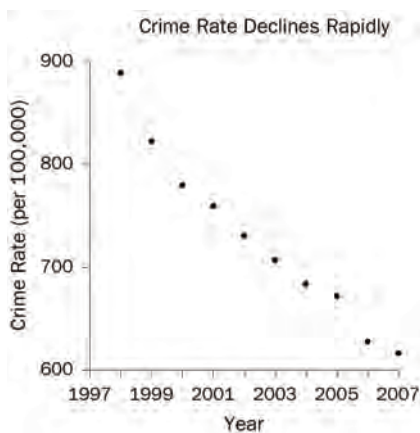
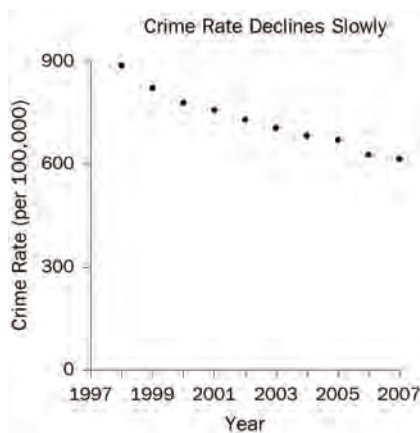


Figure 2



2. How are distributions displayed?

A graph commonly used to display the distribution of data is the histogram. One axis denotes the numbers, and the other indicates how often those fall within

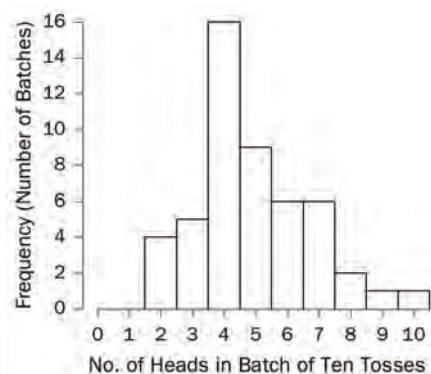
65. Florida Statistical Analysis Center, Florida Department of Law Enforcement, Florida's Crime Rate at a Glance, available at http://www.fdle.state.fl.us/FSAC/Crime_Trends/domestic_violence/index.asp. The data are from the Florida Uniform Crime Report statistics on crimes ranging from simple stalking and forcible fondling to murder and arson. The Web page with the numbers graphed in Figures 1 and 2 is no longer posted, but similar data for all violent crime is available at http://www.fdle.state.fl.us/FSAC/Crime_Trends/Violent-Crime.aspx.

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:⁶⁶

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

The histogram is shown in Figure 3.⁶⁷ A histogram shows how the data are distributed over the range of possible values. The spread can be made to appear larger or smaller, however, by changing the scale of the horizontal axis. Likewise, the shape can be altered somewhat by changing the size of the bins.⁶⁸ It may be worth inquiring how the analyst chose the bin widths.

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



66. 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.

67. In Figure 3, the bin width is 1. There were no 0s or 1s in the data, so the bars over 0 and 1 disappear. There is a bin from 1.5 to 2.5; the four 2s in the data fall into this bin, so the bar over the interval from 1.5 to 2.5 has height 4. There is another bin from 2.5 to 3.5, which catches five 3s; the height of the corresponding bar is 5. 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 \times width) are proportional to the frequencies. In general, a histogram differs from a bar graph in that it represents frequencies by area, not height. See Freedman et al., *supra* note 12, at 31–41.

68. 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.

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 all the numbers and dividing the total by how many numbers were added. By comparison, the median cuts the numbers into halves: half the numbers are larger than the median and half are smaller.⁶⁹ Yet a third statistic is the mode, which is the most common number in the dataset. These statistics are different, although they are not always clearly distinguished.⁷⁰ The mean takes account of all the data—it involves the total of all the numbers; however, particularly with small datasets, a few unusually large or small observations may have too much influence on the mean. The median is resistant to such outliers.

Thus, studies of damage awards in tort cases find that the mean is larger than the median.⁷¹ This is because the mean takes into account (indeed, is heavily influenced by) the magnitudes of the relatively few very large awards, whereas the median merely counts their number. If one is seeking a single, representative number for the awards, the median may be more useful than the mean.⁷² 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.⁷³

69. 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.

70. In ordinary language, the arithmetic mean, the median, and the mode seem to be referred to interchangeably as “the average.” In statistical parlance, however, the average is the arithmetic mean. The mode is rarely used by statisticians, because it is unstable: Small changes to the data often result in large changes to the mode.

71. In a study using a probability sample of cases, the median compensatory award in wrongful death cases was \$961,000, whereas the mean award was around \$3.75 million for the 162 cases in which the plaintiff prevailed. Thomas H. Cohen & Steven K. Smith, U.S. Dep’t of Justice, Bureau of Justice Statistics Bulletin NCJ 202803, Civil Trial Cases and Verdicts in Large Counties 2001, 10 (2004). In *TXO Production Corp. v. Alliance Resources Corp.*, 509 U.S. 443 (1993), briefs portraying the punitive damage system as out of control pointed to mean punitive awards. These were some 10 times larger than the median awards described in briefs defending the system of punitive damages. 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).

72. In passing on proposed settlements in class-action lawsuits, courts have been advised to look to the magnitude of the settlements negotiated by the parties. But the mean settlement will be large if a higher number of meritorious, high-cost cases are resolved early in the life cycle of the litigation. This possibility led the court in *In re Educational Testing Service Praxis Principles of Learning and Teaching, Grades 7-12 Litig.*, 447 F. Supp. 2d 612, 625 (E.D. La. 2006), to regard the smaller median settlement as “more representative of the value of a typical claim than the mean value” and to use this median in extrapolating to the entire class of pending claims.

73. 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

Research also has shown that there is considerable stability in the ratio of punitive to compensatory damage awards, and the Supreme Court has placed great weight on this ratio in deciding whether punitive damages are excessive in a particular case. In *Exxon Shipping Co. v. Baker*,⁷⁴ Exxon contended that an award of \$2.5 billion in punitive damages for a catastrophic oil spill in Alaska was unreasonable under federal maritime law. The Court looked to a “comprehensive study of punitive damages awarded by juries in state civil trials [that] found a median ratio of punitive to compensatory awards of just 0.62:1, but a mean ratio of 2.90:1.”⁷⁵ The higher mean could reflect a relatively small but disturbing proportion of unjustifiably large punitive awards.⁷⁶ Looking to the median ratio as “the line near which cases like this one largely should be grouped,” the majority concluded that “a 1:1 ratio, which is above the median award, is a fair upper limit in such maritime cases [of reckless conduct].”⁷⁷

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.⁷⁸ 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 the range is unstable because it depends entirely on the most extreme values.⁷⁹ The interquartile range is the difference between the 25th and 75th percentiles.⁸⁰ 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.⁸¹

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.)

74. 128 S. Ct. 2605 (2008).

75. *Id.* at 2625.

76. According to the Court, “the outlier cases subject defendants to punitive damages that dwarf the corresponding compensatories,” and the “stark unpredictability” of these rare awards is the “real problem.” *Id.* This perceived unpredictability has been the subject of various statistical studies and much debate. See Anthony J. Sebok, *Punitive Damages: From Myth to Theory*, 92 Iowa L. Rev. 957 (2007).

77. 128 S. Ct. at 2633.

78. 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.

79. Moreover, the range typically depends on the number of units in the sample.

80. 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.

81. When the distribution follows the normal curve, about 68% of the data will be within 1 standard deviation of the mean, and about 95% will be within 2 standard deviations of the mean. For other distributions, the proportions will be different.

There are no hard and fast rules about which statistic is the best. In general, the bigger the measures of spread are, the more the numbers are dispersed.⁸² Particularly in small datasets, the standard deviation can be influenced heavily by a few outlying values. To assess the extent of this influence, the mean and the standard deviation can be recomputed with the outliers discarded. Beyond this, any of the statistics can (and often should) be supplemented with a figure that displays much of the data.

IV. What Inferences Can Be Drawn from the Data?

The inferences that may be drawn from a study depend on the design of the study and the quality of the data (*supra* Section II). The data might not address the issue of interest, might be systematically in error, or might be difficult to interpret because of confounding. Statisticians would group these concerns together under the rubric of “bias.” In this context, bias means systematic error, with no connotation of prejudice. We turn now to another concern, namely, the impact of random chance on study results (“random error”).⁸³

If a pattern in the data is the result of chance, it is likely to wash out when more data are collected. By applying the laws of probability, a statistician can assess the likelihood that random error will create spurious patterns of certain kinds. Such assessments are often viewed as essential when making inferences from data.

Technically, the standard deviation is the square root of the variance; the variance is the mean square deviation from the mean. For example, if the mean is 100, then 120 deviates from the mean by 20, and the square of 20 is $20^2 = 400$. If the variance (i.e., the mean of the squared deviations) is 900, then the standard deviation is the square root of 900, that is, $\sqrt{900} = 30$. Taking the square root gets back to the original scale of the measurements. For example, if the measurements are of length in inches, the variance is in square inches; taking the square root changes back to inches.

82. In *Exxon Shipping Co. v. Baker*, 554 U.S. 471 (2008), along with the mean and median ratios of punitive to compensatory awards of 0.62 and 2.90, the Court referred to a standard deviation of 13.81. *Id.* at 498. These numbers led the Court to remark that “[e]ven to those of us unsophisticated in statistics, the thrust of these figures is clear: the spread is great, and the outlier cases subject defendants to punitive damages that dwarf the corresponding compensatories.” *Id.* at 499-500. The size of the standard deviation compared to the mean supports the observation that ratios in the cases of jury award studies are dispersed. A graph of each pair of punitive and compensatory damages offers more insight into how scattered these figures are. See Theodore Eisenberg et al., *The Predictability of Punitive Damages*, 26 J. Legal Stud. 623 (1997); *infra* Section V.A (explaining scatter diagrams).

83. Random error is also called sampling error, chance error, or statistical error. Econometricians use the parallel concept of random disturbance terms. See Rubinfeld, *supra* note 21. Randomness and cognate terms have precise technical meanings; it is randomness in the technical sense that justifies the probability calculations behind standard errors, confidence intervals, and *p*-values (*supra* Section II.D, *infra* Sections IV.A–B). For a discussion of samples and populations, see *supra* Section II.B.

Thus, statistical inference typically involves tasks such as the following, which will be discussed in the rest of this guide.

- *Estimation.* A statistician draws a sample from a population (*supra* Section II.B) and estimates a parameter—that is, a numerical characteristic of the population. (The average value of a large group of claims is a parameter of perennial interest.) Random error will throw the estimate off the mark. The question is, by how much? The precision of an estimate is usually reported in terms of the standard error and a confidence interval.
- *Significance testing.* A “null hypothesis” is formulated—for example, that a parameter takes a particular value. Because of random error, an estimated value for the parameter is likely to differ from the value specified by the null—even if the null is right. (“Null hypothesis” is often shortened to “null.”) How likely is it to get a difference as large as, or larger than, the one observed in the data? This chance is known as a *p*-value. Small *p*-values argue against the null hypothesis. Statistical significance is determined by reference to the *p*-value; significance testing (also called hypothesis testing) is the technique for computing *p*-values and determining statistical significance.
- *Developing a statistical model.* Statistical inferences often depend 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 suit the occasion, and inferences based on these models will be 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 collection process may then require examination—how good is the analogy? If the model breaks down, that will bias the analysis.
- *Computing posterior probabilities.* Given the sample data, what is the probability of the null hypothesis? The question might be of direct interest to the courts, especially when translated into English; for example, the null hypothesis might be the innocence of the defendant in a criminal case. Posterior probabilities can be computed using a formula called Bayes’ rule. However, the computation often depends on prior beliefs about the statistical model and its parameters; such prior beliefs almost necessarily require subjective judgment. According to the frequentist theory of statistics,⁸⁴

84. The frequentist theory is also called objectivist, by contrast with the subjectivist version of Bayesian theory. In brief, frequentist methods treat probabilities as objective properties of the system being studied. Subjectivist Bayesians view probabilities as measuring subjective degrees of belief. See *infra* Section IV.D and Appendix, Section A, for discussion of the two positions. The Bayesian position is named after the Reverend Thomas Bayes (England, c. 1701–1761). His essay on the subject was published after his death: *An Essay Toward Solving a Problem in the Doctrine of Chances*, 53 Phil. Trans. Royal Soc’y London 370 (1763–1764). For discussion of the foundations and varieties of Bayesian and

prior probabilities rarely have meaning and neither do posterior probabilities.⁸⁵

Key ideas of estimation and testing will be illustrated by courtroom examples, with some complications omitted for ease of presentation and some details postponed (*see infra* Section V.D on statistical models, and the Appendix on the calculations).

The first example, on estimation, concerns the Nixon papers. Under the Presidential Recordings and Materials Preservation Act of 1974, Congress impounded Nixon's presidential papers after he resigned. Nixon sued, seeking compensation on the theory that the materials belonged to him personally. Courts ruled in his favor: Nixon was entitled to the fair market value of the papers, with the amount to be proved at trial.⁸⁶

The Nixon papers were stored in 20,000 boxes at the National Archives in Alexandria, Virginia. It was plainly impossible to value this entire population of material. Appraisers for the plaintiff therefore took a random sample of 500 boxes. (From this point on, details are simplified; thus, the example becomes somewhat hypothetical.) The appraisers determined the fair market value of each sample box. The average of the 500 sample values turned out to be \$2000. The standard deviation (*supra* Section III.E) of the 500 sample values was \$2200. Many boxes had low appraised values whereas some boxes were considered to be extremely valuable; this spread explains the large standard deviation.

A. Estimation

1. What estimator should be used?

With the Nixon papers, it is natural to use the average value of the 500 sample boxes to estimate the average value of all 20,000 boxes comprising the population.

other forms of statistical inference, see, e.g., Richard M. Royall, *Statistical Inference: A Likelihood Paradigm* (1997); James Berger, *The Case for Objective Bayesian Analysis*, 1 *Bayesian Analysis* 385 (2006), available at <http://ba.stat.cmu.edu/journal/2006/vol01/issue03/berger.pdf>; Stephen E. Fienberg, *Does It Make Sense to be an "Objective Bayesian"?* (Comment on Articles by Berger and by Goldstein), 1 *Bayesian Analysis* 429 (2006); 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 Fraesen ed., 1997); see also D.H. Kaye, *What Is Bayesianism?* in *Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism* (Peter Tillers & Eric Green eds., 1988), reprinted in 28 *Jurimetrics J.* 161 (1988) (distinguishing between "Bayesian probability," "Bayesian statistical inference," "Bayesian inference writ large," and "Bayesian decision theory").

85. Prior probabilities of repeatable events (but not hypotheses) can be defined within the frequentist framework. See *infra* note 122. When this happens, prior and posterior probabilities for these events are meaningful according to both schools of thought.

86. *Nixon v. United States*, 978 F.2d 1269 (D.C. Cir. 1992); *Griffin v. United States*, 935 F. Supp. 1 (D.D.C. 1995).

With the average value for each box having been estimated as \$2000, the plaintiff demanded compensation in the amount of

$$20,000 \times \$2,000 = \$40,000,000.$$

In more complex problems, statisticians may have to choose among several estimators. Generally, estimators that tend to make smaller errors are preferred; however, “error” might be quantified in more than one way. Moreover, the advantage of one estimator over another may depend on features of the population that are largely unknown, at least before the data are collected and analyzed. For complicated problems, professional skill and judgment may therefore be required when choosing a sample design and an estimator. In such cases, the choices and the rationale for them should be documented.

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 small amount, because of random error. The standard error gives the likely magnitude of this random error, with smaller standard errors indicating better estimates.⁸⁷ In our example of the Nixon papers, the standard error for the sample average can be computed from (1) the size of the sample—500 boxes—and (2) the standard deviation of the sample values; see *infra* Appendix. Bigger samples give estimates that are more precise. Accordingly, the standard error should go down as the sample size grows, although the rate of improvement slows as the sample gets bigger. (“Sample size” and “the size of the sample” just mean the number of items in the sample; the “sample average” is the average value of the items in the sample.) The standard deviation of the sample comes into play by measuring heterogeneity. The less heterogeneity in the values, the smaller the standard error. For example, if all the values were about the same, a tiny sample would give an accurate estimate. Conversely, if the values are quite different from one another, a larger sample would be needed.

With a random sample of 500 boxes and a standard deviation of \$2200, the standard error for the sample average is about \$100. The plaintiff’s total demand was figured as the number of boxes (20,000) times the sample average (\$2000). Therefore, the standard error for the total demand can be computed as 20,000 times the standard error for the sample average⁸⁸:

87. We distinguish between (1) the standard deviation of the sample, which measures the spread in the sample data and (2) the standard error of the sample average, which measures the likely size of the random error in the sample average. The standard error is often called the standard deviation, and courts generally use the latter term. See, e.g., *Castaneda v. Partida*, 430 U.S. 482 (1977).

88. We are assuming a simple random sample. Generally, the formula for the standard error must take into account the method used to draw the sample and the nature of the estimator. In fact, the Nixon appraisers used more elaborate statistical procedures. Moreover, they valued the material as of

$$20,000 \times \$100 = \$2,000,000.$$

How is the standard error to be interpreted? Just by the luck of the draw, a few too many high-value boxes may have come into the sample, in which case the estimate of \$40,000,000 is too high. Or, a few too many low-value boxes may have been drawn, in which case the estimate is too low. This is random error. The net effect of random error is unknown, because data are available only on the sample, not on the full population. However, the net effect is likely to be something close to the standard error of \$2,000,000. Random error throws the estimate off, one way or the other, by something close to the standard error. The role of the standard error is to gauge the likely size of the random error.

The plaintiff's argument may be open to a variety of objections, particularly regarding appraisal methods. However, the sampling plan is sound, as is the extrapolation from the sample to the population. And there is no need for a larger sample: The standard error is quite small relative to the total claim.

Random errors larger in magnitude than the standard error are commonplace. Random errors larger in magnitude than two or three times the standard error are unusual. Confidence intervals make these ideas more precise. Usually, a confidence interval for the population average is centered at the sample average; the desired confidence level is obtained by adding and subtracting a suitable multiple of the standard error. Statisticians who say that the population average falls within 1 standard error of the sample average will be correct about 68% of the time. Those who say "within 2 standard errors" will be correct about 95% of the time, and those who say "within 3 standard errors" will be correct about 99.7% of the time, and so forth. (We are assuming a large sample; the confidence levels correspond to areas under the normal curve and are approximations; the "population average" just means the average value of all the items in the population.⁸⁹) In summary,

- To get a 68% confidence interval, start at the sample average, then add and subtract 1 standard error.
- To get a 95% confidence interval, start at the sample average, then add and subtract twice the standard error.

1995, extrapolated backward to the time of taking (1974), and then added interest. The text ignores these complications.

89. See *infra* Appendix. The area under the normal curve between -1 and $+1$ is close to 68.3%. Likewise, the area between -2 and $+2$ is close to 95.4%. Many academic statisticians would use ± 1.96 SE 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 a few tenths of a percent. For simplicity, we use ± 1 SE for the 68% confidence level, and ± 2 SE for 95% confidence. The normal curve gives good approximations when the sample size is reasonably large; for small samples, other techniques should be used. See *infra* notes 106–07.

- To get a 99.7% confidence interval, start at the sample average, then add and subtract three times the standard error.

With the Nixon papers, the 68% confidence interval for plaintiff's total demand runs

$$\begin{aligned} &\text{from } \$40,000,000 - \$2,000,000 = \$38,000,000 \\ &\text{to } \$40,000,000 + \$2,000,000 = \$42,000,000. \end{aligned}$$

The 95% confidence interval runs

$$\begin{aligned} &\text{from } \$40,000,000 - (2 \times \$2,000,000) = \$36,000,000 \\ &\text{to } \$40,000,000 + (2 \times \$2,000,000) = \$44,000,000. \end{aligned}$$

The 99.7% confidence interval runs

$$\begin{aligned} &\text{from } \$40,000,000 - (3 \times \$2,000,000) = \$34,000,000 \\ &\text{to } \$40,000,000 + (3 \times \$2,000,000) = \$46,000,000. \end{aligned}$$

To write this more compactly, we abbreviate standard error as SE. Thus, 1 SE is one standard error, 2 SE is twice the standard error, and so forth. With a large sample and an estimate like the sample average, a 68% confidence interval is the range

$$\text{estimate} - 1 \text{ SE to estimate} + 1 \text{ SE.}$$

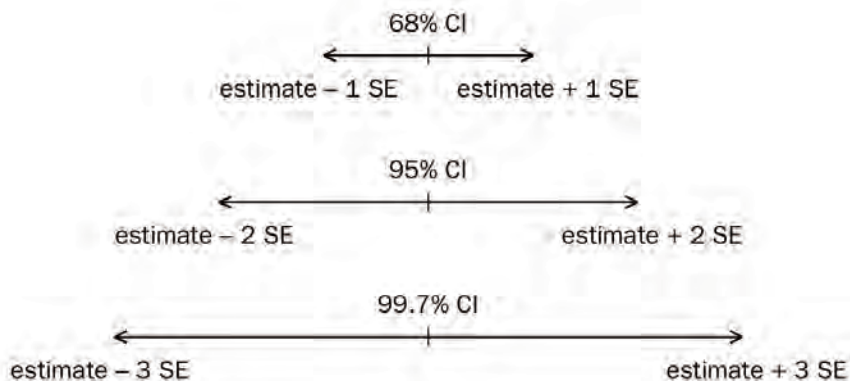
A 95% confidence interval is the range

$$\text{estimate} - 2 \text{ SE to estimate} + 2 \text{ SE.}$$

The 99.7% confidence interval is the range

$$\text{estimate} - 3 \text{ SE to estimate} + 3 \text{ SE.}$$

For a given sample size, increased confidence can be attained only by widening the interval. The 95% confidence level is the most popular, but some authors use 99%, and 90% is seen on occasion. (The corresponding multipliers on the SE are about 2, 2.6, and 1.6, respectively; *see infra* Appendix.) The phrase “margin of error” generally means twice the standard error. In medical journals, “confidence interval” is often abbreviated as “CI.”



The main point is that an estimate based on a sample will differ from the exact population value, because of random error. The standard error gives the likely size of the random error. If the standard error is small, random error probably has little effect. If the standard error is large, the estimate may be seriously wrong. Confidence intervals are a technical refinement, and bias is a separate issue to consider (*infra* Section IV.A.4).

3. How big should the sample be?

There is no easy answer to this sensible question. Much depends on the level of error that is tolerable and the nature of the material being sampled. Generally, increasing the size of the sample will reduce the level of random error (“sampling error”). Bias (“nonsampling error”) cannot be reduced that way. Indeed, beyond some point, large samples are harder to manage and more vulnerable to non-sampling error. To reduce bias, the researcher must improve the design of the study or use a statistical model more tightly linked to the data collection process.

If the material being sampled is heterogeneous, random error will be large; a larger sample will be needed to offset the heterogeneity (*supra* Section IV.A.1). A pilot sample may be useful to estimate heterogeneity and determine the final sample size. Probability samples require some effort in the design phase, and it will rarely be sensible to draw a sample with fewer than, say, two or three dozen items. Moreover, with such small samples, methods based on the normal curve (*supra* Section IV.A.2) will not apply.

Population size (i.e., the number of items in the population) usually has little bearing on the precision of estimates for the population average. This is surprising. On the other hand, population size has a direct bearing on estimated totals. Both points are illustrated by the Nixon papers (*see supra* Section IV.A.2 and *infra* Appendix). To be sure, drawing a probability sample from a large population may

involve a lot of work. Samples presented in the courtroom have ranged from 5 (tiny) to 1.7 million (huge).⁹⁰

4. What are the technical difficulties?

To begin with, “confidence” is a term of art. The confidence level indicates the percentage of the time that intervals from repeated samples would cover the true value. The confidence level does not express the chance that repeated estimates would fall into the confidence interval.⁹¹ With the Nixon papers, the 95% confidence interval should not be interpreted as saying that 95% of all random samples will produce estimates in the range from \$36 million to \$44 million. Moreover, the confidence level does not give the probability that the unknown parameter lies within the confidence interval.⁹² For example, the 95% confidence level should not be translated to a 95% probability that the total value of the papers is in the range from \$36 million to \$44 million. According to the frequentist theory of statistics, probability statements cannot be made about population characteristics: Probability statements apply to the behavior of samples. That is why the different term “confidence” is used.

The next point to make is that for a given confidence level, a narrower interval indicates a more precise estimate, whereas a broader interval indicates less

90. See *Lebrilla v. Farmers Group, Inc.*, No. 00–CC–017185 (Cal. Super. Ct., Orange County, Dec. 5, 2006) (preliminary approval of settlement), a class action lawsuit on behalf of plaintiffs who were insured by Farmers and had automobile accidents. Plaintiffs alleged that replacement parts recommended by Farmers did not meet specifications: Small samples were used to evaluate these allegations. At the other extreme, it was proposed to adjust Census 2000 for undercount and overcount by reviewing a sample of 1.7 million persons. See Brown et al., *supra* note 29, at 353.

91. Opinions reflecting this misinterpretation include *In re Silicone Gel Breast Implants Prods. Liab. Litig.*, 318 F. Supp. 2d 879, 897 (C.D. Cal. 2004) (“a margin of error between 0.5 and 8.0 at the 95% confidence level . . . means that 95 times out of 100 a study of that type would yield a relative risk value somewhere between 0.5 and 8.0.”); *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). The more technically correct statement in the *Silicone Gel* case, for example, would be that “the confidence interval of 0.5 to 8.0 means that the relative risk in the population could fall within this wide range and that in roughly 95 times out of 100, random samples from the same population, the confidence intervals (however wide they might be) would include the population value (whatever it is).”

92. See, e.g., Freedman et al., *supra* note 12, at 383–86; *infra* Section IV.B.1. Consequently, it is misleading to suggest that “[a] 95% confidence interval means that there is a 95% probability that the ‘true’ relative risk falls within the interval” or that “the probability that the true value was . . . within two standard deviations of the mean . . . would be 95 percent.” *DeLuca v. Merrell Dow Pharms., Inc.*, 791 F. Supp. 1042, 1046 (D.N.J. 1992), *aff’d*, 6 F.3d 778 (3d Cir. 1993); *SmithKline Beecham Corp. v. Apotex Corp.*, 247 F. Supp. 2d 1011, 1037 (N.D. Ill. 2003), *aff’d on other grounds*, 403 F.3d 1331 (Fed. Cir. 2005).

precision.⁹³ A high confidence level with a broad interval means very little, but a high confidence level for a small interval is impressive, indicating that the random error in the sample estimate is low. For example, take a 95% confidence interval for a damage claim. An interval that runs from \$34 million to \$44 million is one thing, but –\$10 million to \$90 million is something else entirely. Statements about confidence without mention of an interval are practically meaningless.⁹⁴

Standard errors and confidence intervals are often derived from statistical models for the process that generated the data. The model usually has parameters—numerical constants describing the population from which samples were drawn. When the values of the parameters are not known, the statistician must work backward, using the sample data to make estimates. That was the case here.⁹⁵ Generally, the chances needed for statistical inference are computed from a model and estimated parameter values.

If the data come from a probability sample or a randomized controlled experiment (*supra* Sections II.A–B), 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. With the Nixon papers, the appraisers drew a random sample, and that justified the statistical

93. In *Cimino v. Raymark Industries, Inc.*, 751 F. Supp. 649 (E.D. Tex. 1990), *rev'd*, 151 F.3d 297 (5th Cir. 1998), the district court drew certain random samples from more than 6000 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 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.* This is backward. 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.

94. In *Hilao v. Estate of Marcos*, 103 F.3d 767 (9th Cir. 1996), for example, “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 [9541 facially valid] claims filed.’” *Id.* at 782. 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. The computation 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. For the same reason, it is meaningless to testify, as an expert did in *Ayyad v. Sprint Spectrum, L.P.*, No. RG03-121510 (Cal. Super. Ct., Alameda County) (transcript, May 28, 2008, at 730), that a simple regression equation is trustworthy because the coefficient of the explanatory variable has “an extremely high indication of reliability to more than 99% confidence level.”

95. With the Nixon papers, one parameter is the average value of all 20,000 boxes, and another parameter is the standard deviation of the 20,000 values. These parameters can be used to approximate the distribution of the sample average. See *infra* Appendix. Regression models and their parameters are discussed *infra* Section V and in Rubinfeld, *supra* note 21.

calculations—if not the appraised values themselves. In many contexts, the choice of an appropriate statistical model is less than obvious. When a model does not fit the data collection process, estimates and standard errors will not be probative.

Standard errors and confidence intervals generally ignore systematic errors such as selection bias or nonresponse bias (*supra* Sections II.B.1–2). For example, after reviewing studies to see whether a particular drug caused birth defects, a court observed that mothers of children with birth defects may be more likely to remember taking a drug during pregnancy than mothers with normal children. 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 groups would ignore this bias, as would the confidence interval.⁹⁶

B. Significance Levels and Hypothesis Tests

1. What Is the *p*-value?

In 1969, Dr. Benjamin Spock came to trial in the U.S. District Court for Massachusetts. The charge was conspiracy to violate the Military Service Act. The jury was drawn from a panel of 350 persons selected by the clerk of the court. The panel included only 102 women—substantially less than 50%—although a majority of the eligible jurors in the community were female. The shortfall in women was especially poignant in this case: “Of all defendants, Dr. Spock, who had given wise and welcome advice on child-rearing to millions of mothers, would have liked women on his jury.”⁹⁷

Can the shortfall in women be explained by the mere play of random chance? To approach the problem, a statistician would formulate and test a null hypothesis. Here, the null hypothesis says that the panel is like 350 persons drawn at random from a large population that is 50% female. The expected number of women drawn would then be 50% of 350, which is 175. The observed number of women is 102. The shortfall is $175 - 102 = 73$. How likely is it to find a disparity this large or larger, between observed and expected values? The probability is called *p*, or the *p*-value.

96. *Brock v. Merrell Dow Pharms., Inc.*, 874 F.2d 307, 311–12 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989). 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. This is wrong. Even if the sampling error were nonexistent—which would be the case if one could interview every woman who had a child during 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.

97. Hans Zeisel, *Dr. Spock and the Case of the Vanishing Women Jurors*, 37 U. Chi. L. Rev. 1 (1969). Zeisel’s reasoning was different from that presented in this text. The conviction was reversed on appeal without reaching the issue of jury selection. *United States v. Spock*, 416 F.2d 165 (1st Cir. 1965).

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. In the example, p turns out to be essentially zero. The discrepancy between the observed and the expected is far too large to explain by random chance. Indeed, even if the panel had included 155 women, the p -value would only be around 0.02, or 2%.⁹⁸ (If the population is more than 50% female, p will be even smaller.) In short, the jury panel was nothing like a random sample from the community.

Large p -values indicate that a disparity can easily be explained by the play of chance: The data fall within the range likely to be produced by chance variation. On the other hand, if p is very small, something other than chance must be involved: The data are far away from the values expected under the null hypothesis. Significance testing often seems to involve multiple negatives. This is because a statistical test is an argument by contradiction.

With the Dr. Spock example, the null hypothesis asserts that the jury panel is like a random sample from a population that is 50% female. The data contradict this null hypothesis because the disparity between what is observed and what is expected (according to the null) is too large to be explained as the product of random chance. In a typical jury discrimination case, small p -values help a defendant appealing a conviction by showing that the jury panel is not like a random sample from the relevant population; large p -values hurt. In the usual employment context, small p -values help plaintiffs who complain of discrimination—for example, by showing that a disparity in promotion rates is too large to be explained by chance; conversely, large p -values would be consistent with the defense argument that the disparity is just due to chance.

Because p is calculated by assuming that the null hypothesis is correct, p does not give the chance that the null is true. The p -value merely gives the chance of getting evidence against the null hypothesis as strong as or stronger than the evidence at hand. Chance affects the data, not the hypothesis. According to the frequency theory of statistics, there is no meaningful way to assign a numerical probability to the null hypothesis. The correct interpretation of the p -value can therefore be summarized in two lines:

p is the probability of extreme data given the null hypothesis.
 p is not the probability of the null hypothesis given extreme data.⁹⁹

98. With 102 women out of 350, the p -value is about $2/10^{15}$, where 10^{15} is 1 followed by 15 zeros, that is, a quadrillion. See *infra* Appendix for the calculations.

99. Some opinions present 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 1000 that the phenomenon was attributable to chance”); *Nat’l Abortion Fed. v. Ashcroft*, 330 F. Supp. 2d 436 (S.D.N.Y. 2004), *aff’d in part*, 437 F.3d 278 (2d Cir. 2006), *vacated*, 224 Fed. App’x. 88 (2d Cir. 2007) (“According to Dr. Howell, . . . a ‘P value’ of 0.30 . . . indicates that there is a thirty percent probability that the results of the . . . [s]tudy were merely due to chance alone.”). Such statements confuse the probability of the

To recapitulate the logic of significance testing: If p is small, the observed data are far from what is expected under the null hypothesis—too far to be readily explained by the operations of chance. That discredits the null hypothesis.

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; other times they will not be. Sometimes a difference of two standard errors will imply a p -value of about 5%; other times it will not. In general, the p -value depends on the model, the size of the sample, and the sample statistics.

2. Is a difference statistically significant?

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, as statisticians say, 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. Statistical significance is determined by comparing p to a preset value, called the significance level.¹⁰⁰ The null hypothesis is rejected when p falls below this level.

In practice, statistical analysts typically use levels of 5% and 1%.¹⁰¹ The 5% 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. An unexplained reference to highly significant results probably means that p is less

kind of outcome observed, which is computed under some model of chance, with the probability that chance is the explanation for the outcome—the “transposition fallacy.”

Instances of the transposition fallacy in criminal cases are collected in David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* §§ 12.8.2(b) & 14.1.2 (2d ed. 2011). In *McDaniel v. Brown*, 130 S. Ct. 665 (2010), for example, a DNA analyst suggested that a random match probability of 1/3,000,000 implied a .000033 probability that the DNA was not the source of the DNA found on the victim’s clothing. See David H. Kaye, “False But Highly Persuasive”: *How Wrong Were the Probability Estimates in McDaniel v. Brown?* 108 Mich. L. Rev. First Impressions 1 (2009).

100. 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 a false rejection of the null hypothesis (also called a false positive, a false alarm, or a Type I error). 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.

101. 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 5% and 0.3% when the statistic is normally distributed. The Court’s standard deviation is our standard error.

than 1%. These levels of 5% and 1% have become icons of science and the legal process. In truth, however, such levels are at best useful conventions.

Because the term “significant” is merely a label for a certain kind of p -value, significance is subject to the same limitations as the underlying p -value. Thus, significant differences may be evidence that something besides random error is at work. 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 is negligible—there is no reason to worry about statistical significance.¹⁰²

It is easy to mistake the p -value for the probability of the null hypothesis given the data (*supra* Section IV.B.1). Likewise, if results are significant at the 5% level, it is tempting to conclude that the null hypothesis has only a 5% chance of being correct.¹⁰³ 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 does not 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.

3. Tests or interval estimates?

How can a highly significant difference be practically insignificant? The reason is simple: p depends not only on the magnitude of the effect, but also on the sample size (among other things). With a huge sample, even a tiny effect will be

102. *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.”); *United States v. Henderson*, 409 F.3d 1293, 1306 (11th Cir. 2005) (regardless of statistical significance, excluding law enforcement officers from jury service does not have a large enough impact on the composition of grand juries to violate the Jury Selection and Service Act); *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.”).

103. *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”); *Adams v. Ameritech Serv., Inc.*, 231 F.3d 414, 424 (7th Cir. 2000) (“Two standard deviations is normally enough to show that it is extremely unlikely (. . . less than a 5% probability) that the disparity is due to chance”); *Magistrini v. One Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 605 n.26 (D.N.J. 2002) (a “statistically significant . . . study shows that there is only 5% probability that an observed association is due to chance.”); *cf.* *Giles v. Wyeth, Inc.*, 500 F. Supp. 2d 1048, 1056 (S.D. Ill. 2007) (“While [plaintiff] admits that a p -value of .15 is three times higher than what scientists generally consider statistically significant—that is, a p -value of .05 or lower—she maintains that this “represents 85% certainty, which meets any conceivable concept of preponderance of the evidence.”).

highly significant.¹⁰⁴ For example, suppose that a company hires 52% of male job applicants and 49% of female applicants. With a large enough sample, a statistician could compute an impressively small p -value. This p -value would confirm that the difference does not result from chance, but it would not convert a trivial difference (52% versus 49%) into a substantial one.¹⁰⁵ In short, the p -value does not measure the strength or importance of an association.

A “significant” effect can be small. Conversely, an effect that is “not significant” can be large. By inquiring into the magnitude of an effect, courts can avoid being misled by p -values. To focus attention on more substantive concerns—the size of the effect and the precision of the statistical analysis—interval estimates (e.g., confidence intervals) may be more valuable than tests. Seeing a plausible range of values for the quantity of interest helps describe the statistical uncertainty in the estimate.

4. *Is the sample statistically significant?*

Many a sample has been praised for its statistical significance or blamed for its lack thereof. Technically, this makes little sense. Statistical significance is about the difference between observations and expectations. Significance therefore applies to statistics computed from the sample, but not to the sample itself, and certainly not to the size of the sample. Findings can be statistically significant. Differences can be statistically significant (*supra* Section IV.B.2). Estimates can be statistically significant (*infra* Section V.D.2). By contrast, samples can be representative or unrepresentative. They can be chosen well or badly (*supra* Section II.B.1). They can be large enough to give reliable results or too small to bother with (*supra* Section IV.A.3). But samples cannot be “statistically significant,” if this technical phrase is to be used as statisticians use it.

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:

104. See *supra* Section IV.B.2. Although some opinions seem to equate small p -values with “gross” or “substantial” disparities, most courts recognize the need to decide whether the underlying sample statistics reveal that a disparity is large. *E.g.*, *Washington v. People*, 186 P.3d 594 (Colo. 2008) (jury selection).

105. *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); *Washington*, 186 P.2d 594 (although a jury selection practice that reduced the representation of “African-Americans [from] 7.7 percent of the population [to] 7.4 percent of the county’s jury panels produced a highly statistically significant disparity, the small degree of exclusion was not constitutionally significant.”).

1. The null hypothesis is true.
2. The null is false—but, by chance, the data happened to be of the kind expected under the null.

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 be declared.¹⁰⁶ This chance depends on the size of the effect and the size of the sample. Discerning subtle differences requires large samples; small samples may fail to detect substantial differences.

When a study with low power fails to show a significant effect, the results may therefore be more fairly described as inconclusive than negative. The proof is weak because power is low. 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 nothing much to be found.¹⁰⁷

2. What about small samples?

For simplicity, the examples of statistical inference discussed here (*supra* Sections IV.A–B) were based on large samples. Small samples also can provide useful

106. More precisely, power is the probability of rejecting the null hypothesis when the alternative hypothesis (*infra* Section IV.C.5) is right. Typically, this probability will depend on the values of unknown parameters, as well as the preset significance level α . The power can be computed for any value of α and any choice of parameters satisfying the alternative hypothesis. See *infra* Appendix for an example. Frequentist hypothesis testing keeps the risk of a false positive to a specified level (such as $\alpha = 5\%$) and then tries to maximize power.

Statisticians usually denote power by the Greek letter beta (β). However, some authors use β to denote the probability of *accepting* the null hypothesis when the alternative hypothesis is true; this usage is fairly standard in epidemiology. Accepting the null hypothesis when the alternative holds true is a false negative (also called a Type II error, a missed signal, or a false acceptance of the null hypothesis).

The chance of a false negative may be computed from the power. 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. The argument is fallacious, because α and β do not give the probabilities of the null and alternative hypotheses; see *supra* Sections IV.B.1–2; *supra* note 34. See also 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).

107. Some formal procedures (meta-analysis) are available to aggregate results across studies. See, e.g., *In re Bextra and Celebrex Marketing Sales Practices and Prod. Liab. Litig.*, 524 F. Supp. 2d 1166, 1174, 1184 (N.D. Cal. 2007) (holding that “[a] meta-analysis of all available published and unpublished randomized clinical trials” of certain pain-relief medicine was admissible). In principle, the power of the collective results will be greater than the power of each study. However, these procedures have their own weakness. See, e.g., Richard A. Berk & David A. Freedman, *Statistical Assumptions as Empirical Commitments*, in *Punishment and Social Control: Essays in Honor of Sheldon Messinger* 235, 244–48 (T.G. Blomberg & S. Cohen eds., 2d ed. 2003); Michael Oakes, *Statistical Inference: A Commentary for the Social and Behavioral Sciences* (1986); Diana B. Petitti, *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis Methods for Quantitative Synthesis in Medicine* (2d ed. 2000).

information. Indeed, when confidence intervals and p -values can be computed, the interpretation is the same with small samples as with large ones.¹⁰⁸ The concern with small samples is not that they are beyond the ken of statistical theory, but that

1. The underlying assumptions are hard to validate.
2. Because approximations based on the normal curve generally cannot be used, confidence intervals may be difficult to compute for parameters of interest. Likewise, p -values may be difficult to compute for hypotheses of interest.¹⁰⁹
3. Small samples may be unreliable, with large standard errors, broad confidence intervals, and tests having low power.

3. *One tail or two?*

In many cases, a statistical test can be done either one-tailed or two-tailed; the second method often produces a p -value twice as big as the first method. The methods are easily explained with a hypothetical example. Suppose we toss a coin 1000 times and get 532 heads. The null hypothesis to be tested asserts that the coin is fair. If the null is correct, the chance of getting 532 or more heads is 2.3%. That is a one-tailed test, whose p -value is 2.3%. To make a two-tailed test, the statistician computes the chance of getting 532 or more heads—or $500 - 32 = 468$ heads or fewer. This is 4.6%. In other words, the two-tailed p -value is 4.6%. Because small p -values are evidence against the null hypothesis, the one-tailed test seems to produce stronger evidence than its two-tailed counterpart. However, the advantage is largely illusory, as the example suggests. (The two-tailed test may seem artificial, but it offers some protection against possible artifacts resulting from multiple testing—the topic of the next section.)

Some courts and commentators have argued for one or the other type of test, but a rigid rule is not required if significance levels are used as guidelines rather than as mechanical rules for statistical proof.¹¹⁰ One-tailed tests often make it

108. Advocates sometimes contend that samples are “too small to allow for meaningful statistical analysis,” *United States v. New York City Bd. of Educ.*, 487 F. Supp. 2d 220, 229 (E.D.N.Y. 2007), and courts often look to the size of samples from earlier cases to determine whether the sample data before them are admissible or convincing. *Id.* at 230; *Timmerman v. U.S. Bank*, 483 F.3d 1106, 1116 n.4 (10th Cir. 2007). However, a meaningful statistical analysis yielding a significant result can be based on a small sample, and reliability does not depend on sample size alone (see *supra* Section IV.A.3, *infra* Section V.C.1). Well-known small-sample techniques include the sign test and Fisher’s exact test. *E.g.*, Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* 154–56, 339–41 (2d ed. 2001); see generally E.L. Lehmann & H.J.M. d’Abrera, *Nonparametrics* (2d ed. 2006).

109. With large samples, approximate inferences (e.g., based on the central limit theorem, see *infra* Appendix) may be quite adequate. These approximations will not be satisfactory for small samples.

110. See, e.g., *United States v. State of Delaware*, 93 Fair Empl. Prac. Cas. (BNA) 1248, 2004 WL 609331, *10 n.4 (D. Del. 2004). According to formal statistical theory, the choice between one

easier to reach a threshold such as 5%, at least in terms of appearance. However, if we recognize that 5% is not 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.

4. How many tests have been done?

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. To illustrate the point, consider the problem of deciding whether a coin is biased. The probability that a fair coin will produce 10 heads when tossed 10 times is $(1/2)^{10} = 1/1024$. Observing 10 heads in the first 10 tosses, therefore, would be strong evidence that the coin is biased. Nonetheless, if a fair coin is tossed a few thousand times, it is likely that at least one string of ten consecutive heads will appear. Ten heads in the first ten tosses means one thing; a run of ten heads somewhere along the way to a few thousand tosses of a coin means quite another. A test—looking for a run of ten heads—can be repeated too often.

Artifacts from multiple testing are commonplace. Because research that fails to uncover significance often is not published, reviews of the literature may produce an unduly large number of studies finding statistical significance.¹¹¹ Even a single researcher may examine so many different relationships that a few will achieve statistical significance by mere happenstance. Almost any large dataset—even pages from a table of random digits—will contain some unusual pattern that can be uncovered by 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.

There are statistical methods for dealing with multiple looks at the data, which permit the calculation of meaningful p -values in certain cases.¹¹² 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 models before arriving at the one considered the most satisfactory (*see infra* Section V on regression models). In these situations, courts should not be overly impressed with

tail or two can sometimes be made by considering the exact form of the alternative hypothesis (*infra* Section IV.C.5). *But see* Freedman et al., *supra* note 12, at 547–50. One-tailed tests at the 5% level are viewed as weak evidence—no weaker standard is commonly used in the technical literature. One-tailed tests are also called one-sided (with no pejorative intent); two-tailed tests are two-sided.

111. *E.g.*, Philippa J. Easterbrook et al., *Publication Bias in Clinical Research*, 337 *Lancet* 867 (1991); John P.A. Ioannidis, *Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials*, 279 *JAMA* 281 (1998); 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).

112. *See, e.g.*, Sandrine Dudoit & Mark J. van der Laan, *Multiple Testing Procedures with Applications to Genomics* (2008).

claims that estimates are significant. Instead, they should be asking how analysts developed their models.¹¹³

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. 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.

In *Mapes Casino, Inc. v. Maryland Casualty Co.*,¹¹⁴ 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 sought to quantify its loss from theft. It argued that employees were using an intermediary to cash in chips at other casinos. The casino established that over an 18-month period, the win percentage at its craps tables was 6%, compared to an expected value of 20%. The statistics proved that *something* was wrong at the craps tables—the discrepancy was too big to explain as the product of random chance. But the court was not convinced by plaintiff's alternative hypothesis. The court pointed to other possible explanations (Runyonesque activities such as skimming, scamming, and crossroading) that might have accounted for the discrepancy without implicating the suspect employees.¹¹⁵ In short, rejection of the null hypothesis does not leave the proffered alternative hypothesis as the only viable explanation for the data.¹¹⁶

113. Intuition may suggest that the more variables included in the model, the better. However, this idea often turns out 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. See authorities cited, *supra* note 21.

114. 290 F. Supp. 186 (D. Nev. 1968).

115. *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. 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 “nontraditional” jobs insufficient to rebut prima facie case of gender discrimination); cf. *supra* Section II.A (problem of confounding).

116. *E.g.*, *Coleman v. Quaker Oats Co.*, 232 F.3d 1271, 1283 (9th Cir. 2000) (a disparity with a *p*-value of “3 in 100 billion” did not demonstrate age discrimination because “Quaker never contends that the disparity occurred by chance, just that it did not occur for discriminatory reasons. When other pertinent variables were factored in, the statistical disparity diminished and finally disappeared.”).

D. Posterior Probabilities

Standard errors, p -values, and significance tests are common techniques for assessing random error. These procedures rely on sample data and are justified in terms of the operating characteristics of statistical procedures.¹¹⁷ However, frequentist statisticians generally will not compute the probability that a particular hypothesis is correct, given the data.¹¹⁸ For example, a frequentist may postulate that a coin is fair: There is a 50-50 chance of landing heads, and successive tosses are independent. This is viewed as an empirical statement—potentially falsifiable—about the coin. It is easy to calculate the chance that a fair coin will turn up heads in the next 10 tosses: The answer (*see supra* Section IV.C.4) is 1/1024. Therefore, observing 10 heads in a row brings into serious doubt the initial hypothesis of fairness.

But what of the converse probability: If the coin does land heads 10 times, what is the chance that it is fair?¹¹⁹ 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. In the frequentist theory of inference, such postulates are untenable: Probabilities are objective features of the situation that specify the chances of events or effects, not hypotheses or causes.

By contrast, in the Bayesian approach, probabilities represent subjective degrees of belief about hypotheses or causes rather than objective facts about observations. The observer must quantify beliefs about the chance that the coin is unfair to various degrees—in advance of seeing the data.¹²⁰ 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.

117. Operating characteristics include the expected value and standard error of estimators, probabilities of error for statistical tests, and the like.

118. In speaking of “frequentist statisticians” or “Bayesian statisticians,” we do not mean to suggest that all statisticians fall on one side of the philosophical divide or the other. These are archetypes. Many practicing statisticians are pragmatists, using whatever procedure they think is appropriate for the occasion, and not concerning themselves greatly with foundational issues.

119. We call this a converse probability because it is of the form $P(H_0 | \text{data})$ rather than $P(\text{data} | H_0)$; an equivalent phrase, “inverse probability,” also is used. Treating $P(\text{data} | H_0)$ as if it were the converse probability $P(H_0 | \text{data})$ is the transposition fallacy. For example, most U.S. senators are men, but 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* notes 98, 102. The frequentist p -value, $P(\text{data} | H_0)$, is generally not a good approximation to the Bayesian $P(H_0 | \text{data})$; the latter includes considerations of power and base rates.

120. For example, let p be the unknown probability that the coin lands heads. What is the chance that p exceeds 0.1? 0.6? The Bayesian statistician must be prepared to answer 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); Freedman, *supra* note 84; David Kaye, *The Laws of Probability and the Law of the Land*, 47 U. Chi. L. Rev. 34 (1979).

Prior probabilities can be updated, using Bayes' rule, given data on how the coin actually falls. (The Appendix explains the rule.) In short, a Bayesian statistician can compute posterior probabilities for various hypotheses about the coin, given the data. These posterior probabilities quantify the statistician's confidence in the hypothesis that a coin is fair.¹²¹ Although such posterior probabilities relate directly to hypotheses of legal interest, they are necessarily subjective, for they reflect not just the data but also the subjective prior probabilities—that is, degrees of belief about hypotheses formulated prior to obtaining data.

Such analyses have rarely been used in court, and the question of their forensic value has been aired primarily in the academic literature. Some statisticians favor Bayesian methods, and some commentators have proposed using these methods in some kinds of cases.¹²² The frequentist view of statistics is more conventional; subjective Bayesians are a well-established minority.¹²³

121. 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 D.H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 Cornell L. Rev. 54 (1987).

122. See David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* §§ 12.8.5, 14.3.2 (2d ed. 2010); David H. Kaye, *Rounding Up the Usual Suspects: A Legal and Logical Analysis of DNA Database Trawls*, 87 N.C. L. Rev. 425 (2009). In addition, as indicated in the Appendix, Bayes' rule is crucial in solving certain problems involving conditional probabilities of related events. For example, if the proportion of women with breast cancer in a region is known, along with the probability that a mammogram of an affected woman will be positive for cancer and that the mammogram of an unaffected woman will be negative, then one can compute the numbers of false-positive and false-negative mammography results that would be expected to arise in a population-wide screening program. Using Bayes' rule to diagnose a specific patient, however, is more problematic, because the prior probability that the patient has breast cancer may not equal the population proportion. Nevertheless, to overcome the tendency to focus on a test result without considering the "base rate" at which a condition occurs, a diagnostician can apply Bayes' rule to plausible base rates before making a diagnosis. Finally, Bayes' rule also is valuable as a device to explicate the meaning of concepts such as error rates, probative value, and transposition. See, e.g., David H. Kaye, *The Double Helix and the Law of Evidence* (2010); Wigmore, *supra*, § 7.3.2; David H. Kaye & Jonathan J. Koehler, *The Misquantification of Probative Value*, 27 Law & Hum. Behav. 645 (2003).

123. "Objective Bayesians" use Bayes' rule without eliciting prior probabilities from subjective beliefs. One strategy is to use preliminary data to estimate the prior probabilities and then apply Bayes' rule to that empirical distribution. This "empirical Bayes" procedure avoids the charge of subjectivism at the cost of departing from a fully Bayesian framework. With ample data, however, it can be effective and the estimates or inferences can be understood in frequentist terms. Another "objective" approach is to use "noninformative" priors that are supposed to be independent of all data and prior beliefs. However, the choice of such priors can be questioned, and the approach has been attacked by frequentists and subjective Bayesians. E.g., Joseph B. Kadane, *Is "Objective Bayesian Analysis" Objective, Bayesian, or Wise?*, 1 Bayesian Analysis 433 (2006), available at <http://ba.stat.cmu.edu/journal/2006/vol01/issue03/kadane.pdf>; Jon Williamson, *Philosophies of Probability*, in *Philosophy of Mathematics* 493 (Andrew Irvine ed., 2009) (discussing the challenges to objective Bayesianism).

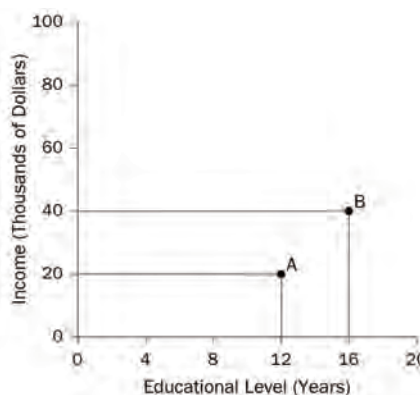
V. Correlation and Regression

Regression models are used by many social scientists to infer causation from association. Such models have been offered in court to prove disparate impact in discrimination cases, to estimate damages in antitrust actions, and for many other purposes. Sections V.A, V.B, and V.C cover some preliminary material, showing how scatter diagrams, correlation coefficients, and regression lines can be used to summarize relationships between variables.¹²⁴ Section V.D explains the ideas and some of the pitfalls.

A. Scatter Diagrams

The relationship between two variables can be graphed in a scatter diagram (also called a scatterplot or scattergram). We begin with data on income and education for a sample of 178 men, ages 25 to 34, residing in Kansas.¹²⁵ Each person in the sample corresponds to one dot in the diagram. As indicated in Figure 5, the horizontal axis shows education, and the vertical axis shows income. Person A completed 12 years of schooling (high school) and had an income of \$20,000. Person B completed 16 years of schooling (college) and had an income of \$40,000.

Figure 5. Plotting a scatter diagram. The horizontal axis shows educational level and the vertical axis shows income.

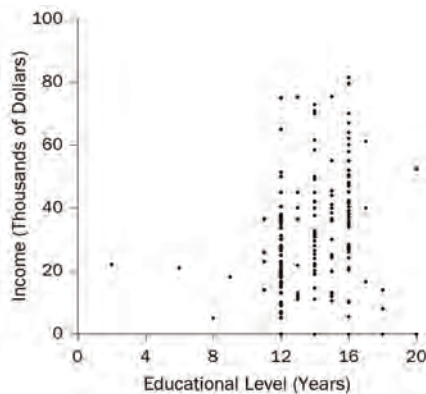


124. The focus is on simple linear regression. See also Rubinfeld, *supra* note 21, and the Appendix, *infra*, and Section II, *supra*, for further discussion of these ideas with an emphasis on econometrics.

125. These data are from a public-use CD, Bureau of the Census, U.S. Department of Commerce, for the March 2005 Current Population Survey. Income and education are self-reported. Income is censored at \$100,000. For additional details, see Freedman et al., *supra* note 12, at A-11. Both variables in a scatter diagram have to be quantitative (with numerical values) rather than qualitative (nonnumerical).

Figure 6 is the scatter diagram for the Kansas 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.

Figure 6. Scatter diagram for income and education: men ages 25 to 34 in Kansas.



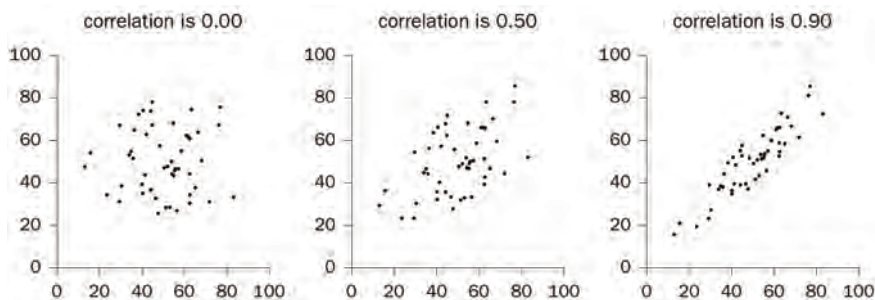
B. Correlation Coefficients

Two variables are positively correlated when their values tend to go up or down together, such as income and education in Figure 5. The correlation coefficient (usually denoted by the letter r) is a single number that reflects the sign of an association and its strength. Figure 7 shows r for three scatter diagrams: In the first, there is no association; in the second, the association is positive and moderate; in the third, the association is positive and strong.

A correlation coefficient of 0 indicates no linear association between the variables. The maximum value for the coefficient is $+1$, indicating a perfect linear relationship: The dots in the scatter diagram fall on a straight line that slopes up. 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 illustrate the idea. 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 that slopes down.

Weak associations are the rule in the social sciences. In Figure 5, the correlation between income and education is about 0.4. The correlation between college grades and first-year law school grades is under 0.3 at most law schools, while the

Figure 7. The correlation coefficient measures the sign of a linear association and its strength.



correlation between LSAT scores and first-year grades is generally about 0.4.¹²⁶ The correlation between heights of fraternal twins is about 0.5. By contrast, the correlation between heights of identical twins is about 0.95.

1. *Is the association linear?*

The correlation coefficient has a number of limitations, to be considered in turn. The correlation coefficient is designed to measure linear association. Figure 8 shows a strong nonlinear pattern with a correlation close to zero. The correlation coefficient is of limited use with nonlinear data.

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 9 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. If there are extreme outliers in the data, the correlation coefficient is unlikely to be meaningful.

3. *Does a confounding variable influence the coefficient?*

The correlation coefficient measures the association between two variables. Researchers—and the courts—are usually more interested in causation. Causation is not the same as association. The association between two variables may be driven by a lurking variable that has been omitted from the analysis (*supra*

126. Lisa Anthony Stilwell et al., Predictive Validity of the LSAT: A National Summary of the 2001–2002 Correlation Studies 5, 8 (2003).

Figure 8. The scatter diagram shows a strong nonlinear association with a correlation coefficient close to zero. The correlation coefficient only measures the degree of linear association.

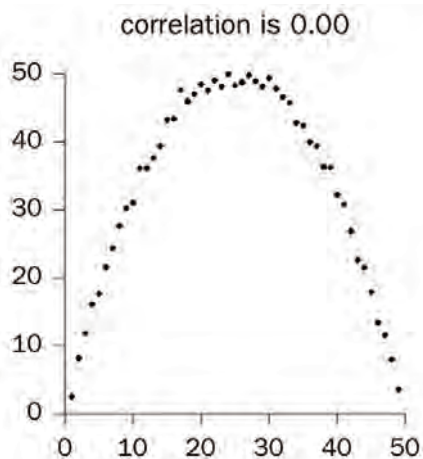
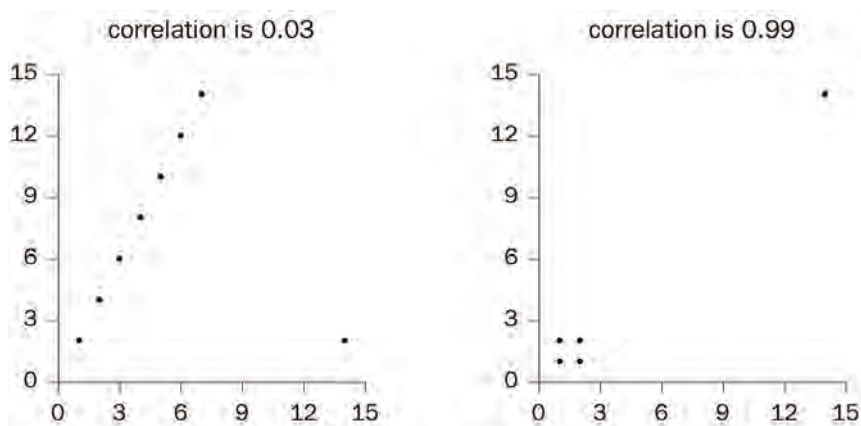


Figure 9. The correlation coefficient can be distorted by outliers.



Section II.A). For an easy example, there is an association between shoe size and vocabulary among schoolchildren. However, learning more words does not cause the feet to get bigger, and swollen feet do not make children more articulate. In this case, the lurking variable is easy to spot—age. In more realistic examples, the lurking variable is harder to identify.¹²⁷

127. Green et al., *supra* note 13, Section IV.C, provides one such example.

In statistics, lurking variables are called confounders or confounding variables. Association often does reflect causation, but a large correlation coefficient is not enough to warrant causal inference. A large value of r only means that the dependent variable marches in step with the independent one: Possible reasons include causation, confounding, and coincidence. Multiple regression is one method that attempts to deal with confounders (*infra* Section V.D).¹²⁸

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 Kansas sample is shown in Figure 10. The height of the line estimates the average income for a given educational level. For example, the average income for people with 8 years of education is estimated at \$21,100, indicated by the height of the line at 8 years. The average income for people with 16 years of education is estimated at \$34,700.

Figure 10. The regression line for income on education and its estimates.

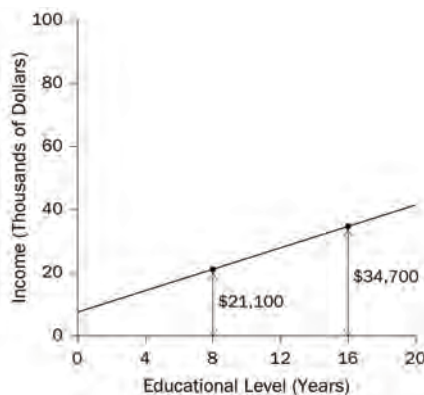
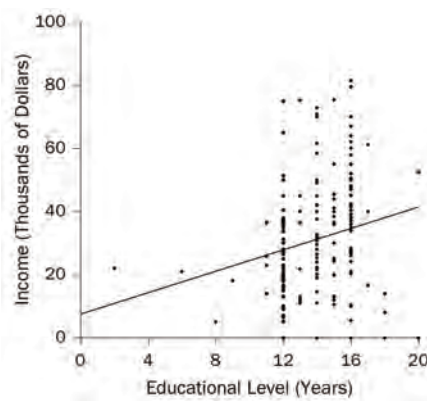


Figure 11 combines the data in Figures 5 and 10: it shows the scatter diagram for income and education, with the regression line superimposed. The line shows the average trend of income as education increases. Thus, the regression line indicates the extent to which a change in one variable (income) is associated with a change in another variable (education).

128. See also Rubinfeld, *supra* note 21. The difference between experiments and observational studies is discussed *supra* Section II.B.

Figure 11. 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 intercept and slope. Often, the slope is the more interesting statistic. In Figure 11, the slope is \$1700 per year. On average, each additional year of education is associated with an additional \$1700 of income. Next, the intercept is \$7500. This is an estimate of the average income for (hypothetical) persons with zero years of education.¹²⁹ Figure 10 suggests this estimate may not be especially good. In general, estimates based on the regression line become less trustworthy as we move away from the bulk of the data.

The slope of the regression line has the same limitations as the correlation coefficient: (1) The slope may be misleading if the relationship is strongly non-linear and (2) the slope may be affected by confounders. With respect to (1), the slope of \$1700 per year in Figure 10 presents each additional year of education as having the same value, but some years of schooling surely are worth more and

129. The regression line, like any straight line, has an equation of the form $y = a + bx$. Here, a is the intercept (the value of y when $x = 0$), and b is the slope (the change in y per unit change in x). In Figure 9, the intercept of the regression line is \$7500 and the slope is \$1700 per year. The line estimates an average income of \$34,700 for people with 16 years of education. This may be computed from the intercept and slope as follows:

$$\$7500 + (\$1700 \text{ per year}) \times 16 \text{ years} = \$7500 + \$22,200 = \$34,700.$$

The slope b is the same anywhere along the line. Mathematically, that is what distinguishes straight lines from other curves. If the association is negative, the slope will be negative too. The slope is like the grade of a road, and it is negative if the road goes downhill. The intercept is like the starting elevation of a road, and it is computed from the data so that the line goes through the center of the scatter diagram, rather than being generally too high or too low.

others less. With respect to (2), the association between education and income is no doubt causal, but there are other factors to consider, including family background. Compared to individuals who did not graduate from high school, people with college degrees usually come from richer and better educated families. Thus, college graduates have advantages besides education. As statisticians might say, the effects of family background are confounded with the effects of education. Statisticians often use the guarded phrases “on average” and “associated with” when talking about the slope of the regression line. This is because the slope has limited utility when it comes to making causal inferences.

2. *What is the unit of analysis?*

If association between characteristics of individuals is of interest, these characteristics should be measured on individuals. Sometimes individual-level data are not to be had, but rates or averages for groups are available. “Ecological” correlations are computed from such rates or averages. These correlations generally overstate the strength of an association. For example, average income and average education can be determined for men living in each state and in Washington, D.C. The correlation coefficient for these 51 pairs of averages turns out to be 0.70. However, states do not go to school and do not earn incomes. People do. The correlation for income and education for men in the United States is only 0.42. The correlation for state averages overstates the correlation for individuals—a common tendency for ecological correlations.¹³⁰

Ecological analysis is often seen in cases claiming dilution in voting strength of minorities. 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.¹³¹ The first requirement is compactness; the second and third define polarized voting.

130. Correlations are computed from the March 2005 Current Population Survey for men ages 25–64. Freedman et al., *supra* note 12, at 149. The ecological correlation uses only the average figures, but within each state there is a lot of spread about the average. The ecological correlation smoothes away this individual variation. Cf. Green et al., *supra* note 13, Section II.B.4 (suggesting that ecological studies of exposure and disease are “far from conclusive” because of the lack of data on confounding variables (a much more general problem) as well as the possible aggregation bias described here); David A. Freedman, *Ecological Inference and the Ecological Fallacy*, in 6 *Int'l Encyclopedia of the Social and Behavioral Sciences* 4027 (Neil J. Smelser & Paul B. Baltes eds., 2001).

131. 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 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 secrecy of the ballot box means that polarized voting cannot be directly observed. Instead, plaintiffs in voting rights cases rely on ecological regression, with scatter diagrams, correlations, and regression lines to estimate voting behavior by groups and demonstrate polarization. The unit of analysis typically is the precinct. For each precinct, public records can be used to determine the percentage of registrants in each demographic group of interest, as well as the percentage of the total vote for each candidate—by voters from all demographic groups combined. Plaintiffs' burden is to determine the vote by each demographic group separately.

Figure 12 shows how the argument unfolds. Each point in the scatter diagram represents data for one 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. The regression line is plotted too. The slope would be interpreted as the difference between the white turnout rate and the black turnout rate for the white candidate. Furthermore, the intercept would be interpreted as the black turnout rate for the white candidate.¹³² The validity of such estimates is contested in the statistical literature.¹³³

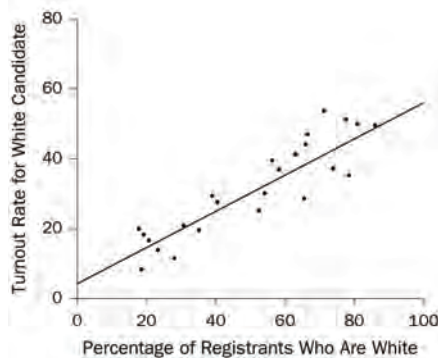
the Voting Rights Act. *E.g.*, *Johnson v. De Grandy*, 512 U.S. 997, 1011 (1994) (“*Gingles* . . . clearly declined to hold [these factors] sufficient in combination, either in the sense that a court’s examination of relevant circumstances was complete once the three factors were found to exist, or in the sense that the three in combination necessarily and in all circumstances demonstrated dilution.”).

132. 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. The intercept of the line in Figure 11 is 4%, and the slope is 0.52. 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.

133. For further discussion of ecological regression in this context, see D. James Greiner, *Ecological Inference in Voting Rights Act Disputes: Where Are We Now, and Where Do We Want to Be?*, 47 *Jurimetrics J.* 115 (2007); Bernard Grofman & Chandler Davidson, *Controversies in Minority Voting: The Voting Rights Act in Perspective* (1992); Stephen P. Klein & David A. Freedman, *Ecological Regression in Voting Rights Cases*, 6 *Chance* 38 (Summer 1993). 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., *Cottier v. City of Martin*, 445 F.3d 1113, 1118 (8th Cir. 2006) (ecological regression is one of the “proven approaches to evaluating elections”); Bruce M. Clarke & Robert Timothy Reagan, *Fed. Judicial Ctr., Redistricting Litigation: An Overview of Legal, Statistical, and Case-Management Issues* (2002); Greiner, *supra*, at 117, 121. 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); *Johnson v. Hamrick*, 296 F.3d 1065, 1080 n.4 (11th Cir. 2002) (“as a general rule, homogenous precinct analysis may be more reliable than ecological regression.”). However, there are problems with both methods. See, e.g., Greiner, *supra*, at 123–39 (arguing that homogeneous precinct analysis is fundamentally flawed and that courts need to be more discerning in dealing with ecological regression).

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 ques-

Figure 12. 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.



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

D. Statistical Models

Statistical models are widely used in the social sciences and in litigation. 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 tax funds.¹³⁴ Other models purport to lift the veil of secrecy from the ballot box, enabling the experts to determine how minority groups have voted—a crucial step in voting rights litigation (*supra* Section V.C). This section discusses the statistical logic of regression models.

A regression model attempts to combine the values of certain variables (the independent variables) 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 be interpreted as showing the effects of changing the corresponding variables. This is justified in some situations, as the next example demonstrates.

tion, but efforts to sustain racially motivated 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).

134. See Brown et al., *supra* note 29; *supra* note 89.

Hooke's law (named after Robert Hooke, England, 1653–1703) describes how a spring stretches in response to a load: Strain is proportional to stress. To verify Hooke's law experimentally, a physicist will make 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 develop a regression model for these data:

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

The error term, denoted by the Greek letter epsilon ϵ , is needed because measured length will not be exactly equal to $a + b \times \text{weight}$. If nothing else, measurement error must be reckoned with. The model takes ϵ as “random error”—behaving like draws 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 the potential errors in the box is assumed to be zero.

Equation (1) has two parameters, a and b . These constants of nature characterize the behavior of the spring: a is length under no load, and b is elasticity (the increase in length per unit increase in weight). By way of numerical illustration, suppose a is 400 and b is 0.05. If the weight is 1, the length of the spring is expected to be

$$400 + 0.05 = 400.05.$$

If the weight is 3, the expected length is

$$400 + 3 \times 0.05 = 400 + 0.15 = 400.15.$$

In either case, the actual length will differ from expected, by a random error ϵ .

In standard statistical terminology, the ϵ 's for different observations on the spring are assumed to be independent and identically distributed, with a mean of zero. Take the ϵ 's for the first two observations. Independence means that the chances for the second ϵ do not depend on outcomes for the first. If the errors are like draws made at random with replacement from a box of tickets, as we assumed earlier, that box will not change from one draw to the next—independence. “Identically distributed” means that the chance behavior of the two ϵ 's is the same: They are drawn at random from the same box. (See *infra* Appendix for additional discussion.)

The parameters a and b in equation (1) are not directly observable, but they can be estimated by the method of least squares.¹³⁵ Statisticians often denote esti-

135. It might seem that a is observable; after all, we can measure the length of the spring with no load. However, the measurement is subject to error, so we observe not a , but $a + \epsilon$. See equation (1). The parameters a and b can be estimated, even estimated very well, but they cannot be observed directly. The least squares estimates of a and b are the intercept and slope of the regression

mates by hats. Thus, \hat{a} is the estimate for a , and \hat{b} is the estimate for b . The values of \hat{a} and \hat{b} are chosen to minimize the sum of the squared prediction errors. These errors are also called residuals. They measure the difference between the actual length of the spring and the predicted length, the latter being $\hat{a} + \hat{b} \times \text{weight}$:

$$\text{actual length} = \hat{a} + \hat{b} \times \text{weight} + \text{residual}. \quad (2)$$

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.¹³⁶ In short, the statistical model corresponds rather closely to the empirical phenomenon.

Equation (1) is a statistical model for the data, with unknown parameters a and b . The error term ϵ is not observable. The model is a theory—and a good one—about how the data are generated. By contrast, equation (2) is a regression equation that is fitted to the data: The intercept \hat{a} , the slope \hat{b} , and the residual can all be computed from the data. The results are useful because \hat{a} is a good estimate for a , and \hat{b} is a good estimate for b . (Similarly, the residual is a good approximation to ϵ .) Without the theory, these estimates would be less useful. Is there a theoretical model behind the data processing? Is the model justifiable? These questions can be critical when it comes to making statistical inferences from the data.

In social science applications, statistical models often are invoked without an independent theoretical basis. We give an example involving salary discrimination in the Appendix.¹³⁷ The main ideas of such regression modeling can be captured in a hypothetical exchange between a plaintiff seeking to prove salary discrimination and a company denying the allegation. Such a dialog might proceed as follows:

1. Plaintiff argues that the defendant company pays male employees more than females, which establishes a prima facie case of discrimination.
2. The company responds that the men are paid more because they are better educated and have more experience.
3. Plaintiff refutes the company's theory by fitting a regression equation that includes a particular, presupposed relationship between salary (the dependent variable) and some measures of education and experience. Plaintiff's expert reports that even after adjusting for differences in education and

line. See *supra* Section V.C.1; Freedman et al., *supra* note 12, at 208–10. The method of least squares was developed by Adrien-Marie Legendre (France, 1752–1833) and Carl Friedrich Gauss (Germany, 1777–1855) to fit astronomical orbits.

136. This is the Gauss model for measurement error. See Freedman et al., *supra* note 12, at 450–52.

137. The Reference Guide to Multiple Regression in this manual describes a comparable example.

experience in this specific manner, men earn more than women. This remaining difference in pay shows discrimination.

4. The company argues that the difference could be the result of chance, not discrimination.
5. Plaintiff replies that because the coefficient for gender in the model is statistically significant, chance is not a good explanation for the data.¹³⁸

In step 3, the three explanatory variables are education (years of schooling completed), experience (years with the firm), and a dummy variable for gender (1 for men and 0 for women). These are supposed to predict salaries (dollars per year). The equation is a formal analog of Hooke's law (equation 1). According to the model, an employee's salary is determined as if by computing

$$a + (b \times \text{education}) + (c \times \text{experience}) + (d \times \text{gender}), \quad (3)$$

and then adding an error ϵ drawn at random from a box of tickets.¹³⁹ The parameters a , b , c , and d , are estimated from the data by the method of least squares.

In step 5, the estimated coefficient d for the dummy variable turns out to be positive and statistically significant and is offered as evidence of disparate impact. Men earn more than women, even after adjusting for differences in background factors that might affect productivity. This showing depends on many assumptions built into the model.¹⁴⁰ Hooke's law—equation (1)—is relatively easy to test experimentally. For the salary discrimination model, validation would be difficult. When expert testimony relies on statistical models, the court may well inquire, what are the assumptions behind the model, and why do they apply to the case at hand? It might then be important to distinguish between two situations:

- The nature of the relationship between the variables is known and regression is being used to make quantitative estimates of parameters in that relationship, or
- The nature of the relationship is largely unknown and regression is being used to determine the nature of the relationship—or indeed whether any relationship exists at all.

138. In some cases, the p -value has been interpreted as the probability that defendants are innocent of discrimination. However, as noted earlier, 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 for gender is zero (see *supra* Section IV.B, *infra* Appendix, Section D.2). Therefore, even if we grant the model, a p -value less than 50% does not demonstrate a preponderance of the evidence against the null hypothesis.

139. Expression (3) is the expected value for salary, given the explanatory variables (education, experience, gender). The error term is needed to account for deviations from expected: Salaries are not going to be predicted very well by linear combinations of variables such as education and experience.

140. See *infra* Appendix.

Regression 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 an issue worth exploration.

In employment discrimination cases, and other contexts too, a wide variety of models can be used. This is only to be expected, because the science does not dictate specific equations. In a strongly contested case, each side will have its own model, presented by its own expert. The experts will reach opposite conclusions about discrimination. The dialog might continue with an exchange about which model is better. Although statistical assumptions 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 example, skill levels or prior evaluations.¹⁴¹ Another model may be challenged because it includes tainted variables reflecting past discriminatory behavior by the firm.¹⁴² The court must decide which model—if either—fits the occasion.¹⁴³

The frequency with which regression models are used is no guarantee that they are the best choice for any particular problem. Indeed, from one perspective, a regression or other statistical model may seem to be a marvel of mathematical rigor. From another perspective, the model is a set of assumptions, supported only by the say-so of the testifying expert. Intermediate judgments are also possible.¹⁴⁴

141. *E.g.*, *Bazemore v. Friday*, 478 U.S. 385 (1986); *In re Linerboard Antitrust Litig.*, 497 F. Supp. 2d 666 (E.D. Pa. 2007).

142. *E.g.*, *McLaurin v. Nat'l R.R. Passenger Corp.*, 311 F. Supp. 2d 61, 65–66 (D.D.C. 2004) (holding that the inclusion of two allegedly tainted variables was reasonable in light of an earlier consent decree).

143. *E.g.*, *Chang v. Univ. of R.I.*, 606 F. Supp. 1161, 1207 (D.R.I. 1985) (“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).

144. *See, e.g.*, David W. Peterson, *Reference Guide on Multiple Regression*, 36 *Jurimetrics J.* 213, 214–15 (1996) (review essay); *see supra* note 21 for references to a range of academic opinion. More recently, some investigators have turned to graphical models. However, these models have serious weaknesses of their own. *See, e.g.*, David A. Freedman, *On Specifying Graphical Models for Causation, and the Identification Problem*, 26 *Evaluation Rev.* 267 (2004).

Appendix

A. Frequentists and Bayesians

The mathematical theory of probability consists of theorems derived from axioms and definitions. Mathematical reasoning is seldom controversial, but there may be disagreement as to how the theory should be applied. For example, statisticians may differ on the interpretation of data in specific applications. Moreover, there are two main schools of thought about the foundations of statistics: frequentist and Bayesian (also called objectivist and subjectivist).¹⁴⁵

Frequentists see probabilities as empirical facts. When a fair coin is tossed, the probability of heads is 1/2; if the experiment is repeated a large number of times, 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.¹⁴⁶ Generally, if a chance experiment can be repeated, the relative frequency of an event approaches (in the long run) its probability. By contrast, a Bayesian considers probabilities as representing not facts but degrees of belief: In whole or in part, probabilities are subjective.

Statisticians of both schools use conditional probability—that is, the probability of one event given that another has occurred. For example, 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 $P(\text{HH} \mid \text{at least one H}) = 1/3$. More generally, the probability of an event C is denoted $P(C)$; the conditional probability of D given C is written as $P(D \mid C)$.

Two events C and D are independent if the conditional probability of D given that C occurs is equal to the conditional probability of D given that C does not occur. Statisticians use “ $\sim C$ ” to denote the event that C does not occur. Thus C and D are independent if $P(D \mid C) = P(D \mid \sim C)$. If C and D are independent, then the probability that both occur is equal to the product of the probabilities:

$$P(C \text{ and } D) = P(C) \times P(D). \quad (\text{A1})$$

145. But see *supra* note 123 (on “objective Bayesianism”).

146. Probabilities may be estimated from relative frequencies, but probability itself is a subtler idea. For example, suppose a computer prints out a sequence of 10 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%. There are difficulties in both the subjectivist and objectivist positions. See Freedman, *supra* note 84.

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

$$P(C \text{ and } D) = P(C) \times P(D | C). \quad (\text{A2})$$

This is the multiplication rule for dependent events.

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. We turn now to Bayes' rule. If H_0 and H_1 are two hypotheses¹⁴⁷ that govern the probability of an event A , a Bayesian can use the multiplication rule (A2) to find that

$$P(A \text{ and } H_0) = P(A | H_0)P(H_0) \quad (\text{A3})$$

and

$$P(A \text{ and } H_1) = P(A | H_1)P(H_1). \quad (\text{A4})$$

Moreover,

$$P(A) = P(A \text{ and } H_0) + P(A \text{ and } H_1). \quad (\text{A5})$$

The multiplication rule (A2) also shows that

$$P(H_1 | A) = \frac{P(A \text{ and } H_1)}{P(A)}. \quad (\text{A6})$$

We use (A4) to evaluate $P(A \text{ and } H_1)$ in the numerator of (A6), and (A3), (A4), and (A5) to evaluate $P(A)$ in the denominator:

$$P(H_1 | A) = \frac{P(A | H_1)P(H_1)}{P(A | H_0)P(H_0) + P(A | H_1)P(H_1)}. \quad (\text{A7})$$

This is a special case of Bayes' rule. It yields the conditional probability of hypothesis H_0 given that event A has occurred.

For a stylized example in a criminal case, H_0 is the hypothesis that blood found at the scene of a crime came from a person other than the defendant; H_1 is the hypothesis that the blood came from the defendant; A is the event that blood from the crime scene and blood from the defendant are both type A. Then $P(H_0)$ is the prior probability of H_0 , based on subjective judgment, while $P(H_0 | A)$ is the posterior probability—updated from the prior using the data.

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

Type A blood occurs in 42% of the population. So $P(A|H_0) = 0.42$.¹⁴⁸ Because the defendant has type A blood, $P(A|H_1) = 1$. Suppose the prior probabilities are $P(H_0) = P(H_1) = 0.5$. According to (A7), the posterior probability that the blood is from the defendant is

$$P(H_1|A) = \frac{1 \times 0.5}{0.42 \times 0.5 + 1 \times 0.5} = 0.70. \quad (\text{A8})$$

Thus, the data increase the likelihood that the blood is the defendant's. The probability went up from the prior value of $P(H_1) = 0.50$ to the posterior value of $P(H_1|A) = 0.70$.

More generally, H_0 and H_1 refer to parameters in a statistical model. For a stylized example in an employment discrimination case, H_0 asserts equal selection rates in a population of male and female applicants; H_1 asserts that the selection rates are not equal; A is the event that a test statistic exceeds 2 in absolute value. In such situations, the Bayesian proceeds much as before. However, the frequentist computes $P(A|H_0)$, and rejects H_0 if this probability falls below 5%. Frequentists have to stop there, because they view $P(H_0|A)$ as poorly defined at best. In their setup, $P(H_0)$ and $P(H_1)$ rarely make sense, and these prior probabilities are needed to compute $P(H_1|A)$: See *supra* equation (A7).

Assessing probabilities, conditional probabilities, and independence is not entirely straightforward, either for frequentists or Bayesians. Inquiry into the basis for expert judgment may be useful, and casual assumptions about independence should be questioned.¹⁴⁹

B. The Spock Jury: Technical Details

The rest of this Appendix provides some technical backup for the examples in Sections IV and V, *supra*. We begin with the *Spock* jury case. On the null hypothesis, a sample of 350 people was drawn at random from a large population that was 50% male and 50% female. The number of women in the sample follows the binomial distribution. For example, the chance of getting exactly 102 women in the sample is given by the binomial formula¹⁵⁰

$$\frac{n!}{j! \times (n-j)!} f^j (1-f)^{n-j}. \quad (\text{A9})$$

148. Not all statisticians would accept the identification of a population frequency with $P(A|H_0)$. Indeed, H_0 has been translated into a hypothesis that the true donor has been selected from the population at random (i.e., in a manner that is uncorrelated with blood type). This step needs justification. See *supra* note 123.

149. For problematic assumptions of independence in litigation, see, e.g., *Wilson v. State*, 803 A.2d 1034 (Md. 2002) (error to admit multiplied probabilities in a case involving two deaths of infants in same family); 1 McCormick, *supra* note 2, § 210; see also *supra* note 29 (on census litigation).

150. The binomial formula is discussed in, e.g., Freedman et al., *supra* note 12, at 255–61.

In the formula, n stands for the sample size, and so $n = 350$; and $j = 102$. The f is the fraction of women in the population; thus, $f = 0.50$. The exclamation point denotes factorials: $1! = 1$, $2! = 2 \times 1 = 2$, $3! = 3 \times 2 \times 1 = 6$, and so forth. The chance of 102 women works out to 10^{-15} . In the same way, we can compute the chance of getting 101 women, or 100, or any other particular number. The chance of getting 102 women or fewer is then computed by addition. The chance is $p = 2 \times 10^{-15}$, as reported *supra* note 98. This is very bad news for the null hypothesis.

With the binomial distribution given by (9), the expected the number of women in the sample is

$$nf = 350 \times 0.5 = 175. \quad (\text{A10})$$

The standard error is

$$\sqrt{n} \times \sqrt{f \times (1-f)} = \sqrt{350} \times \sqrt{0.5 \times 0.5} = 9.35. \quad (\text{A11})$$

The observed value of 102 is nearly 8 SEs below the expected value, which is a lot of SEs.

Figure 13 shows the probability histogram for the number of women in the sample.¹⁵¹ The graph is drawn so that the area between two values is proportional to the chance that the number of women will fall in that range. For example, take the rectangle over 175; its base covers the interval from 174.5 to 175.5. The area of this rectangle is 4.26% of the total area. So the chance of getting exactly 175 women is 4.26%. Next, take the range from 165 to 185 (inclusive): 73.84% of the area falls into this range. This means there is a 73.84% chance that the number of women in the sample will be in the range from 165 to 185 (inclusive).

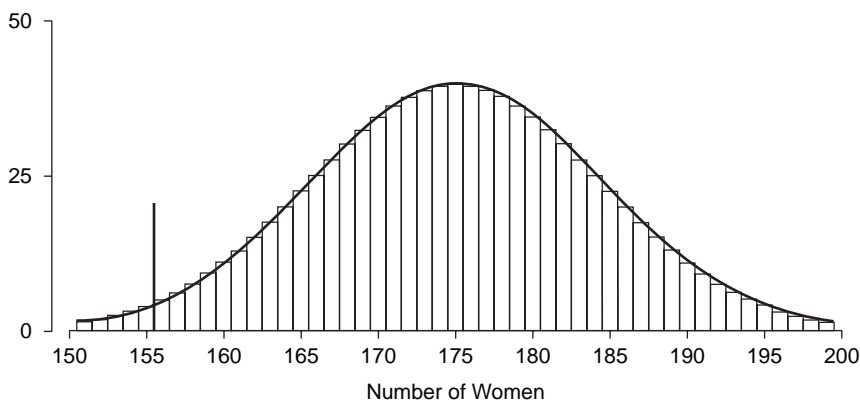
According to a fundamental theorem in statistics (the central limit theorem), the histogram follows the normal curve.¹⁵² Figure 13 shows the curve for comparison: The normal curve is almost indistinguishable from the top of the histogram. For a numerical example, suppose the jury panel had included 155 women. On the null hypothesis, there is about a 1.85% chance of getting 155 women or fewer. The normal curve gives 1.86%. The error is nil. Ordinarily, we would just report $p = 2\%$, as in the text (*supra* Section IV.B.1).

Finally, we consider power. Suppose we reject the null hypothesis when the number of women in the sample is 155 or less. Let us assume a particular alternative hypothesis that quantifies the degree of discrimination against women: The jury panel is selected at random from a population that is 40% female, rather than 50%. Figure 14 shows the probability histogram for the number of women, but now the histogram is computed according to the alternative hypothesis. Again,

151. Probability histograms are discussed in, e.g., *id.* at 310–13.

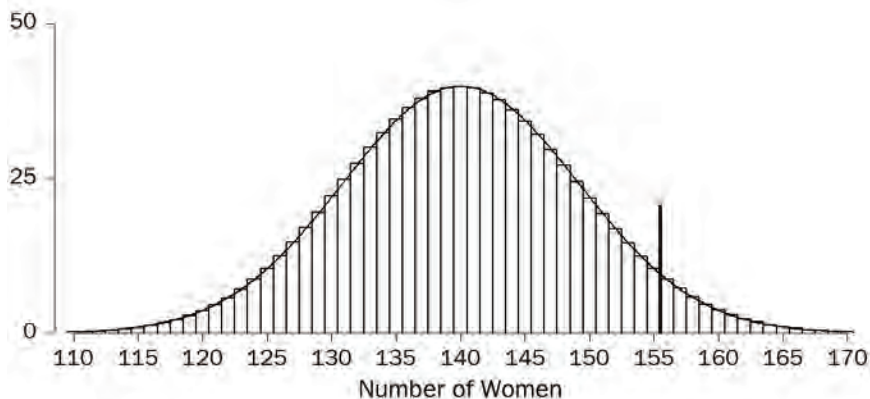
152. The central limit theorem is discussed in, e.g., *id.* at 315–27.

Figure 13. Probability histogram for the number of women in a random sample of 350 people drawn from a large population that is 50% female and 50% male. The normal curve is shown for comparison. About 2% of the area under the histogram is to the left of 155 (marked by a heavy vertical line).



Note: The vertical line is placed at 155.5, and so the area to the left of it includes the rectangles over 155, 154, . . . ; the area represents the chance of getting 155 women or fewer. Cf. Freedman et al., *supra* note 12, at 317. The units on the vertical axis are “percent per standard unit”; cf. *id.* at 80, 315.

Figure 14. Probability histogram for the number of women in a random sample of 350 people drawn from a large population that is 40% female and 60% male. The normal curve is shown for comparison. The area to the left of 155 (marked by a heavy vertical line) is about 95%.



the histogram follows the normal curve. About 95% of the area is to the left of 155, and so power is about 95%. The area can be computed exactly by using the binomial distribution, or to an excellent approximation using the normal curve.

Figures 13 and 14 have the same shape: The central limit theorem is at work. However, the histograms are centered differently. Figure 13 is centered at 175, according to requirements of the null hypothesis. Figure 14 is centered at 140, because the alternative hypothesis is used to determine the center, not the null hypothesis. Thus, 155 is well to the left of center in Figure 13, and well to the right in Figure 14: The figures have different centers. The main point of Figures 13 and 14 is that chances can often be approximated by areas under the normal curve, justifying the large-sample theory presented *supra* Sections IV.A–B.

C. The Nixon Papers: Technical Details

With the Nixon papers, the population consists of 20,000 boxes. A random sample of 500 boxes is drawn and each sample box is appraised. Statistical theory enables us to make some precise statements about the behavior of the sample average.

- The expected value of the sample average equals the population average. Even more tersely, the sample average is an unbiased estimate of the population average.
- The standard error for the sample average equals

$$\sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}. \quad (\text{A12})$$

In (A12), the N stands for the size of the population, which is 20,000; and n stands for the size of the sample, which is 500. The first factor in (A12), with the square root, is the finite sample correction factor. Here, as in many other such examples, the correction factor is so close to 1 that it can safely be ignored. (This is why the size of population usually has no bearing on the precision of the sample average as an estimator for the population average.) Next, σ is the population standard deviation. This is unknown, but it can be estimated by the sample standard deviation, which is \$2200. The SE for the sample mean is therefore estimated from the data as $\$2200/\sqrt{500}$, which is nearly \$100. Plaintiff's total claim is 20,000 times the sample average. The SE for the total claim is therefore $20,000 \times \$100 = \$2,000,000$. (Here, the size of the population comes into the formula.)

With a large sample, the probability histogram for the sample average follows the normal curve quite closely. That is a consequence of the central limit theorem. The center of the histogram is the population average. The SE is given by (A12), and is about \$100.

- What is the chance that the sample average differs from the population average by 1 SE or less? This chance is equal to the area under the probability histogram within 1 SE of average, which by the central limit theorem is almost equal to the area under the standard normal curve between -1 and 1 ; that normal area is about 68%.
- What is the chance that the sample average differs from the population average by 2 SE or less? By the same reasoning, this chance is about equal to the area under the standard normal curve between -2 and 2 , which is about 95%.
- What is the chance that the sample average differs from the population average by 3 SE or less? This chance is about equal to the area under the standard normal curve between -3 and 3 , which is about 99.7%.

To sum up, the probability histogram for the sample average is centered at the population average. The spread is given by the standard error. The histogram follows the normal curve. That is why confidence levels can be based on the standard error, with confidence levels read off the normal curve—for estimators that are essentially unbiased, and obey the central limit theorem (*supra* Section IV.A.2, Appendix Section B).¹⁵³ These large-sample methods generally work for sums, averages, and rates, although much depends on the design of the sample.

More technically, the normal curve is the density of a normal distribution. The standard normal density has mean equal to 0 and standard error equal to 1. Its equation is

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

where $e = 2.71828\dots$ and $\pi = 3.14159\dots$. This density can be rescaled to have any desired mean and standard error. The resulting densities are the famous “normal curves” or “bell-shaped curves” of statistical theory. In Figure 12, the density is scaled to match the probability histogram in terms of the mean and standard error; likewise in Figure 13.

D. A Social Science Example of Regression: Gender Discrimination in Salaries

1. The regression model

To illustrate social science applications of the kind that might be seen in litigation, Section V referred to a stylized example on salary discrimination. A particular

153. See, e.g., *id.* at 409–24. On the standard deviation, see *supra* Section III.E; Freedman et al., *supra* note 12, at 67–72. The finite sample correction factor is discussed in *id.* at 367–70.

regression model was used to predict salaries (dollars per year) of employees in a firm. It had three explanatory variables: education (years of schooling completed), experience (years with the firm), and a dummy variable for gender (1 for men and 0 for women). The regression equation is

$$\text{salary} = a + b \times \text{education} + c \times \text{experience} + d \times \text{gender} + \varepsilon. \quad (\text{A13})$$

Equation (A13) is a statistical model for the data, with unknown parameters a , b , c , and d . Here, a is the intercept and the other parameters are regression coefficients. The ε at the end of the equation is an unobservable error term. In the right-hand side of (A3) and similar expressions, by convention, the multiplications are done before the additions.

As noted in Section V, the equation is a formal analog of Hooke's law (1). According to the model, an employee's salary is determined as if by computing

$$a + b \times \text{education} + c \times \text{experience} + d \times \text{gender} \quad (\text{A14})$$

and then adding an error ε drawn at random from a box of tickets. Expression (A14) is the expected value for salary, given the explanatory variables (education, experience, gender). The error term is needed to account for deviations from expected: Salaries are not going to be predicted very well by linear combinations of variables such as education and experience.

The parameters are estimated from the data using least squares. If the estimated coefficient for the dummy variable turns out to be positive and statistically significant, that would be evidence of disparate impact. Men earn more than women, even after adjusting for differences in background factors that might affect productivity. Suppose the estimated equation turns out as follows:

$$\begin{aligned} \text{predicted salary} &= \$7100 + \$1300 \times \text{education} + \$2200 \\ &\quad \times \text{experience} + \$700 \times \text{gender}. \end{aligned} \quad (\text{A15})$$

According to (A15), the estimated value for the intercept a in (A14) is \$7100; the estimated value for the coefficient b is \$1300, and so forth. According to equation (A15), every extra year of education is worth \$1300. Similarly, every extra year of experience is worth \$2200. And, most important, the company gives men a salary premium of \$700 over women with the same education and experience.

A male employee with 12 years of education (high school) and 10 years of experience, for example, would have a predicted salary of

$$\begin{aligned} & \$7100 + \$1300 \times 12 + \$2200 \times 10 + \$700 \times 1 \\ & = \$7100 + \$15,600 + \$22,000 + \$700 = \$45,400. \end{aligned} \quad (\text{A16})$$

A similarly situated female employee has a predicted salary of only

$$\begin{aligned}
 & \$7100 + \$1300 \times 12 + \$2200 \times 10 + \$700 \times 0 \\
 & = \$7100 + \$15,600 + \$22,000 + \$0 = \$44,700.
 \end{aligned}
 \tag{A17}$$

Notice the impact of the gender variable in the model: \$700 is added to equation (A16), but not to equation (A17).

A major step in proving discrimination is showing that the estimated coefficient of the gender variable—\$700 in the numerical illustration—is statistically significant. This showing depends on the assumptions built into the model. Thus, each extra year of education is assumed to be worth the same across all levels of experience. Similarly, each extra year of experience is worth the same across all levels of education. Furthermore, the premium paid to men does not depend systematically on education or experience. Omitted variables such as ability, quality of education, or quality of experience do not make any systematic difference to the predictions of the model.¹⁵⁴ These are all assumptions made going into the analysis, rather than conclusions coming out of the data.

Assumptions are also made about the error term—the mysterious ϵ at the end of (A13). The errors are assumed to be independent and identically distributed from person to person in the dataset. Such assumptions are critical when computing *p*-values and demonstrating statistical significance. Regression modeling that does not produce statistically significant coefficients will not be good evidence of discrimination, and statistical significance cannot be established unless stylized assumptions are made about unobservable error terms.

The typical regression model, like the one sketched above, therefore involves a host of assumptions. As noted in Section V, Hooke's law—equation (1)—is relatively easy to test experimentally. For the salary discrimination model—equation (A13)—validation would be difficult. That is why we suggested that when expert testimony relies on statistical models, the court may well inquire about the assumptions behind the model and why they apply to the case at hand.

2. Standard errors, *t*-statistics, and statistical significance

Statistical proof of discrimination depends on the significance of the estimated coefficient for the gender variable. Significance is determined by the *t*-test, using the standard error. The standard error measures the likely difference between the estimated value for the coefficient and its true value. The estimated value is \$700—the coefficient of the gender variable in equation (A5); the true value *d* in (A13), remains unknown. According to the model, the difference between the estimated value and the true value is due to the action of the error term ϵ in (A3). Without ϵ , observed values would line up perfectly with expected values,

154. Technically, these omitted variables are assumed to be independent of the error term in the equation.

and estimated values for parameters would be exactly equal to true values. This does not happen.

The t -statistic is the estimated value divided by its standard error. For example, in (A15), the estimate for d is \$700. If the standard error is \$325, then t is $\$700/\$325 = 2.15$. This is significant—that is, hard to explain as the product of random error. Under the null hypothesis that d is zero, there is only about a 5% chance that the absolute value of t is greater than 2. (We are assuming the sample is large.) Thus, statistical significance is achieved (*supra* Section IV.B.2). Significance would be taken as evidence that d —the true parameter in the model (A13)—does not vanish. According to a viewpoint often presented in the social science journals and the courtroom, here is statistical proof that gender matters in determining salaries. On the other hand, if the standard error is \$1400, then t is $\$700/\$1400 = 0.5$. The difference between the estimated value of d and zero could easily result from chance. So the true value of d could well be zero, in which case gender does not affect salaries.

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 the chance that the absolute value of t exceeds 2, given the model and given the null hypothesis that d is zero. However, the 5% is often taken to be the chance of the null hypothesis given the data. This misinterpretation is commonplace in the social science literature, and it appears in some opinions describing expert testimony.¹⁵⁵ For a frequentist statistician, the chance that d is zero given the data makes no sense: Parameters do not exhibit chance variation. For a Bayesian statistician, the chance that d is zero given the data makes good sense, but the computation via the t -test could be seriously in error, because the prior probability that d is zero has not been taken into account.¹⁵⁶

The mathematical terminology in the previous paragraph may need to be deciphered: The “absolute value” of t is the magnitude, ignoring sign. Thus, the absolute value of both +3 and -3 is 3.

155. See *supra* Section IV.B & notes 102 & 116.

156. See *supra* Section IV & *supra* Appendix.

Glossary of Terms

The following definitions are adapted from a variety of sources, including Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* (2d ed. 2001), and David A. Freedman et al., *Statistics* (4th ed. 2007).

absolute value. Size, neglecting sign. The absolute value of +2.7 is 2.7; so is the absolute value of -2.7.

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. 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. In its simplest form, an equation involving conditional probabilities that relates a "prior probability" known or estimated before collecting certain data to a "posterior probability" that reflects the impact of the data on the prior probability. In Bayesian statistical inference, "the prior" expresses degrees of belief about various hypotheses. Data are collected according to some statistical model; at least, the model represents the investigator's beliefs. Bayes' rule combines the prior with the data to yield the posterior probability, which expresses the investigator's beliefs about the parameters, given the data. See Appendix A. Compare frequentist.

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.

between-observer variability. Differences that occur when two observers measure the same thing. Compare within-observer variability.

bias. Also called systematic error. A systematic tendency for an estimate to be too high or too low. An estimate is unbiased if the bias is zero. (Bias does not mean prejudice, partiality, or discriminatory intent.) See nonsampling 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). Called a dummy variable when the two possible values are 0 and 1.

binomial distribution. A distribution for the number of occurrences in repeated, independent "trials" where the probabilities are fixed. For example, the num-

ber of heads in 100 tosses of a coin follows a binomial distribution. When the probability is not too close to 0 or 1 and the number of trials is large, the binomial distribution has about the same shape as the normal distribution. See normal distribution; Poisson distribution.

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 the 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. See histogram; 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 the chi-squared statistic 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, investigators 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*-squared) 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 parameter. For estimates such as averages or rates computed from large samples, 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 central limit theorem; standard error.

confidence level. See confidence interval.

confounding variable; confounder. A confounder is correlated with the independent variable 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.

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 larger, are frowned upon 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. See validity. Compare reliability.

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 in which 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 example, in an experiment to evaluate a new drug, subjects in the treatment group are given the drug, and 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 ensure that any observed difference between the two groups comes from the treatment rather than from preexisting 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 nonrandom sample of units, also called a grab sample. Such samples are easy to take but may suffer from serious bias. Typically, 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 validity.
- data.** Observations or measurements, usually of units in a sample taken from a larger population.
- degrees of freedom.** See *t*-test.
- dependence.** Two events are dependent when the probability of one is affected by the occurrence or non-occurrence of the other. Compare independence; dependent variable.
- dependent variable.** Also called outcome variable. Compare independent variable.
- descriptive statistics.** Like the mean or standard deviation, used to summarize data.
- differential validity.** Differences in validity across different groups of subjects. See validity.
- discrete variable.** A variable that has only a small 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; 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 econometrics, the error term is called a disturbance term.

estimator. A sample statistic used to estimate the value of a population parameter. For example, the sample average commonly is used to estimate the population average. The term “estimator” connotes a statistical procedure, whereas 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.

external validity. See validity.

factors. See independent variable.

Fisher’s exact test. A statistical test for comparing two sample proportions. For example, take the proportions of white and black employees 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 details, see Michael O. Finkelstein and Bruce Levin, *Statistics for Lawyers* 154–56 (2d ed. 2001). See hypergeometric distribution; p -value; significance test; statistical hypothesis.

fitted value. See residual.

fixed significance level. Also alpha; size. A preset level, such as 5% or 1%; 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; relative frequency. Frequency is the number of times that something occurs; relative frequency is the number of occurrences, relative to a total. For example, if a coin is tossed 1000 times and lands heads 517 times, the frequency of heads is 517; the relative frequency is 0.517, or 51.7%.

frequency distribution. Shows how often specified values occur in a dataset.

frequentist. Also called objectivist. Describes statisticians who view probabilities as objective properties of a system that can be measured or estimated. Compare Bayesian. See Appendix.

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.

highly significant. See p -value; practical significance; significance test.

histogram. A plot showing how observed values fall within specified intervals, called bins or class intervals. Generally, matters are arranged so that 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 Probability 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.

identically distributed. Random variables are identically distributed when they have the same probability distribution. For example, consider a box of numbered tickets. Draw tickets at random with replacement from the box. The draws will be independent and identically distributed.

independence. Also, statistical independence. Events are independent when the probability of one is unaffected by the occurrence or non-occurrence of the other. Compare conditional probability; dependence; independent variable; dependent variable.

independent variable. Independent variables (also called explanatory variables, predictors, or risk factors) represent the causes and potential confounders in a statistical study of causation; the dependent variable represents the effect. In an observational study, independent variables may be used to divide the population up into smaller and more homogenous groups (“stratification”). In a regression model, the independent variables are used to predict the dependent variable. For example, 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. The distinction between independent and dependent variables is unrelated to statistical independence. See regression model. Compare dependent variable; dependence; independence.

indicator variable. See dummy variable.

internal validity. See validity.

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 example, $2u + 3v$ is a linear combination of u and v .

list sample. See systematic sample.

loss function. Statisticians may evaluate estimators according to a mathematical formula involving the errors—that is, 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 gives a way to find the center of a batch of numbers: Add 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.

measurement validity. See validity. Compare reliability.

median. The median, like the mean, is a 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 datasets, 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 epidemiological 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 common 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 dataset. 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 5% level, the result is not surprising; indeed, it would 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 -squared. 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.

multistage cluster sample. A probability sample drawn in stages, usually after stratification; the last stage will involve drawing a cluster. See cluster sample; probability sample; stratified random sample.

multivariate methods. Methods for fitting models with multiple variables; in statistics, multiple response variables; in other fields, multiple explanatory variables. See regression model.

natural experiment. An observational study in which treatment and control groups have been formed by some natural development; the assignment of subjects to groups is 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.

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

normal distribution. Also, Gaussian distribution. When the normal distribution has mean equal to 0 and standard error equal to 1, it is said to be “standard normal.” The equation for the density is then

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

where $e = 2.71828\dots$ and $\pi = 3.14159\dots$. The density can be rescaled to have any desired mean and standard error, resulting in the famous “bell-shaped curves” of statistical theory. Terminology notwithstanding, there need be nothing wrong with a distribution that differs from 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.

objectivist. See frequentist.

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 nonsmokers. In an observational study, the groups may differ in important ways that the investigators do not notice; controlled experiments minimize this problem. The critical distinction is that in a controlled experiment, the investigators intervene to manipulate the circumstances of the subjects; in an observational study, the investigators are passive observers. (Of course, running a good observational study is hard work, and may be quite useful.) 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 example, if 10% of all people exposed to a chemical develop a disease, compared with 5% of people who are not exposed, then the odds of the disease in the exposed group are $10/90 = 1/9$, compared with $5/95 = 1/19$ in the unexposed group. The odds ratio is $(1/9)/(1/19) = 19/9 = 2.1$. An odds ratio of 1 indicates no association. Compare relative risk.

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

one-sided test; one-tailed test. See one-sided hypothesis.

outcome variable. See dependent variable.

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. Result from 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 the null hypothesis. However, p does not give the probability that the null hypothesis is true. If p is smaller than 5%, the result is 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 dataset, 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. On the LSAT, a score of 152 places a test taker at the 50th percentile; a score of 164 is at the 90th percentile; a score of 172 is at the 99th percentile. 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. 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.

population size. Also, size of population. Number of units in the population.

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 by 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.

practice effects. Changes in test scores that result from taking the same test twice in succession, or taking two similar tests one after the other.

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. See validity. Compare reliability.

predictor. See independent variable.

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, 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. 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 with a dummy variable. Quantitative variables may be classified as discrete or continuous. Concepts such as 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*-squared varies between 0 (no fit) and 1 (perfect fit). *R*-squared does not measure the extent to which underlying assumptions are justified. See regression model. Compare multiple correlation coefficient; standard error of regression.

random error. Sources of error that are random in their effect, like draws made at random from a box. 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

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

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 a lottery. See controlled experiment. Compare observational study.

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

rate. In an epidemiological study, the number of events, divided by the size of the population; often cross-classified by age and gender. For example, the death rate from heart disease among American men ages 55–64 in 2004 was about three per thousand. Among men ages 65–74, the rate was about seven per thousand. Among women, the rate was about half that for men. Rates adjust for differences in sizes of populations or subpopulations. Often, rates are computed per unit of time, e.g., per thousand persons per year. Data source: Statistical Abstract of the United States tbl. 115 (2008).

regression coefficient. The coefficient of a variable 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, the phrase

“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 such as education and experience to adjust for productivity differences between men and women. The dummy variable would be defined as 1 for the men and 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 frequency. See frequency.

relative risk. A measure of association used in epidemiology. For example, 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 Leon Gordis, *Epidemiology* (4th ed. 2008). Compare odds ratio.

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

representative sample. Not a well-defined technical term. A sample judged to fairly represent the population, or a sample drawn by a process likely to give samples that fairly represent the population, for example, a large probability sample.

resampling. See bootstrap.

residual. The difference between an actual and a predicted value. The predicted value comes typically from a regression equation, and is better called the fitted value, because there is no real prediction going on. See regression model; independent variable.

response variable. See independent variable.

risk. Expected loss. “Expected” means on average, over the various datasets 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.

risk factor. See independent 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. Also, size of sample. The number of units in a sample.

sample weights. See stratified random 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; nonsampling error.

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

sampling interval. See systematic sample.

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

selection bias. Systematic error due to nonrandom 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.

sign test. A statistical test based on counting and the binomial distribution. For example, a Finnish study of twins found 22 monozygotic twin pairs where 1 twin smoked, 1 did not, and at least 1 of the twins had died. That sets up a race to death. In 17 cases, the smoker died first; in 5 cases, the nonsmoker died first. The null hypothesis is that smoking does not affect time to death, so the chances are 50-50 for the smoker to die first. On the null hypothesis, the chance that the smoker will win the race 17 or more times out of 22 is

8/1000. That is the p -value. The p -value can be computed from the binomial distribution. For additional detail, see Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* 339–41 (2d ed. 2001); David A. Freedman et al., *Statistics* 262–63 (4th ed. 2007).

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 preestablished value (α) 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 example, 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-sided 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-sided test would reject the null hypothesis if the sample men score substantially higher or substantially lower than the women.

The one-sided and two-sided 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-sided test would give a p -value only half as large as the two-sided test; that is, the one-sided test would appear to give stronger evidence against the null hypothesis. (“One-sided” and “one-tailed” are synonymous; so are “two-sided” and “two-tailed.”) 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. The investigators take a unit at random (as if by lottery), set it aside, take 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).

skip factor. See systematic sample.

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 example, are correlated—but learning more words will not make the feet grow. Such noncausal correlations are said to be spurious. (Originally, the term seems to have been applied to the correlation between two rates with the same denominator: Even if the numerators are unrelated, the common denominator will create some association.) Compare confounding variable.

standard deviation (SD). Indicates how far a typical element deviates from the average. For example, 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 from 1 SD below the average to 1 SD above the average. 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 3 or 4 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.

standard normal. See normal distribution.

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 example, by adjusting through statistical procedures such as multiple regression. Variables in a multiple regression

equation. See multiple regression; confounding variable; observational study. Compare controlled experiment.

statistical dependence. See dependence.

statistical hypothesis. Generally, a statement about parameters in a probability model for the data. 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 that possibility. See significance test.

statistical independence. See independence.

statistical model. See probability model.

statistical test. See significance test.

statistical significance. See p -value.

stratified random sample. A type of probability sample. The researcher divides the population into relatively homogeneous groups called “strata,” and draws a random sample separately from each stratum. Dividing the population into strata is called “stratification.” Often the sampling fraction will vary from stratum to stratum. Then sampling weights should be used to extrapolate from the sample to the population. For example, if 1 unit in 10 is sampled from stratum A while 1 unit in 100 is sampled from stratum B, then each unit drawn from A counts as 10, and each unit drawn from B counts as 100. The first kind of unit has weight 10; the second has weight 100. See Freedman et al., *Statistics* 401 (4th ed. 2007).

stratification. See independent variable; stratified random sample.

study validity. See validity.

subjectivist. See Bayesian.

systematic error. See bias.

systematic sample. Also, list sample. The elements of the population are numbered consecutively as 1, 2, 3, The investigators choose a starting point and a “sampling interval” or “skip factor” k . Then, every k th element is selected into the sample. If the starting point is 1 and $k = 10$, for example, the sample would consist of items 1, 11, 21, Sometimes the starting point is chosen at random from 1 to k : this is a random-start systematic sample.

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. With a large sample, a t -statistic larger than 2 or 3 in absolute value makes the null hypothesis rather implausible—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 example, 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 $3/1000$ 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.

Consider testing the null hypothesis that the average of a population equals a given value; the population is known to be normal. For small samples, 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. The relevant t -distribution depends on the number of degrees of freedom, which in this context equals the sample size minus one. 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 example, the Gross National Product of the United States from 1945 to 2005.

treatment group. See controlled experiment.

two-sided hypothesis; two-tailed 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-sided (or two-tailed) test. See significance test; statistical hypothesis. Compare one-sided hypothesis.

two-sided test; two-tailed test. See two-sided hypothesis.

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 posi-

tive. For example, 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 example, 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 datasets. 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. Measurement validity is 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 by the correlation coefficient between the test scores and some outcome measure (the criterion variable). See content validity; differential validity; predictive validity. Compare reliability.

Study validity is the extent to which results from a study can be relied upon. Study validity has two aspects, internal and external. A study has high internal validity when its conclusions hold under the particular circumstances of the study. A study has high external validity when its results are generalizable. For example, a well-executed randomized controlled double-blind experiment performed on an unusual study population will have high internal validity because the design is good; but its external validity will be debatable because the study population is unusual.

Validity is used also in its ordinary sense: assumptions are valid when they hold true for the situation at hand.

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.

weights. See stratified random sample.

within-observer variability. Differences that occur when an observer measures the same thing twice, or measures two things that are virtually the same. Compare between-observer variability.

z-statistic. See *t*-statistic.

z-test. See *t*-test.

References on Statistics

General Surveys

David Freedman et al., *Statistics* (4th ed. 2007).

Darrell Huff, *How to Lie with Statistics* (1993).

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

David S. Moore & William I. Notz, *Statistics: Concepts and Controversies* (2005).

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

Statistics: A Guide to the Unknown (Roxy Peck et al. eds., 4th ed. 2005).

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

David W. Barnes & John M. Conley, *Statistical Evidence in Litigation: Methodology, Procedure, and Practice* (1986 & Supp. 1989).

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

Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* (2d ed. 2001).

Modern Scientific Evidence: The Law and Science of Expert Testimony (David L. Faigman et al. eds., Volumes 1 and 2, 2d ed. 2002) (updated annually).

David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* § 12 (2d ed. 2011) (updated annually).

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

Encyclopedia of Statistical Sciences (Samuel Kotz et al. eds., 2d ed. 2005).





Reference Manual on Scientific Evidence: Third Edition


ISBN
978-0-309-21421-6

1038 pages
6 x 9
PAPERBACK (2011)

Committee on the Development of the Third Edition of the Reference Manual on Scientific Evidence; Federal Judicial Center; National Research Council

 Add book to cart

 Find similar titles

 Share this PDF



Visit the National Academies Press online and register for...

- ✓ Instant access to free PDF downloads of titles from the
 - NATIONAL ACADEMY OF SCIENCES
 - NATIONAL ACADEMY OF ENGINEERING
 - INSTITUTE OF MEDICINE
 - NATIONAL RESEARCH COUNCIL
- ✓ 10% off print titles
- ✓ Custom notification of new releases in your field of interest
- ✓ Special offers and discounts

Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press. Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences. Request reprint permission for this book

Reference Guide on Multiple Regression

DANIEL L. RUBINFELD

Daniel L. Rubinfeld, Ph.D., is Robert L. Bridges Professor of Law and Professor of Economics Emeritus, University of California, Berkeley, and Visiting Professor of Law at New York University Law School.

CONTENTS

- I. Introduction and Overview, 305
- II. Research Design: Model Specification, 311
 - A. What Is the Specific Question That Is Under Investigation by the Expert? 311
 - B. What Model Should Be Used to Evaluate the Question at Issue? 311
 - 1. Choosing the dependent variable, 312
 - 2. Choosing the explanatory variable that is relevant to the question at issue, 313
 - 3. Choosing the additional explanatory variables, 313
 - 4. Choosing the functional form of the multiple regression model, 316
 - 5. Choosing multiple regression as a method of analysis, 317
- III. Interpreting Multiple Regression Results, 318
 - A. What Is the Practical, as Opposed to the Statistical, Significance of Regression Results? 318
 - 1. When should statistical tests be used? 319
 - 2. What is the appropriate level of statistical significance? 320
 - 3. Should statistical tests be one-tailed or two-tailed? 321
 - B. Are the Regression Results Robust? 322
 - 1. What evidence exists that the explanatory variable causes changes in the dependent variable? 322
 - 2. To what extent are the explanatory variables correlated with each other? 324
 - 3. To what extent are individual errors in the regression model independent? 325
 - 4. To what extent are the regression results sensitive to individual data points? 326
 - 5. To what extent are the data subject to measurement error? 327

- IV. The Expert, 328
 - A. Who Should Be Qualified as an Expert? 328
 - B. Should the Court Appoint a Neutral Expert? 329
- V. Presentation of Statistical Evidence, 330
 - A. What Disagreements Exist Regarding Data on Which the Analysis Is Based? 330
 - B. Which Database Information and Analytical Procedures Will Aid in Resolving Disputes over Statistical Studies? 331
- Appendix: The Basics of Multiple Regression, 333
 - A. Introduction, 333
 - B. Linear Regression Model, 336
 - 1. Specifying the regression model, 337
 - 2. Regression line, 337
 - C. Interpreting Regression Results, 339
 - D. Determining the Precision of the Regression Results, 340
 - 1. Standard errors of the coefficients and *t*-statistics, 340
 - 2. Goodness-of-fit, 344
 - 3. Sensitivity of least squares regression results, 345
 - E. Reading Multiple Regression Computer Output, 346
 - F. Forecasting, 348
 - G. A Hypothetical Example, 350
- Glossary of Terms, 352
- References on Multiple Regression, 357

I. Introduction and Overview

Multiple regression analysis is a statistical tool used to understand the relationship between or among two or more variables.¹ Multiple regression involves a variable to be explained—called the dependent variable—and additional explanatory variables that are thought to produce or be associated with changes in the dependent variable.² For example, a multiple regression analysis might estimate the effect of the number of years of work on salary. Salary would be the dependent variable to be explained; the years of experience would be the explanatory variable.

Multiple regression analysis is sometimes well suited to the analysis of data about competing theories for which there are several possible explanations for the relationships among a number of explanatory variables.³ Multiple regression typically uses a single dependent variable and several explanatory variables to assess the statistical data pertinent to these theories. In a case alleging sex discrimination in salaries, for example, a multiple regression analysis would examine not only sex, but also other explanatory variables of interest, such as education and experience.⁴ The employer-defendant might use multiple regression to argue that salary is a function of the employee's education and experience, and the employee-plaintiff might argue that salary is also a function of the individual's sex. Alternatively, in an antitrust cartel damages case, the plaintiff's expert might utilize multiple regression to evaluate the extent to which the price of a product increased during the period in which the cartel was effective, after accounting for costs and other variables unrelated to the cartel. The defendant's expert might use multiple

1. A variable is anything that can take on two or more values (e.g., the daily temperature in Chicago or the salaries of workers at a factory).

2. Explanatory variables in the context of a statistical study are sometimes called independent variables. See David H. Kaye & David A. Freedman, Reference Guide on Statistics, Section II.A.1, in this manual. The guide also offers a brief discussion of multiple regression analysis. *Id.*, Section V.

3. Multiple regression is one type of statistical analysis involving several variables. Other types include matching analysis, stratification, analysis of variance, probit analysis, logit analysis, discriminant analysis, and factor analysis.

4. Thus, in *Ottaviani v. State University of New York*, 875 F.2d 365, 367 (2d Cir. 1989) (citations omitted), *cert. denied*, 493 U.S. 1021 (1990), the court stated:

In disparate treatment cases involving claims of gender discrimination, plaintiffs typically use multiple regression analysis to isolate the influence of gender on employment decisions relating to a particular job or job benefit, such as salary.

The first step in such a regression analysis is to specify all of the possible “legitimate” (i.e., non-discriminatory) factors that are likely to significantly affect the dependent variable and which could account for disparities in the treatment of male and female employees. By identifying those legitimate criteria that affect the decisionmaking process, individual plaintiffs can make predictions about what job or job benefits similarly situated employees should ideally receive, and then can measure the difference between the predicted treatment and the actual treatment of those employees. If there is a disparity between the predicted and actual outcomes for female employees, plaintiffs in a disparate treatment case can argue that the net “residual” difference represents the unlawful effect of discriminatory animus on the allocation of jobs or job benefits.

regression to suggest that the plaintiff's expert had omitted a number of price-determining variables.

More generally, multiple regression may be useful (1) in determining whether a particular effect is present; (2) in measuring the magnitude of a particular effect; and (3) in forecasting what a particular effect would be, but for an intervening event. In a patent infringement case, for example, a multiple regression analysis could be used to determine (1) whether the behavior of the alleged infringer affected the price of the patented product, (2) the size of the effect, and (3) what the price of the product would have been had the alleged infringement not occurred.

Over the past several decades, the use of multiple regression analysis in court has grown widely. Regression analysis has been used most frequently in cases of sex and race discrimination⁵ antitrust violations,⁶ and cases involving class cer-

5. Discrimination cases using multiple regression analysis are legion. *See, e.g.*, Bazemore v. Friday, 478 U.S. 385 (1986), *on remand*, 848 F.2d 476 (4th Cir. 1988); Csicseri v. Bowsher, 862 F. Supp. 547 (D.D.C. 1994) (age discrimination), *aff'd*, 67 F.3d 972 (D.C. Cir. 1995); EEOC v. General Tel. Co., 885 F.2d 575 (9th Cir. 1989), *cert. denied*, 498 U.S. 950 (1990); Bridgeport Guardians, Inc. v. City of Bridgeport, 735 F. Supp. 1126 (D. Conn. 1990), *aff'd*, 933 F.2d 1140 (2d Cir.), *cert. denied*, 502 U.S. 924 (1991); Bickerstaff v. Vassar College, 196 F.3d 435, 448–49 (2d Cir. 1999) (sex discrimination); McReynolds v. Sodexho Marriott, 349 F. Supp. 2d 1 (D.C. Cir. 2004) (race discrimination); Hnot v. Willis Group Holdings Ltd., 228 F.R.D. 476 (S.D.N.Y. 2005) (gender discrimination); Carpenter v. Boeing Co., 456 F.3d 1183 (10th Cir. 2006) (sex discrimination); Coward v. ADT Security Systems, Inc., 140 F.3d 271, 274–75 (D.C. Cir. 1998); Smith v. Virginia Commonwealth Univ., 84 F.3d 672 (4th Cir. 1996) (en banc); Hemmings v. Tidyman's Inc., 285 F.3d 1174, 1184–86 (9th Cir. 2000); Mehus v. Emporia State University, 222 F.R.D. 455 (D. Kan. 2004) (sex discrimination); Guterrez v. Johnson & Johnson, 2006 WL 3246605 (D.N.J. Nov. 6, 2006) (race discrimination); Morgan v. United Parcel Service, 380 F.3d 459 (8th Cir. 2004) (racial discrimination). *See also* Keith N. Hylton & Vincent D. Rougeau, *Lending Discrimination: Economic Theory, Econometric Evidence, and the Community Reinvestment Act*, 85 Geo. L.J. 237, 238 (1996) (“regression analysis is probably the best empirical tool for uncovering discrimination”).

6. *E.g.*, United States v. Brown Univ., 805 F. Supp. 288 (E.D. Pa. 1992) (price fixing of college scholarships), *rev'd*, 5 F.3d 658 (3d Cir. 1993); Petruzzi's IGA Supermarkets, Inc. v. Darling-Delaware Co., 998 F.2d 1224 (3d Cir.), *cert. denied*, 510 U.S. 994 (1993); Ohio v. Louis Trauth Dairy, Inc., 925 F. Supp. 1247 (S.D. Ohio 1996); *In re Chicken Antitrust Litig.*, 560 F. Supp. 963, 993 (N.D. Ga. 1980); New York v. Kraft Gen. Foods, Inc., 926 F. Supp. 321 (S.D.N.Y. 1995); Freeland v. AT&T, 238 F.R.D. 130 (S.D.N.Y. 2006); *In re Pressure Sensitive Labelstock Antitrust Litig.*, 2007 U.S. Dist. LEXIS 85466 (M.D. Pa. Nov. 19, 2007); *In re Linerboard Antitrust Litig.*, 497 F. Supp. 2d 666 (E.D. Pa. 2007) (price fixing by manufacturers of corrugated boards and boxes); *In re Polypropylene Carpet Antitrust Litig.*, 93 F. Supp. 2d 1348 (N.D. Ga. 2000); *In re OSB Antitrust Litig.*, 2007 WL 2253418 (E.D. Pa. Aug. 3, 2007) (price fixing of Oriented Strand Board, also known as “waferboard”); *In re TFT-LCD (Flat Panel) Antitrust Litig.*, 267 F.R.D. 583 (N.D. Cal. 2010).

For a broad overview of the use of regression methods in antitrust, see ABA Antitrust Section, *Econometrics: Legal, Practical and Technical Issues* (John Harkrider & Daniel Rubinfeld, eds. 2005). *See also* Jerry Hausman et al., *Competitive Analysis with Differentiated Products*, 34 *Annales D'Économie et de Statistique* 159 (1994); Gregory J. Werden, *Simulating the Effects of Differentiated Products Mergers: A Practical Alternative to Structural Merger Policy*, 5 *Geo. Mason L. Rev.* 363 (1997).

tification (under Rule 23).⁷ However, there are a range of other applications, including census undercounts,⁸ voting rights,⁹ the study of the deterrent effect of the death penalty,¹⁰ rate regulation,¹¹ and intellectual property.¹²

7. In antitrust, the circuits are currently split as to the extent to which plaintiffs must prove that common elements predominate over individual elements. *E.g.*, compare *In Re Hydrogen Peroxide Litig.*, 522 F.2d 305 (3d Cir. 2008) with *In Re Cardizem CD Antitrust Litig.*, 391 F.3d 812 (6th Cir. 2004). For a discussion of use of multiple regression in evaluating class certification, see Bret M. Dickey & Daniel L. Rubinfeld, *Antitrust Class Certification: Towards an Economic Framework*, 66 N.Y.U. Ann. Surv. Am. L. 459 (2010) and John H. Johnson & Gregory K. Leonard, *Economics and the Rigorous Analysis of Class Certification in Antitrust Cases*, 3 J. Competition L. & Econ. 341 (2007).

8. See, *e.g.*, *City of New York v. U.S. Dep't of Commerce*, 822 F. Supp. 906 (E.D.N.Y. 1993) (decision of Secretary of Commerce not to adjust the 1990 census was not arbitrary and capricious), *vacated*, 34 F.3d 1114 (2d Cir. 1994) (applying heightened scrutiny), *rev'd sub nom. Wisconsin v. City of New York*, 517 U.S. 565 (1996); *Carey v. Klutznick*, 508 F. Supp. 420, 432–33 (S.D.N.Y. 1980) (use of reasonable and scientifically valid statistical survey or sampling procedures to adjust census figures for the differential undercount is constitutionally permissible), *stay granted*, 449 U.S. 1068 (1980), *rev'd on other grounds*, 653 F.2d 732 (2d Cir. 1981), *cert. denied*, 455 U.S. 999 (1982); *Young v. Klutznick*, 497 F. Supp. 1318, 1331 (E.D. Mich. 1980), *rev'd on other grounds*, 652 F.2d 617 (6th Cir. 1981), *cert. denied*, 455 U.S. 939 (1982).

9. Multiple regression analysis was used in suits charging that at-large areawide voting was instituted to neutralize black voting strength, in violation of section 2 of the Voting Rights Act, 42 U.S.C. § 1973 (1988). Multiple regression demonstrated that the race of the candidates and that of the electorate were determinants of voting. See *Williams v. Brown*, 446 U.S. 236 (1980); *Rodriguez v. Pataki*, 308 F. Supp. 2d 346, 414 (S.D.N.Y. 2004); *United States v. Vill. of Port Chester*, 2008 U.S. Dist. LEXIS 4914 (S.D.N.Y. Jan. 17, 2008); *Meza v. Galvin*, 322 F. Supp. 2d 52 (D. Mass. 2004) (violation of VRA with regard to Hispanic voters in Boston); *Bone Shirt v. Hazeltine*, 336 F. Supp. 2d 976 (D.S.D. 2004) (violations of VRA with regard to Native American voters in South Dakota); *Georgia v. Ashcroft*, 195 F. Supp. 2d 25 (D.D.C. 2002) (redistricting of Georgia's state and federal legislative districts); *Benavidez v. City of Irving*, 638 F. Supp. 2d 709 (N.D. Tex. 2009) (challenge of city's at-large voting scheme). For commentary on statistical issues in voting rights cases, see, *e.g.*, *Statistical and Demographic Issues Underlying Voting Rights Cases*, 15 Evaluation Rev. 659 (1991); Stephen P. Klein et al., *Ecological Regression Versus the Secret Ballot*, 31 *Jurimetrics J.* 393 (1991); James W. Loewen & Bernard Grofman, *Recent Developments in Methods Used in Vote Dilution Litigation*, 21 *Urb. Law.* 589 (1989); Arthur Lupia & Kenneth McCue, *Why the 1980s Measures of Racially Polarized Voting Are Inadequate for the 1990s*, 12 *Law & Pol'y* 353 (1990).

10. See, *e.g.*, *Gregg v. Georgia*, 428 U.S. 153, 184–86 (1976). For critiques of the validity of the deterrence analysis, see National Research Council, *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates* (Alfred Blumstein et al. eds., 1978); Richard O. Lempert, *Desert and Deterrence: An Assessment of the Moral Bases of the Case for Capital Punishment*, 79 *Mich. L. Rev.* 1177 (1981); Hans Zeisel, *The Deterrent Effect of the Death Penalty: Facts v. Faith*, 1976 *Sup. Ct. Rev.* 317; and John Donohue & Justin Wolfers, *Uses and Abuses of Statistical Evidence in the Death Penalty Debate*, 58 *Stan. L. Rev.* 787 (2005).

11. See, *e.g.*, *Time Warner Entertainment Co. v. FCC*, 56 F.3d 151 (D.C. Cir. 1995) (challenge to FCC's application of multiple regression analysis to set cable rates), *cert. denied*, 516 U.S. 1112 (1996); *Appalachian Power Co. v. EPA*, 135 F.3d 791 (D.C. Cir. 1998) (challenging the EPA's application of regression analysis to set nitrous oxide emission limits); *Consumers Util. Rate Advocacy Div. v. Ark. PSC*, 99 Ark. App. 228 (Ark. Ct. App. 2007) (challenging an increase in nongas rates).

12. See *Polaroid Corp. v. Eastman Kodak Co.*, No. 76-1634-MA, 1990 WL 324105, at *29, *62–63 (D. Mass. Oct. 12, 1990) (damages awarded because of patent infringement), *amended by* No.

Multiple regression analysis can be a source of valuable scientific testimony in litigation. However, when inappropriately used, regression analysis can confuse important issues while having little, if any, probative value. In *EEOC v. Sears, Roebuck & Co.*,¹³ in which Sears was charged with discrimination against women in hiring practices, the Seventh Circuit acknowledged that “[m]ultiple regression analyses, designed to determine the effect of several independent variables on a dependent variable, which in this case is hiring, are an accepted and common method of proving disparate treatment claims.”¹⁴ However, the court affirmed the district court’s findings that the “E.E.O.C.’s regression analyses did not ‘accurately reflect Sears’ complex, nondiscriminatory decision-making processes’” and that the “‘E.E.O.C.’s statistical analyses [were] so flawed that they lack[ed] any persuasive value.’”¹⁵ Serious questions also have been raised about the use of multiple regression analysis in census undercount cases and in death penalty cases.¹⁶

The Supreme Court’s rulings in *Daubert* and *Kumho Tire* have encouraged parties to raise questions about the admissibility of multiple regression analyses.¹⁷ Because multiple regression is a well-accepted scientific methodology, courts have frequently admitted testimony based on multiple regression studies, in some cases over the strong objection of one of the parties.¹⁸ However, on some occasions courts have excluded expert testimony because of a failure to utilize a multiple regression methodology.¹⁹ On other occasions, courts have rejected regression

76-1634-MA, 1991 WL 4087 (D. Mass. Jan. 11, 1991); *Estate of Vane v. The Fair, Inc.*, 849 F.2d 186, 188 (5th Cir. 1988) (lost profits were the result of copyright infringement), *cert. denied*, 488 U.S. 1008 (1989); *Louis Vuitton Malletier v. Dooney & Bourke, Inc.*, 525 F. Supp. 2d 576, 664 (S.D.N.Y. 2007) (trademark infringement and unfair competition suit). The use of multiple regression analysis to estimate damages has been contemplated in a wide variety of contexts. *See, e.g.*, David Baldus et al., *Improving Judicial Oversight of Jury Damages Assessments: A Proposal for the Comparative Additur/Remittitur Review of Awards for Nonpecuniary Harms and Punitive Damages*, 80 Iowa L. Rev. 1109 (1995); Talcott J. Franklin, *Calculating Damages for Loss of Parental Nurture Through Multiple Regression Analysis*, 52 Wash. & Lee L. Rev. 271 (1997); Roger D. Blair & Amanda Kay Esquibel, *Yardstick Damages in Lost Profit Cases: An Econometric Approach*, 72 Denv. U. L. Rev. 113 (1994). Daniel Rubinfeld, *Quantitative Methods in Antitrust*, in 1 *Issues in Competition Law and Policy* 723 (2008).

13. 839 F.2d 302 (7th Cir. 1988).

14. *Id.* at 324 n.22.

15. *Id.* at 348, 351 (quoting *EEOC v. Sears, Roebuck & Co.*, 628 F. Supp. 1264, 1342, 1352 (N.D. Ill. 1986)). The district court commented specifically on the “severe limits of regression analysis in evaluating complex decision-making processes.” 628 F. Supp. at 1350.

16. *See* David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, Sections II.A.3, B.1, in this manual.

17. *Daubert v. Merrill Dow Pharms., Inc.* 509 U.S. 579 (1993); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 147 (1999) (expanding the *Daubert*’s application to nonscientific expert testimony).

18. *See* *Newport Ltd. v. Sears, Roebuck & Co.*, 1995 U.S. Dist. LEXIS 7652 (E.D. La. May 26, 1995). *See also* *Petruzzi’s IGA Supermarkets*, *supra* note 6, 998 F.2d at 1240, 1247 (finding that the district court abused its discretion in excluding multiple regression-based testimony and reversing the grant of summary judgment to two defendants).

19. *See, e.g., In re Executive Telecard Ltd. Sec. Litig.*, 979 F. Supp. 1021 (S.D.N.Y. 1997).

Reference Guide on Multiple Regression

studies that did not have an adequate foundation or research design with respect to the issues at hand.²⁰

In interpreting the results of a multiple regression analysis, it is important to distinguish between correlation and causality. Two variables are correlated—that is, associated with each other—when the events associated with the variables occur more frequently together than one would expect by chance. For example, if higher salaries are associated with a greater number of years of work experience, and lower salaries are associated with fewer years of experience, there is a positive correlation between salary and number of years of work experience. However, if higher salaries are associated with less experience, and lower salaries are associated with more experience, there is a negative correlation between the two variables.

A correlation between two variables does not imply that one event causes the second. Therefore, in making causal inferences, it is important to avoid *spurious correlation*.²¹ Spurious correlation arises when two variables are closely related but bear no causal relationship because they are both caused by a third, unexamined variable. For example, there might be a negative correlation between the age of certain skilled employees of a computer company and their salaries. One should not conclude from this correlation that the employer has necessarily discriminated against the employees on the basis of their age. A third, unexamined variable, such as the level of the employees' technological skills, could explain differences in productivity and, consequently, differences in salary.²² Or, consider a patent infringement case in which increased sales of an allegedly infringing product are associated with a lower price of the patented product.²³ This correlation would be spurious if the two products have their own noncompetitive market niches and the lower price is the result of a decline in the production costs of the patented product.

Pointing to the possibility of a spurious correlation will typically not be enough to dispose of a statistical argument. It may be appropriate to give little weight to such an argument absent a showing that the correlation is relevant. For example, a statistical showing of a relationship between technological skills

20. See *City of Tuscaloosa v. Harcros Chemicals, Inc.*, 158 F.2d 548 (11th Cir. 1998), in which the court ruled plaintiffs' regression-based expert testimony inadmissible and granted summary judgment to the defendants. See also *American Booksellers Ass'n v. Barnes & Noble, Inc.*, 135 F. Supp. 2d 1031, 1041 (N.D. Cal. 2001), in which a model was said to contain "too many assumptions and simplifications that are not supported by real-world evidence," and *Obrey v. Johnson*, 400 F.3d 691 (9th Cir. 2005).

21. See David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, Section V.B.3, in this manual.

22. See, e.g., *Sheehan v. Daily Racing Form Inc.*, 104 F.3d 940, 942 (7th Cir.) (rejecting plaintiff's age discrimination claim because statistical study showing correlation between age and retention ignored the "more than remote possibility that age was correlated with a legitimate job-related qualification"), *cert. denied*, 521 U.S. 1104 (1997).

23. In some particular cases, there are statistical tests that allow one to reject claims of causality. For a brief description of these tests, which were developed by Jerry Hausman, see Robert S. Pindyck & Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts* § 7.5 (4th ed. 1997).

and worker productivity might be required in the age discrimination example, above.²⁴

Causality cannot be inferred by data analysis alone; rather, one must infer that a causal relationship exists on the basis of an underlying causal theory that explains the relationship between the two variables. Even when an appropriate theory has been identified, causality can never be inferred directly. One must also look for empirical evidence that there is a causal relationship. Conversely, the fact that two variables are correlated does not guarantee the existence of a relationship; it could be that the model—a characterization of the underlying causal theory—does not reflect the correct interplay among the explanatory variables. In fact, the absence of correlation does not guarantee that a causal relationship does not exist. Lack of correlation could occur if (1) there are insufficient data, (2) the data are measured inaccurately, (3) the data do not allow multiple causal relationships to be sorted out, or (4) the model is specified wrongly because of the omission of a variable or variables that are related to the variable of interest.

There is a tension between any attempt to reach conclusions with near certainty and the inherently uncertain nature of multiple regression analysis. In general, the statistical analysis associated with multiple regression allows for the expression of uncertainty in terms of probabilities. The reality that statistical analysis generates probabilities concerning relationships rather than certainty should not be seen in itself as an argument against the use of statistical evidence, or worse, as a reason to not admit that there is uncertainty at all. The only alternative might be to use less reliable anecdotal evidence.

This reference guide addresses a number of procedural and methodological issues that are relevant in considering the admissibility of, and weight to be accorded to, the findings of multiple regression analyses. It also suggests some standards of reporting and analysis that an expert presenting multiple regression analyses might be expected to meet. Section II discusses research design—how the multiple regression framework can be used to sort out alternative theories about a case. The guide discusses the importance of choosing the appropriate specification of the multiple regression model and raises the issue of whether multiple regression is appropriate for the case at issue. Section III accepts the regression framework and concentrates on the interpretation of the multiple regression results from both a statistical and a practical point of view. It emphasizes the distinction between regression results that are statistically significant and results that are meaningful to the trier of fact. It also points to the importance of evaluating the robustness

24. See, e.g., *Allen v. Seidman*, 881 F.2d 375 (7th Cir. 1989) (judicial skepticism was raised when the defendant did not submit a logistic regression incorporating an omitted variable—the possession of a higher degree or special education; defendant’s attack on statistical comparisons must also include an analysis that demonstrates that comparisons are flawed). The appropriate requirements for the defendant’s showing of spurious correlation could, in general, depend on the discovery process. See, e.g., *Boykin v. Georgia Pac. Co.*, 706 F.2d 1384 (1983) (criticism of a plaintiff’s analysis for not including omitted factors, when plaintiff considered all information on an application form, was inadequate).

of regression analyses, i.e., seeing the extent to which the results are sensitive to changes in the underlying assumptions of the regression model. Section IV briefly discusses the qualifications of experts and suggests a potentially useful role for court-appointed neutral experts. Section V emphasizes procedural aspects associated with use of the data underlying regression analyses. It encourages greater pretrial efforts by the parties to attempt to resolve disputes over statistical studies.

Throughout the main body of this guide, hypothetical examples are used as illustrations. Moreover, the basic “mathematics” of multiple regression has been kept to a bare minimum. To achieve that goal, the more formal description of the multiple regression framework has been placed in the Appendix. The Appendix is self-contained and can be read before or after the text. The Appendix also includes further details with respect to the examples used in the body of this guide.

II. Research Design: Model Specification

Multiple regression allows the testifying economist or other expert to choose among alternative theories or hypotheses and assists the expert in distinguishing correlations between variables that are plainly spurious from those that may reflect valid relationships.

A. What Is the Specific Question That Is Under Investigation by the Expert?

Research begins with a clear formulation of a research question. The data to be collected and analyzed must relate directly to this question; otherwise, appropriate inferences cannot be drawn from the statistical analysis. For example, if the question at issue in a patent infringement case is what price the plaintiff’s product would have been but for the sale of the defendant’s infringing product, sufficient data must be available to allow the expert to account statistically for the important factors that determine the price of the product.

B. What Model Should Be Used to Evaluate the Question at Issue?

Model specification involves several steps, each of which is fundamental to the success of the research effort. Ideally, a multiple regression analysis builds on a theory that describes the variables to be included in the study. A typical regression model will include one or more dependent variables, each of which is believed to be causally related to a series of explanatory variables. Because we cannot be certain that the explanatory variables are themselves unaffected or independent of the influence of the dependent variable (at least at the point of initial study), the explanatory

variables are often termed *covariates*. Covariates are known to have an association with the dependent or outcome variable, but causality remains an open question.

For example, the theory of labor markets might lead one to expect salaries in an industry to be related to workers' experience and the productivity of workers' jobs. A belief that there is job discrimination would lead one to create a model in which the dependent variable was a measure of workers' salaries and the list of covariates included a variable reflecting discrimination in addition to measures of job training and experience.

In a perfect world, the analysis of the job discrimination (or any other) issue might be accomplished through a controlled "natural experiment," in which employees would be randomly assigned to a variety of employers in an industry under study and asked to fill positions requiring identical experience and skills. In this observational study, where the only difference in salaries could be a result of discrimination, it would be possible to draw clear and direct inferences from an analysis of salary data. Unfortunately, the opportunity to conduct observational studies of this kind is rarely available to experts in the context of legal proceedings. In the real world, experts must do their best to interpret the results of real-world "quasi-experiments," in which it is impossible to control all factors that might affect worker salaries or other variables of interest.²⁵

Models are often characterized in terms of parameters—numerical characteristics of the model. In the labor market discrimination example, one parameter might reflect the increase in salary associated with each additional year of prior job experience. Another parameter might reflect the reduction in salary associated with a lack of current on-the-job experience. Multiple regression uses a sample, or a selection of data, from the population (all the units of interest) to obtain estimates of the values of the parameters of the model. An estimate associated with a particular explanatory variable is an estimated regression coefficient.

Failure to develop the proper theory, failure to choose the appropriate variables, or failure to choose the correct form of the model can substantially bias the statistical results—that is, create a systematic tendency for an estimate of a model parameter to be too high or too low.

1. Choosing the dependent variable

The variable to be explained, the dependent variable, should be the appropriate variable for analyzing the question at issue.²⁶ Suppose, for example, that pay dis-

25. In the literature on natural and quasi-experiments, the explanatory variables are characterized as "treatments" and the dependent variable as the "outcome." For a review of natural experiments in the criminal justice arena, see David P. Farrington, *A Short History of Randomized Experiments in Criminology*, 27 *Evaluation Rev.* 218–27 (2003).

26. In multiple regression analysis, the dependent variable is usually a continuous variable that takes on a range of numerical values. When the dependent variable is categorical, taking on only two or three values, modified forms of multiple regression, such as probit analysis or logit analysis, are

crimination among hourly workers is a concern. One choice for the dependent variable is the hourly wage rate of the employees; another choice is the annual salary. The distinction is important, because annual salary differences may in part result from differences in hours worked. If the number of hours worked is the product of worker preferences and not discrimination, the hourly wage is a good choice. If the number of hours worked is related to the alleged discrimination, annual salary is the more appropriate dependent variable to choose.²⁷

2. Choosing the explanatory variable that is relevant to the question at issue

The explanatory variable that allows the evaluation of alternative hypotheses must be chosen appropriately. Thus, in a discrimination case, the variable of interest may be the race or sex of the individual. In an antitrust case, it may be a variable that takes on the value 1 to reflect the presence of the alleged anticompetitive behavior and the value 0 otherwise.²⁸

3. Choosing the additional explanatory variables

An attempt should be made to identify additional known or hypothesized explanatory variables, some of which are measurable and may support alternative substantive hypotheses that can be accounted for by the regression analysis. Thus, in a discrimination case, a measure of the skills of the workers may provide an alternative explanation—lower salaries may have been the result of inadequate skills.²⁹

appropriate. For an example of the use of the latter, see *EEOC v. Sears, Roebuck & Co.*, 839 F.2d 302, 325 (7th Cir. 1988) (EEOC used logit analysis to measure the impact of variables, such as age, education, job-type experience, and product-line experience, on the female percentage of commission hires).

27. In job systems in which annual salaries are tied to grade or step levels, the annual salary corresponding to the job position could be more appropriate.

28. Explanatory variables may vary by type, which will affect the interpretation of the regression results. Thus, some variables may be continuous and others may be categorical.

29. In *James v. Stockham Valves*, 559 F. 2d 310 (5th Cir. 1977), the Court of Appeals rejected the employer's claim that skill level rather than race determined assignment and wage levels, noting the circularity of defendant's argument. In *Ottaviani v. State University of New York*, 679 F. Supp. 288, 306–08 (S.D.N.Y. 1988), *aff'd*, 875 F.2d 365 (2d Cir. 1989), *cert. denied*, 493 U.S. 1021 (1990), the court ruled (in the liability phase of the trial) that the university showed that there was no discrimination in either placement into initial rank or promotions between ranks, and so rank was a proper variable in multiple regression analysis to determine whether women faculty members were treated differently than men.

However, in *Trout v. Garrett*, 780 F. Supp. 1396, 1414 (D.D.C. 1991), the court ruled (in the damage phase of the trial) that the extent of civilian employees' prehire work experience was not an appropriate variable in a regression analysis to compute back pay in employment discrimination. According to the court, including the prehire level would have resulted in a finding of no sex discrimination, despite a contrary conclusion in the liability phase of the action. *Id.* See also *Stuart v. Roache*, 951 F.2d 446 (1st Cir. 1991) (allowing only 3 years of seniority to be considered as the result of prior

Not all possible variables that might influence the dependent variable can be included if the analysis is to be successful; some cannot be measured, and others may make little difference.³⁰ If a preliminary analysis shows the unexplained portion of the multiple regression to be unacceptably high, the expert may seek to discover whether some previously undetected variable is missing from the analysis.³¹

Failure to include a major explanatory variable that is correlated with the variable of interest in a regression model may cause an included variable to be credited with an effect that actually is caused by the excluded variable.³² In general, omitted variables that are correlated with the dependent variable reduce the probative value of the regression analysis. The importance of omitting a relevant variable depends on the strength of the relationship between the omitted variable and the dependent variable and the strength of the correlation between the omitted variable and the explanatory variables of interest. Other things being equal, the greater the correlation between the omitted variable and the variable of interest, the greater the bias caused by the omission. As a result, the omission of an important variable may lead to inferences made from regression analyses that do not assist the trier of fact.³³

discrimination), *cert. denied*, 504 U.S. 913 (1992). Whether a particular variable reflects “legitimate” considerations or itself reflects or incorporates illegitimate biases is a recurring theme in discrimination cases. See, e.g., *Smith v. Virginia Commonwealth Univ.*, 84 F.3d 672, 677 (4th Cir. 1996) (en banc) (suggesting that whether “performance factors” should have been included in a regression analysis was a question of material fact); *id.* at 681–82 (Luttig, J., concurring in part) (suggesting that the failure of the regression analysis to include “performance factors” rendered it so incomplete as to be inadmissible); *id.* at 690–91 (Michael, J., dissenting) (suggesting that the regression analysis properly excluded “performance factors”); see also *Diehl v. Xerox Corp.*, 933 F. Supp. 1157, 1168 (W.D.N.Y. 1996).

30. The summary effect of the excluded variables shows up as a random error term in the regression model, as does any modeling error. See Appendix, *infra*, for details. *But see* David W. Peterson, *Reference Guide on Multiple Regression*, 36 *Jurimetrics J.* 213, 214 n.2 (1996) (review essay) (asserting that “the presumption that the combined effect of the explanatory variables omitted from the model are uncorrelated with the included explanatory variables” is “a knife-edge condition . . . not likely to occur”).

31. A very low R -squared (R^2) is one indication of an unexplained portion of the multiple regression model that is unacceptably high. However, the inference that one makes from a particular value of R^2 will depend, of necessity, on the context of the particular issues under study and the particular dataset that is being analyzed. For reasons discussed in the Appendix, a low R^2 does not necessarily imply a poor model (and vice versa).

32. Technically, the omission of explanatory variables that are correlated with the variable of interest can cause biased estimates of regression parameters.

33. See *Bazemore v. Friday*, 751 F.2d 662, 671–72 (4th Cir. 1984) (upholding the district court’s refusal to accept a multiple regression analysis as proof of discrimination by a preponderance of the evidence, the court of appeals stated that, although the regression used four variable factors (race, education, tenure, and job title), the failure to use other factors, including pay increases that varied by county, precluded their introduction into evidence), *aff’d in part, vacated in part*, 478 U.S. 385 (1986).

Note, however, that in *Sobel v. Yeshiva University*, 839 F.2d 18, 33, 34 (2d Cir. 1988), *cert. denied*, 490 U.S. 1105 (1989), the court made clear that “a [Title VII] defendant challenging the validity of

Omitting variables that are not correlated with the variable of interest is, in general, less of a concern, because the parameter that measures the effect of the variable of interest on the dependent variable is estimated without bias. Suppose, for example, that the effect of a policy introduced by the courts to encourage husbands to pay child support has been tested by randomly choosing some cases to be handled according to current court policies and other cases to be handled according to a new, more stringent policy. The effect of the new policy might be measured by a multiple regression using payment success as the dependent variable and a 0 or 1 explanatory variable (1 if the new program was applied; 0 if it was not). Failure to include an explanatory variable that reflected the age of the husbands involved in the program would not affect the court's evaluation of the new policy, because men of any given age are as likely to be affected by the old policy as they are the new policy. Randomly choosing the court's policy to be applied to each case has ensured that the omitted age variable is not correlated with the policy variable.

Bias caused by the omission of an important variable that is related to the included variables of interest can be a serious problem.³⁴ Nonetheless, it is possible for the expert to account for bias qualitatively if the expert has knowledge (even if not quantifiable) about the relationship between the omitted variable and the explanatory variable. Suppose, for example, that the plaintiff's expert in a sex discrimination pay case is unable to obtain quantifiable data that reflect the skills necessary for a job, and that, on average, women are more skillful than men. Suppose also that a regression analysis of the wage rate of employees (the dependent variable) on years of experience and a variable reflecting the sex of each employee (the explanatory variable) suggests that men are paid substantially more than women with the same experience. Because differences in skill levels have not been taken into account, the expert may conclude reasonably that the

a multiple regression analysis [has] to make a showing that the factors it contends ought to have been included would weaken the showing of salary disparity made by the analysis," by making a specific attack and "a showing of relevance for each particular variable it contends . . . ought to [be] includ[ed]" in the analysis, rather than by simply attacking the results of the plaintiffs' proof as inadequate for lack of a given variable. *See also* Smith v. Virginia Commonwealth Univ., 84 F.3d 672 (4th Cir. 1996) (en banc) (finding that whether certain variables should have been included in a regression analysis is a question of fact that precludes summary judgment); Freeland v. AT&T, 238 F.R.D. 130, 145 (S.D.N.Y. 2006) ("Ordinarily, the failure to include a variable in a regression analysis will affect the probative value of the analysis and not its admissibility").

Also, in *Bazemore v. Friday*, the Court, declaring that the Fourth Circuit's view of the evidentiary value of the regression analyses was plainly incorrect, stated that "[n]ormally, failure to include variables will affect the analysis' probativeness, not its admissibility. Importantly, it is clear that a regression analysis that includes less than 'all measurable variables' may serve to prove a plaintiff's case." 478 U.S. 385, 400 (1986) (footnote omitted).

34. *See also* David H. Kaye & David A. Freedman, Reference Guide on Statistics, Section V.B.3, in this manual.

wage difference measured by the regression is a conservative estimate of the true discriminatory wage difference.

The precision of the measure of the effect of a variable of interest on the dependent variable is also important.³⁵ In general, the more complete the explained relationship between the included explanatory variables and the dependent variable, the more precise the results. Note, however, that the inclusion of explanatory variables that are irrelevant (i.e., not correlated with the dependent variable) reduces the precision of the regression results. This can be a source of concern when the sample size is small, but it is not likely to be of great consequence when the sample size is large.

4. *Choosing the functional form of the multiple regression model*

Choosing the proper set of variables to be included in the multiple regression model does not complete the modeling exercise. The expert must also choose the proper form of the regression model. The most frequently selected form is the linear regression model (described in the Appendix). In this model, the magnitude of the change in the dependent variable associated with the change in any of the explanatory variables is the same no matter what the level of the explanatory variables. For example, one additional year of experience might add \$5000 to salary, regardless of the previous experience of the employee.

In some instances, however, there may be reason to believe that changes in explanatory variables will have differential effects on the dependent variable as the values of the explanatory variables change. In these instances, the expert should consider the use of a nonlinear model. Failure to account for nonlinearities can lead to either overstatement or understatement of the effect of a change in the value of an explanatory variable on the dependent variable.

One particular type of nonlinearity involves the interaction among several variables. An interaction variable is the product of two other variables that are included in the multiple regression model. The interaction variable allows the expert to take into account the possibility that the effect of a change in one variable on the dependent variable may change as the level of another explanatory variable changes. For example, in a salary discrimination case, the inclusion of a term that interacts a variable measuring experience with a variable representing the sex of the employee (1 if a female employee; 0 if a male employee) allows the expert to test whether the sex differential varies with the level of experience. A significant negative estimate of the parameter associated with the sex variable suggests that inexperienced women are discriminated against, whereas a significant

35. A more precise estimate of a parameter is an estimate with a smaller standard error. See Appendix, *infra*, for details.

negative estimate of the interaction parameter suggests that the extent of discrimination increases with experience.³⁶

Note that insignificant coefficients in a model with interactions may suggest a lack of discrimination, whereas a model without interactions may suggest the contrary. It is especially important to account for interaction terms that could affect the determination of discrimination; failure to do so may lead to false conclusions concerning discrimination.

5. Choosing multiple regression as a method of analysis

There are many multivariate statistical techniques other than multiple regression that are useful in legal proceedings. Some statistical methods are appropriate when nonlinearities are important;³⁷ others apply to models in which the dependent variable is discrete, rather than continuous.³⁸ Still others have been applied predominantly to respond to methodological concerns arising in the context of discrimination litigation.³⁹

It is essential that a valid statistical method be applied to assist with the analysis in each legal proceeding. Therefore, the expert should be prepared to explain why any chosen method, including multiple regression, was more suitable than the alternatives.

36. For further details concerning interactions, see the Appendix, *infra*. Note that in *Ottaviani v. State University of New York*, 875 F.2d 365, 367 (2d Cir. 1989), *cert. denied*, 493 U.S. 1021 (1990), the defendant relied on a regression model in which a dummy variable reflecting gender appeared as an explanatory variable. The female plaintiff, however, used an alternative approach in which a regression model was developed for men only (the alleged protected group). The salaries of women predicted by this equation were then compared with the actual salaries; a positive difference would, according to the plaintiff, provide evidence of discrimination. For an evaluation of the methodological advantages and disadvantages of this approach, see Joseph L. Gastwirth, *A Clarification of Some Statistical Issues in Watson v. Fort Worth Bank and Trust*, 29 *Jurimetrics J.* 267 (1989).

37. These techniques include, but are not limited to, piecewise linear regression, polynomial regression, maximum likelihood estimation of models with nonlinear functional relationships, and autoregressive and moving-average time-series models. See, e.g., Pindyck & Rubinfeld, *supra* note 23, at 117–21, 136–37, 273–84, 463–601.

38. For a discussion of probit analysis and logit analysis, techniques that are useful in the analysis of qualitative choice, see *id.* at 248–81.

39. The correct model for use in salary discrimination suits is a subject of debate among labor economists. As a result, some have begun to evaluate alternative approaches, including urn models (Bruce Levin & Herbert Robbins, *Urn Models for Regression Analysis, with Applications to Employment Discrimination Studies*, *Law & Contemp. Probs.*, Autumn 1983, at 247) and, as a means of correcting for measurement errors, reverse regression (Delores A. Conway & Harry V. Roberts, *Reverse Regression, Fairness, and Employment Discrimination*, 1 *J. Bus. & Econ. Stat.* 75 (1983)). But see Arthur S. Goldberger, *Redirecting Reverse Regressions*, 2 *J. Bus. & Econ. Stat.* 114 (1984); Arlene S. Ash, *The Perverse Logic of Reverse Regression*, in *Statistical Methods in Discrimination Litigation* 85 (D.H. Kaye & Mikel Aickin eds., 1986).

III. Interpreting Multiple Regression Results

Multiple regression results can be interpreted in purely statistical terms, through the use of significance tests, or they can be interpreted in a more practical, nonstatistical manner. Although an evaluation of the practical significance of regression results is almost always relevant in the courtroom, tests of statistical significance are appropriate only in particular circumstances.

A. *What Is the Practical, as Opposed to the Statistical, Significance of Regression Results?*

Practical significance means that the magnitude of the effect being studied is not de minimis—it is sufficiently important substantively for the court to be concerned. For example, if the average wage rate is \$10.00 per hour, a wage differential between men and women of \$0.10 per hour is likely to be deemed practically insignificant because the differential represents only 1% ($\$0.10/\10.00) of the average wage rate.⁴⁰ That same difference could be statistically significant, however, if a sufficiently large sample of men and women was studied.⁴¹ The reason is that statistical significance is determined, in part, by the number of observations in the dataset.

As a general rule, the statistical significance of the magnitude of a regression coefficient increases as the sample size increases. Thus, a \$1.00 per hour wage differential between men and women that was determined to be insignificantly different from zero with a sample of 20 men and women could be highly significant if the sample size were increased to 200.

Often, results that are practically significant are also statistically significant.⁴² However, it is possible with a large dataset to find statistically significant coeffi-

40. There is no specific percentage threshold above which a result is practically significant. Practical significance must be evaluated in the context of a particular legal issue. See also David H. Kaye & David A. Freedman, Reference Guide on Statistics, Section IV.B.2, in this manual.

41. Practical significance also can apply to the overall credibility of the regression results. Thus, in *McCleskey v. Kemp*, 481 U.S. 279 (1987), coefficients on race variables were statistically significant, but the Court declined to find them legally or constitutionally significant.

42. In *Melani v. Board of Higher Education*, 561 F. Supp. 769, 774 (S.D.N.Y. 1983), a Title VII suit was brought against the City University of New York (CUNY) for allegedly discriminating against female instructional staff in the payment of salaries. One approach of the plaintiff's expert was to use multiple regression analysis. The coefficient on the variable that reflected the sex of the employee was approximately \$1800 when all years of data were included. Practically (in terms of average wages at the time) and statistically (in terms of a 5% significance test), this result was significant. Thus, the court stated that “[p]laintiffs have produced statistically *significant* evidence that women hired as CUNY instructional staff since 1972 received *substantially* lower salaries than similarly qualified men.” *Id.* at 781 (emphasis added). For a related analysis involving multiple comparison, see *Csicseri v. Bowsher*,

cients that are practically insignificant. Similarly, it is also possible (especially when the sample size is small) to obtain results that are practically significant but fail to achieve statistical significance. Suppose, for example, that an expert undertakes a damages study in a patent infringement case and predicts “but-for sales”—what sales would have been had the infringement not occurred—using data that predate the period of alleged infringement. If data limitations are such that only 3 or 4 years of preinfringement sales are known, the difference between but-for sales and actual sales during the period of alleged infringement could be practically significant but statistically insignificant. Alternatively, with only 3 or 4 data points, the expert would be unable to detect an effect, even if one existed.

1. *When should statistical tests be used?*

A test of a specific contention, a hypothesis test, often assists the court in determining whether a violation of the law has occurred in areas in which direct evidence is inaccessible or inconclusive. For example, an expert might use hypothesis tests in race and sex discrimination cases to determine the presence of a discriminatory effect.

Statistical evidence alone never can prove with absolute certainty the worth of any substantive theory. However, by providing evidence contrary to the view that a particular form of discrimination has not occurred, for example, the multiple regression approach can aid the trier of fact in assessing the likelihood that discrimination has occurred.⁴³

Tests of hypotheses are appropriate in a cross-sectional analysis, in which the data underlying the regression study have been chosen as a sample of a population at a particular point in time, and in a time-series analysis, in which the data being evaluated cover a number of time periods. In either analysis, the expert may want to evaluate a specific hypothesis, usually relating to a question of liability or to the determination of whether there is measurable impact of an alleged violation. Thus, in a sex discrimination case, an expert may want to evaluate a null hypothesis of no discrimination against the alternative hypothesis that discrimination takes a par-

862 F. Supp. 547, 572 (D.D.C. 1994) (noting that plaintiff’s expert found “statistically significant instances of discrimination” in 2 of 37 statistical comparisons, but suggesting that “2 of 37 amounts to roughly 5% and is hardly indicative of a pattern of discrimination”), *aff’d*, 67 F.3d 972 (D.C. Cir. 1995).

43. See *International Brotherhood of Teamsters v. United States*, 431 U.S. 324 (1977) (the Court inferred discrimination from overwhelming statistical evidence by a preponderance of the evidence); *Ryther v. KARE 11*, 108 F.3d 832, 844 (8th Cir. 1997) (“The plaintiff produced overwhelming evidence as to the elements of a prima facie case, and strong evidence of pretext, which, when considered with indications of age-based animus in [plaintiff’s] work environment, clearly provide sufficient evidence as a matter of law to allow the trier of fact to find intentional discrimination.”); *Paige v. California*, 291 F.3d 1141 (9th Cir. 2002) (allowing plaintiffs to rely on aggregated data to show employment discrimination).

ticular form.⁴⁴ Alternatively, in an antitrust damages proceeding, the expert may want to test a null hypothesis of no legal impact against the alternative hypothesis that there was an impact. In either type of case, it is important to realize that rejection of the null hypothesis does not in itself prove legal liability. It is possible to reject the null hypothesis and believe that an alternative explanation other than one involving legal liability accounts for the results.⁴⁵

Often, the null hypothesis is stated in terms of a particular regression coefficient being equal to 0. For example, in a wage discrimination case, the null hypothesis would be that there is no wage difference between sexes. If a negative difference is observed (meaning that women are found to earn less than men, after the expert has controlled statistically for legitimate alternative explanations), the difference is evaluated as to its statistical significance using the *t*-test.⁴⁶ The *t*-test uses the *t*-statistic to evaluate the hypothesis that a model parameter takes on a particular value, usually 0.

2. *What is the appropriate level of statistical significance?*

In most scientific work, the level of statistical significance required to reject the null hypothesis (i.e., to obtain a statistically significant result) is set conventionally at 0.05, or 5%.⁴⁷ The significance level measures the probability that the null hypothesis will be rejected incorrectly. In general, the lower the percentage required for statistical significance, the more difficult it is to reject the null hypothesis; therefore, the lower the probability that one will err in doing so. Although the 5% criterion is typical, reporting of more stringent 1% significance tests or less stringent 10% tests can also provide useful information.

In doing a statistical test, it is useful to compute an observed significance level, or *p*-value. The *p*-value associated with the null hypothesis that a regression coefficient is 0 is the probability that a coefficient of this magnitude or larger could have occurred by chance if the null hypothesis were true. If the *p*-value were less than or equal to 5%, the expert would reject the null hypothesis in favor of the

44. Tests are also appropriate when comparing the outcomes of a set of employer decisions with those that would have been obtained had the employer chosen differently from among the available options.

45. See David H. Kaye & David A. Freedman, Reference Guide on Statistics, Section IV.C.5, in this manual.

46. The *t*-test is strictly valid only if a number of important assumptions hold. However, for many regression models, the test is approximately valid if the sample size is sufficiently large. See Appendix, *infra*, for a more complete discussion of the assumptions underlying multiple regression.

47. See, e.g., *Palmer v. Shultz*, 815 F.2d 84, 92 (D.C. Cir. 1987) (“the .05 level of significance . . . [is] certainly sufficient to support an inference of discrimination” (quoting *Segar v. Smith*, 738 F.2d 1249, 1283 (D.C. Cir. 1984), *cert. denied*, 471 U.S. 1115 (1985))); *United States v. Delaware*, 2004 U.S. Dist. LEXIS 4560 (D. Del. Mar. 22, 2004) (stating that .05 is the normal standard chosen).

alternative hypothesis; if the p -value were greater than 5%, the expert would fail to reject the null hypothesis.⁴⁸

3. Should statistical tests be one-tailed or two-tailed?

When the expert evaluates the null hypothesis that a variable of interest has no linear association with a dependent variable against the alternative hypothesis that there is an association, a two-tailed test, which allows for the effect to be either positive or negative, is usually appropriate. A one-tailed test would usually be applied when the expert believes, perhaps on the basis of other direct evidence presented at trial, that the alternative hypothesis is either positive or negative, but not both. For example, an expert might use a one-tailed test in a patent infringement case if he or she strongly believes that the effect of the alleged infringement on the price of the infringing product was either zero or negative. (The sales of the infringing product competed with the sales of the infringed product, thereby lowering the price.) By using a one-tailed test, the expert is in effect stating that prior to looking at the data it would be very surprising if the data pointed in the direct opposite to the one posited by the expert.

Because using a one-tailed test produces p -values that are one-half the size of p -values using a two-tailed test, the choice of a one-tailed test makes it easier for the expert to reject a null hypothesis. Correspondingly, the choice of a two-tailed test makes null hypothesis rejection less likely. Because there is some arbitrariness involved in the choice of an alternative hypothesis, courts should avoid relying solely on sharply defined statistical tests.⁴⁹ Reporting the p -value or a confidence interval should be encouraged because it conveys useful information to the court, whether or not a null hypothesis is rejected.

48. The use of 1%, 5%, and, sometimes, 10% levels for determining statistical significance remains a subject of debate. One might argue, for example, that when regression analysis is used in a price-fixing antitrust case to test a relatively specific alternative to the null hypothesis (e.g., price fixing), a somewhat lower level of confidence (a higher level of significance, such as 10%) might be appropriate. Otherwise, when the alternative to the null hypothesis is less specific, such as the rather vague alternative of "effect" (e.g., the price increase is caused by the increased cost of production, increased demand, a sharp increase in advertising, or price fixing), a high level of confidence (associated with a low significance level, such as 1%) may be appropriate. See, e.g., *Vuyanic v. Republic Nat'l Bank*, 505 F. Supp. 224, 272 (N.D. Tex. 1980) (noting the "arbitrary nature of the adoption of the 5% level of [statistical] significance" to be required in a legal context); *Cook v. Rockwell Int'l Corp.*, 2006 U.S. Dist. LEXIS 89121 (D. Colo. Dec. 7, 2006).

49. Courts have shown a preference for two-tailed tests. See, e.g., *Palmer v. Shultz*, 815 F.2d 84, 95–96 (D.C. Cir. 1987) (rejecting the use of one-tailed tests, the court found that because some appellants were claiming overselection for certain jobs, a two-tailed test was more appropriate in Title VII cases); *Moore v. Summers*, 113 F. Supp. 2d 5, 20 (D.D.C. 2000) (reiterating the preference for a two-tailed test). See also David H. Kaye & David A. Freedman, Reference Guide on Statistics, Section IV.C.2, in this manual; *Csicseri v. Bowsher*, 862 F. Supp. 547, 565 (D.D.C. 1994) (finding that although a one-tailed test is "not without merit," a two-tailed test is preferable).

B. Are the Regression Results Robust?

The issue of robustness—whether regression results are sensitive to slight modifications in assumptions (e.g., that the data are measured accurately)—is of vital importance. If the assumptions of the regression model are valid, standard statistical tests can be applied. However, when the assumptions of the model are violated, standard tests can overstate or understate the significance of the results.

The violation of an assumption does not necessarily invalidate a regression analysis, however. In some instances in which the assumptions of multiple regression analysis fail, there are other statistical methods that are appropriate. Consequently, experts should be encouraged to provide additional information that relates to the issue of whether regression assumptions are valid, and if they are not valid, the extent to which the regression results are robust. The following questions highlight some of the more important assumptions of regression analysis.

1. What evidence exists that the explanatory variable causes changes in the dependent variable?

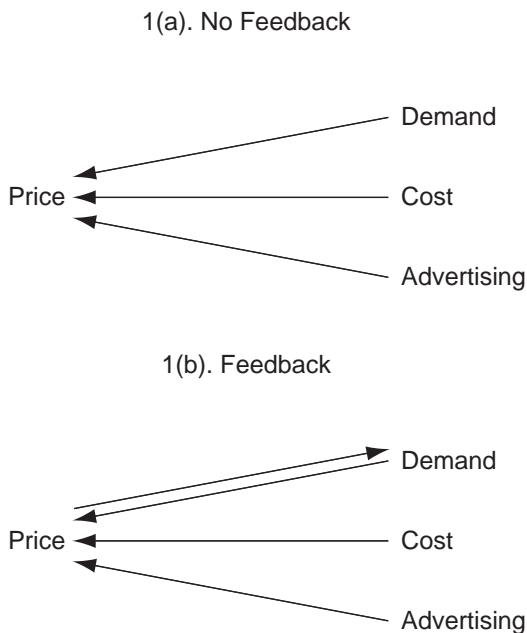
In the multiple regression framework, the expert often assumes that changes in explanatory variables affect the dependent variable, but changes in the dependent variable do not affect the explanatory variables—that is, there is no feedback.⁵⁰ In making this assumption, the expert draws the conclusion that a correlation between a covariate and the dependent outcome variable results from the effect of the former on the latter and not vice versa. Were it the case that the causality was reversed so that the outcome variable affected the covariate, and not vice versa, spurious correlation is likely to cause the expert and the trier of fact to reach the wrong conclusion. Finally, it is possible in some cases that both the outcome variable and the covariate each affect the other; if the expert does not take this more complex relationship into account, the regression coefficient on the variable of interest could be either too high or too low.⁵¹

Figure 1 illustrates this point. In Figure 1(a), the dependent variable, price, is explained through a multiple regression framework by three covariate explanatory variables—demand, cost, and advertising—with no feedback. Each of the three covariates is assumed to affect price causally, while price is assumed to have no effect on the three covariates. However, in Figure 1(b), there is feedback, because price affects demand, and demand, cost, and advertising affect price. Cost and advertising, however, are not affected by price. In this case both price and demand are jointly determined; each has a causal effect on the other.

50. The assumption of no feedback is especially important in litigation, because it is possible for the defendant (if responsible, for example, for price fixing or discrimination) to affect the values of the explanatory variables and thus to bias the usual statistical tests that are used in multiple regression.

51. When both effects occur at the same time, this is described as “simultaneity.”

Figure 1. Feedback.



As a general rule, there are no basic direct statistical tests for determining the direction of causality; rather, the expert, when asked, should be prepared to defend his or her assumption based on an understanding of the underlying behavior evidence relating to the businesses or individuals involved.⁵²

Although there is no single approach that is entirely suitable for estimating models when the dependent variable affects one or more explanatory variables, one possibility is for the expert to drop the questionable variable from the regression to determine whether the variable's exclusion makes a difference. If it does not, the issue becomes moot. Another approach is for the expert to expand the multiple regression model by adding one or more equations that explain the relationship between the explanatory variable in question and the dependent variable.

Suppose, for example, that in a salary-based sex discrimination suit the defendant's expert considers employer-evaluated test scores to be an appropriate explanatory variable for the dependent variable, salary. If the plaintiff were to provide information that the employer adjusted the test scores in a manner that penalized women, the assumption that salaries were determined by test scores and not that test scores were affected by salaries might be invalid. If it is clearly inappropriate,

52. There are statistical time-series tests for particular formulations of causality; see Pindyck & Rubinfeld, *supra* note 23, § 9.2.

the test-score variable should be removed from consideration. Alternatively, the information about the employer's use of the test scores could be translated into a second equation in which a new dependent variable—test score—is related to workers' salary and sex. A test of the hypothesis that salary and sex affect test scores would provide a suitable test of the absence of feedback.

2. *To what extent are the explanatory variables correlated with each other?*

It is essential in multiple regression analysis that the explanatory variable of interest not be correlated perfectly with one or more of the other explanatory variables. If there were perfect correlation between two variables, the expert could not separate out the effect of the variable of interest on the dependent variable from the effect of the other variable. In essence, there are two explanations for the same pattern in the data. Suppose, for example, that in a sex discrimination suit, a particular form of job experience is determined to be a valid source of high wages. If all men had the requisite job experience and all women did not, it would be impossible to tell whether wage differentials between men and women were the result of sex discrimination or differences in experience.

When two or more explanatory variables are correlated perfectly—that is, when there is *perfect collinearity*—one cannot estimate the regression parameters. The existing dataset does not allow one to distinguish between alternative competing explanations of the movement in the dependent variable. However, when two or more variables are highly, but not perfectly, correlated—that is, when there is *multicollinearity*—the regression can be estimated, but some concerns remain. The greater the multicollinearity between two variables, the less precise are the estimates of individual regression parameters, and an expert is less able to distinguish among competing explanations for the movement in the outcome variable (even though there is no problem in estimating the joint influence of the two variables and all other regression parameters).⁵³

Fortunately, the reported regression statistics take into account any multicollinearity that might be present.⁵⁴ It is important to note as a corollary, however, that a failure to find a strong relationship between a variable of interest and

53. See *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971) (The court argued that an education requirement was one rationalization of the data, but racial discrimination was another. If you had put both race and education in the regression, it would have been asking too much of the data to tell which variable was doing the real work, because education and race were so highly correlated in the market at that time.).

54. See *Denny v. Westfield State College*, 669 F. Supp. 1146, 1149 (D. Mass. 1987) (The court accepted the testimony of one expert that “the presence of multicollinearity would merely tend to *overestimate* the amount of error associated with the estimate. . . . In other words, *p*-values will be artificially higher than they would be if there were no multicollinearity present.”) (emphasis added); *In re High Fructose Corn Syrup Antitrust Litig.*, 295 F.3d 651, 659 (7th Cir. Ill. 2002) (refusing to second-guess district court's admission of regression analyses that addressed multicollinearity in different ways).

a dependent variable need not imply that there is no relationship.⁵⁵ A relatively small sample, or even a large sample with substantial multicollinearity, may not provide sufficient information for the expert to determine whether there is a relationship.

3. To what extent are individual errors in the regression model independent?

If the expert calculated the parameters of a multiple regression model using as data the entire population, the estimates might still measure the model's population parameters with error. Errors can arise for a number of reasons, including (1) the failure of the model to include the appropriate explanatory variables, (2) the failure of the model to reflect any nonlinearities that might be present, and (3) the inclusion of inappropriate variables in the model. (Of course, further sources of error will arise if a sample, or subset, of the population is used to estimate the regression parameters.)

It is useful to view the cumulative effect of all of these sources of modeling error as being represented by an additional variable, the error term, in the multiple regression model. An important assumption in multiple regression analysis is that the error term and each of the explanatory variables are independent of each other. (If the error term and an explanatory variable are independent, they are not correlated with each other.) To the extent this is true, the expert can estimate the parameters of the model without bias; the magnitude of the error term will affect the precision with which a model parameter is estimated, but will not cause that estimate to be consistently too high or too low.

The assumption of independence may be inappropriate in a number of circumstances. In some instances, failure of the assumption makes multiple regression analysis an unsuitable statistical technique; in other instances, modifications or adjustments within the regression framework can be made to accommodate the failure.

The independence assumption may fail, for example, in a study of individual behavior over time, in which an unusually high error value in one time period is likely to lead to an unusually high value in the next time period. For example, if an economic forecaster underpredicted this year's Gross Domestic Product, he or she is likely to underpredict next year's as well; the factor that caused the prediction error (e.g., an incorrect assumption about Federal Reserve policy) is likely to be a source of error in the future.

55. If an explanatory variable of concern and another explanatory variable are highly correlated, dropping the second variable from the regression can be instructive. If the coefficient on the explanatory variable of concern becomes significant, a relationship between the dependent variable and the explanatory variable of concern is suggested.

Alternatively, the assumption of independence may fail in a study of a group of firms at a particular point in time, in which error terms for large firms are systematically higher than error terms for small firms. For example, an analysis of the profitability of firms may not accurately account for the importance of advertising as a source of increased sales and profits. To the extent that large firms advertise more than small firms, the regression errors would be large for the large firms and small for the small firms. A third possibility is that the dependent variable varies at the individual level, but the explanatory variable of interest varies only at the level of a group. For example, an expert might be viewing the price of a product in an antitrust case as a function of a variable or variables that measure the marketing channel through which the product is sold (e.g., wholesale or retail). In this case, errors within each of the marketing groups are likely not to be independent. Failure to account for this could cause the expert to overstate the statistical significance of the regression parameters.

In some instances, there are statistical tests that are appropriate for evaluating the independence assumption.⁵⁶ If the assumption has failed, the expert should ask first whether the source of the lack of independence is the omission of an important explanatory variable from the regression. If so, that variable should be included when possible, or the potential effect of its omission should be estimated when inclusion is not possible. If there is no important missing explanatory variable, the expert should apply one or more procedures that modify the standard multiple regression technique to allow for more accurate estimates of the regression parameters.⁵⁷

4. To what extent are the regression results sensitive to individual data points?

Estimated regression coefficients can be highly sensitive to particular data points. Suppose, for example, that one data point deviates greatly from its expected value, as indicated by the regression equation, while the remaining data points show

56. In a time-series analysis, the correlation of error values over time, the “serial correlation,” can be tested (in most instances) using a number of tests, including the Durbin-Watson test. The possibility that some error terms are consistently high in magnitude and others are systematically low, heteroscedasticity can also be tested in a number of ways. See, e.g., Pindyck & Rubinfeld, *supra* note 23, at 146–59. When serial correlation and/or heteroscedasticity are present, the standard errors associated with the estimated coefficients must be modified. For a discussion of the use of such “robust” standard errors, see Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, ch. 8 (4th ed. 2009).

57. When serial correlation is present, a number of closely related statistical methods are appropriate, including generalized differencing (a type of generalized least squares) and maximum likelihood estimation. When heteroscedasticity is the problem, weighted least squares and maximum likelihood estimation are appropriate. See, e.g., *id.* All these techniques are readily available in a number of statistical computer packages. They also allow one to perform the appropriate statistical tests of the significance of the regression coefficients.

little deviation. It would not be unusual in this situation for the coefficients in a multiple regression to change substantially if the data point in question were removed from the sample.

Evaluating the robustness of multiple regression results is a complex endeavor. Consequently, there is no agreed set of tests for robustness that analysts should apply. In general, it is important to explore the reasons for unusual data points. If the source is an error in recording data, the appropriate corrections can be made. If all the unusual data points have certain characteristics in common (e.g., they all are associated with a supervisor who consistently gives high ratings in an equal pay case), the regression model should be modified appropriately.

One generally useful diagnostic technique is to determine to what extent the estimated parameter changes as each data point in the regression analysis is dropped from the sample. An *influential* data point—a point that causes the estimated parameter to change substantially—should be studied further to determine whether mistakes were made in the use of the data or whether important explanatory variables were omitted.⁵⁸

5. *To what extent are the data subject to measurement error?*

In multiple regression analysis it is assumed that variables are measured accurately.⁵⁹ If there are measurement errors in the dependent variable, estimates of regression parameters will be less accurate, although they will not necessarily be biased. However, if one or more independent variables are measured with error, the corresponding parameter estimates are likely to be biased, typically toward zero (and other coefficient estimates are likely to be biased as well).

To understand why, suppose that the dependent variable, salary, is measured without error, and the explanatory variable, experience, is subject to measurement error. (Seniority or years of experience should be accurate, but the type of experience is subject to error, because applicants may overstate previous job responsibilities.) As the measurement error increases, the estimated parameter associated with the experience variable will tend toward zero, that is, eventually, there will be no relationship between salary and experience.

It is important for any source of measurement error to be carefully evaluated. In some circumstances, little can be done to correct the measurement-error prob-

58. A more complete and formal treatment of the robustness issue appears in David A. Belsley et al., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* 229–44 (1980). For a useful discussion of the detection of outliers and the evaluation of influential data points, see R.D. Cook & S. Weisberg, *Residuals and Influence in Regression* (Monographs on Statistics and Applied Probability No. 18, 1982). For a broad discussion of robust regression methods, see Peer J. Rousseeuw & Annick M. Leroy, *Robust Regression and Outlier Detection* (2004).

59. Inaccuracy can occur not only in the precision with which a particular variable is measured, but also in the precision with which the variable to be measured corresponds to the appropriate theoretical construct specified by the regression model.

lem; the regression results must be interpreted in that light. In other circumstances, however, the expert can correct measurement error by finding a new, more reliable data source. Finally, alternative estimation techniques (using related variables that are measured without error) can be applied to remedy the measurement-error problem in some situations.⁶⁰

IV. The Expert

Multiple regression analysis is taught to students in extremely diverse fields, including statistics, economics, political science, sociology, psychology, anthropology, public health, and history. Nonetheless, the methodology is difficult to master, necessitating a combination of technical skills (the science) and experience (the art). This naturally raises two questions:

1. Who should be qualified as an expert?
2. When and how should the court appoint an expert to assist in the evaluation of statistical issues, including those relating to multiple regression?

A. Who Should Be Qualified as an Expert?

Any individual with substantial training in and experience with multiple regression and other statistical methods may be qualified as an expert.⁶¹ A doctoral degree in a discipline that teaches theoretical or applied statistics, such as economics, history, and psychology, usually signifies to other scientists that the proposed expert meets this preliminary test of the qualification process.

The decision to qualify an expert in regression analysis rests with the court. Clearly, the proposed expert should be able to demonstrate an understanding of the discipline. Publications relating to regression analysis in peer-reviewed journals, active memberships in related professional organizations, courses taught on regression methods, and practical experience with regression analysis can indicate a professional's expertise. However, the expert's background and experience with the specific issues and tools that are applicable to a particular case should also be considered during the qualification process. Thus, if the regression methods are being utilized to evaluate damages in an antitrust case, the qualified expert should have sufficient qualifications in economic analysis as well as statistics. An individual whose expertise lies solely with statistics will be limited in his or her ability to evaluate the usefulness of alternative economic models. Similarly, if a case involves

60. See, e.g., Pindyck & Rubinfeld, *supra* note 23, at 178–98 (discussion of instrumental variables estimation).

61. A proposed expert whose only statistical tool is regression analysis may not be able to judge when a statistical analysis should be based on an approach other than regression analysis.

eyewitness identification, a background in psychology as well as statistics may provide essential qualifying elements.

B. Should the Court Appoint a Neutral Expert?

There are conflicting views on the issue of whether court-appointed experts should be used. In complex cases in which two experts are presenting conflicting statistical evidence, the use of a “neutral” court-appointed expert can be advantageous. There are those who believe, however, that there is no such thing as a truly “neutral” expert. In any event, if an expert is chosen, that individual should have substantial expertise and experience—ideally, someone who is respected by both plaintiffs and defendants.⁶²

The appointment of such an expert is likely to influence the presentation of the statistical evidence by the experts for the parties in the litigation. The neutral expert will have an incentive to present a balanced position that relies on broad principles for which there is consensus in the community of experts. As a result, the parties’ experts can be expected to present testimony that confronts core issues that are likely to be of concern to the court and that is sufficiently balanced to be persuasive to the court-appointed expert.⁶³

Rule 706 of the Federal Rules of Evidence governs the selection and instruction of court-appointed experts. In particular:

1. The expert should be notified of his or her duties through a written court order or at a conference with the parties.
2. The expert should inform the parties of his or her findings orally or in writing.
3. If deemed appropriate by the court, the expert should be available to testify and may be deposed or cross-examined by any party.
4. The court must determine the expert’s compensation.⁶⁴
5. The parties should be free to utilize their own experts.

Although not required by Rule 706, it will usually be advantageous for the court to opt for the appointment of a neutral expert as early in the litigation process as possible. It will also be advantageous to minimize any ex parte contact with

62. Judge Posner notes in *In re High Fructose Corn Syrup Antitrust Litig.*, 295 F.2d 651, 665 (7th Cir., 2002), “the judge and jury can repose a degree of confidence in his testimony that it could not repose in that of a party’s witness. The judge and the jury may not understand the neutral expert perfectly but at least they will know that he has no axe to grind, and so, to a degree anyway, they will be able to take his testimony on faith.”

63. For a discussion of the presentation of expert evidence generally, including the use of court-appointed experts, see Samuel R. Gross, *Expert Evidence*, 1991 Wis. L. Rev. 1113 (1991).

64. Although Rule 706 states that the compensation must come from public funds, complex litigation may be sufficiently costly as to require that the parties share the costs of the neutral expert.

the neutral expert; this will diminish the possibility that one or both parties will come to the view that the court's ultimate opinion was unreasonably influenced by the neutral expert.

Rule 706 does not offer specifics as to the process of appointment of a court-appointed expert. One possibility is to have the parties offer a short list of possible appointees. If there was no common choice, the court could select from the combined list, perhaps after allowing each party to exercise one or more peremptory challenges. Another possibility is to obtain a list of recommended experts from a selection of individuals known to be experts in the field.

V. Presentation of Statistical Evidence

The costs of evaluating statistical evidence can be reduced and the precision of that evidence increased if the discovery process is used effectively. In evaluating the admissibility of statistical evidence, courts should consider the following issues:

1. Has the expert provided sufficient information to replicate the multiple regression analysis?
2. Are the expert's methodological choices reasonable, or are they arbitrary and unjustified?

A. What Disagreements Exist Regarding Data on Which the Analysis Is Based?

In general, a clear and comprehensive statement of the underlying research methodology is a requisite part of the discovery process. The expert should be encouraged to reveal both the nature of the experimentation carried out and the sensitivity of the results to the data and to the methodology.

The following suggestions are useful requirements that can substantially improve the discovery process:

1. To the extent possible, the parties should be encouraged to agree to use a common database. Even if disagreement about the significance of the data remains, early agreement on a common database can help focus the discovery process on the important issues in the case.
2. A party that offers data to be used in statistical work, including multiple regression analysis, should be encouraged to provide the following to the other parties: (a) a hard copy of the data when available and manageable in size, along with the underlying sources; (b) computer disks or tapes on which the data are recorded; (c) complete documentation of the disks or tapes; (d) computer programs that were used to generate the data (in hard

copy if necessary, but preferably on a computer disk or tape, or both); and (e) documentation of such computer programs. The documentation should be sufficiently complete and clear so that the opposing expert can reproduce all of the statistical work.

3. A party offering data should make available the personnel involved in the compilation of such data to answer the other parties' technical questions concerning the data and the methods of collection or compilation.
4. A party proposing to offer an expert's regression analysis at trial should ask the expert to fully disclose (a) the database and its sources,⁶⁵ (b) the method of collecting the data, and (c) the methods of analysis. When possible, this disclosure should be made sufficiently in advance of trial so that the opposing party can consult its experts and prepare cross-examination. The court must decide on a case-by-case basis where to draw the disclosure line.
5. An opposing party should be given the opportunity to object to a database or to a proposed method of analysis of the database to be offered at trial. Objections may be to simple clerical errors or to more complex issues relating to the selection of data, the construction of variables, and, on occasion, the particular form of statistical analysis to be used. Whenever possible, these objections should be resolved before trial.
6. The parties should be encouraged to resolve differences as to the appropriateness and precision of the data to the extent possible by informal conference. The court should make an effort to resolve differences before trial.

These suggestions are motivated by the objective of improving the discovery process to make it more informative. The fact that these questions may raise some doubts or concerns about a particular regression model should not be taken to mean that the model does not provide useful information. It does, however, take considerable skill for an expert to determine the extent to which information is useful when the model being utilized has some shortcomings.

*B. Which Database Information and Analytical Procedures Will Aid in Resolving Disputes over Statistical Studies?*⁶⁶

To help resolve disputes over statistical studies, experts should follow the guidelines below when presenting database information and analytical procedures:

65. These sources would include all variables used in the statistical analyses conducted by the expert, not simply those variables used in a final analysis on which the expert expects to rely.

66. For a more complete discussion of these requirements, see *The Evolving Role of Statistical Assessments as Evidence in the Courts*, app. F at 256 (Stephen E. Fienberg ed., 1989) (Recommended

1. The expert should state clearly the objectives of the study, as well as the time frame to which it applies and the statistical population to which the results are being projected.
2. The expert should report the units of observation (e.g., consumers, businesses, or employees).
3. The expert should clearly define each variable.
4. The expert should clearly identify the sample for which data are being studied,⁶⁷ as well as the method by which the sample was obtained.
5. The expert should reveal if there are missing data, whether caused by a lack of availability (e.g., in business data) or nonresponse (e.g., in survey data), and the method used to handle the missing data (e.g., deletion of observations).
6. The expert should report investigations into errors associated with the choice of variables and assumptions underlying the regression model.
7. If samples were chosen randomly from a population (i.e., probability sampling procedures were used),⁶⁸ the expert should make a good-faith effort to provide an estimate of a sampling error, the measure of the difference between the sample estimate of a parameter (such as the mean of a dependent variable under study), and the (unknown) population parameter (the population mean of the variable).⁶⁹
8. If probability sampling procedures were not used, the expert should report the set of procedures that was used to minimize sampling errors.

Standards on Disclosure of Procedures Used for Statistical Studies to Collect Data Submitted in Evidence in Legal Cases).

67. The sample information is important because it allows the expert to make inferences about the underlying population.

68. In probability sampling, each representative of the population has a known probability of being in the sample. Probability sampling is ideal because it is highly structured, and in principle, it can be replicated by others. Nonprobability sampling is less desirable because it is often subjective, relying to a large extent on the judgment of the expert.

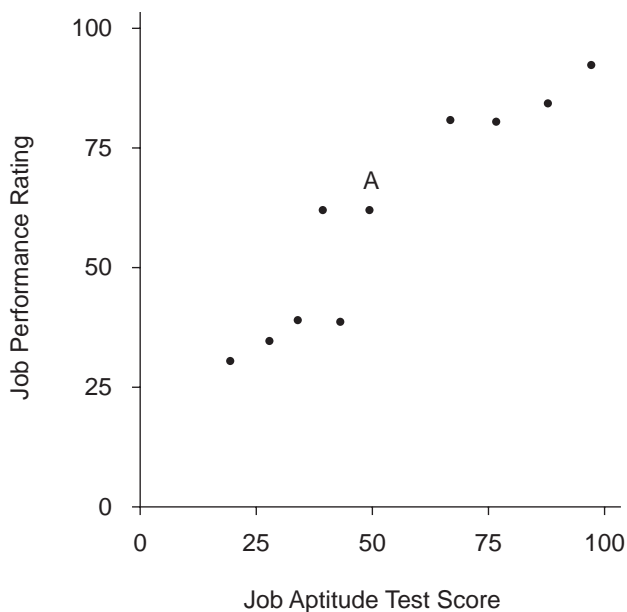
69. Sampling error is often reported in terms of standard errors or confidence intervals. See Appendix, *infra*, for details.

Appendix: The Basics of Multiple Regression

A. Introduction

This appendix illustrates, through examples, the basics of multiple regression analysis in legal proceedings. Often, visual displays are used to describe the relationship between variables that are used in multiple regression analysis. Figure 2 is a scatterplot that relates scores on a job aptitude test (shown on the x -axis) and job performance ratings (shown on the y -axis). Each point on the scatterplot shows where a particular individual scored on the job aptitude test and how his or her job performance was rated. For example, the individual represented by Point A in Figure 2 scored 49 on the job aptitude test and had a job performance rating of 62.

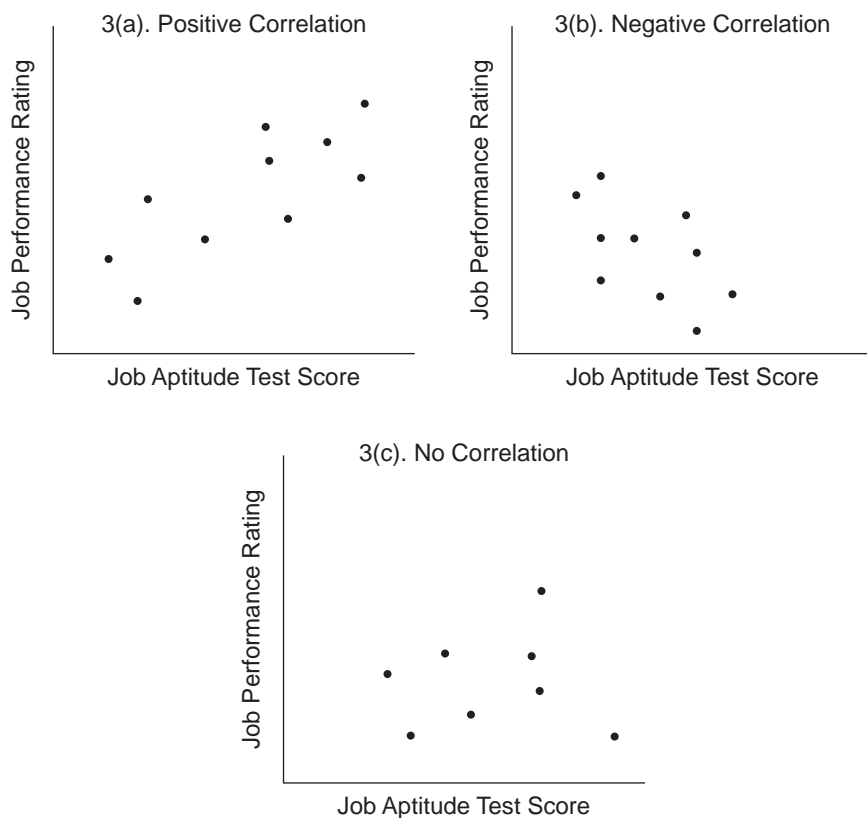
Figure 2. Scatterplot of scores on a job aptitude test relative to job performance rating.



The relationship between two variables can be summarized by a correlation coefficient, which ranges in value from -1 (a perfect negative relationship) to $+1$ (a perfect positive relationship). Figure 3 depicts three possible relationships between the job aptitude variable and the job performance variable. In Figure 3(a), there is a positive correlation: In general, higher job performance ratings are associated with higher aptitude test scores, and lower job performance ratings are associated with lower aptitude test scores. In Figure 3(b), the correlation is

negative: Higher job performance ratings are associated with lower aptitude test scores, and lower job performance ratings are associated with higher aptitude test scores. Positive and negative correlations can be relatively strong or relatively weak. If the relationship is sufficiently weak, there is effectively no correlation, as is illustrated in Figure 3(c).

Figure 3. Correlation between the job aptitude variable and the job performance variable: (a) positive correlation, (b) negative correlation, (c) weak relationship with no correlation.



Multiple regression analysis goes beyond the calculation of correlations; it is a method in which a regression line is used to relate the average of one variable—the dependent variable—to the values of other explanatory variables. As a result, regression analysis can be used to predict the values of one variable using the values of others. For example, if average job performance ratings depend on aptitude test scores, regression analysis can use information about test scores to predict job performance.

A regression line is the best-fitting straight line through a set of points in a scatterplot. If there is only one explanatory variable, the straight line is defined by the equation

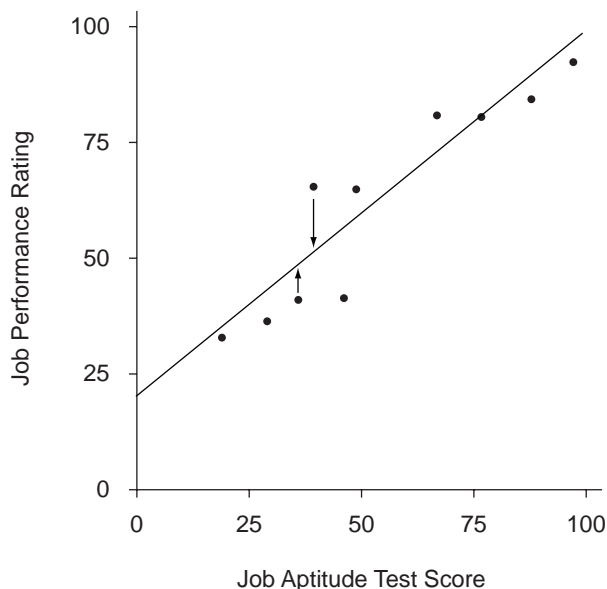
$$Y = a + bX. \quad (1)$$

In equation (1), a is the intercept of the line with the y -axis when X equals 0, and b is the slope—the change in the dependent variable associated with a 1-unit change in the explanatory variable. In Figure 4, for example, when the aptitude test score is 0, the predicted (average) value of the job performance rating is the intercept, 18.4. Also, for each additional point on the test score, the job performance rating increases .73 units, which is given by the slope .73. Thus, the estimated regression line is

$$Y = 18.4 + .73X. \quad (2)$$

The regression line typically is estimated using the standard method of least squares, where the values of a and b are calculated so that the sum of the squared deviations of the points from the line are minimized. In this way, positive deviations and negative deviations of equal size are counted equally, and large deviations are counted more than small deviations. In Figure 4 the deviation lines are verti-

Figure 4. Regression line.



cal because the equation is predicting job performance ratings from aptitude test scores, not aptitude test scores from job performance ratings.

The important variables that systematically might influence the dependent variable, and for which data can be obtained, typically should be included explicitly in a statistical model. All remaining influences, which should be small individually, but can be substantial in the aggregate, are included in an additional random error term.⁷⁰ Multiple regression is a procedure that separates the systematic effects (associated with the explanatory variables) from the random effects (associated with the error term) and also offers a method of assessing the success of the process.

B. Linear Regression Model

When there are an arbitrary number of explanatory variables, the linear regression model takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon \quad (3)$$

where Y represents the dependent variable, such as the salary of an employee, and $X_1 \dots X_k$ represent the explanatory variables (e.g., the experience of each employee and his or her sex, coded as a 1 or 0, respectively). The error term, ϵ , represents the collective unobservable influence of any omitted variables. In a linear regression, each of the terms being added involves unknown parameters, $\beta_0, \beta_1, \dots, \beta_k$,⁷¹ which are estimated by “fitting” the equation to the data using least squares.

Each estimated coefficient β_k measures how the dependent variable Y responds, on average, to a change in the corresponding covariate X_k , after “controlling for” all the other covariates. The informal phrase “controlling for” has a specific statistical meaning. Consider the following three-step procedure. First, we calculate the residuals from a regression of Y on all covariates other than X_k . Second, we calculate the residuals of a regression of X_k on all the other covariates. Third, and finally, we regress the first residual variable on the second residual variable. The resulting coefficient will be identically equal to β_k . Thus, the coeffi-

70. It is clearly advantageous for the random component of the regression relationship to be small relative to the variation in the dependent variable.

71. The variables themselves can appear in many different forms. For example, Y might represent the logarithm of an employee’s salary, and X_1 might represent the logarithm of the employee’s years of experience. The logarithmic representation is appropriate when Y increases exponentially as X increases—for each unit increase in X , the corresponding increase in Y becomes larger and larger. For example, if an expert were to graph the growth of the U.S. population (Y) over time (t), the following equation might be appropriate:

$$\log(Y) = \beta_0 + \beta_1 \log(t).$$

cient in a multiple regression represents the slope of the line “ Y , adjusted for all covariates other than X_k versus X_k adjusted for all the other covariates.”⁷²

Most statisticians use the least squares regression technique because of its simplicity and its desirable statistical properties. As a result, it also is used frequently in legal proceedings.

1. Specifying the regression model

Suppose an expert wants to analyze the salaries of women and men at a large publishing house to discover whether a difference in salaries between employees with similar years of work experience provides evidence of discrimination.⁷³ To begin with the simplest case, Y , the salary in dollars per year, represents the dependent variable to be explained, and X_1 represents the explanatory variable—the number of years of experience of the employee. The regression model would be written

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon. \quad (4)$$

In equation (4), β_0 and β_1 are the parameters to be estimated from the data, and ε is the random error term. The parameter β_0 is the average salary of all employees with no experience. The parameter β_1 measures the average effect of an additional year of experience on the average salary of employees.

2. Regression line

Once the parameters in a regression equation, such as equation (3), have been estimated, the fitted values for the dependent variable can be calculated. If we denote the estimated regression parameters, or regression coefficients, for the model in equation (3) by $\beta_0, \beta_1, \dots, \beta_k$, the fitted values for Y , denoted \hat{Y} , are given by

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k. \quad (5)$$

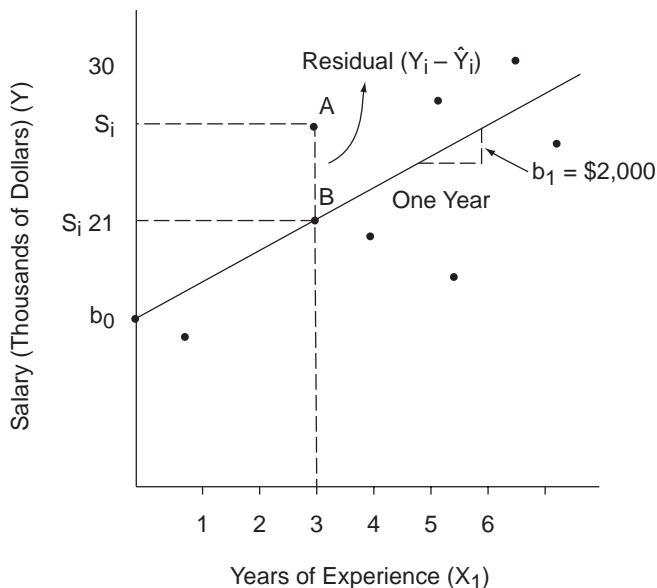
Figure 5 illustrates this for the example involving a single explanatory variable. The data are shown as a scatter of points; salary is on the vertical axis, and years of experience is on the horizontal axis. The estimated regression line is drawn through the data points. It is given by

$$\hat{Y} = \$15,000 + \$2000X_1. \quad (6)$$

72. In econometrics, this is known as the **Frisch–Waugh–Lovell theorem**.

73. The regression results used in this example are based on data for 1715 men and women, which were used by the defense in a sex discrimination case against the *New York Times* that was settled in 1978. Professor Orley Ashenfelter, Department of Economics, Princeton University, provided the data.

Figure 5. Goodness of fit.



Thus, the fitted value for the salary associated with an individual's years of experience X_{1i} is given by

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} \text{ (at Point B).} \quad (7)$$

The intercept of the straight line is the average value of the dependent variable when the explanatory variable or variables are equal to 0; the intercept β_0 is shown on the vertical axis in Figure 5. Similarly, the slope of the line measures the (average) change in the dependent variable associated with a unit increase in an explanatory variable; the slope β_1 also is shown. In equation (6), the intercept \$15,000 indicates that employees with no experience earn \$15,000 per year. The slope parameter implies that each year of experience adds \$2000 to an "average" employee's salary.

Now, suppose that the salary variable is related simply to the sex of the employee. The relevant indicator variable, often called a dummy variable, is X_2 , which is equal to 1 if the employee is male, and 0 if the employee is female. Suppose the regression of salary Y on X_2 yields the following result: $Y = \$30,449 + \$10,979X_2$. The coefficient \$10,979 measures the difference between the average salary of men and the average salary of women.⁷⁴

74. To understand why, note that when X_2 equals 0, the average salary for women is $\$30,449 + \$10,979 \cdot 0 = \$30,449$. Correspondingly, when $X_2 = 1$, the average salary for men is $\$30,449 + \$10,979 \cdot 1 = \$41,428$. The difference, $\$41,428 - \$30,449$, is \$10,979.

a. Regression residuals

For each data point, the regression residual is the difference between the actual values and fitted values of the dependent variable. Suppose, for example, that we are studying an individual with 3 years of experience and a salary of \$27,000. According to the regression line in Figure 5, the average salary of an individual with 3 years of experience is \$21,000. Because the individual's salary is \$6000 higher than the average salary, the residual (the individual's salary minus the average salary) is \$6000. In general, the residual e associated with a data point, such as Point A in Figure 5, is given by $e_i = Y_i - \hat{Y}_i$. Each data point in the figure has a residual, which is the error made by the least squares regression method for that individual.

b. Nonlinearities

Nonlinear models account for the possibility that the effect of an explanatory variable on the dependent variable may vary in magnitude as the level of the explanatory variable changes. One useful nonlinear model uses interactions among variables to produce this effect. For example, suppose that

$$S = \beta_1 + \beta_2 \text{SEX} + \beta_3 \text{EXP} + \beta_4 (\text{EXP})(\text{SEX}) + \epsilon \quad (8)$$

where S is annual salary, SEX is equal to 1 for women and 0 for men, EXP represents years of job experience, and ϵ is a random error term. The coefficient β_2 measures the difference in average salary (across all experience levels) between men and women for employees with no experience. The coefficient β_3 measures the effect of experience on salary for men (when $\text{SEX} = 0$), and the coefficient β_4 measures the difference in the effect of experience on salary between men and women. It follows, for example, that the effect of 1 year of experience on salary for men is β_3 , whereas the comparable effect for women is $\beta_3 + \beta_4$.⁷⁵

C. Interpreting Regression Results

To explain how regression results are interpreted, we can expand the earlier example associated with Figure 5 to consider the possibility of an additional explanatory variable—the square of the number of years of experience, X_3 . The X_3 variable is designed to capture the fact that for most individuals, salaries increase with experience, but eventually salaries tend to level off. The estimated regression line using the third additional explanatory variable, as well as the first explanatory variable for years of experience (X_1) and the dummy variable for sex (X_2), is

75. Estimating a regression in which there are interaction terms for all explanatory variables, as in equation (8), is essentially the same as estimating two separate regressions, one for men and one for women.

$$\hat{Y} = \$14,085 + \$2323X_1 + \$1675X_2 - \$36X_3. \quad (9)$$

The importance of including relevant explanatory variables in a regression model is illustrated by the change in the regression results after the X_3 and X_1 variables are added. The coefficient on the variable X_2 measures the difference in the salaries of men and women while controlling for the effect of experience. The differential of \$1675 is substantially lower than the previously measured differential of \$10,979. Clearly, failure to control for job experience in this example leads to an overstatement of the difference in salaries between men and women.

Now consider the interpretation of the explanatory variables for experience, X_1 and X_3 . The positive sign on the X_1 coefficient shows that salary increases with experience. The negative sign on the X_3 coefficient indicates that the rate of salary increase decreases with experience. To determine the combined effect of the variables X_1 and X_3 , some simple calculations can be made. For example, consider how the average salary of women ($X_2 = 0$) changes with the level of experience. As experience increases from 0 to 1 year, the average salary increases by \$2251, from \$14,085 to \$16,336. However, women with 2 years of experience earn only \$2179 more than women with 1 year of experience, and women with 1 year of experience earn only \$2127 more than women with 2 years. Furthermore, women with 7 years of experience earn \$28,582 per year, which is only \$1855 more than the \$26,727 earned by women with 6 years of experience.⁷⁶ Figure 6 illustrates the results: The regression line shown is for women's salaries; the corresponding line for men's salaries would be parallel and \$1675 higher.

D. Determining the Precision of the Regression Results

Least squares regression provides not only parameter estimates that indicate the direction and magnitude of the effect of a change in the explanatory variable on the dependent variable, but also an estimate of the reliability of the parameter estimates and a measure of the overall goodness of fit of the regression model. Each of these factors is considered in turn.

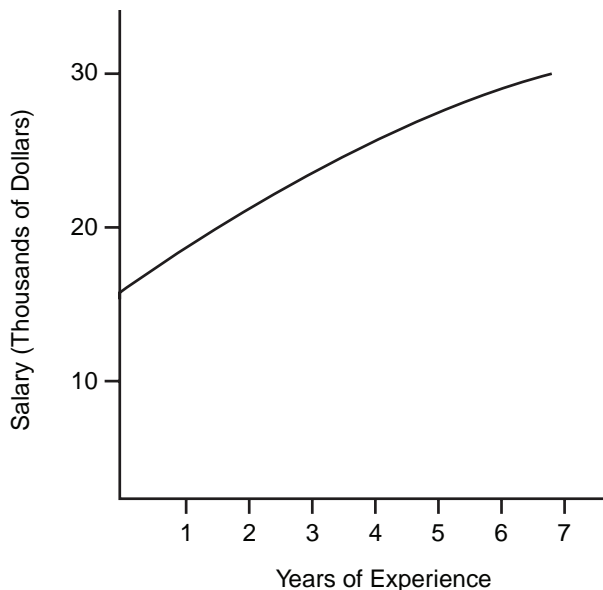
1. Standard errors of the coefficients and *t*-statistics

Estimates of the true but unknown parameters of a regression model are numbers that depend on the particular sample of observations under study. If a different sample were used, a different estimate would be calculated.⁷⁷ If the expert continued to collect more and more samples and generated additional estimates, as might happen when new data became available over time, the estimates of each

76. These numbers can be calculated by substituting different values of X_1 and X_3 in equation (9).

77. The least squares formula that generates the estimates is called the least squares estimator, and its values vary from sample to sample.

Figure 6. Regression slope for women's salaries and men's salaries.



parameter would follow a probability distribution (i.e., the expert could determine the percentage or frequency of the time that each estimate occurs). This probability distribution can be summarized by a mean and a measure of dispersion around the mean, a standard deviation, which usually is referred to as the standard error of the coefficient, or the standard error (SE).⁷⁸

Suppose, for example, that an expert is interested in estimating the average price paid for a gallon of unleaded gasoline by consumers in a particular geographic area of the United States at a particular point in time. The mean price for a sample of 10 gas stations might be \$1.25, while the mean for another sample might be \$1.29, and the mean for a third, \$1.21. On this basis, the expert also could calculate the overall mean price of gasoline to be \$1.25 and the standard deviation to be \$0.04.

Least squares regression generalizes this result, by calculating means whose values depend on one or more explanatory variables. The standard error of a regression coefficient tells the expert how much parameter estimates are likely to vary from sample to sample. The greater the variation in parameter estimates from sample to sample, the larger the standard error and consequently the less reliable the regression results. Small standard errors imply results that are likely to

78. See David H. Kaye & David A. Freedman, Reference Guide on Statistics, Section IV.A, in this manual.

be similar from sample to sample, whereas results with large standard errors show more variability.

Under appropriate assumptions, the least squares estimators provide “best” determinations of the true underlying parameters.⁷⁹ In fact, least squares has several desirable properties. First, least squares estimators are unbiased. Intuitively, this means that if the regression were calculated repeatedly with different samples, the average of the many estimates obtained for each coefficient would be the true parameter. Second, least squares estimators are consistent; if the sample were very large, the estimates obtained would come close to the true parameters. Third, least squares is efficient, in that its estimators have the smallest variance among all (linear) unbiased estimators.

If the further assumption is made that the probability distribution of each of the error terms is known, statistical statements can be made about the precision of the coefficient estimates. For relatively large samples (often, thirty or more data points will be sufficient for regressions with a small number of explanatory variables), the probability that the estimate of a parameter lies within an interval of 2 standard errors around the true parameter is approximately .95, or 95%. A frequent, although not always appropriate, assumption in statistical work is that the error term follows a normal distribution, from which it follows that the estimated parameters are normally distributed. The normal distribution has the property that the area within 1.96 standard errors of the mean is equal to 95% of the total area. Note that the normality assumption is not necessary for least squares to be used, because most of the properties of least squares apply regardless of normality.

In general, for any parameter estimate b , the expert can construct an interval around b such that there is a 95% probability that the interval covers the true parameter. This 95% confidence interval⁸⁰ is given by⁸¹

$$b \pm 1.96 (\text{SE of } b). \quad (10)$$

The expert can test the hypothesis that a parameter is actually equal to 0 (often stated as testing the null hypothesis) by looking at its t -statistic, which is defined as

$$t = \frac{b}{\text{SE}(b)}. \quad (11)$$

79. The necessary assumptions of the regression model include (a) the model is specified correctly, (b) errors associated with each observation are drawn randomly from the same probability distribution and are independent of each other, (c) errors associated with each observation are independent of the corresponding observations for each of the explanatory variables in the model, and (d) no explanatory variable is correlated perfectly with a combination of other variables.

80. Confidence intervals are used commonly in statistical analyses because the expert can never be certain that a parameter estimate is equal to the true population parameter.

81. If the number of data points in the sample is small, the standard error must be multiplied by a number larger than 1.96.

Reference Guide on Multiple Regression

If the t -statistic is less than 1.96 in magnitude, the 95% confidence interval around b must include 0.⁸² Because this means that the expert cannot reject the hypothesis that β equals 0, the estimate, whatever it may be, is said to be not statistically significant. Conversely, if the t -statistic is greater than 1.96 in absolute value, the expert concludes that the true value of β is unlikely to be 0 (intuitively, b is “too far” from 0 to be consistent with the true value of β being 0). In this case, the expert rejects the hypothesis that β equals 0 and calls the estimate statistically significant. If the null hypothesis β equals 0 is true, using a 95% confidence level will cause the expert to falsely reject the null hypothesis 5% of the time. Consequently, results often are said to be significant at the 5% level.⁸³

As an example, consider a more complete set of regression results associated with the salary regression described in equation (9):

$$\begin{array}{rcccc} \hat{Y} & = & \$14,085 & + & \$2323X_1 & + & \$1675X_2 & - & \$36X_3 \\ & & (1577) & & (140) & & (1435) & & (3.4) \\ t & = & 8.9 & & 16.5 & & 1.2 & & -10.8. \end{array} \quad (12)$$

The standard error of each estimated parameter is given in parentheses directly below the parameter, and the corresponding t -statistics appear below the standard error values.

Consider the coefficient on the dummy variable X_2 . It indicates that \$1675 is the best estimate of the mean salary difference between men and women. However, the standard error of \$1435 is large in relation to its coefficient \$1675. Because the standard error is relatively large, the range of possible values for measuring the true salary difference, the true parameter, is great. In fact, a 95% confidence interval is given by

$$\$1675 \pm \$1435 \cdot 1.96 = \$1675 \pm \$2813. \quad (13)$$

In other words, the expert can have 95% confidence that the true value of the coefficient lies between -\$1138 and \$4488. Because this range includes 0, the effect of sex on salary is said to be insignificantly different from 0 at the 5% level. The t value of 1.2 is equal to \$1675 divided by \$1435. Because this t -statistic is less than 1.96 in magnitude (a condition equivalent to the inclusion of a 0 in the above confidence interval), the sex variable again is said to be an insignificant determinant of salary at the 5% level of significance.

82. The t -statistic applies to any sample size. As the sample gets large, the underlying distribution, which is the source of the t -statistic (Student's t -distribution), approximates the normal distribution.

83. A t -statistic of 2.57 in magnitude or greater is associated with a 99% confidence level, or a 1% level of significance, that includes a band of 2.57 standard deviations on either side of the estimated coefficient.

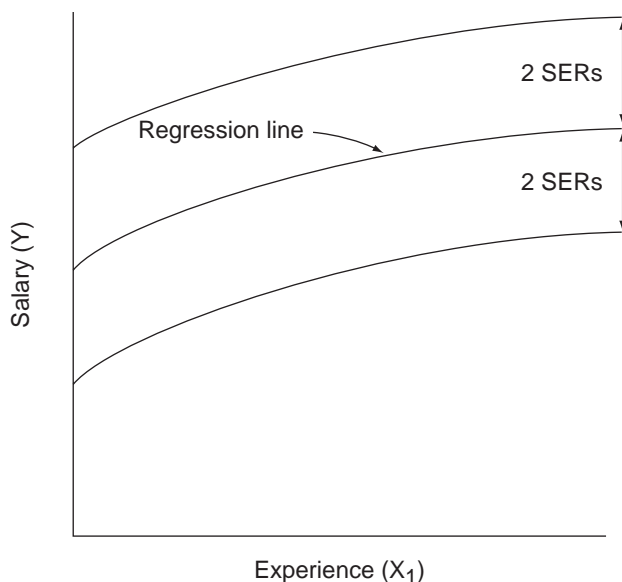
Note also that experience is a highly significant determinant of salary, because both the X_1 and the X_3 variables have t -statistics substantially greater than 1.96 in magnitude. More experience has a significant positive effect on salary, but the size of this effect diminishes significantly with experience.

2. Goodness of fit

Reported regression results usually contain not only the point estimates of the parameters and their standard errors or t -statistics, but also other information that tells how closely the regression line fits the data. One statistic, the standard error of the regression (SER), is an estimate of the overall size of the regression residuals.⁸⁴ An SER of 0 would occur only when all data points lie exactly on the regression line—an extremely unlikely possibility. Other things being equal, the larger the SER, the poorer the fit of the data to the model.

For a normally distributed error term, the expert would expect approximately 95% of the data points to lie within 2 SERs of the estimated regression line, as shown in Figure 7 (in Figure 7, the SER is approximately \$5000).

Figure 7. Standard error of the regression.



84. More specifically, it is a measure of the standard deviation of the regression error ϵ . It sometimes is called the root mean squared error of the regression line.

R -squared (R^2) is a statistic that measures the percentage of variation in the dependent variable that is accounted for by all the explanatory variables.⁸⁵ Thus, R^2 provides a measure of the overall goodness of fit of the multiple regression equation. Its value ranges from 0 to 1. An R^2 of 0 means that the explanatory variables explain none of the variation of the dependent variable; an R^2 of 1 means that the explanatory variables explain all of the variation. The R^2 associated with equation (12) is .56. This implies that the three explanatory variables explain 56% of the variation in salaries.

What level of R^2 , if any, should lead to a conclusion that the model is satisfactory? Unfortunately, there is no clear-cut answer to this question, because the magnitude of R^2 depends on the characteristics of the data being studied and, in particular, whether the data vary over time or over individuals. Typically, an R^2 is low in cross-sectional studies in which differences in individual behavior are explained. It is likely that these individual differences are caused by many factors that cannot be measured. As a result, the expert cannot hope to explain most of the variation. In time-series studies, in contrast, the expert is explaining the movement of aggregates over time. Because most aggregate time series have substantial growth, or trend, in common, it will not be difficult to “explain” one time series using another time series, simply because both are moving together. It follows as a corollary that a high R^2 does not by itself mean that the variables included in the model are the appropriate ones.

As a general rule, courts should be reluctant to rely solely on a statistic such as R^2 to choose one model over another. Alternative procedures and tests are available.⁸⁶

3. Sensitivity of least squares regression results

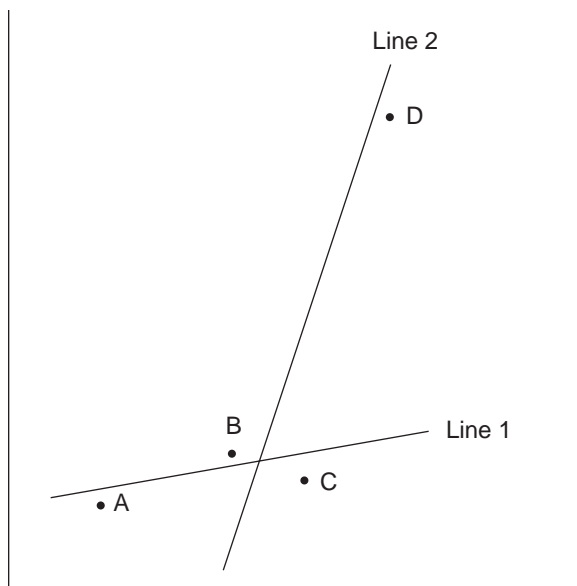
The least squares regression line can be sensitive to extreme data points. This sensitivity can be seen most easily in Figure 8. Assume initially that there are only three data points, A, B, and C, relating information about X_1 to the variable Y . The least squares line describing the best-fitting relationship between Points A, B, and C is represented by Line 1. Point D is called an *outlier* because it lies far from the regression line that fits the remaining points. When a new, best-fitting least squares line is reestimated to include Point D, Line 2 is obtained. Figure 8 shows that the outlier Point D is an *influential* data point, because it has a dominant effect on the slope and intercept of the least squares line. Because least squares attempts to minimize the sum of squared deviations, the sensitivity of the line to individual points sometimes can be substantial.⁸⁷

85. The variation is the square of the difference between each Y value and the average Y value, summed over all the Y values.

86. These include F -tests and specification error tests. See Pindyck & Rubinfeld, *supra* note 23, at 88–95, 128–36, 194–98.

87. This sensitivity is not always undesirable. In some instances it may be much more important to predict Point D when a big change occurs than to measure the effects of small changes accurately.

Figure 8. Least squares regression.



What makes the influential data problem even more difficult is that the effect of an outlier may not be seen readily if deviations are measured from the final regression line. The reason is that the influence of Point D on Line 2 is so substantial that its deviation from the regression line is not necessarily larger than the deviation of any of the remaining points from the regression line.⁸⁸ Although they are not as popular as least squares, alternative estimation techniques that are less sensitive to outliers, such as robust estimation, are available.

E. Reading Multiple Regression Computer Output

Statistical computer packages that report multiple regression analyses vary to some extent in the information they provide and the form that the information takes. Table 1 contains a sample of the basic computer output that is associated with equation (9).

88. The importance of an outlier also depends on its location in the dataset. Outliers associated with relatively extreme values of explanatory variables are likely to be especially influential. See, e.g., *Fisher v. Vassar College*, 70 F.3d 1420, 1436 (2d Cir. 1995) (court required to include assessment of “service in academic community,” because concept was too amorphous and not a significant factor in tenure review), *rev’d on other grounds*, 114 F.3d 1332 (2d Cir. 1997) (en banc).

Table 1. Regression Output

Dependent variable: Y					
	SSE	62346266124	F -test	174.71	
	DFE	561	Prob > F	0.0001	
	MSE	111134164	R^2	0.556	
Variable	DF	Parameter Estimate	Standard Error	t -Statistic	Prob > $ t $
Intercept	1	14,084.89	1577.484	8.9287	.0001
X_1	1	2323.17	140.70	16.5115	.0001
X_2	1	1675.11	1435.422	1.1670	.2437
X_3	1	-36.71	3.41	-10.7573	.0001

Note: SSE = sum of squared errors; DFE = degrees of freedom associated with the error term; MSE = mean squared error; DF = degrees of freedom; Prob = probability.

In the lower portion of Table 1, note that the parameter estimates, the standard errors, and the t -statistics match the values given in equation (12).⁸⁹ The variable “Intercept” refers to the constant term b_0 in the regression. The column “DF” represents degrees of freedom. The “1” signifies that when the computer calculates the parameter estimates, each variable that is added to the linear regression adds an additional constraint that must be satisfied. The column labeled “Prob > $|t|$ ” lists the two-tailed p -values associated with each estimated parameter; the p -value measures the observed significance level—the probability of getting a test statistic as extreme or more extreme than the observed number if the model parameter is in fact 0. The very low p -values on the variables X_1 and X_3 imply that each variable is statistically significant at less than the 1% level—both highly significant results. In contrast, the X_2 coefficient is only significant at the 24% level, implying that it is insignificant at the traditional 5% level. Thus, the expert cannot reject with confidence the null hypothesis that salaries do not differ by sex after the expert has accounted for the effect of experience.

The top portion of Table 1 provides data that relate to the goodness of fit of the regression equation. The sum of squared errors (SSE) measures the sum of the squares of the regression residuals—the sum that is minimized by the least squares procedure. The degrees of freedom associated with the error term (DFE) are given by the number of observations minus the number of parameters that were estimated. The mean squared error (MSE) measures the variance of the error term (the square of the standard error of the regression). MSE is equal to SSE divided by DFE.

89. Computer programs give results to more decimal places than are meaningful. This added detail should not be seen as evidence that the regression results are exact.

The R^2 of 0.556 indicates that 55.6% of the variation in salaries is explained by the regression variables, X_1 , X_2 , and X_3 . Finally, the F -test is a test of the null hypothesis that all regression coefficients (except the intercept) are jointly equal to 0—that there is no linear association between the dependent variable and any of the explanatory variables. This is equivalent to the null hypothesis that R^2 is equal to 0. In this case, the F -ratio of 174.71 is sufficiently high that the expert can reject the null hypothesis with a very high degree of confidence (i.e., with a 1% level of significance).

F. Forecasting

In general, a forecast is a prediction made about the values of the dependent variable using information about the explanatory variables. Often, ex ante forecasts are performed; in this situation, values of the dependent variable are predicted beyond the sample (e.g., beyond the time period in which the model has been estimated). However, ex post forecasts are frequently used in damage analyses.⁹⁰ An ex post forecast has a forecast period such that all values of the dependent and explanatory variables are known; ex post forecasts can be checked against existing data and provide a direct means of evaluation.

For example, to calculate the forecast for the salary regression discussed above, the expert uses the estimated salary equation

$$\hat{Y} = \$14,085 + \$2323X_1 + \$1675X_2 - \$36X_3. \quad (14)$$

To predict the salary of a man with 2 years' experience, the expert calculates

$$\hat{Y}(2) = \$14,085 + (\$2323 \cdot 2) + \$1675 - (\$36 \cdot 2) = \$20,262. \quad (15)$$

The degree of accuracy of both ex ante and ex post forecasts can be calculated provided that the model specification is correct and the errors are normally distributed and independent. The statistic is known as the standard error of forecast (SEF). The SEF measures the standard deviation of the forecast error that is made within a sample in which the explanatory variables are known with certainty.⁹¹ The

90. Frequently, in cases involving damages, the question arises, what the world would have been like had a certain event not taken place. For example, in a price-fixing antitrust case, the expert can ask what the price of a product would have been had a certain event associated with the price-fixing agreement not occurred. If prices would have been lower, the evidence suggests impact. If the expert can predict how much lower they would have been, the data can help the expert develop a numerical estimate of the amount of damages.

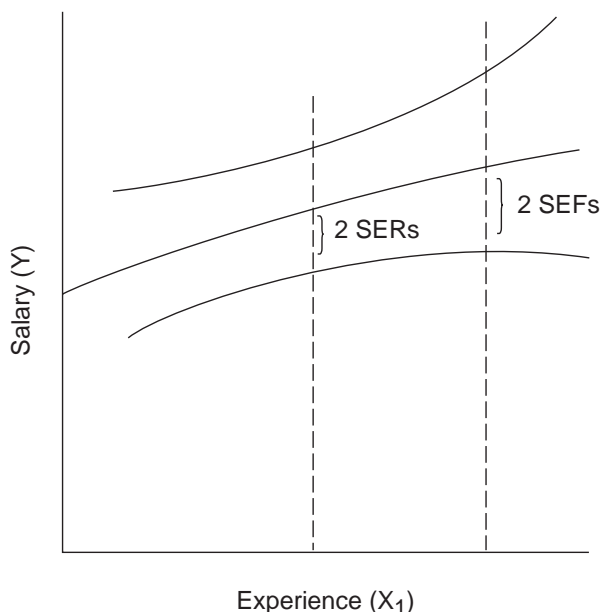
91. There are actually two sources of error implicit in the SEF. The first source arises because the estimated parameters of the regression model may not be exactly equal to the true regression parameters. The second source is the error term itself; when forecasting, the expert typically sets the error equal to 0 when a turn of events not taken into account in the regression model may make it appropriate to make the error positive or negative.

SEF can be used to determine how accurate a given forecast is. In equation (15), the SEF associated with the forecast of \$20,262 is approximately \$5000. If a large sample size is used, the probability is roughly 95% that the predicted salary will be within 1.96 standard errors of the forecasted value. In this case, the appropriate 95% interval for the prediction is \$10,822 to \$30,422. Because the estimated model does not explain salaries effectively, the SEF is large, as is the 95% interval. A more complete model with additional explanatory variables would result in a lower SEF and a smaller 95% interval for the prediction.

A danger exists when using the SEF, which applies to the standard errors of the estimated coefficients as well. The SEF is calculated on the assumption that the model includes the correct set of explanatory variables and the correct functional form. If the choice of variables or the functional form is wrong, the estimated forecast error may be misleading. In some instances, it may be smaller, perhaps substantially smaller, than the true SEF; in other instances, it may be larger, for example, if the wrong variables happen to capture the effects of the correct variables.

The difference between the SEF and the SER is shown in Figure 9. The SER measures deviations within the sample. The SEF is more general, because it calculates deviations within or without the sample period. In general, the difference between the SEF and the SER increases as the values of the explanatory variables increase in distance from the mean values. Figure 9 shows the 95% prediction interval created by the measurement of two SEFs about the regression line.

Figure 9. Standard error of forecast.



G. A Hypothetical Example

Jane Thompson filed suit in federal court alleging that officials in the police department discriminated against her and a class of other female police officers in violation of Title VII of the Civil Rights Act of 1964, as amended. On behalf of the class, Ms. Thompson alleged that she was paid less than male police officers with equivalent skills and experience. Both plaintiff and defendant used expert economists with econometric expertise to present statistical evidence to the court in support of their positions.

Plaintiff's expert pointed out that the mean salary of the 40 female officers was \$30,604, whereas the mean salary of the 60 male officers was \$43,077. To show that this difference was statistically significant, the expert put forward a regression of salary (SALARY) on a constant term and a dummy indicator variable (FEM) equal to 1 for each female and 0 for each male. The results were as follows:

	SALARY = \$43,077 - \$12,373*FEM	
Standard Error	(\$1528)	(\$2416)
p-value	<.01	<.01
$R^2 = .22$		

The $-\$12,373$ coefficient on the FEM variable measures the mean difference between male and female salaries. Because the standard error is approximately one-fifth of the value of the coefficient, this difference is statistically significant at the 5% (and indeed at the 1%) level. If this is an appropriate regression model (in terms of its implicit characterization of salary determination), one can conclude that it is highly unlikely that the difference in salaries between men and women is due to chance.

The defendant's expert testified that the regression model put forward was the wrong model because it failed to account for the fact that males (on average) had substantially more experience than females. The relatively low R^2 was an indication that there was substantial unexplained variation in the salaries of male and female officers. An examination of data relating to years spent on the job showed that the average male experience was 8.2 years, whereas the average for females was only 3.5 years. The defense expert then presented a regression analysis that added an additional explanatory variable (i.e., a covariate), the years of experience of each police officer (EXP). The new regression results were as follows:

	SALARY = \$28,049 - \$3860*FEM + \$1833*EXP		
Standard Error	(2513)	(\$2347)	(\$265)
p-value	<.01	<.11	<.01
$R^2 = .47$			

Experience is itself a statistically significant explanatory variable, with a p-value of less than .01. Moreover, the difference between male and female

Reference Guide on Multiple Regression

salaries, holding experience constant, is only \$3860, and this difference is not statistically significant at the 5% level. The defense expert was able to testify on this basis that the court could not rule out alternative explanations for the difference in salaries other than the plaintiff's claim of discrimination.

The debate did not end here. On rebuttal, the plaintiff's expert made three distinct points. First, whether \$3860 was statistically significant or not, it was practically significant, representing a salary difference of more than 10% of the mean female officers' salaries. Second, although the result was not statistically significant at the 5% level, it was significant at the 11% level. If the regression model were valid, there would be approximately an 11% probability that one would err by concluding that the mean salary difference between men and women was a result of chance.

Third, and most importantly, the expert testified that the regression model was not correctly specified. Further analysis by the expert showed that the value of an additional year of experience was \$2333 for males on average, but only \$1521 for females. Based on supporting testimonial experience, the expert testified that one could not rule out the possibility that the mechanism by which the police department discriminated against females was by rewarding males more for their experience than females. The expert made this point clear by running an additional regression in which a further covariate was added to the model. The new variable was an interaction variable, INT, measured as the product of the FEM and EXP variables. The regression results were as follows:

$$\begin{array}{l} \text{SALARY} = \$35,122 - \$5250 \cdot \text{FEM} + \$2333 \cdot \text{EXP} - \$812 \cdot \text{FEM} \cdot \text{EXP} \\ \text{St. Error} \quad (\$2825) \quad (\$347) \quad (\$265) \quad (\$185) \\ p\text{-value} \quad <.01 \quad <.11 \quad <.01 \quad <.04 \\ R^2 = .65 \end{array}$$

The plaintiff's expert noted that for all males in the sample, FEM = 0, in which case the regression results are given by the equation

$$\text{SALARY} = \$35,122 + \$2333 \cdot \text{EXP}$$

However, for females, FEM = 1, in which the corresponding equation is

$$\text{SALARY} = \$29,872 + \$1521 \cdot \text{EXP}$$

It appears, therefore, that females are discriminated against not only when hired (i.e., when EXP = 0), but also in the reward they get as they accumulate more and more experience.

The debate between the experts continued, focusing less on the statistical interpretation of any one particular regression model, but more on the model choice itself, and not simply on statistical significance, but also with regard to practical significance.

Glossary

The following terms and definitions are adapted from a variety of sources, including *A Dictionary of Epidemiology* (John M. Last et al., eds., 4th ed. 2000) and Robert S. Pindyck & Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts* (4th ed. 1998).

alternative hypothesis. See hypothesis test.

association. The degree of statistical dependence between two or more events or variables. Events are said to be associated when they occur more frequently together than one would expect by chance.

bias. Any effect at any stage of investigation or inference tending to produce results that depart systematically from the true values (i.e., the results are either too high or too low). A biased estimator of a parameter differs on average from the true parameter.

coefficient. An estimated regression parameter.

confidence interval. An interval that contains a true regression parameter with a given degree of confidence.

consistent estimator. An estimator that tends to become more and more accurate as the sample size grows.

correlation. A statistical means of measuring the linear association between variables. Two variables are correlated positively if, on average, they move in the same direction; two variables are correlated negatively if, on average, they move in opposite directions.

covariate. A variable that is possibly predictive of an outcome under study; an explanatory variable.

cross-sectional analysis. A type of multiple regression analysis in which each data point is associated with a different unit of observation (e.g., an individual or a firm) measured at a particular point in time.

degrees of freedom (DF). The number of observations in a sample minus the number of estimated parameters in a regression model. A useful statistic in hypothesis testing.

dependent variable. The variable to be explained or predicted in a multiple regression model.

dummy variable. A variable that takes on only two values, usually 0 and 1, with one value indicating the presence of a characteristic, attribute, or effect (1), and the other value indicating its absence (0).

efficient estimator. An estimator of a parameter that produces the greatest precision possible.

error term. A variable in a multiple regression model that represents the cumulative effect of a number of sources of modeling error.

- estimate.** The calculated value of a parameter based on the use of a particular sample.
- estimator.** The sample statistic that estimates the value of a population parameter (e.g., a regression parameter); its values vary from sample to sample.
- ex ante forecast.** A prediction about the values of the dependent variable that go beyond the sample; consequently, the forecast must be based on predictions for the values of the explanatory variables in the regression model.
- explanatory variable.** A variable that is associated with changes in a dependent variable.
- ex post forecast.** A prediction about the values of the dependent variable made during a period in which all values of the explanatory and dependent variables are known. Ex post forecasts provide a useful means of evaluating the fit of a regression model.
- F-test.** A statistical test (based on an F -ratio) of the null hypothesis that a group of explanatory variables are jointly equal to 0. When applied to all the explanatory variables in a multiple regression model, the F -test becomes a test of the null hypothesis that R^2 equals 0.
- feedback.** When changes in an explanatory variable affect the values of the dependent variable, and changes in the dependent variable also affect the explanatory variable. When both effects occur at the same time, the two variables are described as being determined simultaneously.
- fitted value.** The estimated value for the dependent variable; in a linear regression, this value is calculated as the intercept plus a weighted average of the values of the explanatory variables, with the estimated parameters used as weights.
- heteroscedasticity.** When the error associated with a multiple regression model has a nonconstant variance; that is, the error values associated with some observations are typically high, while the values associated with other observations are typically low.
- hypothesis test.** A statement about the parameters in a multiple regression model. The null hypothesis may assert that certain parameters have specified values or ranges; the alternative hypothesis would specify other values or ranges.
- independence.** When two variables are not correlated with each other (in the population).
- independent variable.** An explanatory variable that affects the dependent variable but that is not affected by the dependent variable.
- influential data point.** A data point whose deletion from a regression sample causes one or more estimated regression parameters to change substantially.
- interaction variable.** The product of two explanatory variables in a regression model. Used in a particular form of nonlinear model.

intercept. The value of the dependent variable when each of the explanatory variables takes on the value of 0 in a regression equation.

least squares. A common method for estimating regression parameters. Least squares minimizes the sum of the squared differences between the actual values of the dependent variable and the values predicted by the regression equation.

linear regression model. A regression model in which the effect of a change in each of the explanatory variables on the dependent variable is the same, no matter what the values of those explanatory variables.

mean (sample). An average of the outcomes associated with a probability distribution, where the outcomes are weighted by the probability that each will occur.

mean squared error (MSE). The estimated variance of the regression error, calculated as the average of the sum of the squares of the regression residuals.

model. A representation of an actual situation.

multicollinearity. When two or more variables are highly correlated in a multiple regression analysis. Substantial multicollinearity can cause regression parameters to be estimated imprecisely, as reflected in relatively high standard errors.

multiple regression analysis. A statistical tool for understanding the relationship between two or more variables.

nonlinear regression model. A model having the property that changes in explanatory variables will have differential effects on the dependent variable as the values of the explanatory variables change.

normal distribution. A bell-shaped probability distribution having the property that about 95% of the distribution lies within 2 standard deviations of the mean.

null hypothesis. In regression analysis the null hypothesis states that the results observed in a study with respect to a particular variable are no different from what might have occurred by chance, independent of the effect of that variable. See *hypothesis test*.

one-tailed test. A hypothesis test in which the alternative to the null hypothesis that a parameter is equal to 0 is for the parameter to be either positive or negative, but not both.

outlier. A data point that is more than some appropriate distance from a regression line that is estimated using all the other data points in the sample.

p-value. The significance level in a statistical test; the probability of getting a test statistic as extreme or more extreme than the observed value. The larger the p-value, the more likely that the null hypothesis is valid.

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

perfect collinearity. When two or more explanatory variables are correlated perfectly.

population. All the units of interest to the researcher; also, universe.

practical significance. Substantive importance. Statistical significance does not ensure practical significance, because, with large samples, small differences can be statistically significant.

probability distribution. The process that generates the values of a random variable. A probability distribution lists all possible outcomes and the probability that each will occur.

probability sampling. A process by which a sample of a population is chosen so that each unit of observation has a known probability of being selected.

quasi-experiment (or natural experiment). A naturally occurring instance of observable phenomena that yield data that approximate a controlled experiment.

R-squared (R^2). A statistic that measures the percentage of the variation in the dependent variable that is accounted for by all of the explanatory variables in a regression model. R -squared is the most commonly used measure of goodness of fit of a regression model.

random error term. A term in a regression model that reflects random error (sampling error) that is the result of chance. As a consequence, the result obtained in the sample differs from the result that would be obtained if the entire population were studied.

regression coefficient. Also, regression parameter. The estimate of a population parameter obtained from a regression equation that is based on a particular sample.

regression residual. The difference between the actual value of a dependent variable and the value predicted by the regression equation.

robust estimation. An alternative to least squares estimation that is less sensitive to outliers.

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

sample. A selection of data chosen for a study; a subset of a population.

sampling error. A measure of the difference between the sample estimate of a parameter and the population parameter.

scatterplot. 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.

serial correlation. The correlation of the values of regression errors over time.

- slope.** The change in the dependent variable associated with a one-unit change in an explanatory variable.
- spurious correlation.** When two variables are correlated, but one is not the cause of the other.
- standard deviation.** The square root of the variance of a random variable. The variance is a measure of the spread of a probability distribution about its mean; it is calculated as a weighted average of the squares of the deviations of the outcomes of a random variable from its mean.
- standard error of forecast (SEF).** An estimate of the standard deviation of the forecast error; it is based on forecasts made within a sample in which the values of the explanatory variables are known with certainty.
- standard error of the coefficient; standard error (SE).** A measure of the variation of a parameter estimate or coefficient about the true parameter. The standard error is a standard deviation that is calculated from the probability distribution of estimated parameters.
- standard error of the regression (SER).** An estimate of the standard deviation of the regression error; it is calculated as the square root of the average of the squares of the residuals associated with a particular multiple regression analysis.
- statistical significance.** A test used to evaluate the degree of association between a dependent variable and one or more explanatory variables. If the calculated p -value is smaller than 5%, the result is said to be statistically significant (at the 5% level). If p is greater than 5%, the result is statistically insignificant (at the 5% level).
- t -statistic.** A test statistic that describes how far an estimate of a parameter is from its hypothesized value (i.e., given a null hypothesis). If a t -statistic is sufficiently large (in absolute magnitude), an expert can reject the null hypothesis.
- t -test.** A test of the null hypothesis that a regression parameter takes on a particular value, usually 0. The test is based on the t -statistic.
- time-series analysis.** A type of multiple regression analysis in which each data point is associated with a particular unit of observation (e.g., an individual or a firm) measured at different points in time.
- two-tailed test.** A hypothesis test in which the alternative to the null hypothesis that a parameter is equal to 0 is for the parameter to be either positive or negative, or both.
- variable.** Any attribute, phenomenon, condition, or event that can have two or more values.
- variable of interest.** The explanatory variable that is the focal point of a particular study or legal issue.

References on Multiple Regression

- Jonathan A. Baker & Daniel L. Rubinfeld, *Empirical Methods in Antitrust: Review and Critique*, 1 Am. L. & Econ. Rev. 386 (1999).
- Gerald V. Barrett & Donna M. Sansonetti, *Issues Concerning the Use of Regression Analysis in Salary Discrimination Cases*, 41 Personnel Psychol. 503 (2006).
- Thomas J. Campbell, *Regression Analysis in Title VII Cases: Minimum Standards, Comparable Worth, and Other Issues Where Law and Statistics Meet*, 36 Stan. L. Rev. 1299 (1984).
- Catherine Connolly, *The Use of Multiple Regression Analysis in Employment Discrimination Cases*, 10 Population Res. & Pol'y Rev. 117 (1991).
- Arthur P. Dempster, *Employment Discrimination and Statistical Science*, 3 Stat. Sci. 149 (1988).
- Michael O. Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases*, 80 Colum. L. Rev. 737 (1980).
- Michael O. Finkelstein & Hans Levenbach, *Regression Estimates of Damages in Price-Fixing Cases*, Law & Contemp. Probs., Autumn 1983, at 145.
- Franklin M. Fisher, *Multiple Regression in Legal Proceedings*, 80 Colum. L. Rev. 702 (1980).
- Franklin M. Fisher, *Statisticians, Econometricians, and Adversary Proceedings*, 81 J. Am. Stat. Ass'n 277 (1986).
- Joseph L. Gastwirth, *Methods for Assessing the Sensitivity of Statistical Comparisons Used in Title VII Cases to Omitted Variables*, 33 Jurimetrics J. 19 (1992).
- Note, *Beyond the Prima Facie Case in Employment Discrimination Law: Statistical Proof and Rebuttal*, 89 Harv. L. Rev. 387 (1975).
- Daniel L. Rubinfeld, *Econometrics in the Courtroom*, 85 Colum. L. Rev. 1048 (1985).
- Daniel L. Rubinfeld & Peter O. Steiner, *Quantitative Methods in Antitrust Litigation*, Law & Contemp. Probs., Autumn 1983, at 69.
- Daniel L. Rubinfeld, *Statistical and Demographic Issues Underlying Voting Rights Cases*, 15 Evaluation Rev. 659 (1991).
- The Evolving Role of Statistical Assessments as Evidence in the Courts (Stephen E. Fienberg ed., 1989).

The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics

Joshua D. Angrist and Jörn-Steffen Pischke

Just over a quarter century ago, Edward Leamer (1983) reflected on the state of empirical work in economics. He urged empirical researchers to “take the con out of econometrics” and memorably observed (p. 37): “Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone else’s data analysis seriously.” Leamer was not alone; Hendry (1980), Sims (1980), and others writing at about the same time were similarly disparaging of empirical practice. Reading these commentaries as late-1980s Ph.D. students, we wondered about the prospects for a satisfying career doing applied work. Perhaps credible empirical work in economics is a pipe dream. Here we address the questions of whether the quality and the credibility of empirical work have increased since Leamer’s pessimistic assessment. Our views are necessarily colored by the areas of applied microeconomics in which we are active, but we look over the fence at other areas as well.

Leamer (1983) diagnosed his contemporaries’ empirical work as suffering from a distressing lack of robustness to changes in key assumptions—assumptions he called “whimsical” because one seemed as good as another. The remedy he proposed was sensitivity analysis, in which researchers show how their results vary with changes in specification or functional form. Leamer’s critique had a refreshing emperor’s-new-clothes earthiness that we savored on first reading and still enjoy today. But we’re happy to report that Leamer’s complaint that “hardly anyone takes anyone else’s data analysis seriously” no longer seems justified.

■ *Joshua D. Angrist is Ford Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Jörn-Steffen Pischke is Professor of Economics, London School of Economics, London, United Kingdom. Their e-mail addresses are (angrist@mit.edu) and (s.pischke@lse.ac.uk).*

doi=10.1257/jep.24.2.3

Empirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact. Sensitivity analysis played a role in this, but as we see it, the primary engine driving improvement has been a focus on the quality of empirical research designs. This emphasis on research design is in the spirit of Leamer's critique, but it did not feature in his remedy.

The advantages of a good research design are perhaps most easily apparent in research using random assignment, which not coincidentally includes some of the most influential microeconomic studies to appear in recent years. For example, in a pioneering effort to improve child welfare, the *Progresa* program in Mexico offered cash transfers to randomly selected mothers, contingent on participation in prenatal care, nutritional monitoring of children, and the children's regular school attendance (Gertler, 2004, and Schultz, 2004, present some of the main findings). In the words of Paul Gertler, one of the original investigators (quoted in Ayres, 2007, p. 86), "*Progresa* is why now thirty countries worldwide have conditional cash transfer programs." *Progresa* is emblematic of a wave of random assignment policy evaluations sweeping development economics (Duflo and Kremer, 2008, provide an overview).

Closer to home, the *Moving to Opportunity* program, carried out by the U.S. Department of Housing and Urban Development, randomly selected low-income families in Baltimore, Boston, Chicago, Los Angeles, and New York City to be offered housing vouchers specifically limited to low-poverty areas (Kling, Liebman, and Katz, 2007). The program has produced surprising and influential evidence weighing against the view that neighborhood effects are a primary determinant of low earnings by the residents of poor neighborhoods.

Structural econometric parameters, such as the intertemporal substitution elasticity (a labor supply elasticity that measures the response to transitory wage changes), have also been the focus of randomized experiments. For example, Fehr and Goette (2007) randomized the pay of bicycle messengers, offering one group and then another a temporarily higher wage. This cleverly designed study shows how wages affect labor supply in an environment where lifetime wealth is unchanged. The result is dramatic and convincing: holding wealth constant, workers shift hours into high-wage periods, with an implied intertemporal substitution elasticity of about unity.

Such studies offer a powerful method for deriving results that are defensible both in the seminar room and in a legislative hearing. But experiments are time consuming, expensive, and may not always be practical. It's difficult to imagine a randomized trial to evaluate the effect of immigrants on the economy of the host country. However, human institutions or the forces of nature can step into the breach with informative natural or quasi-experiments. For example, in an influential paper, Card (1990a) used the Mariel boatlift from Cuba to Florida, when Cuban émigré's increased Miami's labor force by about 7 percent in a period of three months, as a natural experiment to study immigration. More recently, paralleling the *Moving to Opportunity* experimental research agenda, Jacob (2004) studied the causal effects of public housing on housing project residents by exploiting the

fact that public housing demolition in Chicago was scheduled in a manner unrelated to the characteristics of the projects and their residents.

Like the results from randomized trials, quasi-experimental findings have filtered quickly into policy discussions and become part of a constructive give-and-take between the real world and the ivory tower, at least when it comes to applied microeconomics. Progress has been slower in empirical macro, but a smattering of design-based empirical work appears to be generating a limited though useful consensus on key concerns, such as the causal effect of monetary policy on inflation and output. Encouragingly, the recent financial crisis has spurred an effort to produce credible evidence on questions related to banking. Across most fields (although industrial organization appears to be an exception, as we discuss later), applied economists are now less likely to pin a causal interpretation of the results on econometric methodology alone. Design-based studies are distinguished by their *prima facie* credibility and by the attention investigators devote to making both an institutional and a data-driven case for causality.

Accounting for the origins of the credibility revolution in empirical economics is like trying to chart the birth of rock and roll. Early influences are many, and every fan has a story. But from the trenches of empirical labor economics, we see an important impetus for better designs and more randomized trials coming from studies questioning the reliability of econometric evaluations of subsidized government training programs. A landmark here is Lalonde (1986), who compared the results from an econometric evaluation of the National Supported Work demonstration with those from a randomized trial. The econometric results typically differed quite a bit from those using random assignment. Lalonde argued that there is little reason to believe that statistical comparisons of alternative models (specification testing) would point a researcher in the right direction. Two observational studies of training effects foreshadowed the Lalonde results: Ashenfelter (1978) and Ashenfelter and Card (1985), using longitudinal data to evaluate federal training programs without the benefit of a quasi-experimental research design, found it difficult to construct specification-robust estimates. Ashenfelter (1987) concluded that randomized trials are the way to go.

Younger empiricists also began to turn increasingly to quasi-experimental designs, often exploiting variation across U.S. states to get at causal relationships in the fields of labor and public finance. An early example of work in this spirit is Solon (1985), who estimated the effects of unemployment insurance on the duration of unemployment spells by comparing the change in job-finding rates in states that had recently tightened eligibility criteria for unemployment insurance, to the change in rates in states that had not changed their rules. Gruber's (1994) influential study of the incidence of state-mandated maternity benefits applies a similar idea to a public finance question. Angrist (1990) and Angrist and Krueger (1991) illustrated the value of instrumental variables identification strategies in studies of the effects of Vietnam-era military service and schooling on earnings. Meyer's (1995) methodological survey made many applied microeconomists aware of the quasi-experimental tradition embodied in venerable texts on social science

research methods by Campbell and Stanley (1963) and Cook and Campbell (1979). These texts, which emphasize research design and threats to validity, were well known in some disciplines, but distinctly outside the econometric canon.¹

In this essay, we argue that a clear-eyed focus on research design is at the heart of the credibility revolution in empirical economics. We begin with an overview of Leamer's (1983) critique and his suggested remedies, based on concrete examples of that time. We then turn to the key factors we see contributing to improved empirical work, including the availability of more and better data, along with advances in theoretical econometric understanding, but especially the fact that research design has moved front and center in much of empirical micro. We offer a brief digression into macroeconomics and industrial organization, where progress—by our lights—is less dramatic, although there is work in both fields that we find encouraging. Finally, we discuss the view that the design pendulum has swung too far. Critics of design-driven studies argue that in pursuit of clean and credible research designs, researchers seek good answers instead of good questions. We briefly respond to this concern, which worries us little.

The Leamer Critique and His Proposed Remedies

Naive Regressions and Extreme Bounds Analysis

Leamer (1983) presented randomized trials—a randomized evaluation of fertilizer, to be specific—as an ideal research design. He also argued that randomized experiments differ only in degree from nonexperimental evaluations of causal effects, the difference being the extent to which we can be confident that the causal variable of interest is independent of confounding factors. We couldn't agree more. However, Leamer went on to suggest that the best way to use nonexperimental data to get closer to the experimental ideal is to explore the fragility of nonexperimental estimates. Leamer did not advocate *doing* randomized trials or, for that matter, looking for credible natural experiments.

The chief target of Leamer's (1983) essay was naive regression analysis. In fact, none of the central figures in the Leamer-inspired debate had much to say about research design. Rather, these authors (like McAleer, Pagan, and Volker, 1985, and Cooley and LeRoy, 1986, among others) appear to have accepted the boundaries of established econometric practice, perhaps because they were primarily interested in addressing traditional macroeconomic questions using time series data.

After making the tacit assumption that useful experiments are an unattainable ideal, Leamer (1983, but see also 1978, 1985) proposed that the whimsical nature of key assumptions in regression analysis be confronted head-on through a process of

¹ Many of the applied studies mentioned here have been the subjects of critical re-examinations. This back and forth has mostly been constructive. For example, in an influential paper that generated wide-ranging methodological work, Bound, Jaeger, and Baker (1995) argue that the use of many weak instrumental variables biases some of the estimates reported in Angrist and Krueger (1991). For a recent discussion of weak instruments problems, see our book Angrist and Pischke (2009).

sensitivity analysis. Sims (1988) threw his weight behind this idea as well. The general heading of sensitivity analysis features an explicitly Bayesian agenda. Recognizing the severe demands of Bayesian orthodoxy, such as a formal specification of priors and their incorporation into an elaborate multivariate framework, Leamer also argued for a more *ad hoc* but intuitive approach called “extreme bounds analysis.” In a nutshell, extreme bounds analysis amounts to the estimation of regressions with many different sets of covariates included as controls; practitioners of this approach are meant to report a range of estimates for the target parameter.

The Deterrent Effect of Capital Punishment

We sympathize with Leamer’s (1983) view that much of the applied econometrics of the 1970s and early 1980s lacked credibility. To make his point, and to illustrate the value of extreme bounds analysis, Leamer picked an inquiry into whether capital punishment deters murder. This question had been analyzed in a series of influential papers by Isaac Ehrlich, one exploiting time series variation (Ehrlich, 1975a) and one using cross sections of states (Ehrlich, 1977b). Ehrlich concluded that the death penalty had a substantial deterrent effect. Leamer (1983) did not try to replicate Ehrlich’s work, but reported on an independent time-series investigation of the deterrence hypothesis using extreme bounds analysis, forcefully arguing that the evidence for deterrence is fragile at best (although Ehrlich and Liu, 1999, disputed this).

It’s hard to exaggerate the attention this topic commanded at the time. The U.S. Supreme Court decision in *Furman v. Georgia* (408 U.S. 153 [1972]) had created a de facto moratorium on the death penalty. This moratorium lasted until *Gregg v. Georgia* (428 U.S. 153 [1976]), at which time the high court decided that the death penalty might be allowable if capital trials were bifurcated into separate guilt–innocence and sentencing phases. Gary Gilmore was executed not long after, in January 1977. Part of the intellectual case for restoration of capital punishment was the deterrent effect (against a backdrop of high and increasing homicide rates at that time). Indeed, the U.S. Supreme Court cited Ehrlich’s (1975a) paper in its *Gregg v. Georgia* decision reinstating capital punishment.

Ehrlich’s work was harshly criticized by a number of contemporaries in addition to Leamer, most immediately Bowers and Pierce (1975) and Passell and Taylor (1977). Ehrlich’s results appeared to be sensitive to changes in functional form, inclusion of additional controls, and especially to changes in sample. Specifically, his finding of a significant deterrent effect seemed to depend on observations from the 1960s. The critics argued that the increase in murder rates in the 1960s may have been driven by factors other than the sharp decline in the number of executions during this period. Ehrlich (1975b, 1977a) disputed the critics’ claims about functional form and argued that the 1960s provided useful variation in executions that should be retained.

Ehrlich’s contemporaneous critics failed to hit on what we think of as the most obvious flaw in Ehrlich’s analysis. Like other researchers studying deterrent effects, Ehrlich recognized that the level of the murder rate might affect the number of

executions as well as vice versa and that his results might be biased by omitted variables (especially variables with a strong trend). Ehrlich sought to address problems of reverse causality and omitted variables bias by using instrumental variables in a two-stage least squares procedure. He treated the probabilities of arrest, conviction, and execution as endogenous in a simultaneous-equations set-up. His instrumental variables were lagged expenditures on policing, total government expenditure, population, and the fraction of the population nonwhite. But Ehrlich did not explain why these are good instruments, or even how and why these variables are correlated with the right-hand-side endogenous variables.²

Ehrlich's work on capital punishment seems typical of applied work in the period about which Leamer (1983) was writing. Most studies of this time used fairly short time series samples with strong trends common to both dependent and independent variables. The use of panel data to control for year and fixed effects—even panels of U.S. states—was still rare. The use of instrumental variables to uncover causal relationships was typically mechanical, with little discussion of why the instruments affected the endogenous variables of interest or why they constitute a “good experiment.” In fact, Ehrlich was ahead of many of his contemporaries in that he recognized the need for something other than naive regression analysis. In our view, the main problem with Ehrlich's work was the lack of a credible research design. Specifically, he failed to isolate a source of variation in execution rates that is likely to reveal causal effects on homicide rates.

The Education Production Function

Other examples of poor research design from this time period come from the literature on education production. This literature (surveyed in Hanushek, 1986) is concerned with the causal effect of school inputs, such as class size or per-pupil expenditure, on student achievement. The systematic quantitative study of school inputs was born with the report by Coleman et al. (1966), which (among other things) used regression techniques to look at the proportion of variation in student outputs that can be accounted for in an R^2 sense by variation in school inputs. Surprisingly to many at the time, the Coleman report found only a weak association between school inputs and achievement. Many subsequent regression-based studies replicated this finding.

The Coleman Report was one of the first investigations of education production in a large representative sample. It is also distinguished by sensitivity analysis, in that it discusses results from many specifications (with and without controls for family background, for example). The problem with the Coleman report and many of the studies in this mold that followed is that they failed to separate variation in inputs from confounding variation in student, school, or community characteristics. For example, a common finding in the literature on education

² Ehrlich's (1977b) follow-up cross-state analysis did not use two-stage least squares. In later work, Ehrlich (1987, 1996) discussed his choice of instruments and the associated identification problems at greater length.

production is that children in smaller classes tend to do worse on standardized tests, even after controlling for demographic variables. This apparently perverse finding seems likely to be at least partly due to the fact that struggling children are often grouped into smaller classes. Likewise, the relationship between school spending and achievement is confounded by the fact that spending is often highest in a mixture of wealthy districts and large urban districts with struggling minority students. In short, these regressions suffer from problems of reverse causality and omitted variables bias.

Many education production studies from this period also ignored the fact that inputs like class size and per-pupil expenditure are inherently linked. Because smaller classes cannot be had without spending more on teachers, it makes little sense to treat total expenditure (including teacher salaries) as a control variable when estimating the causal effect of class size (a point noted by Krueger, 2003). Finally, the fact that early authors in the education production literature explored many alternative models was not necessarily a plus. In what was arguably one of the better studies of the period, Summers and Wolfe (1977) report only the final results of an exhaustive specification search in their evaluation of the effect of school resources on achievement. To their credit, Summers and Wolfe describe the algorithm that produced the results they chose to report, and forthrightly caution (p. 642) that “the data have been mined, of course.” As we see it, however, the main problem with this literature is not data mining, but rather the weak foundation for a causal interpretation of whatever specification authors might have favored.

Other Empirical Work in the Age of Heavy Metal

The 1970s and early 1980s saw rapid growth in mainframe computer size and power. Stata had yet to appear, but magnetic tape jockeys managed to crunch more and more numbers in increasingly elaborate ways. For the most part, however, increased computing power did not produce more credible estimates. For example, the use of randomized trials and quasi-experiments to study education production was rare until fairly recently (a history traced in Angrist, 2004). Other areas of social science saw isolated though ambitious efforts to get at key economic relationships using random assignment. A bright spot was the RAND Health Insurance Experiment, initiated in 1974 (Manning, Newhouse, Duan, Keeler, and Leibowitz, 1987). This experiment looked at the effects of deductibles and copayments on health care usage and outcomes. Unfortunately, many of the most ambitious (and expensive) social experiments were seriously flawed: the Seattle/Denver and Gary Income Maintenance Experiments, in which the government compared income-support plans modeled on Milton Friedman’s idea of a negative income tax, were compromised by sample attrition and systematic income misreporting (Ashenfelter and Plant, 1990; Greenberg and Halsey, 1983). This fact supports Leamer’s (1983) contention that the difference between a randomized trial and an observational study is one of degree. Indeed, we would be the first to admit that a well-done observational study can be more credible and persuasive than a poorly executed randomized trial.

There was also much to complain about in empirical macroeconomics. An especially articulate complaint came from Sims (1980), who pointed out that macroeconomic models of that time, typically a system of simultaneous equations, invoked identification assumptions (the division of variables into those that are jointly determined and exogenous) that were hard to swallow and poorly defended. As an alternative to the simultaneous equations framework, Sims suggested the use of unrestricted vector autoregressions (VARs) to describe the relation between a given set of endogenous variables and their lags. But Sims's complaint did not generate the same kind of response that grew out of concerns about econometric program evaluation in the 1980s among labor economists. Macroeconomists circled their wagons but did not mobilize an identification posse.

Sims's argument came on the heels of a closely related and similarly influential stab at the heart of empirical macro known as the Lucas critique. Lucas (1976) and Kydland and Prescott (1977) argued via theoretical examples that in a world with forward-looking optimizing agents, *nothing* can be learned from past policy changes. Lucas held out the hope that we might instead try to recover the empirical response to changes in policy rules by estimating the structural parameters that lie at the root of economic behavior, such as those related to technology or preferences (Lucas saw these parameters as stable or at least policy invariant). But Kydland and Prescott—invoking Lucas—appeared willing to give up entirely on conventional empirical work (1977, p. 487): “If we are not to attempt to select policy optimally, how should it be selected? Our answer is, as Lucas (1976) proposed, that economic theory be used to evaluate alternative policy rules and that one with good operating characteristics be selected.” This view helped to lay the intellectual foundations for a sharp turn toward theory in macro, though often informed by numbers via “calibration.”

Our overview of empirical work in the Leamer era focuses on shortcomings. But we should also note that the best applied work from the 1970s and early 1980s still holds up today. A well-cited example is Feldstein and Horioka (1980), which argues that the strong link between domestic savings and investment weighs against the notion of substantial international capital mobility. The Feldstein and Horioka study presents simple evidence in favor of a link between domestic savings and investment, discusses important sources of omitted variables bias and simultaneity bias in these estimates, and tries to address these concerns. Obstfeld's (1995) extensive investigation of the Feldstein and Horioka (1980) framework essentially replicates their findings for a later and longer period.

Why There's Less Con in Econometrics Today

Improvements in empirical work have come from many directions. Better data and more robust estimation methods are part of the story, as is a reduced emphasis on econometric considerations that are not central to a causal interpretation of the main findings. But the primary force driving the credibility revolution has been a vigorous push for better and more clearly articulated research designs.

Better and More Data

Not unusually for the period, Ehrlich (1975a) analyzed a time series of 35 annual observations. In contrast, Donohue and Wolfers (2005) investigate the capital punishment question using a panel of U.S. states from 1934 to 2000, with many more years and richer within-state variation due to the panel structure of the data. Better data often engenders a fresh approach to long-standing research questions. Grogger's (1990) investigation of the deterrent effect of executions on daily homicide rates, inspired by sociologist Phillips (1980), is an example.³ Farther afield, improvements have come from a rapidly expanding reservoir of micro data in many countries. The use of administrative records has also grown.

Fewer Distractions

Bower's and Pierce (1975) devoted considerable attention to Ehrlich's (1975a) use of the log transformation, as well as to his choice of sample period. Passell and Taylor (1977) noted the potential for omitted variables bias, but worried as much about *F*-tests for temporal homogeneity and logs. The methodological appendix to Ehrlich's (1977b) follow-up paper discusses the possibility of using a Box-Cox transformation to implement a flexible functional form, tests for heteroskedasticity, and uses generalized least squares. Ehrlich's (1975b) reply to Bowers and Pierce focused on the statistical significance of trend terms in samples of different lengths, differences in computational procedures related to serial correlation, and evidence for robustness to the use of logs. Ehrlich's (1977a) reply to Passell covers the sample period and logs, though he also reports some of his (1977b) cross-state estimates. Ehrlich's rejoinders devoted little attention to the core issue of whether the sources of variation in execution used by his statistical models justify a causal interpretation of his estimates, but Ehrlich's contemporaneous critics did not hit this nail on the head either. Even were the results insensitive to the sample, the same in logs and levels, and the residuals independent and identically distributed, we would remain unsatisfied. In the give and take that followed Ehrlich's original paper, the question of instrument validity rarely surfaced, while the question of omitted variables bias took a back seat to concerns about sample break points and functional form.⁴

As in the exchange over capital punishment, others writing at about the same time often seemed distracted by concerns related to functional form and generalized least squares. Today's applied economists have the benefit of a less dogmatic understanding of regression analysis. Specifically, an emerging grasp of the sense in which regression and two-stage least squares produce average effects even when the underlying relationship is heterogeneous and/or nonlinear has made

³ The decline in the use of time series and the increase in the use of panel data and researcher-originated data are documented for the field of labor economics in Table 1 of Angrist and Krueger (1999).

⁴ Hoenack and Weiler's (1980) critical re-examination of Ehrlich (1975a) centered on identification problems, but the alternative exclusion restrictions Hoenack and Weiler proposed were offered without much justification and seem just as hard to swallow as Ehrlich's (for example, the proportion nonwhite is used as an instrument).

functional form concerns less central. The linear models that constitute the workhorse of contemporary empirical practice usually turn out to be remarkably robust, a feature many applied researchers have long sensed and that econometric theory now does a better job of explaining.⁵ Robust standard errors, automated clustering, and larger samples have also taken the steam out of issues like heteroskedasticity and serial correlation. A legacy of White's (1980a) paper on robust standard errors, one of the most highly cited from the period, is the near death of generalized least squares in cross-sectional applied work. In the interests of replicability, and to reduce the scope for errors, modern applied researchers often prefer simpler estimators though they might be giving up asymptotic efficiency.

Better Research Design

Leamer (1983) led his essay with the idea that experiments—specifically, randomized trials—provide a benchmark for applied econometrics. He was not alone among econometric thought leaders of the period in this view. Here is Zvi Griliches (1986, p. 1466) at the beginning of a chapter on data in *The Handbook of Econometrics*: “If the data were perfect, collected from well-designed randomized experiments, there would hardly be room for a separate field of econometrics.” Since then, empirical researchers in economics have increasingly looked to the ideal of a randomized experiment to justify causal inference. In applied micro fields such as development, education, environmental economics, health, labor, and public finance, researchers seek real experiments where feasible, and useful natural experiments if real experiments seem (at least for a time) infeasible. In either case, a hallmark of contemporary applied microeconomics is a conceptual framework that highlights specific sources of variation. These studies can be said to be *design based* in that they give the research design underlying any sort of study the attention it would command in a real experiment.

The econometric methods that feature most prominently in quasi-experimental studies are instrumental variables, regression discontinuity methods, and differences-in-differences-style policy analysis. These econometric methods are not new, but their use has grown and become more self-conscious and sophisticated since the 1970s. When using instrumental variables, for example, it's no longer enough to mechanically invoke a simultaneous equations framework, labeling some variables endogenous and others exogenous, without substantially justifying the exclusion restrictions and as-good-as-randomly-assigned assumptions that make instruments valid. The best of today's design-based studies make a strong institutional case, backed up with empirical evidence, for the variation thought to generate a useful natural experiment.

⁵For this view of regression, see, for example, White (1980b), Chamberlain's (1984) chapter in the *Handbook of Econometrics*, Goldberger's (1991) econometrics text, or our book Angrist and Pischke (2009) for a recent take. Angrist and Imbens (1995) show how conventional two-stage least squares estimates can be interpreted as an average causal effect in models with nonlinear and heterogeneous causal effects.

The Card and Krueger (1992a, b) school quality studies illustrate this and arguably mark a turning point in the literature on education production. The most important problem in studies of school quality is omitted variables bias. On one hand, students who attend better-resourced schools often end up in those schools by virtue of their ability or family background, while on the other, weaker students may receive disproportionately more inputs (say, smaller classes). Card and Krueger addressed this problem by focusing on variation in resources at the state-of-birth-by-cohort level, which they link to the economic returns to education estimated at the same level. For example, they used Census data to compare the returns to education for residents of Northern states educated in the North with the returns to education for residents of Northern states educated in more poorly resourced Southern schools.

The Card and Krueger papers show that the economic returns to schooling are higher for those from states and cohorts with more resources (controlling for cohort and state fixed effects and for state of residence). They implicitly use state-level variation in education spending as a natural experiment: aggregation of individual data up to the cohort/state level is an instrumental variables procedure where the instruments are state-of-birth and cohort dummy variables. (In Angrist and Pischke, 2009, we show why aggregation in this way works as an instrumental variable.) State-by-cohort variation in the returns to schooling is unlikely to be driven by selection or sorting, because individuals do not control these variables. State-by-cohort variation in school resources also appears unrelated to omitted factors such as family background. Finally, Card and Krueger took advantage of the fact that school resources increased dramatically in the South when the Southerners in their sample were school age. The Card and Krueger school quality studies are not bulletproof (Heckman, Layne-Farrar, and Todd, 1996 offer a critique), but their findings on class size (the strongest set of results in Card and Krueger, 1992a) have been replicated in other studies with good research designs.

Angrist and Lavy (1999) illustrate the regression discontinuity research design in a study of the effects of class size on achievement. The regression discontinuity approach can be used when people are divided into groups based on a certain cutoff score, with those just above or just below the cutoff suddenly becoming eligible for a different treatment. The Angrist–Lavy research design is driven by the fact that class size in Israel is capped at 40, so a cohort of 41 is usually split into two small classes, while a cohort of 39 is typically left in a single large class. This leads to a series of notional experiments: comparisons of schools with enrollments just above and below 40, 80, or 120, in which class sizes vary considerably. In this setting, schools with different numbers of students may be quite similar in other characteristics. Thus, as school enrollment increases, a regression capturing the relationship between number of students and academic achievement should show discontinuities at these break points. The Angrist–Lavy design is a version of what is known as the “fuzzy” regression discontinuity design, in which the fuzziness comes from the fact that class size is not a deterministic function of the kinks or discontinuities in

the enrollment function. Regression discontinuity estimates using Israeli data show a marked increase in achievement when class size falls.⁶

The key assumption that drives regression discontinuity estimation of causal effects is that individuals are otherwise similar on either side of the discontinuity (or that any differences can be controlled using smooth functions of the enrollment rates, also known as the “running variable,” that determine the kink points). In the Angrist–Lavy study, for example, we would like students to have similar family backgrounds when they attend schools with grade enrollments of 35–39 and 41–45. One test of this assumption, illustrated by Angrist and Lavy (and Hoxby, 2000) is to estimate effects in an increasingly narrow range around the kink points; as the interval shrinks, the jump in class size stays the same or perhaps even grows, but the estimates should be subject to less and less omitted variables bias. Another test, proposed by McCrary (2008), looks for bunching in the distribution of student background characteristics around the kink. This bunching might signal strategic behavior—an effort by some families, presumably not a random sample, to sort themselves into schools with smaller classes. Finally, we can simply look for differences in mean pre-treatment characteristics around the kink.

In a recent paper, Urqiola and Verhoogen (2009) exploit enrollment cutoffs like those used by Angrist and Lavy in a sample from Chile. The Chilean data exhibit an enticing first stage, with sharp drops (discontinuities) in class size at the cutoffs (multiples of 45). But household characteristics also differ considerably across these same kinks, probably because the Chilean school system, which is mostly privatized, offers both opportunities and incentives for wealthier students to attend schools just beyond the cutoffs. The possibility of such a pattern is an important caution for users of regression discontinuity methods, though Urqiola and Verhoogen note that the enrollment manipulation they uncover for Chile is far from ubiquitous and does not arise in the Angrist–Lavy study. A large measure of the attraction of the regression discontinuity design is its experimental spirit and the ease with which claims for validity of the design can be verified.

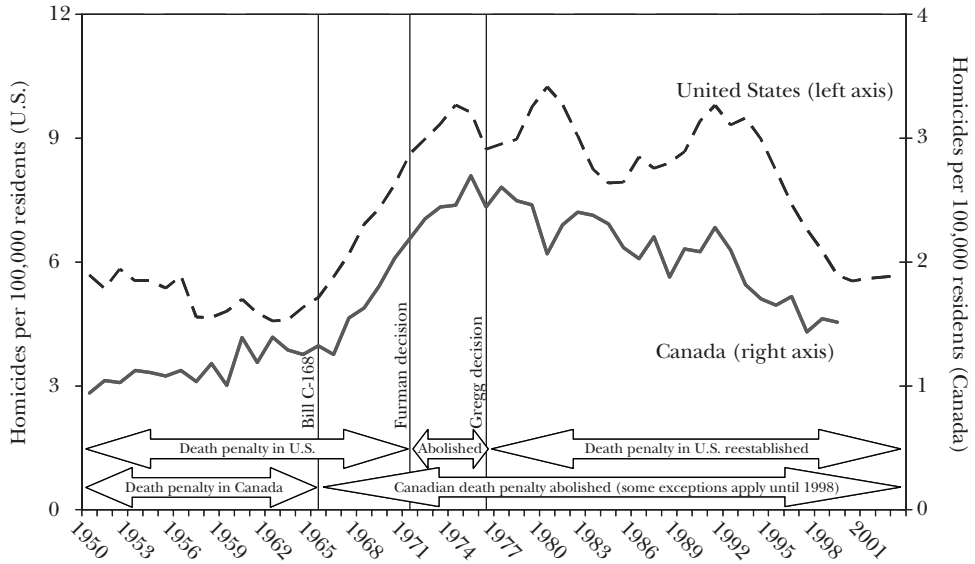
The last arrow in the quasi-experimental quiver is differences-in-differences, probably the most widely applicable design-based estimator. Differences-in-differences policy analysis typically compares the evolution of outcomes in groups affected more and less by a policy change. The most compelling differences-in-differences-type studies report outcomes for treatment and control observations for a period long enough to show the underlying trends, with attention focused on how deviations from trend relate to changes in policy. Figure 1, from Donohue and Wolfers (2005), illustrates this approach for the death penalty question. This figure plots homicide rates in Canada and the United States for over half a century, indicating

⁶ Fuzzy regression discontinuity designs are most easily analyzed using instrumental variables. In the language of instrumental variables, the relationship between achievement and kinks in the enrollment function is the reduced form, while the change in class size at the kinks is the first stage. The ratio of reduced form to first-stage effects is an instrumental variable estimate of the causal effect of class size on test scores. Imbens and Lemieux (2008) offer a practitioners’ guide to the use of regression discontinuity designs in economics.

Figure 1

Homicide Rates and the Death Penalty in the United States and Canada

(U.S. and Canada rates on the left and right y-axes, respectively)



Source: Donohue and Wolfers (2005).

periods when the death penalty was in effect in the two countries. The point of the figure is not to focus on Canada’s consistently lower homicide rate, but instead to show that Canadian and U.S. homicide rates move roughly in parallel, suggesting that America’s sharp changes in death penalty policy were of little consequence for murder. The figure also suggests that the deterrent effect would have to be large to be visible against the background noise of yearly fluctuations in homicide rates.

Paralleling the growth in quasi-experimental experiment designs, the number and scope of real experiments has increased dramatically, with a concomitant increase in the quality of experimental design, data collection, and statistical analysis. While 1970s-era randomized studies of the negative income tax were compromised by misreporting and differential attrition in treatment and control groups, researchers today give these concerns more attention and manage them more effectively. Such problems are often solved by a substantial reliance on administrative data, and a more sophisticated interpretation of survey data when administrative records are unavailable.

A landmark randomized trial related to education production is the Tennessee STAR experiment. In this intervention, more than 10,000 students were randomly assigned to classes of different sizes from kindergarten through third grade. Like the negative income tax experiments, the STAR experiment had its flaws. Not all subjects contributed follow-up data and some self-selected into smaller classes after random assignment. A careful analysis by Krueger (1999), however, shows clear

evidence of achievement gains in smaller classes, even after taking attrition and self-selection into account.⁷

Economists are increasingly running their own experiments as well as processing the data from experiments run by others. A recent randomized trial of a microfinance scheme, an important policy tool for economic development, is an ambitious illustration (Banerjee, Duflo, Glennerster, and Kinnan, 2009). This study evaluates the impact of offering small loans to independent business owners living in slums in India. The Banerjee et al. study randomizes the availability of microcredit across over 100 Indian neighborhoods, debunking the claim that realistic and relevant policy interventions cannot be studied with random assignment.

With the growing focus on research design, it's no longer enough to adopt the language of an orthodox simultaneous equations framework, labeling some variables endogenous and others exogenous, without offering strong institutional or empirical support for these identifying assumptions. The new emphasis on a credibly exogenous source of variation has also filtered down to garden-variety regression estimates, in which researchers are increasingly likely to focus on sources of omitted variables bias, rather than a quixotic effort to uncover the "true model" generating the data.⁸

More Transparent Discussion of Research Design

Over 65 years ago, Haavelmo submitted the following complaint to the readers of *Econometrica* (1944, p. 14): "A design of experiments (a prescription of what the physicists call a 'crucial experiment') is an essential appendix to any quantitative theory. And we usually have some such experiment in mind when we construct the theories, although—unfortunately—most economists do not describe their design of experiments explicitly."

In recent years, the notion that one's identification strategy—in other words, research design—must be described and defended has filtered deeply into empirical practice. The query "What's your identification strategy?" and others like it are now heard routinely at empirical workshops and seminars. Evidence for this claim comes from the fact that a full text search for the terms "empirical strategy," "identification strategy," "research design," or "control group" gets only 19 hits in Econlit from 1970–1989, while producing 742 hits from 1990–2009. We acknowledge that just because the author uses the term "research design" does not mean that he or she has a good one! Moreover, some older studies incorporate quality designs

⁷ A related development at the forefront of education research is the use of choice lotteries as a research tool. In many settings where an educational option is over-subscribed, allocation among applicants is by lottery. The result is a type of institutional random assignment, which can then be used to study school vouchers, charter schools, and magnet schools (for example, Rouse, 1998, who looks at vouchers).

⁸ The focus on omitted variables bias is reflected in a burgeoning literature using matching and the propensity score as an alternative (or complement) to regression. In the absence of random assignment, such strategies seek to eliminate observable differences between treatment and control groups, with little or no attention devoted to modeling the process determining outcomes. See Imbens and Wooldridge (2009) for an introduction.

without using today's language. Still, the shift in emphasis is dramatic and reflects a trend that's more than semantic.

Good designs have a beneficial side effect: they typically lend themselves to a simple explanation of empirical methods and a straightforward presentation of results. The key findings from a randomized experiment are typically differences in means between treatment and controls, reported before treatment (to show balance) and after treatment (to estimate causal effects). Nonexperimental results can often be presented in a manner that mimics this, highlighting specific contrasts. The Donohue and Wolfers (2005) differences-in-differences study mentioned above illustrates this by focusing on changes in American law as a source of quasi-experimental variation and documenting the parallel evolution of outcomes in treatment and control groups in a comparison of the United States and Canada.

Whither Sensitivity Analysis?

Responding to what he saw as the fragility of naive regression analysis, Leamer (1983) proposed extreme bounds analysis, which focuses on the distribution of results generated by a variety of specifications. An extreme version of extreme bounds analysis appears in Sala-i-Martin's (1997) paper reporting two million regressions related to economic growth. Specifically, in a variation on a procedure first proposed in this context by Levine and Renelt (1992), Sala-i-Martin computes two million of the many possible growth regressions that can be constructed from 62 explanatory variables. He retains a fixed set of three controls (GDP, life expectancy, and the primary school enrollment rate in 1960), leaving 59 possible "regressors of interest." From these 59, sets of three additional controls are chosen from 58 while the 59th is taken to be the one of interest. This process is repeated until every one of the 59 possible regressors of interest has played this role in equations with all possible sets of three controls, generating 30,857 regressions per regressor of interest. The object of this exercise is to see which variables are robustly significant across specifications.

Sala-i-Martin's (1997) investigation of extreme bounds analysis must have been fun. Happily, however, this kind of agnostic specification search has not emerged as a central feature of contemporary empirical work. Although Sala-i-Martin succeeds in uncovering some robustly significant relations (the "fraction of the population Confucian" is a wonderfully robust predictor of economic growth), we don't see why this result should be taken more seriously than the naive capital punishment specifications criticized by Leamer. Are these the right controls? Are six controls enough? How are we to understand sources of variation in one variable when the effects of three others, arbitrarily chosen, are partialled out? Wide-net searches of this kind offer little basis for a causal interpretation.

Design-based studies typically lead to a focused and much narrower specification analysis, targeted at *specific* threats to validity. For example, when considering results from a randomized trial, we focus on the details of treatment assignment and the evidence for treatment-control balance in pre-treatment variables. When using instrumental variables, we look at whether the instrument might have causal

effects on the outcome in ways other than through the channel of interest (in simultaneous equations lingo, this is an examination of the exclusion restriction). With differences-in-differences, we look for group-specific trends, since such trends can invalidate a comparison of changes across groups. In a regression discontinuity design, we look at factors like bunching at the cutoff point, which might suggest that the cutoff directly influenced behavior. Since the nature of the experiment is clear in these designs, the tack we should take when assessing validity is also clear.

Mad About Macro

In an essay read to graduating University of Chicago economics students in 1988, Robert Lucas (1988) described what, as he sees it, economists do. Lucas used the specific question of the connection between monetary policy and economic depression to frame his discussion, which is very much in the experimentalist spirit: “One way to demonstrate that I understand this connection—I think the only really convincing way—would be for me to engineer a depression in the United States by manipulating the US money supply.”

Ruling out such a national manipulation as immoral, Lucas (1988) describes how to create a depression by changing the money supply at Kennywood Park, an amusement park near Pittsburgh that is distinguished by stunning river views, wooden roller coasters, and the fact that it issues its own currency. Lucas’s story is evocative and compelling (the Kennywood allegory is a version of Lucas, 1973). We’re happy to see a macroeconomist of Lucas’s stature use an experimental benchmark to define causality and show a willingness to entertain quasi-experimental evidence on the effects of a change in the money supply. Yet this story makes us wonder why the real world of empirical macro rarely features design-based research.

Many macroeconomists have abandoned traditional empirical work entirely, focusing instead on “computational experiments,” as described in this journal by Kydland and Prescott (1996). In a computational experiment, researchers choose a question, build a (theoretical) model economy, “calibrate” the model so that its behavior mimics the real economy along some key statistical dimensions, and then run a computational experiment by changing model parameters (for example, tax rates or the money supply rule) to address the original question. The last two decades have seen countless studies in this mold, often in a dynamic stochastic general equilibrium framework. Whatever might be said in defense of this framework as a tool for clarifying the implications of economic models, it produces no direct evidence on the magnitude or existence of causal effects. An effort to put reasonable numbers on theoretical relations is harmless and may even be helpful. But it’s still theory.

Some rays of sunlight poke through the grey clouds of dynamic stochastic general equilibrium. One strand of empirical macro has turned away from modeling outcome variables such as GDP growth, focusing instead on the isolation of useful variation in U.S. monetary and fiscal policy. A leading contribution here

is Romer and Romer (1989), who, in the spirit of Friedman and Schwartz (1963), review the minutes of Federal Reserve meetings and try to isolate events that look like good monetary policy “experiments.” Their results suggest that monetary contractions have significant and long-lasting effects on the real economy. Later, in Romer and Romer (2004), they produced similar findings for the effects of policy shocks conditional on the Fed’s own forecasts.⁹

The Romers’ work is design based in spirit and, for the most part, in detail. Although a vast literature models Federal Reserve decision making, until recently, surprisingly few studies have made an institutional case for policy experiments as the Romers’ study does. Two recent monetary policy studies in the Romer spirit, and perhaps even closer to the sort of quasi-experimental work we read and do, are Richardson and Troost (2009), who exploit regional differences in Fed behavior during the Depression to study liquidity effects, and Velde (2009), who describes the results of an extreme monetary experiment much like the one Lucas envisioned (albeit in eighteenth-century France). Romer and Romer (2007) use methods similar to those they used for money to study fiscal policy, as do Ramey and Shapiro (1998) and Barro and Redlick (2009), who investigate the effects of large fiscal shocks due to wars.

The literature on empirical growth has long suffered from a lack of imagination in research design, but here too the picture has recently improved. The most influential design-based study in this area has probably been Acemoglu, Johnson, and Robinson (2001), who argue that good political institutions are a key ingredient in the recipe for growth, an idea growth economists have entertained for many decades. The difficulty here is that better institutions might be a luxury that richer countries can enjoy more easily, leading to a vexing reverse causality problem. Acemoglu, Johnson, and Robinson (2001) try to overcome this problem by using the differential mortality rates of European settlers in different colonies as an instrument for political institutions in the modern successor countries. Their argument goes: where Europeans faced high mortality rates, they couldn’t settle, and where Europeans couldn’t settle, colonial regimes were more extractive, with little emphasis on property rights and democratic institutions. Where European immigrants could settle, they frequently tried to emulate the institutional set-up of their home countries, with stronger property rights and more democratic institutions. This approach leads to an instrumental variables strategy where the instrument for the effect of institutions on growth is settler mortality.¹⁰

Acemoglu, Johnson, and Robinson (2001) are in the vanguard of promising research on the sources of economic growth using a similar style. Examples include Bleakley (2007), who looks at the effect of hookworm eradication on income in the American South; and Rodrik and Wacziarg (2005) and Persson and Tabellini

⁹ Angrist and Kuersteiner (2007) implement a version of the Romer and Romer (2004) research design using the propensity score and an identification argument cast in the language of potential outcomes commonly used in microeconomic program evaluation.

¹⁰ Albouy (2008) raises concerns about the settler mortality data that Acemoglu, Johnson, and Robinson (2001) used to construct instruments. See Acemoglu, Johnson, and Robinson (2006) for a response to earlier versions of Albouy’s critique.

(2008), who investigate interactions between democracy and growth using differences-in-differences type designs.

With these examples accumulating, macroeconomics seems primed for a wave of empirical work using better designs. Ricardo Reis, a recently tenured macroeconomist at Columbia University, observed in the wake of the 2008 financial crisis: “Macroeconomics has taken a turn towards theory in the last 10–15 years. Most young macroeconomists are more comfortable with proving theorems than with getting their hands on any data or speculating on current events.”¹¹ The charge that today’s macro agenda is empirically impoverished comes also from older macro warhorses like Mankiw (2006) and Solow (2008). But the recent economic crisis, fundamentally a macroeconomic and policy-related affair, has spawned intriguing design-based studies of the crisis’s origins in the mortgage market (Keys, Mukherjee, Seru, and Vig, 2010; Bubb and Kaufman, 2009). The theory-centric macro fortress appears increasingly hard to defend.

Industrial Disorganization

An important question at the center of the applied industrial organization agenda is the effect of corporate mergers on prices. One might think, therefore, that studies of the causal effects of mergers on prices would form the core of a vast micro-empirical literature, the way hundreds of studies in labor economics have looked at union relative wage effects. We might also have expected a large parallel literature evaluating merger policy, in the way that labor economists have looked at the effect of policies like right-to-work laws. But it isn’t so. In a recent review, Ashenfelter, Hosken, and Weinberg (2009) found only about 20 empirical studies evaluating the price effects of consummated mergers directly; for example, Borenstein (1990) compares prices on airline routes out of hubs affected to differing degrees by mergers. Research on the aggregate effects of merger policy seems to be even more limited; see the articles by Baker (2003) and Crandall and Winston (2003) in this journal for a review and conflicting interpretations.

The dominant paradigm for merger analysis in modern academic studies, sometimes called the “new empirical industrial organization,” is an elaborate exercise consisting of three steps: The first estimates a demand system for the product in question, often using the discrete choice/differentiated products framework developed by Berry, Levinsohn, and Pakes (1995). Demand elasticities are typically identified using instrumental variables for prices; often, the instruments are prices in other markets (as in Hausman, 1996). Next, researchers postulate a model of market conduct, say, Bertrand–Nash price-based competition between different brands or products. In the context of this model, the firms’ efforts to maximize profits lead to a set of relationships between prices

¹¹ As quoted by Justin Wolfers (2008) in his *New York Times* column “Freakonomics” (<http://freakonomics.blogs.nytimes.com/2008/03/31/more-on-the-missing-macroeconomists/>).

and marginal costs for each product, with the link provided by the substitution matrix estimated in the initial step. Finally, industry behavior is simulated with and without the merger of interest.

Nevo (2000) uses this approach to estimate the effect of mergers on the price of ready-to-eat breakfast cereals in a well-cited paper. Nevo's study is distinguished by careful empirical work, attention to detail, and a clear discussion of the superstructure of assumptions upon which it rests. At the same time, this elaborate superstructure should be of concern. The postulated demand system implicitly imposes restrictions on substitution patterns and other aspects of consumer behavior about which we have little reason to feel strongly. The validity of the instrumental variables used to identify demand equations—prices in other markets—turns on independence assumptions across markets that seem arbitrary. The simulation step typically focuses on a single channel by which mergers affect prices—the reduction in the number of competitors—when at least in theory a merger can lead to other effects like cost reductions that make competition tougher between remaining producers. In this framework, it's hard to see precisely which features of the data drive the ultimate results.

Can mergers be analyzed using simple, transparent empirical methods that trace a shorter route from facts to findings? The challenge for a direct causal analysis of mergers is to use data to describe a counterfactual world in which the merger didn't occur. Hastings (2004) does this in a study of the retail gasoline market. She analyzes the takeover of independent Thrifty stations by large vertically integrated station owner ARCO in California, with an eye to estimating the effects of this merger on prices at Thrifty's competitors. Hastings' research design specifies a local market for each station: treatment stations are near a Thrifty station, control stations are not. She then compares prices around the time of the merger using a straightforward differences-in-differences framework.

A drawback of the Hastings (2004) analysis is that it captures the effects of a merger on Thrifty's competitors, but not on the former Thrifty stations. Still, it seems likely that anticompetitive effects would turn up at any station operating in affected markets. We therefore see the Hastings approach as a fruitful change in direction. Her estimates have clear implications for the phenomenon of interest, while their validity turns transparently on the quality of the control group, an issue that can be assessed using pre-merger observations to compare price trends. Hastings's paper illustrates the power of this approach by showing almost perfectly parallel price trends for treatment and control stations in two markets (Los Angeles and San Diego) in pre-treatment months, followed by a sharp uptick in Thrifty competitor pricing after the merger.¹²

¹² As with most empirical work, Hastings's (2004) analysis has its problems and her conclusions may warrant qualification. Taylor, Kreisle, and Zimmerman (2007) fail to replicate Hastings's findings using an alternative data source. Here as elsewhere, however, a transparent approach facilitates replication efforts and constructive criticism.

For policy purposes, of course, regulators must evaluate mergers before they have occurred; design-based studies necessarily capture the effects of mergers after the fact. Many new empirical industrial organization studies forecast counterfactual outcomes based on models and simulations, without a clear foundation in experience. But should antitrust regulators favor the complex, simulation-based estimates coming out of the new empirical industrial organization paradigm over a transparent analysis of past experience? At a minimum, we'd expect such a judgment to be based on evidence showing that the simulation-based approach delivers reasonably accurate predictions. As it stands, the proponents of this work seem to favor it as a matter of principle.

So who can you trust when it comes to antitrust? Direct Hastings (2004)-style evidence, or structurally derived estimates as in Nevo (2000)? We'd be happy to see more work trying to answer this question by contrasting credible quasi-experimental estimates with results from the new empirical industrial organization paradigm. A pioneering effort in this direction is Hausman and Leonard's (2002) analysis contrasting "direct" (essentially, differences-in-differences) and "indirect" (simulation-based) estimates of the equilibrium price consequences of a new brand of toilet paper. They evaluate the economic assumptions underlying alternative structural models (for example, Nash-Bertrand competition) according to whether the resulting structural estimates match the direct estimates. This is reminiscent of Lalonde's (1986) comparison of experimental and nonexperimental training estimates, but instead of contrasting model-based estimates with those from a randomized trial, the direct estimates are taken to provide a benchmark that turns on fewer assumptions than the structural approach. Hausman and Leonard conclude that one of their three structural models produces estimates "reasonably similar" to the direct estimates. Along the same lines, Peters (2006) looks at the predictive value of structural analyses of airline mergers, and finds that structural simulation methods yield poor predictions of post-merger ticket prices. Likewise, Ashenfelter and Hosken (2008) compare differences-in-differences-type estimates of the effects of the breakfast cereals merger to those reported by Nevo (2000). Ashenfelter and Hoskens conclude that transparently identified design-based results differ markedly from those produced by the structural approach.

A good structural model might tell us something about economic mechanisms as well as causal effects. But if the information about mechanisms is to be worth anything, the structural estimates should line up with those derived under weaker assumptions. Does the new empirical industrial organization framework generate results that match credible design-based results? So far, the results seem mixed at best. Of course, the question of which estimates to prefer turns on the quality of the relevant quasi-experimental designs and our faith in the ability of a more elaborate theoretical framework to prop up a weakly identified structural model. We find the empirical results generated by a good research design more compelling than the conclusions derived from a good theory, but we also hope to see industrial organization move towards stronger and more transparent identification strategies in a structural framework.

Has the Research Design Pendulum Swung Too Far?

The rise of the experimentalist paradigm has provoked a reaction, as revolutions do. The first counterrevolutionary charge raises the question of external validity—the concern that evidence from a given experimental or quasi-experimental research design has little predictive value beyond the context of the original experiment. The second charge is that experimentalists are playing small ball while big questions go unanswered.

External Validity

A good research design reveals a particular truth, but not necessarily the whole truth. For example, the Tennessee STAR experiment reduced class sizes from roughly 25 to 15. Changes in this range need not reveal the effect of reductions from 40 students to 30. Similarly, the effects might be unique to the state of Tennessee. The criticism here—made by a number of authors including Heckman (1997); Rosenzweig and Wolpin (2000); Heckman and Urzua (2009); and Deaton (2009)—is that in the quest for internal validity, design-based studies have become narrow or idiosyncratic.

Perhaps it's worth restating an obvious point. Empirical evidence on any given causal effect is always local, derived from a particular time, place, and research design. Invocation of a superficially general structural framework does not make the underlying variation or setting more representative. Economic theory often suggests general principles, but extrapolation of causal effects to new settings is always speculative. Nevertheless, anyone who makes a living out of data analysis probably believes that heterogeneity is limited enough that the well-understood past can be informative about the future.

A constructive response to the specificity of a given research design is to look for more evidence, so that a more general picture begins to emerge. For example, one of us (Angrist) has repeatedly estimated the effects of military service, with studies of veterans of World War II, the Vietnam era, the first Gulf War, and periods in between. The cumulative force of these studies has some claim to external validity—that is, they are helpful in understanding the effects of military service for those who served in any period and therefore, hopefully, for those who might serve in the future. In general, military service tends to depress civilian earnings, at least for whites, a finding that is both empirically consistent and theoretically coherent. The primary theoretical channel by which military service affects earnings is human capital, particularly in the form of lost civilian experience. In a design-based framework, economic theory helps us understand the picture that emerges from a constellation of empirical findings, but does not help us paint the picture. For example, the human capital story is not integral to the validity of instrumental variable estimates using draft lottery numbers as instruments for Vietnam-era military service (as in Angrist, 1990). But human capital theory provides a framework that reconciles larger losses early in a veteran's career (when experience profiles tend to be steeper) with losses dissipating after many years (as shown in Angrist and Chen, 2008).

The process of accumulating empirical evidence is rarely sexy in the unfolding, but accumulation is the necessary road along which results become more general (Imbens, 2009, makes a similar point). The class size literature also illustrates this process at work. Reasonably well-identified studies from a number of advanced countries, at different grade levels and subjects, and for class sizes ranging anywhere from a few students to about 40, have produced estimates within a remarkably narrow band (Krueger, 1999; Angrist and Lavy, 1999; Rivkin, Hanushek, and Kain, 2005; Heinesen, forthcoming). Across these studies, a ten-student reduction in class size produces about a 0.2 to 0.3 standard deviation increase in individual test scores. Smaller classes do not always raise test scores, so the assessment of findings should be qualified (see, for example, Hoxby, 2000). But the weight of the evidence suggests that class size reductions generate modest achievement gains, albeit at high cost.

Applied micro fields are not unique in accumulating convincing empirical findings. The evidence on the power of monetary policy to influence the macro economy also seems reasonably convincing. As we see it, however, the most persuasive evidence on this point comes not from elaborate structural models, which only tell us that monetary policy does or does not affect output depending on the model, but from credible empirical research designs, as in some of the work we have discussed. Not surprisingly, the channels by which monetary policy affects output are less clear than the finding that there is an effect. Questions of why a given effect appears are usually harder to resolve than the questions of whether it appears or how large it is. Like most researchers, we have an interest in mechanisms as well as causal effects. But inconclusive or incomplete evidence on mechanisms does not void empirical evidence of predictive value. This point has long been understood in medicine, where clinical evidence of therapeutic effectiveness has for centuries run ahead of the theoretical understanding of disease.

Taking the “Econ” out of Econometrics too?

Related to the external validity critique is the claim that the experimentalist paradigm leads researchers to look for good experiments, regardless of whether the questions they address are important. In an engaging account in *The New Republic*, Scheiber (2007) argued that young economists have turned away from important questions like poverty, inequality, and unemployment to study behavior on television game shows. Scheiber quotes a number of distinguished academic economists who share this concern. Raj Chetty comments: “People think about the question less than the method . . . so you get weird papers, like sanitation facilities in Native American reservations.” James Heckman is less diplomatic: “In some quarters of our profession, the level of discussion has sunk to the level of a *New Yorker* article.”

There is no shortage of academic triviality. Still, Scheiber’s (2007) critique misses the mark because he equates triviality with narrowness of context. For example, he picks on DellaVigna and Malmendier (2006), who look at the attendance and renewal decisions of health club members, and on Conlin, O’Donoghue, and Vogelsang (2007), who study catalog sales of winter clothing. Both studies are concerned with the behavioral economics notion of present-oriented biases, an

issue with far-reaching implications for economic policy and theory. The market for snow boots seems no less interesting in this context than any other retail market, and perhaps more so if the data are especially good. We can look to these design-based studies to validate the findings from more descriptive empirical work on bigger-ticket items. For example, DellaVigna and Paserman (2005) look for present-oriented biases in job search behavior.

In the empirical universe, evidence accumulates across settings and study designs, ultimately producing some kind of consensus. Small ball sometimes wins big games. In our field, some of the best research designs used to estimate labor supply elasticities exploit natural and experimenter-induced variation in specific labor markets. Oettinger (1999) analyzes stadium vendors' reaction to wage changes driven by changes in attendance, while Fehr and Goette (2007) study bicycle messengers in Zurich who, in a controlled experiment, received higher commission rates for one month only. These occupations might seem small and specialized, but they are no less representative of today's labor market than the durable manufacturing sector that has long been of interest to labor economists.

These examples also serve to refute the claim that design-based empirical work focuses on narrow policy effects and cannot uncover theoretically grounded structural parameters that many economists care about. Quasi-experimental labor supply studies such as Oettinger (1999) and Fehr and Goette (2007) try to measure the intertemporal substitution elasticity, a structural parameter that can be derived from a stochastic dynamic framework. Labor demand elasticities, similarly structural, can also be estimated using quasi-experiments, as in Card (1990b), who exploits real wage variation generated by partial indexation of union contracts.

Quasi-experimental empirical work is also well suited to the task of contrasting competing economic hypotheses. The investigations of present-oriented biases mentioned above focus on key implications of alternative models. In a similarly theory-motivated study, Karlan and Zinman (2009) try to distinguish moral hazard from adverse selection in the consumer credit market using a clever experimental design involving two-stage randomization. First, potential borrowers were offered different interest rates before they applied for loans. Their initial response to variation in interest rates is used to gauge adverse selection. Some of the customers who took loans were then randomly given rates lower than the rates initially offered. This variation is used to identify moral hazard in a sample where everyone has already committed to borrow.

What about grand questions that affect the entire world or the march of history? Nunn (2008) uses a wide range of historical evidence, including sailing distances on common trade routes, to estimate the long-term growth effects of the African slave trade. Deschênes and Greenstone (2007) use random year-to-year fluctuations in temperature to estimate effects of climate change on energy use and mortality. In a study of the effects of foreign aid on growth, Rajan and Subramanian (2008) construct instruments for foreign aid from the historical origins of donor-recipient relations. These examples and many more speak eloquently for the wide applicability of a design-based approach. Good research designs complement good questions. At

the same time, in favoring studies that feature good designs, we accept an incremental approach to empirical knowledge in which well-designed studies get the most weight while other evidence is treated as more provisional.

Conclusion

Leamer (1983) drew an analogy between applied econometrics and classical experimentation, but his proposal for the use of extreme bounds analysis to bring the two closer is not the main reason why empirical work in economics has improved. Improvement has come mostly from better research designs, either by virtue of outright experimentation or through the well-founded and careful implementation of quasi-experimental methods. Empirical work in this spirit has produced a credibility revolution in the fields of labor, public finance, and development economics over the past 20 years. Design-based revolutionaries have notched many successes, putting hard numbers on key parameters of interest to both policymakers and economic theorists. Imagine what could be learned were a similar wave to sweep the fields of macroeconomics and industrial organization.

■ *We thank Guido Imbens for suggesting this topic and for feedback; Daron Acemoglu, Olivier Blanchard, John Donohue, Isaac Ehrlich, Glenn Ellison, Jeff Grogger, Radha Iyengar, Larry Katz, Alan Krueger, Ethan Ilzetzki, Guido Lorenzoni, Albert Marcet, Aviv Nevo, Alan Manning, Bruce Meyer, Parag Pathak, Gary Solon, Matt Weinberg, and Justin Wolfers for helpful comments and discussions; and the JEP editors—David Autor, James Hines, Charles Jones, and Timothy Taylor—for comments on earlier drafts. Remaining errors and omissions are our own.*

References

- Acemoglu, Daron, Simon Johnson, and James A. Robinson.** 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, 91(5): 1369–1401.
- Acemoglu, Daron, Simon Johnson, and James A. Robinson.** 2006. "Reply to the Revised (May 2006) Version of David Albouy's 'The Colonial Origins of Comparative Development: An Investigation of the Settler Mortality Data.'" Available at: <http://econ-www.mit.edu/faculty/acemoglu/paper>.
- Albouy, David Y.** 2008. "The Colonial Origins of Comparative Development: An Investigation of the Settler Mortality Data." NBER Working Paper 14130.
- Angrist, Joshua D.** 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*, 80(3): 313–36.
- Angrist, Joshua D.** 2004. "Education Research Changes Tack." *Oxford Review of Economic Policy*, 20(2): 198–212.
- Angrist, Joshua D., and Stacey Chen.** 2008. "Long-term Economic Consequences of Vietnam-Era Conscription: Schooling, Experience and Earnings." IZA Discussion Paper 3628.
- Angrist, Joshua D., and Guido W. Imbens.** 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association*, 90(430): 431–42.
- Angrist, Joshua D., and Alan B. Krueger.** 1991. "Does Compulsory School Attendance Affect

- Schooling and Earnings?" *Quarterly Journal of Economics*, 106(4): 976–1014.
- Angrist, Joshua D., and Alan B. Krueger.** 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, vol. 3, ed. O. Ashenfelter and D. Card, 1277–1366. Amsterdam: North-Holland.
- Angrist, Joshua D., and Guido Kuersteiner.** 2007. "Semiparametric Causality Tests Using the Policy Propensity Score." NBER Working Paper 10975.
- Angrist, Joshua D., and Victor Lavy.** 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics*, 114(2): 533–75.
- Angrist, Joshua D., and Jörn-Steffen Pischke.** 2009. *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton: Princeton University Press.
- Ashenfelter, Orley.** 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics*, 60(1): 47–57.
- Ashenfelter, Orley.** 1987. "The Case for Evaluating Training Programs with Randomized Trials." *Economics of Education Review*, 6(4): 333–38.
- Ashenfelter, Orley, and David Card.** 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics*, 67(4): 648–60.
- Ashenfelter, Orley, and Daniel Hosken.** 2008. "The Effect of Mergers on Consumer Prices: Evidence from Five Selected Case Studies." NBER Working Paper 13859.
- Ashenfelter, Orley, Daniel Hosken, and Matthew Weinberg.** 2009. "Generating Evidence to Guide Merger Enforcement?" NBER Working Paper 14798.
- Ashenfelter, Orley, and Mark W. Plant.** 1990. "Nonparametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs." *Journal of Labor Economics*, 8(1, Part 2): S396–S415.
- Ayres, Ian.** 2007. *Super Crunchers*. New York: Bantam Books.
- Baker, Jonathon B.** 2003. "The Case for Antitrust Enforcement." *Journal of Economic Perspectives*, 17(4): 27–50.
- Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan.** 2009. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." Unpublished manuscript, MIT Department of Economics, May.
- Barro, Robert J., and Charles J. Redlick.** 2009. "Macroeconomic Effects from Government Purchases and Taxes." NBER Working Paper 15369.
- Berry, Steven, James Levinsohn, and Ariel Pakes.** 1995. "Automobile Prices in Market Equilibrium." *Econometrica*, 63(4): 841–90.
- Bleakley, Hoyt.** 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *Quarterly Journal of Economics*, 122(1): 73–117.
- Borenstein, Severin.** 1990. "Airline Mergers, Airport Dominance, and Market Power." *American Economic Review*, 80(2): 400–404.
- Bound, John, David Jaeger, and Regina Baker.** 1995. "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association*, 90(430): 443–50.
- Bowers, William J., and Glenn L. Pierce.** 1975. "The Illusion of Deterrence in Isaac Ehrlich's Research on Capital Punishment." *Yale Law Journal*, 85(2): 187–208.
- Bubb, Ryan, and Alex Kaufman.** 2009. "Securitization and Moral Hazard: Evidence from a Lender Cutoff Rule." Federal Reserve Bank of Boston Public Policy Discussion Paper No. 09-5.
- Campbell, Donald, and Julian Stanley.** 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Card, David.** 1990a. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, 43(2): 245–57.
- Card, David.** 1990b. "Unexpected Inflation, Real Wages, and Employment Determination in Union Contracts." *American Economic Review*, 80(4): 669–88.
- Card, David, and Alan B. Krueger.** 1992a. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy*, 100(1): 1–40.
- Card, David, and Alan B. Krueger.** 1992b. "School Quality and Black-White Relative Earnings: A Direct Assessment." *Quarterly Journal of Economics*, 107(1): 151–200.
- Chamberlain, Gary.** 1984. "Panel Data." In *Handbook of Econometrics*, vol. 2, ed. Zvi Griliches and Michael D. Intriligator, 1248–1318. Amsterdam: North-Holland.
- Coleman, James S., et al.** 1966. *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.
- Conlin, Michael, Ted O'Donoghue, and Timothy J. Vogelsang.** 2007. "Projection Bias in Catalog Orders." *American Economic Review*, 97(4): 1217–1249.
- Cook, Thomas D., and Donald T. Campbell.** 1979. *Quasi-Experimentation: Design and Analysis for Field Settings*. Chicago: Rand McNally.
- Cooley, Thomas F., and Stephen F. LeRoy.** 1986. "What Will Take the Con Out of Econometrics? A Reply to McAleer, Pagan, and Volker." *American Economic Review*, 76(3): 504–507.

- Crandall, Robert W., and Clifford Winston.** 2003. "Does Antitrust Policy Improve Consumer Welfare? Assessing the Evidence." *The Journal of Economic Perspectives*, 17(4): 3–26.
- Deaton, Angus.** 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." NBER Working Paper 14690.
- DellaVigna, Stefano, and Ulrike Malmendier.** 2006. "Paying Not to Go to the Gym." *American Economic Review*, 96(3): 694–719.
- DellaVigna, Stefano, and Daniele Paserman.** 2005. "Job Search and Impatience." *Journal of Labor Economics*, 23(3): 527–88.
- Deschênes, Olivier, and Michael Greenstone.** 2007. "Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US." NBER Working Paper 13178.
- Donohue, John J., and Justin Wolfers.** 2005. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review*, vol. 58, pp. 791–845.
- Duflo, Esther, and Michael Kremer.** 2008. "Use of Randomization in the Evaluation of Development Effectiveness." In *Evaluating Development Effectiveness*, World Bank Series on Evaluation and Development, vol. 7, pp. 93–120. Transaction Publishers.
- Ehrlich, Isaac.** 1975a. "The Deterrent Effect of Capital Punishment: A Question of Life and Death." *American Economic Review*, 65(3): 397–417.
- Ehrlich, Isaac.** 1975b. "Deterrence: Evidence and Inference." *Yale Law Journal*, 85(2): 209–27.
- Ehrlich, Isaac.** 1977a. "The Deterrent Effect of Capital Punishment: Reply." *American Economic Review*, 67(3): 452–58.
- Ehrlich, Isaac.** 1977b. "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence." *Journal of Political Economy*, 85(4): 741–88.
- Ehrlich, Isaac.** 1987. "On the Issue of Causality in the Economic Model of Crime and Law Enforcement: Some Theoretical Considerations and Experimental Evidence." *American Economic Review*, 77(2): 99–106.
- Ehrlich, Isaac.** 1996. "Crime, Punishment, and the Market for Offenses." *Journal of Economic Perspectives*, 10(1): 43–67.
- Ehrlich, Isaac, and Zhiqiang Liu.** 1999. "Sensitivity Analyses of the Deterrence Hypothesis: Let's Keep the Econ in Econometrics." *Journal of Law & Economics*, 42(1): 455–87.
- Fehr, Ernst, and Lorenz Goette.** 2007. "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review*, 97(1): 298–317.
- Feldstein, Martin, and Charles Horioka.** 1980. "Domestic Saving and International Capital Flows." *Economic Journal*, 90(358): 314–29.
- Friedman, Milton, and Anna J. Schwartz.** 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press for the National Bureau of Economic Research.
- Gertler, Paul.** 2004. "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment." *American Economic Review*, 94(2): 336–41.
- Goldberger, Arthur S.** 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.
- Greenberg, David, and Harlan Halsey.** 1983. "Systematic Misreporting and Effects of Income Maintenance Experiments on Work Effort: Evidence from the Seattle–Denver Experiment." *Journal of Labor Economics*, 1(4): 380–407.
- Griliches, Zvi.** 1986. "Economic Data Issues." In *Handbook of Econometrics*, vol. 3, ed. Zvi Griliches and Michael D. Intriligator, 1465–1514. Amsterdam: North-Holland.
- Grogger, Jeffrey.** 1990. "The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts." *Journal of the American Statistical Association*, 85(410): 295–303.
- Gruber, Jonathan.** 1994. "The Incidence of Mandated Maternity Benefits." *American Economic Review*, 84(3): 662–41.
- Haavelmo, Trygve.** 1944. "The Probability Approach in Econometrics." *Econometrica*, 12(Supplement): 1–115.
- Hanushek, Eric A.** 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature*, 24(3): 1141–77.
- Hastings, Justine S.** 2004. "Vertical Relationships and Competition in Retail Gasoline Markets: Empirical Evidence from Contract Changes in Southern California." *American Economic Review*, 94(1): 317–28.
- Hausman, Jerry A.** 1996. "Valuation of New Goods under Perfect and Imperfect Competition." In *The Economics of New Goods*, ed. Timothy F. Bresnahan and Robert J. Gordon, 209–247. Chicago: National Bureau of Economic Research.
- Hausman, Jerry A., and Gregory K. Leonard.** 2002. "The Competitive Effects of a New Product Introduction: A Case Study." *Journal of Industrial Economics*, 50(3): 237–63.
- Heckman, James J.** 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources*, 32(3): 441–62.
- Heckman, James J., and Sergio Urzua.** 2009. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." NBER Working Paper 14706.

- Heckman, James J., Anne Layne-Farrar, and Petra Todd.** 1996. "Does Measured School Quality Really Matter?" In *Does Money Matter?: The Effect of School Resources on Student Achievement and Adult Success*, ed. Gary Burtless, 192–289. Washington, DC: Brookings Institution Press.
- Heinesen, Eskil.** Forthcoming. "Estimating Class-Size Effects Using Within-School Variation in Subject-Specific Classes." *Economic Journal*.
- Hendry, David F.** 1980. "Econometrics—Alchemy or Science?" *Economica*, 47(188): 387–406.
- Hoenack, Stephen A., and William C. Weiler.** 1980. "A Structural Model of Murder Behavior and the Criminal Justice System." *American Economic Review*, 70(3): 327–41.
- Hoxby, Caroline M.** 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics*, 115(4): 1239–85.
- Imbens, Guido W.** 2009. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." NBER Working Paper 14896.
- Imbens, Guido W., and Thomas Lemieux.** 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142(2): 615–35.
- Imbens, Guido W., and Jeffrey M. Wooldridge.** 2009. "Recent Developments in the Econometrics of Program Development." *Journal of Economic Literature*, 47(1): 5–86.
- Jacob, Brian A.** 2004. "Public Housing, Housing Vouchers and Student Achievement: Evidence from Public Housing Demolitions in Chicago." *American Economic Review*, 94(1): 233–58.
- Karlan, Dean, and Jonathan Zinman.** 2009. "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment." *Econometrica*, 77(6): 1993–2008.
- Keys, Benjamin, Tanmoy Mukherjee, Amit Seru, and Vikrant Vig.** 2010. "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Quarterly Journal of Economics*, 125(1): 307–62.
- Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz.** 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.
- Krueger, Alan B.** 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics*, 114(2): 497–532.
- Krueger, Alan B.** 2003. "Economic Considerations and Class Size." *Economic Journal*, 113(485): F34–F63.
- Kydland, Finn E., and Edward C. Prescott.** 1977. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy*, 85(3): 473–92.
- Kydland, Finn E., and Edward C. Prescott.** 1996. "The Computational Experiment: An Econometric Tool." *Journal of Economic Perspectives*, 10(1): 69–85.
- LaLonde, Robert J.** 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, 76(4): 604–620.
- Leamer, Edward.** 1978. *Specification Searches: Ad Hoc Inference with Non Experimental Data*. New York: John Wiley and Sons.
- Leamer, Edward.** 1983. "Let's Take the Con Out of Econometrics." *American Economic Review*, 73(1): 31–43.
- Leamer, Edward.** 1985. "Sensitivity Analyses Would Help." *American Economic Review*, 75(3): 308–313.
- Levine, Ross, and David Renelt.** 1992. "A Sensitivity Analysis of Cross-Country Growth Regressions." *American Economic Review*, 82(4): 942–63.
- Lucas, Robert E.** 1973. "Some International Evidence on Output–Inflation Tradeoffs." *American Economic Review*, 63(3): 326–34.
- Lucas, Robert E.** 1976. "Econometric Policy Evaluation: A Critique." In *Carnegie-Rochester Conference Series on Public Policy*, vol. 1, pp. 19–46.
- Lucas, Robert E.** 1988. "What Economists Do." Unpublished.
- Mankiw, Gregory N.** 2006. "The Macroeconomist as Scientist and Engineer." *Journal of Economic Perspectives*, 20(4): 29–46.
- Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, and Arleen Leibowitz.** 1987. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." *American Economic Review*, 77(3): 251–77.
- McAleer, Michael, Adrian R. Pagan, Paul A. Volker.** 1985. "What Will Take the Con Out of Econometrics?" *American Economic Review*, 75(3): 293–307.
- McCrary, Justin.** 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698–714.
- Meyer, Bruce D.** 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business & Economic Statistics*, 13(2): 151–61.
- Nevo, Aviv.** 2000. "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry." *The RAND Journal of Economics*, 31(3): 395–42.
- Nunn, Nathan.** 2008. "The Long-Term Effects of Africa's Slave Trades." *Quarterly Journal of Economics*, 123(1): 139–76.
- Obstfeld, Maurice.** 1995. "International Capital Mobility in the 1990s." In *Understanding*

Interdependence: The Macroeconomics of the Open Economy, ed. Peter B. Kenen, 201–261. Princeton: Princeton University Press.

Oettinger, Gerald S. 1999. “An Empirical Analysis of the Daily Labor Supply of Stadium Vendors.” *Journal of Political Economy*, 107(2): 360–92.

Passell, Peter, and John B. Taylor. 1977. “The Deterrent Effect of Capital Punishment: Another View.” *American Economic Review*, 67(3): 445–51.

Persson Torsten, and Guido Tabellini. 2008. “The Growth Effect of Democracy: Is it Heterogeneous and How can It be Estimated?” Chap. 13 in *Institutions and Economic Performance*, ed. E. Helpman. Cambridge, MA: Harvard University Press.

Peters, Craig. 2006. “Evaluating the Performance of Merger Simulation: Evidence from the US Airline Industry.” *Journal of Law and Economics*, 49(2): 627–49.

Phillips, David P. 1980. “The Deterrent Effect of Capital Punishment: New Evidence on an Old Controversy.” *American Journal of Sociology*, 86(1): 139–48.

Rajan, Raghuram G., and Arvind Subramanian. 2008. “Aid and Growth: What Does the Cross-Country Evidence Really Show?” *Review of Economics and Statistics*, 90(4): 643–65.

Ramey, Valerie, and Matthew D. Shapiro. 1998. “Costly Capital Reallocation and the Effects of Government Spending.” *Carnegie-Rochester Conference Series on Public Policy*, 48(1): 145–94.

Richardson, Gary, and William Troost. 2009. “Monetary Intervention Mitigated Banking Panics during the Great Depression: Quasi-Experimental Evidence from a Federal Reserve District Border, 1929–1933.” *Journal of Political Economy*, 117(6): 1031–73.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools, and Academic Achievement.” *Econometrica*, 73(2): 417–58.

Rodrik, Dani, and Romain Wacziarg. 2005. “Do Democratic Transitions Produce Bad Outcomes?” *American Economic Review*, 95(2): 50–55.

Romer, Christina D., and David H. Romer. 1989. “Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz.” *NBER Macroeconomics Annual*, vol. 4, pp. 121–70.

Romer, Christina D., and David H. Romer. 2004. “A New Measure of Monetary Shocks: Derivation and Implications.” *American Economic Review*, 94(4): 1055–1084.

Romer, Christina D., and David H. Romer. 2007. “The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of

Fiscal Shocks.” NBER Working Paper 13264.

Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. “Natural ‘Natural Experiments’ in Economics.” *Journal of Economic Literature*, 38(4): 827–74.

Rouse, Cecilia. 1998. “Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program.” *Quarterly Journal of Economics*, 113(2): 553–602.

Sala-i-Martin, Xavier. 1997. “I Just Ran Two Million Regressions.” *American Economic Review*, 87(2): 178–83.

Scheiber, Noam. 2007. “Freaks and Geeks. How Freakonomics Is Ruining the Dismal Science.” *The New Republic*, April 2, pp. 27–31.

Schultz, T. Paul. 2004. “School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program.” *Journal of Development Economics*, 74(1): 199–250.

Sims, Christopher A. 1980. “Macroeconomics and Reality.” *Econometrica*, 48(1): 1–48.

Sims, Christopher A. 1988. “Uncertainty across Models.” *American Economic Review*, 78(2): 163–67.

Solon, Gary. 1985. “Work Incentive Effects of Taxing Unemployment Insurance.” *Econometrica*, 53(2): 295–306.

Solow, Robert. 2008. “The State of Macroeconomics.” *Journal of Economic Perspectives*, 22(1): 243–249.

Summers, Anita A., and Barbara L. Wolfe. 1977. “Do Schools Make a Difference?” *American Economic Review*, 67(4): 639–52.

Taylor, Christopher, Nicholas Kreisle, and Paul Zimmerman. 2007. “Vertical Relationships and Competition in Retail Gasoline Markets: Comment.” The Federal Trade Commission, Bureau of Economics Working Paper 291.

Urquiola, Miguel, and Eric Verhoogen. 2009. “Class-size Caps, Sorting, and the Regression-Discontinuity Design.” *American Economic Review*, 99(1): 179–215.

Velde, Francois. 2009. “Chronicles of a Deflation Unforetold.” *Journal of Political Economy*, 117(4): 591–634.

White, Halbert. 1980a. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, 48(4): 817–38.

White, Halbert. 1980b. “Using Least Squares to Approximate Unknown Regression Functions.” *International Economic Review*, 21(1): 149–70.

Wolfers, Justin. 2008. “More on the Missing Macroeconomists.” “Freakonomics” column of the *New York Times*, March 31. <http://freakonomics.blogs.nytimes.com/2008/03/31/more-on-the-missing-macroeconomists/>.

This article has been cited by:

1. Ruopeng An, Roland Sturm. 2012. School and Residential Neighborhood Food Environment and Diet Among California Youth. *American Journal of Preventive Medicine* **42**:2, 129-135. [[CrossRef](#)]
2. Madhu Sudan Mohanty. 2012. Effects of Positive Attitude and Optimism on Wage and Employment: A Double Selection Approach. *Journal of Socio-Economics* . [[CrossRef](#)]
3. Allyson Pollock, Azeem Majeed, Alison Macfarlane, Ian Greener, Graham Kirkwood, Howard Mellett, Sylvia Godden, Sean Boyle, Carol Morelli, Petra Brhlikova. 2011. In defence of our research on competition in England's National Health Service – Authors' reply. *The Lancet* **378**:9809, 2065-2066. [[CrossRef](#)]
4. G. Andrew Karolyi. 2011. The Ultimate Irrelevance Proposition in Finance?. *Financial Review* **46**:4, 485-512. [[CrossRef](#)]
5. Kevin Milligan. 2011. The design of tax policy in Canada: thoughts prompted by Richard Blundell's 'Empirical evidence and tax policy design'. *Canadian Journal of Economics/Revue canadienne d'économique* **44**:4, 1184-1194. [[CrossRef](#)]
6. Nicholas Bloom, Zack Cooper, Martin Gaynor, Stephen Gibbons, Simon Jones, Alistair McGuire, Rodrigo Moreno-Serra, Carol Propper, John Van Reenen, Stephan Seiler. 2011. In defence of our research on competition in England's National Health Service. *The Lancet* . [[CrossRef](#)]
7. Jose G. Montalvo. 2011. Re-examining the evidence on the electoral impact of terrorist attacks: The Spanish election of 2004. *Electoral Studies* . [[CrossRef](#)]
8. Degnet Abebaw. 2011. INFANT AND CHILD HEALTH IN ETHIOPIA: REFLECTIONS ON REGIONAL PATTERNS AND CHANGES. *Journal of International Development* n/a-n/a. [[CrossRef](#)]
9. Ole Dahl Rasmussen, Nikolaj Malchow-Møller, Thomas Barnebeck Andersen. 2011. Walking the talk: the need for a trial registry for development interventions. *Journal of Development Effectiveness* 1-18. [[CrossRef](#)]
10. Valerie A. Ramey. 2011. Can Government Purchases Stimulate the Economy?. *Journal of Economic Literature* **49**:3, 673-685. [[Abstract](#)] [[View PDF article](#)] [[PDF with links](#)]
11. François Claveau. 2011. Evidential variety as a source of credibility for causal inference: beyond sharp designs and structural models. *Journal of Economic Methodology* **18**:3, 233-253. [[CrossRef](#)]
12. Michael A. Clemens, Gabriel Demombynes. 2011. When does rigorous impact evaluation make a difference? The case of the Millennium Villages. *Journal of Development Effectiveness* **3**:3, 305-339. [[CrossRef](#)]
13. Joel Slemrod, Caroline Weber. 2011. Evidence of the invisible: toward a credibility revolution in the empirical analysis of tax evasion and the informal economy. *International Tax and Public Finance* . [[CrossRef](#)]
14. Jeffrey Zabel, Maurice Dalton. 2011. The impact of minimum lot size regulations on house prices in Eastern Massachusetts. *Regional Science and Urban Economics* . [[CrossRef](#)]
15. Judea Pearl. 2011. Statistics and Causality: Separated to Reunite-Commentary on Bryan Dowd's "Separated at Birth". *Health Services Research* **46**:2, 421-429. [[CrossRef](#)]
16. Hilmar Schneider, Arne Uhlendorff, Klaus F. Zimmermann. 2011. Mit Workfare aus der Sozialhilfe? Lehren aus einem Modellprojekt. *Zeitschrift für ArbeitsmarktForschung* . [[CrossRef](#)]
17. Arnold Kling. 2011. MACROECONOMETRICS: THE SCIENCE OF HUBRIS. *Critical Review* **23**:1, 123-133. [[CrossRef](#)]

18. Ole Rogeberg, Hans Olav Melberg. 2011. Acceptance of unsupported claims about reality: a blind spot in economics. *Journal of Economic Methodology* **18**:1, 29-52. [[CrossRef](#)]
19. Henk Folmer, Olof Johansson-Stenman. 2011. Does Environmental Economics Produce Aeroplanes Without Engines? On the Need for an Environmental Social Science. *Environmental and Resource Economics* **48**:3, 337-361. [[CrossRef](#)]
20. Orley Ashenfelter, Daniel Hosken, Michael Vita, Matthew Weinberg. 2011. Retrospective Analysis of Hospital Mergers. *International Journal of the Economics of Business* **18**:1, 5-16. [[CrossRef](#)]
21. Gregory Leonard, G. Steven Olley. 2011. What Can Be Learned About the Competitive Effects of Mergers from “Natural Experiments”? *International Journal of the Economics of Business* **18**:1, 103-107. [[CrossRef](#)]
22. Allan Dafoe. 2011. Statistical Critiques of the Democratic Peace: Caveat Emptor. *American Journal of Political Science* no-no. [[CrossRef](#)]
23. Daniel E. Ho, Donald B. Rubin. 2010. Credible Causal Inference for Empirical Legal Studies. *Annual Review of Law and Social Science* **7**:1, 110301100413081. [[CrossRef](#)]
24. Robert J. Sampson. 2010. Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology. *Journal of Quantitative Criminology* **26**:4, 489-500. [[CrossRef](#)]
25. James Fenske. 2010. THE CAUSAL HISTORY OF AFRICA: A RESPONSE TO HOPKINS. *Economic History of Developing Regions* **25**:2, 177-212. [[CrossRef](#)]
26. Changhui Kang. 2010. Confronting the shadow education system: what government policies for what private tutoring?. *Education Economics* **18**:3, 373-375. [[CrossRef](#)]
27. Peter Arcidiacono, Paul B. Ellickson. 2010. Practical Methods for Estimation of Dynamic Discrete Choice Models. *Annual Review of Economics* **3**:1, 110301095653089. [[CrossRef](#)]
28. Angus Deaton. 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* **48**:2, 424-455. [[Abstract](#)] [[View PDF article](#)] [[PDF with links](#)]
29. James J. Heckman. 2010. Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature* **48**:2, 356-398. [[Abstract](#)] [[View PDF article](#)] [[PDF with links](#)]