ROB BONTA
Attorney General of California
MARK R. BECKINGTON
Supervising Deputy Attorney General
ROBERT L. MEYERHOFF
Deputy Attorney General
State Bar No. 298196
300 South Spring Street, Suite 1702
Los Angeles, CA 90013-1230
Telephone: (213) 269-6177
Fax: (916) 731-2144
E-mail: Robert.Meyerhoff@doj.ca.gov
Attorneys for Defendant Rob Bonta in his official capacity as Attorney General of the State of California

IN THE UNITED STATES DISTRICT COURT

FOR THE SOUTHERN DISTRICT OF CALIFORNIA

CIVIL DIVISION

VIRGINIA DUNCAN, RICHARD LEWIS, PATRICK LOVETTE, DAVID MARGUGLIO, CHRISTOPHER WADDELL, and CALIFORNIA RIFLE & PISTOL ASSOCIATION, INC., a California corporation,

Plaintiffs,

v.

ROB BONTA, in his official capacity as Attorney General of the State of California; and DOES 1-10,

Defendants.

Case No. 3:17-cv-01017-BEN-JLB

COMPENDIUM OF WORKS CITED IN SUPPLEMENTAL DECLARATION OF LUCY P. ALLEN

Courtroom: 5A
Judge: Hon. Roger T. Benitez
Action Filed: May 17, 2017

INDEX

DATA VISUALIZATION Firearms

Defensive Gun Uses in the U.S.

Updated October 7, 2022

2019 2020 2021 2022 PAST 90 DAYS INCIDENTS: 2,714

RESET ZOOM

DATE	LOCATION	CONTEXT	CONCEALED-CARRY PERMIT?
–	–	–	–

MULTIPLE ASSAILANTS?	SHOTS FIRED?	FIREARM TYPE	DETAILS
–	–	–	–

NOTES: This page was originally published on April 13, 2020. Locations are approximate.

SOURCE: Heritage Foundation research.

DESIGN AND DEVELOPMENT: Graphic produced by John W. Fleming and Jay Simon. Data compiled by Amy Swearer.

All of the law-abiding citizens featured in this database successfully defended their liberties, lives, or livelihoods with the lawful use of a firearm. These cases are not based on hearsay, but on verifiable reports found through public sources.

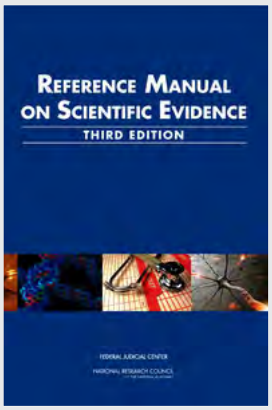
According to the Centers for Disease Control and Prevention, almost every major study on defensive gun use has found that Americans use their firearms defensively between 500,000 and 3 million times each year. There's good reason to believe that most defensive gun uses are never reported to law enforcement, much less picked up by local or national media outlets.

This database, therefore, is not intended to be comprehensive. Instead, it highlights just a fraction of the incredible number of times Americans relied on the Second Amendment—not the government getting there in time—to protect their inalienable rights. Despite the limitations on data, these confirmed cases of defensive gun use help prove that the "good guy with a gun" is not a myth, but an integral part of American society.

Click on a circle to see more information in the map above. Scroll to zoom and click + drag to pan.

Related

- The Latest Crime News Provides Evidence in Favor of Armed Citizens
- New Cases of Armed Citizens Stopping Criminals in February
- Latest News of Self-Defense With Firearms Contradicts Gun Control
- The Second Amendment Saved These Gun Owners' Lives in April
- Gun Rights Made All the Difference for These Intended Victims
- Why These Defensive Uses of Firearms Should Disarm Second Amendment Skeptics
- Guns Saved These Americans From Assault and Robbery in July
- Lawful Gun Owners Defended Their Lives and Livelihoods in August
- These Gun Owners Were Able to Confront Criminals in September
- These Law-Abiding People Used Guns to Defend Themselves in October



Reference Manual on Scientific Evidence: Third Edition (2011)

DETAILS

1034 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-21421-6 | DOI 10.17226/13163

CONTRIBUTORS

Committee on the Development of the Third Edition of the Reference Manual on Scientific Evidence; Committee on Science, Technology, and Law; Policy and Global Affairs; Federal Judicial Center; National Research Council

SUGGESTED CITATION

National Research Council 2011. *Reference Manual on Scientific Evidence: Third Edition*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13163.

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

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 *Vuyanich v. Republic National Bank*, 505 F. Supp. 224, 319 (N.D. Tex. 1980), *vacated*, 723 F.2d 1195 (5th Cir. 1984), defendant's statistical expert criticized the plaintiffs' statistical model for an implicit, but restrictive, assumption about male and female salaries. The district court trying the case accepted the model because the plaintiffs' expert had a "very strong guess" about the assumption, and her expertise included labor economics as well as statistics. *Id.* It is doubtful, however, that economic knowledge sheds much light on the assumption, and it would have been simple to perform a less restrictive analysis.

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. Rubinfeld, 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—nonresponse 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

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

The entries in Table 4 add up to the entries in Table 3. 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

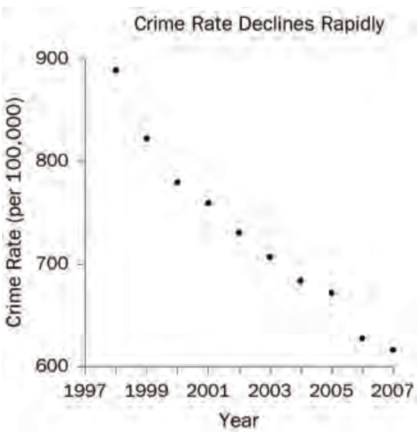
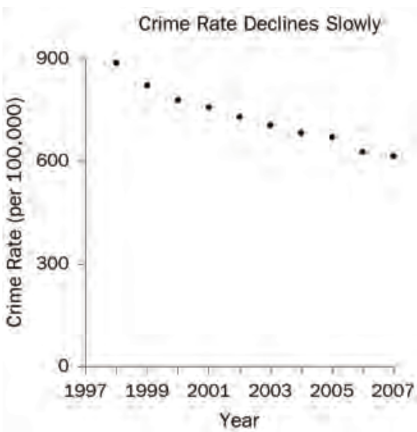


Figure 2



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

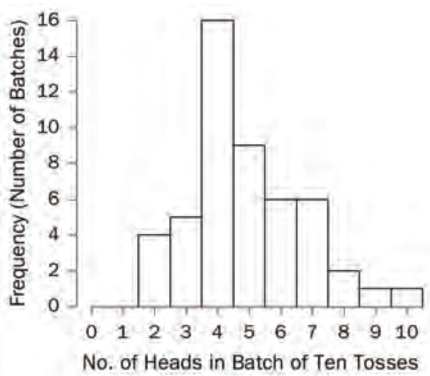
65. Florida Statistical Analysis Center, Florida Department of Law Enforcement, Florida's Crime Rate at a Glance, *available at* http://www.fdle.state.fl.us/FSAC/Crime_Trends/domestic_violence/index.asp. The data are from the Florida Uniform Crime Report statistics on crimes ranging from simple stalking 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 × 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 Fraasen 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 average[88]:

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

from \$40,000,000 – \$2,000,000 = \$38,000,000
to \$40,000,000 + \$2,000,000 = \$42,000,000.

The 95% confidence interval runs

from \$40,000,000 – (2 × \$2,000,000) = \$36,000,000
to \$40,000,000 + (2 × \$2,000,000) = \$44,000,000.

The 99.7% confidence interval runs

from \$40,000,000 – (3 × \$2,000,000) = \$34,000,000
to \$40,000,000 + (3 × \$2,000,000) = \$46,000,000.

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

estimate – 1 SE to estimate + 1 SE.

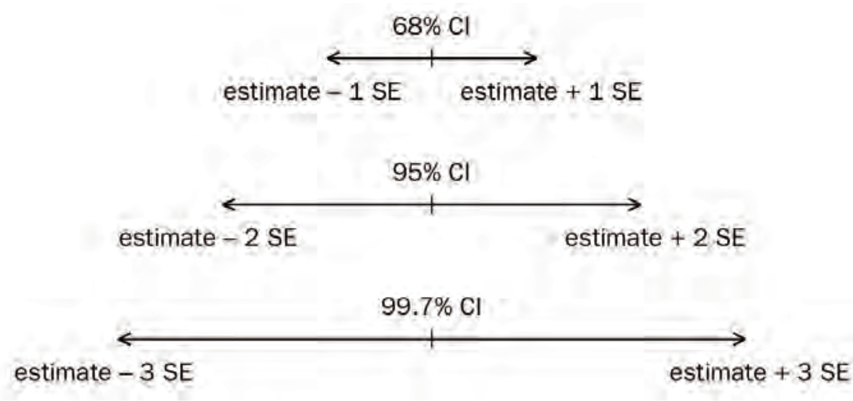
A 95% confidence interval is the range

estimate – 2 SE to estimate + 2 SE.

The 99.7% confidence interval is the range

estimate – 3 SE to estimate + 3 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 α = 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).

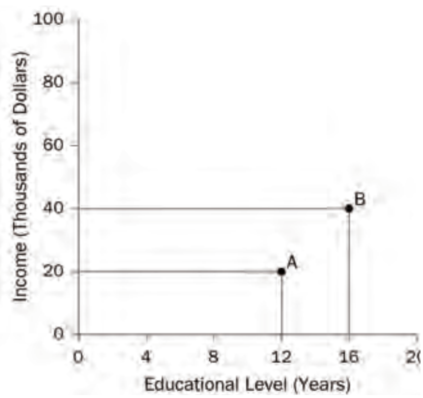
V. Correlation and Regression

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.

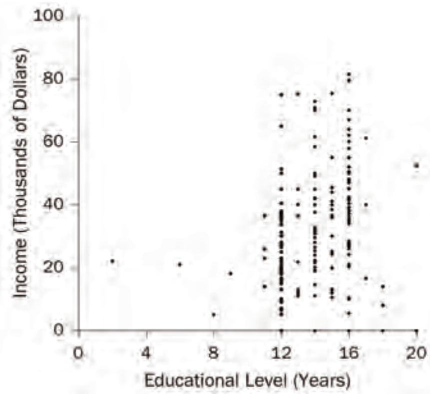


124. The focus is on simple linear regression. *See also* Rubinfeld, *supra* note 21, and the Appendix, *infra,* and Section II, *supra,* for further discussion of these ideas with an emphasis on econometrics.

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.



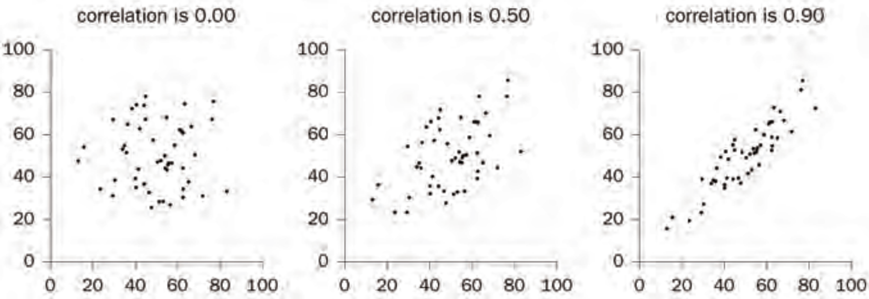
B. Correlation Coefficients

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.

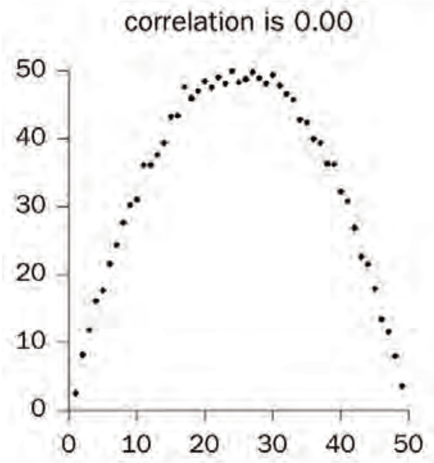
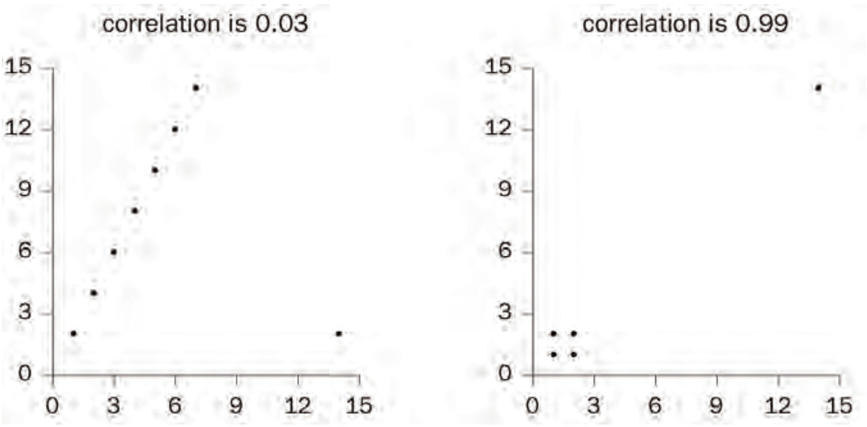


Figure 9. The correlation coefficient can be distorted by outliers.



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.

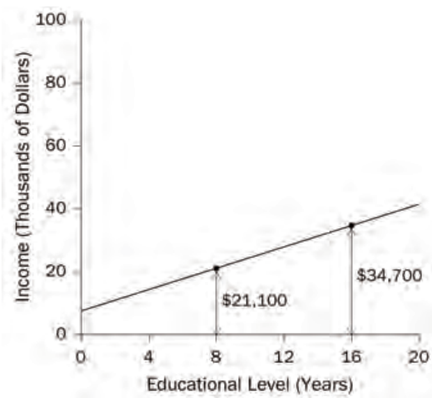
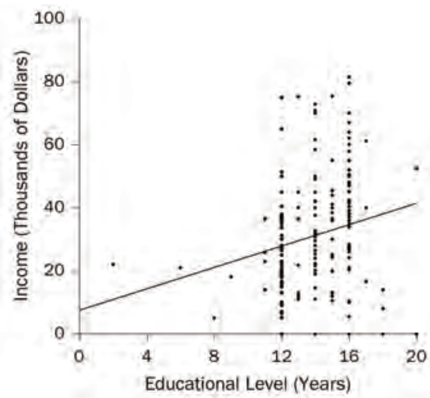


Figure 11 combines the data in Figures 5 and 10: it shows the scatter diagram for income and education, with the regression line superimposed. The line shows the average trend of income as education increases. Thus, the regression line indicates the extent to which a change in one variable (income) is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

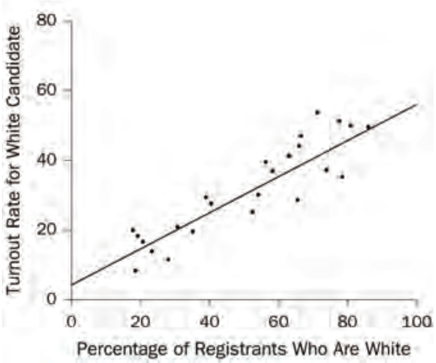
the Voting Rights Act. *E.g.*, Johnson v. De Grandy, 512 U.S. 997, 1011 (1994) ("*Gingles* . . . clearly declined to hold [these factors] sufficient in combination, either in the sense that a court's examination of relevant circumstances was complete once the three factors were found to exist, or in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

$$\text{length} = a + b \times \text{weight} + \varepsilon. \tag{1}$$

The error term, denoted by the Greek letter epsilon ε, is needed because measured length will not be exactly equal to $a + b \times$ weight. If nothing else, measurement error must be reckoned with. 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 $a + \varepsilon$. *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 $\hat{a} + \hat{b} \times$ weight:

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

Of course, no one really imagines there to be a box of tickets hidden in the spring. However, the variability of physical measurements (under many but by no means all circumstances) does seem to be remarkably like the variability in draws from a box.[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

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

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(HH | 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 | 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 "~C" to denote the event that C does not occur. Thus C and D are independent if P(D | C) = P(D | ~C). If C and D are independent, then the probability that both occur is equal to the product of the probabilities:

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

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:

$$P(C \text{ and } D) = P(C) \times P(D \mid C). \tag{A2}$$

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 hypotheses[147] that govern the probability of an event A, a Bayesian can use the multiplication rule (A2) to find that

$$P(A \text{ and } H_0) = P(A \mid H_0)P(H_0) \tag{A3}$$

and

$$P(A \text{ and } H_1) = P(A \mid H_1)P(H_1). \tag{A4}$$

Moreover,

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

The multiplication rule (A2) also shows that

$$P(H_1 \mid A) = \frac{P(A \text{ and } H_1)}{P(A)}. \tag{A6}$$

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:

$$P(H_1 \mid A) = \frac{P(A \mid H_1)P(H_1)}{P(A \mid H_0)P(H_0) + P(A \mid H_1)P(H_1)}. \tag{A7}$$

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 $P(H_0 \mid A)$ 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

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

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 formula[150]

$$\frac{n!}{j! \times (n-j)!} f^j \left(1-f\right)^{n-j}. \tag{A9}$$

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

$$nf = 350 \times 0.5 = 175. \tag{A10}$$

The standard error is

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

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

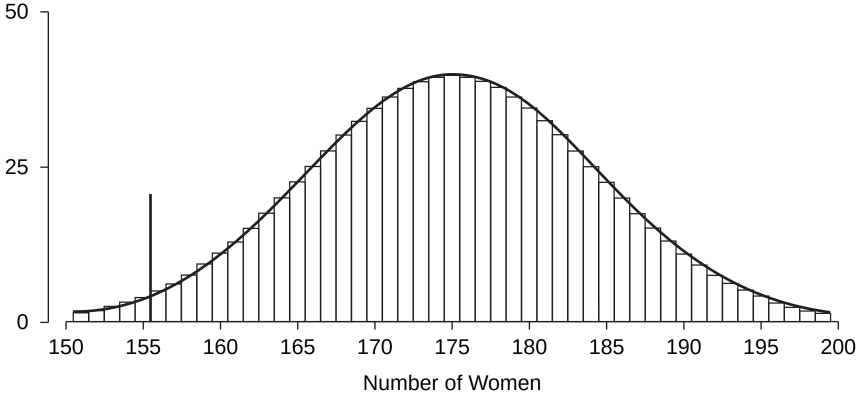
Figure 13 shows the probability histogram for the number of women in the sample.[151] The graph is drawn so that the area between two values is proportional to the chance that the number of women will fall in that range. For example, take the rectangle over 175; its base covers the interval from 174.5 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,

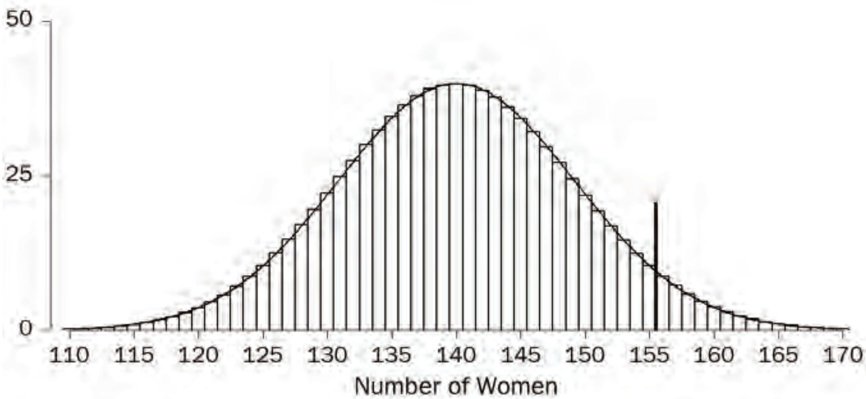
151. Probability histograms are discussed in, e.g., *id.* at 310–13.
152. The central limit theorem is discussed in, e.g., *id.* at 315–27.

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

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

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 × $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... and π = 3.14159... . 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:

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

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

$$\$7100 + \$1300 \times 12 + \$2200 \times 10 + \$700 \times 0$$
$$= \$7100 + \$15{,}600 + \$22{,}000 + \$0 = \$44{,}700. \qquad \text{(A17)}$$

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 number 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... and π = 3.14159... . 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

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

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 kth 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 positive. 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).

Multiple Regression in Legal Proceedings

Franklin M. Fisher *

Multiple regression analysis is a device for making precise and quantitative estimates of the effects of different factors on some variable of interest. It is not a new tool, going back in its origins to Carl Friedrich Gauss, an extremely important mathematician born about 200 years ago. Nevertheless, the practical use of multiple regression has grown very substantially over the past twenty-five years or so. This growth is due partly to the development of modern statistical methods, partly to increasing availability of decent statistical data, and perhaps most of all to the development of the electronic computer. Some of the increasing use of multiple regression and related techniques has occurred in connection with legal proceedings of various kinds, although lawyers and judges have often tended to view such use with general (and occasionally healthy) distrust.

In light of the increasing prominence of multiple regression analysis, it is important for lawyers to understand what it is, how it works, and what it properly can be used for. Perhaps the single most important legal use of multiple regression thus far has been the analyses of the deterrent effects of the death penalty on murder, cited by the Solicitor General in his amicus brief before the Supreme Court in the death penalty cases.[1] The fact that the studies relied on by the Solicitor General were, in my opinion, fatally flawed[2] only adds to the importance of understanding the methodology involved. On a less grand level, multiple regression studies have figured in a number of other legal proceedings, and while the ones with which I am most familiar have been regulatory proceedings, there is no reason why multiple regression should not be used in other litigation as well.[3]

This Article first explains, on a basic level, the concept of multiple regression analysis, its basic properties, and the fundamental assumptions upon which its validity rests.[4] I will also discuss methods of measuring the

---

* Professor of Economics, Massachusetts Institute of Technology. A.B. 1956, M.A. 1957, Ph.D. 1960, Harvard University.

This Article was adapted from a paper delivered before the Association of the Bar of the City of New York (Special Committee on Empirical Data in Legal Decision Making) in May 1979. I am indebted to Michael O. Finkelstein for helpful criticism but retain the usual responsibility for error.

1. Fowler v. North Carolina, 428 U.S. 904 (1976); Woodson v. North Carolina, 428 U.S. 280 (1976); Jurek v. Texas, 428 U.S. 262 (1976); Proffitt v. Florida, 428 U.S. 242 (1976); Gregg v. Georgia, 428 U.S. 153 (1976). See generally Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates (A. Blumstein, J. Cohen & D. Nagin eds. 1978) [hereinafter cited as Deterrence and Incapacitation], which contains, among other things, some devastating discussion of the studies involved. (In particular, see the paper by Klein, Forst & Filatov, The Deterrent Effect of Capital Punishment: An Assessment of the Estimates, id. at 336.)

2. See text accompanying notes 38-46 infra.

3. For an excellent discussion of proceedings using multiple regression studies, see Finkelstein, Regression Models in Administrative Proceedings, 86 Harv. L. Rev. 1442 (1973).

4. I have been responsible for several multiple regression studies used in legal proceedings and, because I know them best, it is those studies on which I shall draw for examples for

accuracy and reliability of estimates generated by multiple regression. The second part of the Article explores in greater depth the proper use of multiple regression in legal proceedings by focusing on three areas in which multiple regression studies might play a role—the examination of wage discrimination, the determination of antitrust damages, and the evaluation of punishment as a deterrent to crime.

I. Multiple Regression Analysis

A. *Uses of Multiple Regression*

The two primary uses of multiple regression analysis are best illustrated through an examination of actual situations in which multiple regression studies were employed. Consider the following two cases:

> 1. For many years after the disappearance of coal-burning locomotives, there was a perennial labor dispute concerning the preservation of the jobs of railroad firemen. Whatever the merits of that dispute (ultimately resolved, I believe, through negotiation), one of the issues in it concerned the question of whether the presence of a fireman on trains contributed to railroad safety. A study of that issue, using multiple regression, was presented in testimony before a Presidential emergency board in 1970.[5]
>
> 2. Cable television systems (CATVs) have been the subject of repeated rulemaking proceedings by the Federal Communications Commission. Among the issues involved in such proceedings is the effect of the entry and activity of CATVs on the profits and growth of broadcast television stations. This issue involves such questions as the influence of CATVs on the viewing audience reached by particular broadcast stations and the effect of changes in a station's audience on the revenue it receives.[6] In general, as one would expect, cable operators have claimed such effects to be small and broadcast stations have insisted they are large. The problem has been studied repeatedly by multiple regression methods, most recently in a study of the relationship between audience size and revenues authored in part by me and submitted to the FCC in 1978-79.[7]

---

much of this Article, hoping thereby to put some more interest into what otherwise might degenerate into a fairly dry and technical discussion.

5. The study is most conveniently reported in Fisher & Kraft, The Effect of the Removal of the Firemen on Railroad Accidents 1962-1967, 2 Bell J. Econ. & Management Sci. 470 (1971).

6. These are important questions for the FCC since they bear on the extent to which regulation of cable television is needed to foster the growth of new UHF stations or to maintain the profits that subsidize the public service and other programming of local broadcast stations.

7. Charles River Associates, The Audience-Revenue Relationship for Local Television Stations (1978) (FCC Docket No. 21284); Charles River Associates, The Value of Different

In the first case, the issue is whether or not a particular variable (presence or absence of firemen) has *any* effect on some other variable (railroad safety). In the second case (the audience-revenue relationship), there is not much doubt that audience size affects television revenue—viewer attention is what stations sell to advertisers, and all parties are vitally interested in audience statistics; the problem is rather one of measuring the effect. These two uses of multiple regression are what statisticians call "testing hypotheses" on the one hand and "parameter estimation" on the other. In the first type, one wishes to be able to state whether or not something is true. In the second, one is more interested in the precise magnitude of the effects involved. Obviously, the two questions are closely related.

There is a third, but less widespread, use to which multiple regression analysis can be put: to *forecast* the values of some variable. A multiple regression analysis shows how certain independent variables affect a dependent variable. From that analysis, and from a forecast of the values of the independent variables (obtained from some other source), one can generate a forecast of the dependent variable. This type of "unconditional forecast" is not always useful—which is fortunate, since such unconditional forecasts tend to be relatively inaccurate. Far more often what is of interest is a "conditional forecast"—a prediction of what will happen to the dependent variable if another variable is changed or, looking retrospectively, what would have happened to the dependent variable had the value of an independent variable been different.

Consider the two examples already described. The question in the case of the railroad firemen did not really involve predicting the number of railroad accidents. Rather, it involved trying to decide whether the number of those accidents would be significantly greater if the railroad firemen were no longer employed. Similarly, while prediction of television station revenue would be desirable for some regulatory ends, the primary issue in the audience-revenue study was systematic measurement of the effects on revenue of changes in the size and socio-economic characteristics of a station's audience.

The firemen example best brings out the problems involved in making such predictions. By their nature, railroad accidents involve random, chance events. Even the accident *rate* (however measured) is subject to such chance fluctuations. Simply determining whether the presence or absence of firemen makes a significant difference to the railroad accident rate may be easier than predicting the rate itself with great precision. One of the distinctive characteristics of multiple regression analysis is that it is able to provide information about the effects of the variable of interest (in this case the employment of

---

Day Parts in the Audience-Revenue Relationship for Local Television Stations (1979) (FCC Docket No. 21284). The first study of the problem was, I believe, the one that I gave as written testimony to the FCC in 1964. It is most conveniently reported in Fisher, Ferrall, Jr., et al., Community Antenna Television Systems and Local Television Station Audience, 80 Q.J. Econ. 227 (1966). Both studies were done on behalf of the National Association of Broadcasters.

the firemen) on the dependent variable (here, railroad accidents) without necessarily being able to predict the dependent variable itself with great accuracy.

In a way, one might describe multiple regression as a method used to extract a systematic signal from the noise presented by data. There are two primary problems involved in extracting such a signal. First, it is typically the case that the factor whose influence one wishes to test or measure is not the only major factor affecting the dependent variable—for example, the amount of traffic on the railroads has something to do with accidents as well. Second, even if one can somehow account for the effects of the other important systematic factors, there typically remain chance components.

If we could make controlled experiments, it would be relatively easy to quantify the relationship being investigated. A controlled experiment in the audience-revenue case, for example, would vary station audiences and the other variables expected to influence revenue one at a time, holding everything else constant and observing the resulting revenue. Obviously, this is impossible—there is no way we can tinker with station audiences. This means that we must be content with analyzing, as it were, the experiments performed by nature, in which more than one of the variables deemed likely to affect revenue move at the same time.

Moreover, even if we could control station audiences and hold constant the variables that we believed to be important, we would not know enough about the audience-revenue relationship to be sure of holding constant all the variables that actually affect the revenues of an individual station. It may be, for example, that the personality and effectiveness of the stations' sales representatives or the advertising policies or publishing quality of competing newspapers affect revenue. These variables are hard to measure, let alone hold constant.[8]

Inability to perform well-controlled experiments is not uncommon. Indeed, it occurs even when one is making so-called controlled experiments in the natural sciences. The difference there is that one can be fairly sure that the uncontrolled effects that one does *not* know about in detail are extremely small. When dealing with observations from the economic system (or, indeed, from any system in which the experiments are performed by nature rather than by the experimenter), there is likely to be a nontrivial, residual element of unexplained effects on the variable of interest, even after one has taken account of the major systematic effects. Multiple regression is a way of dealing with these difficulties.

B. *How Multiple Regression Works*

1. *An Overall View.* In multiple regression, one first specifies the major variables that are believed to influence the dependent variable. In our

8. Similarly, in the case of the firemen, even if we could experiment with firemen employment, we could not hold railroad traffic constant. Moreover, other variables affecting safety (the ones we call "chance") are never known precisely.

examples, this means specifying the important or systematic influences that may affect station revenue or railroad safety. There inevitably remain minor influences, each one perhaps very small, but creating in combination a non-negligible effect. These minor influences are treated by placing them in what is called a random disturbance term and assuming that their joint effect is not systematically related to the effects of the major variables being investigated—in other words by treating their effects as due to chance.[9] Obviously, it is very desirable to have the random part of the relationship small, particularly relative to the systematic part. Indeed, the size of the random part provides an indication of how correctly one has judged what the systematic part is. Multiple regression thus provides a means not only for extracting the systematic effects from the data but also for assessing how well one has succeeded in doing so in the presence of the remaining random effects.

The relationship between the dependent variable and the independent variable of interest is then estimated by extracting the effects of the other major variables (the systematic part). When this has been done, one has the best available substitute for controlled experimentation. The results of multiple regressions can be read as showing the effects of each variable on the dependent variable, holding the others constant. Moreover, those results allow one to make statements about the probability that the effect described has merely been observed as a result of chance fluctuation.

2. *Estimating Multiple Regressions.* Suppose that the relationship to be examined is to include only two variables, the dependent variable (Y) and one independent variable (X). Suppose further (for simplicity of exposition) that it is believed that the relationship between these variables is a straight line.[10] Such a relationship could be expressed mathematically as:

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

or, diagrammatically, as in Figure 1. The problem for the investigator is to discover the values of the parameters, a and b (i.e., the intercept and slope of the line). If the relationship really were exact—if there were no random influences at all—this would be extremely easy to do. One would need only to observe two points with different values of X. Since two points determine a line, it would require only routine arithmetic calculation to find the line they determine.

In real life, however, the relationships to be fitted are not exact. Rather

9. The disturbances (the random or unsystematic part) will then affect the dispersion of the true values of the dependent variable around the values that would be predicted from the systematic part alone.

10. I chose the straight line case as the easiest to understand, but the theory is not so restricted. There is nothing to prevent one or more of the variables in equation (1) from being a square, a logarithm, or the ratio or two other variables. Many (not all) mathematical relations can be cast into the form of equation (1) by transforming or redefining the variables. Furthermore, most nonlinear relationships can be at least approximated by straight lines.

there are random influences on the dependent variable, as described above. Hence, the correct relationship is not equation (1) but rather:

$$(2) \quad Y = a + bX + u$$

where u represents the random influences. Different values of u will produce different values of Y which will be either above or below the line; indeed, they will produce a scatter of points such as that shown in Figure 1. The task for the investigator is to cut through the noise generated by these random influences and extract the "signal," namely, the line around which the points

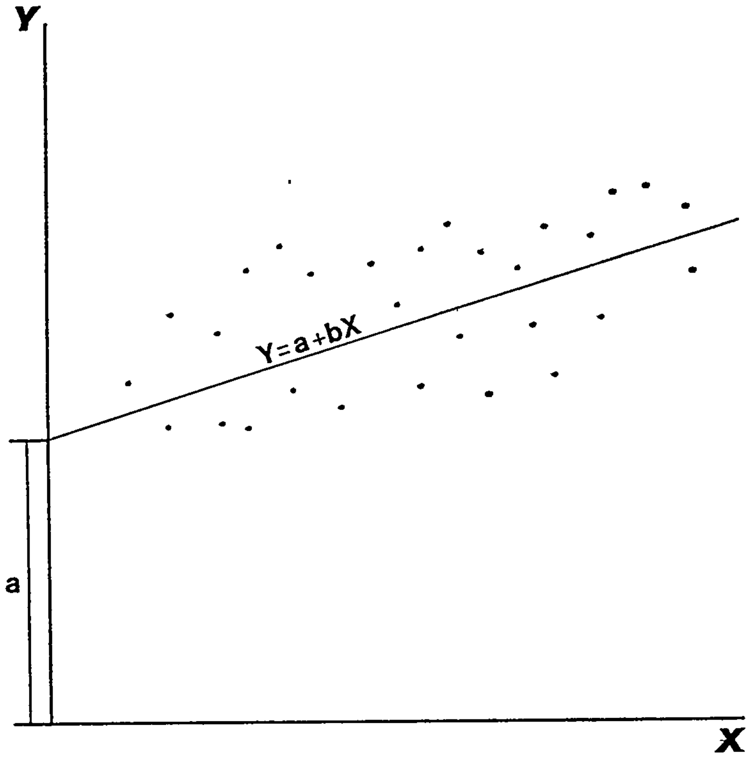


Figure 1

are scattered. This is done by picking the line that best fits the scatter of points in the sense that the sum of the squared deviations between predicted and actual Y values is minimized.[11] This is called "least squares regression." (The adjective "multiple" is used when there is more than one X.)

---

11. Using the sum of squared deviations gives equal weight to positive and negative deviations. Further, in a multi-dimensional diagram (not drawn) it can be shown that there is a sense in which minimizing the sum of squared deviations amounts to minimizing the

3. *Assumptions of Least Squares Regression.* In practice, least squares regression is not done diagrammatically but numerically (generally by computer), resulting in numerical estimates of a and b. The relation that such estimates are likely to have to the true values of a and b depends on the assumptions one is willing to make about the random disturbance term, u, and its relationship with the independent variable, X, whose effect on Y (represented by b) is to be measured.

There are essentially three major assumptions involved: (a) that the effects of the random disturbance term are independent of the effects of the independent variable; (b) that the values of the random term for different observations are not systematically related and that the average squared size of the random effect has no systematic tendency to change over observations; and (c) that the sum of random effects embodied in the disturbance term is distributed normally, in the "bell curve" generally characteristic of the distribution of the sum of independent random effects.[12]

The validity of these assumptions bears on the effectiveness and reliability of least squares analysis. Various properties of multiple regression depend on the accuracy of the assumptions, different properties involving different assumptions. Moreover, the dependence is cumulative: if the early assumptions are invalid, the properties associated with the later assumptions are not likely to be present. In situations where the assumptions may fail, the use of multiple regression analysis is likely to be inappropriate.[13]

a. *Independence of the Disturbance Term.* The fundamental assumption of least squares regression is that the uncontrolled effects of the random disturbance (u) are in an appropriate technical sense independent of the controlled effects of the independent variable (X). (Alternatively, this can be expressed as the assumption that the disturbance term has a zero mean whatever the value of X. In repeated samples, the disturbance term for any given X is neither positive nor negative on the average.) If this were not so, then attempting to determine the effects of X on Y could not be done simply by observing different X's and trying to average out the effects of u. In such a case, movements in X would be systematically associated with movements in u and, without a great deal of care, the estimates of b would include not merely the direct effects of X on Y, but also the associated effects of movements in the disturbance term, u.

When is such an assumption likely to fail? The simplest case to understand occurs when some large and systematic factor, other than X, has been

---

distance between the point which represents the actual values of the dependent variable and the point which represents the values one would predict from the regression. Average squared values are the standard statistical measure of dispersion.

12. The word "normal" here is a term of art referring to the shape of the distribution. The name indicates that the distribution involved is characteristic of many random variables. Most important, if a random variable is composed of the sum of other random variables acting independently, that sum tends to be distributed normally. This makes the assumption of normality the obvious one unless there is a compelling reason to depart from it.

13. As a general rule, there are methods of testing for and dealing with the failure of such assumptions, but they involve the more advanced tools of econometrics rather than least squares regression.

left out of the analysis; this is called misspecification. In the revenue-audience study, for example, it turns out that average household income as well as size of audience affects television station revenue. Suppose, however, that we had not thought of this but had simply tried to estimate the effect of audience size on revenue. (Here, revenue would be Y and audience size would be X.) In effect, this would mean that we were placing household income in the disturbance term. Yet, if household income across television markets is positively associated with audience size, then part of what we would attribute to larger audience size would in fact be attributable to higher income. In other words, we would have failed to control for income levels and the lack of such control would matter.

Obviously, the assumption that one has controlled for all the important influences is basic to any attempt to measure those influences correctly. There are, however, other ways in which the assumption of independence between random disturbance and included factors can be violated. In general, this will happen when there exist relations between the dependent and independent variables in addition to the relation being estimated. I shall discuss specific examples of such cases in part II.

If the assumption of independence between u and X is warranted, then least squares estimates of the parameters (a and b) will have some desirable properties. First, the estimates will be *unbiased*—they will be correct on the average. This means that if one did the calculations for a sample of a particular size, and were then to repeat the procedure on numerous samples of the same size, each time obtaining different estimates for a and b, the average of the estimates so obtained would be the true values of a and b. To put it a little differently, least squares estimates have no tendency to err systematically on either the high side or the low side.

Further, if the assumption of independence is correct, least squares estimates will be *consistent*. The property of consistency means that, as the sample size increases, the probability of obtaining least squares estimates that differ from the true values by more than any given amount goes to zero. Thus, as more data become available it will become easier to extract the true values of a and b from the noise presented by the random part.

b. *Behavior of the Disturbance Term.* Consistency is the minimal property that one wants an estimator to have. But there are many consistent estimators and, in some situations, even many unbiased ones. Moreover, unbiasedness assures only that the estimator is right on the average; it does not indicate how far off it is likely to be on any given sample. Similarly, consistency guarantees only that one will get close to the true values of the parameters if one knows enough; it cannot determine how much one needs to know to get close. It is clearly desirable to have measures of reliability—that is, measures of how far off one can generally expect estimates to be. Moreover, within the class of unbiased or consistent estimators, it is obviously desirable to choose the one likely to be most reliable.

With an additional assumption, least squares regression turns out to be such an estimator and will itself generate estimates of its reliability. This assumption concerns the nature of the random disturbance term (u), rather than an assumption concerning its relation with X. The assumption can be divided into two parts.

First, it is assumed that if one had information about the value of u for some observations, one would not thereby gain any information about its value for other observations. For example, if the observations are on the variables over time, an unusually high and positive value for u should not be followed by a tendency for u to be high the next year. Rather, successive values of u should be independent of each other. One can see why this is likely to matter. Least squares regression is a generalized form of averaging. Averaging is an excellent way to take care of random noise, provided that one is averaging over independent events. If the random disturbances from different observations are not mutually independent, however, then the averaging involved in least squares regression will not defuse the random effects. In such a case one could do better by expressly assuming that a high disturbance term in one period indicates something about the value of the disturbance term in the following period, and then using this information to attempt to factor the disturbance out of the equation.

Second, it is assumed that there is no systematic tendency for the random disturbance (u) to be either big or small.[14] To put it differently, one assumes that the chances of a large random effect versus a small one are the same for all observations.[15] Again, one can see why this will matter. If some observations tended to have larger random effects than others, then the observations with large random effects would contain less reliable information than would the observations with small random effects. In any averaging procedure, one would want to give more weight to the latter. Since least squares regression will treat all observations equally, it will not take this into account.[16]

These assumptions will be violated if, when dealing with a series of observations over time, the disturbance term includes the effects of variables that behave systematically over time. Certainly, this is a serious possibility

---

14. We have already assumed that the random disturbance term has no systematic tendency to be high or low—that is, that it has a mean, or expected value, of 0 for all values of X. ("Expected value" is to be thought of as the population mean. Roughly speaking it is the average value one expects to obtain if one takes a large enough sample.) That assumption involves the algebraic sign of the random disturbance term. The present assumption, on the other hand, has to do with the absolute magnitude of the disturbance term, regardless of sign. Put more precisely, the dispersion of a random variable is measured by the average or expected value of the squared deviation from its mean. This is called the "variance." Its square root is called the "standard deviation." The assumption previously made in the text was that the mean of the random disturbance term is not systematically related to X. The assumption now being made is that the variance or standard deviation of the disturbance term is not so related and is, in fact, the same for all observations.

15. Technically, this is the property that the variance of the disturbance term should be the same for all observations.

16. There are ways of taking this failure of assumption into account: not surprisingly, the technique involved is called "weighted least squares," a variety of "generalized least squares."

in econometric models. Similarly, if the observations are of individual entities, such as firms, it may very well be that the effects of particular uncontrolled events (such as political events) will be larger for large firms than for small ones. In such a case, the second part of the assumption would be violated. As with all the assumptions of least squares regression, however, one would want to be sure that the violations are really important before abandoning regression analysis. In the cases posited above, small departures from the assumptions would have small effects. Furthermore, the properties of least squares associated with the assumptions are so strong as to make least squares regression superior to the alternative estimators that would result from trying to cure such small departures.

Given the validity of the assumptions under discussion, least squares estimates will be *efficient*. This means that, within a wide class of unbiased and consistent estimators, least squares estimates will have the smallest variation. Thus, if one could take repeated samples, the variation of the least squares estimates around the true values of a and b would be less than the variation of other unbiased and consistent estimators; in short, the least squares estimates will be more reliable.

c. *Normality of Distribution.* The last assumption of least squares imposes greater restrictions on the random disturbance, u, than the ones already discussed. The assumption is that u, for all values of X, follows the normal distribution (bell curve),[17] with a mean of zero, as already assumed. This, however, is not as restrictive as it may appear. As a general proposition of statistics, the normal distribution is characteristic of large averages of independent random effects. To the extent that the error term is made up of the sum of small random effects, that sum will tend to be distributed normally.[18]

The normality assumption, in addition to bolstering least squares' property of efficiency, implies the ability to make precise probability statements concerning how far off the least squares estimates are likely to be.[19]

4. *Multiple Independent Variables.* In practice, one does not usually work with relationships involving only two variables, but rather with relationships in which a dependent variable is influenced by many independent ones (railroad traffic as well as firemen employment; audience income as well as audience size). Denoting the independent variables as $X_1, X_2, \ldots, X_k$, the relationship to be estimated (assuming linearity) [20] can be expressed as:

(3) $Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + u$

17. See note 12 supra.

18. See note 12 supra. The "normal" distribution is completely characterized by its mean and variance. It is hard to construct practical examples in which one would be inclined to question normality without also questioning the other assumptions about the random disturbance term. Hence, while there are tests for departure from normality, they are hardly ever used.

19. See text accompanying notes 24-28 infra.

20. Again, I have chosen a linear form here. Least squares theory runs mostly in terms of such forms, but this is not as restrictive as it might appear, since many nonlinear forms

If there is only one independent variable, this is the case already considered, the case of a straight line. When there are two independent variables, one is fitting a plane to a scatter of points in space. When there are more than two independent variables, one is fitting a hyperplane (the generalization of a plane to more than three dimensions), but the principles are still the same, although the visualization is no longer immediate. Least squares still retains all the properties listed for the simple case above.

Least squares regression takes advantage of the fact that the independent variables seldom move in perfect step together but rather move (as the name suggests) independently. By determining how the dependent variable changes when the independent variables move in a variety of different ways, the effect of each of the independent variables is extracted.

This kind of systematic extraction of the effects of each variable is important. Examination of raw data leads to facile, and sometimes erroneous, conclusions. Over time, for example, removal of firemen and increased numbers of accidents both occurred. That these events were causally connected cannot be concluded if both are also associated with increases in a third variable (railroad traffic) that plausibly affects railroad accidents. Only by systematically using the fact that railroad traffic, while associated with fireman employment in the data, is not perfectly so associated, can one find out about the independent effect of the firemen. Not controlling for railroad traffic would place it in the disturbance term of equation (3) and violate the fundamental assumption of least squares that disturbance terms and independent variables are independent.

---

can be cast into a linear form similar to equation (3) by appropriate transformations of the variables.

The basic assumption involved in linearity is that the effect of each independent variable on the dependent variable is independent of the level of the other independent variables. Thus, in the firemen example, linearity would imply that the effect of the presence of firemen on the number of railroad accidents was the same at high levels of traffic as at low levels. It would also imply that the effect was the same regardless of whether there were other crew members substituting for the firemen. Obviously, these are not assumptions on which one necessarily wants to rely.

Fortunately, it is not necessary to rely on them. If one thought, for example, that two of the variables—say X_1 and X_2—interacted, then one could define a new variable X_3 as the product of X_1 and X_2. Least squares regression would then proceed as if X_3 were simply a different variable, but its coefficient would tell you something about the importance of such interaction.

To take a different example, it is often not very plausible to suppose (as linearity does) that the effect on the dependent variable of changing an independent variable by one unit should be the same in absolute terms for all levels of the independent variable. It is frequently more plausible to assume that a one percent change in an independent variable has a constant percentage effect on the dependent variable. Such cases can be treated within the framework of linearity by entering into equation (3) not the original variables themselves, but rather their logarithms. This is frequently done and has the advantage, as well, of assuming that the effect of the random error on different observations is likely to be of the same size in percentage rather than absolute terms, a matter that came up above in the discussion of one of the least squares assumptions. See text accompanying notes 14-16 supra.

In general, the choice of the form in which to enter the variables or, more generally, the form of the relationship requires serious thinking about the way in which the relationship being estimated is likely to work. As with deciding which variable to include in the relationship in the first place, this must be done in large part by thinking about the problem rather than by hoping that the data will provide the answer. In any case, relations such as equation (3) are substantially more general than might appear at first sight.

As this description suggests, it is very important that variables do in fact move somewhat independently. Suppose, for example, that in the revenue-audience study one wished to investigate the separate effects on station revenue of audiences close to a station (X_1) and audiences located farther away (X_2). Suppose, as is not the case, that whenever the nearby audience increased by ten percent, as one went from station to station, the far-away audience also increased by ten percent. Then, although one would be able to determine the influence on revenue of the *total* audience, one could not find out the separate effects on revenue of the two subdivisions of that audience. No "experiment performed by nature" would have separated those effects in any way.

Such an extreme situation is not generally encountered in practice; rather what is encountered is something close to it. Suppose that every time the nearby audience went up by ten percent, the far-away audience went up by amounts that varied only slightly up or down from ten percent. In that case, it would be possible to estimate the separate effects generated by each subdivision of audience size, but one would be very uncertain about the estimate. Nature would not be performing experiments calculated to separate those effects with any high degree of accuracy. Such a circumstance is called *multicollinearity*—so called because it involves an additional linear relationship between the variables on the right hand side of the equation.

Obviously, the less multicollinearity is present, the better able one will be to separate out the effects of interest. Unless multicollinearity is perfect, however, multiple regression will be able to separate the effects to some extent and, again, will do so more precisely than any other method, producing estimates with the properties discussed above as well as measures of the reliability of these estimates. The effects of multicollinearity will show up in such reliability measures (standard errors), as discussed below.[21]

5. *Erroneous Inclusion or Exclusion of Variables.* The discussion thus far has presumed that the true systematic relationship is the one being estimated. To put it another way, we have already seen in discussing unbiasedness that multiple regression retains the desirable properties associated with it only if one has in fact included all the variables likely to have a large effect on the dependent variable and can safely assume that the remaining effects are not correlated with the independent variables included. In the audience-revenue study it was thus necessary to control for household income and not place it in the disturbance term. It is therefore important to

21. See text accompanying notes 24-28 infra.
Note that the problem here occurs when two of the independent variables move together in an approximately linear fashion. If they move together nonlinearly, there will not be so severe a problem. If what is involved is not another relation between two or more of the independent variables but another relation between the *dependent* variable and an independent variable, then the basic assumption of least squares will be violated and we will have a situation involving simultaneous equations as discussed below. See text accompanying notes 34, 35 & 43-45 infra.

proceed by including at some stage all the variables that one might think could possibly have a significant effect on the dependent variable. In general, one does this by first examining those variables that one thinks are actually important and then asking what happens when additional variables are included.

Note that this must be done by specifying *in advance* what variables are thought to be important. To proceed by first looking at the data and then including those factors that appear correlated with the dependent variable is a recipe for spurious results. It leads to a situation where no true test of the estimated relationship can be made. In addition, it is likely to leave out variables that truly belong in and thus lead to invalid as well as untested results. The measurement provided by least squares regression is a way of making theoretical assumptions precise or of testing them; it is not a substitute for thought.

I mention this emphatically because a number of packaged computer programs that are sometimes used involve what is known as "step-wise regression." Such programs build up multiple regressions in ways similar to the following. First, the program finds the independent variable in the list most correlated with the dependent variable and does a regression involving it. It then looks at the sample deviations from the regression (the differences between actual and predicted values) and asks whether those deviations are correlated with another independent variable. If so, it puts in the variable most correlated with those errors and so forth. This is not recommended. In the first place, even if none of the independent variables have anything to do with the dependent variable, proceeding in this fashion is very likely to produce the appearance of a high correlation in a particular sample. Second, variables that in fact belong in the relationship but that are correlated with the independent variables uscd early in the procedure tend never to get in. In general, such computer programs suffer from the same problems as attempts to look by eye at bilateral relationships that in fact involve the influence of many variables: they are likely to attribute the effects of the omitted variables to the included ones and result in biased estimates.

The opposite of building regressions up one variable at a time is to put many variables in and then see whether some of them should come out. This is a somewhat better method. Whereas there is a major effect from excluding a variable whose true coefficient is far from zero, the effect of erroneously including a variable whose true coefficient is zero is of very little consequence. Such a variable can be thought of as actually present in the relationship, with the zero coefficient simply indicating that the variable has little or no effect. The multiple regression technique then estimates that coefficient along with the other true coefficients; thus, the regression technique must extract one more parameter from the same number of observations. This is equivalent to having one less observation with which to extract the nonzero parameters.[22] If the sample size is large (there are more than five hundred television stations in the United States, for example), there will be only a very small effect on the estimates of the remaining coefficients and on the prediction of the dependent variable (unless the inclusion of the extra variable adds to multicollinearity). The reliability measures and the measures of "goodness-of-fit"[23] will take full account of the slight reduction in information involved. Where possible, therefore, it may be best to start with an overly complex model and build down.

Nevertheless, it is important to realize that such building down cannot be done without an antecedent theory; the use of computer programs that do "backwards step-wise regression" is not recommended. Without some theory about which variables are likely to matter, throwing a great number of variables into the hopper is likely to lead to spurious results. If one tries enough combinations of variables, then, in a particular sample, one will tend to get some relationship that appears to fit well. Therefore, a properly done study begins with a decent theoretical idea of what variables are likely to be important. It then can proceed to test well-defined hypotheses about additional variables. But a study that casts about for a good-looking relationship by trying all sorts of possibilities is very likely to come up with relationships where none exist.

This leads directly to two comments relevant to lawyers. First, when having a study done by an expert, one should not be too insistent about covering every possibility at once. Rather, one should make sure that the expert proceeds by estimating a reasonable model including the major variables and then goes on to test other possibilities. If one insists that all possible variables are likely to be of equal importance, one is likely to end up with a rather doubtful result.

Second, when faced with an opposing expert who has done a regression study, one should find out how the expert decided on the variables he included and how many different combinations of variables and models he tried before settling on the one that is being presented. If the basic model was tried relatively early and variations were then tried simply to see if anything else seemed to matter, the study may be sound. If, however, the basic model being presented is the end result of vast amounts of computer work, particularly mindless and mechanical computer work, then one may have a legitimate point of attack.

---

22. This is because, for the purpose of assessing reliability of the regression estimate, what matters is the number of "degrees of freedom"—the excess of the number of observations over the number of parameters to be estimated. The following conveys some idea of what is involved. One can always fit a line to two observations, but there are no degrees of freedom and no way of assessing the reliability of the result. If one has a third observation, then one cannot always fit a line exactly but some notion of reliability can be gained from observing how close one comes in fact. Add another variable with a coefficient to be estimated, however, and one is estimating a plane that can be fitted precisely to three observations. Thus, the addition of another coefficient to be estimated has the same effect as the removal of one observation.

23. See text accompanying notes 24-28 infra.

C. *Measuring "Goodness-of-Fit"*

As I have already mentioned several times, least squares regression not only estimates the effects of the variables involved in the model but also measures the certainty or accuracy of such estimates. In addition, it provides overall measures of how well the model fits the data as a whole. There are several different measures involved and because they each measure different things, it is important to be clear on the differences among them.

1. *Standard Errors of Coefficients and t-Statistics.* Associated with the estimated value of each regression coefficient (a and b in the above equations) is a figure known as the standard error[24] of that coefficient, which measures the coefficient's reliability. In general, the larger the standard error, the less reliable or the less accurate is the estimated value of the coefficient.

Speaking somewhat loosely, in large samples the chances are nineteen out of twenty that the true coefficient lies within approximately two standard errors of the estimated coefficient. The chances are ninety-nine out of one hundred that it lies within approximately two and one half standard errors of the coefficient.[25] (In small samples the bounds tend to be wider.) Thus, for example, if the estimated coefficient is ten with a standard error of two, the chances are nineteen out of twenty that the true coefficient lies between six and fourteen and ninety-nine out of one hundred that it lies between five and fifteen. To say that the chances are nineteen out of twenty that the true coefficient lies between six and fourteen, however, does not mean that the true coefficient is equally likely to be in any part of that range. The single most probable figure is ten. The probability of matching the correct figure decreases as one moves away from ten and, as the slight difference between the six-to-fourteen and five-to-fifteen ranges indicates, that probability decreases very fast as one moves substantially away from the middle estimate.

It is conventional to use the standard error of an estimated coefficient to make a statistical test of the hypothesis that the true coefficient is actually zero—i.e., that the variable to which it corresponds really has no effect on the dependent variable. Essentially, such statements are constructed by asking how likely it is that ranges of the sort just described will include zero. This is done by taking the ratio of the estimated coefficient to its standard error. Such a ratio is called a t-statistic.

24. As explained in note 14 supra, the two basic measures of dispersion of a random variable are its variance, the average square deviation around its mean, and its standard deviation, the square root of the variance. The standard error of a statistic (here, the standard error of a regression coefficient) is, in a rough sense, its expected standard deviation. More precisely, it is the square root of the average squared deviation that one would expect to obtain if one used the same estimating procedure over and over again. It is a convenient measure of the reliability of the statistic with which it is associated since the probability that the statistic differs from the true value by any given amount depends directly on the number of standard errors that the amount represents.

25. This will depend on the normality assumption, discussed at text accompanying notes 17-19 supra.

In large samples, a t-statistic of approximately two means that the chances are less than one in twenty that the true coefficient is actually zero and that we are observing a larger coefficient just by chance. In such a case, the coefficient is said to be "significant at the five percent level." A t-statistic of approximately two and one half means that the chances are only one in one hundred that the true coefficient is zero; in that case, the coefficient is "significant at the one percent level." [26] (In small samples, t-statistics must be larger for a given significance level.) In the numerical example given, the t-statistic would be five (ten divided by two) and the probability that the true coefficient is zero is extremely small. The coefficient would be significant at much better than the one percent level.

Significance levels of five percent and one percent are generally used by statisticians in testing hypotheses. That is, given a significance level of five percent· (or one percent for a stricter researcher) it is safe to assume that the true coefficient is not zero and that therefore the variable being tested has some effect on the dependent variable in question. Some lawyers might question whether the use of such levels imposes too severe a standard. Why rejcct the hypothesis that a certain coefficient is zero only if the probability that the results obtained are due to chance is five percent or less? Where the hypothesis involved is of legal importance (for example, when a nonzero coefficient would indicate the presence of sex discrimination in wages), would it not make more sense to use a "preponderence of the evidence" standard and require only significance at fifty percent?

Such an approach, however, would reflect a flawed understanding of what significance levels really mean. In particular, a significance level of fifty percent would not correspond to a "preponderence of the evidence" standard. The significance level tells us only the probability of obtaining the measured coefficient value *if* the true value is zero; it does *not* give the probability that the coefficient's true value *is* zero, nor does subtracting the significance level from one hundred percent give the probability that the hypothesis is not true. Because, even with a large sample, it is quite possible to obtain results differing from a coefficient's true value, it is conventionally thought that there must be a very high probability that the coefficient is not zero before it can be conclusively claimed that the variable associated with the coefficient has a definite effect on the dependent variable.

This does not mean that only results significant at the five percent

---

26. The examples of significance given in the text are for what is known as a "two-tailed test." For example, the significance level of five percent associated with a t-statistic of about two is the probability of obtaining an estimated coefficient as large as that actually obtained, *either positive or negative*, if the true coefficient is actually zero. In many situations, for example, there is no issue as to whether or not a particular coefficient is positive or negative; rather, the only issue may be whether it is positive or zero. In such a circumstance, the appropriate test is a "one-tailed test" in which five percent would represent the probability of observing some *positive* coefficient if the true value were really zero. The t-statistic required for significance at a given level on a one-tailed test is less than that required for the same level on a two-tailed test. In the case of five percent, for example, what is required is approximately 1.6 rather than 2.

level should be presented or considered. Less significant results may be suggestive, even if not probative, and suggestive evidence is certainly worth something. In multiple regressions, one should never eliminate a variable that there is firm theoretical foundation for including just because its estimated coefficient happens not to be significant in a particular sample.

Nevertheless, the computation of the standard errors of the coefficients or the corresponding t-statistics is a matter of considerable importance. It is routinely done by all professionals, with the five and one percent significance levels generally accepted as the point at which the zero hypothesis is rejected. Failure to report such measures of reliability is a clear signal that the study is suspect.

2. *The Standard Error of Estimate.* Another statistic often reported with the results of least squares regression is the "standard error of estimate" or "standard error of the regression." This is not to be confused with the standard errors of the coefficients. The standard error of estimate is one of the summary measures reflecting the degree to which the estimated regression line or plane fits the data. In terms of the discussion given earlier, it is an estimate of how widely the points are scattered around the line.

More precisely, the standard error of estimate describes the average deviation of the actual values of the dependent variable in the sample from the values that would be predicted from the regression.[27] Thus a standard error of zero would correspond to a perfect fit. The larger the standard error of estimate, the poorer is the fit, in the sense that the more important is the random component not being explained.

The size of the standard error of estimate will depend upon the units in which the variables are measured. For example, if we were to measure the dependent variable in pennies rather than in dollars, the standard error of estimate would also be in pennies rather than in dollars and would therefore be multiplied by one hundred. To judge whether the standard error of estimate is large or small, therefore, one must compare it with something else. One such comparison involves computation of the correlation coefficient, discussed below. Other comparisons involve looking at, for example, the mean value of the dependent variable and determining what percentage of that value the standard error is. In general, the standard error of estimate can be used to make probability statements about how far off forecasts from the model are likely to be. Around the mean of the sample (if the sample is of considerable size), forecasts are likely to be off by more than approximately two standard errors of estimate only once in twenty times.[28]

27. It is in fact not computed as an arithmetic average. Rather, it is the square root of the average squared deviation in the sample (with an adjustment for degrees of freedom, see note 22 supra).

28. Related to the standard error of estimate, but not identical to it, is the standard error of forecast. This is a measure of how reliable forecasts made from the regression equation are likely to be. More precisely, it is the square root of the expected squared difference between the actual value of the dependent variable and its forecast value. The

It is very important, however, to realize that a large standard error of estimate does not tell one anything at all about the accuracy with which the effects of the independent variables are measured. Similarly, a large standard error of estimate says nothing at all about the probability that the effects of those variables are really zero and one is observing only chance effects. (Those propositions are assessed by means of the standard errors of the coefficients and the t-statistics as described above.) The standard error of estimate is a way of assessing how important the random part of the model is; it does not tell one how large the effects of such randomness are on one's ability to measure the systematic part.

An example may make this clear. Suppose that a group of workers are all paid the same per-hour wage, w, for each hour worked. Suppose, in addition, that workers are employed for different numbers of hours. Now suppose that at the end of each week each worker takes his pay and engages in a high-stakes roulette game. Then the income of each worker will be the sum of his pay from his job and his winnings or losings in the roulette game.

Now suppose that we are trying to estimate the common per-hour wage, w, from data on the number of hours worked and total income, but that we cannot observe take-home pay directly. We could do this by a regression in which the dependent variable was total income and the independent variable was hours worked; the coefficient of hours worked would be our estimate of the per-hour wage, w. The influence of the roulette game, of course, would be the random part of the model.

How would we measure the accuracy of our estimate of the per-hour wage? This would be measured in terms of the standard error *of the estimated coefficient* (w). If we had a large enough sample, that standard error would be very small. (This is the consistency property of least squares.) Despite this, we would still find a large standard error *of estimate* because no matter what we did, we would be unable systematically to estimate the effects of the unsystematic roulette game. In such a circumstance, we would be entitled to conclude that there were large unsystematic effects that affected our ability to predict total income. However, under no circumstances would we be entitled to conclude from that fact that we had a biased or unreliable estimate of the per-hour wage. Still less would we be entitled to conclude that changing the number of hours worked had no effect on income (i.e., that the true wage was equal to zero) or, to take the most extreme case, that workers should be

---

standard error of forecast and the standard error of estimate differ for the following reason. Whereas the standard error of estimate measures the extent of deviation in the sample period around the relationship as estimated, forecast errors will involve not only deviation from the estimated relationship but also the fact that the estimated relationship itself may deviate from the true relationship.

The way in which these two standard errors differ is somewhat instructive. In general, one expects to be surest about where the true relationship is for points that fall inside the range of points already observed in the sample. One would be less sure of points less typical of the sample. The standard error of forecast does depend on how far from typical sample values the values of the independent variable for the forecast period happen to fall. It is larger the farther away from the sample are such values. Given the location of the independent variables for the forecast period, however, the standard error of forecast is proportional to the standard error of estimate, which does not vary with such location.

indifferent about whether or not they are laid off. Statements of this sort would be signaled by very large standard errors of the estimated per-hour wage, the regression coefficient of hours worked, not large standard errors of estimate of the regression.

Thus, a large standard error of estimate of the regression tells you that you do not know *everything*. This is not the same as telling you that you do not know *anything*. This is important in practice. In the case of the firemen what is involved is the difference between being able to predict the number of accidents well and being sure that employment of firemen affected that number. While related, these are not the same thing and they are measured differently.

3. *The Correlation Coefficient*. The most common way of normalizing the standard error of estimate for different units is to compare it (or more properly, its square) with a measure of the total variation of the dependent variable. What such a comparison does is to split the variation of the dependent variable around its mean into a part that is explained by movements of the independent variable (the systematic part) and a part that is not so explained (the unsystematic part). The squared multiple correlation coefficient, R^2, measures the percentage of that variation that is explained by the systematic part.[29]

How should values of R^2 be interpreted? Obviously, a value of zero means that movements in the independent variables do not explain movements in the dependent variable at all. The higher R^2, the greater the association between movements in the dependent and independent variables. A value of unity means that the entire variation in the dependent variable is explained by the model.[30] Beyond that, this commonly used measure must be approached with a fair amount of caution, since R^2 can be affected by otherwise trivial changes in the way in which the problem is set up.[31]

II. The Appropriate and Inappropriate Use of Multiple Regression in Legal Proceedings

So far, this Article on "Multiple Regression in Legal Proceedings" has been primarily about multiple regression. The time has come to talk about

29. The reasons for writing the correlation coefficient as a square need not detain us here.

30. How high a value of R^2 is to be expected depends on the number of degrees of freedom. (See note 22 supra). When one has two observations with which to fit a line, for example, such a fit will always be exact and R^2 always equal to unity. Where the line must fit many observations, then an R^2 near unity would be more impressive evidence that movements in the dependent variable are explained by movements in the independent variables.

31. Thus, for example, suppose that in the audience-revenue relationship, we had decided that the true relationship was logarithmic, with the logarithm of revenue as the dependent variable. Suppose also that one of the independent variables was the log of audience size. Suppose then that we subtracted the log of audience size from both sides, making the dependent variable the log of revenue per viewer (equal to log of revenue minus log of audience size). Obviously, the only substantive thing that this would do would be to subtract one from the coefficient of the log of the audience. But it would also change R^2, which would now measure how much of the variation in the log of revenue *per viewer* we were explaining rather than how much of the variation of the log of revenue itself. The resulting value of R^2 might thus be either higher or lower than the original value.

legal proceedings. I shall do this by discussing three areas where multiple regression analysis has figured: the examination of wage discrimination against women; the determination of damages in price-fixing cases; and the evaluation of punishment as a deterrent to crime. These three examples will illustrate a number of the technical points already made as well as providing some lessons concerning what multiple regression analysis can and cannot do. I believe multiple regression analysis to be an entirely appropriate tool for the examination of possible discrimination in wages, but I am very dubious about its utility in price-fixing cases and I believe it to be dangerously misleading in the examination of deterrence.

A. *Discrimination in Wages*

In this example, a case is brought against a firm on behalf of a group of its women employees. They charge that the firm discriminates by paying women less than men. The object of the statistical study is to test whether this is indeed so.

Let us suppose that the facts are such that it appears to be so. The wage paid the average female employee is less than that paid the average male employee. To make things simple, let us suppose that we are considering only women and men in similar jobs.[32] The firm defends (or is likely to defend) by claiming that the women are on the average not as qualified as the men. In particular, they are less well educated and have less job experience. They also score lower on certain aptitude tests.

This is obviously a reasonable defense, if in fact it is true. For it to be true, however, it must not only be the case that women, on the average, are less qualified according to these various measures, but also that the difference in qualifications accounts for the difference in pay. If the firm does not pay well-educated men more than less-educated men, then it can hardly claim that this is the basis for the difference between male and female wages.

Multiple regression is well suited to answer this sort of question fairly precisely. Moreover, without a multiple regression study it is difficult to see how it could be decided. The raw comparison of average wages for women and for men may make one suspicious, but it cannot tell one anything definite. Indeed, it can be misleading in either direction. For example, it would be entirely possible in a different setting that women are paid on the average just as much as men but that a multiple regression analysis would show that there is indeed discrimination because women are *more* highly qualified in the measures that account for the variation in male pay.[33]

Returning to the original problem, how can this be set up in a multiple

---

32. Controlling for job classification is an obvious thing to do and might be done by multiple regression.

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

regression framework? We begin by doing something that may seem needlessly cumbersome but will pay off later. We define a variable, S, as follows:

$$(4) \quad S = \begin{cases} 0 \text{ if the employee is a woman} \\ 1 \text{ if the employee is a man} \end{cases}$$

S is what is called a "dummy" variable, used in situations where one wants to examine discrete rather than continuous variations—in particular, classification into categories. Consider the regression equation:

$$(5) \quad Y = a + bS + u$$

where Y denotes the income paid to a particular employee. It is not hard to see that estimating equation (5) by least squares regression is simply another way of computing the difference in average pay between men and women. If $S = 0$, then, on the average, pay will be given by a; this will be the average pay of female employees. On the other hand, if $S = 1$, then, on the average, pay will be given by $(a + b)$; this will be the average pay of male employees. The difference in the averages is thus b, the coefficient of S, and testing whether that coefficient is significantly different from zero tests whether men are indeed paid more than women.

But of course, such a test is only a test of the original proposition, that men, on the average, are paid more than women and that the difference in pay is not accounted for only by random fluctuations. Such a test is better than simply looking at the difference in pay, but we have not yet tackled the problem of controlling for other variables, namely qualifications.

Such controlling is fairly easily done. For example, suppose for a moment that there were only one measure of qualifications (say, aptitude test scores), denoted by A. Consider the following modification of equation (5):

$$(6) \quad Y = a + bS + cA + u$$

Estimation of this equation by multiple regression will give an answer to the question of whether sex affects wages, with aptitude test scores constant. This may be seen diagramatically in Figure 2.

In Figure 2, employee income is plotted against aptitude test scores. Points denoting male employees are indicated by M; points denoting female employees are indicated by F. I have drawn a case in which male employees are obviously paid more than female employees on the average, but in which, again on the average, female employees score lower on aptitude tests than do male employees. Examination of the average wages without correcting for aptitude tests (equivalent to least squares regression estimation of equation (5)) amounts to drawing a horizontal line in the diagram (horizontal because aptitude is assumed to have no effect in equation (5)) at the level of average male income and another one at the level of average female income. These are relatively far apart. Correcting for aptitude test scores by estimating equation (6), on the other hand, amounts to drawing two

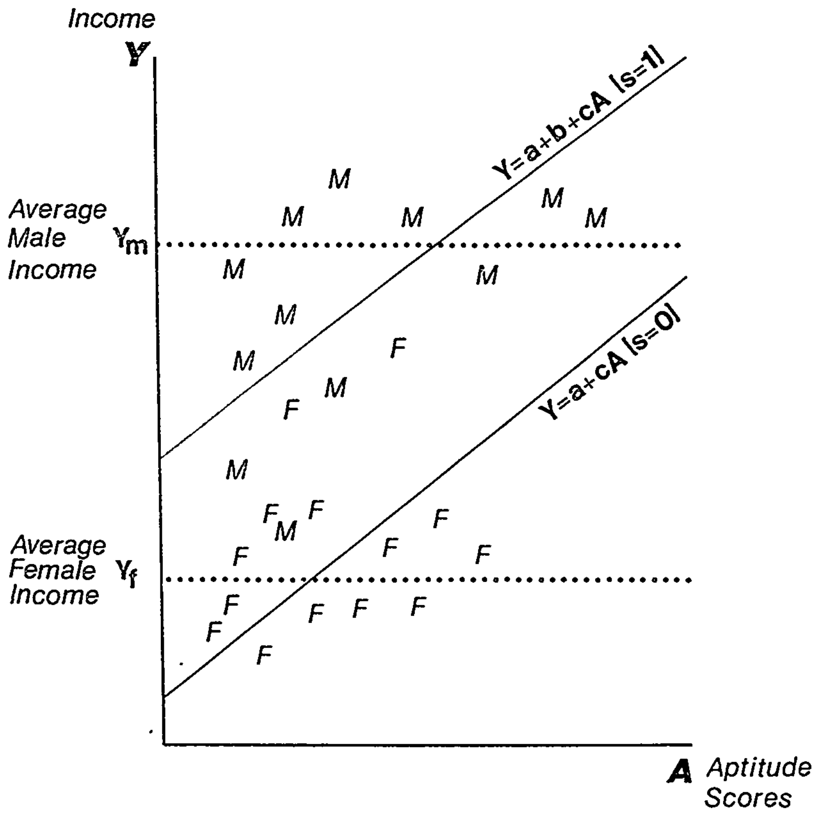


Figure 2

parallel lines through the male and female points respectively. The fact that the lines are parallel indicates the assumption that aptitude tests should have the same effect on wages for males and females if there is no discrimination. The difference in the intercepts is the coefficient of S, a measure of the remaining difference in wages after aptitude scores have been controlled for.

The proposition that males systematically earn more than females even after controlling for aptitude test scores can now be directly tested. This would be done using the t-statistic associated with b (the coefficient of S) to see whether that coefficient is significantly different from zero. (Since no one supposes that women earn systematically *more* than men, the appropriate test would be a one-tailed test.) "Significance at the five percent level" would require a t-statistic of a little more than 1.6.

This example can be extended in a few ways that are worth discussing. In the first place, there is no reason why only one measure of qualifications—aptitude test scores—should be controlled. I chose that case because the resulting diagram was easy to draw. If there are several possible measures of qualifications, then all of them can be included in the regression as new variables. One of the great advantages in this problem is that there are not many variables that plausibly explain wages, and thus interest centers simply on whether sex is one of them. There is little need to thrash about for various different combinations of variables that might be included. Rather, having found an apparent effect in the raw data, the only question is whether that effect is caused by failure to control for other plausible variables.

I have set up the problem in equation (6) as though the only issue was whether a man with given aptitude was paid a fixed number of dollars more than a woman with the same aptitude. This is indicated in Figure 2 by the constant distance between the two sloping lines. According to the equation, women are at a constant dollar handicap whatever their aptitude, and the question is whether or not that handicap is zero. But of course, this may not be the most likely possibility. It is at least as plausible that women are at a constant *percentage* handicap, so that the difference in dollar terms is greatest for women with high aptitudes. This is easily accommodated in the analysis. I shall not attempt to draw the resulting diagram, but all that would be required would be the use of the logarithm of income instead of income itself as the dependent variable in equation (6).

One might also try a somewhat subtler variation. I have set up equation (6) (or its logarithmic equivalent) so that what is tested is the hypothesis that women are at a disadvantage, *given* that aptitude test scores affect wages in the same way for men and for women (the sloping lines in Figure 2 are drawn parallel). This is a good way to do it, but it is not the only way. One could estimate two separate regression equations—one for men and one for women—in which income would be regressed on aptitude. One could then test to see whether the regression coefficients for the two equations were the same *in all respects*. After all, it would be evidence of discrimination if the effect of aptitude tests on wages was not the same for men as for women. It is possible to construct cases in which b in equation (6) turns out to be zero, but in which separately estimated equations would yield significantly different values of b for men and women. On the other hand, trying to examine several things at once (i.e., whether whole sets of coefficients are the same for men and women) will produce less powerful tests than will examining each one of them individually.

Two other features of this example deserve comment. First, I have deliberately used aptitude test scores as a measure of aptitude. It is common knowledge that such tests do not provide perfect measures of ability. However, this may not make any difference in the validity of the regression model. To the extent that true aptitude has different dimensions, the crudeness of aptitude test scores as a measure may be corrected for by the other variables to be introduced into equation (6)—variables such as years of education or work experience. Second, what matters in the current problem is what the employer can observe in distinguishing aptitude. The defendant in this case will look relatively weak if he claims only that he had an unmeasurable way of evaluating aptitude and that all measurable methods are subject to error. In effect, what is important in this problem is not some underlying measure of aptitude but the measure that the employer can see and reward. An argument that aptitude tests are subject to error ought to be challenged by demand for some more reliable but objective measure.

Putting this aside, however, the crudeness of aptitude scores might make a substantial difference if the true variable (aptitude) were measured only by aptitude test scores with a random error. In such a case, it is possible to show that the estimates of c, the coefficient of aptitude test scores in equation (6), would be biased toward zero. This is perhaps what one would expect, since putting in variables that contain a lot of "noise" is likely to result in estimates suggesting that those variables do not have much systematic effect. More important, however, the bias will not be restricted to the coefficient of the variable that is subject to the error. In the present problem, the variable S (describing sex differences) is correlated with the variable A (denoting aptitude test scores), reflecting the fact that, in the sample of employees, women tend to score lower than men on aptitude tests. Such correlation means that the coefficient of S will also be biased and this coefficient is the one that is of interest. Unfortunately, it is not possible to say (without more assumptions) in what direction that coefficient will be biased. Under some circumstances, there are steps that can be taken to guard against the effects of measurement error, but it would take me too far afield to discuss them here.

The final point to be made about this example is that accurate prediction of the dependent variable, income, is not required for successful resolution of the problem. Rather what is involved here is a direct test of the significance of a particular coefficient. The precision of that test (technically its "power") will depend on the standard error of that coefficient and not directly on how well the equation can be expected to do in predicting the dependent variable. Generally, tests like these are likely to be more successful than tests that depend directly on predictions.

What makes the wage discrimination example so suitable for multiple regression is its simplicity and the readiness with which it can be cast into the mold of a test of the significance of a particular regression coefficient. Notice in particular the following feature: whether there is discrimination or not, one would expect the expanded version of equation (6) to fit well. What is being done there is to imbed in a theory of wage determination the difference that discrimination does or does not make. At least at this level, the question of what factors other than discrimination determine wages can be considered without regard to whether or not there is in fact discrimination. Further, the presence or absence of discrimination makes a clearly definable difference in the result one would expect to find. These features stand in contrast to those of the next example.

B. *Antitrust Damages in Price-Fixing Cases*

In this example, the defendants have lost on the issue of liability in a price-fixing case, and the issue to be decided is the extent of damages. The defendants prepare a study attempting to show that the effect of fixing the price was minimal, in that the price would have been the same (or higher) without the conspiracy.[34] There are a number of ways in which this might be done, but I am very dubious about the usefulness of any of them.

One way to proceed is to take a leaf from the discrimination example just discussed. In that example, the study proceeded by controlling for several variables and, in effect, estimating what income would have been if there were no discrimination. Why not systematically estimate what prices would have been without price fixing? We might think of doing this as follows. Under competition, price is determined by the intersection of supply and demand curves. Let us assume, for simplicity, that there are no close substitutes for the product in question, so that demand depends only on the income of consumers (or the output of industrial customers) as well as on price. Supply will depend on price and on costs, which in turn depend on the prices of the factors of production. This suggests that we ought to be able to explain price by a regression equation involving consumer income and factor prices.

Although one might assume that *quantity* should be included as one of the variables that may have an impact on price, it is more appropriate to treat price and quantity independently since, in fact, the same market forces control both. This is evident from an examination of the specific equations (supply and demand curves) that determine supply and demand in the market.

Quantity, like price, is determined by the intersection of the supply and demand curves. Assuming linearity, for convenience, we can write the demand curve as:

(7) $Q = a + bP + cY + u$

Here, Q denotes quantity, P denotes price and Y denotes consumer income. As before, u is a random disturbance. Similarly we can write the supply curve as:

(8) $Q = d + eP + fW + v$

Here, W is a measure of factor prices and v is another random disturbance.

Equations (7) and (8) form what are called the "structural equations" of a "simultaneous equation system." Such a system involves the interaction of more than one equation—equations that can be solved simultaneously. The fact that price is determined by the intersection of supply and demand is reflected by the fact that P and Q must have the same value in both equations. We can thus solve both equations together for those two variables

34. Since, under the per se rule, the ineffectiveness of a price-fixing conspiracy is not a defense, such a showing would be irrelevant to the issue of liability.

by equating the values that each equation predicts for the "quantity" variable. To do this, we create new coefficients (π_0, π_1, π_2, etc.) that depend on all the coefficients of the supply and demand curves. When this is done, the solution for price will look as follows:

$$(9) \quad P = \pi_0 + \pi_1 y + \pi_2 w + u^*$$

u^* is the random disturbance, which depends on some of the coefficients in the supply and demand curves, as well as on u and v. (Precisely, it is equal to $(u - v)/(e - b)$.) The exact algebra need not detain us. There will be a similar solution for Q.

Equation (9) and its companion for Q are called the "reduced form" of the model. They show price and quantity directly in terms of those variables that are determined by forces other than those being modelled (Y, W, u, and v). Such reduced-form equations can be estimated by least squares regression.

It would be a mistake, however, to include Q in the equation for P. It does not appear in equation (9) for the very good reason that quantity and price are jointly determined by the same forces, and it cannot be said that one of them determines the other. A regression that includes quantity on one side and price on the other might be interpreted as an attempt to estimate either equation (7) or equation (8) directly, but this cannot be done consistently by least squares. The easiest way to see this is as follows. A movement in the disturbance term in equation (7), u, affects quantity, Q; this is essentially a random shift of the demand curve. But random shifts of the demand curve affect not only quantity but also price. Hence, shifts in u are associated with movements in P, as can be seen directly from equation (9) and the fact that u^* depends on u. This means that, in estimating equation (7), the fundamental assumption of least squares—that random disturbances move independently of the independent variables—is violated. Equation (7) can be estimated, but least squares is not the way to do it.

Thus, trying to determine what price would have been in a competitive market by regressing price on a set of variables including quantity is doomed to failure. Suppose, however, that we were more sensible and simply regressed price on income and factor price (Y and W), thus estimating equation (9) directly and using that equation to predict price absent the price-fixing agreement.

This is better, but still not adequate. The problem here is that there will not be a clear distinction between the results that one would obtain if the market was affected by the price-fixing scheme and the results that one would obtain if it was not. If the market was not competitive but was seriously affected by price fixing, price was not determined by the intersection of competitive supply and demand curves. Rather, price was determined largely by the price fixers. But the price fixers presumably did not set arbitrary prices but rather set prices to maximize their profits to the extent that they could.

Without going into great detail, it is not hard to see that profit maximization would have required consideration of the position and shape of the demand curve (7) as well as consideration of the costs of production. In the standard terms of economists, profit maximization requires the equating of marginal revenue and marginal cost. Marginal revenue will depend directly on demand and marginal cost directly on factor prices. The price and quantity that equate marginal revenue and marginal cost will, just as in equation (9), depend on income and factor costs. Indeed, for price, one is quite likely to end up with an equation identical to equation (9); the difference that price fixing makes is that the coefficients in equation (9) will be different under price fixing than under competition.

This means, however, that there is no point to estimating equation (9) directly and using it to forecast price. Equation (9) would be valid whether or not there was price fixing and one will not be able to tell whether the predictions that it generates are competitive or noncompetitive. The case was quite different in the wage discrimination example. There the issue was sharply defined as whether a certain coefficient was zero or nonzero. Here the issue might be described as involving differences in a certain set of coefficients (the π's in equation (9)), but we can estimate those coefficients only once and there is thus no way that we can compare the values we obtain with the unknown values that we would have obtained under either the competitive or the noncompetitive hypothesis.

Does this mean there is nothing that can be done? No, but it comes close. We might proceed in a somewhat more sophisticated manner and try to estimate equation (7), the demand curve, which is the same under both regimes. We might then ask what the competitive supply curve would have looked like. Theoretically this could be done, but in practice it is probably impossible. The demand curve (equation (7)) can be estimated. As we have seen, it cannot be estimated by least squares under the hypothesis of competition, but there are other methods of estimating it, and those methods would remain valid, in general, even under a scheme of price fixing.[35] However, in order to find out what price would have been under competitive conditions, it will be necessary to estimate the competitive supply curve. One cannot do that directly from the observations because to do so is to assume that the observations were generated by competitive supply and demand. That, however, is what one wants to prove. Hence, one will have to look elsewhere. In general this will mean estimating the cost curve of the producers and calculating marginal cost. Even if the defendants are

35. If one were willing to admit that the price-fixing agreement did have a substantial impact on price (which, presumably, one is not), least squares estimation of the demand curve might become easier, essentially because prices would have been determined in a controlled manner. On this point, see my study of aluminum demand, F.M. Fisher, A Priori Information and Time Series Analysis 93-117 (1962).

willing to give up the information required for this calculation, it is likely to prove extraordinarily difficult to estimate. Once we move away from simple one-product examples, the cost calculations (and indeed the estimation of the various demand curves as well) become quite complicated. What is involved here is a major undertaking requiring a great deal of data, most of it unlikely to be in usable form, and generating only a thin promise at the other end. Indeed, if one is going to look directly at cost information, it might be better to make a direct showing that prices approximated marginal costs. To do that, one would not need to look at demand.

There remains one possibility in this area that looks slightly more promising. Many of the problems just discussed occur because one wants to know how competition would have looked without directly assuming that competition in fact existed. If, however, there is agreement that the price-fixing conspiracy was in effect only for a limited time, then one might consider estimating the reduced form equataion for price (equation (9)) and the companion equation for quantity, using only data from the competitive period. One conld then use those equations to forecast price for the price-fixing period and study the difference in results.

This sort of program is feasible, at least in principle.[36] Unfortunately, it is unlikely to pay off in practice. One will be using the estimated equations to forecast out of the sample period. If conditions have changed (and over time they usually do) this is going to mean forecasting away from the mean of the sample. Even if the model is entirely correct, one is not going to be able to make this sort of forecast with a great deal of certainty. One is likely to find that the price at a given moment during the price-fixing period is not significantly higher than that which would be predicted by the competitive model, but that the standard error of that prediction is large. Thus, although it will be possible to test whether the difference in price is significant, it will probably be very hard to decide how much of that difference is due to random error.[37] Furthermore, variations in price in either direction can be explained away, by either plaintiffs or defendants, in terms of shifts in demand or cost conditions. Hence, if what is involved is prediction over a long time, this forecasting may be worth trying, but it is not likely to be useful. As opposed to the other approaches already discussed, however, it does have the merit of providing a clear comparison of the two hypotheses involved.

36. There may be some technical problems concerning whether to estimate equation (9) directly by multiple regression or to use sophisticated simultaneous equation techniques to estimate equations (7) and (8) directly, but they need not detain us.

37. This would generally be tested by a so-called "Chow" test. See Fisher, Tests of Equality Between Sets of Coefficients in Two Linear Regressions: An Expository Note, 38 Econometrica 361-66 (1970). This would also be the test used to determine whether the entire regression of income on aptitude was the same for men and for women in the discrimination example above.

C. *Punishment as a Deterrent to Crime*

The last topic that I will discuss is the use of studies that purport to examine the effect of punishment as a general deterrent to crime—that is, as a deterrent to persons other than those being punished. I have already mentioned the death penalty studies referred to by the Solicitor General. In addition, there are a number of studies of other categories of crimes and types of punishment. This is not the occasion to discuss these studies in great detail; such discussions can be found elsewhere.[38] However, a consideration of some of the reasons why these studies are unsatisfactory will illustrate points that are generally applicable to the use of multiple regression analysis.

At first glance, the problem seems to be eminently suitable for regression analysis. Nearly any examination of data in which punishment varies also shows crime varying in the opposite direction. Yearly data on murders committed in the United States (a "time series") show the number of murders rising in years with no executions. With respect to other crimes, cross-section data show that jurisdictions with less severe sentences tend to be the jurisdictions with higher crime rates. It plainly appears that there is a negative correlation between severity of punishment and crime rate and that the problem is merely that of assessing the magnitude of the deterrent effect.

Unfortunately, while I agree that there probably is something significant in these data, the problem of measurement turns out to be very severe. This is true for more than one reason. First, there is a problem because we do not have a very good theory of what causes crime, and thus we do not really know what other variables should be controlled for in deriving a crime equation. Second, one has to control not only for other variables in the same equation but also for the presence of additional relations between those variables and crime. Add to this the doubtful nature of much of the data and one has a serious problem.

Let me begin by considering the death penalty studies.[39] The primary study[40] used time-series data on the United States as a whole for the years 1933-1969. This is a sample of thirty-seven observations, although data on some of the variables had to be constructed. However, it turns out that the results depend almost entirely on the years after 1962. This is, perhaps, no surprise; it was primarily in these years and in the early 1970's that many jurisdictions experimented with the abolition of capital punishment. It does mean, however, that there is only a relatively limited amount of data to use in controlling for other effects, despite the seemingly large sample size. Furthermore, these same years coincide with a general upsurge in crime,

38. See, e.g., Deterrence and Incapacitation, supra note 1.
39. For a more detailed discussion and references, see Klein, Forst & Filatov, supra note 1.
40. Ehrlich, The Deterrent Effect of Capital Punishment: A Question of Life and Death, 65 Am. Econ. Rev. 397 (1975).

not just in those crimes subject to capital punishment. Therefore, we cannot be sure that the results of the study do not simply depend on poorly understood phenomena concerning the causes of crime.

There are lessons to be learned here. First, when faced with a multiple regression study, one should try to determine whether the results crucially depend on certain of the years chosen or whether they stand up to variations in the sample. If the results do depend on certain years, one should try to decide whether there are other characteristics specially associated with these years that might have affected the results. Second, and perhaps more important, one must try to determine whether enough is known about the phenomenon being investigated (here, the causes of crime) to estimate it in terms of the model selected. If not, there will be other plausible explanations for the results achieved.

The death penalty study also turns out to depend rather crucially on the form of the equation used. There is a big difference in its results depending on whether the equation is estimated in linear or logarithmic form.[41] Of course, if one had reason to believe that the correct form of the equation was one or the other, one would simply use that form. But one does not know which form is "correct." Results that depend on the use of a particular version of the equation may not be valid; they depend on an unsupported assumption.[42] When one is deciding whether to execute a man, it ought to concentrate the mind wonderfully. In such matters, the studies to be relied on ought not depend on particular sample periods or choice of specifications.

Many of the problems with the capital punishment study arise because of the limited nature of the available data. An obvious alternative set of experiments would involve the use of data concerning various crimes and drawn from different jurisdictions, in order to get a large sample and a lot of variation.[43] The trouble here is as follows.

Obviously, there are reasons other than variations in punishment why crime rates vary over jurisdictions. It is therefore necessary to control for such reasons. Some possibilities for such variables are unemployment rate, percentage of urban population, and so forth. Multiple regression might in fact do this.

Unfortunately, there are also reasons why punishment levels vary over jurisdictions. One of the reasons suggested in the literature has to do with crime rates. It is easy to see how this might happen. Jurisdictions

41. See notes 10 & 20 supra.

42. There are ways of testing whether one form is better than another. Often, however, it is hard to tell from the results.

43. Ehrlich has also performed cross-section analyses of murder, but I am less familiar with these than with his study of noncapital crimes. The latter is Ehrlich, Participation in Illegitimate Activities: A Theoretical and Empirical Investigation, 81 J. Pol. Econ. 521 (1973). The following comments are expanded in Nagin, General Deterrence: A Review of the Empirical Evidence, in Deterrence and Incapacitation, supra note 1, at 95, and Fisher & Nagin, On the Feasibility of Identifying the Crime Function in a Simultaneous Model of Crime Rate and Sanction Levels, id. at 361.

with higher crime rates may adopt "get tough" policies. Alternatively (and this is the suggestion in much of the literature), jurisdictions with high crime rates may overload their punishment facilities and thus may come to tolerate relatively common offenses somewhat more than do jurisdictions with low crime rates. In any event, there is a serious possibility that the variation in punishment levels over jurisdictions can be accounted for, at least in part, by the variation in crime rates. In this circumstance, as in part of the supply and demand example given above, the problem is not merely that one has to control for other variables, but that one has to control for the presence of another equation. To see the kind of problem that arises, consider the following vastly simplified example.

Assume, for the moment, that the *only thing* that affects crime rates is punishment. Assuming linearity, for convenience only, the crime rate equation to be estimated could then be written as:

(10) $C = a + bS + u$

Here, C is the measure of crime rate and S is a measure of punishment or

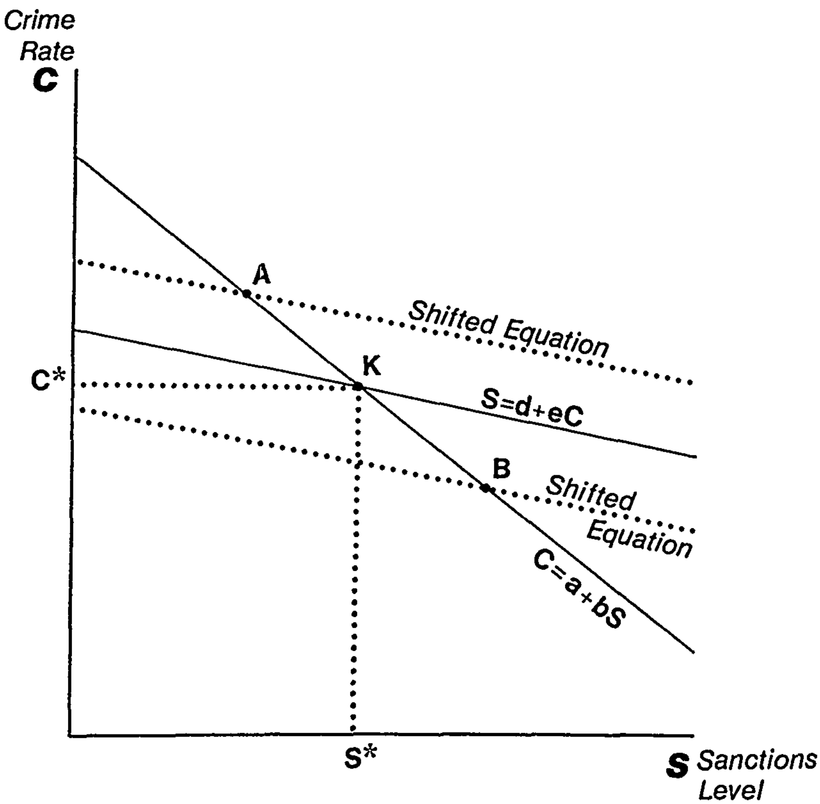


Figure 3

sanctions levels. The coefficient b would represent the deterrent effect of increasing sanctions.

Suppose, however, that sanctions also depended on the crime rate and only on the crime rate. Then the equation that shows how sanction levels are determined can be written (again assuming linearity):

(11) $S = d + eC + v$

In these equations u and v are random disturbances.

Given the interrelation between these two equations, one could not effectively estimate the crime equation (equation (10)) by least squares regression. The fundamental assumption of least squares regression is that the random disturbance term operates independently of the independent variable. All of the properties of least squares depend on this. In the present instance this would require that u and S be uncorrelated. This cannot be the case, however, because the model itself (just as in the supply and demand example given above) implies that it is not so. An upward shift in u, according to equation (10) itself, will mean an upward shift in the crime rate, C. But an upward shift in the crime rate, C, will, according to equation (11), cause a shift in the sanctions level, S. Hence, shifts in u cannot be independent of shifts of S and least squares regression will fail. (This may also be seen by solving equations (10) and (11) for C and S to obtain the reduced form of the system, as was done in the supply and demand example.)

The problem is worse than this, however. To see this, ignore the random disturbances, for a moment, and suppose that equations (10) and (11) were exact. I have graphed those equations in Figure 3. In such a situation, the crime rate and the sanctions level would be entirely determined by the simultaneous solution of the two nonrandom equations—the intersection of the two lines in Figure 3 at K. (The resemblance to a supply and demand graph is not accidental.) If this were really the case, the only point we would ever observe would correspond to that intersection at sanctions level denoted by S * and crime rate denoted by C *. But if that point were the only one observed, there would be no way of recovering equations (10) and (11). In terms of the graph, we could not tell the true crime function (the more steeply sloped line) apart from the sanctions function (the less steeply sloped line) or, indeed, from any other line that went through that same point; each line could vary, in an infinite number of ways, around the point K.

Even if we put random disturbances back in, we would not get anywhere. The effect of random disturbances would be to produce a cluster of points surrounding the intersection drawn in the graph, but again, it would not be possible to recover the two underlying lines that generated this cluster or to tell the two lines apart even if we could recover them. In this circumstance, the crime and sanctions equations are said to be "not identifiable."

This problem, one of *identification,* is a well-studied subject in econometrics.[44] I have deliberately chosen an extreme case. Unfortunately, the identification problem continues in the deterrence studies even when the extreme assumptions are relaxed.

Suppose, for example, that there was some variable that shifted sanctions levels over jurisdictions but did not affect crime rate. This would mean that there would be an additional significant variable in equation (11) that was not also a variable in an expanded version of equation (10). Leaving equation (10) as it is, the effect would be to shift the sanctions equation in Figure 3 up and down. (This is illustrated by dashed lines parallel to the solid line corresponding to the sanctions equation in Figure 3 and marked "shifted equation.") If this happened, we would observe not merely one intersection of the sanctions equation and the crime equation but several intersections, points such as A and B, for example. Those points would all lie on the crime equation and, indeed, as the sanctions equation shifted back and forth because of the presence of the additional variable, the points of intersection would trace out the crime equation.

In such a situation, as the diagram suggests, there is a technique for recovering the crime equation from the data. That technique, however, is not least squares regression, because the correlation between the disturbance term and the independent variable in equation (10) would generate invalid results. Moreover, it will still not be possible to recover the sanctions equation itself.

Because of the identification problem it is necessary to find a variable that shifts one equation of the model but not the equation to be identified. However, it is not only bad practice to attempt to find such variables from the data, it is literally impossible. No amount of manipulation of data generated by the model will reveal such variables; the selection of such a variable must be done as a matter of prior theory.

It is easy to see from Figure 3 why this should be so. If there is a variable shifting the sanctions equation but not the crime equation, then the observed points will be like the points A, B, and K in the diagram. But such a pattern of intersection could also be produced by a variable shifting the crime equation but not the sanctions equation. More generally, it could be produced by shifts in both equations. Only if we *know* from theoretical, nondata-generated considerations that it is the sanctions equation that shifts can we be sure that it is the crime equation that is traced out.

In most situations, such theoretical considerations may readily be found. (For example, consumer income enters demand but not supply curves; factor costs affect supply but not demand.) This is not so in the present case, however. While there are a number of variables that may enter the sanctions equation, it is difficult, if not impossible, to think of such

44. See F.M. Fisher, The Identification Problem in Econometrics (1966) (reissued 1976).

a variable that would not also enter the crime equation.[45] The existing studies have tried to get around this by casually assuming that variables such as unemployment influence sanctions levels but not crime. This is plainly wrong. In the present state of our knowledge, we simply do not know enough about the structure of the system generating the observations to be able validly to estimate the crime equation.

This problem has some general implications for the use of regression analysis. First, it is important to be very careful not only about controlling for additional variables, but also about the possibility that one must control for the existence of additional relationships between the dependent and independent variables. If there are such relationships, least squares will not be an appropriate estimator, and it is at least possible that no appropriate estimator will exist (although this is not common). Second, if there is another equation involved, one must find out how the expert really did his estimation. If he explored the data by multiple regression and then, having decided on a model, altered it with another estimation technique, the results are quite suspect.[46]

Finally, one should make sure that the model used is constructed on sound hypotheses based on theoretical considerations generated from outside the model itself. While multiple regression and related econometric techniques are powerful tools for analyzing data, their proper use presupposes an underlying theory of the structure generating those data. While some hypotheses concerning that structure can be tested with these tools, the theory itself cannot be discovered by computer runs and data experimentation. Thus, the expert making the study must not only understand the proper uses of the statistical tools, he also must learn something about the phenomena and hypotheses being investigated.

Conclusion

Multiple regression analysis can play a vital role in legal proceedings. Used properly, it is an accurate and reliable method of determining the relationships between two or more variables, and it can be a valuable tool for resolving factual disputes. In order for this to happen, however, multiple

45. On the other hand, it is not difficult to think of variables that enter the crime equation but that would not directly influence the choice of sanctions. Unemployment, for example, is far more likely to influence the crime rate than to influence sanctions. Other examples might include measures of income disparity or expenditures on security systems. If such variables really do influence crime rate, but not sanctions, then including them in the crime equation would shift that equation relative to the sanctions equation. The points of intersection traced out would all lie on the sanctions equation, which would then be identifiable and could be estimated (although still not by least squares).

46. Consider the following all too common procedure. Since multiple regression is easy to do, one experiments with multiple regression until one has a version of the estimated equation that corresponds to one's own predilections. Then one reestimates the equation by an appropriate simultaneous equation technique. If the results look very different from the least squares version one goes on exploring. This is not a way to produce consistent results.

regression must be better understood by the legal community; in particular, there must be an understanding of both the potential and the limits of the technique.

It is not necessary that lawyers understand the mechanics of multiple regression in terms of what goes on inside the computer. It is necessary, however, that they understand the regression model and the assumptions being used in any given regression study, how the results of the regression bear on the hypothesis to be tested, and how the results distinguish this particular hypothesis from other hypotheses. The expert constructing the analysis should be able to explain all of this to the attorney who employs him, and an expert who cannot explain such things is likely to fall apart on cross-examination.

Lawyers will increasingly find themselves in a position where it would be profitable to use a regression analysis or where they must confront a regression study produced by an opponent. When that happens, a basic knowledge of multiple regression may be a valuable asset.

NRA-ILA
INSTITUTE for LEGISLATIVE ACTION

APPEARS IN GUN LAWS

Gas station clerk scares off robber, The Times Picayune, New Orleans, La. 09/02/15

WEDNESDAY, SEPTEMBER 9, 2015

A man entered a Shell station in New Orleans, La. and attempted to rob the cashier, by claiming he was carrying a gun. The cashier responded by retrieving a gun and leveling it at the thief, prompting the criminal to flee. (The Times Picayune, New Orleans, La. 09/02/15)

http://www.nola.com/crime/index.ssf/2015/09/gas_station_clerk_pulled_gun_o.html

NRAEXPLORE **MORE LIKE THIS FROM AROUND THE NRA**

BRUCE PIATT WINS 2022 CRAWFISH CUP NRA ACTION PISTOL REGIONAL

An NRA Shooting Sports Journal

LOUISIANA FALLS JUST SHORT OF GETTING CONSTITUTIONAL CARRY

An Official Journal Of The NRA

LOUISIANA, TEXAS MOVING CLOSER TO CONSTITUTIONAL CARRY

An Official Journal Of The NRA

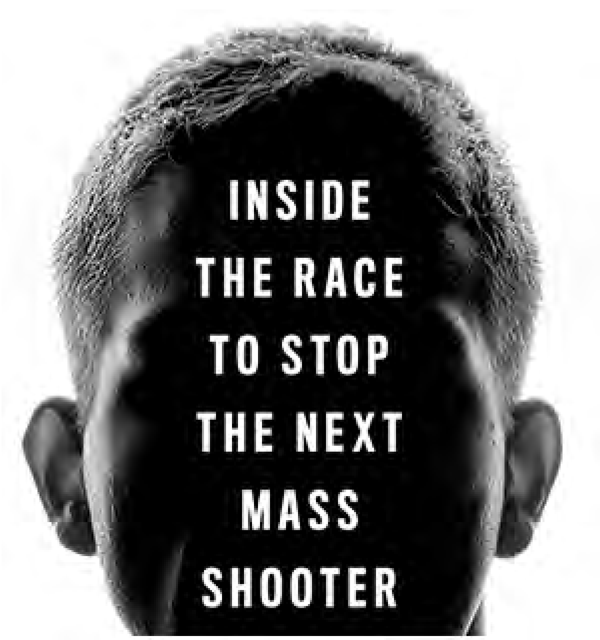
Mother Jones

A Guide to Mass Shootings in America

There have been at least 134 in the past four decades—and most of the killers got their guns legally.

MARK FOLLMAN, GAVIN ARONSEN, AND DEANNA PAN UPDATED: OCT. 14, 2022

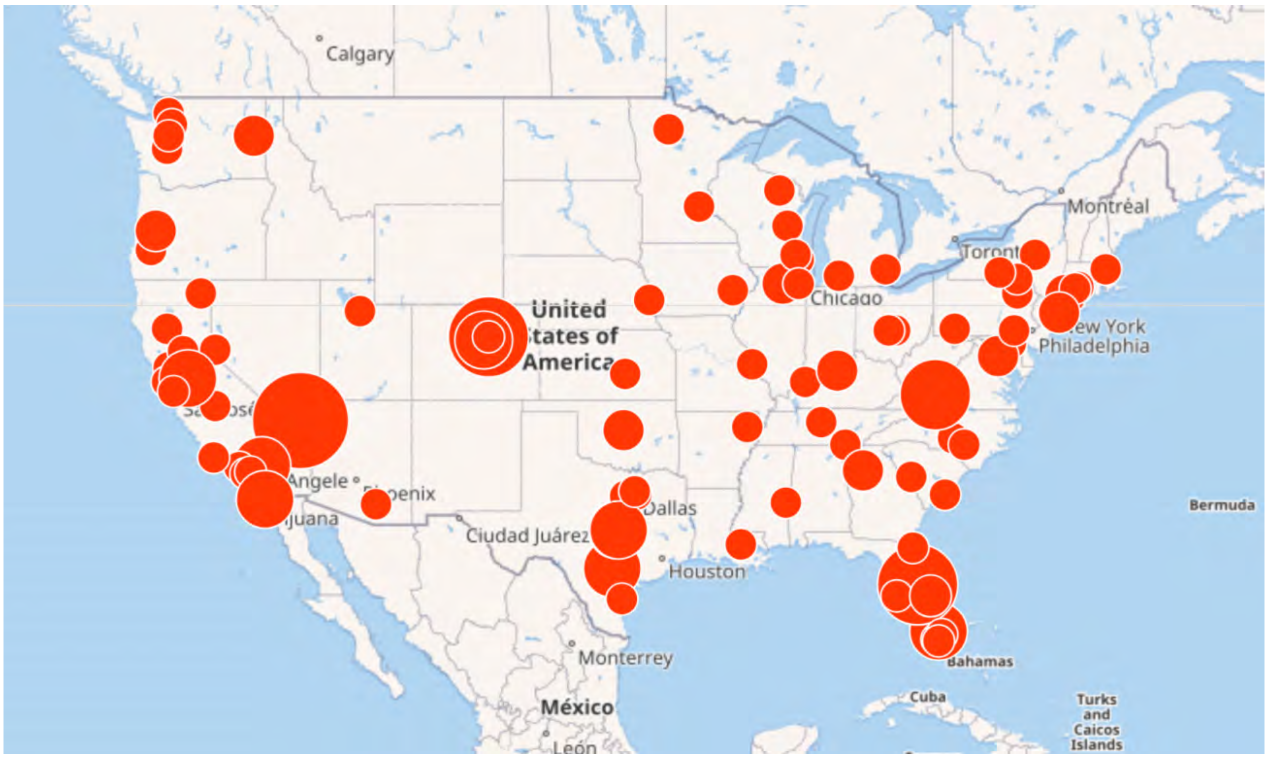
In July 2012, in the aftermath of the movie theater massacre in Aurora, Colorado, *Mother Jones* created a first-of-its-kind **open-source database documenting mass shootings** in the United States. Our research focused on indiscriminate rampages in public places resulting in four or more victims killed by the attacker. We exclude shootings stemming from more conventionally motivated crimes such as armed robbery or gang violence. (Or in which the perpetrators have not been identified.) Other news outlets and researchers have since published larger tallies that include a wide range of gun crimes in which four or more people have been either wounded or killed. While those larger datasets of multiple-victim shootings are useful for studying the broader problem of gun violence, our investigation provides an in-depth look at a distinct phenomenon—from the firearms used and mental health factors to the growing copycat problem. Tracking mass shootings is complex; we believe ours is the most useful approach for studying this specific phenomenon.



Can the next attack be prevented?

Since we began, our interactive map below and the downloadable database behind it have been expanded with 72 additional cases from 2013-2022. Dating back to at least 2005, the FBI and leading criminologists essentially defined a mass shooting as a single attack in a public place in which four or more victims were killed. We adopted that baseline for fatalities when we gathered data in 2012 on three decades worth of cases.

(It is important to note that there have been many similar indiscriminate gun rampages in public places—but resulting in fewer fatalities—that would otherwise be included in our dataset. In that regard, ours is a conservative measure of the problem.) In January 2013, a mandate for federal investigation of mass shootings authorized by President Barack Obama lowered that baseline to three or more victims killed. Accordingly, we include attacks dating from January 2013 in which three or more victims were killed. (Any analysis of the frequency of mass shootings using our database should account for this.) Our original analysis, which covers cases from 1982-2012 with four or more victims killed, follows below. The cases we've documented since then using the revised federal baseline reaffirm our major analytical findings.



This heat map depicting mass shooting casualties across the United States was generated in 2018. To view, explore, or download the updated database, **click here**.

It is perhaps too easy to forget how many times this has happened. The gun massacre at a movie theater in Aurora, Colorado, in July 2012, another at a Sikh temple in Wisconsin that August, another at a manufacturer in Minneapolis that September —and then the unthinkable nightmare at a Connecticut elementary school that December—were some of the latest in an epidemic of such gun violence over the past three-plus decades. Since 1982, there have been at least 134 public mass shootings across the country, with the killings unfolding in 37 states, from Massachusetts to Hawaii. They are occurring more often: An analysis of this database by researchers at Harvard University, further corroborated by a separate study from the FBI, determined that mass shootings have tripled in frequency in recent years.

We've gathered detailed data on four decades worth of cases, including information on the attackers' profiles, the types of weapons they used, and the number of victims they injured and killed. [***Editor's note:*** The following analysis covers our original dataset comprised of 62 cases from 1982-2012.]

Weapons: Of the 143 guns possessed by the killers, more than three quarters were obtained legally. They included dozens of assault weapons and semi-automatic handguns with high-capacity magazines. (See charts below.) Just as a perpetrator used a .40-caliber Glock to slaughter students in Red Lake, Minnesota, in 2005, so too did the one in Aurora, along with an AR-15 assault rifle, when blasting away at his victims in a darkened movie theater. In Newtown, Connecticut, the attacker wielded a .223 Bushmaster semi-automatic assault rifle as he massacred 20 school children and six adults.

The perpetrators: More than half of the cases involved school or workplace shootings (12 and 20, respectively); the other 30 cases took place in locations including shopping malls, restaurants, and religious and government buildings. Forty-four of the killers were white males. Only one was a woman. (See Goleta, Calif., in 2006.) The average age of the killers was 35, though the youngest among them was a mere 11 years old. (See Jonesboro, Ark., in 1998.) A majority were mentally troubled—and many displayed signs of mental health problems before setting out to kill. Explore the above map and database for further details—we do not consider it to be all-inclusive, but based on the criteria we used, we believe that we have produced the most comprehensive rundown available on this particular type of violence. (Mass shootings represent a small fraction of America's overall gun violence.) For the stories of the 151 shooting rampage victims of 2012, click here, and for our groundbreaking investigation into the economic costs of the nation's gun violence, including mass shootings, click here.

Here is a description of the criteria we use:

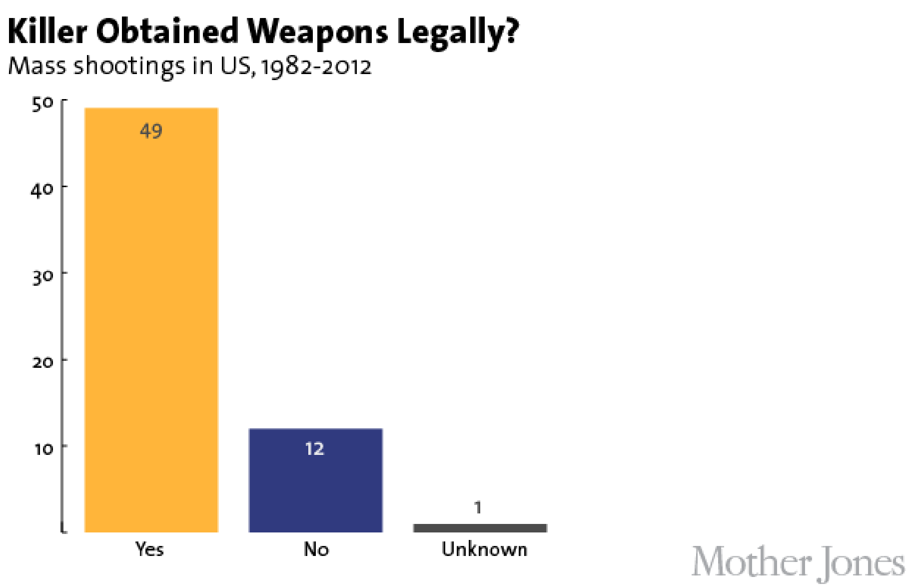
- **The perpetrator took the lives of at least four people**. A 2008 FBI report identifies an individual as a mass murderer—versus a spree killer or a serial killer—if he kills four or more people in a single incident (not including himself), typically in a single location. (*In 2013, the US government's fatality baseline was revised down to three; our database reflects this change beginning from Jan. 2013, as detailed above.)
- **The killings were carried out by a lone shooter**. (Except in the case of the Columbine massacre and the Westside Middle School killings, which involved two shooters.)
- **The shootings occurred in a public place.** (Except in the case of a party on private property in Crandon, Wisconsin, and another in Seattle, where crowds of strangers had gathered, essentially constituting a

public crowd.) Crimes primarily related to gang activity or armed robbery are not included, nor are mass killings that took place in private homes (often stemming from domestic violence).

- **Perpetrators who died or were wounded** during the attack are not included in the victim tallies.
- **We included a handful of cases also known as "spree killings"**—cases in which the killings occurred in more than one location, but still over a short period of time, that otherwise fit the above criteria.

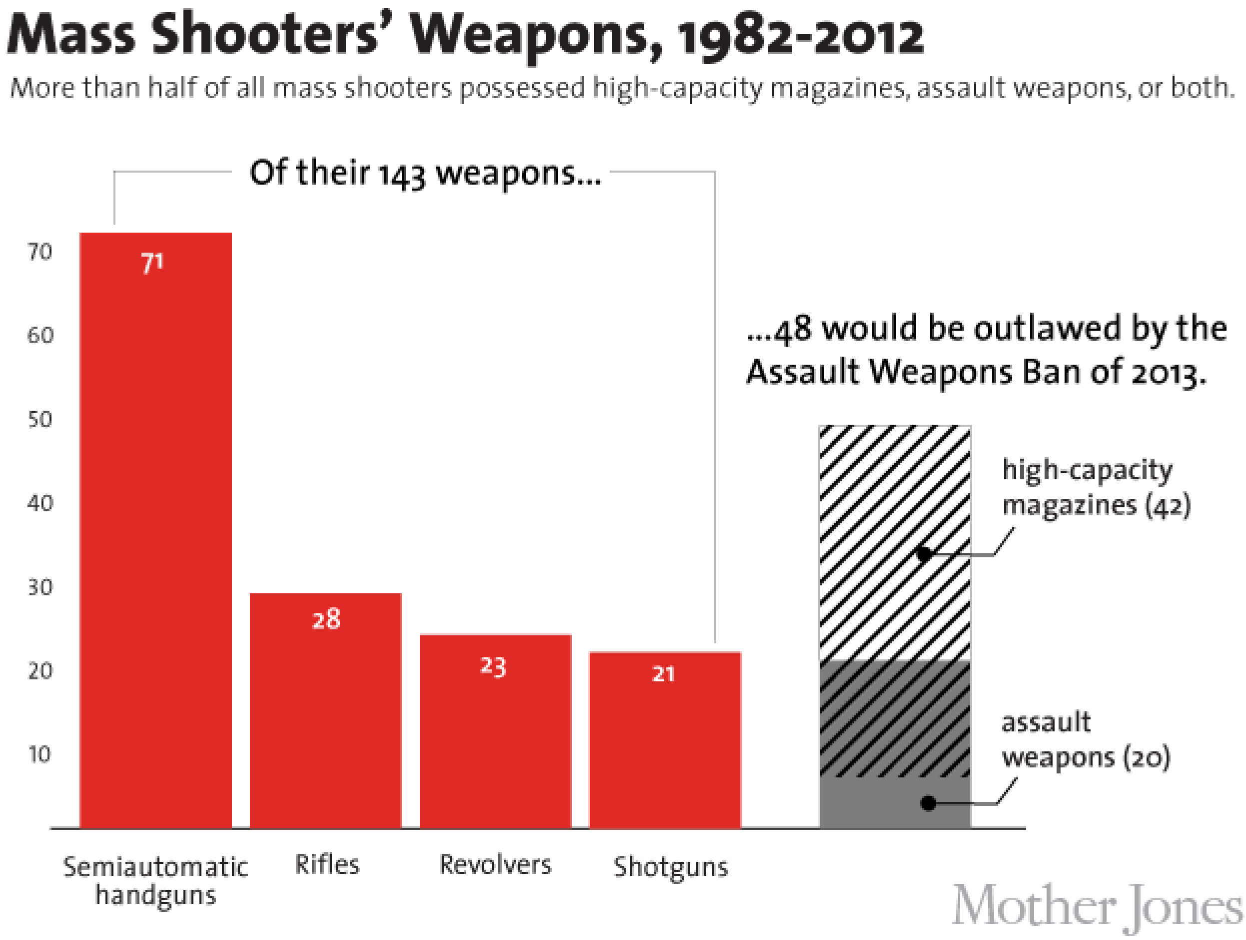
For more on the thinking behind our criteria, see these two explanatory pieces. Plus: more on the mental health factor and on state laws rolling back gun restrictions across the US. And: Explore the full data set behind our investigation.

Here are two charts detailing the killers' weapons:



This guide was first published on July 20, 2012. Since then, we've updated and expanded it numerous times with additional research and reporting. The analysis and charts above cover the data through 2012 (comprising 62 cases); additional data and analysis on the shooters' weapons are in this story. Information on additional mass shootings from 2013-2022 is included in our full data set here. For much more of our reporting on mass shootings, gun violence, and gun laws, see our special investigations: America Under the Gun and The True Cost of Gun Violence.

Additional reporting and production contributed by: ***AJ Vicens, Olivia Exstrum, Tasneem Raja, Jaeah Lee,*** *and* ***Maggie Caldwell.***



Mass Shooters' Weapons, 1982-2012
More than half of all mass shooters possessed high-capacity magazines, assault weapons, or both.
Of their 143 weapons...
70
60
50
40
30
20
10
71
28
23
21
Semiautomatic handguns
Rifles
Revolvers
Shotguns
...48 would be outlawed by the Assault Weapons Ban of 2013.
high-capacity magazines (42)
assault weapons (20)
Mother Jones

The Heritage Foundation's Defensive Gun Use Database

The data set from the Heritage Foundation's Defensive Gun Use Database, analyzed in the Supplemental Declaration of Lucy P. Allen (Dkt. No. 137-1 at ¶¶ 6–12 & n.2), cannot be reproduced in this compendium due to formatting limitations. The data set is not publicly available online. Defendant will deliver a copy of the data set in native Excel format via e-mail to chambers (efile_benitez@casd.uscourts.gov) and plaintiffs' counsel (cmichel@michellawyers.com).

Research Article

Justice Research and Policy
2016, Vol. 17(1) 28-47

DOI: 10.1177/1525107116674926
journals.sagepub.com/home/jrx
SAGE

Large-Capacity Magazines and the Casualty Counts in Mass Shootings: The Plausibility of Linkages

Gary Kleck[1]

Abstract

Do bans on large-capacity magazines (LCMs) for semiautomatic firearms have significant potential for reducing the number of deaths and injuries in mass shootings? The most common rationale for an effect of LCM use is that they allow mass killers to fire many rounds without reloading. LCMs are known to have been used in less than one third of 1% of mass shootings. News accounts of 23 shootings in which more than six people were killed or wounded and LCMs were known to have been used, occurring in the United States in 1994–2013, were examined. There was only one incident in which the shooter may have been stopped by bystander intervention when he tried to reload. In all of these 23 incidents, the shooter possessed either multiple guns or multiple magazines, meaning that the shooter, even if denied LCMs, could have continued firing without significant interruption by either switching loaded guns or changing smaller loaded magazines with only a 2- to 4-seconds delay for each magazine change. Finally, the data indicate that mass shooters maintain such slow rates of fire that the time needed to reload would not increase the time between shots and thus the time available for prospective victims to escape.

Keywords

mass shootings, gun control, large-capacity magazines

[1] College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL, USA

Corresponding Author:
Gary Kleck, College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL 32306, USA.
Email: gkleck@fsu.edu

Introduction—Mass Shootings and Large-Capacity Magazines (LCMs)

There have been at least 23 shootings in which more than six victims were shot and one or more LCMs were known to have been used in the United States in the period 1994–2013. One of the most common political responses to mass shootings has been to propose new gun control measures, commonly focusing on "assault weapons" and LCMs. LCMs are detachable ammunition magazines used in semiautomatic firearms that are capable of holding more than a specified number (most commonly 10 or 15) rounds. For example, the 1994 federal assault weapons ban prohibited both (a) certain kinds of guns defined as assault weapons and (b) magazines able to hold more than 10 rounds (Koper, 2004). At least eight states and the District of Columbia similarly ban magazines with a large capacity, and still other states are considering bills to enact such restrictions (Brady Campaign to Prevent Gun Violence, 2013).

Theory—The Rationale for LCM Bans

When supporters of bans on LCMs provide an explicit rationale for these measures, they stress the potential for such restrictions to reduce the death toll in mass shootings. And indeed there is a statistical association between LCM use and the casualty count in mass shootings (Koper, 2004), though it is unknown whether this reflects an effect of LCM use or is merely a spurious association reflecting the offender's stronger intention to harm many people. If there is a causal effect, how would it operate? Does possession of LCMs somehow enable aggressors to shoot more victims, above and beyond the ability conferred by the use of semiautomatic guns equipped with smaller capacity detachable magazines? (A semiautomatic firearm is a gun that fires a single shot for each pull of the gun's trigger, but automatically causes a fresh round to be loaded into the gun's firing chamber.)

Possession of LCMs is largely irrelevant to ordinary gun crimes, that is, those with fewer victims than mass shootings, because it is extremely rare that the offenders in such attacks fire more rounds than can be fired from guns with ordinary ammunition capacities. For example, only 2.5% of handgun crimes in Jersey City, NJ, in 1992–1996 involved over 10 rounds being fired (Reedy & Koper, 2003, p. 154). Even among those crimes in which semiautomatic pistols were used, and some of the shooters were therefore likely to possess magazines holding more than 10 rounds, only 3.6% of the incidents involved over 10 rounds fired. Thus, if LCMs have any effect on the outcomes of violent crimes, it is more likely to be found among mass shootings with many victims, which involve unusually large numbers of rounds being fired.

Koper (2004) noted that "one of the primary considerations motivating passage of the ban on [LCMs]" was the belief that

> semiautomatic weapons with LCMs enable offenders to fire high numbers of shots rapidly, thereby potentially increasing both the number of persons wounded per gunfire incident . . . and the number of gunshot victims suffering multiple wounds, both of which would increase deaths and injuries from gun violence. (p. 80)

This summary was as much a rationale for restricting semiautomatic guns as it was for limits on magazine capacity, but Koper also concluded that "an LCM is arguably the most important feature of an AW. Hence, use of guns with LCMs is probably more consequential than use of guns with other military-style features" (p. 80). He then went on: "By forcing AW and LCM offenders to substitute non-AWs with small magazines, the ban might reduce the number of shots fired per gun, thereby reducing both victims shot per gunfire incident and gunshot victims sustaining multiple wounds" (p. 81).

It is reasonable to expect fewer people shot if fewer rounds were fired, but Koper did not explain why, for example, the use of three 10-round magazines would result in fewer shots fired than if a 30-round magazine were used. After all, three 10-round magazines and one 30-round magazine both contain 30 cartridges and thus allow 30 shots to be fired. Semiautomatic guns do not fire any faster when they have a larger magazine inserted in them than when they have a smaller magazine, nor is the lethality of any one shot affected by the size of the magazine from which it came. A limit on the number of cartridges that the shooter could fit into any *one* magazine would not limit the total number of rounds of ammunition that a would-be mass shooter could bring to the scene of their crime, or even the total number loaded into multiple detachable magazines.

The main difference between a 30-round magazine and three 10-round magazines, however, is that a shooter equipped with three 10-round magazines would have to change magazines twice in order to fire 30 rounds, while a shooter with a 30-round magazine would not have to change magazines at all. This presumably is what Koper (2004) meant when he wrote that "semiautomatic weapons with LCMs enable offenders to fire high numbers of shots rapidly" (p. 80).

Thus, it could be the *additional magazine changes* necessitated by the use of smaller magazines that might reduce the number of people hurt in mass shootings. Advocates of LCM bans argue that, if LCMs were not available, would-be mass murderers would shoot fewer people because they *would have to reload more often* due to the more limited capacities of the magazines that would then be legally available. A spokesperson for the Violence Policy Center (2011), for example, argued that "High-capacity ammunition magazines facilitate mass shootings by giving attackers the ability to fire numerous rounds without reloading."

It is not, however, self-evident why this should be so. Skilled shooters can change detachable magazines in 2 seconds or less, and even relatively unskilled persons can, with minimal practice, do so in 4 seconds (for a demonstration, see the video at https://www.youtube.com/watch?v=ZRCjY-GtROY, which shows a 2-seconds magazine change by an experienced shooter). Certainly, additional magazine changes do not increase the time needed to fire a given number of rounds by much.

Why, then, might inducing more magazine changes reduce casualty counts? Two explanations have been offered. First, during an additional interval when the shooter was forced to change magazines, *bystanders might tackle the shooter and prevent any further shooting*. Bystanders are presumably more willing to tackle a shooter while the shooter was reloading because it would be safer to do so—a shooter armed with only one loaded gun would not be able to shoot those seeking to intervene during the effort to reload. A shooter equipped only with smaller capacity magazines would have to change magazines sooner and would therefore presumably shoot fewer people before he was tackled by the bystanders.

Second, additional magazine changes could extend the time interval between some of the shots, thereby *allowing more prospective victims to safely escape the scene* than otherwise would have been the case had the possession of LCMs enabled the shooter to reload less often.

These scenarios are plausible as logical possibilities, but have they actually occurred in the past often enough for it to be plausible that they would happen with some nonnegligible frequency in the future? If the past is any guide to the future, the credibility of any expectation of future benefits from LCM restrictions would rely heavily on how often these scenarios have actually played out in past mass shootings. This research is intended to test the plausibility of these possible causal linkages between LCM use and the casualty counts of mass shootings by closely examining the relevant details of such crimes. In particular, it was intended to estimate the share of mass shootings in which LCM use could plausibly have affected the casualty count.

Prior Research on LCMs

No one has actually tested whether mass shooters with LCMs fire more rounds than those without LCMs. We only have evidence indirectly bearing on this issue. Koper reported data showing that there are more *gunshot wound victims* in incidents in which the offender used an LCM (Koper, 2004, p. 86). The meaning of this statistical association, however, is unclear since one would expect it to exist even if LCM use had no causal effect on either the number of shots fired or the number of victims shot. The association is at least partly spurious if the deadliness of the shooter's intentions affects both his selection of weaponry (including magazines) and the number of shots he fires or persons he wounds.

It is a virtual tautology that the deadliness of the shooter's intentions affects the number of people hurt, unless one is prepared to assert that there is no relationship whatsoever between violent intentions and outcomes. While it is certainly true that outcomes do not match intentions perfectly, it is unlikely that there is no correlation at all.

The deadliness of a would-be mass shooter's intentions, however, is also likely to affect preparations for the shooting, such as accumulating many rounds of ammunition, acquiring multiple guns and multiple magazines, and selecting larger magazines rather than smaller ones. Accounts of mass shootings with high death tolls routinely describe the shooters making elaborate plans for their crimes, well in advance of the attacks, and stockpiling weaponry and ammunition (e.g., see Office of the State's Attorney 2013, regarding the Sandy Creek elementary school shootings; *Washington Post* "Pa. Killer had Prepared for 'Long Siege,'" October 4, 2006, regarding the Amish school killings in Lancaster, PA; Virginia Tech Review Panel, 2007, especially pp. 25–26, regarding the shootings at Virginia Tech; "Before gunfire, hints of bad news," *New York Times* August 27, 2012, regarding the Aurora Colorado movie theater shootings). In short, people who intend to shoot many people are not only more likely to end up doing so but also prepare for doing so by acquiring equipment that they believe is better suited to this task.

The most direct indication that the intentions of mass shooters are more deadly than those of the average gun aggressor, aside from the number of casualties inflicted itself, is the percentage of wounded victims who were killed rather than nonfatally wounded. The data gathered for the present study indicate that in 23 LCM-involved mass shooting incidents, a total of 197 gunshot victims were killed and 298 were nonfatally wounded, for a fatality rate of 40.0%. In contrast, Cook (1985, p. 96) reported that police reports on general samples of shootings indicated that about only 15% of those wounded by gunshot were killed. Thus, the lethality of gunshot wounds inflicted by mass shooters is about 2.7 times as high as for shootings in general. Any one shot fired from a gun equipped with a larger capacity magazine is no more deadly or accurate than one fired from a gun with a smaller capacity magazine, so it is implausible that LCMs affect this fatality rate (deaths/persons wounded) by enabling shooters to more accurately hit vital areas of a victim's body where wounds are more likely to be fatal. Indeed, if those who suggest that shooters with LCMs fire faster than other shooters are correct, accuracy would be worse in LCM-involved shootings.

Thus, it is more likely that the high fatality rate in mass shootings is a product of the aggressor's stronger intentions to shoot more people, though it could also be partly a product of the greater use of rifles and shotguns in mass shootings (25 of the 66 guns used in these incidents [38%] of known gun type were rifles or shotguns; in comparison, only 8% of all U.S. gun homicides in 2014 were committed with rifles or shotguns—U.S. Federal Bureau of Investigation [FBI], 2015). This too could be an indication of greater shooter lethality, since rifles and shotguns are, on average, more lethal than handguns (Kleck, 1984). In sum, mass shooters appear to have more lethal intentions as aggressors, apart from any advantages they may gain from use of LCMs.

There is therefore sound reason to question whether a simple bivariate association between LCM use and number of shots fired, or victims wounded, in a mass shooting reflects a causal effect of LCM use. Unfortunately, there is no known way to directly measure the lethality of shooters' intentions at the time of their shootings, so we cannot simply statistically control for lethality of intentions in order to isolate the effect of LCM use. On the other hand, it would become more plausible to conclude that LCM use made its own contribution to the casualty count of shootings, above and beyond the effects of the apparently more lethal intentions of their users, if there was some evidence that either (a) significant numbers of mass shootings were disrupted by bystanders intervening when the shooters attempted to reload detachable magazines or (b) magazine changes increase the time intervals between shots fired, thus potentially allowing more prospective victims to escape to safety. This article provides a close examination of the details of mass shootings so as to cast light on these and related issues.

Method

Definition of Eligible Incidents

We tried to identify, as comprehensively as possible, all mass shootings that occurred in the United States in the 20-year period from 1994 through 2013 inclusive and that were known to have involved an LCM. An LCM was defined as a magazine holding more than 10 rounds of ammunition. A mass shooting was defined as one in which more than six people were shot, either fatally or nonfatally, in a single incident. Any specific numerical cutoff is necessarily somewhat arbitrary, but some are less arbitrary than others. The six-victim cutoff was used because an offender could shoot as many as six persons using a typical old-fashioned six-shot revolver of the sort that has been around since the 19th century, and our goal was to identify all incidents in which it was plausible that use of an LCM (always used in connection with modern semiautomatic firearms) affected the number of casualties. It is less likely that LCMs affect the casualty count in incidents in which few people were shot, and generally fewer rounds were fired, since the rationale for banning LCMs is that they permit shooters to fire many rounds without reloading, and thereby kill or injure more victims (Koper, 2004). Thus, had the numerical cutoff been set lower, the sample of incidents would have included more cases in which LCM use was unlikely to have affected the number of victims. In that way, we have intentionally biased the sample in favor of the hypothesis that LCM use causes a higher casualty count.

We partly relied on a list compiled by the staff of the Violence Policy Center (2015) to identify LCM-involved mass shootings. Because this organization advocates bans on LCMs (Violence Policy Center, 2011), we are confident its staff were well motivated to compile as comprehensive a list as possible so as to better document the need to restrict magazine capacities. Our search of NewsBank and the other compilations of mass shootings that we cite (see Data Sources section) did not uncover any additional qualifying incidents. It is nevertheless logically impossible to know for certain that all qualifying incidents were included.

We did not employ the oft-used definition of "mass murder" as a homicide in which four or more victims were killed, because most of these involve just four to six victims (Duwe, 2007), which could therefore have involved as few as six rounds fired, a number that shooters using even ordinary revolvers are capable of firing without reloading. LCMs obviously cannot help shooters who fire no more rounds than could be fired without LCMs, so the inclusion of "nonaffectable" cases with only four to six victims would dilute the sample, reducing the percentage of sample incidents in which an LCM might have affected the number of casualties. Further, had we studied only homicides with four or more dead victims, drawn from the FBI's Supplementary Homicide Reports (SHR), we would have missed cases in which huge numbers of people were shot, and huge numbers of rounds were fired, but three or fewer of the victims died. For example, in one widely publicized shooting carried out in Los Angeles on February 28, 1997, two bank robbers shot a total of 18 people—surely a mass shooting by any reasonable standard (Table 1). Yet, because none of the people they shot died, this incident would not qualify as a mass murder (or even murder of

Table 1. Mass Shootings in Which Shooter(s) Used Magazines With a Capacity Over 10 Rounds, United States, 1994–2013.[a]

Shooter(s)	Date	Number of Shooters	Number of Guns	Number of Magazines	Capacity of Largest Magazine	Shooter(s) Reloaded?	Number of Shots Fired	Seconds Per Shot	Number Killed	Number Nonfatally Wounded
Dean Allen Mellberg	June 20, 1994	1	2	4	70	?	43–56	<6	4	23
Larry Phillips, Jr., and Emil Matasareanu	February 28, 1997	2	6	9+	100	Yes	1,101	2.40	0	18
Mitchell Johnson and Andrew Golden	March 24, 1998	2	13	3	30	?	30	?	5	11
Kip Kinkel	May 21, 1998	1	3	3+	50	Yes	51	?	2	15
Dylan Klebold and Eric Harris	April 20, 1999	2	4	16	52	Yes	188	15.64	13	21
Larry Gene Ashbrook	September 15, 1999	1	2	6	15	Yes	>100	6.00	7	7
Byran Koji Uyesugi	November 2, 1999	1	1	3	15	?	10	180.0	7	0
Michael McDermott	December 26, 2000	1	3	4+	30	Yes	37	10.54	7	0
Terry Ratzmann	March 12, 2005	1	1	3	15?	Yes	22	<2.7	7	4
Seung-Hui Cho	April 16, 2007	1	2	19	15	Yes	174	53.79	32	23
Robert Hawkins	December 5, 2007	1	1	2	30	?	>30	12.00	8	5
Steven Kazmierczak	February 14, 2008	1	4	6+	33	Yes	56	5.36	5	21
Jiverley Wong	April 3, 2009	1	2	3	30	Yes	99	?	13	4
George Sodini	August 4, 2009	1	4	3+	30	?	50	?	3	9
Nidal Hasan	November 5, 2009	1	2	15	30	Yes	214	?	13	38
Timothy Hendron	January 7, 2010	1	4	3+	Probable LCM	?	115	c. 18	3	5
Omar Thorton	August 3, 2010	1	2	4	17	?	19	9.47	8	2
Jared Loughner	January 8, 2011	1	1	4	33	No[b]	31	0.45	6	13
Eduardo Sancion	September 6, 2011	1	3	3	30	Yes	60+	1.42	4	14
James Holmes	July 20, 2012	1	4	4	100	Yes	76	4.74	12	58
Michael Page	August 5, 2012	1	1	3	19	Yes	33+	?	6	3
Andrew Engeldinger	September 27, 2012	1	1	2	15	Yes	46+	16.3	6	2
Adam Lanza	December 14, 2012	1	4	12+	30	Yes	154+	1.56	26	2

Note. Details of these incidents and citations to news accounts used as sources may be found in the appendix to an extended version of this article, with the same title, on the Social Science Research Network, at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2700166. LCM = large-capacity magazine; c = circa, i.e. approximately; ? = unknown.
[a]Number of guns is the number in the shooter's immediate possession, not necessarily the number fired. Number of magazines is the number of detachable magazines in the shooter's immediate possession. The number includes magazines in loaded semiautomatic firearms. "Seconds per shot" is the average time interval between shots through the period of shooting. [b]Shooter was prevented from reloading a defective magazine by bystanders tackling him.

any kind). Exclusion of such incidents would bias the sample against the proposition that LCM use increases the number of victims by excluding incidents with large numbers of victims.

We also excluded shootings in which more than six persons were shot over the entire course of the incident, but the shootings occurred in multiple locations with no more than six people shot in any one of the locations, and substantial periods of time intervened between episodes of shooting. An example is the series of killings committed by Rodrick Dantzler on July 7, 2011. He killed seven people and wounded two others, but did so in three different locations over a 5-hr period, shooting no more than four people in any one of the locations. Since shooters in these types of incidents have ample time to reload between sets of shots even without LCMs, use of an LCM is less likely to be relevant to the casualty counts than in a mass shooting as defined herein.

It is not possible to compare shootings involving LCMs with shootings not involving LCMs, because no source of information on shooting incidents, whether news media reports or police offense reports, systematically establishes which shootings did *not* involve LCMs. Thus, it is impossible to distinguish (a) shootings in which the perpetrator did not use an LCM from (b) shootings in which the perpetrator *did* use an LCM, but this fact was not mentioned in the account of the incident. Consequently, we are necessarily limited to describing incidents that were affirmatively identified as involving LCMs. In any case, since our purpose was to establish how often LCM use affects casualty counts in mass shootings, even if we could identify incidents that definitely did not involve LCMs, they would be irrelevant to this narrow purpose because they are obviously cases in which LCM use could not have affected casualty counts.

Data Sources

We relied on news stories to identify mass shootings and get information on their details. Relying on news outlets has obvious limits, since some mass shootings get little news coverage beyond a few stories by news outlets near the shooting location, and it is possible that none of the writers of these few stories used even one of the common words and phrases we used in our database searches. Further, even multiple news accounts of widely reported incidents may not include crucial details of the incidents, especially the number of shots fired and the duration of the shooting. Also, early news accounts of shootings are sometimes inaccurate in their details (Huff-Corzine, Corzine, Jarvis, Tetzlaff-Bemiller, Weller, & Landon, 2014), so we consulted later stories on a given incident (often pertaining to the trial of the shooter) in addition to early ones. Excluding the early news stories, we found that reported details of mass shootings were extremely consistent across stories. Fortunately, the known biases of news coverage of crime mostly work in favor of our goal of covering shootings in which many shots were fired, since news coverage is biased in favor of reporting incidents with larger numbers of victims (Duwe, 2000).

The alternative of using police reports was not feasible because such reports are not publicly available for a large share of homicides. Relying on the FBI's SHR would be even worse than news accounts for our purposes, because this source says nothing about the number of rounds fired, number of guns used, details about the guns used (beyond whether they were handguns, rifles, or shotguns), number of magazines used, or the capacity of magazines used for *any* homicide incidents, whereas news stories provide such information for many mass shootings. These same deficiencies apply to data from the FBI's National Incident-based Reporting System, which have the additional disadvantage of covering only part of the nation.

A variety of sources were used to identify eligible incidents. First, as previously noted, we consulted "Mass Shootings in the United States Involving High-Capacity Ammunition Magazines," a fact sheet compiled by the Violence Policy Center, available online at http://www.vpc.org/fact_sht/VPCshootinglist.pdf. This source only covers incidents known to involve magazines with a capacity of 10 or more rounds.

Second, we searched the NewsBank Infoweb online database which covers hundreds of print, broadcast, and online news outlets, including newspapers, news magazines, transcripts of television news programs, and online-only news providers, in every state in the nation. We searched for articles whose text (including headlines) included any of the following phrases: "mass shooting," "massacre," mass murder, "shooting spree," or "rampage" for the 20-year period from January 1, 1994, through December 31, 2013.

Third, we consulted the following existing compilations of mass shootings, mass murders, and "active shooter incidents" (and the sources they cited) to identify potentially relevant shooting incidents:

- "US Mass Shootings, 1982–2012: Data from Mother Jones' (2013) Investigation," created by the staff of *Mother Jones* magazine, available online at http://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data. This source only covers incidents in public places with four or more dead, and therefore misses those with many victims shot but three or fewer of them fatally as well as incidents occurring in private places. It also includes some spree shootings in which only a few victims were shot in any one location.
- "Analysis of Recent Mass Shootings" (September 2013), compiled by Mayors Against Illegal Guns, and available online at http://www.demandaction.org/detail/2013-09-updated-analysis-of-recent-mass-shootings. This covers incidents only for January 2009 to September 2013, and only those with four or more dead victims, thereby excluding those with many victims shot, but three or fewer shot fatally.
- Bjelopera, Bagalman, Caldwell, Finklea, and McCallion (March 18, 2013). *Public Mass Shootings in the United States: Selected Implications for Federal Public Health and Safety Policy*. Washington, DC: Congressional Research Service. This source only covers incidents occurring in public places and with four or more deaths, thereby excluding cases with many victims shot but three or fewer fatally as well as those occurring in private places.
- Citizens Crime Commission of New York City. "Mass Shooting Incidents in America (1984–2012)," at http://www.nycrimecommission.org/mass-shoot

ing-incidents-america.php, accessed January 15, 2014. This source covers shootings with four or more persons killed, with a magazine capable of holding more than 10 rounds. It excludes cases with no known use of LCMs, and incidents with many victims shot but three or fewer killed.

Notwithstanding the use of these multiple sources, we cannot be certain of achieving absolutely complete coverage of all LCM-involved mass shootings. Most of the sources rely, directly or indirectly, on news media accounts of the incidents, and some of these shootings received little coverage beyond local news outlets and perhaps an Associated Press state wire service story. The fewer news stories reporting an incident, the more likely it is that there were no stories containing any of the commonly used phrases for which we searched. The mass shootings most likely to receive little news coverage are those with fewer than four victims killed. Most of the lightly covered incidents we discovered also involved fewer than 10 victims shot, fatally or nonfatally.

On the other hand, it is unlikely that we missed many large-scale shootings, because these are likely to be well covered by multiple news outlets. Since those we missed are likely to involve fewer victims, it is also less likely that an LCM was needed for shooting as many people as were shot in these incidents. Omission of these cases, therefore, biases the sample in favor of the hypothesis that LCMs affect casualty counts.

As a check on the completeness of coverage of our methods, we used the FBI's SHRs data to identify all SHR-covered U.S. homicides that involved more than six dead victims and the use of firearms (not just those involving LCMs). These SHR data sets cover about 90% of U.S. homicides. For the period 1994–2013, we identified 17 qualifying incidents in the SHR data sets. We then checked to see if our search methods would have identified these cases. We found that searches of the NewsBank database alone identified all 17 of these incidents. Thus, shootings with many dead victims clearly are completely covered by the news media.

Once eligible incidents were identified, we searched through news accounts for details related to whether the use of LCMs could have influenced the casualty counts. Specifically, we searched for (1) the number of magazines in the shooter's immediate possession, (2) the capacity of the largest magazine, (3) the number of guns in the shooter's immediate possession during the incident, (4) the types of guns possessed, (5) whether the shooter reloaded during the incident, (6) the number of rounds fired, (7) the duration of the shooting from the first shot fired to the last, and (8) whether anyone intervened to stop the shooter.

Findings

How many mass shootings were known to have been committed using LCMs? We identified 23 total incidents in which more than six people were shot at a single time and place in the United States from 1994 through 2013 and that were known to involve use of any magazines with capacities over 10 rounds. Table 1 summarizes key details of the LCM-involved mass shootings relevant to the issues addressed in this article.

What fraction of all mass shootings are known to involve LCMs? There is no comprehensive listing of all mass shootings available for the entire 1994–2013 period, but the most extensive one currently available is the one at the Shootingtracker.com website, which only began its coverage in 2013. For 2013, this database identified 31 incidents in which more than six victims were supposedly killed or injured. This source includes deaths or injuries of perpetrators in their counts of "victim" deaths and injuries and also counts as victims' persons who were shot at, but not hit. Correcting these flaws eliminated six of the incidents as mass shootings, while another three incidents were spree shootings. Eliminating these nine ineligible incidents left 22 genuine mass shootings. The Shootingtracker database itself does not record LCM use, but examination of news media accounts indicated that none of these 22 incidents in 2013 were known to involve use of an LCM. For 2013, the Violence Policy Center (2015) identified just one shooting with more than six victims killed or injured that involved an LCM, but this incident was a spree shooting in which eight people were shot in three different widely spaced locations, with no more than three shot in any one of the locations (the June 7, 2013, incident in Santa Monica, CA). Thus, there apparently were zero mass shootings in 2013 known to involve LCMs.

To put these numbers in perspective, for the United States as a whole in 2013, there were an estimated 14,196 people killed in murders and nonnegligent manslaughters (MNNM) involving any weapon types, 9,795 of them killed with firearms (U.S. FBI, 2014b). There were an estimated 13,349 mnnm incidents,[1] of which just 3 involved more than six dead victims, 12,675 involved a single dead victim, and 13,346 involved six or fewer dead victims (U.S. Department of Justice Federal Bureau of Investigation, 2015). The 22 qualifying shooting incidents identified by Shooting Tracker as involving more than six victims therefore accounted for less than one sixth of 1% of homicide incidents and victims killed in those incidents claimed less than one tenth of 1% of homicide victims.

One might speculate that there were significant numbers of mass shootings in which LCMs were used, but not a single news account mentioned the LCM use. The use of LCMs has been a major focus of gun control advocacy groups and national news outlets since at least 1989, when a Stockton California schoolyard shooting lead to the nation's first state-level assault weapons ban (Kleck, 1997, chap. 4). In this light, it seems unlikely that LCM use in a mass shooting would go completely unreported in all news accounts, but it cannot be ruled out as a logical possibility. It is, however, irrelevant to our analyses unless shootings with unmentioned LCM use are systematically different from those that explicitly mentioned LCM use—a speculation we cannot test.

LCMs are sometimes defined as magazines holding over 10 rounds, sometimes as those holding over 15 rounds (Koper, 2004). For our entire 20-year study period of 1994–2013, 23 mass shootings were known to involve LCMs using the more inclusive cutoff of 10 rounds, that is, at least one round was fired during the incident from a gun equipped with a magazine capable of holding more than 10 rounds. Using the more stringent cutoff of more than 15 rounds, 20 incidents were known to involve LCMs.

Thus, LCM-involved mass shootings are known to have occurred an average of once per year in the United States over this 20-year period.

How often have bystanders intervened while a mass shooter was trying to reload? How many times people have disrupted a mass shooting while the shooter was trying to load a detachable magazine into a semiautomatic gun? Note that it is irrelevant whether interveners have stopped a shooter while trying to reload some other type of gun, using other kinds of magazines, since we are addressing the potential significance of restrictions on the capacity of detachable magazines that are used only with semiautomatic firearms. Thus, bystander intervention directed at shooters using other types of guns that take much longer to reload than a semiautomatic gun using detachable magazines could not provide any guidance as to the likelihood of bystander intervention when the shooter was using a semiautomatic gun equipped with detachable magazines that can be reloaded very quickly. Prospective interveners would presumably be more likely to tackle a shooter who took a long time to reload than one who took only 2- to 4-s to do so. Likewise, bystander interventions that occurred at a time when the shooter was *not* reloading (e.g., when he was struggling with a defective gun or magazine) are irrelevant, since that kind of bystander intervention could occur regardless of what kinds of magazines or firearms the shooter was using. It is the need to reload detachable magazines sooner and more often that differentiates shooters using smaller detachable magazines from those using larger ones.

For the period 1994–2013 inclusive, we identified three mass shooting incidents (with or without LCM use) in which it was claimed that interveners disrupted the shooting by tackling the shooter while he was trying to reload. In only one of the three cases, however, did interveners actually tackle the shooter while he may have been reloading a semiautomatic firearm. In one of the incidents, the weapon in question was a shotgun that had to be reloaded by inserting one shotshell at a time into the weapon (*Knoxville News Sentinel* "Takedown of Alleged Shooter Recounted" July 29, 2008, regarding a shooting in Knoxville, TN on July 27, 2008), and so the incident is irrelevant to the effects of detachable LCMs. In another incident, occurring in Springfield, OR, on May 21, 1998, the shooter, Kip Kinkel, was using a semiautomatic gun, and he was tackled by bystanders, but not while he was reloading. After exhausting the ammunition in one gun, the shooter started firing another loaded gun, one of the three firearms he had with him. The first intervener was shot in the hand in the course of wresting this still-loaded gun away from the shooter (*The (Portland) Oregonian*, May 23, 1998).

The final case occurred in Tucson, AZ, on January 8, 2011. This is the shooting in which a man named Jared Loughner attempted to assassinate Representative Gabrielle Giffords. The shooter was using a semiautomatic firearm and was tackled by bystanders, purportedly while trying to reload a detachable magazine. Even in this case, however, there were important uncertainties. According to one news account, one bystander "grabbed a full magazine" that the shooter dropped, and two others helped subdue him (Associated Press, January 9, 2011). It is not, however, clear whether this bystander intervention was facilitated because (1) the shooter was reloading or because (2) the shooter stopping firing when his gun or magazine failed to function properly. Eyewitness testimony, including that of the interveners, was inconsistent as to exactly why or how the intervention transpired in the Giffords shooting. One intervener insisted that he was sure the shooter had exhausted the ammunition in the first magazine (and thus was about to reload) because he saw the gun's slide locked back—a condition he believed could only occur with this particular firearm after the last round is fired. In fact, this can also happen when the gun jams, that is, fails to chamber the next round (Morrill, 2014; Salzgeber, 2014).

Complicating matters further, the *New York Times* reported that the spring on the second magazine was broken, presumably rendering it incapable of functioning. Their story's headline and text characterized this mechanical failure as "perhaps the only fortunate event of the day" (*New York Times* "A Single, Terrifying Moment: Shots, Scuffle, Some Luck," January 10, 2011, p. A1). If the *New York Times* account was accurate, the shooter would not have been able to continue shooting with that magazine even if no one had stopped him from loading it into his gun. Detachable magazines of any size can malfunction, which would at least temporarily stop a prospective mass shooter from firing, and thereby provide an opportunity for bystanders to stop the shooter. It is possible that the bystander intervention in the Tucson case could have occurred regardless of what size magazines the shooter possessed, since a shooter struggling with a defective small-capacity magazine would be just as vulnerable to disruption as one struggling with a defective LCM. Thus, it remains unclear whether the shooter was reloading a functioning magazine when the bystanders tackled him.

The real significance of LCM use in the Gabrielle Giffords shooting is that the first magazine that the shooter used had a capacity of 33 rounds, and the shooter fired 31 times before being tackled. Had he possessed only a 15-round magazine, and bystanders were willing to intervene when the shooter either reloaded or struggled with a defective magazine, he would have been able to fire at most 16 rounds (including one in the firing chamber)—15 fewer than the 31 he actually fired before he was stopped, for whatever reason. Consequently, instead of the 19 people he shot (6 fatally, 13 nonfatally), it would be reasonable to estimate that he would have shot only about half as many victims. Thus, the absence of an LCM might have prevented three killings and six or seven nonfatal gunshot woundings in this incident.

The bystander intervention in the Giffords shooting was, however, unique, and occurred only because there were extraordinarily courageous and quick-thinking bystanders willing and able to tackle the shooter. Over a 20-year period in the United States, the Tucson incident appears to be the only known instance of a mass shooter using a semiautomatic firearm and detachable magazines in which the shooter was stopped by bystanders while the shooter may have been trying to reload such a magazine. All other mass shootings have instead stopped only when the shooter chose to stop and left the scene, the shooter committed suicide, or armed police arrived and forced the shooter to stop (see U.S. FBI, 2014a).

The use of multiple guns and multiple magazines. Restrictions on LCMs obviously could not have affected mass shootings in which no LCMs were used, so it is just those that

Table 2. Summary of Key Characteristics of Mass Shootings (>6 Shot) With Large-Capacity Magazines, United States, 1994–2013.

	Mass Shootings With Magazines Over 10 Rounds (n = 23)			Mass Shootings With Magazines Over 15 Rounds (n = 20)		
Key Characteristics of the Incidents	Yes	No	Not Reported	Yes	No	Not Reported
Multiple guns	17 (74/74%)	6	0	15 (75/75%)	5	0
Multiple magazines	23 (100/100%)	0	0	20 (100/100%)	0	0
Both multiple guns and multiple magazines	17 (74/74%)	6	0	15 (75/75%)	5	0
Either multiple guns or multiple magazines	23 (100/100%)	0	0	20 (100/100%)	0	0
Shooter reloaded	14 (88/61%)	2	7	12 (86/60%)	2	6

Note. First number in parentheses after each frequency is the percentage of incidents with nonmissing information that had the indicated attribute. The second number in parentheses is the percentage of all incidents, including those for which the relevant information was missing, that had the indicated attribute.

involved LCMs that are relevant to judging the benefits that might have accrued had LCMs been unavailable at the beginning of the study period. As previously noted, there is considerable evidence that people who commit large-scale shootings, unlike most ordinary aggressors, devote considerable advance planning to their crimes. Part of their preparations entails cumulating multiple guns, multiple magazines, and many rounds of ammunition. The significance of this is that, in cases where the shooter has more than one loaded gun, he can continue firing, without significant pause, even without LCMs, simply by switching to a loaded gun. Alternatively, if he has multiple small magazines rather than LCMs, the shooter can continue firing many rounds with only a 2- to 4-s pause between shots for switching magazines.

Table 2 displays how often LCM-involved mass shootings involved shooters using either multiple guns or multiple magazines. Of 23 such incidents using the "more-than-10-rounds" criterion, the shooters possessed more than one gun in 17 incidents (74%), leaving six cases in which it was known that the shooter possessed just one gun. Of 20 incidents using the more-than-15-rounds criterion, the shooters possessed more than one gun in 15 incidents (75%), leaving five cases in which it was known that the shooter possessed just one gun.

Of 23 mass shootings with LCMs (>10 rounds), offenders were known to possess multiple detachable magazines in all 23 incidents (100%). Likewise, of the 20 mass shootings with magazines holding over 15 rounds, all 20 involved shooters with multiple magazines.

The average number of magazines in the immediate possession of offenders in incidents in which magazines with a capacity greater than 10 were possessed was at least 5.78 (Table 1). These offenders could have continued firing, even if they had possessed only one gun, with only the interruptions of 2–4 s that it would take for each magazine change.

In sum, there were no mass shootings in the United States in 1994–2013 known to have involved LCMs in which the shooter did not possess either multiple guns or multiple detachable magazines. In all mass shootings in which the shooters were known to have possessed one or more LCMs, the shooters could have either continued firing many rounds without any interruption at all simply by switching loaded guns or could have fired many rounds with only very brief interruptions of 2–4 s to change detachable magazines.

The offenders in LCM-involved mass shootings were also known to have *reloaded* during 14 of the 23 (61%) incidents with magazine holding over 10 rounds. The shooters were known to have *not* reloaded in another 2 of these 20 incidents, and it could not be determined if they reloaded in the remaining seven incidents. Thus, even if the shooters had been denied LCMs, we know that most of them definitely would have been able to reload smaller detachable magazines without interference from bystanders since they in fact did change magazines. The fact that this percentage is less than 100% should not, however, be interpreted to mean that the shooters were *unable* to reload in the other nine incidents. It is possible that the shooters could also have reloaded in many of these nine shootings, but chose not to do so, or did not need to do so in order to fire all the rounds they wanted to fire. This is consistent with the fact that there has been at most only one mass shooting in 20 years in which reloading a semiautomatic firearm might have been blocked by bystanders intervening and thereby stopping the shooter from doing all the shooting he wanted to do. All we know is that in two incidents, the shooter did not reload, and news accounts of seven other incidents did not mention whether the offender reloaded.

Do more magazine changes allow more prospective victims to escape? An alternative rationale for why limiting aggressors to smaller magazines would result in fewer casualties in mass shootings is that the increased number of magazine changes necessitated by use of smaller magazines would create additional pauses in the shooting, allowing more potential victims to escape than would otherwise escape. For example, a story in the *Hartford Courant* about the Sandy Hook elementary school killings in 2012 was headlined "Shooter Paused, and Six Escaped," the text asserting that as many as six children may have survived because the shooter paused to reload (December 23, 2012). The author of the story, however, went on to concede that this was just a speculation by an unnamed source, and that it was also possible that some children simply escaped when the killer was shooting other children. There was no reliable evidence that the pauses were due to the shooter reloading, rather than his guns jamming or the shooter simply choosing to pause his shooting while his gun was still loaded.

The plausibility of the "victims escape" rationale depends on the average rates of fire that shooters in mass shootings typically maintain. If they fire very fast, the 2–4 s it takes to change box-type detachable magazines could produce a slowing of the rate of fire that the shooters otherwise would have maintained without the magazine changes, increasing the average time between rounds fired and potentially allowing more victims to escape during the between-shot intervals. On the other hand, if mass

Table 3. Known Rates of Fire in Mass Shootings, 1994–2013.

Date of Incident	Shots Fired[a]	Time of Firing (Minutes)[a]	Average Shots Per Minute	Average Seconds Per Shot	Number of Guns
June 20, 1994	>50	c. 5	>10	<6.0	2
February 28, 1997	1,101	44	25	2.4	4
April 20, 1999	188	49	3.8	15.8	4
September 15, 1999	>100	10	>10.0	<6.0	2
September 2, 1999	10	<30	>0.33	<180.0	1
May 24, 2000	c. 7	<90	>0.08	<771.4	1
September 22, 2000	9+	<10	>0.9	<66.7	1
December 26, 2000	37	5–8 (6.5)	5.7	10.5	3
February 5, 2001	25–30 (27.5)	8–15 (11.5)	2.4	25.1	4
March 5, 2001	c. 24	6	c. 4.0	c. 15.0	1
March12, 2005	22	<1	>22.0	<2.7	1
March 21, 2005	45	9	5.0	12.0	3
March 25, 2006	9+	c. 5	>1.6	<33.3	2
October 2, 2006	17–18 (17.5)	c. 2	c. 8.75	c. 6.9	2
April 16, 2007	c. 174	156	c. 1.11	c. 53.8	2
October 7, 2007	30	c. 1	c. 30.0	c. 2.0	3
December 5, 2007	>30	c. 6	>5.0	<12.0	1
February 14, 2008	56	5	11.1	5.4	4
January 7, 2010	115	30	3.8	15.7	4
August 3, 2010	19	3	6.3	9.5	2
January 8, 2011	31	0.25	125	0.48	1
September 6, 2011	60+	1.42	42.3+	1.4	3
July 20, 2012	76	c. 6	12.7	4.74	4
September 27, 2012	46+	14	>3.3	<18.3	1
December 14, 2012	154+	4	38.5+	1.6	3

Note. c = circa.
[a]Where a range was provided in news accounts, the midpoint of the range (shown in parentheses) of shots fired or time of firing was used in rate-of-fire computations.

shooters fire their guns with the average interval between shots lasting *more* than 2–4 s, the pauses due to additional magazine changes would be no longer than the pauses the shooter typically took between shots even when not reloading. In that case, there would be no more opportunity for potential victims to escape than there would have been without the additional magazine changes.

Table 3 displays data on rates of fire for LCM-involved mass shootings in 1994–2013. Information on both the duration of the firing and the number of rounds fired was available for 17 of the 23 incidents shown in Table 1 plus another 8 mass shootings for which the necessary information was available but that did not involve any known LCM use. Reliable information on duration of fire may well be unavailable from any source for many mass shootings. There are rarely audio recordings that would provide precise information on the duration of fire (as there were in the 2012 Aurora Colorado movie theater shooting), so eyewitness estimates are usually the basis for establishing this. On the other hand, there is often quite reliable information on the number of rounds fired, since semiautomatic firearms eject an empty shell casing after each round is fired. When shooters use such guns, crime scene investigators can (absent removal of the evidence by the offender or souvenir hunters) establish the number of rounds fired by counting cartridge casings recovered at the scene.

Average rate of fire was computed as the average number of seconds between shots. In the 25 incidents for which average rates of fire could be determined, shooters never maintained an average rate of fire anywhere as fast as that at which their firearms were capable of firing. Shooters firing as fast as the gun allows can easily fire three rounds per second with a typical semiautomatic firearm, that is, with only about one third of a second between rounds. In only three incidents were mass shooters known to have averaged less than *2* s between rounds. This is no more than one sixth of the maximum rate of fire of which semiautomatic guns are capable (see Table 3, incidents occurring on January 8, 2011, September 6, 2011, and December 14, 2012). This means that taking 2 s to reload a detachable magazine would not have slowed the shooters' average rate of fire at all in 22 of the 25 incidents for which rate of fire could be established and would have only slightly slowed the rate in the remaining three incidents.

It cannot be assumed, however, that in the three incidents in which usually high rates of fire were maintained, use of smaller magazines would have slowed the rate of fire due to a need to change magazines more often. Shooters possessed multiple guns in two of these three relatively rapid fire incidents (those occurring on September 6, 2011 and December 13, 2012), which means that, rather than needing to change magazines to continue shooting, the aggressors could simply have switched guns, from one firearm emptied of rounds to another loaded firearm, without pausing in their shooting at all. Over the 20-year study period, there was just one LCM-involved mass shooting incident in the United States in which a shooter maintained an average rate of fire with less than 2 s elapsing between shots, *and* possessed only a single gun—the shooting involving Jared Loughner (on January 8, 2011), who was stopped from further shooting when he was tackled by bystanders.

In sum, in nearly all LCM-involved mass shootings, the time it takes to reload a detachable magazine is no greater than the average time between shots that the shooter takes anyway when not reloading. Consequently, there is no affirmative evidence that reloading detachable magazines slows mass shooters' rates of fire, and thus no affirmative evidence that the number of victims who could escape the killers due to additional pauses in the shooting is increased by the shooter's need to change magazines.

Conclusions

In light of the foregoing information, it is unlikely that the larger number of rounds fired in the average LCM-linked mass shooting found by Koper (2004) was in any sense caused by the use of LCMs. In all but one of such cases in the period from 1994 through 2013, there was nothing impossible or even difficult about the shooter firing equally large numbers of rounds even if he had possessed only smaller capacity magazines, since the same number of rounds could easily have been fired with smaller detachable magazines of the sort that would remain legally available under LCM bans. Instead, the larger number of rounds fired by LCM-using shooters is more likely to reflect the more lethal intentions prevailing among such shooters, just as their planned use of multiple guns and multiple magazines, and the unusually high fatality rate (deaths over total woundings) of their attacks are outward indications of a desire to shoot many people. Unfortunately, there are no known methods for reliably measuring the lethality of shooters' intentions independent of the outcomes of their crimes, making it impossible to statistically control for this factor in a multivariate statistical analysis and thereby isolate the effects of LCM use.

One cannot prove a negative, and it is possible that mass shooters in the future might be different from those in the past, and that would-be mass shooters, unlike those of the past, would not obtain multiple guns or multiple smaller capacity magazines as substitutes for LCMs. One might also speculate that incidents that did *not* end up with many shooting victims turned out that way because the shooter did *not* use an LCM. At this point, however, there is little sound affirmative empirical basis for expecting that fewer people would be killed or injured if LCM bans were enacted.

Focusing gun control efforts on mass shootings makes sense from a political standpoint, since support for gun control is elevated following highly publicized gun crimes. Such efforts, however, are less sensible for purposes of reducing the death toll from gun violence, especially if they focus on technologies rarely used in gun crime as a whole. Controls aimed at reducing ordinary forms of firearm violence, such as shootings with just one or a few victims, are more likely to have large impacts on the aggregate gun violence death toll for the simple reason that nearly all victims of gun violence are hurt in incidents with a small number of victims. For example, less than 1% of U.S. homicide incidents in 2013 involved more than two victims killed (U.S. Department of Justice Federal Bureau of Investigation, 2015).

Most types of gun control focus on preventing more dangerous people from acquiring, possessing, or using *any* type of gun, and therefore have potential to prevent a wide array of gun crimes. A prime example is a law requiring background checks on persons seeking to buy guns. Gun laws with a background check component, such owner license and purchase permit laws, have been found to be potentially effective in reducing homicide (Kleck & Patterson, 1993, p. 274). There is already a federal law requiring background checks, but it only applies to purchases from licensed gun dealers. Extending these checks to cover private gun transfers—that is, implementing a federal universal background check (Kleck, 1991, pp. 433–435)—is far more likely to prevent significant numbers of gun crimes than measures aimed at rarely used gun technologies like LCMs and extremely rare types of violent incidents like mass shootings.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplementary Material

The online appendices are available at http://journals.sagepub.com/doi/suppl/10.1177/1525107116674926

Note

1. Supplementary Homicide Reports (SHR) data for 2013 indicate that there were an average of 1.063 victims per SHR-covered homicide incident, implying 13,349 incidents.

References

Bjelopera, J. P., Bagalman, E., Caldwell, S. W., Finklea, K. M., & McCallion, G. (2013, March 18). *Public mass shootings in the United States: Selected implications for Federal Public Health and Safety Policy*. Washington, DC: Congressional Research Service.

Brady Campaign to Prevent Gun Violence. (2013). 2013 State Scorecard. Retrieved from http://www.bradycampaign.org/2013-state-scorecard

Citizens Crime Commission of New York City. (2013). Mass shooting incidents in America (1984–2012). Retrieved January 1, 2014, from http://www.nycrimecommission.org/mass-shooting-incidents-america.php

Cook, P. (1985). The case of the missing victims: Gunshot woundings in the National Crime Survey. *Journal of Quantitative Criminology*, *1*, 91–102.

Duwe, G. (2000). Body-count journalism. *Homicide Studies*, *4*, 364–399.

Duwe, G. (2007). *Mass murder in the United States*. Jefferson, NC: McFarlane.

Huff-Corzine, L., Corzine, J., Jarvis, J. P., Tetzlaff-Bemiller, M. J., Weller, M., & Landon, M. (2014). Shooting for accuracy: Comparing data sources on mass murder. *Homicide Studies*, *18*, 105–124.

Kleck, G. (1984). Handgun-only gun control. In D. B. Kates, Jr (Ed.), *Firearms and violence: Issues of public policy* (pp. 167–199). San Francisco, CA: Pacific Institute.

Kleck, G. (1991). *Point blank*. NY: Aldine de Gruyter.

Kleck, G. (1997). *Targeting guns*. NY: Aldine de Gruyter.

Kleck, G., & Patterson, E. B. (1993). The impact of gun control and gun ownership levels on violence rates. *Journal of Quantitative Criminology*, *9*, 249–287.

Koper, C. S. (2004). *An updated assessment of the federal assault weapons ban: Impacts on gun markets and gun violence, 1994–2003*. Philadelphia: University of Pennsylvania.

Mayors Against Illegal Guns. (2013). *Analysis of recent mass shootings*. Retrieved from http://www.demandaction.org/detail/2013-09-updated-analysis-of-recent-mass-shootings

Morrill, L. (2014). Statement of Colorado Assistant Attorney General Leann Morrill, stipulating that the slide on a Glock 19 pistol can ‘‘stick back’’ due to a jam. P. 1832, transcript of ninth day of testimony, April 10, 2014, in Colorado Outfitters Association, et al., v. John W. Hickenlooper, United States District Court for the District of Colorado Civil Action No. 13-Cv-1300-Msk-Mjw. Retrieved from http://coloradoguncase.org

Mother Jones. (2013). *US mass shootings, 1982–2012: Data from Mother Jones' investigation*. Retrieved from http://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data

Reedy, D. C., & Koper, C. S. (2003). Impact of handgun types on gun assault outcomes: A comparison of gun assaults involving semiautomatic pistols and revolvers. *Injury Prevention*, *9*, 151–155.

Salzgeber, R. (2014). Trial testimony of Roger Salzgeber, April 8, 2014, in Colorado Outfitters Association, et al., v. John W. Hickenlooper, The United States District Court for the District of Colorado Civil Action No. 13-Cv-1300-Msk-Mjw. Retrieved from http://coloradoguncase.org/

U.S. Department of Justice Federal Bureau of Investigation. (2015, April 24). *Uniform crime reporting program data: Supplementary Homicide Reports, 2013* (ICPSR36124-v1). Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor]. doi:10.3886/ICPSR36124.v1

U.S. Federal Bureau of Investigation. (2014a). *A study of active shooter incidents in the United States between 2000 and 2013*. Retrieved March 2, 2016, from https://www.fbi.gov/about-us/office-of-partner-engagement/active-shooter-incidents/a-study-of-active-shooter-incidents-in-the-u.s.-2000-2013

U.S. Federal Bureau of Investigation. (2014b). *Crime in the United States—Uniform Crime Reports*. Retrieved March 1, 2016, from https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/expanded-homicidedata/expanded_homicide_data_table_8_murder_victims_by_weapon_2010-2014.xls

U.S. Federal Bureau of Investigation. (2015). *2014 Crime in the United States*. Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/cius-home

Violence Policy Center. (2011). *High-capacity ammunition magazines: The common thread that runs through mass shootings*. Press release dated January 11, 2011. Washington, DC: Author.

Violence Policy Center. (2015). *Mass shootings in the United States involving high-capacity ammunition magazines*. Retrieved from http://www.vpc.org/fact_sht/VPCshootinglist.pdf

Virginia Tech Review Panel. (2007). *Mass shootings at Virginia Tech*. Retrieved from http://scholar.lib.vt.edu/prevail/docs/April16ReportRev20091204.pdf

Author Biography

Gary Kleck is the Emeritus David J. Bordua Professor of Criminology and Criminal Justice at Florida State University, having retired after 38 years at FSU. He has won the Michael J. Hindelang Award for Point Blank, testified to Congress and state legislatures on gun control, and served on numerous national task forces and panels. He is currently completing a book, with Brion Sever, on the effects of legal punishment on crime.

J Urban Health (2018) 95:313–321
DOI 10.1007/s11524-017-0205-7

Criminal Use of Assault Weapons and High-Capacity Semiautomatic Firearms: an Updated Examination of Local and National Sources

Christopher S. Koper · William D. Johnson · Jordan L. Nichols · Ambrozine Ayers · Natalie Mullins

Published online: 2 October 2017

Abstract Policies restricting semiautomatic assault weapons and large-capacity ammunition magazines are intended to reduce gunshot victimizations by limiting the stock of semiautomatic firearms with large ammunition capacities and other military-style features conducive to criminal use. The federal government banned such weaponry from 1994 to 2004, and a few states currently impose similar restrictions. Recent debates concerning these weapons have highlighted their use in mass shootings, but there has been little examination of their use in gun crime more generally since the expiration of the federal ban. This study investigates current levels of criminal activity with assault weapons and other high-capacity semiautomatics in the USA using several local and national data sources including the following: (1) guns recovered by police in ten large cities, (2) guns reported by police to federal authorities for investigative tracing, (3) guns used in murders of police, and (4) guns used in mass murders. Results suggest assault weapons (primarily assault-type rifles) account for 2–12% of guns used in crime in general (most estimates suggest less than 7%) and 13–16% of guns used in murders of police. Assault weapons and other high-capacity semiautomatics together generally account for 22 to 36% of crime guns, with some estimates upwards of 40% for cases involving serious violence including murders of police. Assault weapons and other high-capacity semiautomatics appear to be used in a higher share of firearm mass murders (up to 57% in total), though data on this issue are very limited. Trend analyses also indicate that high-capacity semiautomatics have grown from 33 to 112% as a share of crime guns since the expiration of the federal ban—a trend that has coincided with recent growth in shootings nationwide. Further research seems warranted on how these weapons affect injuries and deaths from gun violence and how their regulation may impact public health.

Keywords Firearms · Assault weapons · Violence

C. S. Koper (✉) · W. D. Johnson · J. L. Nichols · A. Ayers · N. Mullins
Center for Evidence-Based Crime Policy, Department of Criminology, Law and Society, George Mason University, Fairfax, VA, USA
e-mail: ckoper2@gmu.edu

Introduction

Firearm violence imposes a significant burden on public health in the USA. From 2010 through 2012, the nation experienced an annual average of 11,256 firearm homicides and 48,534 non-fatal assault-related gunshot victimizations that cost society nearly $22 billion a year in lifetime medical and work-related costs [1]. One type of policy response to reduce gun violence involves restricting or mandating design changes in particular types of firearms that are considered to be especially dangerous and/or attractive for criminal use.

Restrictions on assault weapons (AWs) represent one particularly controversial and highly contested form of such legislation that has featured prominently in gun policy debates in recent decades. In general, AW laws restrict manufacturing, sales, and ownership of semiautomatic firearms with large ammunition capacities and other military-style features that appear useful in military and criminal applications but unnecessary in shooting sports or self-defense [2]. Examples of such features include pistol grips on rifles, flash hiders, folding rifle stocks, threaded barrels for attaching silencers, and barrel shrouds on pistols. AW laws also commonly include restrictions on large-capacity magazines (LCMs), which are typically defined as ammunition feeding devices holding more than ten rounds of ammunition (some laws have higher limits). LCM restrictions are arguably the most important components of AW laws in that they also apply to the larger class of high-capacity semiautomatic firearms without military-style features. In the broadest sense, AW-LCM laws are thus intended to reduce gunshot victimizations by limiting the stock of semiautomatic firearms with large ammunition capacities and other features conducive to criminal use. The federal government enacted a national ban on AWs and LCMs in 1994 but allowed it to expire in 2004. Currently, eight states and the District of Columbia have AW and/or LCM restrictions, as do some additional localities [3].

Recent discussion and debates concerning these weapons have largely focused on their use in mass shootings. However, there has been little examination of the use of AWs and LCMs in gun crime more generally since the expiration of the federal ban. Studies conducted around the time of the federal ban found that AWs accounted for up to 8% of guns used in crime (generally between 1 and 6% and averaging around 2%) and that the broader class of firearms equipped with LCMs (including AWs and other semiautomatic firearms equipped with LCMs) accounted for up to a quarter [2, 4–12]. Criminal use of such weaponry declined during the years of the federal ban [2, 13, 14], but trends since then have only been examined in the state of Virginia, where LCM use rose following the ban's expiration [14]. Semiautomatic weapons with LCMs and/or other military-style features are common among models produced in the contemporary gun market [15, 16], but precise estimates of their production and ownership are unavailable. Growth in the use of such weapons could have important implications for public health as these weapons tend to produce more lethal and injurious outcomes when used in gun violence [2, 17]. This study provides an updated examination of the AW issue by investigating current levels of criminal activity with AWs and other LCM firearms as measured in a variety of national and local data sources.

Data and Methods

There is no national data source that can be used to count the numbers of homicides, non-fatal shootings, or other crimes committed with AWs and other LCM firearms. Therefore, criminal use of these weapons was approximated by examining and triangulating across several local and national data sources on guns used in different types of crimes.

Local Data Sources

The local-level analyses are based on guns recovered by police over multiple years (defined below) in a convenience sample of ten cities including Hartford (CT), Rochester (NY), Syracuse (NY), Baltimore (MD), Richmond (VA), Minneapolis (MN), Milwaukee (WI), Kansas City (MO), Seattle (WA), and Sacramento (CA). Large cities were selected for the analysis (these cities range in size from roughly 124,000 to 684,500) due to the concentration of gun violence in urban areas [18, 19]. Patterns and trends in these particular cities may not be indicative of those elsewhere; further, some (Baltimore, Hartford, Rochester, Syracuse, and Sacramento) are covered by state AW and LCM restrictions that were in effect during all or portions of the study period (this study does not attempt to evaluate the implementation and effects of these laws or variations therein). Nonetheless, these cities constitute a geographically diverse set of ban and non-ban locations, thus strengthening generalizations. The data were obtained from law enforcement authorities in these jurisdictions except where otherwise noted. Information available in most of the police databases included the type, make, model, and caliber of each confiscated firearm; the date when it was recovered; and the type of crime with which it was associated.

Guns recovered by police (often referred to as "crime guns") are the only readily available data with which to study patterns and trends in the types of guns used in crime across jurisdictions, and they are commonly used in research on gun markets, gun violence, and gun policy [2, 9, 20–37]. Guns confiscated by police include guns recovered in violent crime investigations as well as those recovered in connection with weapon offenses (illegal possession, carrying, and discharges), drug violations, property crimes, and other incidents. These samples thus represent guns known to have been used in violence as well as guns possessed and/or carried by criminal and otherwise high-risk persons. As others have noted, they represent a sample from the population of guns that are at greatest risk of misuse [24] and thereby provide a probable sample of guns used to commit crimes [21]. As caveats, nonetheless, it should be noted that police do not recover all guns used and possessed illegally, and it is possible that the types of guns they confiscate differ from those of unrecovered guns linked to illegal possessors and users. The analyses highlighted below are based on all confiscated firearms in the study jurisdictions. Additional analyses conducted with just those guns clearly connected to a violent offense, which represented at least 13 to 19% of guns across the cities, produced very similar results except where noted (separate offense-type analyses could not be conducted with the Syracuse and Rochester gun data or the Richmond LCM data).

National Data Sources

National-level analyses were conducted using three data sources and compilations. The first consists of information on firearms recovered by law enforcement agencies throughout the nation and reported to the federal Bureau of Alcohol, Tobacco, Firearms, and Explosives (ATF) for investigative tracing of their sale histories. Guns reported to ATF provide a national sample of crime guns numbering in the hundreds of thousands annually (predominantly from urban jurisdictions), but they do not constitute a statistically representative sample for the nation given that gun tracing is voluntary (agencies trace guns as needed for specific investigations and/or analysis of illegal gun markets) and varies between agencies and over time [24, 27, 38–40]. Further, publicly available data on traced guns are limited to aggregate figures on basic types and calibers of the weapons, thus limiting the analyses that could be conducted as described below. The other national data sources included information on guns used in murders of police officers and mass murder incidents. Prior research has shown that AWs and LCM firearms are used in a higher share of these crimes, due presumably to their lethality and attractiveness to the types of offenders who commit these offenses [2, 4], and this has been a prominent issue in the AW debate. Information on firearms used in murders of police, including the type, make, model, and caliber of each weapon, was obtained from the Federal Bureau of Investigation (FBI), which compiles these data from reports by police agencies throughout the country. Information on firearms used in mass murder shooting incidents was collected from lists and reports compiled by several organizations since there is no single official data source that regularly provides detailed and comprehensive information on mass murders and the guns used in these incidents [41–50]. Consistent with many prior studies of this issue, firearm mass murders were defined as incidents in which four or more people were murdered with a firearm, not including the death of the shooter if applicable and irrespective of the number of additional victims shot but not killed. This increased the number of sources that could be used to gather information. As described below, however, detailed weapon information could not be found in public sources for many of the cases.

Methods

There is no universal definition of an AW that applies across current and past AW laws. For example, the expired federal ban and some current state laws define AWs as having two military-style features, whereas other state bans and a recent (2013) proposal for a new federal ban use a one feature criterion [2, 51]. For this study, AWs were defined based on the weapons that have most commonly been identified as such based on the old federal ban, current state laws, and the recently proposed federal ban. This list included more than 200 make-model combinations covered by either of the federal lists (2004 and 2013) or at least two of the state laws. Based on preliminary analyses showing that most recovered AWs are assault rifles (as opposed to assault pistols or assault shotguns), an additional ceiling estimate of AW use was calculated based on the prevalence of semiautomatic rifles. This was also done to compensate for imprecision in the AW estimates (due, for example, to missing or partial gun model data, lack of information about the specific features or configurations of the weapons that could affect their AW status, and possible omissions from the operational AW list).

Use of guns with LCMs could only be measured precisely for the Syracuse, Baltimore, and Richmond analyses, which are based on data sources having an indicator for magazine capacity (which is typically missing from police gun databases), and some of the mass murder incidents. For most analyses, use of LCM firearms was approximated based on recoveries of semiautomatics that are commonly manufactured and sold with LCMs, referred to below as LCM-compatible firearms. Identification of these models was based on gun catalogs (such as the *Blue Book of Gun Values* and *Gun Digest*) and examination of gun manufacturers' websites. This method likely overstates LCM use to some degree since many LCM compatible firearms can also be equipped with smaller magazines. As a rough guide, inspection of all recoveries of a small number of LCM-compatible handgun models in the Baltimore data revealed that approximately four of five were equipped with LCMs. Conversely, LCM use can also be undercounted for guns that were missing complete model information or equipped with aftermarket LCMs, which are available for some guns not sold with LCMs at retail. LCM use was not estimated for Rochester and Sacramento since New York and California have had longstanding restrictions on magazines with more than ten rounds (hence, it seems less likely that LCM-compatible guns recovered in those jurisdictions were actually equipped with LCMs).

Data were collected from 2014 through 2016. Current estimates of AW and LCM use were developed using the most recent 2–3 years of data from the local police databases and ATF data. Data spanning the most recent 5–6 years were used to generate contemporary estimates of AW and LCM use in murders of police and mass murders due to the rarity of these events. As described below, some data sources were also used to estimate trends in the use of semiautomatic rifles and LCM firearms since the expiration of the federal ban. Reported figures highlight AWs and LCM firearms as a share of crime guns in order to control for differences in the volume of gun crime and overall gun recoveries between places and over time. Other noteworthy aspects of the data and analyses are discussed below.

Results

Local Analyses

Results of the local analyses are presented in Table 1. For each site, estimates are based on data spanning different portions of the 2011–2014 period. The number of guns analyzed ranged from 281 in Syracuse to 4994 in Kansas City and totaled 21,551 across all data sources.

Estimates of the prevalence of AWs among crime guns ranged from a low of 2.4% in Baltimore to a high of 8.5% in Syracuse. Assault rifles (e.g., variations of the AR-15 or AK-47) accounted for the majority of AWs in all sites and more than three-quarters in all but one (Richmond). The remaining AWs consisted entirely (or nearly so) of assault pistols (e.g., the TEC-9 or TEC-22). The share of crime guns consisting of semiautomatic rifles of any sort is also displayed in Table 1 for localities that had gun databases with gun-type designations (i.e., handgun/rifle/shotgun, semiautomatic/non-semiautomatic). These estimates ranged from a low of 4.1% in Hartford to 12.4% in Rochester but were less than 9% for most cities. (The Milwaukee estimate is based on the percentage of crime guns that were rifles of any sort as semiautomatic/non-semiautomatic designations were unavailable.) As noted, the semiautomatic rifle estimates, which include both AW-type and non-AW-type rifles, provide a likely ceiling for estimates of AW prevalence.

The percentage of crime guns clearly equipped with an LCM (including AWs and other high-capacity semiautomatics, most of which are pistols) was 16.5% in Baltimore during the 2012–2014 period, but this figure rose to 21.5% for guns that were connected to a violent crime. These findings are similar to those from a recent news report (involving a separate and independent analysis of Baltimore data) indicating that 18.4% of guns recovered in Baltimore had LCMs for the period of 2010 through 2016 [52]. In Richmond, 22% of crime guns were equipped with LCMs during 2008 and 2009 based on data collected by the Virginia State Police and initially reported by *The Washington Post* [14] (the *Post's* reported figures have been reanalyzed here to focus on the most recent available years and to assess trends). Crime guns were least likely to be equipped with LCMs in Syracuse (14.6%), where New York State LCM restrictions have been in effect since the early 2000s.

For the other sites, the prevalence of LCM-compatible guns ranged from 22.2% in Hartford to 36.2% in both Kansas City and Seattle, with the majority of the estimates (3 of 5) higher than one-third. In most of these cities, the prevalence of LCM guns was similar whether focusing on all guns or those connected to a violent crime. In Hartford, however, 30% of violent crime guns were LCM compatible in contrast to 22.2% for all guns. Further, a supplemental analysis of guns linked to assault-

Table 1 Prevalence of assault weapons, semiautomatic rifles, and semiautomatics with large-capacity magazines among guns recovered by police: estimates for selected cities and years

Location and sample	Assault weapons as % of guns	Semiautomatic rifles as % of guns	Semiautomatics with large-capacity magazines as % of guns
Hartford, CT (2011–2012, *N* = 854)	2.6%	4.1%	22.2% overall, 30% for guns linked to violent crime
Rochester, NY (2012–July 2014, *N* = 1687)	4.9%	12.4%	Not estimated
Syracuse, NY (2012–May 2014, *N* = 281)	8.5%	12.1%	14.6%
Baltimore, MD (2012–Sep. 2014, *N* = 4680)	2.4%	5.4%	16.5% overall, 21.5% for guns linked to violent crime
Richmond, VA (AW analysis: 2012–2013, *N* = 1180) (LCM analysis: 2008–2009, *N* = 1960)	2.7%	Not estimated	22.0%
Minneapolis, MN (2012–Aug. 2014, *N* = 2178)	3.4%	6.4%	25.1% overall, 46.3% for guns linked to shootings
Milwaukee, WI (Jul. 2013–Jun. 2014, *N* = 1868)	4.6%	< 9.4%	35.5%
Kansas City, MO (2012–Aug. 2014, *N* = 4994)	6.1%	6.3%	36.2%
Seattle, WA (2012–July 2014, *N* = 596 guns linked to violent crimes or weapons violations)	6.4%	7.9%	36.2%
Sacramento, CA (Aug. 2013–Jul. 2014, *N* = 1273)	6.0%	Not estimated	Not estimated

Estimates are based on general gun recovery samples except where noted. Estimates were similar for guns known to have been connected to violent crimes except where noted. Large-capacity magazine (LCM) estimates for Syracuse, Baltimore, and Richmond are based on known LCM recoveries (the Richmond estimates are based on Virginia State Police data initially reported by *The Washington Post*). Other LCM estimates are based on recoveries of LCM compatible firearm models. The Milwaukee semiautomatic rifle estimate is based on the prevalence of all rifles

related shootings in Minneapolis (using gunshot victimization data provided by Minneapolis police) revealed that 46.3% were LCM compatible, though this was based on a small sample (*n* = 80 guns).

National Analyses

Results of the national analyses are presented in Table 2. AW prevalence was approximated in the national ATF tracing data for 2012 and 2013 (*n* = 481,632) based on traces of guns in calibers .223, 5.56, and 7.62 mm. These are common calibers for AW-type semiautomatic rifles, though not all firearms in these calibers are AWs, and not all AWs fall into these calibers. This method nonetheless yielded an estimate of 5%, which is within the range of estimates provided by the local analyses. Further estimates of semiautomatic rifles and LCM firearms were not possible given the limitations of published tracing data.

Guns used in murders of police were analyzed for the years 2009 through 2013 (*n* = 219, excluding cases involving the officers' own weapons, which are often LCM firearms). AWs accounted for an estimated 13.2% of the firearms used in these crimes overall and varied between 8 and 18% from year to year. Virtually all of the AWs (97%) were assault rifles. Semiautomatic rifles overall accounted for 15.5% of the firearms used in these cases and ranged from 5 to 23% annually. LCM-compatible firearms more generally constituted 40.6% of the murder weapons, ranging from 35 to 48% annually.

AW and LCM use in firearm mass murders was examined for a sample of 145 incidents that occurred from 2009 through 2015 but could only be estimated within broad ranges due to high levels of missing weapons data in public accounts. AWs were used in at least 10.3% of these incidents. However, only 42 incidents had sufficiently detailed weapon information to make a definitive determination regarding AW use; among these cases, 35.7% involved AW use. All but one AW case involved an assault rifle. (A separate estimate for semiautomatic rifle use is not presented because only two additional cases clearly involved a semiautomatic rifle with an unclear or non-AW designation.) LCM firearms overall were involved in at least 18.6% of the incidents based on cases that involved clear possession of LCMs, AWs, or other LCM-compatible models. Although many additional cases involved semiautomatic firearms, an LCM coding could

Table 2 Prevalence of assault weapons, semiautomatic rifles, and semiautomatics with large-capacity magazines among national samples of guns recovered by police, guns used in murders of police, and guns used in mass murders

Data source and sample	Assault weapons as % of guns	Semiautomatic rifles as % of guns	Semiautomatics with large-capacity magazines as % of guns
Federal Bureau of Alcohol, Tobacco, Firearms, and Explosives (ATF): guns recovered by police and reported to ATF for investigative tracing (2012–2013, $N = 481{,}632$)	5%	Not estimated	Not estimated
Federal Bureau of Investigation: guns used in murders of police (2009–2013, $N = 219$)	13.2%	15.5%	40.6%
Public reports of firearm mass murders (4+ killed) (2009–2015, $N = 145$)	10.3–35.7%	Not estimated	18.6–57.4%

Assault weapon estimate for ATF data is based on reported firearms in calibers .223, 5.56, and 7.62 mm. LCM estimates are based on recoveries of LCM compatible firearm models in the FBI data and recoveries of both LCMs and LCM compatible firearms in the mass murder data

only be made for 47 cases, 57.4% of which involved an LCM firearm. The identified AW and LCM cases typically occurred in public locations (80%) and resulted in more than twice as many people shot on average as did other incidents (13.7 victims on average for AW-LCM cases versus 5.2 for other cases; t test p level < 0.01).

Trend Analyses

Trends in the use of AWs and LCM firearms since the end of the federal AW ban or the early post-ban years were also estimated using selected data sources that had sufficiently detailed weapon information and spanned the period of interest. First, trends in recoveries of semiautomatic rifles were used to approximate trends in crime with AWs using the FBI national data on police murders (2003–2013) and data from the following cities and time periods: Baltimore (2004–2014), Rochester (2004–2014), Syracuse (2004–2014), Milwaukee (2006–2014, based on all rifles), Seattle (2008–2014), Minneapolis (2006–2014), and Kansas City (2008–2014). In summary, these analyses (not shown) revealed little evidence of upward trends in the use of semiautomatic rifles across sites.

Second, trends in crimes with LCM firearms were estimated based on guns used in murders of police (2003–2013) as well as guns recovered in Baltimore (2004–2014), Richmond (2003–2009), and Minneapolis (2006–2014). Table 3 shows changes over time in the percentage of guns that were LCM firearms using the earliest and latest years of each data source. In relative terms, the prevalence of LCM firearms increased from 33 to 49% in the Baltimore, Minneapolis, and national (FBI) data (note that Maryland restricted LCMs with more than 20 rounds throughout this period and extended these restrictions to LCMs with more than 10 rounds in late 2013). The largest increase occurred in Richmond, where LCM firearms increased 111.5%, rising from 10.4% of recovered guns in 2003–2004 (the final years of the federal AW ban) to 22% in 2008–2009. Similar trends have also been reported for the state of Virginia overall [14]. All of these changes were statistically significant ($p < 0.05$) based on chi-square tests of the equality of proportions.

Discussion

Subject to caveats noted above, this examination of several national and local data sources suggests that AWs are used in between 2 and 9% of gun crimes in general with most estimates being less than 7%. Upper bound estimates of AW use based on semiautomatic rifles range from 4 to 12% in most data sources and are typically less than 9%. These estimates are broadly similar to those generated in the early 1990s prior to the federal AW ban [2], though they are perhaps somewhat higher on average. However, comparisons of these estimates with others should be made cautiously, as operational definitions of an AW have varied across studies and estimates presented here are based on the most contemporary definitions of AWs. One clearly notable

Table 3 Changes in prevalence of semiautomatics with LCMs: estimates for selected local and national data sources and time frames, 2003–2014

Data source/location	LCM firearm prevalence: early time period	LCM firearm prevalence: late time period	Change in LCM firearm prevalence
Baltimore crime guns	11.1% (2004, 2006, $N = 5369$ total firearms)	16.5% (2012–Sep. 2014, $N = 4381$ total firearms)	+ 48.6%**
Richmond, VA crime guns	10.4% (2003–2004, $N = 2413$ total firearms)	22.0% (2008–2009, $N = 1960$ total firearms)	+ 111.5%**
Minneapolis crime guns	16.8% (2006–2007, $N = 2564$ total firearms)	25.1% (2012–Aug. 2014, $N = 2178$ total firearms)	+ 49.4%**
National (FBI): guns used in murders of police	30.4% (2003–2007, $N = 224$ total firearms)	40.6% (2009–2013, $N = 219$ total firearms)	+ 33.6%*

Change in proportions statistically significant at $p < 0.05$ (*) or $p < 0.01$ (**)

Estimates are based on general gun recovery samples except where noted. LCM estimates for Baltimore and Richmond are based on known LCM recoveries (the Richmond estimates are based on Virginia State Police data initially reported by *The Washington Post*). The early period estimate for Baltimore excludes the year 2005 due to an unusually large number of guns appearing that year within the buyback/turn-in/safekeeping category. Other LCM estimates are based on recoveries of LCM compatible firearm models

recent change is that assault rifles, rather than assault pistols, now account for a substantial majority of AWs used in crime in contrast to prior estimates [2]. This implies an increase over time in the average lethality of AWs used in violence.

LCM firearms, which include AWs as well as other high-capacity semiautomatics, appear to account for 22 to 36% of crime guns in most places, with some estimates upwards of 40% for cases involving serious violence. These estimates are comparable to or higher than earlier estimates of LCM use. However, the higher-end estimates may overstate LCM use somewhat as most are based on measurement of LCM-compatible guns that may not all have been equipped with LCMs.

Consistent with prior research, this study also finds that AWs and LCM firearms are more heavily represented among guns used in murders of police and mass murders. AWs account for 13–16% of guns used in murders of police, while LCM weapons overall account for about 41% of these weapons. Estimates for firearm mass murders are very imprecise due to lack of data on the guns and magazines used in these cases, but available information suggests that AWs and other high-capacity semiautomatics are involved in as many as 57% of such incidents. Further, they are particularly prominent in public mass shootings and those resulting in the highest casualty counts.

Importantly, trend analyses suggest that LCM firearms have grown substantially as a share of crime guns since the expiration of the federal ban on AWs and LCMs. This implies possible increases in the level of gunfire and injury per gun attack during this time. Consistent with this inference, national statistics from the Centers for Disease Control and Prevention (CDC) and the FBI show that the ratio of gun homicides and assaultive non-fatal shootings to overall reported violent gun crimes (homicides, assaults, and robberies) rose from an average of 0.163 for 2003–2005 to an average of 0.21 for 2010–2012 (calculated from CDC [53] and FBI [54] data). This change was driven by non-fatal shootings, which have been trending upward since the early 2000s and recently reached their highest levels since 1995 [1]. The findings presented in this study suggest the possibility that greater use of high-capacity semiautomatics has contributed to this upward trend in shootings.

Further study would seem warranted on LCM use trends with additional jurisdictions and data sources. Research on this issue could be facilitated by more systematic efforts to collect detailed information on crime guns and magazines in local police databases as well as through national data collection systems like the Supplemental Homicide Reports and the National Violent Death Reporting System. Study of these weapons is also hampered by lack of public data on production of LCMs and LCM-compatible firearms. The need for better data on this issue may become more pressing if there continue to be significant changes in the lethality of commercially available firearms.

Additional research is also needed to quantify the effects that LCM use has on injuries and deaths from gun attacks—and by extension on the costs to society from gun violence. Research suggests that gunfire attacks involving semiautomatics produce more lethal and injurious outcomes [2, 10, 17, 55] and that 4–5% of assault-related gunshot victims are wounded in attacks involving more than ten shots fired [2]. However, such evidence is extremely limited at present. Studies of this issue, combined with evaluation research on the effects of current state and local LCM laws, could provide additional insights into the efficacy of expanding LCM restrictions at the local, state, and/or national levels. Research illuminating the public health and safety benefits of AW-LCM restrictions could also inform the courts as they continue to adjudicate recent challenges to the constitutionality of these statutes. Although this study does not directly evaluate any AW-LCM law, it provides further evidence that the federal ban curbed the spread of high-capacity semiautomatic weapons when it was in place and, in so doing, may have had preventive effects on gunshot victimizations.

Acknowledgments The authors thank the police agencies that provided data for this study: the Hartford (CT) Police Department, the New York State Police, the Baltimore Police Department, the Richmond (VA) Police Department, the Minneapolis Police Department, the Milwaukee Police Department, the Kansas City (MO) Police Department, the Sacramento Police Department, the Seattle Police Department, and the Federal Bureau of Investigation. The authors also thank Grace Beya, Mark Ecleo, and Thomas Prifti for additional research assistance. The opinions expressed in this manuscript are those of the authors and should not be attributed to any of the aforementioned organizations or individuals.

References

1. Fowler KA, Dahlberg LL, Haileyesus T, Annest JL. Firearm injuries in the United States. *Prev Med*. 2015;79:5–14.
2. Koper CS. *An Updated Assessment of the Federal Assault Weapons Ban: Impacts on Gun Markets and Gun Violence*, 1994-2003. Report to the National Institute of Justice, U.S. Department of Justice. Philadelphia, PA: Jerry Lee Center of Criminology, University of Pennsylvania; 2004.
3. Law center to prevent gun violence. http://smartgunlaws.org. Accessed May 2016.
4. Adler WC, Bielke FM, Doi DJ, Kennedy JF. *Cops under fire: law enforcement officers killed with assault weapons or guns with high capacity magazines*. Washington, DC: Handgun Control, Inc.; 1995.
5. Beck A, Gilliard, D, Greenfeld L, et al. *Survey of state prison inmates*, 1991. Washington, DC: Bureau of Justice Statistics, U.S. Department of Justice. 1993.
6. Hargarten SW, Karlson TA, O'Brien M, Hancock J, Quebbeman E. Characteristics of firearms involved in fatalities. *JAMA*. 1996;275:42–5.
7. Hutson HR, Anglin D, Pratts MJ Jr. Adolescents and children injured or killed in drive-by shootings in Los Angeles. *N Engl J Med*. 1994;330:324–7.
8. Hutson HR, Anglin D, Kyriacou DN, Hart J, Spears K. The epidemic of gang-related homicides in Los Angeles county from 1979 through 1994. *JAMA*. 1995;274:1031–6.
9. Kleck G. *Targeting guns: firearms and their control.* New York: NY; Aldine de Gruyter; 1997.
10. McGonigal MD, Cole J, Schwab CW, Kauder DR, Rotondo MF, Angood PB. Urban firearm deaths: a five-year perspective. *J Trauma.* 1993;35:532–7.
11. New York State Division of Criminal Justice Services. *Assault weapons and homicide in New York City.* Albany, NY: Author; 1994.
12. Zawitz MW. *Guns used in crime.* Washington, DC: Bureau of Justice Statistics, U.S. Department of Justice; 1995.
13. Koper CS. America's experience with the federal assault weapons ban, 1994–2004: key findings and implications. In: Webster DW, Vernick JS, editors. *Reducing gun violence in America: informing policy with evidence and analysis,* vol. 2013. Baltimore MD: Johns Hopkins University Press; 2013. p. 157–71.
14. Fallis D VA. Data show drop in criminal firepower during assault gun ban. *The Washington Post,* 2011; January 23.
15. Lee J, editor. *Gun digest 2015.* Iola, WI: Krause Publications; 2014.
16. Violence Policy Center. *The militarization of the U.S. civilian firearms market.* Washington, DC: Author; 2011.
17. Reedy DC, Koper CS. Impact of handgun types on gun assault outcomes: a comparison of gun assaults involving semiautomatic pistols and revolvers. *Injury Prevention.* 2003;9:151–5.
18. Planty M, Truman JL. *Firearm violence, 1993–2011.* Washington, DC: Bureau of Justice Statistics, U.S. Department of Justice; 2013.
19. Smith EL, Cooper A. *Homicides in the U.S. known to law enforcement, 2011.* Washington, DC: Bureau of Justice Statistics, U.S. Department of Justice; 2013.
20. Bureau of Alcohol, Tobacco, and Firearms. Crime gun trace reports. In: *National report.* Washington, DC: United States Department of the Treasury; 2000. p. 2002.
21. Brill S. *Firearm abuse: a research and policy report.* Washington, DC: Police Foundation; 1977.
22. Braga AA, Wintemute GJ, Pierce GL, Cook PJ, Ridgeway G. Interpreting the empirical evidence on illegal gun market dynamics. *J Urban Health: Bull New York Acad Med.* 2012;89:779–93.
23. Braga AA, Pierce GL. Disrupting illegal firearms markets in Boston: the effects of operation ceasefire on the supply of new handguns to criminals. *Criminol Public Policy.* 2005;4(4):717–48.
24. Cook PJ, Braga AA. Comprehensive firearms tracing: strategic and investigative uses of new data on firearms markets. *Arizona Law Review.* 2001;43:277–309.
25. Koper CS. Federal legislation and gun markets: how much have recent reforms of the federal firearms licensing system reduced criminal gun suppliers? *Criminol Public Policy.* 2002;1:151–78.
26. Koper CS. Crime gun risk factors: buyer, seller, firearm, and transaction characteristics associated with gun trafficking and criminal gun use. *J Quant Criminol.* 2014;30:285–315.

27. Pierce GL, Braga AA, Hyatt RR Jr, Koper CS. Characteristics and dynamics of illegal firearms markets: implications for a supply-side enforcement strategy. *Justice Q.* 2004;21:391–422.
28. Vernick JS, Webster DW, Hepburn LM. Effects of Maryland's law banning saturday night special handguns on crime guns. *Injury Prev.* 1999;5:259–63.
29. Webster DW, Vernick JS, Hepburn LM. Effects of Maryland's law banning saturday night special handguns on homicides. *Am J Epidemiol.* 2002;155:406–12.
30. Webster DW, Bulzacchelli MT, Zeoli AM, Vernick JS. Effects of undercover police stings of gun dealers on the supply of new guns to criminals. *Injury Prev.* 2006;12:255–30.
31. Webster DW, Vernick JS, Bulzacchelli MT. Effects of a gun dealer's change in sales practices on the supply of guns to criminals. *J Urban Health: Bull New York Acad Med.* 2006;83:778–87.
32. Webster DW, Vernick JS, Hepburn LM. Relationship between licensing, registration, and other gun sales laws and the source state of crime guns. *Injury Prev.* 2006;7:184–9.
33. Webster DW, Vernick JS, Bulzacchelli MT. Effects of state-level firearm seller accountability policies on firearm trafficking. *J Urban Health: Bull New York Acad Med.* 2009;86:525–37.
34. Wintemute GJ. *Ring of fire: the handgun makers of southern California.* Davis, CA: Violence Prevention Research Program, University of California, Davis; 1994.
35. Wintemute GJ, Romero MP, Wright MA, Grassel KM. The life cycle of crime guns: a description based on guns recovered from young people in California. *Ann Emerg Med.* 2004;43:733–42.
36. Wintemute GJ, Cook PJ, Wright MA. Risk factors among retail handgun dealers for frequent and disproportionate sales of guns used in violent and firearm related crimes. *Injury Prev.* 2005;11:357–63.
37. Wright MA, Wintemute GJ, Webster DW. Factors affecting a recently purchased handgun's risk for use in crime under circumstances that suggest gun trafficking. *J Urban Health: Bull New York Acad Med.* 2010;87:352–64.
38. Koper CS. Purchase of multiple firearms as a risk factor for criminal gun use: implications for gun policy and enforcement. *Criminol Public Policy.* 2005;4:749–78.
39. Kleck GBATF. Gun trace data and the role of organized gun trafficking in supplying guns to criminals. *Saint Louis Univ Public Law Rev.* 1999;18:23–45.
40. National Research Council. *Firearms and violence*: a critical review. Washington, DC: The national academies press; 2005.
41. Citizens Crime Commission of New York City. *Mass shooting incidents in America* (1984–2012). http://www.nycrimecommission.org/mass-shooting-incidents-america.php. Accessed March 2015.
42. Everytown for Gun Safety. *Analysis of recent mass shootings.* New York: NY; Author; 2014.
43. Everytown for Gun Safety. *Mass shootings in the United States:* 2009-2016. New York: Author; 2017.
44. Mass shooting tracker, gun violence archive. http://www.shootingtracker.com/. Accessed Aug. 2016.
45. Federal Bureau of Investigation. *A study of active shooter incidents in the United States between 2000 and 2013.* Washington, DC: U.S. Department of Justice.
46. Kaminski Leduc JL. *Weapons used in mass shootings.* Report 2013-R-0057. Hartford: Connecticut; Office of Legislative Research, Connecticut General Assembly; 2013.
47. Aronsen G, Follman M, Pan D. A guide to mass shootings in America. *Mother Jones.* http://www.motherjones.com/politics/2012/07/mass-shootings-map. Accessed Feb. 2015, Jul. 2016.
48. New York City Police Department. *Active shooter: recommendations and analysis for risk mitigation. New York: counterterrorism bureau,* New York City Police Department; New York, NY; 2012.
49. Violence Policy Center. *Mass shootings in the United States involving high-capacity ammunition magazines.* http://www.vpc.org/fact_sht/VPCshootinglist.pdf. Accessed June 18, 2017.
50. Berkowitz B, Gamio L, Lu D, Uhrmacher K, and Lindeman T. The math of mass shootings. *The Washington Post* https://www.washingtonpost.com/graphics/national/mass-shootings-in-america/. Accessed Aug. 2016.
51. Assault Weapons Act of 2013. OLL13047, 113th Cong. (2013).
52. Freskos B. Baltimore police are recovering more guns loaded with high-capacity magazines, despite ban on sales. *The trace.* 2017; Apr. 1. https://www.thetrace.org/2017/03/high-capacity-magazine-ban-baltimore-police/ Accessed Apr. 1, 2017.
53. Centers for Disease Control and Prevention. Injury Center. https://www.cdc.gov/injury/wisqars/. Accessed March 2015.
54. Federal Bureau of Investigation. Uniform crime reporting. https://ucr.fbi.gov/ucr. Accessed March 2015.
55. Richmond TS, Branas CC, Cheney RA, Schwab CW. The case for enhanced data collection of gun type. *J Trauma.* 2004;57:1356–60.

The Effect of Large-Capacity Magazine Bans on High-Fatality Mass Shootings, 1990–2017

Louis Klarevas, PhD, Andrew Conner, BS, David Hemenway, PhD

Objectives. To evaluate the effect of large-capacity magazine (LCM) bans on the frequency and lethality of high-fatality mass shootings in the United States.

Methods. We analyzed state panel data of high-fatality mass shootings from 1990 to 2017. We first assessed the relationship between LCM bans overall, and then federal and state bans separately, on (1) the occurrence of high-fatality mass shootings (logit regression) and (2) the deaths resulting from such incidents (negative binomial analysis). We controlled for 10 independent variables, used state fixed effects with a continuous variable for year, and accounted for clustering.

Results. Between 1990 and 2017, there were 69 high-fatality mass shootings. Attacks involving LCMs resulted in a 62% higher mean average death toll. The incidence of high-fatality mass shootings in non–LCM ban states was more than double the rate in LCM ban states; the annual number of deaths was more than 3 times higher. In multivariate analyses, states without an LCM ban experienced significantly more high-fatality mass shootings and a higher death rate from such incidents.

Conclusions. LCM bans appear to reduce both the incidence of, and number of people killed in, high-fatality mass shootings. (*Am J Public Health.* 2019;109:1754–1761. doi: 10.2105/AJPH.2019.305311)

The recent spate of gun massacres in the United States has re-energized the debate over how to prevent such tragedies.[1] A common response to high-profile acts of gun violence is the promotion of tighter gun legislation, and there is some evidence that laws imposing tighter restrictions on access to firearms have been associated with lower levels of mass shootings.[2] One proposal that has received renewed interest involves restricting the possession of large-capacity magazines (LCMs).[3–5] This raises an important question: what has been the impact of LCM bans on high-fatality mass shootings?

In an attempt to arrest an uptick in mass shooting violence in the early 1990s, Congress in 1994 enacted the federal assault weapons ban, which, among other things, restricted ownership of certain ammunition-feeding devices.[6,7] The law, which contained a sunset provision, was allowed to expire a decade later. Pursuant to that ban (18 USC §921(a) [1994]; repealed), it was illegal to possess LCMs—defined as any ammunition-feeding device holding more than 10 bullets—unless the magazines were manufactured before the enactment of the ban. LCM restrictions are arguably the most important component of assault weapons bans because they also apply to semiautomatic firearms without military-style features.[8,9]

Beginning with New Jersey in 1990, some states implemented their own regulations on LCMs. Today, 9 states and the District of Columbia restrict the possession of LCMs. The bans vary along many dimensions, including maximum bullet capacity of permissible magazines, grandfathering of existing LCMs, and applicable firearms. Moreover, overlaps sometimes exist between assault weapons bans and LCM bans, but not in all states. For example, California instituted a ban on assault weapons in 1989, but LCMs remained unregulated in the state until 1994, when the federal ban went into effect. In 2000, California's own statewide ban on LCMs took effect as a safeguard in the event the federal ban expired, which happened in 2004.[10,11]

LCMs provide a distinct advantage to active shooters intent on murdering numerous people: they increase the number of rounds that can be fired at potential victims before having to pause to reload or switch weapons. Evidence shows that victims struck by multiple rounds are more likely to die, with 2 studies finding that, when compared with the fatality rates of gunshot wound victims who were hit by only a single bullet, the fatality rates of those victims hit by more than 1 bullet were more than 60% higher.[12,13] Being able to strike human targets with more than 1 bullet increases shooters' chances of killing their victims. Analyses of gunshot wound victims at level I trauma centers have suggested that this multiple-impact capability is often attributable to the use of LCMs.[14,15]

In addition, LCMs provide active shooters with extended cover.[16] During an attack, perpetrators are either firing their guns or not firing their guns. While gunmen are firing, it is extremely difficult for those in the line of fire to take successful defensive maneuvers. But if gunmen run out of bullets, there are lulls in the shootings, as the perpetrators are forced to pause their attacks to reload or change weapons. These pauses provide opportunities for people to intervene and disrupt a shooting. Alternatively, they provide individuals in

ABOUT THE AUTHORS

Louis Klarevas is with the Teachers College, Columbia University, New York, NY. Andrew Conner is with the Frank H. Netter, MD, School of Medicine, Quinnipiac University, North Haven, CT. David Hemenway is with the Harvard T. H. Chan School of Public Health, Harvard University, Boston, MA.

Correspondence should be sent to Louis Klarevas, Research Professor, Office of the Provost, Teachers College, Columbia University, 525 W 120th St, New York, NY 10027 (e-mail: ljk2149@tc.columbia.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted July 22, 2019.
doi: 10.2105/AJPH.2019.305311

harm's way with a chance to flee or hide. Legislative endeavors that restrict access to LCMs are implemented with the express objective of reducing an active shooter's multiple-impact capability and extended cover.[10]

Although mass shootings have received extensive study, there has been little scholarly analysis of LCM bans.[17–24] The studies undertaken that have broached the subject of ammunition capacity have primarily concentrated on the effect of LCM bans on violent crimes other than mass shootings or on the impact of the assault weapons bans on mass shootings.[25–27]

Evidence suggests that firearms equipped with LCMs are involved in a disproportionate share of mass shootings.[10,20,28] Proponents of LCM bans believe that without LCMs, fewer people will be killed in a mass shooting, other things equal. In turn, fewer shootings will cross the threshold required to be classified as what we call a "high-fatality mass shooting" (≥6 victims shot to death). If LCM bans are effective, we should expect to find that high-fatality mass shootings occur at a lower incidence rate when LCM bans are in place, and fewer people are killed in such attacks. But have LCM bans actually saved lives in practice? To our knowledge, the impact of LCM bans has never been systematically assessed. This study fills that void.

METHODS

Mass shootings have been defined in a variety of ways, with some analyses setting the casualty threshold as low as 2 people wounded or killed and others requiring a minimum of 7 gunshot victims.[18,22,29] We focused on high-fatality mass shootings—the deadliest and most disturbing of such incidents—which are defined as intentional crimes of gun violence with 6 or more victims shot to death, not including the perpetrators.[20,30,31] After an exhaustive search, we identified 69 such incidents in the United States between 1990 and 2017. We then discerned whether each high-fatality mass shooting involved a LCM—unless otherwise stated, defined consistent with the 1994 federal ban as a detachable ammunition-feeding device capable of holding more than 10 bullets. (See Table 1 for a list of incidents and for additional details on the search and identification strategy we employed.)

The first state to enact an LCM ban was New Jersey in 1990. Since then, another 8 states and the District of Columbia have enacted LCM bans (Table A, available as a supplement to the online version of this article at http://www.ajph.org).[10] With no LCM bans in effect before 1990, a priori we chose that year to begin our analysis to avoid inflating the impact of the bans. Our data set extends 28 years, from 1990 through 2017. As a secondary analysis, we used a 13-year data set, beginning in 2005, the first full year after the federal assault weapons ban expired.

Our primary outcome measures were the incidence of high-fatality mass shootings and the number of victims killed. We distinguished between high-fatality mass shootings occurring with and without a ban in effect. Because the federal ban was in effect nationwide from September 13, 1994, through September 12, 2004, we coded every state as being under an LCM ban during that 10-year timeframe.

Our interest was in the effect of LCM bans. We ran regression analyses to determine if any relationship between LCM bans and high-fatality mass shootings can be explained by other factors. In our state–year panel multivariate analyses, the outcome variables were (1) whether an LCM-involved high-fatality mass shooting occurred, (2) whether any high-fatality mass shooting occurred, (3) the number of fatalities in an LCM-involved high-fatality mass shooting, and (4) the number of fatalities in any high-fatality mass shooting. Our analyses first combined and then separated federal and state LCM bans.

Consistent with the suggestions and practices of the literature on firearm homicides and mass shootings, our explanatory variables are population density; proportion of population aged 19 to 24 years, aged 25 to 34 years, that is Black, and with a college degree; real per-capita median income; unemployment rate; and per-capita prison population.[2,26,27,32] We also added a variable for percentage of households with a firearm. All regression models controlled for total state population. When the dependent variable reflected occurrences of incidents (ordered choice data), we used logit regression; we ran probit regression as a sensitivity analysis. We had multiple observations for individual states. To control for this, we utilized cluster-robust standard errors to account for the clustering of observations. When the dependent variable reflected deaths (count data), we used negative binomial regression; Gius used a Poisson regression, and we used that approach as a sensitivity analysis.[26] We included state fixed effects. We used a continuous variable for year because the rate of high-fatality mass shootings has increased over time. For purposes of sensitivity analysis, we also replaced the linear yearly trend with a quadratic function. We performed multivariate statistical analyses by using Stata/IC version 15.1 (StataCorp LP, College Station, TX).

Population data came from the US Census Bureau, unemployment data came from the Bureau of Labor Statistics, and imprisonment data came from the Bureau of Justice Statistics. The percentage of households with a firearm was a validated proxy (the percentage of suicides that are firearm suicides) derived from Centers for Disease Control and Prevention National Vital Statistics Data.[33]

RESULTS

Between 1990 and 2017, there were 69 high-fatality mass shootings (≥6 victims shot to death) in the United States. Of these, 44 (64%) involved LCMs, 16 did not (23%), and for 9 (13%) we could not determine whether LCMs were used (Table 1). The mean number of victims killed in the 44 LCM-involved high-fatality mass shootings was 11.8; including the unknowns resulted in that average falling to 11.0 (not shown). The mean number of victims killed in high-fatality mass shootings in which the perpetrator did not use an LCM was 7.3 (Table B, available as a supplement to the online version of this article at http://www.ajph.org); including the unknowns resulted in that average falling to 7.1 (not shown). When we excluded unknown cases, the data indicated that utilizing LCMs in high-fatality mass shootings resulted in a 62% increase in the mean death toll.

Data sets of mass shooting fatalities by their nature involve truncated data, with the mode generally being the baseline number of fatalities required to be included in the data set (6 fatalities in the current study). Our data

TABLE 1—High-Fatality Mass Shootings in the United States, 1990–2017

Incident	Date	City	State	LCM	Deaths, No.	State LCM Ban	Federal Assault Weapons Ban
1	Jun 18, 1990	Jacksonville	FL	Y	9	N	N
2	Jan 26, 1991	Chimayo	NM	N	7	N	N
3	Aug 9, 1991	Waddell	AZ	N	9	N	N
4	Oct 16, 1991	Killeen	TX	Y	23	N	N
5	Nov 7, 1992	Morro Bay and Paso Robles	CA	N	6	N	N
6	Jan 8, 1993	Palatine	IL	N	7	N	N
7	May 16, 1993	Fresno	CA	Y	7	N	N
8	Jul 1, 1993	San Francisco	CA	Y	8	N	N
9	Dec 7, 1993	Garden City	NY	Y	6	N	N
10	Apr 20, 1999	Littleton	CO	Y	13	Y	Y
11	Jul 12, 1999	Atlanta	GA	U	6	Y	Y
12	Jul 29, 1999	Atlanta	GA	Y	9	Y	Y
13	Sep 15, 1999	Fort Worth	TX	Y	7	Y	Y
14	Nov 2, 1999	Honolulu	HI	Y	7	Y	Y
15	Dec 26, 2000	Wakefield	MA	Y	7	Y	Y
16	Dec 28, 2000	Philadelphia	PA	Y	7	Y	Y
17	Aug 26, 2002	Rutledge	AL	N	6	Y	Y
18	Jan 15, 2003	Edinburg	TX	U	6	Y	Y
19	Jul 8, 2003	Meridian	MS	N	6	Y	Y
20	Aug 27, 2003	Chicago	IL	N	6	Y	Y
21	Mar 12, 2004	Fresno	CA	N	9	Y	Y
22	Nov 21, 2004	Birchwood	WI	Y	6	N	N
23	Mar 12, 2005	Brookfield	WI	Y	7	N	N
24	Mar 21, 2005	Red Lake	MN	Y	9	N	N
25	Jan 30, 2006	Goleta	CA	Y	7	Y	N
26	Mar 25, 2006	Seattle	WA	Y	6	N	N
27	Jun 1, 2006	Indianapolis	IN	Y	7	N	N
28	Dec 16, 2006	Kansas City	KS	N	6	N	N
29	Apr 16, 2007	Blacksburg	VA	Y	32	N	N
30	Oct 7, 2007	Crandon	WI	Y	6	N	N
31	Dec 5, 2007	Omaha	NE	Y	8	N	N
32	Dec 24, 2007	Carnation	WA	U	6	N	N
33	Feb 7, 2008	Kirkwood	MO	Y	6	N	N
34	Sep 2, 2008	Alger	WA	U	6	N	N
35	Dec 24, 2008	Covina	CA	Y	8	Y	N
36	Jan 27, 2009	Los Angeles	CA	N	6	Y	N
37	Mar 10, 2009	Kinston, Samson, and Geneva	AL	Y	10	N	N
38	Mar 29, 2009	Carthage	NC	N	8	N	N
39	Apr 3, 2009	Binghamton	NY	Y	13	Y	N
40	Nov 5, 2009	Fort Hood	TX	Y	13	N	N
41	Jan 19, 2010	Appomattox	VA	Y	8	N	N

Continued

TABLE 1—*Continued*

Incident	Date	City	State	LCM	Deaths, No.	State LCM Ban	Federal Assault Weapons Ban
42	Aug 3, 2010	Manchester	CT	Y	8	N	N
43	Jan 8, 2011	Tucson	AZ	Y	6	N	N
44	Jul 7, 2011	Grand Rapids	MI	Y	7	N	N
45	Aug 7, 2011	Copley Township	OH	N	7	N	N
46	Oct 12, 2011	Seal Beach	CA	N	8	Y	N
47	Dec 25, 2011	Grapevine	TX	N	6	N	N
48	Apr 2, 2012	Oakland	CA	N	7	Y	N
49	Jul 20, 2012	Aurora	CO	Y	12	N	N
50	Aug 5, 2012	Oak Creek	WI	Y	6	N	N
51	Sep 27, 2012	Minneapolis	MN	Y	6	N	N
52	Dec 14, 2012	Newtown	CT	Y	27	N	N
53	Jul 26, 2013	Hialeah	FL	Y	6	N	N
54	Sep 16, 2013	Washington	DC	N	12	Y	N
55	Jul 9, 2014	Spring	TX	Y	6	N	N
56	Sep 18, 2014	Bell	FL	U	7	N	N
57	Feb 26, 2015	Tyrone	MO	U	7	N	N
58	May 17, 2015	Waco	TX	Y	9	N	N
59	Jun 17, 2015	Charleston	SC	Y	9	N	N
60	Aug 8, 2015	Houston	TX	U	8	N	N
61	Oct 1, 2015	Roseburg	OR	Y	9	N	N
62	Dec 2, 2015	San Bernardino	CA	Y	14	Y	N
63	Feb 21, 2016	Kalamazoo	MI	Y	6	N	N
64	Apr 22, 2016	Piketon	OH	U	8	N	N
65	Jun 12, 2016	Orlando	FL	Y	49	N	N
66	May 27, 2017	Brookhaven	MS	U	8	N	N
67	Sep 10, 2017	Plano	TX	Y	8	N	N
68	Oct 1, 2017	Las Vegas	NV	Y	58	N	N
69	Nov 5, 2017	Sutherland Springs	TX	Y	25	N	N

Note. LCM = large-capacity magazine; N = no; U = unknown; Y = yes. From September 13, 1994, until and including September 12, 2004, each and every state, including the District of Columbia, was subject to a ban on LCMs pursuant to the federal assault weapons ban. To collect the data in Table 1, we searched the following news media resources for every shooting that resulted in 6 or more fatalities: America's Historical Newspapers, EBSCO, Factiva, Gannett Newsstand, Google News Archive, Lexis-Nexis, Newspaper Archive, Newspaper Source Plus, Newspapers.com, Newswires, ProQuest Historical Newspapers, and ProQuest Newsstand. We also reviewed mass shooting data sets maintained by *Mother Jones*, the *New York Times*, and *USA Today*. In addition to news media sources, we reviewed reports on mass shootings produced by think tank, policy advocacy, and governmental organizations, including the US Federal Bureau of Investigation Supplementary Homicide Reports, the crowdsourced Mass Shooting Tracker, and the open-source databases maintained by the Gun Violence Archive and the Stanford University Geospatial Center. Finally, when it was relevant, we also reviewed court records as well as police, forensic, and autopsy reports. As a general rule, when government sources were available, they were preferred over other sources. Furthermore, when media sources conflicted on the number of casualties or the weaponry involved, the later sources were privileged (as later reporting is often more accurate).

set of high-fatality mass shootings was no exception. As such, the median average number of fatalities for each subset of incidents—those involving and those not involving LCMs—was necessarily lower than the mean average. Nevertheless, like the mean average, the median average was higher when LCMs were employed—a median average of 8 fatalities per incident compared with 7 fatalities per incident for attacks not involving LCMs.

For the 60 incidents in which it was known if an LCM was used, in 44 the perpetrator used an LCM. Of the 44 incidents in which the perpetrators used LCMs, 77% (34/44) were in nonban states. In the 16 incidents in which the perpetrators did not use LCMs, 50% (8/16) were in nonban states (Table B, available as a supplement to the online version of this article at http://www.ajph.org). Stated differently, in nonban states, 81% (34/42) of high-fatality mass shooting perpetrators used LCMs; in LCM-ban states, only 55% (10/18) used LCMs.

The rate of high-fatality mass shootings increased considerably after September 2004 (when the federal assault weapons ban expired). In the 10 years the federal ban was in effect, there were 12 high-fatality mass shootings and 89 deaths (an average of 1.2 incidents and 8.9 deaths per year). Since then, through 2017, there have been 48 high-fatality mass shootings and 527 deaths (an average of 3.6 incidents and 39.6 deaths per year in these 13.3 years).

Of the 69 high-fatality mass shootings from 1990 to 2017, 49 occurred in states without an LCM ban in effect at the time and 20 in states with a ban in effect at the time. The annual incidence rate for high-fatality mass shootings in states without an LCM ban was 11.7 per billion population; the annual incidence rate for high-fatality mass shootings in states with an LCM ban was 5.1 per billion population. In that 28-year period, the rate of high-fatality mass shootings per capita was 2.3 times higher in states without an LCM ban (Table 2).

Non–LCM ban states had not only more incidents but also more deaths per incident (10.9 vs 8.2). The average annual number of high-fatality mass shooting deaths per billion population in the non–LCM ban states was 127.4. In the LCM ban states, it was 41.6 (Table 2).

For the time period beginning with the first full calendar year following the expiration of the federal assault weapons ban (January 1, 2005–December 31, 2017), there were 47 high-fatality mass shootings in the United States. Of these, 39 occurred in states where an LCM ban was not in effect, and 8 occurred in LCM ban locations. The annual incidence rate for high-fatality mass shootings in states without an LCM ban was 13.2 per billion population; for states with an LCM ban, it was 7.4 per billion population (Table 2). During this period, non–LCM ban states had not only more incidents but also more deaths per incident (11.4 vs 9.4). In terms of high-fatality mass shooting deaths per billion population, the annual number of deaths in the non–LCM ban states was 150.6; in the LCM ban states it was 69.2 (Table 2).

When we limited the analysis solely to high-fatality mass shootings that definitely involved LCMs, the differences between ban and nonban states became larger. For example, for the entire period of 1990 to 2017, of the 44 high-fatality mass shootings that involved LCMs, the annual incidence rate for LCM-involved high-fatality mass shootings in nonban states was 8.1 per billion population; in LCM-ban states it was 2.5 per billion population. The annual rate of high-fatality mass shooting deaths in the non–LCM ban states was 102.1 per billion population; in the LCM ban states it was 23.3. In terms of LCM-involved high-fatality mass shootings, we also found comparable wide differences in incidence and fatality rates between ban and nonban states for the post–federal assault weapons ban period (2005–2017; Table 2).

We found largely similar results in the multivariate analyses (1990–2017). States that did not ban LCMs were significantly more likely to experience LCM-involved high-fatality mass shootings as well as more likely to experience any high-fatality mass shootings (regardless of whether an LCM was involved). States that did not ban LCMs also experienced significantly more deaths from high-fatality mass shootings, operationalized as the absolute number of fatalities (Table 3).

When the LCM bans were separated into federal and state bans, both remained significantly related to the incidence of LCM-involved high-fatality mass shooting events and to the number of LCM-involved high-fatality mass shooting deaths. The associations between federal and state bans and

TABLE 2—High-Fatality Mass Shootings (≥ 6 Victims Shot to Death) by Whether LCM Bans Were in Effect: United States, 1990–2017

	Average Annual Population, No. (Millions)	Total Incidents, No.	Annual Incidents per Billion Population, No.	Total Deaths, No.	Annual Deaths per Billion Population, No.	Deaths per Incident, No.
All high-fatality mass shootings, 1990–2017 (28 y)						
Non–LCM ban states	149.7	49	11.7	534	127.4	10.9
LCM ban states	140.7	20	5.1	164	41.6	8.2
All high-fatality mass shootings, 2005–2017 (13 y)						
Non–LCM ban states	227.8	39	13.2	446	150.6	11.4
LCM ban states	83.4	8	7.4	75	69.2	9.4
LCM-involved high-fatality mass shootings, 1990–2017 (28 y)						
Non–LCM ban states	149.7	34	8.1	428	102.1	12.6
LCM ban states	140.7	10	2.5	92	23.3	9.2
LCM-involved high-fatality mass shootings, 2005–2017 (13 y)						
Non–LCM ban states	227.8	28	9.5	369	124.6	13.2
LCM ban states	83.4	4	3.7	42	38.7	10.5
Non-LCM high-fatality mass shootings, 1990–2017 (28 y)						
Non–LCM ban states	149.7	8	1.9	56	13.4	7.0
LCM ban states	140.7	8	2.0	60	15.2	7.5

Note. LCM = large-capacity magazine.

TABLE 3—Multivariate Results of the Relationship Between LCM Bans and High-Fatality Mass Shootings (≥ 6 Victims Shot to Death), 1990–2017 Combined Federal and State Large Capacity Magazine Bans: United States

	LCM-Involved High-Fatality Mass Shootings, b (95% CI)		All High-Fatality Mass Shootings, b (95% CI)	
	Incidents[a]	No. Deaths[b]	Incidents[a]	No. Deaths[b]
All LCM bans (federal and state)	−2.217 (−3.493, −0.940)	−5.912 (−9.261, −2.563)	−1.283 (−2.147, −0.420)	−3.660 (−5.695, −1.624)
Population density	−0.011 (−0.052, 0.031)	0.013 (−0.068, 0.095)	0.001 (−0.003, 0.006)	0.011 (−0.005, 0.026)
% aged 19–24 y	−0.480 (−1.689, 0.730)	−2.496 (−5.893, 0.901)	0.283 (−0.599, 1.164)	−0.585 (−2.666, 1.495)
% aged 25–34 y	−0.801 (−1.512, −0.089)	−2.390 (−4.391, −0.388)	−0.337 (−0.871, 0.197)	−1.114 (−2.463, 0.235)
% Black	−0.227 (−1.062, 0.607)	−0.654 (−2.831, 1.522)	−0.163 (−0.703, 0.377)	−0.261 (−1.391, 0.870)
% with a bachelor's degree or higher	−0.009 (−0.492, 0.474)	−0.469 (−1.590, 0.652)	0.143 (−0.214, 0.501)	0.183 (−0.715, 1.081)
Percentage of households with a firearm (proxy)	−0.047 (−0.195, 0.101)	−0.147 (−0.546, 0.251)	−0.020 (−0.131, 0.091)	−0.084 (−0.368, 0.200)
Median household income	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Unemployment rate	−0.072 (−0.293, 0.149)	−0.476 (−1.081, 0.129)	0.041 (−0.135, 0.216)	−0.182 (−0.628, 0.263)
Imprisonment rate (per 100 000 population)	−0.006 (−0.012, 0.001)	−0.007 (−0.017, 0.004)	−0.001 (−0.006, 0.003)	−0.003 (−0.012, 0.007)
Total population	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Pseudo R^2	0.31	0.16	0.26	0.11

Note. CI = confidence interval; LCM = large-capacity magazine. There were a total of 1428 observations in state-years (51 jurisdictions—all 50 states plus Washington, DC—over a 28-year period). Mean variance inflation factor = 3.49.

[a]Logit regression.

[b]Negative binomial regression.

the overall incidence of all high-fatality mass shootings as well as the total number of victims in these events remained strongly negative but was only sometimes statistically significant (Table 4).

In terms of sensitivity analyses, using probit instead of logit gave us similar results (not shown). When the outcome variable was the number of high-fatality mass shooting deaths, we obtained largely similar results concerning the association between LCM bans and the outcome variables, regardless of whether we used Poisson or negative binominal regression (not shown). Moreover, replacing the linear yearly trend with a quadratic function did not change the major results of the analyses (not shown). Variance inflation factors for all the independent variables never exceeded 10.0, with the variance inflation factor for LCM ban variables always being less than 2.0, indicating that there were no significant multicollinearity issues (Tables 3 and 4).

DISCUSSION

In the United States, LCMs are disproportionately used in high-fatality mass shootings (incidents in which ≥ 6 victims are shot to death). In at least 64% of the incidents since 1990, perpetrators used LCMs. (For 23%, we determined that they did not involve LCMs, and a determination could not be made for the remaining 13%.) Previous research has shown that LCM firearms are used in a high share of mass murders (typically defined as ≥ 4 homicides) and murders of police.[9]

We could not find reliable estimates of LCM firearms in the US gun stock. However, it is likely much lower than 64%, given that commonly owned firearms such as revolvers, bolt-action rifles, and shotguns are not typically designed to be LCM-capable. During the decade the federal assault weapons ban was in effect, no firearms were legally manufactured with LCMs for sale in the United States. In the postban era, semiautomatic firearms, especially pistols, are often sold with factory-issue LCMs, but firearms that are not semiautomatic are not sold with such magazines.

Why do we find LCMs so prominent among high-fatality mass shootings? We suspect there are 2 main reasons. The first is that perpetrators probably deliberately select LCMs because they facilitate the ability to fire many rounds without having to stop to reload. The second reason is that the ability of shooters to kill many victims—especially the 6 victims required to be included in our data set—may be reduced if LCMs are not available. In other words, the first explanation is that shooters perceive LCMs to be more effective at killing many people; the second explanation is that LCMs are indeed more effective at killing many people.

High-fatality mass shootings are not common, even in the United States. Between 1990 and 2017, there has been an average of 2.5 incidents per year, with an average of 25 people killed annually in such attacks. However, the number of incidents and the number of people killed per incident have been increasing since the end of the federal assault weapons ban.

In our study, we found that bans on LCMs were associated with both lower incidence of high-fatality mass shootings and lower fatality tolls per incident. The difference in incidence and overall number of fatalities between states, with and without bans, was even greater for LCM-involved high-fatality mass shootings.

The multivariate results are largely consistent with these bivariate associations. When we controlled for 10 independent variables often associated with overall crime rates, as well as state and year effects, states with LCM bans had lower rates of high-fatality mass shootings and fewer high-fatality mass shooting deaths. When we investigated federal and state bans separately in the multiple

TABLE 4—Multivariate Results of the Relationship Between Large Caliber Magazine Bans and High-Fatality Mass Shootings (≥ 6 Victims Shot to Death), 1990–2017 Separate Federal and State Large Caliber Magazine Bans: United States

	LCM-Involved High-Fatality Mass Shootings, b (95% CI)		All High-Fatality Mass Shootings, b (95% CI)	
	Incidents[a]	No. Deaths[b]	Incidents[a]	No. Deaths[b]
Federal LCM ban	−1.434 (−2.622, −0.245)	−3.571 (−7.103, −0.038)	−0.895 (−1.806, 0.016)	−2.570 (−4.902, −0.238)
State LCM bans	−2.603 (−4.895, −0.311)	−8.048 (−15.172, −0.925)	−1.277 (−2.977, 0.422)	−3.082 (−7.227, 1.064)
Population density	−0.012 (−0.055, 0.030)	−0.001 (−0.085, 0.083)	0.001 (−0.003, 0.006)	0.009 (−0.007, 0.024)
% aged 19–24 y	−0.311 (−1.499, 0.878)	−2.589 (−6.057, 0.879)	0.342 (−0.551, 1.236)	−0.531 (−2.759, 1.698)
% aged 25–34 y	−0.812 (−1.532, −0.093)	−2.660 (−4.848, −0.471)	−0.323 (−0.864, 0.217)	−0.848 (−2.236, 0.539)
% Black	−0.229 (−1.101, 0.643)	−0.770 (−3.232, 1.693)	−0.150 (−0.698, 0.398)	−0.154 (−1.321, 1.013)
% with a bachelor's degree or higher	−0.031 (−0.447, 0.509)	−0.479 (−1.577, 0.618)	0.156 (−0.199, 0.511)	0.269 (−0.567, 1.106)
Percentage of households with a firearm (proxy)	−0.055 (−0.210, 0.101)	−0.227 (−0.651, 0.196)	−0.019 (−0.133, 0.094)	−0.107 (−0.399, 0.186)
Median household income	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Unemployment rate	−0.061 (−0.284, 0.162)	−0.420 (−1.041, 0.201)	0.046 (−0.132, 0.224)	−0.157 (−0.619, 0.305)
Imprisonment rate (per 100 000 population)	−0.006 (−0.013, 0.000)	−0.012 (−0.026, 0.002)	−0.002 (−0.007, 0.003)	−0.003 (−0.014, 0.007)
Total population	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Pseudo R^2	0.30	0.15	0.26	0.11

Note. CI = confidence interval; LCM = large-capacity magazine. There were a total of 1428 observations in state-years (51 jurisdictions—all 50 states plus Washington, DC—over a 28-year period). Mean variance inflation factor = 3.45.
[a]Logit regression.
[b]Negative binomial regression.

regressions, both were significantly associated with the incidence of LCM-involved high-fatality mass shootings as well as the number of victims in LCM-involved attacks. The relationship between these bans, considered separately, and all high-fatality mass shooting incidence and deaths is often not statistically significant, although this may be attributable to lack of statistical power (number of observations) to find a statistically significant effect.

Our analysis provides answers to 4 important questions:

1. How often are LCMs used in high-fatality mass shootings? At minimum, 64% of high-fatality mass shootings perpetrated between 1990 and 2017 involved LCMs.
2. Are more people killed when LCMs are used? Yes, and the difference in our data set is substantial and statistically significant (11.8 vs 7.3). We should add that our results likely underestimate the difference because we have a truncated sample (we only examined incidents with at least 6 victim fatalities), compounded by the fact that the number of homicide incidents fell as the number of victims increased.
3. Do states with LCM bans experience high-fatality mass shootings involving LCMs at a lower rate and a lower fatality count than those states with no such bans in effect? Yes. In fact, the effect is more pronounced for high-fatality mass shootings involving LCMs than for those not involving LCMs.
4. Do states with LCM bans experience high-fatality mass shootings (regardless of whether they involve LCMs) at a lower rate and a lower fatality count than states with no such bans in effect? Yes.

Limitations

Our study had various limitations. First, although we carefully searched for every high-fatality mass shooting, it is possible that we might have missed some. Nevertheless, we suspect that this is unlikely, because it would mean that others who compiled lists have also missed the same ones, for we checked our list against multiple sources.

Second, our definition of a high-fatality mass shooting is a shooting that results in 6 or more fatal victims. A different threshold criterion (e.g., 6 or more people shot; 5 or more victims killed), might lead to somewhat different results. We expect that as the number of victims in a shooting increases, the likelihood that the perpetrator used an LCM also increases. Indeed, of the 13 high-fatality mass shootings with 10 or more fatalities in our data set, 12 (92%) involved an LCM.

Third, although many high-fatality mass shootings tend to be highly publicized, in 13% of the incidents we reviewed, we could not determine whether an LCM was used. As a sensitivity analysis, we assessed the assumptions that all of the unknown cases first did, and then did not, involve LCMs. Neither assumption appreciably changed our main results (not shown).

Fourth, as a general rule, clustering standard errors is most appropriate when there is a large number of treated units. Although during the decade of the federal assault weapons bans all 50 states plus the District of Columbia regulated LCMs, during the remaining time periods under examination, only 8 jurisdictions regulated LCMs. As a result, there is the possibility that the standard errors were underestimated in our analyses.[34]

Fifth, there were only 69 events that met our criterion for a "high-fatality mass shooting." Although 69 is a horrific number of incidents, for statistical purposes, it is a relatively small number and limits the power to detect significant associations. For example, we did not have the statistical power (and thus did not even try) to determine whether different aspects of the various LCM laws might have differential effects on the incidence of high-fatality mass shootings. Moreover, because of suboptimal statistical power, there is also the possibility that the magnitude of the effects detected was overestimated.[35]

Public Health Implications

LCMs increase the ability to fire large numbers of bullets without having to pause to reload. Any measure that can force a pause in an active shooting—creating opportunities for those in the line of fire to flee, take cover, or physically confront a gunman—offers a possibility of reducing the number of victims in such an attack. To put it in different terms, if the only firearms available were 18th-century muskets, it is doubtful that mass shootings would be the social problem they are today.

The impact of individual state firearm laws is reduced by the fact that guns often move across state lines—occasionally purchased in locales with more permissive laws and taken to states with more restrictive laws. This is partly why efforts aimed at reducing the frequency and lethality of mass shootings must necessarily be multifaceted and multidisciplinary. Legal restrictions on firearms are merely a part of this broader, public health approach. That being said, the theory behind reducing the availability of LCMs to reduce the number of victims in mass shootings makes sense, and our empirical results, consistent with much of the limited literature on mass shootings, suggest that LCM bans have been effective in saving lives. AJPH

CONTRIBUTORS

L. Klarevas and D. Hemenway designed the study, collected the data, and contributed equally to all parts of the study. A. Conner ran the statistical analyses and helped construct the tables that report the results of the multivariate analyses. All authors approved the final article as submitted.

ACKNOWLEDGMENTS

The authors would like to thank John Berrigan, research assistant at the Harvard Injury Control Research Center, for his assistance with the undertaking of this study.

CONFLICTS OF INTEREST

L. Klarevas has, in the past 2 years, served as an expert to the states of Colorado and California in civil litigation that involved the constitutionality of state restrictions on large-capacity magazines. The authors have no additional conflicts of interest to report.

HUMAN PARTICIPANT PROTECTION

No protocol approval was needed because no human participants were involved in this study.

REFERENCES

1. Wintemute GJ. How to stop mass shootings. *N Engl J Med*. 2018;379(13):1193–1196.

2. Reeping PM, Cerda M, Kalesan B, Wiebe DJ, Galea S, Branas CC. State gun laws, gun ownership, and mass shootings in the US: cross sectional time series. *BMJ*. 2019; 364(8190):1542–1548.

3. Sanger-Katz M, Bui Q. How to reduce mass shooting deaths? Experts rank gun laws. *New York Times*. October 5, 2017. Available at: https://www.nytimes.com/interactive/2017/10/05/upshot/how-to-reduce-mass-shooting-deaths-experts-say-these-gun-laws-could-help.html. Accessed June 10, 2019.

4. Kamerow D. Guns don't kill crowds, people with semi-automatics do. *BMJ*. 2011;342(1):d477.

5. Barry CL, Webster DW, Stone E, Crifasi CK, Vernick JS, McGinty EE. Public support for gun violence prevention policies among gun owners and non–gun owners in 2017. *Am J Public Health*. 2018;108(7): 878–881.

6. Lenett MG. Taking a bite out of violent crime. *Univ Dayton Law Rev*. 1995;20(2):573–617.

7. Muchnick JY. The assault weapons ban: saving lives. *Univ Dayton Law Rev*. 1995;20(2):641–651.

8. Koper CS, Roth JA. The impact of the 1994 federal assault weapon ban on gun violence outcomes: an assessment of multiple outcome measures and some lessons for policy evaluation. *J Quant Criminol*. 2001;17(1):33–74.

9. Koper CS, Johnson WD, Nicholas JL, Ayers A, Mullins N. Criminal use of assault weapons and high-capacity semiautomatic firearms: an updated examination of local and national sources. *J Urban Health*. 2018;95(3):313–321.

10. Giffords Law Center to Prevent Gun Violence. Large capacity magazines. Available at: http://lawcenter.giffords.org/gun-laws/policy-areas/hardware-ammunition/large-capacity-magazines. Accessed June 10, 2019.

11. Giffords Law Center to Prevent Gun Violence. Assault weapons. Available at: https://lawcenter.giffords.org/gun-laws/policy-areas/hardware-ammunition/assault-weapons. Accessed June 10, 2019.

12. Webster DW, Champion HR, Gainer PS, Sykes L. Epidemiologic changes in gunshot wounds in Washington, DC, 1983–1990. *Arch Surg*. 1992;127(6): 694–698.

13. Koper CS, Roth JA. A priori assertions versus empirical inquiry: a reply to Kleck. *J Quant Criminol*. 2001; 17(1):81–88.

14. Livingston DH, Lavery RF, Lopreiato MC, Lavery DF, Passannante MR. Unrelenting violence: an analysis of 6,322 gunshot wound patients at a level I trauma center. *J Trauma Acute Care Surg*. 2014;76(1):2–9.

15. Manley NR, Fabian TC, Sharpe JP, Magnotti LJ, Croce MA. Good news, bad news: an analysis of 11,294 gunshot wounds (GSWs) over two decades in a single center. *J Trauma Acute Care Surg*. 2017;84(1):58–65.

16. McCullough J. *The Ultimate Guide to US Army Combat Skills, Tactics and Techniques*. New York, NY: Skyhorse; 2012.

17. Fox JA, DeLateur MJ. Mass shootings in America: moving beyond Newtown. *Homicide Stud*. 2014;18(1): 125–145.

18. Krouse WJ, Richardson DJ. Mass murder with firearms: incidents and victims, 1999–2013. CRS Report R44126. Washington, DC: Congressional Research Service; 2015.

19. Wilson LC. *The Wiley Handbook of the Psychology of Mass Shootings*. Hoboken, NJ: Wiley; 2016.

20. Klarevas L. *Rampage Nation: Securing America From Mass Shootings*. Amherst, NY: Prometheus; 2016.

21. Schildkraut J, Elsass HJ. *Mass Shootings: Media, Myths, and Realities*. Denver, CO: Praeger; 2016.

22. Schildkraut J. *Mass Shootings in America: Understanding the Debates, Causes, and Responses*. Denver, CO: ABC-CLIO; 2018.

23. Rocque M, Duwe G. Rampage shootings: an historical, empirical, and theoretical overview. *Curr Opin Psychol*. 2018;19(1):28–33.

24. Phaneuf SW. Mass shootings: understanding the complexities. In: Hilinski-Rosick CM, Lee DR, eds. *Contemporary Issues in Victimology: Identifying Patterns and Trends*. New York, NY: Lexington; 2018.

25. Moody CE, Marvell TB. Clustering and standard error bias in fixed effects panel data regressions. *J Quant Crim*; 2018:1–23.

26. Gius M. The impact of state and federal assault weapons bans on public mass shootings. *Appl Econ Lett*. 2015;22(4):281–284.

27. Blau BM, Gorry DH, Wade C. Guns, laws and public shootings in the United States. *Appl Econ*. 2016;48(49): 4732–4746.

28. Follman M, Aronsen G, Pan D. A guide to mass shootings in America. *Mother Jones*. May 31, 2019. Available at: https://www.motherjones.com/politics/2012/07/mass-shootings-map. Accessed June 10, 2019.

29. Kleck G. Large-capacity magazines and the casualty counts in mass shootings: the plausibility of linkages. *Justice Res Policy*. 2016;17(1):28–47.

30. Webster DW, Donohue JJ, Klarevas L, et al. Firearms on college campuses: research evidence and policy implications. Report of the Johns Hopkins University Center for Gun Policy and Research. Baltimore, MD: Johns Hopkins Bloomberg School of Public Health; 2016.

31. Kleck G. *Targeting Guns: Firearms and Their Control*. Hawthorne, NY: Aldine de Gruyter; 1997.

32. Hemenway D. *Private Guns, Public Health*. Ann Arbor, MI: University of Michigan Press; 2017.

33. Azrael D, Cook PJ, Miller M. State and local prevalence of firearms ownership: measurement, structure and trends. *J Quant Criminol*. 2004;20(1):43–62.

34. Conley TG, Taber CR. Inference with "difference in differences" with a small number of policy changes. *Rev Econ Stat*. 2011;93(1):113–125.

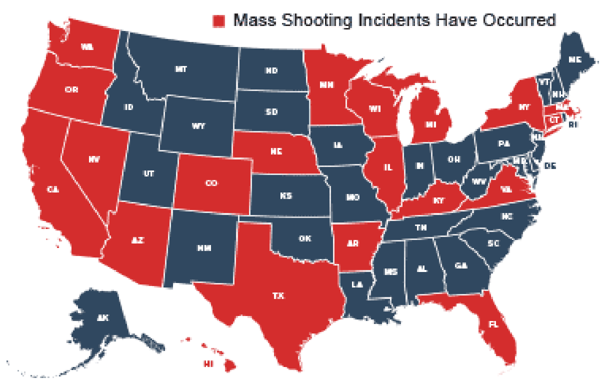
35. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience [erratum *Nat Rev Neurosci*. 2013;14(6):451]. *Nat Rev Neurosci*. 2013;14(5):365–376.

CITIZENS CRIME COMMISSION
OF NEW YORK CITY

[A non-partisan non-profit organization](#) working to make criminal justice
and public safety policies and practices more effective through innovation,
research, and education.

INITIATIVE

Mass Shooting Incidents in America (1984-2012)



Mass shootings are a unique feature of American life which have occurred consistently throughout history in every region of the country. The increased lethality of such incidents is made possible by the use of large capacity ammunition magazines (defined as more than 10-rounds) which enable a shooter to rapidly fire off as many as 100-rounds without having to reload the firearm. Designed for military use to kill greater numbers of people more effectively, large capacity ammunition magazines have facilitated some of the worst mass murders ever committed in the United States. As these incidents occur in every region of the country, restricting civilian access to these weapons is not a state specific problem. The federal government needs to take action to protect all Americans by reinstating the ban on large capacity ammunition magazines.

This database provides an overview of significant mass shooting incidents in America (defined by the FBI as four or more victims killed), all of which involved large capacity ammunition magazines. *

December 14, 2012

Newtown, CT

Shooter
Adam Lanza, 20

Ammo Magazine Capacity
30-rounds

Shots Fired >154
Killed 27 (plus shooter = 28)
Wounded unknown

Sandy Hook Elementary School

Incident
On December 14, 2012, Adam Lanza armed with a .22-caliber rifle killed his mother in her home in Newtown, CT. Lanza then stocked his mother's car with firearms and drove to Sandy Hook Elementary School. He shot his way into the school and opened fire with a Bushmaster XM15 .223-caliber semiautomatic assault rifle equipped with a 30-round large capacity ammunition magazine, killing 26, including 20 students' ages six and seven. As police closed in Lanza committed suicide by shooting himself with a GLOCK 10mm handgun. He fired over 154 shots in less than five minutes.

Weapons
An unknown make and model .22-caliber rifle, a Bushmaster XM15 .223-caliber semiautomatic assault rifle equipped with a 30-round large capacity ammunition magazine, and a GLOCK 10mm handgun were used. According to the Danbury State's Attorney, police also recovered in Lanza's possession a SIG SAUER P226 9mm handgun and three loaded 30-round large capacity ammunition magazines for the Bushmaster. Six additional 30-round large capacity ammunition magazines were recovered at the scene. A loaded unknown make and model 12-gauge shotgun was found in the passenger compartment of the car (later moved to the trunk by police). All of the guns used in the shooting were purchased by Lanza's mother.

Outcome
Suicide.

September 27, 2012

Minneapolis, MN

Accent Signage Systems

Incident
On September 27, 2012, after working his shift at Accent Signage Systems, Andrew Engeldinger was told by two company managers that he was being fired for chronic tardiness

Shooter
Andrew John Engeldinger, 36

Ammo Magazine Capacity
15-rounds

Shots Fired >46
Killed 6 (plus shooter = 7)
Wounded 2

and poor performance. Upon hearing this news, Engeldinger pulled out a semiautomatic handgun equipped with a 15-round large capacity ammunition magazine, the managers tried to get the gun from him, unable to both mangers were shot. The large capacity ammunition magazine was dropped during the struggle; Engeldinger reinserted the magazine into the firearm and began to move through the office, shooting at some employees but not others. Over approximately 15 minutes, Engeldinger shot seven employees and a UPS driver before turning the gun on himself. Four victims died at the scene, two died at the hospital (one the following day and the other two weeks later), and two others were injured.

Weapons
GLOCK 19 9mm semiautomatic pistol equipped with a 15-round large capacity ammunition magazine. Engeldinger purchased the firearm one year before the shooting at KGS Guns and Ammo in Minneapolis after passing a background check and obtaining a permit to purchase. Police reportedly found packaging for 10,000 rounds of ammunition and another handgun in Engeldinger's home.

Outcome
Suicide.

August 5, 2012

Oak Creek, WI

Shooter
Wade Michael Page, 40

Ammo Magazine Capacity
19-rounds

Shots Fired unknown
Killed 6 (plus shooter = 7)
Wounded 3

Sikh Temple of Wisconsin

Incident
Around 10:30 AM, Wade Michael Page, a U.S. Army veteran, opened fire in the parking lot of a Sikh temple, then entered the building shooting congregants gathering for Sunday meditation. Police officers arrived on the scene in response to 911 calls, and exchanged fire with the shooter. Page killed six and injured three, including a responding officer, before committing suicide.

Weapons
Springfield Armory XD(M) 9mm semiautomatic handgun equipped with a 19-round large capacity ammunition magazine. Weeks before the shooting, Wade legally purchased the handgun and three 19-round large capacity ammunition magazines from a federal firearms licensed dealer in nearby West Allis, WI. According to media reports, Wade served in the U.S. Army from 1992 until 1998, when he was given an other-than-honorable discharge or general discharge. In 1994, while stationed at Fort Bliss in Texas, he was arrested by El Paso police, and pled guilty to a misdemeanor charge of criminal mischief. Federal law does not prohibit persons with convictions for misdemeanors other than domestic violence misdemeanors or persons who have been discharged from the military for reasons other than "dishonorably" from purchasing firearms.

Outcome
Wade committed suicide after being shot by police at the scene. The FBI is leading the investigation which is being treated as a possible act of domestic terrorism.

July 20, 2012

Aurora, CO

Shooter
James Holmes, 24

Ammo Magazine Capacity
100-rounds

Shots Fired >80
Killed 12
Wounded 70

The Dark Knight Rises: Movie theatre Shooting

Incident
Shortly after the start of the midnight premiere screening of *Batman: The Dark Knight Rises* on July 20, 2012, at the Century Aurora 16 movie theatre in Aurora, CO, James Holmes exited the theatre through an emergency exit. He returned through the propped open emergency exit door, clad in ballistic body armor, wearing a gas mask, and armed with multiple firearms. After tossing two canisters of tear gas into the theatre he began firing upon the audience. He first used an AR-15-type assault rifle equipped with a 100-round drum large capacity ammunition magazine, after the assault rifle jammed, he then continued with a 12-gauge shotgun and a handgun--killing 12 and wounding 70 (including three wounded when bullets went through a wall into an adjacent theatre).

Weapons
A Smith & Wesson M&P15 assault rifle equipped with a 100-round drum large capacity ammunition magazine, a Remington Model 870 12-gauge pump shotgun, and two GLOCK .40-caliber handguns, were recovered at the scene by police. In the months leading to the shooting, Holmes purchased the weapons and 6,000-rounds of ammunition at gun shops and over the Internet. In addition to the weapons used in the shooting, Holmes booby-trapped his apartment, rigging trip wire to detonate 30 plastic shells stuffed with gunpowder, several glass jars filled

with gasoline and gunpowder, and 10 gallons of gasoline in canisters.

Outcome
Holmes was apprehended by the police in the theatre's rear parking lot within seven minutes of the first 911 calls from moviegoers. On July 30, 2012, Holmes appeared before the District Court of Arapahoe County, CO for formal charging on 142 counts. Later in the court process, the prosecution amended the charges to include 24 counts of murder in the first degree (two counts for each of the 12 victims killed); 140 counts of attempted murder in the first degree (two counts for each of the 70 victims injured); one count of possession of explosive or incendiary devices; and one count of unlawful use of a deadly weapon in the commission of a violent crime. On June 4, 2013, Holmes changed his original plea of not guilty to a plea of not guilty by reason of insanity. Trial began on April 27, 2015, and on July 16, 2015, the jurors found Holmes guilty on 24 counts of murder in the first degree, 134 counts of attempted murder in the first degree, 6 counts of the lesser included offense of attempted murder in the second degree, one count of possession of explosive or incendiary devices; and one count of unlawful use of a deadly weapon in the commission of a violent crime. On August 27, 2015, Holmes was sentenced to 12 consecutive life imprisonment sentences without the possibility of parole plus 3,318 years imprisonment.

September 6, 2011

Carson City, NV

Shooter
Eduardo Sencion, 32

Ammo Magazine Capacity
30-rounds

Shots Fired unknown
Killed 4 (plus shooter = 5)
Wounded 7

Carson City IHOP

Incident
At about 9 AM, Sencion entered an IHOP restaurant and began shooting at a table of uniformed National Guard members. He hit all 5 of the members, in addition to 5 civilians inside the restaurant. He eventually moved out into the parking lot, where he shot one woman before turning the gun on himself. Though his eight-minute rampage seemed focused on the Guardsmen, Sencion had no known association with the military and his motives remain unknown. He had no criminal record, but his family has indicated that he had a history of mental illness.

Weapons
AK-47 type assault rifle equipped with a 30-round large capacity ammunition magazine. Two additional guns and two more magazines were found in his vehicle.

Outcome
Suicide.

July 7, 2011

Grand Rapids, MI

Shooter
Rodrick Shonte Dantzler, 34

Ammo Magazine Capacity
30-rounds

Shots Fired >10
Killed 7 (plus shooter = 8)
Wounded 2

Grand Rapids

Incident
On a Thursday afternoon, Dantzler went to two homes on a shooting rampage, killing two ex-girlfriends and members of their families, including his own ten-year-old daughter and another child. He then led police on a high-speed chase, shooting two bystanders before crashing his car into an embankment. Dantzler fled, forced his way inside a nearby home, and held three occupants hostage for four hours before shooting himself in the head at about 11:30 PM. He had been arrested once before for assault with intent to do great bodily harm.

Weapons
GLOCK 9mm semiautomatic pistol (unknown model) equipped with a 30-round large capacity ammunition magazine.

Outcome
Suicide.

January 8, 2011

Tucson, AZ

Shooter
Jared Lee Loughner, 22

U.S. Rep. Gabriel Giffords Congress on Your Corner

Incident
During an outdoor constituent meet-and-greet at a Tucson grocery store, Loughner allegedly attempted to assassinate Rep. Giffords, and in the process murdered 6 and wounded 12 others. He first shot Rep. Giffords in the head from about three feet away and then turned to the crowd,

Ammo Magazine Capacity
33-rounds
15-rounds

Shots Fired 33
Killed 6
Wounded 13

firing over 30 rounds in just 15 seconds. Among those killed include a federal judge, Hon. John M. Roll, congressional staff, and civilians ranging in age from 9 to 79.

Weapons
GLOCK 19 9mm semiautomatic pistol equipped with a 33-round large capacity ammunition magazine. Loughner was also carrying two 15-round large capacity ammunition magazines, and a knife. The ATF determined Loughner legally purchased the GLOCK pistol with an extended magazine and one box of Winchester ammunition on November 30, 2010, from Sportsman's Warehouse in Tucson.

Outcome
Loughner was tackled while attempting to reload his firearm with another large capacity ammunition magazine. He was later taken into custody by Sheriff's deputies at the scene. The day following the shooting, Loughner was charged with five federal counts to which he pleaded not guilty. On March 4, 2011, he was charged with an additional 49 federal charges, to which he also pleaded not guilty. On May 25, 2011, Loughner was found not mentally competent to stand trial. A federal judge ruled on September 28, 2011, that efforts to treat him for mental illness in a federal facility should continue until he is mentally fit to be tried. Loughner was diagnosed with and treated for schizophrenia. After he was found mentally competent to stand trial, Loughner pleaded guilty on August 7, 2012, to 19 counts related to the date of the shooting. On November 8, 2012, Loughner was sentenced to seven consecutive life terms, plus 140 years in prison without the possibility of parole (one life term for the attempted assassination of Congresswoman Gabrielle Giffords; two life terms for the murder of two federal employees; four life terms for the murders of four participants at the event; two 20 year terms for the attempted murders of two federal employees; and ten 10 year terms for causing the injuring through the use of a firearm of ten participants at the event).

August 3, 2010

Manchester, CT

Shooter
Omar Thornton, 34

Ammo Magazine Capacity
17-rounds

Shots Fired >11
Killed 8 (plus shooter = 9)
Wounded 2

Hartford Beer Distributor

Incident
Thornton arrived at work early in the morning for a meeting with his employers. During the meeting he was shown video surveillance which proved he had been stealing beer from the company. Thornton was offered the choice to either resign from his position as a truck driver or be fired. Following the meeting, Thornton went into the employee kitchen to retrieve two handguns equipped with 17-round large capacity ammunition magazines he had previously hidden. He then traveled through the Distributor warehouse shooting deliberately. During the rampage, he murdered eight co-workers and wounded two more. Thornton eventually hid in a far office where he called the police to explain his motive prior to committing suicide. In his 911 call, Thornton claimed that the Hartford Beer Distributor was a "racist place." As he told the 911 dispatcher, "They treat me bad over here and they treat all the other black employees bad over here too."

Weapons
Two Ruger SR9 9mm semiautomatic pistols equipped with 17-round magazines. Thornton purchased both firearms legally from an East Windsor, CT gun dealer.

Outcome
Suicide.

November 5, 2009

Fort Hood, TX

Shooter
Nidal Malik Hasan, 39

Ammo Magazine Capacity
30-rounds
20-rounds

Shots Fired 214
Killed 13

Fort Hood

Incident
On the afternoon of November 5, 2009, Major Nidal Malik Hasan—an army psychiatrist—walked into a medical processing center and began firing upon those inside. The rampage began at 1:20 pm, and lasted for about four minutes, during which Hasan fired off about 214 shots, killing 13 and wounding 32 more. After running outside the building to chase down a wounded soldier, Hasan was confronted by a police officer. Engaging in a brief firefight, the officer managed to down Hasan with a shot to the torso. Reports have linked the incident to domestic terrorism.

Weapons
FN Herstal 5.7 Tactical Pistol equipped with 20-round large capacity ammunition magazine. When Hasan was apprehended, investigators found in his possession 177-rounds in 30-round

Wounded 32

and 20-round large capacity ammunition magazines, another handgun, a revolver, and two gunsights (for different lighting conditions). Hasan purchased the FN Herstal 5.7 Tactical Pistol legally at Guns Galore, a shop in Killeen, TX.

Outcome
After he was shot, Hasan was arrested. In 2009, he was charged with 13 counts of premeditated murder and 32 counts of attempted murder under the Uniform Code of Military Justice. In August 2013, following a 22-day court-martial, during which he represented himself, Hasan was convicted of all charges. He was sentenced to the death penalty.

April 3, 2009

Binghamton, NY

Shooter
Jiverly Wong, 41

Ammo Magazine Capacity
30-rounds
15-rounds

Shots Fired 99
Killed 13 (plus shooter = 14)
Wounded 4

American Civic Association

Incident
Armed with two handguns and 30- and 15-round large capacity ammunition magazines, Wong drove to the American Civic Association building, where he previously took classes. He first barricaded the back entrance of the building with a borrowed car, then entered through the front entrance and began firing. He first opened fire on the association's receptionists, killing one and wounding the other. The surviving receptionist, Shirley DeLucia, feigned death and, after Wong moved further into the building, called 911. Meanwhile, Wong entered a classroom and resumed fire, killing 12 and wounding 3 students and association workers, before eventually turning his gun on himself. His exact motives remain unclear; however, a letter he wrote a month prior to the attack indicates great frustration both with the police and with his lack of employment.

Weapons
Beretta .45-caliber semiautomatic pistol, Beretta 9mm semiautomatic pistol (models unknown), and two 30-round large capacity ammunition magazines and two 15-round large capacity ammunition magazines.

Outcome
Suicide.

February 14, 2008

DeKalb, IL

Shooter
Steven Phillip Kazmierczak, 27

Ammo Magazine Capacity
33-rounds
15-rounds

Shots Fired 54
Killed 5 (plus shooter = 6)
Wounded 21

Northern Illinois University

Incident
Armed with four firearms and 33- and 15-round large capacity ammunition magazines, graduate student Steven Kazmierczak kicked in the door of a Cole Hall lecture room and began firing on the 162-person class. Firing approximately 54 shots, he killed 5 students and wounded 17 others, before taking his own life. Kazmierczak had a history of mental illness, erratic behavior, and self-mutilation, and had reportedly stopped taking his medication in the weeks leading up to the shooting.

Weapons
SIG SAUER Kurz 9mm semiautomatic pistol, Hi-Point CF380 .380 caliber semiautomatic pistol, GLOCK 19 9mm semiautomatic pistol, Remington Sportsman 48 12-gauge shotgun, and 33-round and 15-round large capacity ammunition magazines. Kazmierczak purchased all four weapons from Tony's Gun & Ammo in Champaign, IL between August 3, 2007 and February 9, 2008. Kazmierczak also purchased gun accessories from a website operated by TGSCOM, Inc., the same company patronized by the VA Tech shooter.

Outcome
Suicide.

December 5, 2007

Omaha, NE

Shooter
Robert Hawkins, 19

Westroads Mall

Incident
Armed with an assault rifle and two 30-round large capacity ammunition magazines, Hawkins opened fire from the third floor balcony of the Westroads Mall. He killed six employees and two customers, and wounded five more, before taking his own life. Police arrived on the scene about six minutes after the shooting began, by which time it was already over. Hawkins had a history of mental illness and a criminal record. Police say the shooting was random.

Ammo Magazine Capacity
30-rounds

Weapons
WASR-10 semiautomatic assault rifle and two 30-round large capacity ammunition magazines.

Shots Fired >14
Killed 8 (plus shooter = 9)
Wounded 5

Outcome
Suicide.

April 16, 2007

Virginia Polytechnic Institute and State University

Blacksburg, VA

Shooter
Seung-Hui Cho, 23

Ammo Magazine Capacity
15-rounds

Shots Fired 176
Killed 32 (plus shooter = 33)
Wounded 17

Incident
At about 7 AM, Cho entered West Ambler Johnston dormitory, shot and killed two students, then returned to his dormitory to change out of his bloody clothes. At approximately 9:40 AM, he entered Norris Hall and began shooting at students and faculty in classrooms on the second floor. The rampage—during which 30 more people were killed and 17 wounded—lasted until approximately 9:51 AM, when Cho committed suicide. Exact motives remain unclear. Cho had a long history of mental and physical illness, depression, selective mutism, and wrote "dark and troubling" papers for his classes, which included fantasies about the Columbine shooting.

Weapons
GLOCK 19 9mm semiautomatic pistol and Walther P22 .22-caliber semiautomatic pistol. Investigators found a total of 17 empty magazines at the scene of the shooting, a mix of several 15-round, and 10-round magazines loaded with hollow-point rounds (bullets with the tip hollowed out, designed to expand upon impact). He possessed over 400 rounds of ammunition. Cho ordered the Walther P22 from a website operated by TGSCOM, Inc. Kazmierczak patronized the same company before the NIU shooting. On February 9, 2007, Cho picked up the pistol from J-N-D Pawn-brokers, located across the street from the VA Tech campus. In compliance with the state law limiting handgun purchases to one every 30 days, Cho purchased the GLOCK 19 on March 13, 2007. He also purchased five 10-round magazines from eBay in March. Cho's purchase of these firearms was in violation of federal law; he was disqualified from purchasing or possessing a firearm and ammunition, because a special justice of the Montgomery County General District Court had found him to be a danger to himself on December 14, 2005.

Outcome
Suicide.

January 30, 2006

Santa Barbara Postal Processing and Distribution Center

Goleta, CA

Shooter
Jennifer San Marco, 44

Ammo Magazine Capacity
15-rounds

Shots Fired unknown
Killed 7 (plus shooter = 8)
Wounded 0

Incident
On the night of January 30, 2006, Jennifer San Marco sneaked into a Santa Barbara condominium where she shot and killed a former neighbor. Less than an hour later, her rampage continued at the Santa Barbara Postal Processing and Distribution Center where she had worked for about six years. Armed with a semiautomatic handgun equipped with a 15-round large capacity ammunition magazine, San Marco shot six postal employees (two in the parking lot and four in the building), before turning the gun on herself. Five victims died at the scene and one died in the hospital two days later. San Marco's employment at the postal facility ended in 2003 when she was placed on retirement disability for psychological reasons. No suicide note was left to explain her motive, but police reportedly found writings in San Marco's New Mexico home (where she moved in 2004) alluding to a conspiracy plot involving the postal facility where the shooting occurred, a local medical facility, and the Santa Barbara County Sheriff's Department.

Weapons
Smith & Wesson 915 9mm semiautomatic handgun equipped with a 15-round large capacity ammunition magazine. San Marco purchased the firearm at a pawn shop in New Mexico in August 2005.

Outcome
Suicide.

November 21, 2004	Hunting Camp
Meteor, WI	Incident On a hunting trip in Northwest Wisconsin, at about noon on a Sunday, Vang was sitting in a hunting stand used to look out for deer, when he encountered a group of other hunters who informed him that he was trespassing on private property. Police report that Vang began to walk away, then turned, and opened fire. During the course of the shooting, he shot nine people, five of whom died during the incident (the sixth victim succumbed to the gunshot wounds the following day). One of the wounded victims recorded the hunting license number posted on Vang's orange vest and supplied it to police.
Shooter Chai Vang, 36	
Ammo Magazine Capacity 20-rounds	
Shots Fired 20 Killed 6 Wounded 3	Weapons SKS 7.62mm semiautomatic assault rifle equipped with a 20-round large capacity ammunition magazine. Outcome At about 5 PM that same day, police arrested Vang. At Vang's preliminary hearing, he pleaded not guilty to six counts of murder and three counts of attempted murder. During the trial, which lasted from September 11 to 18, 2005, Vang's defense argued that he had felt "under siege" from the other hunters, and that they had been using racial slurs against him. Vang was convicted of murder and eventually sentenced to six life sentences without the possibility of parole.

December 26, 2000	Edgewater Technology Office
Wakefield, MA	Incident Armed with multiple firearms and a 60-round large capacity ammunition magazine, McDermott arrived at his workplace at about 9 AM. After about two hours, he began his rampage by walking to the reception desk and shooting and killing the office manager. He moved throughout the building continuing to shoot at specific coworkers, firing 37 shots over the course of five to six minutes before he stopped firing, returned to the reception area and sat down. Authorities speculated that McDermott's motive centered on anger that his wages were to be collected by the IRS for the payment of back taxes.
Shooter Michael McDermott, 42	
Ammo Magazine Capacity 60-rounds	
Shots Fired 37 Killed 7 Wounded 0	Weapons AK-47-type semiautomatic assault rifle, unknown make and model 12-gauge shotgun, unknown make and model .32-caliber semiautomatic pistol, and 60-round large capacity ammunition magazine. Outcome McDermott was arrested at the scene. He was charged with seven counts of murder, to which he pleaded not guilty. Over the course of a 14-day trial in April 2002, McDermott's defense was based on insanity. During his testimony, he expressed a belief that he had been sent back in time to kill Nazis, a move which the prosecution claimed to be a fabricated "psychic alibi." At the end of the trial, McDermott was convicted of seven counts of murder and received seven life sentences.

November 2, 1999	Xerox Office Building
Honolulu, HI	Incident Armed with a handgun and three 17-round large capacity ammunition magazines, Uyesugi entered offices of the Xerox corporation in Honolulu and commenced firing. After firing approximately 28 shots, killing 7 people (he missed an 8th), Uyesugi promptly left and drove to the Hawaii Nature Center. After a 5-hour standoff with police, he surrendered. Uyesugi is said to have been a disgruntled employee—with a history of anger issues—who at the time was feeling work-related pressure.
Shooter Byran Uyesugi, 40	
Ammo Magazine Capacity 17-rounds	
Shots Fired 28 Killed 7 Wounded 0	Weapons GLOCK 17 9mm semiautomatic pistol and three 17-round large capacity ammunition magazines, loaded with hollow point bullets (bullets with the tip hollowed out, designed to expand upon impact). Uyesugi legally purchased the GLOCK in 1989. Outcome On November 9, 1999, Uyesugi was indicted on nine felony counts, including one count of first

degree murder, seven counts of murder in the second degree, and one count of attempted murder in the second degree. On May 15, 2000, the trial against Uyesugi began. He pleaded not guilty by reason of insanity, but the jury rejected that plea and found him guilty. Uyesugi was sentenced to life without the possibility of parole. In 2002, he appealed his conviction but the State of Hawai'i Supreme Court upheld his conviction.

September 15, 1999

Fort Worth, TX

Shooter
Larry Gene Ashbrook, 47

Ammo Magazine Capacity
15-rounds

Shots Fired 30
Killed 7 (plus shooter = 8)
Wounded 7

Wedgwood Baptist Church

Incident
Armed with two handguns and three 15-round large capacity ammunition magazines, Ashbrook walked into Wedgwood Baptist Church during a teen rally and began shooting. He killed 7 people (three of whom were teenagers) and wounded 7 more. Over the course of the attack, he fired approximately 30 shots and threw a pipe bomb in the church. Ashbrook then committed suicide. According to witnesses, during the shooting Ashbrook was yelling anti-religious invectives. In addition, a news report described him as one who "seethed with hostility, distrusted neighbors, and sometimes victimized the vulnerable."

Weapons
Ruger P85 9mm semiautomatic pistol, unknown make and model .380 caliber semiautomatic pistol, and three 15-round large capacity ammunition magazines. Ashbrook legally acquired both weapons from federally licensed firearms dealers in 1992.

Outcome
Suicide.

April 20, 1999

Littleton, CO

Shooter
Eric Harris, 18
Dylan Klebold, 17

Ammo Magazine Capacity
52-rounds
32-rounds
28-rounds

Shots Fired 188
Killed 13 (plus shooters = 15)
Wounded 24

Columbine High School

Incident
On the morning of April 20th, Harris and Klebold entered Columbine High School and placed two propane bombs in the cafeteria. They then returned to their cars, awaiting detonation. After the bombs failed to detonate, Harris and Klebold gathered their guns and large capacity ammunition magazines ranging from 28- to 52-rounds, they then approached the school's west entrance. At approximately 11:20 AM, they begin shooting at students outside the school. After entering the school, they commenced shooting and throwing pipe bombs at random, eventually proceeding to the library where they killed 10 and injured 12 more. Leaving the library, they continued wandering about the school, occasionally firing through windows at law enforcement, until—at around noon—they committed suicide. In less than an hour, Harris and Klebold killed 13 and wounded 24.

Weapons
Savage Springfield 67H 12-gauge pump-action shotgun, Savage Stevens 311D 12-gauge sawed-off shotgun, Hi-Point 995 9mm semiautomatic rifle, INTRATEC TEC-DC9 9mm semiautomatic pistol, and thirteen 10-round magazines, one 52-, one 32-, one 28-round large capacity ammunition magazines. Harris and Klebold illegally acquired the shotguns and Hi-Point rifle through a "straw purchase" (a transaction in which a legal buyer makes a purchase for someone who cannot legally purchase the firearm). Their friend, Robyn Anderson, purchased the three firearms at the Tanner Gun Show from unlicensed sellers in December of 1998. A pizza shop employee, Mark Manes, illegally sold them the INTRATEC TEC-DC9.

Outcome
Suicide.

May 20-21, 1998

Springfield, OR

Shooter
Kipland Philip "Kip" Kinkel, 15

Thurston High School

Incident
At about 3 PM, Kinkel, who had earlier been suspended from school for illegal possession of a firearm, loaded a .22-caliber rifle and shot his father in the back of the head. Roughly 3 hours later, Kinkel's mother returned home and he fatally shot her six times. The next morning, Kinkel armed himself with multiple weapons including a 50-round large capacity ammunition magazine, then drove to his school, arriving at about 7:55 AM. Walking through a school

Ammo Magazine Capacity
50-rounds

Shots Fired >50
Killed 4
Wounded 25

hallway, he shot 27 students, killing 2 of them, before he was finally tackled to the ground by other students while trying to reload.

Weapons
GLOCK 19 9mm semiautomatic pistol, Ruger (unknown model) .22-caliber semiautomatic pistol, Ruger (unknown model) .22-caliber rifle, and a 50-round large capacity ammunition magazine. The GLOCK and rifle were legally purchased by Kinkel's father.

Outcome
Kinkel was taken into custody by the police at the scene. On the 16th of June, Kinkel was indicted on 58 charges, 4 of which were for aggravated murder. In September of the following year, Kinkel pleaded guilty to the aggravated murder charges and 25 counts of attempted murder, and pleaded no contest to one attempted murder count. During his sentencing hearing, psychiatrists testified that Kinkel showed signs of schizophrenia. Evidence was also presented that he expressed admiration for the Westside Middle School shooting which occurred two months earlier. On November 2nd, Kinkel was sentenced to 111 years and 8 months in prison without the possibility of parole. In 2002, he appealed his sentence, but the Court of Appeals of Oregon found the sentence did not violate the Oregon Constitution. In 2007, he petitioned for a new trial, but a Marion County judge denied the motion. Kinkel then appealed that decision but on January 12, 2011, the Oregon Court of Appeals affirmed the trial court decision denying his motion for a new trial.

March 24, 1998

Jonesboro, AR

Shooter
Andrew Golden, 11
Mitchell Johnson, 13

Ammo Magazine Capacity
30-rounds
15-rounds

Shots Fired >26
Killed 5
Wounded 10

Westside Middle School

Incident
On the morning of March 24, Golden and Johnson stole a van owned by the Johnson family, drove to Golden's grandparents' house to acquire weaponry, including multiple 30- and 15-round large capacity ammunition magazines, and then continued on to Westside Middle School. Golden entered the school and pulled the fire alarm, then ran back outside to wait with Johnson. As students and teachers came running out of the school, the two boys opened fire, killing 5 (one of whom was a teacher) and wounding 10 (9 students and 1 teacher). Johnson claims Golden came up with the plan just to scare the kids who had bullied him.

Weapons
Universal M1 Carbine .30-caliber replica, Davis Industries .38-caliber two-shot derringer, Double Deuce Buddie .22-caliber two-shot derringer, Charter Arms .38-caliber revolver, Star .380-caliber pistol, FIE .380-caliber pistol, Ruger Security Six .357-caliber revolver, Ruger .44 magnum rifle, Smith & Wesson .38-caliber revolver, Remington 742 .30-06-caliber rifle, 15-round large capacity ammunition magazines, three 30-round large capacity ammunition magazines, and over 150-rounds of ammunition.

Outcome
After the shooting, Golden and Johnson ran into the woods and were eventually caught by police. The boys were convicted as juveniles to the maximum sentence possible under state law, imprisonment until they turned 18. Prior to their 18th birthdays, they were convicted of a federal crime for bringing a gun to school. They were then transferred to federal prisons until their 21st birthdays. Upon release they would have no criminal record, making them legally eligible to purchase a firearm. Johnson was released on August 11, 2005, and Golden was released on May 25, 2007.

March 6, 1998

Newington, CT

Shooter
Matthew Beck, 35

Ammo Magazine Capacity
19-rounds

Shots Fired >5
Killed 4 (plus shooter = 5)

Connecticut State Lottery Headquarters

Incident
Nearly two weeks after retuning to work following several months of "stress-related" medical leave, Beck, a State Lottery employee, arrived at work armed with a handgun equipped with a 19-round large capacity ammunition magazine. He shot and killed four of his bosses. As police arrived, Beck shot and killed himself. Beck had a history of depression and was disgruntled with his employer over a salary dispute and being passed over for a promotion.

Weapons
GLOCK model unknown 9mm semiautomatic pistol equipped with a 19-round large capacity ammunition magazine. Beck had a permit for the 9mm pistol used in the shooting.

Wounded 0

Outcome
Suicide.

December 18, 1997

Orange, CA

Shooter
Arturo Reyes Torres, 41

Ammo Magazine Capacity
30-rounds

Shots Fired 144
Killed 4 (plus shooter = 5)
Wounded 2

Caltrans Maintenance Yard

Incident
Armed with an assault rifle and five 30-round large capacity ammunition magazines, Torres fired 144 rounds in just over two minutes upon his former co-workers. He killed four, including his former supervisor, and wounded two more. Torres had recently been accused of stealing and selling government-owned materials and subsequently fired from his job at Caltrans. He is believed to have been seeking revenge against his former supervisor, who Torres felt set him up.

Weapons
Chinese-made AK-47-type 7.62mm semiautomatic assault rifle and five 30-round large capacity ammunition magazines. Torres legally purchased the rifle on April 30, 1988, from B&B Gun Sales in Orange County, CA.

Outcome
Torres was shot and killed by police.

June 20, 1994

Fairchild Air Force Base, WA

Shooter
Dean Allen Mellberg, 20

Ammo Magazine Capacity
75-rounds

Shots Fired unknown
Killed 5 (plus shooter = 6)
Wounded 23

Fairchild Air Force Base

Incident
Weeks after receiving an involuntary honorable discharge from the Air Force, Dean Allen Mellberg took a cab to the Fairchild Air Force Base hospital armed with a Mak-90 semiautomatic assault rifle equipped with a 75-round drum large capacity ammunition magazine. He shot and killed two doctors, who he reportedly blamed for his discharge from the military. Mellberg then fired upon others in the hospital, chasing some outside the building. Once outside he encountered a military police officer who fatally shot him. In the few minutes Mellberg was shooting, he killed 5 and wounded 23.

Weapons
Chinese-made Mak-90 semiautomatic assault rifle equipped with a 75-round drum large capacity ammunition magazine. He purchased the assault rifle on June 15, 1994, five days before the shooting, and the following day purchased 80 rounds of 7.62x39mm ammunition and a 75-round drum large capacity ammunition magazine.

Outcome
Shot and killed by military police.

December 7, 1993

Long Island, NY

Shooter
Colin Ferguson, 35

Ammo Magazine Capacity
15-rounds

Shots Fired 30
Killed 6
Wounded 19

Long Island Railroad

Incident
Armed with a handgun and four 15-round large capacity ammunition magazines, Ferguson boarded a 5:33 PM Long Island bound commuter train from NYC's Pennsylvania Station. During the journey he began firing on passengers. He emptied approximately 30 rounds upon 25 people, killing 6. Ferguson's motives for the shooting are believed to stem from a variety of complaints. Police discovered a notebook in which Ferguson vented his hatred for "Caucasians and Uncle Tom Negroes," then-Governor Mario Cuomo, and the state Workers' Compensation Board.

Weapons
Ruger P89 9mm semiautomatic pistol and four 15-round large capacity ammunition magazines. Ferguson legally acquired the weapon in California at an outlet of Turner's Outdoorsman.

Outcome
Stopping to reload, Ferguson was tackled by three train passengers. Ferguson was indicted on January 19, 1994. A lengthy and controversial trial ensued, during which Ferguson's lawyers—William Kunstler and Ronald Kuby—insisted that he was overcome with "black rage." Ferguson

rejected that defense and eventually dismissed Kunstler and Kuby. Maintaining his plea of not guilty, Ferguson was finally convicted of murder on February 17, 1995.

July 1, 1993

San Francisco, CA

Shooter
Gian Luigi Ferri, 55

Ammo Magazine Capacity
50-rounds
40-rounds

Shots Fired >75
Killed 8 (plus shooter = 9)
Wounded 6

101 California Street Office of Pettit & Martin Law Firm

Incident
Armed with three firearms and 40- and 50-round large capacity ammunition magazines, Ferri opened fire on the offices of the law firm Pettit & Martin on the 34th floor of a San Francisco high-rise. He fired between 75 to 100 rounds, killing eight and wounding six, before killing himself. Ferri—a real estate speculator undergoing major financial trouble—had previously hired the law firm. His exact motives remain unclear, but police found a letter written by Ferri indicating frustrations with Pettit & Martin over real estate advice they had given him in 1981.

Weapons
Two INTRATEC TEC-DC9 semiautomatic pistols, Colt (unknown model) .45-caliber semiautomatic pistol, and 40-round and 50-round large capacity ammunition magazines loaded with a mix of Black Talon and standard ammunition. According to the Las Vegas Metropolitan Police Department, Ferri purchased the pistols from two stores in Las Vegas: Super Pawn and Pacific Tactical Weapons.

Outcome
Suicide.

October 16, 1991

Killeen, TX

Shooter
George Hennard, 35

Ammo Magazine Capacity
17-rounds
15-rounds

Shots Fired 100
Killed 23 (plus shooter = 24)
Wounded 20

Luby's Cafeteria

Incident
Armed with two handguns and 17-round and 15-round large capacity ammunition magazines, Hennard crashed his pickup into Luby's Cafeteria during a busy lunch hour. Stepping out of the vehicle, he began shooting randomly, killing 23 and wounding 20. After firing approximately 100 shots over 10 minutes, Hennard shot himself in the head. His motives remain unclear, but neighbors described him as "combative and unstable."

Weapons
GLOCK 17 9mm semiautomatic pistol, Ruger P89 semiautomatic pistol, and 17-round and 15-round large capacity ammunition magazines. Hennard legally purchased the weapons from Mike's Gun Shop in Henderson, NV, in February and March of 1991.

Outcome
Suicide.

June 18, 1990

Jacksonville, FL

Shooter
James Edward Pough, 42

Ammo Magazine Capacity
30-rounds

Shots Fired >14
Killed 9 (plus shooter = 10)
Wounded 4

General Motors Acceptance Corporation Office

Incident
Armed with two firearms and a 30-round large capacity ammunition magazine, Pough opened fire in offices of General Motors. He killed nine and wounded four before taking his own life. It is believed Pough was angered by having his 1988 Pontiac Grand Am repossessed by the Corporation.

Weapons
Universal M1 .30-caliber semiautomatic assault rifle, unknown make and model .38-caliber revolver, and a 30-round large capacity ammunition magazine.

Outcome
Suicide.

September 14, 1989

Standard Gravure Corporation

Louisville, KY

Shooter
Joseph Wesbecker, 47

Ammo Magazine Capacity
30-rounds

Shots Fired >21
Killed 8 (plus shooter = 9)
Wounded 12

Incident
Armed with a duffle-bag full of firearms and 30-round large capacity ammunition magazines, Wesbecker opened fire at the offices of his former employer, shooting and killing 8 and wounding 12, before taking his own life. Wesbecker had been placed on permanent disability leave due to mental illness.

Weapons
Chinese-made AK-47-type semiautomatic assault rifle, two INTRATEC MAC-11 semiautomatic assault pistols, SIG SAUER unknown model 9mm semiautomatic pistol, unknown make and model .38-caliber revolver, and 30-round large capacity ammunition magazines. Wesbecker legally purchased the AK-47-type assault rifle from Tilford's Gun Sales in Louisville.

Outcome
Suicide.

January 17, 1989

Stockton, CA

Shooter
Patrick Purdy, 24

Ammo Magazine Capacity
75-rounds
35-rounds

Shots Fired 106
Killed 5 (plus shooter = 6)
Wounded 30

Cleveland Elementary School

Incident
Armed with two firearms and multiple 75- and 35-round large capacity magazines, Purdy first set his car on fire in the parking lot of Cleveland Elementary School. He then entered school grounds and began shooting. Over the course of the rampage, Purdy killed 5 students and wounded 30 others, including one teacher. After firing approximately 106 shots with an AK-47-type assault rifle over less than two minutes, he shot himself in the head with a pistol. Purdy's former acquaintances reported that he "developed a hate for everybody" including an intense dislike of Asian Americans. Of the five fatalities incurred during the Cleveland School Massacre, four were born in Cambodia and one in Vietnam.

Weapons
Chinese-made AK-47-type semiautomatic assault rifle, Taurus unknown model 9mm semiautomatic pistol, a 75-round large capacity ammunition drum magazine, a 75-round large capacity ammunition rotary magazine, and four 35-round large capacity ammunition banana magazines. Purdy legally purchased the AK-47-type rifle at Sandy Trading Post, in Sandy, OR on August 3, 1988, and the Taurus 9mm pistol at Hunter Loan and Jewelry Co. in Stockton, CA on December 28, 1988.

Outcome
Suicide.

April 23, 1987

Palm Bay, FL

Shooter
William Cruse, Jr., 59

Ammo Magazine Capacity
30-rounds

Shots Fired unknown
Killed 6
Wounded 10

Palm Bay shopping center

Incident
On April 23, 1987, William Cruse, Jr., loaded his car with a Strum, Ruger Mini-14 semiautomatic assault rifle equipped with a 30-round large capacity ammunition magazine, five 30-round large capacity ammunition magazines, 180 rounds of ammunition, a shotgun, and a pistol, and began to drive to a local shopping center. He first stopped at a neighbor's driveway, opened the car window, picked up his shotgun and opened fire upon two brothers and their father and mother, wounding one of the brothers. Cruse then continued on to the Palm Bay Center where he shot and killed three people and wounded three others with the assault rifle. He then drove across the street to the Sabal Palm Square shopping center, exited his car and again opened fire. As officers approached, Cruse reloaded his assault rifle and fired into the police car killing an officer. Another officer arrived and exited his police car, Cruse continued firing upon the officers, killing another officer. Cruse then fled into a grocery store firing upon the shoppers inside, killing one and wounding several more. He then found two women hiding in the restroom; he let one out of the store to negotiate with police and kept the other hostage. After several hours, Cruse released the hostage. Police then fired tear gas and stun grenades into the store, forcing Cruse out of the store and allowing officers to take him into custody. During the over 7 hour rampage, Cruse killed 6, including 2 police officers, and wounded 10 more. Police officers were so outgunned that a neighbor provided police an AR-15 assault rifle to help match Cruse's firepower.

Weapons

Strum, Ruger Mini-14 semiautomatic assault rifle equipped with a 30-round large capacity ammunition magazine, five 30-round large capacity ammunition magazines, 180 rounds of ammunition, a shotgun (unknown make and model), and a pistol (unknown make and model). Cruse ordered the assault rifle on March 21, 1987. On April 17, 1987, he purchased 100-rounds of ammunition and six 30-round large capacity ammunition magazines.

Outcome
Cruse was arrested at the scene. He pleaded not guilty by reason of insanity. In 2009, a jury in Polk County, FL, convicted Cruse of 6 counts of first-degree murder, 22 counts of attempted first-degree murder, 2 counts of attempted second-degree murder, 1 count of false imprisonment, and 1 count of kidnapping. In 1989, Curse was sentenced to the death penalty for the murders of the two officers and sentenced to consecutive life sentences for the other four murders and attempted murders. While on death row, Cruse died of natural causes in 2009.

July 18, 1984

San Ysidro, CA

Shooter
James Oliver Huberty, 41

Ammo Magazine Capacity
25-rounds

Shots Fired 257
Killed 21 (plus shooter = 22)
Wounded 19

McDonald's Restaurant

Incident
Armed with multiple firearms and 25-round large capacity ammunition magazines, Huberty entered the McDonald's restaurant and opened fire. He shot 40 people, killing 21 and wounding 19. He expended 257 rounds over 77 minutes, before being killed by a police sniper. No motive has been established. Prior to the shooting, Huberty told his wife, "I'm going hunting humans."

Weapons
Browning P-35 9mm semiautomatic pistol, Winchester 1200 pump-action 12-gauge shotgun, Israeli Military Industries 9mm Model A Carbine (Uzi), and 25-round large capacity ammunition magazines.

Outcome
Huberty was shot and killed by police.

June 29, 1984

Dallas, TX

Shooter
Abdelkrim Belachheb, 39

Ammo Magazine Capacity
14-rounds

Shots Fired unknown
Killed 6
Wounded 1

Ianni's Club

Incident
On June 29, 1984, after offending his dancing partner at a Dallas night club, Abdelkrim Belachheb, a Moroccan in the U.S. illegally, left the club and returned with a Smith & Wesson 9mm semiautomatic pistol equipped with a 14-round large capacity ammunition magazine. He emptied the magazine into his dance partner, reloaded and fired into the crowd. Belachheb killed his dance partner, five others, and wounded one more.

Weapons
Smith & Wesson (unknown model) 9mm semiautomatic pistol and two 14-round large capacity ammunition magazines.

Outcome
Belachheb surrendered to police hours later. He pleaded not guilty by reason of insanity. On November 15, 1984, a jury found Belachheb guilty of the six murders. He was sentenced to six consecutive life sentences plus 20 years, and $70,000 in fines.

* *Disclaimer*: Information for this database has been compiled from publicly available news sources. Every effort has been made to obtain the most accurate information, however, contradictions may exist between this database and other sources. As the ATF does not require police departments to collect data related to the capacity of a firearm's ammunition magazine, this database is not an exhaustive list of mass shootings involving large capacity ammunition magazines.

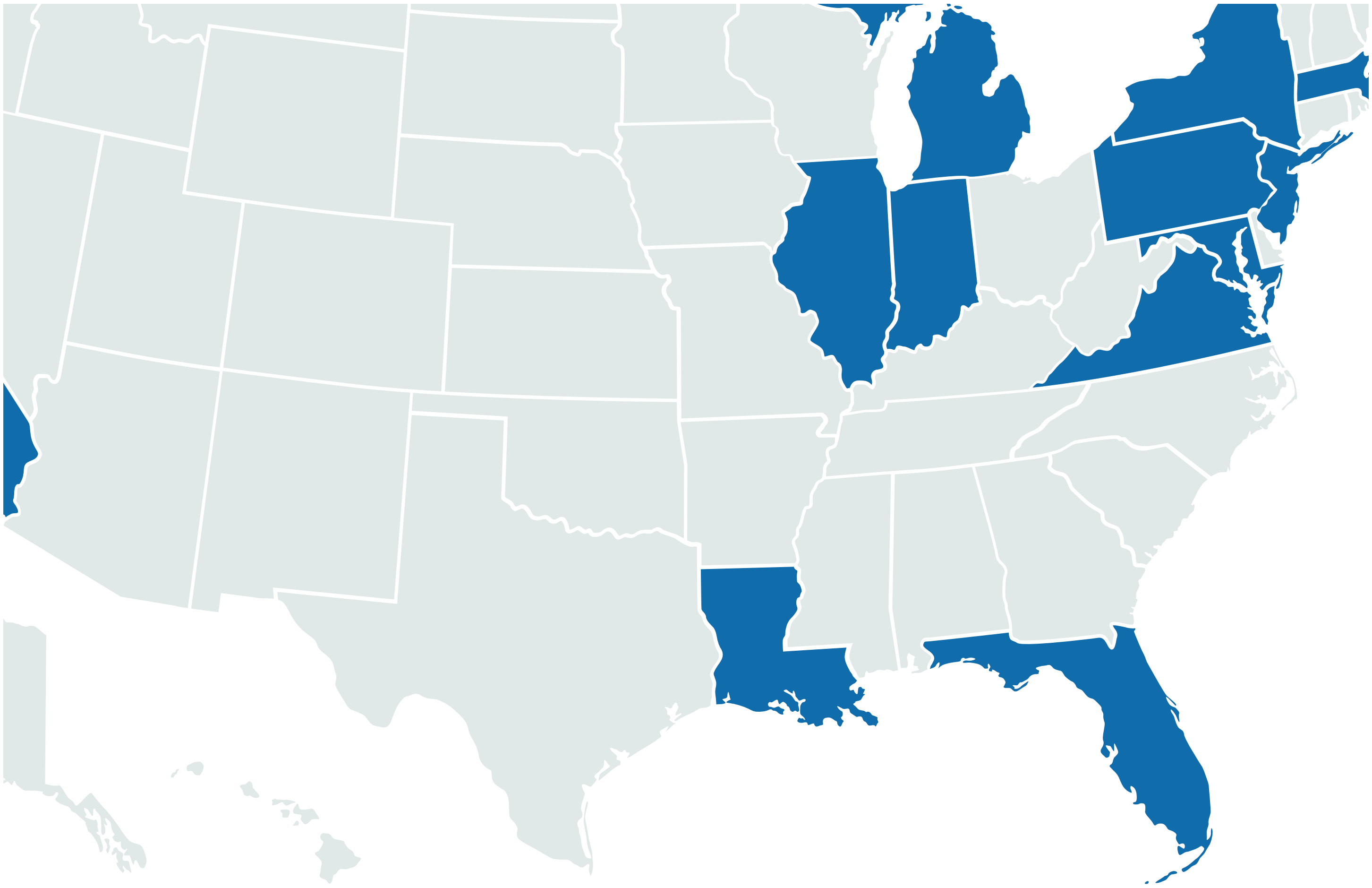
CURRENT LITIGATION

Donate

MENU

NRA CIVIL RIGHTS DEFENSE FUND

NRA Civil Rights Defense Fund



Filter Articles

Select a State

Search

RESET

Guns Save Lives, et al. v. Kwame Raoul.

This case, filed in Sangamon County, challenges the Illinois Firearm Owners Identification Card Act, which requires Illinois residents to secure a license and pay a fee in order to be permitted to own a firearm, even in their own homes.

Altman v. Santa Clara.

Brandy v. Villanueva.

In response to the COVID-19 outbreak, California and Los Angeles County issued an order closing down all non-essential business, which included gun shops. NRA (along with SAF and FPC) filed this lawsuit challenging the closure under the Second Amendment and Due Process Clause.

Doe, et al. v. Bonta.

Whenever a Californian purchases a firearm or ammunition or applies for a concealed carry license, they must undergo a background check and submit their name, address, place of birth, telephone number, occupation, California driver's license or ID number, race, sex, height, weight, hair color, eye color, and, in some instances, social security number to the California DoJ. The California DoJ then maintains that personal information in a database. Up until September 2021, California law only allowed that personal information to be used for law enforcement investigations. But in September 2021, the assembly passed and Governor Newsome signed AB 173 into law. That law mandates that the California DoJ disclose all of the personal information that it has on gun owners to the California Firearm Violence Research Center at UC Davis and any other "bona fide research institution." This lawsuit challenges AB 173 under the federal constitutional rights to privacy, the anti-retroactivity doctrine, the Second Amendment, and the Supremacy Clause (federal preemption).

Duncan, et al, v. Becerra

This case was filed in May 2017 in the Southern District of the United States District Court, challenging both a state bill and Proposition 63 which placed a ban on the possession of magazines that have a capacity of more than ten (10) rounds. The lawsuit challenges California's regulatory scheme against standard capacity magazines on the grounds that it violates the Second Amendment, Due Process Clause, and Takings Clause of the United States Constitution.

Ninth Circuit, on August 17, 2016, a lawsuit was filed in the United States District Court for the Central District, Western Division, of California, challenging California's prohibition of the right to publicly bear arms for self-defense by denying law-abiding citizens the license to carry concealed weapons nor allowing open carry. The lawsuit seeks to force the court to decide whether or not California's entire regulatory scheme prohibiting both open and concealed carry violates the Second Amendment.

Rhode, et al. v. Becerra.

California enacted ammunition sales restrictions, including requirements that all sales be conducted via face-to-face transactions, all ammunition sales be recorded with California's Department of Justice, and purchasers undergo a background check.

Rupp, et al. v. Becerra.

California prohibits the manufacture, possession, sale, transfer, or import of "assault weapons" within the state. In 2016, the law was expanded to include firearms whose magazines were detached by a "bullet button."

Rupp, et al. v. Becerra. National African American Gun Association, Inc. Amicus Curiae Brief.

The amicus curiae brief was filed on January 31, 2020.

City of Weston v. Scott; Daley v. Florida; Broward County v. Florida.

Florida law broadly preempts the regulation of firearms and ammunition by municipalities, and it imposes penalties on local officials and municipalities who violate the preemption statute. These three consolidated cases are brought by local officials and municipalities challenging the penalty provisions of Florida law.

Donate

MENU

NRA CIVIL RIGHTS DEFENSE FUND NRA Civil Rights Defense Fund

Learn More About the NRA Civil Rights Defense Fund

Browse through the site to discover how the NRA Civil Rights Defense Fund has provided millions of dollars in support of cases involving individuals and organizations defending the individual right to keep and bear arms and to support legal research and education

Current Litigation

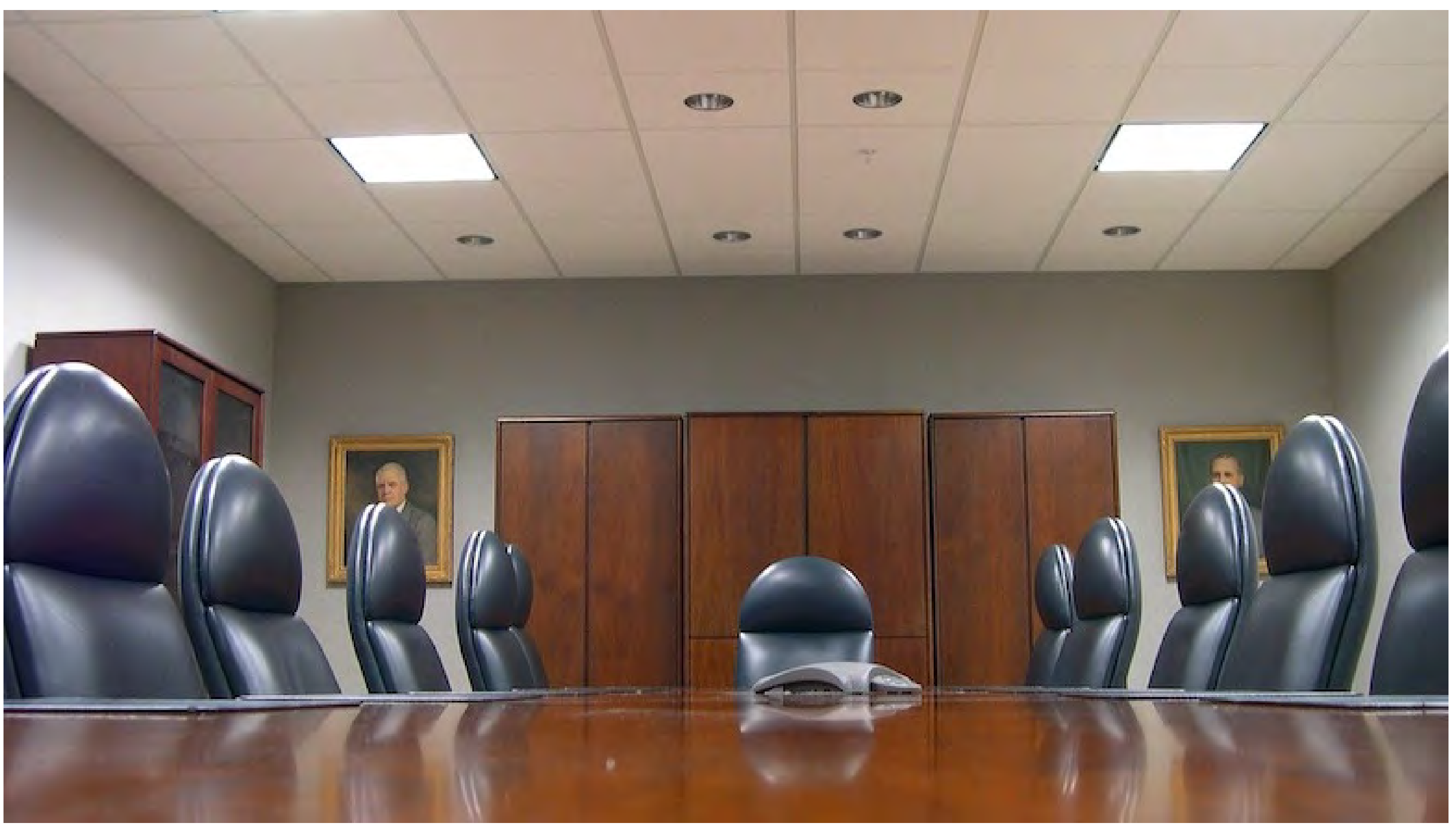
Donate

MENU

NRA CIVIL RIGHTS DEFENSE FUND

NRA Civil Rights Defense Fund

Eligible Cases



Officers and Board of Trustees

Donate

MENU

NRA CIVIL RIGHTS DEFENSE FUND NRA Civil Rights Defense Fund

Contact us by phone (703) 267-1250, fax 703-267-3985 or email

nracrdf@nrahq.org

The Washington Post

National • Analysis

The terrib [...] hat grow with ea [...] ting

Analysis Interpretation of the news based on evidence, including data, as well as anticipating how events might unfold based on past events.

MAY 9, 2021

Six people killed at a birthday party in Colorado Springs

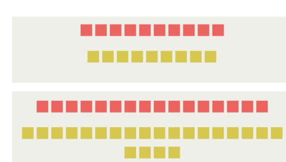
By Bonnie Berkowitz and Chris Alcantara
Updated May 12, 2021

***Editor's Note**: This page is no longer being updated.*

The places change, the numbers change, but the choice of weapon remains the same. In the United States, people who want to kill a lot of other people most often do it with guns.

Public mass shootings account for a tiny fraction of the country's gun deaths, but they are uniquely terrifying because they occur without warning in the most mundane places. Most of the victims are chosen not for what they have done but simply for where they happen to be.

RELATED

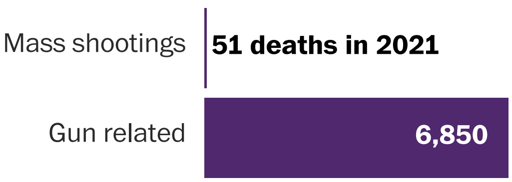


The number of U.S. mass shootings

depends on how you count ➔

There is no universally accepted definition of a public mass shooting, and this piece defines it narrowly. It looks at the **189 shootings** in which four or more people were killed, usually by a lone shooter. It does not include shootings tied to robberies that went awry, and it does not include domestic shootings that took place exclusively in private homes. A broader definition would yield much higher numbers.

Public mass shootings are a small slice of gun deaths



Source: Gun Violence Archive. Excludes the roughly 22,000 annual gun suicides, which are not publicly reported in real time.

This tally begins Aug. 1, 1966, when a student sniper fired down on passersby from the observation deck of a clock tower at the University of Texas. By the time police killed him, 17 other people were dead or dying. As Texas Monthly's Pamela Colloff wrote, the shooting "ushered in the notion that any group of people, anywhere — even walking around a university campus on a summer day — could be killed at random by a stranger."

Search for details of a particular shooting. The most recent is selected.

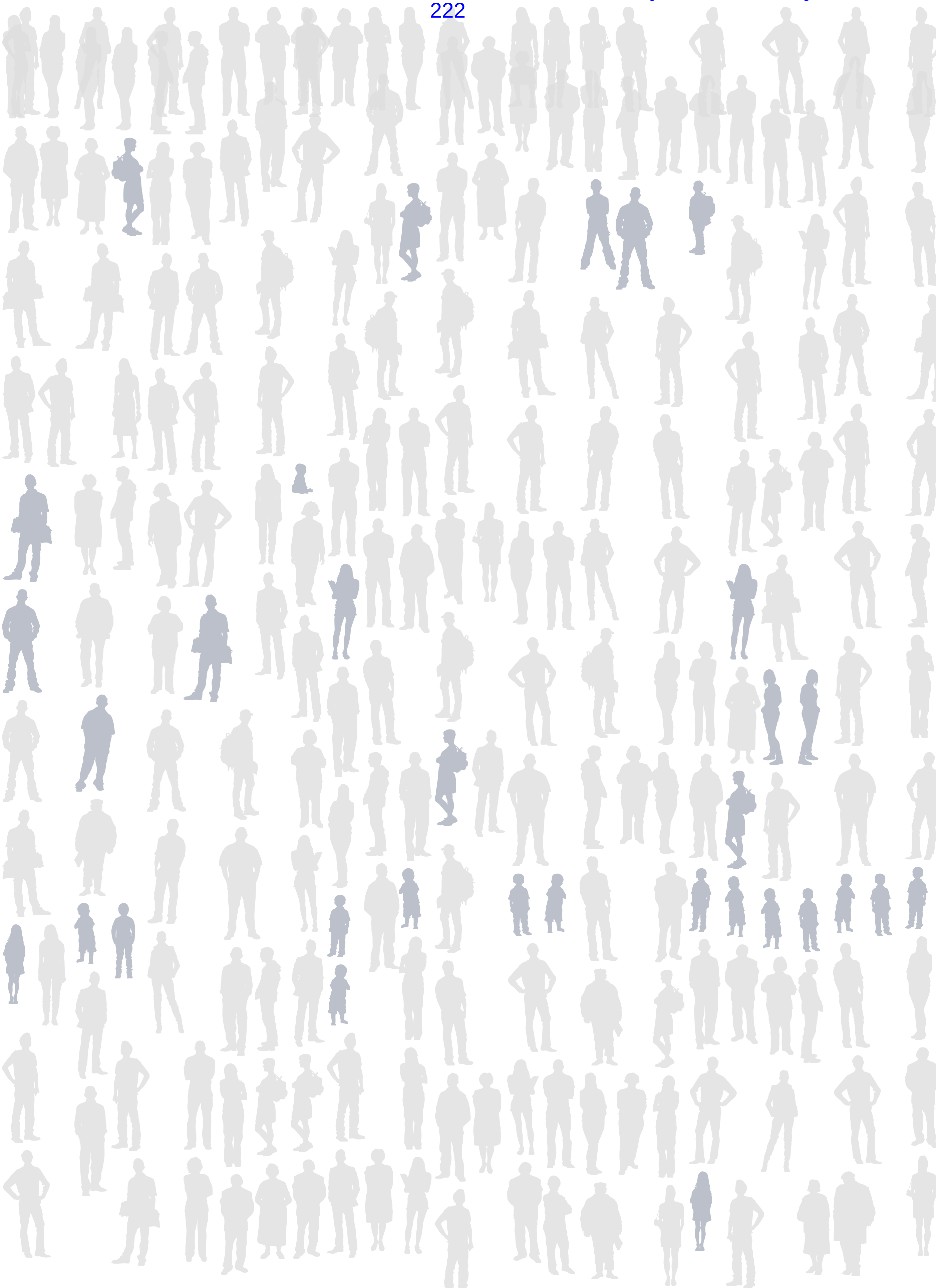
COLORADO SPRINGS – MAY 9, 2021
Canterbury Mobile Home Park shooting

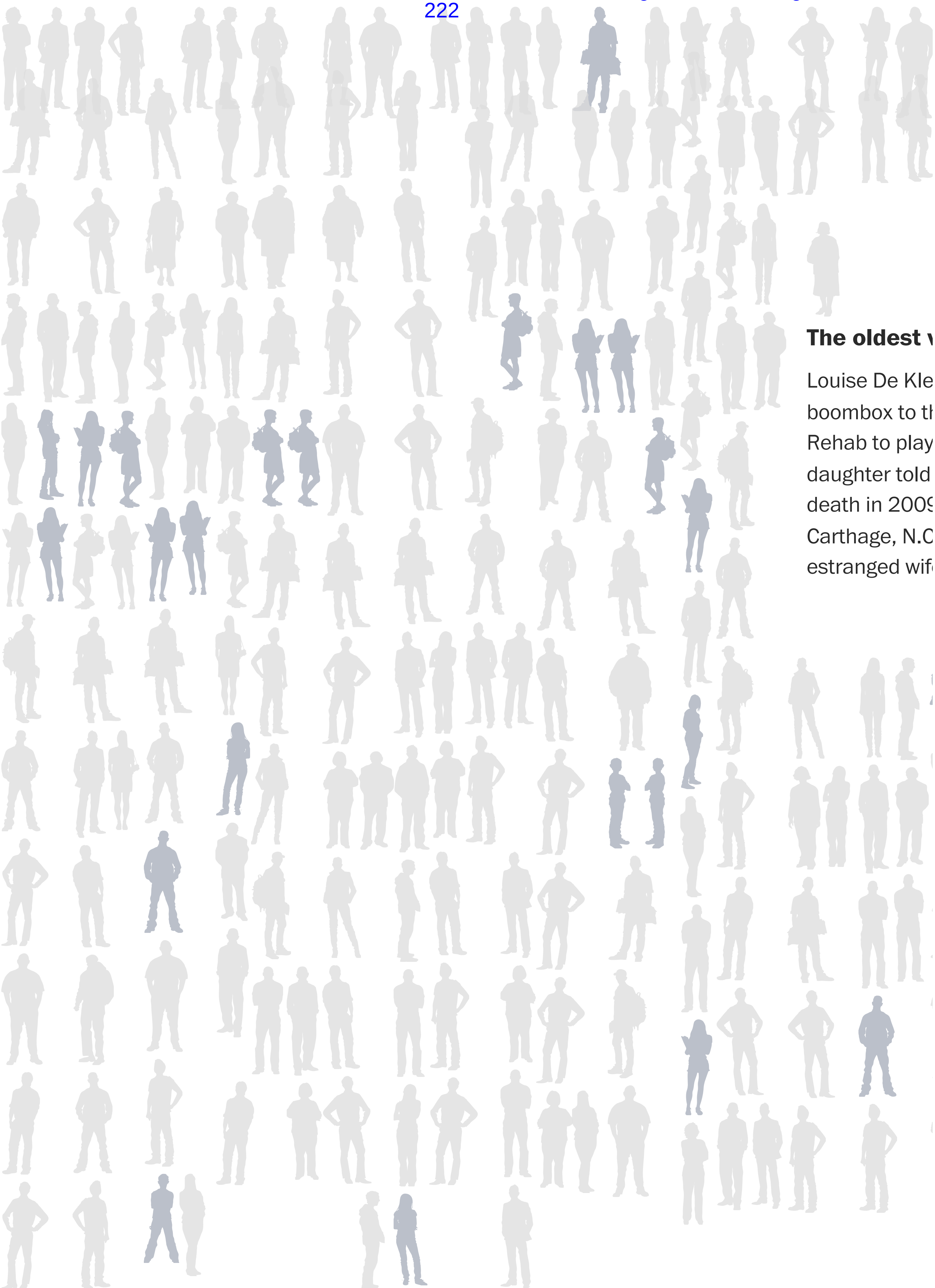
Victims: **6** killed and **0** injured Guns: **1** Shooters: **1 dead** **Read details**

1,322 killed

The people who were killed came from nearly every imaginable race, religion and socioeconomic background. Their ages range from the unborn to the elderly; **210 were children and teenagers**. In addition, thousands of survivors were left with devastating injuries, shattered families and psychological scars.

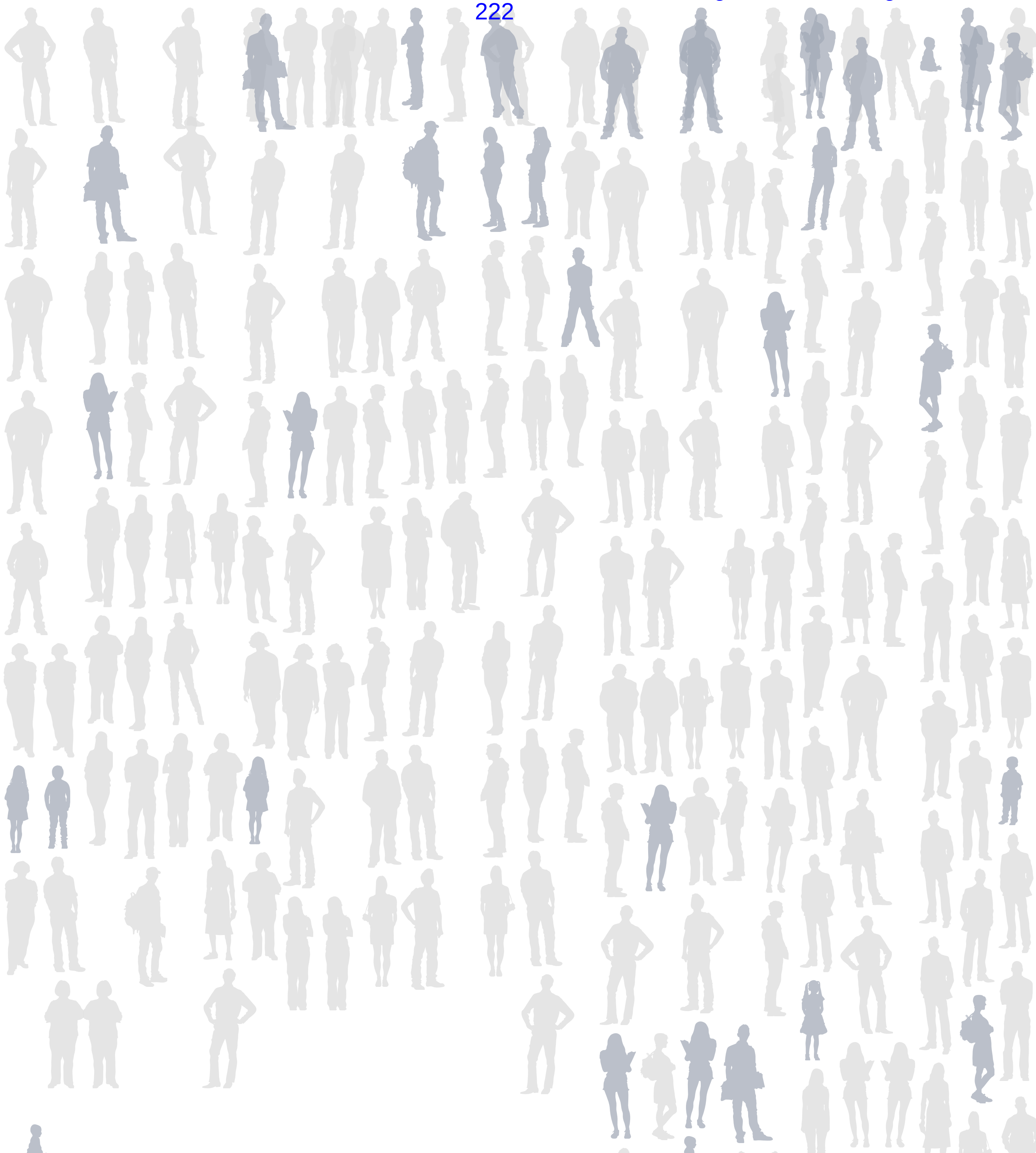
CLICK ON AN ICON FOR DETAILS ABOUT EACH VICTIM.





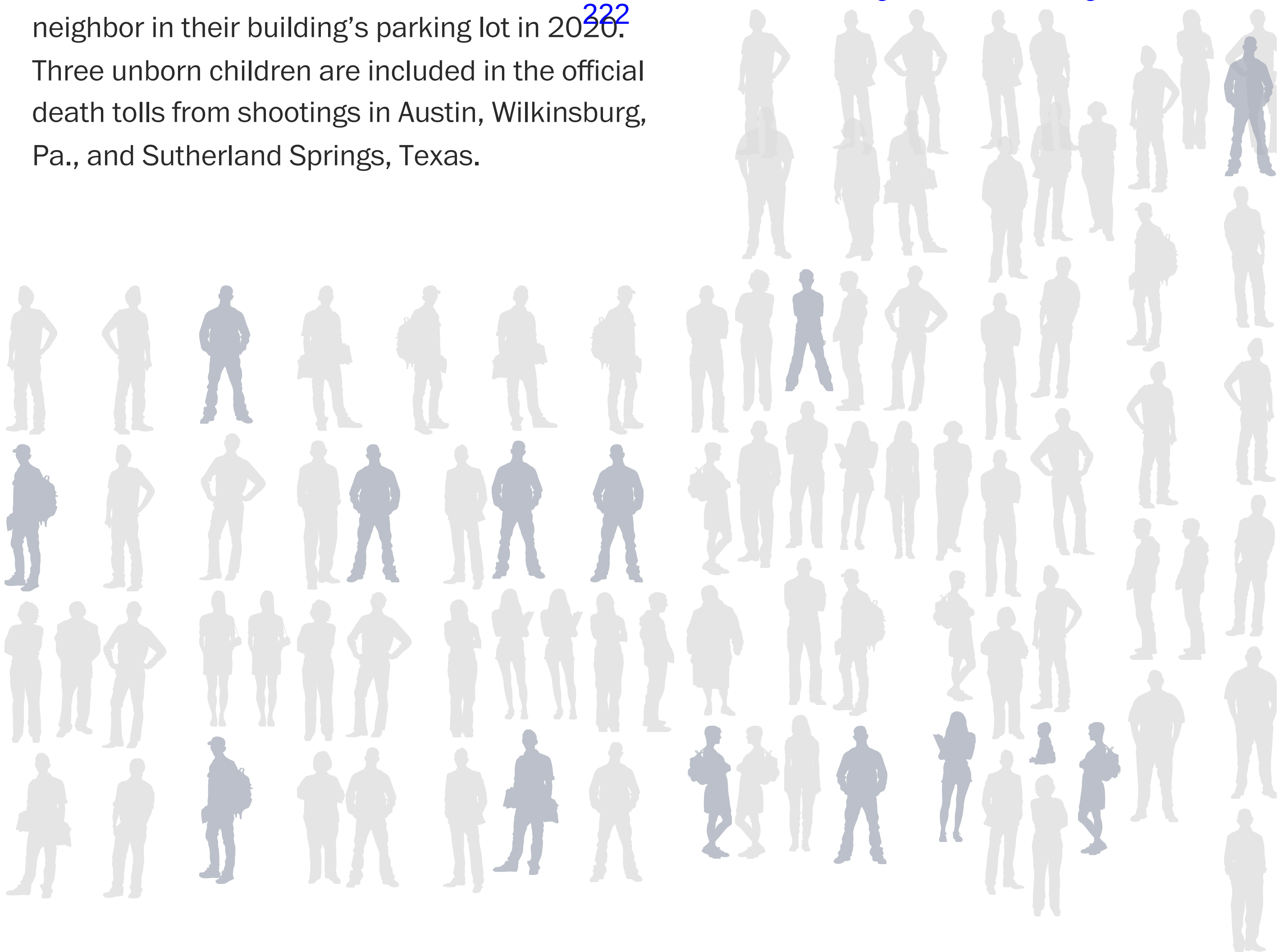
The oldest v

Louise De Kle
boombox to t
Rehab to play
daughter told
death in 200
Carthage, N.C
estranged wif



The youngest victims

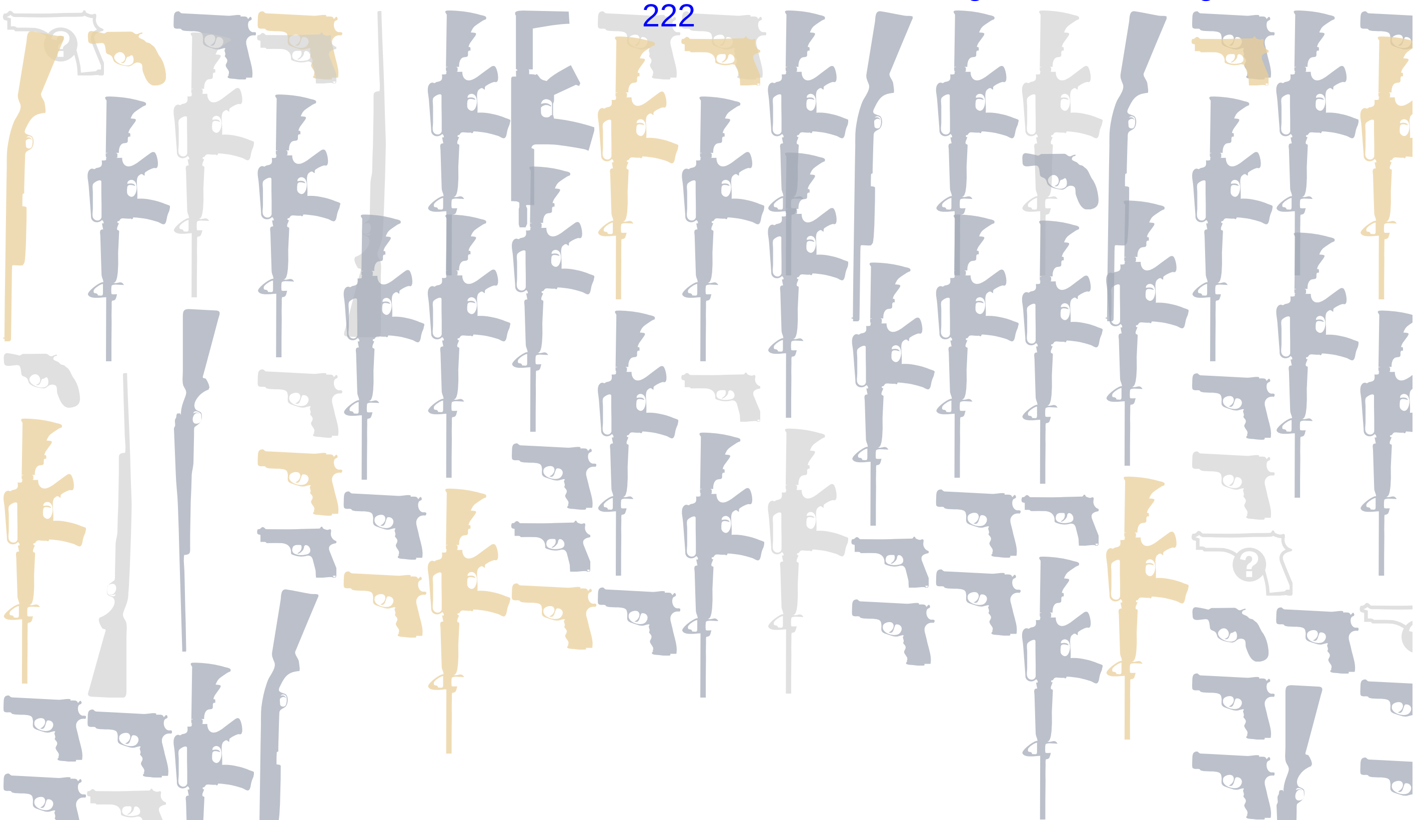
Eight-month-old Carlos Reyes was buried in a casket with his mother, Jackie, who had tried to shield him as an unemployed father of two opened fire at a busy McDonald's in San Ysidro, Calif., in 1984. Glory Tucker, 5 months, was killed along with her three siblings by her mother, who had just fatally shot a neighbor in their building's parking lot in 2020. Three unborn children are included in the official death tolls from shootings in Austin, Wilkinsburg, Pa., and Sutherland Springs, Texas.



360 guns

Shooters often carried more than one weapon; one was found with 24. At least **188 of mass shooters' weapons were obtained legally** and **62 were obtained illegally**. It's unclear how **110 weapons** were acquired.

SILHOUETTES REPRESENT A BASIC TYPE OF GUN RATHER THAN EXACT MAKES OR MODELS. CLICK ON AN ICON FOR DETAILS ABOUT EACH GUN.



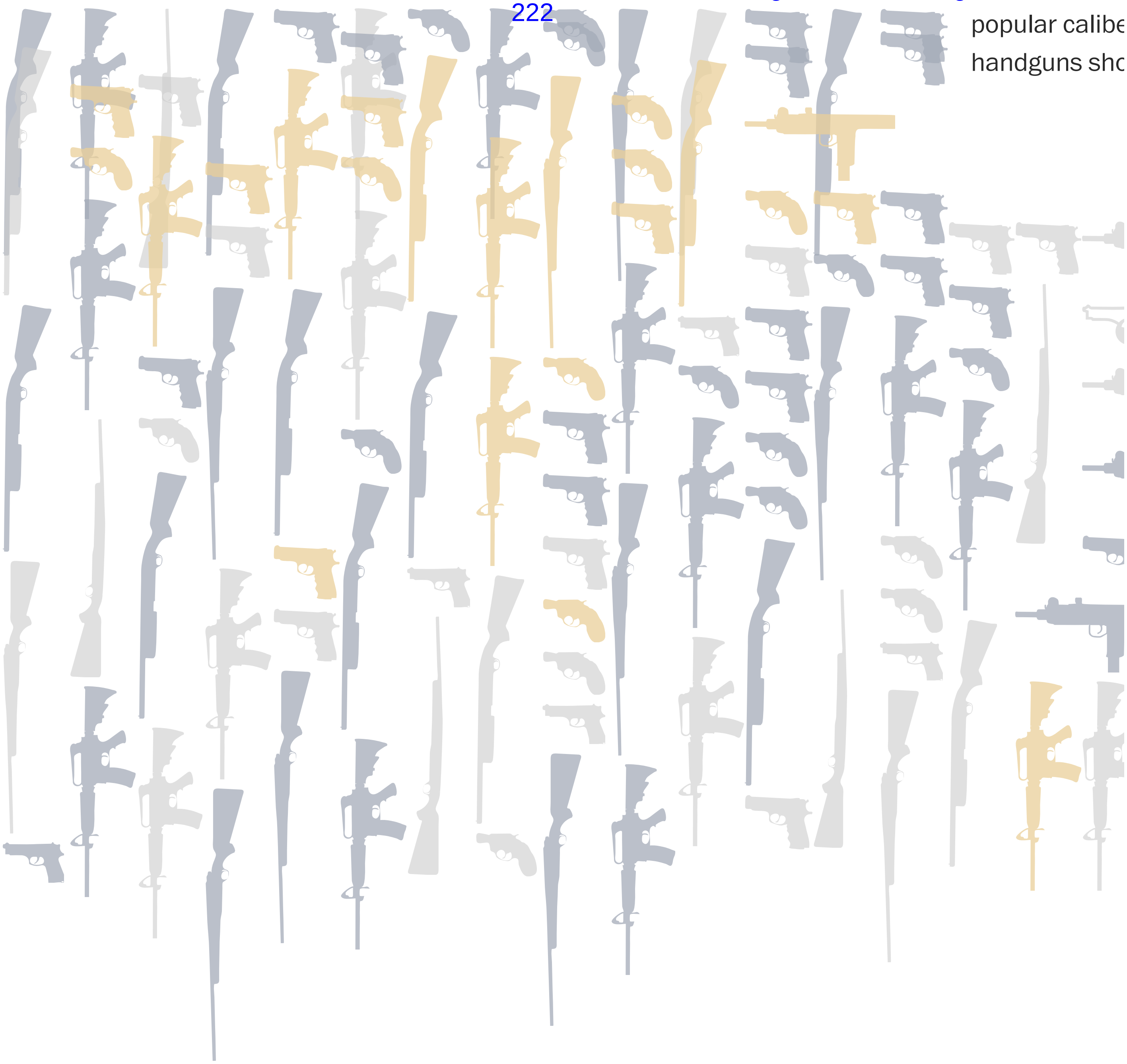
Semiautomatic rifles

Semiautomatic rifles have been used in some of the country's deadliest shootings, such as those in Newtown, Orlando, San Bernardino and Las Vegas. The AR-15, a lightweight, customizable version of the military's M16, soared in popularity after a 10-year federal ban on assault weapons expired in 2004. Some of the Las Vegas shooter's guns had been fitted with legal devices called "bump-fire stocks," which allow semiautomatic rifles to fire as quickly as automatic ones.

Semiautoma

The country's
semiautomati
enforcement
inexpensive, e
fire as quickly
gunman who
Virginia Tech i
Glock 19 (and

popular calibe
handguns sho



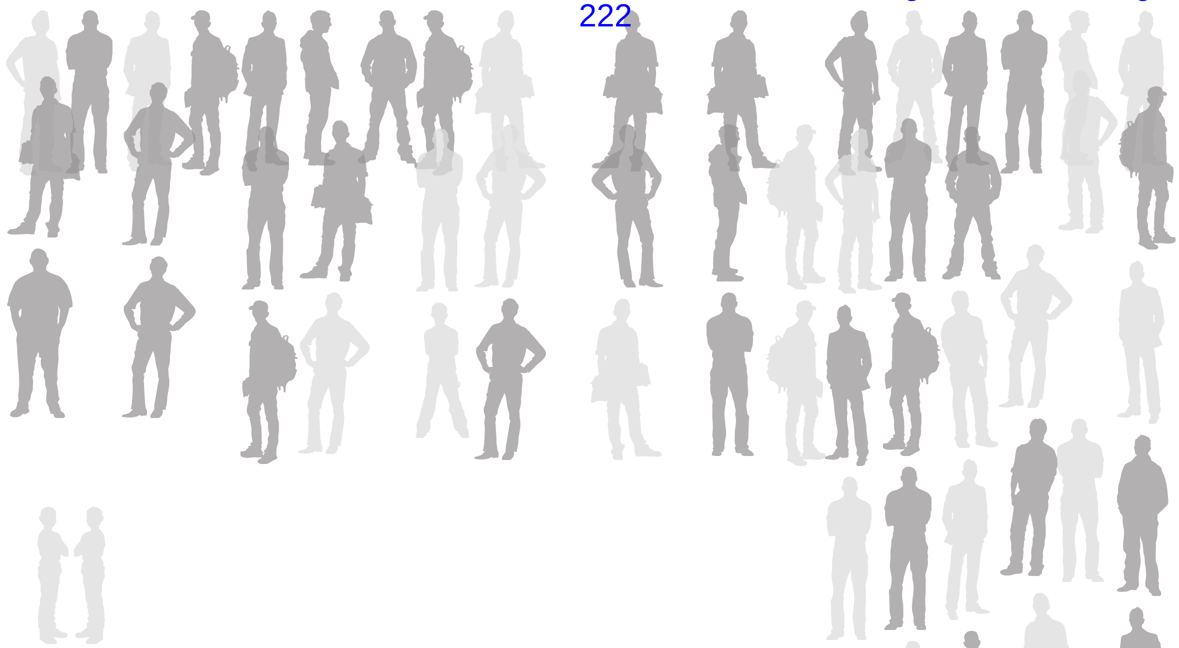
202 shooters

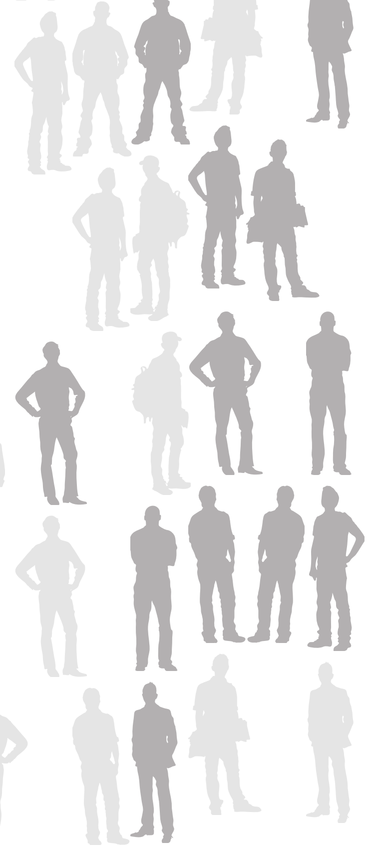
Some of these mass shooters were known to have violent tendencies or criminal pasts. Others seemed largely fine until they attacked. All but 5 were male. The vast majority were between the ages of 20 and 49. More than half — **108 of them** — died at or near the scene of the shooting, often by killing themselves.

CLICK ON AN ICON FOR DETAILS ABOUT EACH SHOOTER.

Women

Five women are on this list, and two of them partnered with men. Francine Graham and her boyfriend carried out the 2019 shooting in the Jersey City Kosher Supermarket. Pakistani mother Tashfeen Malik and her husband killed 14 partygoers at his workplace in San Bernardino, Calif., in 2015. Brittany Tucker killed a neighbor in a parking lot and then her four children in Monroe, La., in 2020. Cherie Lash Rhodes, a former tribal council chairwoman, killed her brother and three others at an eviction hearing in Alturas, Calif., in 2014, and ex-postal worker Jennifer San Marco killed seven in a Goleta, Calif., mail facility in 2006.





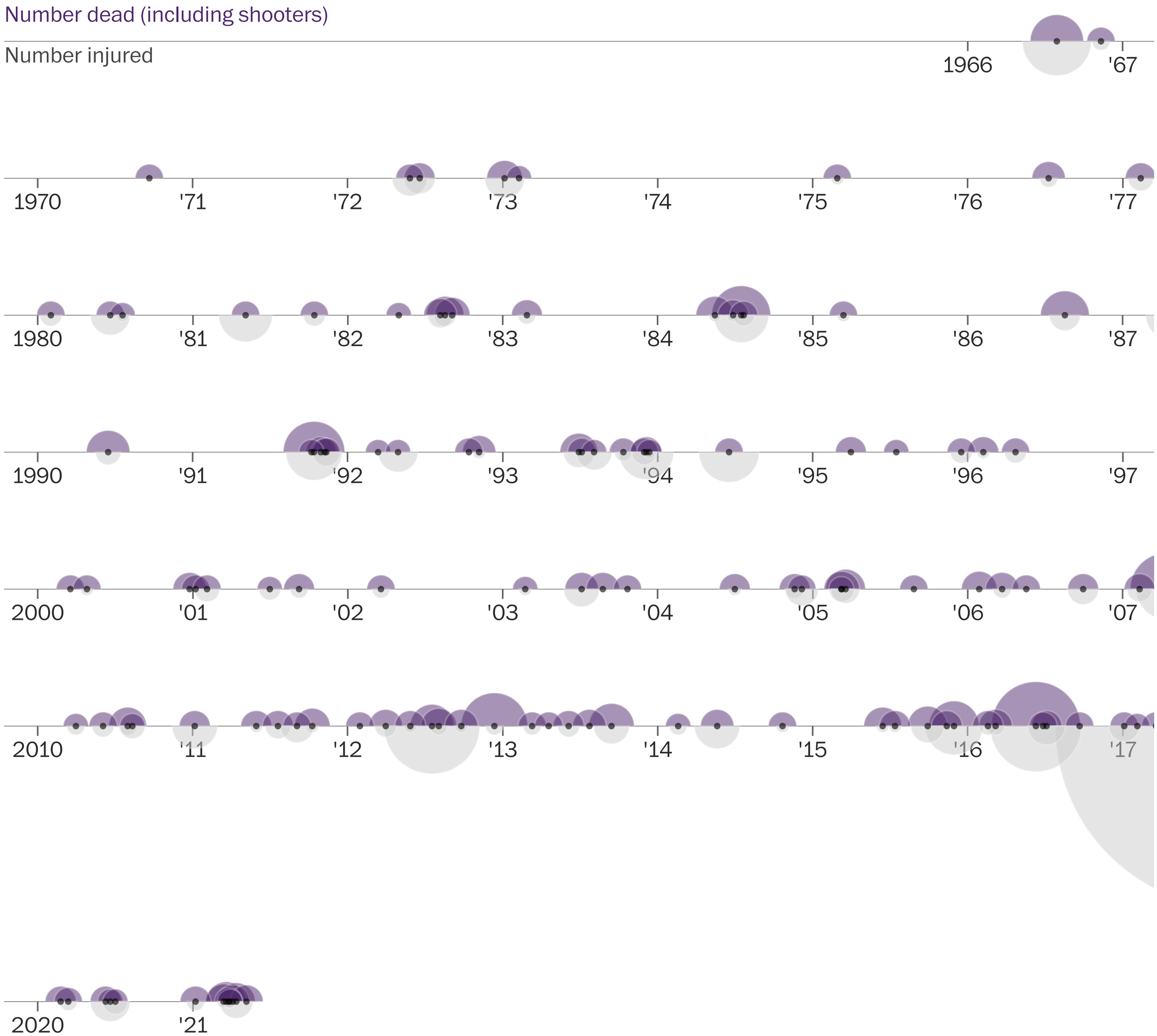
Middle-schoolers

Andrew Golden, 11, and Mitchell Johnson, 13, pulled a fire alarm to flush students and teachers out of their Jonesboro, Ark., middle school in 1998, and began shooting from a wooded perch nearby. They killed four girls and a teacher and wounded 10 others.

189 shootings

In the 50 years before the Texas tower shooting, there were just 25 public mass shootings in which four or more people were killed, according to author and criminologist Grant Duwe. Since then, the number has risen dramatically, and many of the deadliest shootings have occurred within the past few years.

HOVER FOR DETAILS ABOUT EACH SHOOTING.

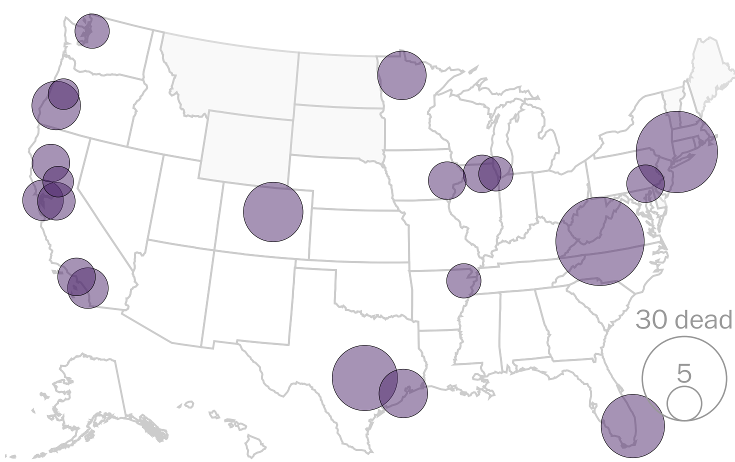


42 states and the District

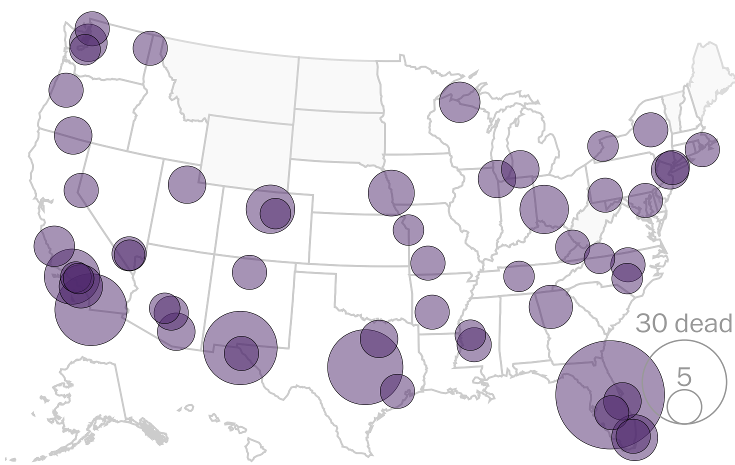
Shootings in schools and houses of worship tend to stand out in our minds, but they make up a relatively small portion of public mass shootings. More common are those in offices and retail establishments such as restaurants and stores. California has had more of these public mass shootings than any other state, with 32.

HOVER FOR INFO ABOUT EACH SHOOTING.

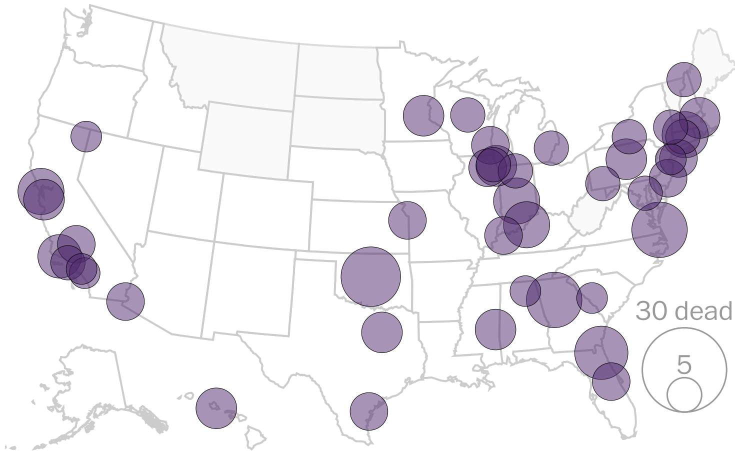
SCHOOLS



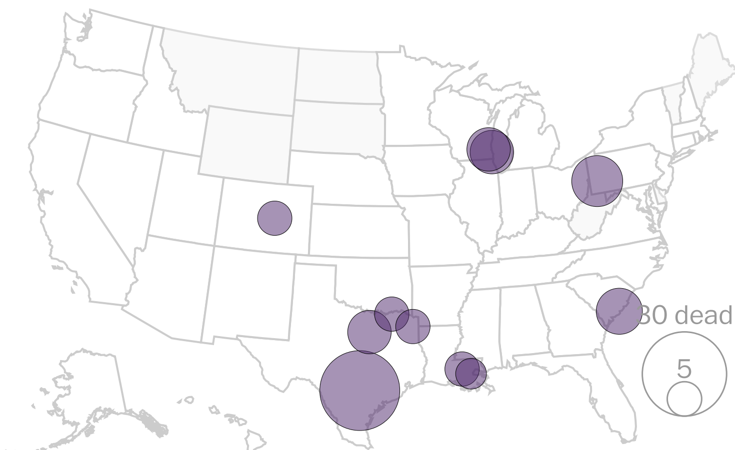
STORES, RESTAURANTS AND BARS



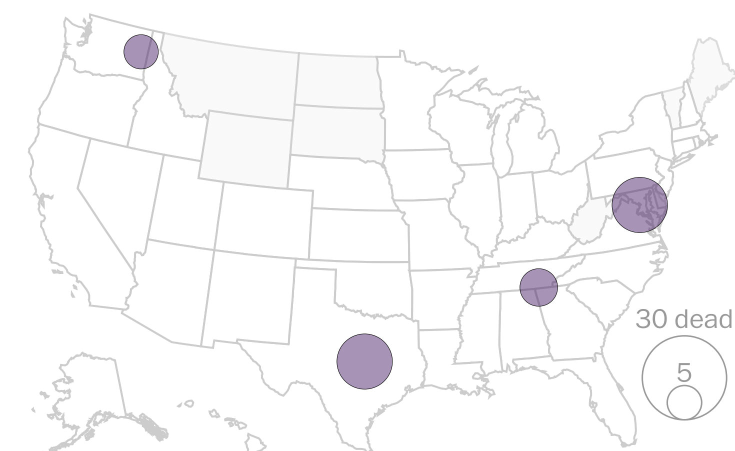
OFFICES



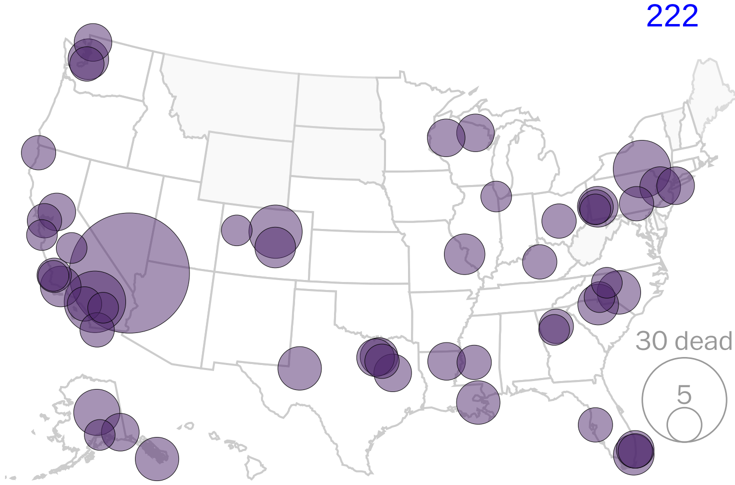
PLACE OF WORSHIP



MILITARY BASES



OTHER PLACES



Some locations have simply become shorthand for the horrors that occurred there — Columbine, Aurora, Sandy Hook. And some have added other tragic phrases to the national vocabulary.

“Going postal”

One of the most notorious workplace shootings was carried out by an ex-Marine in an Edmond, Okla., post office in 1986. He killed 14 and wounded six before killing himself. It was the deadliest in a string of rage-fueled killings by current and former postal employees that gave rise to the phrase “going postal.”

“Active shooter”

The 1999 siege by two seniors at Columbine High School in Colorado became a turning point after which school shootings could no longer be considered unthinkable aberrations. After a confused response that played out over several hours while a wounded teacher bled to death, U.S. law enforcement agencies overhauled procedures and officer training to create protocols for stopping an “active shooter.”

“Lockdown drill”

After Columbine, many schools created safety plans so that children and educators would know what to do during an attack. After Sandy Hook, “lockdown drills” became as common as fire drills. No children were killed at the Rancho Tehama Elementary School shooting in California in 2017, when fast-acting educators and students executed lockdown procedures that kept the gunman out of the school.

Lazaro Gamio, Alex Horton, Denise Lu, Richard Johnson, Ted Mellnik and Kevin Uhrmacher contributed to this report.

About this story

This data is compiled from Grant Duwe, author of “Mass Murder in the United States: A History,” Mother Jones and Washington Post research.

Death tolls include victims killed by shooters within a day of the main shooting, including any who were killed in another way. Totals also include people who later died from injuries received during the shootings. Injuries include everyone reportedly hurt in the event, not just gunshot injuries. A gun purchase that should have been rejected but was allowed because of a bureaucratic or reporting glitch is considered illegal. Reports disagree on some ages in this dataset.

Additional sources: Violence Policy Center, Gun Violence Archive; FBI 2014 Study of Active Shooter Incidents; published reports.

This is an updated version of a piece originally published in December 2015. Updates and corrections are made frequently as new information becomes available.

Originally published Feb. 14, 2018.

Comments

More stories

A gunman opens fire in your building. What do you do?

What would you do if someone walked into the building you are in right now and started shooting?

Mass shootings: How U.S. gun culture compares with the rest of the world

After mass shooting events, much debate centers around Americans’ relatively easy access to guns.

Most Read

Follow Post Graphics

Twitter	Facebook	Tumblr

Terms of use

Policies and Standards
Digital Products Terms of Sale
Print Products Terms of Sale
Terms of Service
Policies and Standards
Privacy Policy
Submissions and Discussion Policy
RSS Terms of Service
Ad Choices
Do Not Sell My Info

Mother Jones

CRIME AND JUSTICE UPDATED: OCT. 14, 2022

US Mass Shootings, 1982–2022: Data From Mother Jones' Investigation

The full data set from our in-depth investigation into mass shootings.

MARK FOLLMAN, GAVIN ARONSEN, and DEANNA PAN

This database originally covered cases from 1982 to 2012 and has since been updated and expanded numerous times. For analysis and context on this data—including how we built the database, and a change to the baseline for victim fatalities with cases dating from January 2013—see our **Guide to Mass Shootings in America,** which includes an interactive map documenting all of the cases.

[**Editor's note, 4/24/22:** Readers may wonder why this database does not include the New York City subway shooting on April 12, the school shooting in Washington, DC, on April 22, or other such attacks in which fewer than three victims died; for additional context on the challenges of defining and tracking mass shootings, and on our approach, see this piece and this piece.]

You can scan the underlying spreadsheet by clicking here and **download the database in its entirety (in CSV format) by clicking here.**

Mother Jones - ... Saved to Drive

File Edit View Insert Format Data Tools Extensions

100% View only Share

A1 case

	A	B	C	
1	case	location	date	summary
2	Raleigh spree shooting	Hedingham, North Carc	10/13/22	Austin Thompson
3	Greenwood Park Mall shooting	Greenwood, Indiana	7/17/22	Jonathan Sapirma
4	Highland Park July 4 parade shooting	Highland Park, Illinois	7/4/22	Suspected gunma
5	Church potluck dinner shooting	Birmingham, Alabama	6/16/22	Robert Findlay Sr
6	Concrete company shooting	Smithsburg, Maryland	6/9/22	The suspected 23
7	Tulsa medical center shooting	Tulsa, Oklahoma	6/1/22	Michael Louis, 45
8	Robb Elementary School massacre	Uvalde, Texas	5/24/22	Salvador Ramos,
9	Buffalo supermarket massacre	Buffalo, New York	5/14/22	Payton S. Gendrc
10	Sacramento County church shooting	Sacramento, California	2/28/22	"A man believed t
11	Oxford High School shooting	Oxford, Michigan	11/30/21	Ethan Crumbley,
12	San Jose VTA shooting	San Jose, California	5/26/21	Samuel Cassidy,
13	FedEx warehouse shooting	Indianapolis, Indiana	4/15/21	Brandon Scott Ho
14	Orange office complex shooting	Orange, California	3/31/21	Aminadab Gaxiol
15	Boulder supermarket shooting	Boulder, Colorado	3/22/21	Ahmad Al Aliwi Al
16	Atlanta massage parlor shootings	Atlanta, Georgia	3/16/21	Robert Aaron Lon
17	Springfield convenience store shooting	Springfield, Missouri	3/16/20	Joaquin S. Roma
18	Molson Coors shooting	Milwaukee, Wisconsin	2/26/20	Anthony Ferrill, 5
19	Jersey City kosher market shooting	Jersey City, New Jersey	12/10/19	David N. Anderso
20	Pensacola Naval base shooting	Pensacola, Florida	12/6/19	Ahmed Mohamm

21	Odessa-Midland shooting spree	Odessa, Texas	8/31/19	Seth A. Ator, 36, f
22	Dayton entertainment district shooting	Dayton, Ohio	8/4/19	Connor Betts, 24,
23	El Paso Walmart mass shooting	El Paso, Texas	8/3/19	Patrick Crusius, 2
24	Gilroy garlic festival shooting	Gilroy, California	7/28/19	Santino William L
25	Virginia Beach municipal building shoo	Virginia Beach, Virginia	5/31/19	DeWayne Craddo
26	Harry Pratt Co. warehouse shooting	Aurora, Illinois	2/15/19	Gary Martin, 45, v
27	Pennsylvania hotel bar shooting	State College, Pennsylv	1/24/19	Jordan Witmer, 21
28	SunTrust bank shooting	Sebring, Florida	1/23/19	Zephen A. Xaver,
29	Mercy Hospital shooting	Chicago, Illinois	11/19/18	Juan Lopez, 32, c
30	Thousand Oaks nightclub shooting	Thousand Oaks, Califor	11/7/18	Ian David Long, 2
31	Tree of Life synagogue shooting	Pittsburgh, Pennsylvani	10/27/18	Robert D. Bowers
32	Rite Aid warehouse shooting	Perryman, Maryland	9/20/18	Snochia Moseley,
33	T&T Trucking shooting	Bakersfield, California	9/12/18	Javier Casarez, 5
34	Fifth Third Center shooting	Cincinnati, Ohio	9/6/18	Omar Enrique Sa
35	Capital Gazette shooting	Annapolis, Maryland	6/28/18	Jarrod W. Ramos
36	Santa Fe High School shooting	Santa Fe, Texas	5/18/18	Dimitrios Pagourt
37	Waffle House shooting	Nashville, Tennessee	4/22/18	Travis Reinking, 2
38	Yountville veterans home shooting	Yountville, California	3/9/18	Army veteran Alb
39	Marjory Stoneman Douglas High Scho	Parkland, Florida	2/14/18	Nikolas J. Cruz, 1
40	Pennsylvania carwash shooting	Melcroft, Pennsylvania	1/28/18	Timothy O'Brien S
41	Rancho Tehama shooting spree	Rancho Tehama, Califo	11/14/17	Kevin Janson Nea
42	Texas First Baptist Church massacre	Sutherland Springs, Tex	11/5/17	Devin Patrick Kell
43	Walmart shooting in suburban Denver	Thornton, Colorado	11/1/17	Scott Allen Ostrer
44	Edgewood businees park shooting	Edgewood, Maryland	10/18/17	Radee Labeeb Pr
45	Las Vegas Strip massacre	Las Vegas, Nevada	10/1/17	Stephen Craig Pa
46	San Francisco UPS shooting	San Francisco, Californ	6/14/17	Jimmy Lam, 38, fa
47	Pennsylvania supermarket shooting	Tunkhannock, Pennsylv	6/7/17	Randy Stair, a 24-
48	Florida awning manufacturer shooting	Orlando, Florida	6/5/2017	John Robert Neur
49	Rural Ohio nursing home shooting	Kirkersville, Ohio	5/12/2017	Thomas Hartless,
50	Fresno downtown shooting	Fresno, California	4/18/2017	Kori Ali Muhamma
51	Fort Lauderdale airport shooting	Fort Lauderdale, Florida	1/6/2017	Esteban Santiago
52	Cascade Mall shooting	Burlington, Washington	9/23/2016	Arcan Cetin, 20, k
53	Baton Rouge police shooting	Baton Rouge, Lousiana	7/17/2016	Gavin Long, 29, a

Sheet1

case	location	date	summary	fatalities	injured	total_victim	location	age_of_	prior_signs_mental_h	mental_health_details	weapons_obtained_legally	where_obtaine	weapon_type	weapon_details	race	gender	sources	mental_health_sources	sources_additi	latitude	longitude	type	year
Raleigh spree shooting	Hedingham, North Carc	10/13/22	Austin Thompson, 15, went on a rampage in the H	5	2	7	Other	15	-	-	-	-	semiautomatic rifle, semiautomati		White	M	https://www.new	-	-	-	-	Spree	2022
Greenwood Park Mall shooting	Greenwood, Indiana	7/17/22	Jonathan Sapirman, 20, opened fire in a mall fooc	3	2	5	workplace	20	-	-	yes	in Greenwood	semiautomatic ri	Sig Sauer M400	White	M	https://apnews.c	-	-	-	-	Mass	2022
Highland Park July 4 parade shooting	Highland Park, Illinois	7/4/22	Suspected gunman Robert "Bobby" Crimo, 21, all	7	46	53	Other	21	-	-	yes	legally purchase	semiautomatic ri	AR-15 style rifle,	White	M	https://www.cbs	-	-	-	-	Mass	2022
Church potluck dinner shooting	Birmingham, Alabama	6/16/22	Robert Findlay Smith, 70, opened fire with a hand	3	0	3	religious	70	-	-	yes	-	semiautomatic h	-	White	M	https://apnews.c	-	-	-	-	Mass	2022
Concrete company shooting	Smithsburg, Maryland	6/9/22	The suspected 23-year-old gunman shot four cow	3	1	4	workplace	23	-	-	-	-	-	-	-	M	https://www.was	-	-	-	-	Mass	2022
Tulsa medical center shooting	Tulsa, Oklahoma	6/1/22	Michael Louis, 45, killed four, including two doctor	4	r than 10"	TK	workplace	-	-	-	yes	-	semiautomatic ri	AR-15 style rifle	Black	M	https://www.nyti	-	-	-	-	Mass	2022
Robb Elementary School massacre	Uvalde, Texas	5/24/22	Salvador Ramos, 18, was identified by authorities	21	17	38	School	18	yes	-	yes	-	semiautomatic ri	-	Latino	M	https://abcnews	-	-	-	-	Mass	2022
Buffalo supermarket massacre	Buffalo, New York	5/14/22	Payton S. Gendron, 18, committed a racially moti	10	3	13	workplace	18	yes	previous threats and a me	yes	-	semiautomatic ri	Bushmaster XM-	White	M	https://www.nyti	https://www.nytimes.com/2022/05/15/nyreg	-	-	-	Mass	2022
Sacramento County church shooting	Sacramento, California	2/28/22	"A man believed to be meeting his three children	4	0	4	Religious	-	-	-	-	-	-	-	-	M	https://www.nyti	-	-	38.60111019	-121.4189659	Mass	2022
Oxford High School shooting	Oxford, Michigan	11/30/21	Ethan Crumbley, a 15-year-old student at Oxford	4	7	11	School	15	-	-	-	-	semiautomatic h	Sig Sauer 9mm	-	M	https://www.was	-	-	42.84410784	-83.25992831	Mass	2021
San Jose VTA shooting	San Jose, California	5/26/21	Samuel Cassidy, 57, a Valley Transportation Auth	9	0	9	Workplace	57	yes	Perpetrator had a history	-	-	semiautomatic h	-	-	M	https://www.sfch	https://www.nytimes.com/2021/05/26/us/wh	-	37.316097	-121.888533	Mass	2021
FedEx warehouse shooting	Indianapolis, Indiana	4/15/21	Brandon Scott Hole, 19, opened fire around 11 p.	8	7	15	Workplace	19	yes	Perpetrator had been rep	yes	-	semiautomatic ri	-	White	M	https://www.indy	https://www.nytimes.com/2021/04/19/us/inc	-	39.68663	-86.32313	Mass	2021
Orange office complex shooting	Orange, California	3/31/21	Aminadab Gaxiola Gonzalez, 44, allegedly opene	4	1	5	Workplace	-	-	-	-	-	semiautomatic h	-	-	M	https://www.latir	-	-	33.83542	-117.85379	Mass	2021
Boulder supermarket shooting	Boulder, Colorado	3/22/21	Ahmad Al Aliwi Alissa, 21, carried out a mass sho	10	0	10	Workplace	21	yes	Brother described him as	Yes	-	semiautomatic ri	Ruger AR-556; v	-	M	https://www.den	https://www.thedailybeast.com/boulder-colc	-	39.986961	-105.251168	Mass	2021
Atlanta massage parlor shootings	Atlanta, Georgia	3/16/21	Robert Aaron Long, 21, killed eight people at thre	8	1	9	Workplace	21	-	-	-	-	semiautomatic h	-	White	M	https://www.ajc.	-	-	34.111653	-84.580376	Spree	2021
Springfield convenience store shooting	Springfield, Missouri	3/16/20	Joaquin S. Roman, 31, went on a rampage culmi	4	0	4	Workplace	31	-	-	-	-	-	-	-	M	https://www.kiro	-	-	37.210432	-93.23686	Mass	2020
Molson Coors shooting	Milwaukee, Wisconsin	2/26/20	Anthony Ferrill, 51, an employee armed with two	5	0	5	Workplace	51	-	-	-	-	semiautomatic h	-	Black	M	https://www.jsor	-	-	43.044511	-87.962537	Mass	2020
Jersey City kosher market shooting	Jersey City, New Jersey	12/10/19	David N. Anderson, 47, and Francine Graham, 50	4	3	7	Other	-	-	-	-	-	-	-	Black	Male & Female	https://www.nyti	-	-	40.707363	-74.083609	Spree	2019
Pensacola Naval base shooting	Pensacola, Florida	12/6/19	Ahmed Mohammed al-Shamrani, A Saudi Arabia	3	8	11	Military	-	-	-	-	-	semiautomatic h	-	-	M	https://www.was	-	-	30.364707	-87.288567	Mass	2019
Odessa-Midland shooting spree	Odessa, Texas	8/31/19	Seth A. Ator, 36, fired at police officers who stopp	7	25	32	Other	36	yes	"One friend of the family s	-	-	semiautomatic ri	-	White	M	https://www.was	https://www.nytimes.com/2019/09/02/us/te	-	31.925974	-102.2796	Spree	2019
Dayton entertainment district shooting	Dayton, Ohio	8/4/19	Connor Betts, 24, died during the attack, following	9	27	36	Other	24	-	-	Yes	-	semiautomatic ri	AR-15-style rifle,	White	M	https://www.nyti	-	-	39.757312	-84.184947	Mass	2019
El Paso Walmart mass shooting	El Paso, Texas	8/3/19	Patrick Crusius, 21, who was apprehended by pol	22	26	48	Workplace	21	-	-	Yes	-	semiautomatic ri	AK-47-style rifle,	White	M	https://www.was	-	-	31.771068	-106.375655	Mass	2019
Gilroy garlic festival shooting	Gilroy, California	7/28/19	Santino William LeGan, 19, fired indiscriminately i	3	12	15	Other	19	TBD	-	Yes	Nevada, on July	semiautomatic ri	AK-47-style rifle,	-	M	https://www.cbs	-	-	36.997191	-121.584819	Mass	2019
Virginia Beach municipal building shoo	Virginia Beach, Virginia	5/31/19	DeWayne Craddock, 40, a municipal city worker v	12	4	16	Workplace	40	TBD	-	Yes	-	2 handguns	.45-caliber hand	Black	M	https://www.was	-	-	36.75442	-76.060378	Mass	2019
Harry Pratt Co. warehouse shooting	Aurora, Illinois	2/15/19	Gary Martin, 45, went on a rampage inside the wa	5	6	11	Workplace	45	Yes	(pending)	No	-	handgun	Smith & Wesson	Black	M	https://www.was	-	-	41.753725	-88.331057	Mass	2019
Pennsylvania hotel bar shooting	State College, Pennsylv	1/24/19	Jordan Witmer, 21, shot three people at a Ramad	3	1	4	Other	21	-	-	Yes	-	handgun	-	White	M	http://www.wistv	-	-	40.785142	-77.839411	Spree	2019
SunTrust bank shooting	Sebring, Florida	1/23/19	Zephen A. Xaver, 21, fatally shot five women insic	5	0	5	Workplace	21	Yes	Xaver was reported to ha	Yes	-	handgun	9 mm handgun	White	M	https://wsbt.com	https://wsbt.com/news/local/classmate-says	-	27.471043	-81.45847	Mass	2019
Mercy Hospital shooting	Chicago, Illinois	11/19/18	Juan Lopez, 32, confronted his former fiancé, ER	3	0	3	Workplace	32	-	-	-	-	semiautomatic h	9mm handgun	Latino	M	https://www.cnn	-	-	41.847667	-87.622009	Mass	2018
Thousand Oaks nightclub shooting	Thousand Oaks, Califor	11/7/18	Ian David Long, 28, dressed in black and armed v	12	22	34	Other	28	Yes	Per the LATimes: "Sheriff	Yes	A gun store in Si	semiautomatic h	Glock 21, .45 ca	-	M	http://www.latim	http://www.latimes.com/local/lanow/la-me-l	-	34.176946	-118.874793	Mass	2018
Tree of Life synagogue shooting	Pittsburgh, Pennsylvani	10/27/18	Robert D. Bowers, 46, shouted anti-Semitic slurs	11	6	17	Religious	46	-	-	Yes	(Unclear; investi	semiautomatic ri	AR-15; Glock .3!	White	M	https://www.nyti	-	-	40.443898	-79.921398	Mass	2018
Rite Aid warehouse shooting	Perryman, Maryland	9/20/18	Snochia Moseley, 26, reportedly a disgruntled em	3	3	6	Workplace	26	-	-	Yes	-	semiautomatic h	Glock 9 mm	Black	F	http://www.baltir	-	-	39.455658	-76.208485	Mass	2018
T&T Trucking shooting	Bakersfield, California	9/12/18	Javier Casarez, 54, who was going through a bitte	5	0	5	Workplace	54	-	-	Yes	-	handgun	Smith & Wesson	Latino	M	http://www.latim	-	-	35.349388	-118.916335	Spree	2018
Fifth Third Center shooting	Cincinnati, Ohio	9/6/18	Omar Enrique Santa Perez, 29, walked into the gr	3	2	5	Workplace	29	Yes	Perez filed a "delusional"	Yes	A local gun store	semiautomatic h	9mm handgun	Latino	M	https://www.cnn	-	-	39.101981	-84.511782	Mass	2018
Capital Gazette shooting	Annapolis, Maryland	6/28/18	Jarrod W. Ramos, 38, shot through the glass door	5	2	7	Workplace	38	-	-	Yes	-	shotgun	12-gauge pump-	White	M	http://www.capit	-	-	38.994548	-76.543657	Mass	2018
Santa Fe High School shooting	Santa Fe, Texas	5/18/18	Dimitrios Pagourtzis, a 17-year-old student, opene	10	13	23	School	17	-	-	TBD	Father's weapon	shotgun; .38 rev	-	White	M	https://www.chro	-	https://www.chro	29.392825	-95.141972	Mass	2018
Waffle House shooting	Nashville, Tennessee	4/22/18	Travis Reinking, 29, opened fire around 3:30 a.m.	4	4	8	Other	29	Yes	Reinking had a history of	Yes	Family member	semiautomatic ri	AR-15	White	M	https://www.was	https://www.buzzfeed.com/briannasacks/th	https://www.wasl	36.052521	-86.616944	Mass	2018
Yountville veterans home shooting	Yountville, California	3/9/18	Army veteran Albert Cheung Wong, 36, stormed a	3	0	3	Workplace	36	Yes	Wong had served in Afgh	Yes	TBD	semiautomatic ri	-	Asian	M	https://www.cnn	https://www.cnn.com/2018/03/10/us/califorr	https://www.cnn.	38.392496	-122.366528	Mass	2018
Marjory Stoneman Douglas High Scho	Parkland, Florida	2/14/18	Nikolas J. Cruz, 19, heavily armed with an AR-15,	17	17	34	School	19	Yes	Cruz had a long history of	Yes	A Florida pawn s	semiautomatic ri	AR-15	White	M	https://www.nyti	https://www.nytimes.com/2018/02/15/us/nik	https://www.nytir	26.30483	-80.269511	Mass	2018
Pennsylvania carwash shooting	Melcroft, Pennsylvania	1/28/18	Timothy O'Brien Smith, 28, wearing body armor a	4	1	5	Other	28	TBD	-	TBD	TBD	semiautomatic ri	-	White	M	http://www.wpxi	-	http://www.wpxi.	40.052151	-79.389166	Mass	2018
Rancho Tehama shooting spree	Rancho Tehama, Califo	11/14/17	Kevin Janson Neal, 44, went on an approximately	5	10	15	Other	44	TBD	-	No	TBD	semiautomatic ri	Two illegally mo	White	M	https://www.nbc	-	https://www.nbcr	40.018759	-122.393089	Spree	2017
Texas First Baptist Church massacre	Sutherland Springs, Tex	11/5/17	Devin Patrick Kelley, a 26-year-old ex-US Air Forc	26	20	46	Religious	26	Yes	Kelley had a history of do	Kelley passed federal crimin	Purchased in Ap	semiautomatic ri	Ruger AR-556; I	White	M	https://www.was	http://www.expressnews.com/news/local/ar	https://www.wasl	29.273282	-98.056488	Mass	2017
Walmart shooting in suburban Denver	Thornton, Colorado	11/1/17	Scott Allen Ostrem, 47, walked into a Walmart in a	3	0	3	Other	47	Unclear	-	TBD	-	semiautomatic h	-	White	M	https://www.nyti	-	https://www.nytir	39.876374	-104.986132	Mass	2017
Edgewood businees park shooting	Edgewood, Maryland	10/18/17	Radee Labeeb Prince, 37, fatally shot three peopl	3	3	6	Workplace	37	Unclear	-	No	Unclear	handgun	.380-caliber; mal	Black	M	http://www.baltir	-	http://www.baltim	39.452189	-76.309988	Mass	2017
Las Vegas Strip massacre	Las Vegas, Nevada	10/1/17	Stephen Craig Paddock, 64, fired a barrage of rap	58	546	604	Other	64	TBD	Perpetrator's history uncle	Yes	Two gun shops i	23 firearms, mos	AR-15-style and	White	M	https://www.lvm 700"	https://www.nytimes.com/2017/10/13/us/ste	https://www.lvmp 700"	36.095739	-115.171544	Mass	2017
San Francisco UPS shooting	San Francisco, Californ	6/14/17	Jimmy Lam, 38, fatally shot three coworkers and	3	2	5	Workplace	38	Yes	Lam had a history of dom	No	Unclear; the firea	two handguns	MAC-10-style "a	Asian	M	http://www.nbcb	http://www.ktvu.com/news/ktvu-local-news/	http://www.nbcba	37.765947	-122.406087	Mass	2017
Pennsylvania supermarket shooting	Tunkhannock, Pennsylv	6/7/17	Randy Stair, a 24-year-old worker at Weis grocery	3	0	3	Workplace	24	Unclear	-	TBD	-	shotguns	-	White	M	http://www.press	-	http://www.press	41.529546	-75.94722	Mass	2017
Florida awning manufacturer shooting	Orlando, Florida	6/5/2017	John Robert Neumann, Jr., 45, a former employee	5	0	5	Workplace	45	Unclear	-	TBD	-	semiautomatic h	-	-	M	http://www.posta	-	http://www.posta	28.580295	-81.294086	Mass	2017
Rural Ohio nursing home shooting	Kirkersville, Ohio	5/12/2017	Thomas Hartless, 43, shot and killed a former girl	3	0	3	Workplace	43	Yes	Hartless had a violent crir	TBD	-	handgun, shotgu	-	White	M	http://abc6onyou	http://www.cbsnews.com/news/64-guns-se	http://abc6onyou	39.959034	-82.596508	Mass	2017
Fresno downtown shooting	Fresno, California	4/18/2017	Kori Ali Muhammad, 39, opened fire along a stree	3	0	3	Other	39	Unclear	-	Unknown	-	handgun	.357 revolver	Black	M	http://www.daily	-	http://www.dailyr	36.746378	-119.800319	Mass	2017
Fort Lauderdale airport shooting	Fort Lauderdale, Florida	1/6/2017	Esteban Santiago, 26, flew from Alaska to Fort La	5	6	11	Airport	26	Yes	Among other signs, Santi	Yes	-	semiautomatic h	Walther 9mm se	Latino	M	http://www.nytim	http://www.nytimes.com/2017/01/07/us/este	http://www.nytim	26.072751	-80.143382	Mass	2017
Cascade Mall shooting	Burlington, Washington	9/23/2016	Arcan Cetin, 20, killed a teen girl and three wome	5	0	5	Other	20	Yes	According to the Cetin's s	TBD	-	Rifle	Ruger .22-calibe	-	M	http://www.nytim	http://www.nbcnews.com/news/us-news/ar	http://www.nytim	48.461367	-122.337918	Mass	2016
Baton Rouge police shooting	Baton Rouge, Lousiana	7/17/2016	Gavin Long, 29, a former Marine who served in Ira	3	3	6	Other	29	Yes	Unclear	Unknown	-	Two semiautoma	IWI Tavor SAR 5	Black	M	http://time.com/4	-	http://time.com/4	30.433601	-91.081403	Spree	2016
Dallas police shooting	Dallas, Texas	7/7/2016	Micah Xavier Johnson, a 25-year-old Army vetera	5	11	16	Other	25	Unclear	Unclear	Yes	online and or gu	Semiautomatic r	Izhmash-Saiga 5	Black	M	http://www.nytim	-	http://www.nytim	32.7801052	-96.8000082	Mass	2016
Orlando nightclub massacre	Orlando, Florida	6/12/2016	Omar Mateen, 29, attacked the Pulse nighclub in	49	53	102	Other	29	Unclear	Unclear	Yes	Shooting center	Semiautomatic r	Sig Sauer MCX (	Other	M	http://www.moth	-	http://www.mothe	28.519718	-81.376777	Mass	2016
Excel Industries mass shooting	Hesston, Kansas	2/25/2016	Cedric L. Ford, who worked as a painter at a man	3	14	17	Workplace	38	Unclear	Unclear	Yes	-	Semiautomatic r	Zastava Serbia /	Black	M	http://www.nytim	-	http://www.nytim	38.135992	-97.425145	Mass	2016
Kalamazoo shooting spree	Kalamazoo County, Mic	2/20/2016	Jason B. Dalton, a driver for Uber, apparently sele	6	2	8	Other	45	Unclear	Unclear	Yes	-	Semiautomatic h	9 mm handgun (	White	M	http://www.nytim	-	http://www.nytim	42.236689	-85.674795	Mass	2016
San Bernardino mass shooting	San Bernardino, Califor	12/2/2015	Syed Rizwan Farook left a Christmas party held a	14	21	35	Workplace	28	Unclear	Unclear	Yes	The suspects pu	Two assault rifles	Two semiautoma	Other	Male & Female	http://www.moth	-	http://www.mothe	34.075961	-117.27789	Mass	2015
Planned Parenthood clinic	Colorado Springs, Colo	11/27/2015	Robert Lewis Dear, 57, shot and killed a police off	3	9	12	Workplace	57	Unclear	The judge in the case has	Unknown	Unclear	Long gun	Reportedly an Al	White	M	http://www.nytim	-	http://www.nytim	38.881031	-104.849057	Mass	2015
Colorado Springs shooting rampage	Colorado Springs, Colo	10/31/2015	Noah Harpham, 33, shot three people before dea	3	0	3	Other	33	Unclear	Prior to the rampage, Har	Yes	Unclear	Two handguns a	AR-15 rifle, a 9 r	White	M	http://www.moth	-	http://www.mothe	38.83755	-104.814251	Mass	2015
Umpqua Community College shooting	Roseburg, Oregon	10/1/2015	26-year-old Chris Harper Mercer opened fire at U	9	9	18	School	26	Unclear	Harper-Mercer's mother s	Yes	From the home I	Five pistols, one	9 mm Glock pist	Other	Male	http://www.nytim	http://www.latimes.com/nation/nationnow/la	http://www.nytim	43.289538	-123.333193	Mass	2015
Chattanooga military recruitment cente	Chattanooga, Tennesse	7/16/2015	Kuwaiti-born Mohammod Youssuf Abdulazeez, 24	5	2	7	Military	24	Unclear	Abdulazeez "had sufferec	Yes ("some of the weapons	On the internet,	2 assault rifles; s	AK-47, AR-15, a	Other	Male	http://www.reute	-	http://www.reuter	35.047157	-85.311819	Mass	2015
Charleston Church Shooting	Charleston, South Caro	6/17/2015	Dylann Storm Roof, 21, shot and killed 9 people a	9	1	10	Religious	21	Unclear	-	Yes	Shooter's Choice	Handgun	.45-caliber Glock	White	Male	http://www.moth	-	http://www.mothe	32.788387	-79.933143	Mass	2015
Trestle Trail bridge shooting	Menasha, Wisconsin	6/11/2015	Sergio Valencia del Toro, 27, in what officials say	3	1	4	Other	27	Yes	While serving in the Air F	Yes	Unclear	Two handguns	Details unclear, I	Latino	M	http://fox6now.c	-	http://fox6now.cc	44.204124	-88.467541	Mass	2015
Marysville-Pilchuck High School shooti	Marysville, Washington	10/24/2014	Jaylen Fryberg, 15, using a .40-caliber Berretta, s students at Marysville High School, including two	5	1	6	School	15	Unclear	Fryberg was well-liked an	No	Gun was his fath	Handgun	Beretta .40-calib	Native American	Male	http://www.seatt	http://www.newyorker.com/science/maria-k	http://www.seattl	48.050824	-122.176918	Mass	2014
Isla Vista mass murder	Santa Barbara, Californ	5/23/2014	Elliot Rodger, 22, shot three people to death in the	6	13	19	School	22	Yes	Rodger was never hospit	Yes	-	Three semi-auto	Two Sig Sauer P	White	M	http://www.sbsh	-	http://www.sbshe	34.436283	-119.8714406	Mass	2014
Fort Hood shooting 2	Fort Hood, Texas	4/3/2014	Army Specialist Ivan Lopez, 34, opened fire at the	3	12	15	Military	34	Unclear	Lopez "had a long history	Yes	Local gun store i	handgun	.45-caliber Smith	Latino	M	http://www.cnn.c	-	http://www.cnn.c	31.141716	-97.777559	Mass	2014
Alturas tribal shooting	Alturas, California	2/20/2014	Cherie Lash Rhoades, 44, opened fire at the Ced	4	2	6	Other	44	Unknown	-	Unknown	-	Two handguns a	9mm semi-autom	Native American	Female	http://www.sacb	-	http://www.sacbe	41.487104	-120.542237	Spree	2014
Washington Navy Yard shooting	Washington, D.C.	9/16/2013	Aaron Alexis, 34, a military veteran and contracto	12	8	20	Military	34	Yes	Had told Rhode Island po	Yes	Sharpshooters S	Sawed-off shotg	Remington 870 I	Black	Male	http://www.nytim	http://bigstory.ap.org/article/13-killed-washi	http://www.nytim	38.874981	-76.99453	Mass	2013
Hialeah apartment shooting	Hialeah, Florida	7/26/2013	Pedro Vargas, 42, set fire to his apartment, killed	7	0	7	Other	42	Unclear	His mother told authoritie	Yes	Florida Gun Cen	9mm semi-auton	Glock 17	Latino	Male	http://www.mian	http://www.miamiherald.com/2013/08/03/v-	http://www.miam	25.864338	-80.311775	Mass	2013
Santa Monica rampage	Santa Monica, Californi	6/7/2013	John Zawahri, 23, armed with a homemade assau	6	3	9	Other	23	Yes	He was known as a solita	Yes	Assembled a rifl	Assault rifle, high	.223-caliber sem	White	Male	http://www.cbsn	http://www.cbsnews.com/8301-504083_162	http://www.cbsne	34.008617	-118.494754	Mass	2013
Pinewood Village Apartment shooting	Federal Way, Washingt	4/21/2013	Dennis Clark III, 27, shot and killed his girlfriend ir	5	0	5	Other	27	No	-	Yes	Unknown	Semiautomatic h	.40 caliber semi-	Black	Male	http://seattletime	-	http://seattletime	47.3129607	-122.3393665	Mass	2013
Mohawk Valley shootings	Herkimer County, New	3/13/2013	Kurt Myers, 64, shot six people in neighboring tow	5	2	7	Other	64	No	-	Yes	Frank's Guns in	Shotgun	Unknown	White	Male	https://www.nys	http://poststar.com/news/local/state-and-re	https://www.nysp	43.045601	-74.984891	Mass	2013
Sandy Hook Elementary massacre	Newtown, Connecticut	12/14/2012	Adam Lanza, 20, shot his mother dead at their ho	27	2	29	School	20	Yes	Lanza had a history of se	No	Stolen from moth	Two semiautoma	10mm Glock, 9m	white	Male	http://usnews.nt	http://usnews.nbcnews.com/_news/2012/12	http://usnews.nb	41.4123225	-73.31142358	Mass	2012
Accent Signage Systems shooting	Minneapolis, Minnesota	9/27/2012	Andrew Engeldinger, 36, upon learning he was be	7	1	8	Workplace	36	Yes	His family worried about h	Yes	Unknown	One semiautoma	9mm Glock sem	white	Male	http://www.startr	http://www.startribune.com/local/17177446	http://www.startri	44.977425	-93.310408	Mass	2012
Sikh temple shooting	Oak Creek, Wisconsin	8/5/2012	U.S. Army veteran Wade Michael Page, 40, open	7	3	10	Religious	40	Yes	His Army friends once bro	Yes	Unknown	One semiautoma	9mm Springfield	white	Male	http://www.jsonl	http://www.jsonline.com/news/milwaukee/fr	http://www.jsonli	42.8858503	-87.8631362	Mass	2012
Aurora theater shooting	Aurora, Colorado	7/20/2012	James Holmes, 24, opened fire in a movie theater	12	70	82	Other	24	Yes	He saw at least three mer	Yes	Gander Mountain	Two semiautoma	Two .40-caliber (	white	Male	http://www.cbsn	http://www.cbsnews.com/8301-201_162-57	http://www.cbsne	39.706038	-104.820594	Spree	2012
Seattle cafe shooting	Seattle, Washington	5/20/2012	Ian Stawicki, 40, gunned down four patrons at a c	6	1	7	Other	40	Yes	His family said he was me	Yes	Bull's Eye Shoot	Two semiautoma	Two .45-caliber s	white	Male	http://usnews.nt	http://usnews.nbcnews.com/_news/2012/0	http://usnews.nb	47.6038321	-122.3300624	Mass	2012
Oikos University killings	Oakland, California	4/2/2012	One L. Goh, 43, a former student, opened fire in a	7	3	10	School	43	Yes	A former instructor at Oik	Yes	Bullseye in Cast	One semiautoma	.45-caliber semi	Asian	Male	http://blog.sfgat	http://berkeley.patch.com/articles/one-l-goh	http://blog.sfgate	37.8043808	-122.2708166	Mass	2012
Su Jung Health Sauna shooting	Norcross, Georgia	2/21/2012	Jeong Soo Paek, 59, returned to a Korean spa frc	5	0	5	Other	59	Yes	His sister worried about h	Yes	Unknown	One semiautoma	.45-caliber semi	Asian	Male	http://www.gwin	http://www.gwinnettdailypost.com/news/201	http://www.gwinr	33.9412127	-84.2135309	Mass	2012
Seal Beach shooting	Seal Beach, California	10/12/2011	Scott Evans Dekraai, 42, opened fire inside a hair	8	1	9	Other	42	Yes	He suffered from bipolar	Yes	Unknown	Two semiautoma	.45-caliber Heck	white	Male	http://laist.com/2	http://laist.com/2011/10/13/seal_beach_sh	http://laist.com/2	33.741176	-118.1046356	Mass	2011
IHOP shooting	Carson City, Nevada	9/6/2011	Eduardo Sencion, 32, opened fire at an Internatio	5	7	12	Other	32	Yes	He was diagnosed with pa	Yes	Purchased from	Two rifles (both a	AK-47 Norinco A	Latino	Male	http://www.huffir	http://www.huffingtonpost.com/2011/11/03/	http://www.huffin	39.1637984	-119.7674034	Mass	2011
Tucson shooting	Tucson, Arizona	1/8/2011	Jared Loughner, 22, opened fire outside a Safewa	6	13	19	Other	22	Yes	His symptoms pointed to	Yes	Sportsmen's Wa	One semiautoma	9mm Glock 19 s	white	Male	http://www.time.	http://www.time.com/time/magazine/article/	http://www.time.c	32.335941	-110.975132	Mass	2011
Hartford Beer Distributor shooting	Manchester, Connecticu	8/3/2010	Omar S. Thornton, 34, shot up his Hartford Beer D	9	2	11	Workplace	34	No	He apparently was driven	Yes	Gun dealer in Ea	Two semiautoma	Two 9mm Ruger	black	Male	http://www.cbsn	http://www.cbsnews.com/8301-504083_162	http://www.cbsne	41.798764	-72.570068	Mass	2010
Coffee shop police killings	Parkland, Washington	11/29/2009	Maurice Clemmons, 37, a felon who was out on b	4	1	5	Other	37	Yes	He had a history of erratic	No	Stolen from an ir	One semiautoma	9mm Glock 17 s	black	Male	http://seattletime	http://seattletimes.com/html/localnews/201	http://seattletime	47.15277	-122.467308	Mass	2009
Fort Hood massacre	Fort Hood, Texas	11/5/2009	Army psychiatrist Nidal Malik Hasan, 39, opened	13	31	44	Military	39	Unclear	Medical officials at Walter	Yes	Guns Galore in F	One semiautoma	FN Five-seven s	Other	Male	http://seattletime	http://www.npr.org/templates/story/story.ph	http://seattletime	31.135557	-97.783664	Mass	2009
Binghamton shootings	Binghamton, New York	4/3/2009	Jiverly Wong, 41, opened fire at an American Civi	14	4	18	Other	41	Yes	He apparently harbored g	Yes	Gander Mountain	Two semiautoma	9mm Beretta, .45	Asian	Male	http://www.nytim	http://www.nytimes.com/2009/04/12/nyregi	http://www.nytim	42.099802	-75.917723	Mass	2009
Carthage nursing home shooting	Carthage, North Carolin	3/29/2009	Robert Stewart, 45, opened fire at a nursing home	8	3	11	Other	45	Yes	His estranged wife told he	Yes	Local sporting go	One revolver, on	Winchester 1300	white	Male	http://www.wral.	http://www.wral.com/news/local/story/9845	http://www.wral.c	35.333434	-79.414592	Mass	2009
Atlantis Plastics shooting	Henderson, Kentucky	6/25/2008	Disgruntled employee Wesley Neal Higdon, 25, sl	6	1	7	Workplace	25	No	He called his girlfriend tw	Yes	Unknown	One semiautoma	.45-caliber Hi-Po	white	Male	http://www.foxne	http://www.foxnews.com/story/0,2933,3712	http://www.foxne	37.76721	-87.5573742	Spree	2008
Northern Illinois University shooting	DeKalb, Illinois	2/14/2008	Steven Kazmierczak, 27, opened fire in a lecture	5	21	26	School	27	Yes	He had a long history of m	Yes	Online and gun s	Three semiautor	9mm Glock 19, I	white	Male	http://www.huffir	http://www.cbsnews.com/stories/2008/02/1	http://www.huffin	41.9294736	-88.7503647	Mass	2008
Kirkwood City Council shooting	Kirkwood, Missouri	2/7/2008	Charles "Cookie" Lee Thornton, 52, went on a ran	6	2	8	Other	52	No	He was known for histrion	No	Taken in burglary	One semiautoma	.40-caliber Smith	black	Male	http://www.stltoc	http://www.stltoday.com/news/local/crime-a	http://www.stltod	38.580093	-90.40691	Mass	2008
Westroads Mall shooting	Omaha, Nebraska	12/5/2007	Robert A. Hawkins, 19, opened fire inside Westro	9	4	13	Other	19	Yes	He had been treated in th	No	Stolen from gran	One rifle (assaul	WASR-10 Centu	white	Male	http://www.guar	http://www.guardian.co.uk/world/2007/dec/	http://www.guard	41.265719	-96.067495	Spree	2007
Crandon shooting	Crandon, Wisconsin	10/7/2007	Off-duty sheriff's deputy Tyler Peterson, 20, opene	6	1	7	Other	20	Unclear	The families of victims file	Yes	Issued by Forest	One rifle (assaul	AR-15 SWAT se	white	Male	http://www.river	http://www.rivernewsonline.com/main.asp?	http://www.rivern	45.5719072	-88.9028922	Mass	2007
Virginia Tech massacre	Blacksburg, Virginia	4/16/2007	Virginia Tech student Seung-Hui Cho, 23, opened	32	23	55	School	23	Yes	A district court ruled Cho	Yes	Purchased vario	Two semiautoma	9mm Glock 19, .	Asian	Male	http://www.nytim	http://abcnews.go.com/US/story?id=30522	http://www.nytim	37.2295733	-80.4139393	Mass	2007

case	location	date	summary	fatalities	injured	total_victim	location	age_of_	prior_signs_mental_h	mental_health_details	weapons_obtained_legally	where_obtained	weapon_type	weapon_details	race	gender	sources	mental_health_sources	sources_additi	latitude	longitude	type	year
Trolley Square shooting	Salt Lake City, Utah	2/12/2007	Sulejman Talović , 18, rampaged through the sh	6	4	10	Other	18	Unclear	According to one relative,	Unknown	Purchased vario	One revolver, on	Mossberg Maver	white	Male	http://www.dese	http://www.deseretnews.com/article/66020	http://www.deser	40.7606467	-111.89109	Mass	2007
Amish school shooting	Lancaster County, Penn	10/2/2006	Charles Carl Roberts, 32, shot 10 young girls in a	6	5	11	School	32	No	He told his wife that he'd	Yes	Local stores in N	One semiautoma	Springfield semi	white	Male	http://news.goog	http://www.cnn.com/2006/US/10/03/amish.	http://news.goog	39.9589	-76.0806	Mass	2006
Capitol Hill massacre	Seattle, Washington	3/25/2006	Kyle Aaron Huff, 28, opened fire at a rave afterpa	7	2	9	Other	28	No	Police were unable to find	Yes	Various sporting	Two semiautoma	.40-caliber Ruge	white	Male	http://seattletime	http://seattletimes.com/html/localnews/200	http://seattletime	47.6229	-122.3165	Mass	2006
Goleta postal shootings	Goleta, California	1/30/2006	Former postal worker Jennifer Sanmarco, 44, sho	8	0	8	Workplace	44	Yes	She was placed on retirer	Yes	Pawn shops in G	One semiautoma	9mm Smith & W	white	Female	http://www.msnb	http://www.msnbc.msn.com/id/11167920/ns	http://www.msnb	34.425571	-119.866069	Mass	2006
Red Lake massacre	Red Lake, Minnesota	3/21/2005	Jeffrey Weise, 16, murdered his grandfather, who	10	5	15	School	16	Yes	He voluntarily visited a a	No	Glock and Remi	Two semiautoma	.40-caliber Glock	Native American	Male	http://news.goog	http://www.cbsnews.com/stories/2005/03/2	http://news.goog	47.876346	-95.0169401	Mass	2005
Living Church of God shooting	Brookfield, Wisconsin	3/12/2005	Living Church of God member Terry Michael Ratz	7	4	11	Religious	44	Yes	Neighbors said he suffere	Yes	Gun dealer in W	One semiautoma	9mm Beretta ser	white	Male	http://www.cbsn	http://www.cbsnews.com/2100-201_162-67	http://www.cbsne	43.0605671	-88.1064787	Spree	2005
Damageplan show shooting	Columbus, Ohio	12/8/2004	Nathan Gale, 25, possibly upset about the breaku	5	7	12	Other	25	Yes	He was discharged from t	Yes	Received as a gi	One semiautoma	9mm Beretta 92I	white	Male	http://www.cbsn	http://www.cbsnews.com/2100-201_162-65	http://www.cbsne	39.9622601	-83.0007065	Mass	2004
Lockheed Martin shooting	Meridian, Mississippi	7/8/2003	Assembly line worker Douglas Williams, 48, open	7	8	15	Workplace	48	Yes	His cousin said he was de	Yes	Purchased from	One semiautoma	.45-caliber Ruge	white	Male	http://usatoday3	http://usatoday30.usatoday.com/news/natic	http://usatoday30	32.410842	-88.634539	Mass	2003
Navistar shooting	Melrose Park, Illinois	2/5/2001	Fired employee William D. Baker, 66, opened fire	5	4	9	Workplace	66	No	He had a criminal past, in	Yes	Some purchased	Two rifles, one re	SKS 1954R, .30-	black	Male	http://articles.ch	http://articles.chicagotribune.com/2001-02-	http://articles.chi	41.908163	-87.879908	Mass	2001
Wakefield massacre	Wakefield, Massachuse	12/26/2000	Michael McDermott, 42, opened fire on co-worker	7	0	7	Workplace	42	Yes	Psychiatrist guessed he h	Yes	Unknown	One semiautoma	.32-caliber Retol	white	Male	http://www.time.	http://articles.cnn.com/2002-04-22/justice/c	http://www.time.c	42.500429	-71.075913	Mass	2000
Hotel shooting	Tampa, Florida	12/30/1999	Hotel employee Silvio Leyva, 36, gunned down fo	5	3	8	Workplace	36	Yes	His brother called him "ur	Yes	One purchased f	One semiautoma	9mm Lorcin sem	Latino	Male	http://www.sptin	http://www.sptimes.com/News/123000/new	http://www.sptim	27.966479	-82.570586	Spree	1999
Xerox killings	Honolulu, Hawaii	11/2/1999	Byran Koji Uyesugi, 40, a Xerox service technicia	7	0	7	Workplace	40	Yes	A psychiatrist, testifying f	Yes	Hunting Supplies	One semiautoma	9mm Glock 17 s	Asian	Male	http://archives.s	http://archives.starbulletin.com/2000/06/02/	http://archives.st	21.320063	-157.876462	Spree	1999
Wedgwood Baptist Church shooting	Fort Worth, Texas	9/15/1999	Larry Gene Ashbrook, 47, opened fire inside the W	8	7	15	Religious	47	Yes	His siblings decribed him	Yes	Trader's Village f	Two semiautoma	.380-caliber, 9mi	white	Male	http://www.nytin	http://www.nytimes.com/1999/09/18/us/dea	http://www.nytim	32.664511	-97.384246	Mass	1999
Atlanta day trading spree killings	Atlanta, Georgia	7/29/1999	Day trader Mark O. Barton, 44, who had recently	9	13	22	Workplace	44	Yes	In letters, he details his de	Yes	Glock purchased	Three semiautor	.45-caliber Colt 1	white	Male	http://www.indep	http://www.independent.co.uk/news/i-dont-	http://www.indep	33.850116	-84.377839	Mass	1999
Columbine High School massacre	Littleton, Colorado	4/20/1999	Eric Harris, 18, and Dylan Klebold, 17, opened fire	13	24	37	School	17	Yes	Harris was an alleged psy	No	Purchased vario	One semiautoma	9mm Intratec DC	white	Male	http://www.usatc	http://www.slate.com/articles/news_and_pc	http://www.usato	39.604034	-105.074103	Mass	1999
Thurston High School shooting	Springfield, Oregon	5/21/1998	After he was expelled for having a gun in his locke	4	25	29	School	15	Yes	Doctors testified that he w	No	Purchased vario	Two semiautoma	9mm Glock, .22-	white	Male	http://www.katu.	http://articles.cnn.com/2000-01-21/us/kinke	http://www.katu.c	44.0462362	-123.0220289	Mass	1998
Westside Middle School killings	Jonesboro, Arkansas	3/24/1998	Mitchell Scott Johnson, 13, and Andrew Douglas	5	10	15	School	11	No	Boys displayed unruly an	No	Stolen from gran	Two semiautoma	FIE 380, .380-ca	white	Male	http://www.vpc.c	http://www.nytimes.com/1998/03/29/us/fron	http://www.vpc.o	35.8209895	-90.6682606	Mass	1998
Connecticut Lottery shooting	Newington, Connecticut	3/6/1998	Lottery worker Matthew Beck, 35, gunned down fo	5	1	6	Workplace	35	Yes	He had been hospitalized	Yes	Unknown	One semiautoma	9mm semiautom	white	Male	http://www.nytin	http://www.nytimes.com/2000/04/11/us/hole	http://www.nytim	41.6856325	-72.72983827	Mass	1998
Caltrans maintenance yard shooting	Orange, California	12/18/1997	Former Caltrans employee Arturo Reyes Torres, 4	5	2	7	Workplace	41	No	He was disgruntled after l	Yes	B&B Gun Sales	One rifle (assaul	7.62mm AK-47 (	Latino	Male	http://articles.lat	http://articles.latimes.com/1997/dec/20/new	http://articles.lati	33.7877944	-117.8531119	Spree	1997
R.E. Phelon Company shooting	Aiken, South Carolina	9/15/1997	Ex-con Hastings Arthur Wise, 43, opened fire at th	4	3	7	Workplace	43	No	An ex-con, he had been f	No	Unknown	One semiautoma	9mm semiautom	black	Male	http://www.vpc.c	http://chronicle.augusta.com/stories/1997/0	http://www.vpc.o	33.5598586	-81.721952	Mass	1997
Fort Lauderdale revenge shooting	Fort Lauderdale, Florida	2/9/1996	Fired city park employee Clifton McCree, 41, oper	6	1	7	Workplace	41	Yes	Co-workers complained a	Yes	Unknown	One semiautoma	9mm Glock sem	black	Male	http://www.nytin	http://articles.sun-sentinel.com/1996-02-11/	http://www.nytim	26.119269	-80.104119	Spree	1996
Walter Rossler Company massacre	Corpus Christi, Texas	4/3/1995	Disgruntled former metallurgist James Daniel Sim	6	0	6	Workplace	28	No	He was likely angry becau	Yes	Unknown	One semiautoma	9mm Ruger sem	unclear	Male	http://web.caller	http://web.caller.com/2000/april/03/today/lo	http://web.caller.	27.828025	-97.548198	Mass	1995
Air Force base shooting	Fairchild Air Force Base	6/20/1994	Former airman Dean Allen Mellberg, 20, opened f	5	23	28	Military	20	Yes	He was repeatedly diagno	Yes	Gun dealer in Sp	One rifle (assaul	MAK-90 semiaut	white	Male	http://articles.lat	http://articles.latimes.com/1994-06-22/news	http://articles.lati	47.61864486	-117.6483587	Mass	1994
Chuck E. Cheese's killings	Aurora, Colorado	12/14/1993	Nathan Dunlap, 19, a recently fired Chuck E. Che	4	1	5	Workplace	19	Unclear	While he was in prison av	Unknown	Unknown	One semiautoma	.25-caliber semi	black	Male	http://www.denv	http://www.5280.com/magazine/2008/12/pc	http://www.denve	39.675599	-104.844845	Mass	1993
Long Island Rail Road massacre	Garden City, New York	12/7/1993	Colin Ferguson, 35, opened fire on an eastbound	6	19	25	Other	35	Yes	Psychiatrists and others s	Yes	Turner's Outdoo	One semiautoma	9mm Ruger P89	black	Male	http://www.nytin	http://www.nytimes.com/1993/12/12/nyregi	http://www.nytim	40.7267682	-73.6342955	Spree	1993
Luigi's shooting	Fayetteville, North Caro	8/6/1993	Army Sgt. Kenneth Junior French, 22, opened fire	4	8	12	Other	22	No	He had an abusive father	Yes	Unknown	One rifle, two sh	.22-caliber rifle; t	white	Male	http://articles.lat	http://news.google.com/newspapers?id=0A	http://articles.lati	35.0529931	-78.8787058	Mass	1993
101 California Street shootings	San Francisco, Californ	7/1/1993	Failed businessman Gian Luigi Ferri, 55, opened	9	6	15	Other	55	No	He was down on his luck	No	Super Pawn and	Three semiautor	Two Intratec DC-	white	Male	http://articles.lat	http://articles.latimes.com/1993-07-03/news	http://articles.lati	37.792968	-122.397973	Mass	1993
Watkins Glen killings	Watkins Glen, New York	10/15/1992	John T. Miller, 50, killed four child-support workers	5	0	5	Other	50	Yes	The day before the shoot	Yes	Mumford Sports	One semiautoma	9mm Llama sem	white	Male	http://www.nytin	http://www.nytimes.com/1992/10/24/nyregi	http://www.nytim	42.3810555	-76.8705777	Mass	1992
Lindhurst High School shooting	Olivehurst, California	5/1/1992	Former Lindhurst High School student Eric Housto	4	10	14	School	20	No	He suffered violent physic	Yes	Local gun retaile	One rifle, one sh	.22-caliber sawe	white	Male	http://www.scho	http://www.schoolshooters.info/PL/Subject-	http://www.schoo	39.07868761	-121.5475762	Mass	1992
Royal Oak postal shootings	Royal Oak, Michigan	11/14/1991	Laid-off postal worker Thomas McIlvane, 31, oper	5	5	10	Workplace	31	Yes	Police revoked his CCW p	Yes	Local gun store	One rifle	.22-caliber Ruge	white	Male	http://www.nytin	http://www.nytimes.com/1991/11/15/us/ex-p	http://www.nytim	42.4894801	-83.1446485	Mass	1991
University of Iowa shooting	Iowa City, Iowa	11/1/1991	Former graduate student Gang Lu, 28, went on a	6	1	7	School	28	Unclear	He was described as dark	Yes	Fin & Feather in	One revolver	.38-caliber Tauru	Asian	Male	http://www.nytin	http://www.nytimes.com/1991/11/03/us/gun	http://www.nytim	41.6606893	-91.5302214	Spree	1991
Luby's massacre	Killeen, Texas	10/16/1991	George Hennard, 35, drove his pickup truck into a	24	20	44	Other	35	No	Acquaintances described	Yes	Mike's Gun Shop	Two semiautoma	9mm Glock 17, 9	white	Male	http://www.nytin	http://www.nytimes.com/1991/10/20/weekir	http://www.nytim	31.1171194	-97.7277959	Mass	1991
GMAC massacre	Jacksonville, Florida	6/18/1990	James Edward Pough, 42, opened fire at a Gener	10	4	14	Other	42	No	Police speculated he had	Yes	Unknown	One rifle, one rev	.30-caliber Unive	black	Male	http://www.nytin	http://www.nytimes.com/1990/06/20/us/haz	http://www.nytim	30.3321838	-81.655651	Mass	1990
Standard Gravure shooting	Louisville, Kentucky	9/14/1989	Joseph T. Wesbecker, 47, gunned down eight peo	9	12	21	Workplace	47	Yes	Prior to the shooting, he h	Yes	AK-47 purchase	Three semiautor	Two Intratec MA	white	Male	http://nl.newsba	http://www.nytimes.com/1989/09/16/us/dist	http://nl.newsbar	38.2542376	-85.759407	Mass	1989
Stockton schoolyard shooting	Stockton, California	1/17/1989	Patrick Purdy, 26, an alcoholic with a police recor	6	29	35	School	26	Yes	He told a mental health p	Yes	Sandy Trading P	One semiautoma	9mm Taurus sen	white	Male	http://www.recor	http://www.recordnet.com/apps/pbcs.dll/art	http://www.recor	37.9577016	-121.2907796	Mass	1989
ESL shooting	Sunnyvale, California	2/16/1988	Former ESL Incorporated employee Richard Farle	7	4	11	Workplace	39	Yes	He stalked and harassed	Yes	Various sporting	Two semiautoma	.380 ACP Brown	white	Male	http://articles.lat	http://books.google.com/books?id=JiQUkw	http://articles.lati	37.3688301	-122.0363496	Mass	1988
Shopping centers spree killings	Palm Bay, Florida	4/23/1987	Retired librarian William Cruse, 59, was paranoid	6	14	20	Other	59	Yes	He suffered from paranoi	Yes	Gun store in Nor	One rifle, one rev	Sturm, Ruger Mi	white	Male	http://articles.lat	(Supreme Court of Florida Document) http:	http://articles.lati	28.0331886	-80.6429695	Spree	1987
United States Postal Service shooting	Edmond, Oklahoma	8/20/1986	Postal worker Patrick Sherrill, 44, opened fire at a	15	6	21	Workplace	44	Unclear	He was worried he had in	Yes	Issued by Oklah	Three semiautor	.22-caliber, two .	white	Male	http://news.goog	http://newsok.com/sherrill-feared-mental-ill	http://news.goog	35.6672015	-97.42937037	Mass	1986
San Ysidro McDonald's massacre	San Ysidro, California	7/18/1984	James Oliver Huberty, 41, opened fire in a McDor	22	19	41	Other	41	Yes	The day before the shoot	Yes	Unknown	One semiautoma	9mm Browning F	white	Male	http://www.utsar	http://www.nctimes.com/news/local/article_	http://www.utsan	32.5520013	-117.0430813	Mass	1984
Dallas nightclub shooting	Dallas, Texas	6/29/1984	Abdelkrim Belachheb, 39, opened fire at an upsca	6	1	7	Other	39	Yes	During his last meal with	No	Hines Boulevard	One semiautoma	9mm Smith & W	white	Male	http://books.goo	http://books.google.com/books?id=Hr3OBv	http://books.goog	32.925166	-96.838676	Mass	1984
Welding shop shooting	Miami, Florida	8/20/1982	Junior high school teacher Carl Robert Brown, 51	8	3	11	Other	51	Yes	His second wife left him b	Yes	Garcia Gun Cen	One shotgun	Mossberg 500 P	white	Male	http://www.nytin	http://news.google.com/newspapers?id=uu	http://www.nytim	25.796491	-80.226683	Mass	1982

Mother Jones

CRIME AND JUSTICE AUGUST 24, 2012

What Exactly Is a Mass Shooting?

MARK FOLLMAN

Update, December 16, 2012: *In the wake of the massacre at Sandy Hook Elementary in Newtown, a story from the* Associated Press *suggested that mass shootings have not increased in the United States in recent years. But the* AP *cited research that uses broader criteria than the criteria we used for our investigation, which found an increase. Here is our approach, explained:*

What is a mass shooting?

Broadly speaking, the term refers to an incident involving multiple victims of gun violence. But there is no official set of criteria or definition for a mass shooting, according to criminology experts and FBI officials contacted by *Mother Jones*.

Generally, there are three terms you'll see to describe a perpetrator of this type of gun violence: mass murderer, spree killer, or serial killer. An FBI crime classification report from 2005 identifies an individual as a mass murderer **if he kills four or more people in a single incident (not including himself), typically in a single location.** (The baseline of four fatalities is key—more on that just below.)

The primary distinction between a mass murderer and a spree killer, according to the FBI, is that the latter strikes in multiple locations, though still in a relatively short time frame. The third type, a serial killer, is distinguished by striking over a longer time frame, in multiple locations, with opportunity for what the FBI report refers to as "cooling-off periods" in between attacks.

How often do mass shootings occur?

Beginning in July, after the movie theater slaughter in Aurora, Colorado, we documented and analyzed 62 mass shootings from the last 30 years. As we delved into the research, we realized that robust data on this subject was hard to come by, in part due to the lack of clear criteria. We were focused on the question of how many times Aurora-like events had actually happened. We honed our criteria accordingly:

- The attack must have occurred essentially in a single incident, in a public place;
- We excluded crimes of armed robbery, gang violence, or domestic violence in a home, focusing on cases in which the motive appeared to be indiscriminate mass murder;
- The killer, in accordance with the FBI criterion, had to have taken the lives of at least four people.

The traumatic events included in our guide to mass shootings are the kind that tend to grab national attention—school and workplace shootings, attacks in shopping malls or government buildings—but they represent only a sliver of America's gun violence, which results in approximately 30,000 deaths annually.

Interactive guide: 30 Years of Mass Shootings in America.

Since the 1980s, the baseline of four fatalities has generally been used for studying mass murder, according to Professor James Alan Fox of Northeastern University, who has written multiple books on the subject. But as Fox agreed when we spoke, that baseline ultimately is arbitrary. Was it not a "mass shooting" in 2008, for example, when a man walked into a church in Tennessee and opened fire with a shotgun, killing two and injuring seven? Dropping the number of fatalities by just one, or including motives of armed robbery, gang violence, or domestic violence, would add many, many more cases to the list.

According to a recent report in Time magazine (available only to subscribers, and whose criteria is unclear), there've been "nearly 20 mass shootings" *every year* on average during the last three and a half decades.

Why didn't you include the infamous DC Beltway sniper attacks on your mass shootings map?
We've been asked this question numerous times. The man who killed 10 and wounded 3 others a decade ago (along with a young accomplice) was a serial killer: He committed multiple attacks over several weeks, in different locations. It was a particularly tense time for people living in the DC metro area—the shooter "terrorized our neighborhood," as one person wrote to me in an email—but the case did not fit the criteria described above.

Is *Mother Jones* focusing on this stuff as part of a conspiracy to take away Americans' gun rights?
No. One of our lead reporters on this beat, Adam Weinstein, who covered the Trayvon Martin killing and investigated how the National Rifle Association helped spread "Stand Your Ground" laws nationwide, is a Navy veteran and third-generation gun owner. Multiple other *Mother Jones* staffers are experienced with guns.

The debate over guns in the United States is extremely contentious and polarizing, and we think that the more reporting and clear data available about guns, the better. That mass shootings keep happening is an undeniable fact. Why they do, and how to stop them, is a matter for further investigation.

Update, January 8, 2013: Where can I learn more about MoJo's investigation?
See our recently published America Under the Gun: a Special Report on gun laws and the rise of mass shootings, which contains interactive maps, charts, and dozens of stories from over the last year.