Oasis or Mirage?

by

David A. Freedman

Statistics Department

UC Berkeley, CA 94720-3860

This paper will review the design of statistical studies, and comment on the difficulty of drawing causal inferences from non-experimental data. The most basic design is a comparison of rates for two groups of subjects. Subjects in one group get the treatment of interest; subjects in the other group are the controls. The difficulty is ensuring that the groups are similar, apart from the treatment.

Experiments versus observational studies

In a *randomized controlled experiment*, the investigators assign the subjects to treatment or control, for instance, by tossing a coin. In an *observational study*, the subjects assign themselves. The difference is crucial, because of *confounding*. Confounding means a difference between the treatment group and the control group, other than the causal factor of primary interest. The confounder may be responsible for some or all of the observed effect that is of interest.

In a randomized controlled experiment, near enough, chance will balance the two groups. Thus, confounding is rarely a problem. In an observational study, however, there often are important differences between the treatment and control groups. That is why experiments provide a more secure basis for causal inference than observational studies. When there is a conflict, experiments usually trump observational studies. However, experiments are hard to do, and well-designed observational studies can be informative. In social science and medicine, a lot of what we know—or think we know—comes from observational studies.

Most studies on smoking are observational. Taken together, they make a powerful case that smoking kills. A great many lives have been saved by tobacco control measures which were prompted by the observational studies. There are a few experiments; the treatments (like counseling) were aimed at getting smokers to quit. Paradoxically, results from the experiments are inconclusive.

Even with studies on smoking, confounding can be a problem. Smokers die at higher rates from cirrhosis than non-smokers. Cigarettes cause heart disease and cancer, but they do not cause cirrhosis. What explains the association? The confounder is drinking. Smokers drink more, and alcohol causes cirrhosis. Here, confounding can be handled by sorting people into groups by the amount they drink and the amount they smoke. At each level of drinking, there will be little association between smoking and the death rate from cirrhosis. By contrast, at each level of smoking, there will be a strong association between drinking and cirrhosis.

This sort of analysis can be done by *cross-tabulation*. The key idea is making comparisons within smaller and more homogeneous groups of subjects. However, large samples are required. Furthermore, confounding variables are often hard to spot. (Generally, a confounder has to be associated with the effect, and with the factor thought to be causal.) With more variables, cross-tabulation gets complicated and the sample gets used up rather quickly. Applied workers may then try to control for the confounders by statistical modeling, which is our chief topic.

What is a statistical model?

A statistical model assumes a relationship between the effect and (i) the primary variable the investigators think of as the cause, as well as (ii) potential confounders. The objective is to get statistical (if not experimental) control over the confounders, isolating the effect of the primary variable. Applied workers tend to use relationships that are familiar and tractable; linearity often plays a key role. Certain numerical features of the model are estimated from the data, for instance, coefficients in a linear combination of variables. The investigators determine whether such coefficients are *statistically significant*, that is, hard to explain by chance. If the coefficient of the primary variable is significant and has the right sign, the causal hypothesis has been "verified" by the data analysis.

The search for significance

Significance-testing is an integral part of modeling. This creates problems, because significant findings can be due to chance. If investigators test at the 5% level, and nothing is going on except chance variation, then 5% of the "significant" findings will be due to chance. In short, with many studies and many tests, significant findings are bound to crop up. Journals often look for significant findings; authors oblige. There is tacit agreement to ignore contradictory results that are found along the way, and search efforts are seldom reported.

In consequence, published significance levels are very difficult to interpret. Given the null hypothesis, the chance of a spurious but significant finding can be held to a desired level, like 5%. Given a significant finding, however, the chance of the null hypothesis being true is ill-defined—especially when publication is driven by the search for significance.

The validity of the models

Fitting models is justified when the models derive from strong prior theory, or the models can be validated by data analysis. (This is trickier than it sounds: high R^2's and low P-values don't do much to justify causal inference.) In social science and medicine, the picture is often untidy. Do we have the right variables in the model? Are variables measured with reasonable accuracy? Is the functional form correct? What about assumptions on error terms? Does causation run in the direction assumed by the model? These questions seldom have satisfactory answers, and the list of difficulties can be extended.

Assumptions behind models are rarely articulated, let alone defended. The problem is exacerbated because journals tend to favor a mild degree of novelty in statistical procedures. Modeling, the search for significance, the preference for novelty, and lack of interest in assumptions—these norms are likely to generate a flood of non-reproducible results. The next section will discuss some examples of current interest.

Case studies

(i) *Vitamins, fruits, vegetables, and a low-fat diet* protect in various combinations against cancer, heart disease, and cognitive decline, according to many observational studies. After controlling for confounders by modeling, investigators find reductions in risk that are statistically significant. Dozens of big experiments have been done to confirm the findings. Surprisingly—or not—the experiments generally contradict the observational data.

On balance, vitamin supplements are not beneficial. The low-fat diet rich in fruits and vegetables is not beneficial. (By contrast, there is good evidence to show that the Mediterranean diet does protect against heart failure.) The chief problem with the observational studies seems to be confounding. People who eat five helpings of fruits and vegetables a day are different from the rest of us, in ways that are hard to model.

(ii) *Hormone replacement therapy* protects against heart disease. According to its proponents,

> "Consistent evidence from over 40 epidemiologic studies demonstrates that postmenopausal women who use estrogen therapy after the menopause have significantly lower rates of heart disease than women who do not take estrogen."

However, large-scale experiments show that hormone replacement therapy is at best neutral. Again, the most plausible explanation is confounding. Women who take hormones are different from other women, in ways that are not picked up by the models.

Medical opinion changes only slowly in response to data. Believers in hormone replacement therapy claim that if you adjust using the "right" model, the observational studies agree with the experiments. Skeptics might reply that without the experiments, the modelers wouldn't know when to stop. In any case, the degree of agreement between the two kinds of studies is rather imperfect, even after the modelers have done what they can.

(iii) *Get-out-the-vote campaigns*. There are many attempts to mobilize voters by non-partisan telephone canvassing. Do these campaigns increase the rate at which people vote? Statistical modeling suggests a big effect, but the experimental evidence goes the other way.

(iv) *Welfare programs*. For two decades, investigators have compared experimental and non-experimental methods for evaluating job training programs and the like. There are substantial discrepancies, and there does not seem to be any analytic method that reliably eliminates the biases in the observational data.

(v) *Meta-analysis* is often proposed as a way to distill the truth from a body of disparate studies. Systematic reviews of the literature can be very useful. However, formal meta-analysis of observational studies—with effect sizes, confidence intervals, and P-values—often comes down to an unconvincing model for results from other models. Even at that, measurement problems can be intractable because many published reports lack critical detail. Such difficulties are rarely acknowledged. And what about unpublished reports? Conventional solutions to the "file drawer" problem depend on another layer of unconvincing assumptions.

From a critical perspective, a great deal of meta-analysis appears to be problematic. From another perspective, the software is easy to use, and results are welcome in the journals—especially when effects are highly significant and go in the right direction. Needless to say, proponents of vitamins, low fat diets, hormone replacement therapy, matching, modeling, and meta-analysis will disagree with every syllable (this sentence apart).

How should experimental data be analyzed?

Experimental data are frequently analyzed through the prism of models. This is a mistake, because randomization does not guarantee the validity of the assumptions. Bias is likely unless the sample is large, and standard errors are liable to be wrong. Before investigators turn to modeling, they should be comparing rates or averages in the treatment group and the control group.

As is only to be expected, experiments have problems of their own. One is *crossover*. Subjects assigned to treatment may refuse, while subjects assigned to control may insist on doing something else. The *intention-to-treat* analysis compares results for those assigned to the treatment group with those assigned to the control group. This should be the primary analysis, because it avoids bias by taking advantage of the experimental design.

Intention-to-treat measures the effect of assignment rather than treatment itself. Under some circumstances, the effect of treatment can be estimated, even if there is crossover. After the intention-to-treat tables, there is room for secondary analyses that might illuminate the results or generate hypotheses for future investigation. For instance, investigators might fit models, or look for subgroups of subjects with unusually strong responses to treatment.

What about subgroup analysis?

Richard Peto says that you should always do subgroup analysis and never believe the results. (The same might be true of modeling.) In a large-scale study, whether experimental or observational, the principal tables should be specified in advance of data collection. After that, of course, investigators should look at their data and report what they see. However, the analyses that were not pre-specified should be clearly differentiated from the ones that were. Such recommendations might even apply to small-scale studies, although repeating the study may be an adequate corrective.

Whose data are they anyway?

Social scientists are often generous in sharing their data. However, replicating the results—in the narrow sense of reproducing the coefficient estimates—is seldom possible. The data have been cleaned up, the exact form of the equations has been lost. . . . In the medical sciences, it is seldom possible to get the data, or even any detail on the modeling, except for the version number of the statistical package. Studies are often run according to written protocols, which is good, but deviations from protocol are rarely noted. Some agencies make data available to selected researchers for further analysis, but the conditions are very restrictive. There is no excuse for this state of affairs.

The equations and the data should be archived and publicly available. The scope of data analysis should not be restricted. Confidentiality of personal information needs to be protected, but this is a tension that can be resolved. Deviations from protocol should be disclosed, like the search efforts. Assumptions should be identified. Journal articles should explain which assumptions have been tested, and why untested assumptions are plausible. Authors have responsibilities here; so do the journals and the funding agencies. Empirical studies are expensive, and they are usually funded by tax dollars. Why shouldn't taxpayers have access to the data? Even if crass arguments about money are set aside, isn't transparency a basic scientific norm?

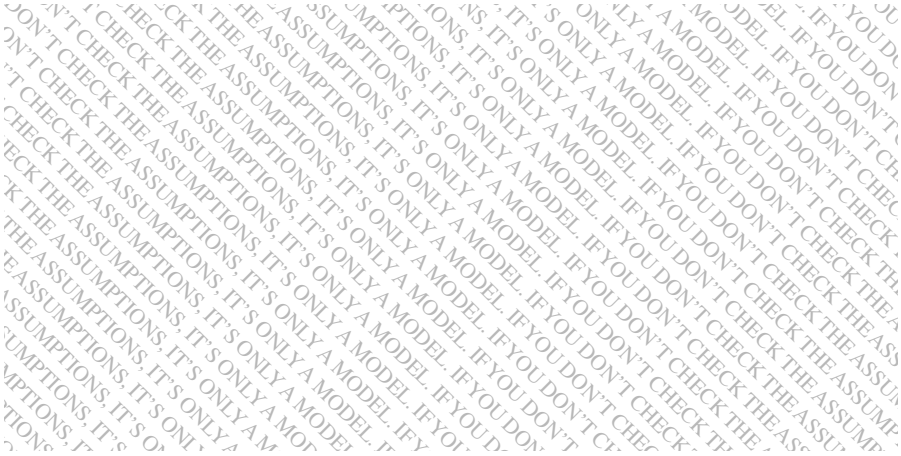
Peer review takes care of it?

Some experts think that peer review validates published research. For those of us who have been editors, associate editors, reviewers, or the targets of peer review, this argument may ring hollow. Even for careful readers of journal articles, the argument may seem a little farfetched.

The perfect is the enemy of the good?

There are defenders of the research practices criticized here. What do they say? Some maintain there is no need to worry about multiple comparisons. Indeed, worrying in public reduces the power of the studies, and impedes (i) the ability of the epidemiologist to protect the public, or (ii) the ability of the social scientist to assist the decision-maker. Only one thing is missing: an explanation of what the P-values might mean.

Other defenders love to quote variations on George Box's old saw. No model is perfect, but some models are useful. A moment's thought generates some uncomfortable questions. Useful to whom, and for what? How would the rest of us know? If models could be calibrated in some way, objections to their use would be much diminished. (As ongoing scholarly disagreements might suggest, calibration is not the easiest of tasks in medicine and the social sciences.) Until a happier future arrives, imperfections in models require further thought, and routine disclosure of imperfections would be helpful. A watermark might be a good interim step.



A public-domain watermark

References

Altman, D.G., Schulz, K.F., Moher, D., et al. (2001). "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration," *Annals of Internal Medicine*, 134, 663–694. Documents flaws in reporting randomized trials and makes detailed suggestions for improvement.

Arceneaux, K., Gerber, A.S., and Green, D.P. (2006). "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment," *Political Analysis*, 14, 37–62. Shows that matching substantially over-estimates the effect of a voter mobilization campaign on turnout.

Austin, P.C., Mamdani, M.M., Juurlink, D.N., and Hux, J.E. (2006). "Testing Multiple Statistical Hypotheses Resulted in Spurious Associations: A Study of Astrological Signs and Health," *Journal of Clinical Epidemiology*, 59, 964–969. An elegant demonstration that the search for significance takes the searcher into uncharted territory.

Benson, K. and Hartz, A.J. (2000). "A Comparison of Observational Studies and Randomized, Controlled Trials," *New England Journal of Medicine*, 342, 1878–1886. Argues that a well-conducted observational study is as good as an experiment. Also see Concato and Horowitz (2000), below.

Berk, R.A. and Freedman, D.A. (2003). "Statistical Assumptions as Empirical Commitments," in *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd ed., ed. T.G. Blomberg and S. Cohen, New York: Aldine de Gruyter, 235–254. Pages 244–248 state the assumptions behind meta-analysis, and give a critique; little merit is found in typical meta-analyses of observational material; meta-analysis of experimental studies can be useful, especially when testing the null hypothesis that treatment has no effect.

Briggs, D.C. (2005). "Meta-Analysis: A Case Study," *Evaluation Review*, 29, 87–127. Discusses problems in meta-analysis. Focus is measurements issues, replicability. Demonstrates the fragility of one prominent meta-analysis.

Chlebowski, R.T., Blackburn, G.L., Thomson, C.A., et al. (2006). "Dietary Fat Reduction and Breast Cancer Outcome: Interim Efficacy Results from the Women's Intervention Nutrition Study," *Journal of the National Cancer Institute*, 98, 1767–76. Demonstrates a protective effect by statistical modeling; also see Pierce et al. below.

Concato, J.S. and Horowitz, R. (2000). "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs," *New England Journal of Medicine*, 342, 1887–1892. Like Benson and Hartz (2000), above.

Freedman, D.A. (1999). "From Association to Causation: Some Remarks on the History of Statistics," *Statistical Science*, 14, 243–258. Reprinted in *Journal de la Société Française de Statistique* (1999) 140, 5–32 and in *Stochastic Musings: Perspectives from the Pioneers of the Late 20th Century*, ed. J. Panaretos (2003), Hillsdale, N.J.: Lawrence Erlbaum Associates, 45–71. Discusses experiments, observational studies, confounding; the evidence on smoking; problems with models.

Freedman, D.A. (2005). *Statistical Models: Theory and Practice*, New York: Cambridge University Press. Chapter 1 discusses experiments, observational studies, confounding, the intention-to-treat principle, the evidence on smoking, causal inference from observational data. Pages 64–65 discuss the search for significance. Pages 85–96 lay out the assumptions needed to infer causation by regression. There are many examples to illustrate the limits of modeling.

Freedman, D.A. (2006). "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review*, 30, 691–713. Discusses Neyman's model for causal inference and methods that correct for crossover; shows that randomization does not justify the assumptions behind regression models (or logits or probits). Compares intention-to-treat, per-protocol, and treatment-received analyses; identifies the parameters being estimated.

Freedman, D.A. (2007). "On Regression Adjustments to Experimental Data," In press, *Advances in Applied Mathematics*. Asymptotic results for regression estimates applied to experimental data.

Freedman, D.A. and Petitti, D.B. (2001). "Salt and Blood Pressure," *Evaluation Review*, 25, 267–87. Discusses epidemiology of salt, impact of protocol violations and publication bias.

Freedman, D.A., Petitti, D.B., and Robins, J.M. (2004). "On the Efficacy of Screening for Breast Cancer," *International Journal of Epidemiology*, 33, 43–73 (with discussion). Correspondence, 1404–1406. Dissects one meta-analysis that attracted a lot of attention. Pages 72–73 give an example of correction for crossover.

Freedman, D.A., Pisani, R., and Purves, R.A. (2007). *Statistics*. 4th ed., New York: W.W. Norton & Company. Chapters 1–2 introduce the design of experiments, confounding, observational studies. There are reference to the smoking literature and the literature on vitamins. Chapter 29 discusses pitfalls in testing, including the search for significance.

Glazerman, S., Levy, D.M., and Myers, D. (2003). "Nonexperimental versus Experimental Estimates of Earnings Impacts," *Annals of the American Academy of Political and Social Science*, 589, 63–93. Empirical study of matching and modeling versus experiments.

International Journal of Epidemiology, June 2004, vol. 33, no. 3. Reprints a classic observational study claiming a protective effect for hormone replacement therapy, with commentary from many perspectives—including the author of the study, who still says he got it right.

Ioannidis, J.P. (2005). "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association*, 294, 218–228. Results often do not replicate well, especially for observational studies; reasons are suggested; also see Ioannidis (2006), below.

Ioannidis, J.P. (2006). "Evolution and Translation of Research Findings: From Bench to Where?" *PLoS Clinical Trials*, www.plosclinicaltrials.org. Why a lot of medical research doesn't go anywhere.

Journal of the American Medical Association, Feb. 8, 2006, vol. 295, pp. 629ff. Results from the Women's Health Initiative, comparing experiments and observational studies: is a low fat diet protective against colon cancer, breast cancer, or heart disease?

Journal of Econometrics, March–April 2005, vol. 125, nos. 1–2. Empirical study of matching and modeling versus experiments; matching and modeling do sometimes work. Also see *Review of Economics and Statistics*, February 2004, below.

Kunz, R. and Oxman, A.D. (1998). "The Unpredictability Paradox: Review of Empirical Comparisons of Randomised and Non-Randomised Clinical Trials," *British Medical Journal*, 317, 1185–1190. Observational studies tend to be over-enthusiastic.

Lawlor, D.A., Davey Smith, G., Kundu, D., et al. (2004). "Those Confounded Vitamins: What Can We Learn from the Differences between Observational vs Randomised Trial Evidence," *Lancet*, 363, 1724–1727. Discusses the contrast between the observational studies and the experiments.

de Lorgeril, M., Salen, P., Martin, J.L., et al. (1999). "Mediterranean Diet, Traditional Risk Factors, and the Rate of Cardiovascular Complications after Myocardial Infarction: Final Report of the Lyon Diet Heart Study," *Circulation*, 99, 779–785. Shows protective effect of Mediterranean diet in a randomized controlled experiment.

McMahon, J.A., Green, T.J., Skeaff, C.M., et al. (2006). "A Controlled Trial of Homocysteine Lowering and Cognitive Performance," *New England Journal of Medicine*, 354, 2764–2772. Experiment on vitamin supplements contradicts results of observational studies.

Miller, E.R., Pastor-Barriuso R., Dalal, D., et al. (2005). "Meta-Analysis: High-Dosage Vitamin E Supplementation May Increase All-Cause Mortality," *Annals of Internal Medicine*, 142, 37–46. Reviews experiments on vitamin E supplements; shows there are no benefits. The author's demonstration that vitamins are harmful is strongly model-dependent.

Oakes, M.W. (1990). *Statistical Inference*. Chestnut Hill, MA.: Epidemiology Resources Inc. Section 7.2 discusses problems of meta-analysis. Explains why the conventional solution to the "file drawer problem" is unconvincing.

Petitti, D.B. (2002). "Hormone Replacement Therapy for Prevention," *Journal of the American Medical Association*, 288, 99–101. Describes how the experiments on hormone replacement therapy contradict the observational studies.

Petitti, D.B. and Freedman, D.A. (2005). "How Far Can Epidemiologists Get with Statistical Adjustment?" *American Journal of Epidemiology*, 162, 1–4. Discusses attempts to reconcile experiments and observational studies by additional modeling.

Pierce, J.P., Natarajan, L., Caan, B.J., et al. (2007). "Influence of a Diet Very High in Vegetables, Fruit, and Fiber and Low in Fat on Prognosis Following Treatment for Breast Cancer: The Women's Healthy Eating and Living (WHEL) Randomized Trial," *Journal of the American Medical Association*, 298, 289–98. Demonstrates no protective effect; also see Chlebowski et al. above.

Pocock, S.J., Collier, T.J., Dandreo, K.J., et al. (2004). "Issues in the Reporting of Epidemiological Studies: A Survey of Recent Practice," *British Medical Journal*, 329, 883–887. "This survey raises concerns regarding inadequacies in the analysis and reporting of epidemiological publications in mainstream journals." Authors have developed standards for reporting observational studies. See http://www.strobe-statement.org. Compare Altman et al. (2001) above on the CONSORT statement. For additional discussion, see *Epidemiology*, November 2007, vol. 18, no. 6.

Review of Economics and Statistics, February 2004, vol. 86, no. 1. Like *Journal of Econometrics*, March-April 2005, above.

Rothman, K.J. (1990). "No Adjustments Are Needed for Multiple Comparisons," *Epidemiology*, 1, 43–46. Explains that adjustment for multiple testing would reduce power; does not explain the meaning of unadjusted P-values.

Shapiro, S. (1994). "Meta-Analysis, Shmeta-Analysis," *American Journal of Epidemiology*, 140, 771–791 (with discussion). Title says it all.

U.S. Preventive Services Task Force (2003). "Routine Vitamin Supplementation to Prevent Cancer and Cardiovascular Disease: Recommendations and Rationale," *Annals of Internal Medicine*, 139, 51–55. A review of the literature "could not determine the balance of benefits and harms of routine use of supplements of vitamins A, C, or E; multivitamins with folic acid; or antioxidant combinations for the prevention of cancer or cardiovascular disease."

Writing Group for the Women's Health Initiative Investigators (2002). "Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results from the Women's Health Initiative Randomized Controlled Trial," *Journal of the American Medical Association*, 288, 321–333.

Vivekananthan, D.P., Penn, M.S., Sapp, S.K., et al. (2003). "Use of Antioxidant Vitamins for the Prevention of Cardiovascular Disease: Meta-Analysis of Randomised Trials," *Lancet*, 361, 2017–2023. No protective effect from vitamin E; risk of harm from beta-carotene.

About the author

David A. Freedman received his B.Sc. degree from McGill and his Ph.D. from Princeton. He is professor of statistics at U.C. Berkeley, and a former chair of the department. He has been Sloan Professor and Miller Professor, and is now a member of the American Academy of Arts and Sciences. In 2003, he received the John J. Carty Award for the Advancement of Science from the National Academy of Sciences. He has written several books, including a widely-used elementary

text, as well as many papers in probability and statistics. He has worked on the foundations of statistics, procedures for testing and evaluating models, epidemiology, statistics and the law. He has worked as a consultant for the Carnegie Commission, the City of San Francisco, and the Federal Reserve, as well as several departments of the U.S. Government. He has testified on employment discrimination, voting rights, census adjustment, and agricultural loan policies.