TEACHING CAUSAL INFERENCE IN EXPERIMENTS AND OBSERVATIONAL STUDIES

Donald B. Rubin, Harvard University
I Oxford St., 6th fl., Cambridge, MA 02138

Key Words: potential outcomes, Rubin Causal Model (RCM), Fisherian inference, Neymanian inference, Bayesian inference, experiments, observational studies, instrumental variables, noncompliance, statistical education

Inference for causal effects is a critical activity in many branches of science and public policy. The field of statistics is the one field most suited to address such problems, whether from designed experiments or observational studies. Consequently, it is arguably essential that departments of statistics teach courses in causal inference to both graduate and undergraduate students. This presentation will discuss some aspects of such courses based on: a graduate level course taught at Harvard for a half dozen years, sometimes jointly with the Department of Economics (with Professor Guido Imbens, now at UCLA), and current plans for an undergraduate core course at Harvard University. An expanded version of this brief document will outline the courses' contents more completely. Moreover, a textbook by Imbens and Rubin, due to appear in 2000, will cover the basic material needed in both courses.

The current course at Harvard begins with the definition of causal effects through potential outcomes. Causal estimands are comparisons of the outcomes that would have been observed under different exposures of units to treatments. This approach is commonly referred to as 'Rubin's Causal Model - RCM" (Holland, 1986), but the formal notation in the context of randomization-based inference in randomized experiments goes back to Neyman (1923), and the intuitive idea goes back centuries in various literatures; see also Fisher (1918), Tinbergen (1930) and Haavelmo (1944). The label "RCM" arises because of extensions (e.g., Rubin, 1974, 1977, 1978) that allowed a formal framework beyond randomized experiments and beyond randomization-based inference; it also allowed the formal consideration of complications, such as unintended missing data and noncompliance. Because of the importance of making this very intuitive and transparent definition of causal effects central to students' thinking about causality, the course devotes some brief time to the history of this idea, including the history of needed assumptions, such as “no-interference-between-units” (Cox, 1958) and the more encompassing “stable-unit-treatment-value” assumption (Rubin, 1984). Without some such “exclusion restrictions”, to use the language of economists, causal inference is impossible, and understanding this limitation is critical.

This RCM framework is now rather generally accepted in many fields. For example, in psychology, see Wilkinson et al. (1999); in economics see the transition to adopt it reflected by comparing Heckman (1979) to Heckman (1989), and Pratt and Schlaifer (1984) to Pratt and Schlaifer (1988), after discussion by Holland (1989) and Rosenbaum and Rubin (1984), respectively. Also see Baker, (1998), Dempster (1990), Efron and Feldman (1991), Gelman and King (1991), Greenland and Pool (1988), Greenland, Robins, and Pearl (1999), Holland (1988a, b, 1989), Holland and Rubin (1983), Kadane and Seidenfeld (1990), Robins (1987, 1989), Rosenbaum (1987), Smith and Sugden (1988), Sobel (1990, 1995, 1996), Sugden (1988), and their references. A recent article exploring whether the full potential outcomes framework can be avoided when conducting causal inference is Dawid (2000) with discussion.

The essential role of the assignment mechanism is then introduced: without a model for how treatments get assigned to units, formal causal inference, as least using probabilistic statements, is impossible. It is also critical that students appreciate this, especially because most articles purporting to do causal inference in many areas of application of statistics never even consider the assignment mechanism unless the study was randomized. Examples of assignment mechanisms and a classification of them is introduced: classic randomized experiments, unconfounded and regular designs, more general ignorable designs, and nonignorable designs (Rubin, 1976). At this point, we do not NEED any more assumptions to proceed with some forms of causal inference: those based solely on the randomization distribution induced by randomized assignment. Unless students understand how to analyze

randomized experiments validly and why their analysis is relatively simple, it is impossible to teach them how to analyze observational data validly for causal effects. Thus, they need to understand very well the basis of causal inference in randomized experiments, and then to use this foundation to learn how to draw causal inferences in nonrandomized studies.

There are two distinct forms of randomization inference, one due to Neyman (1923) and the other due

to Fisher (1925). Fisher's is the more direct conceptually and is introduced next. It is closely related to the mathematical idea of proof by contradiction. It basically is a stochastic "proof' by contradiction giving the significance level (or p-value) of the null hypothesis of absolutely no effect whatsoever: the probability of a result (represented by the value of an a priori defined statistic, such as the difference of observed average treatment outcome minus observed average control outcome) this rare or more rare if the null hypothesis were true, where the probability is over the distribution induced by the assignment mechanism. This form of inference is very elegant but very limited: how much can be learned from finding, with high probability, a model that does not fit the data, especially as measured by some possibly arbitrary or mathematically convenient statistic?

Neyman's form of randomization-based inference can be viewed as drawing inferences by evaluating the expectations of statistics over the distribution induced by the assignment mechanism the essential idea is the same as in Neyman's (1934) classic article on randomization-based (now often called 'designed-based") inference in surveys. Typically, an unbiased estimator of the causal estimand is created, and an unbiased, or upwardly biased estimator, of the variance of that unbiased estimator is found (bias and variance both defined with respect to the randomization distribution). Then an appeal is made to the central limit theorem for the normality of the estimator over its randomization distribution, whence a confidence interval for the causal estimand is obtained. This form of inference is less direct than Fisher's; it is really aimed at evaluations of procedures, and only indirectly at inference from the data at hand. Nevertheless, it forms the basis for most of what is done in important areas of application (e.g., the world of pharmaceutical development and approval, the world of randomized clinical trials in medicine), and therefore, once again, it is critical that students understand the framework.

The third form of statistical inference for causal effects is Bayesian, where the model for the assignment mechanism is supplemented with a model for the data (Rubin, 1978). A causal inference is obtained as the posterior distribution of the causal estimand, which follows from the posterior predictive distribution of the unobserved potential outcomes, which in turn follows by Bayes theorem from the observed data and the models for the assignment mechanism and the data. All calculations are described in terms of simulating this posterior predictive distribution - basically by multiply-imputing the missing potential outcomes. This approach opens the door for the much more demanding computational methods used in the latter part of the graduate course, which would be absent from the undergraduate core course. The Bayesian approach, which is the foundation for typical linear and loglinear models, is by far the most intuitive, direct and flexible of the modes of inference for causal effects, but achieves these benefits by postulating a distribution for the data, which the randomization-based approaches avoid. Such a distribution can be very useful, but is like a loaded gun in the hands of a child, fraught with danger for the naive data analyst. This inherent danger is an important message to convey to students, especially those who have been exposed to complicated automatic causal modelling software and so may be more willing to accept and use procedures they do not fully understand.

In practice, one should be willing to use the best features of all approaches. In simple classical randomized experiments with normal-like data, the three approaches give very similar practical answers, but they do not in more difficult cases, where each perspective provides different strengths. At this point, the relationships between these three conceptually cogent formulations and the more common linear equation, regression and path diagram approaches are made. Care must be taken here, especially in an undergraduate course, to avoid confusing students through the introduction of generally inappropriate methods, which can achieve appropriate answers in the simplest settings, but which lead to inappropriate answers in realistic settings.

The course then turns to 'regular" designs, which are like classical randomized experiments except that the probabilities of treatment assignment are allowed to depend on observed covariates and so can vary from unit to unit (e.g., older males have probability .8 of being assigned the new treatment; younger males,, .6; older females, .5; and younger females, .2. Horvitz and Thompson (1952; Cochran, 1963) estimation is considered with known assignment probabilities, and Fisherian and Neymanian inferences are derived and compared with Bayesian model-based estimation for robustness and efficiency. This leads to estimated propensity score methods (Rosenbaum and Rubin, 1983), and their use in combination with models. These techniques are then applied to observational studies.

The key idea is to conceptualize the observational study as if it were a regular design, and therefore an ignorable design, and to use this template to draw inferences. Many illustrative examples are given, using matching, subclassification, and model building (e.g., Rosenbaum and Rubin, 1984, 1985;

Reinisch et al., 1995; Rubin, 1997; Smith, 1997). Using real examples, including ones where there exist nearly parallel randomized and nonrandomized studies (e.g., as in LaLonde, 1986; Dehejia and Wahba, 1999), it becomes clear that in general the combination of both propensity score methods and Bayesian model building is superior to either alone for the objective of achieving essentially the same answer from an observational study as from the parallel randomized experiment. In the graduate course, some attention is paid to theoretical results on propensity scores, such as in Rubin and Thomas (1992a,b) and on the combination of propensity score methods and modeling (Rubin and Thomas, 2000), but the most important ideas are easily conveyed by the real examples, and so the theory would be avoided in an undergraduate course, although not in the graduate course.

Because the conceptualization of an observational data set as arising from a regular (and thus, ignorable) assignment mechanism is an assumption, it is important to consider deviations from that assumption. Sensitivity of conclusions to the assumption of ignorability is then studied using methods such as those of Rosenbaum and Rubin (1985) and Rosenbaum (1995), which display how point and interval estimates change as a function of assumptions. Bounds on point estimates (e.g., Manski and Nagin, 1998; Horowitz and Manski, 1999) are also considered but usually as a special case of sensitivity analysis. Such analyses are relatively easy to explain to students at all levels. Other issues about ignorability (e.g., tests for it as in Rosenbaum, 1984) can be considered in a graduate course but not in an undergraduate course.

The topic of noncompliance in randomized studies is then introduced because it is a relatively well-understood island between the shores of the perfect randomized experiment and the uncontrolled observational study. For example, a random half of doctors are encouraged to give their patients a flu shot, but some patients in each random group do and do not take the flu shot (Hirano, Imbens, Rubin and Zhou, 1999). Such studies are ignorable for the assigned treatment (e.g., encouragement to take the flu shot) but nonignorable for the received treatment (getting the flu shot), and comprise a more general template for the analysis of observational studies than do ignorable designs. The approach taken to such studies in the course bridges classic econometric and statistical approaches to causal inference as described in Angrist, Imbens, and Rubin (1996). The most direct result is the method-of-moments "instrumental variables' estimate for the "complier average causal effect', and used, for example, by Sommers and Zeger (1991) to estimate the effect of vitamin A on infant survival. The much preferred approach to the analysis of such data, however, is likelihood/Bayesian, as described in Imbens and Rubin (1997), but this requires the introduction of iterative maximization and simulation techniques, such as the EM algorithm (Dempster, Laird and Rubin, 1977) and data augmentation (Tanner and Wong, 1987), which is appropriate for a graduate course, but is too demanding for an undergraduate core course. Severd examples of this kind of approach are given, including actual randomized experiments with noncompliance (e.g., Sommer and Zeger, 1991; Goetghebeur and Molenberghs, 1996; Baker and Lindeman, 1994; Frangakis, Rubin and Zhou, 1999) and observations studies where some version of the instrumenw variables approach seems plausible (e.g., Ettner, 1996).

The graduate course continues with the consideration of how to deal with some common complications with randomized experiments with noncompliance, such as missing outcome data (Frangakis and Rubin, 1999), randomization in clustered groups (Frangakis, Rubin and Zhou, 1999), and missing covariates and multivariate outcomes (Barnard, Frangakis, Hill and Rubin, 2000), which are all beyond the scope of an undergraduate course, as is the topic of multiple treatments in a longitudinal setting (e.g., Robins, 1997).

The final lecture is much like this little article: a review of how to apply fundamentally sound statistical thinking about causal inference to estimate causal effects in experiments and observational studies.

As stated in the abstract, I firmly believe that some such course is arguably an essential ingredient of any statistics departments' offerings, and moreover is an important component of the education of anyone who intends to draw conclusions about causal effects.

REFERENCES

Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996). "Identification of Causal Effects Using Instrumental Variables.' Journal of the American Statistical Association, 91, 434, as
Applications Invited Discussion Article with discussion and rejoinder, pp. 444-472.

Baker, S.G. (1998). "Analysis of Survival Data From a Randomized Trial With All-or-None Compliance &: Estimating the Cost-Effectiveness of a Cancer Screening Program.' Journal of the American Statistical Association, 93, pp. 929-934.

Baker, S.G. and Lindeman, K.S. (1994). "The Paired Availability Design: A Proposal for Evaluating Epidural Analgesia During Labor.' Statistics in Medicine, 13, pp. 2269-2278.

Bamard, J., Frangakis, C., Hill, J. and Rubin. D.B. (2000). "School Choice in NY City: A Bayesian Analysis of an Imperfect Randomized Experiment." To appear in Case Studies in Bayesian Statistics, V. 5, Kass et alia (eds.). New York: Springer-Verlag.

Cochran, W.G. (1963). SaMRIini! Technigues, 2nd ed. New York: Wiley.

Cox, D.R. (1958). Planning, of Experiments. New York: John Wiley.

Dawid, A.P. (2000). "Causal Inference Without Counterfactuals". To appear in Journal of the American Statistical Association, June 2000. With discussion.

Dehejia, R.H. and Wahba, S. (1999). 'Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs.' To appear in Journal of the American Statistical Association.

Dempster, A.P. (1990). "Causality and Statistics.' Journal of Statistical Planning and Inference, 25, pp. 261-278.

Dempster, A.P., Laird, N. and Rubin, D.B. (1977).
'Maximum Likelihood from Incomplete Data Via the EM Algorithm." The Journal of the Royal Statistical Society, Series B, 39, 1, pp. 1-38. With discussion and reply.

Efron, B. and Feldman, D. (1991). 'Compliance as an Explanatory Variable in Clinical Trials." Journal of the American Statistical Association, 86, pp. 9-17.

Ettner, S.L. (1996). "The Timing of Preventive Services for Women and Children: The Effect of Having a Usual Source of Care.' American Journal of Public Health, 86, 17481754.

Fisher, R.A. (1918). "The Causes of Human Variability." Eugenics Review, 10, pp. 213220.

Fisher, R.A. (1925). Statistical Methods for Research Workers. London: Oliver & Boyd.

Frangakis, C. and Rubin, D.E. (1999). "Addressing Complications of Intention-To-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes." Biometrika, 86, 2, pp. 366-379.

Frangakis, C., Rubin, D.B. and Zhou, X.-H. (1999). "The Clustered-Encouragement-Design."
Proceedinjzs of the Biometrics Session of the American Statistical Association, pp. 71-79.

Gelman, A. and King, G. (1991). "Estimating Incumbency Advantage Without Bias.' American Journal of Political Science, 3-4, pp. 1142-1164.

Goetghebeur, E. and Molenberghs, G. (1996). "Causal Inference in a Placebo-Controlled Clinical Trial with Binary Outcome and Ordered Compliance." Journal of the American Statistical Association 91, pp. 928-934.

Greenland, S. and Poole, C. (1988). 'Invariants and Noninvariants in the Concept of Interdependent Effects." Scandinavian Journal of Work and Environmental Health, 14, pp. 125-129.

Greenland, S., Robins, J.M. and Pearl, J. (1999). 'Confounding and Collapsibility in Causal Inference." Statistical Science, 1-4, pp. 29-46.

Haavelmo, T. (1944). 'The Probability Approach in Econometrics.' Econometrica, 15, pp. 413419.

Heckman, J.J. (1979). "Sample Selection Bias as a Specification Error.' Econometrica, 47, pp. 153-161.

Heckman, J.J. (1989). 'Causal Inference and
Nonrandom Samples.' Journal of Educational Statistics, 14, pp. 159-168.

Heckman, J.J. (1996). Comment on "Identification of Causal Effects Using Instrumental Variables." Journal of the American Statistical Asso 91, pp. 459-462.

Hirano, K., Imbens, G., Rubin, D.B. and Zhou, X.H. (1999). 'Estimating the Effect of an Influenza Vaccine in an Encouragement Design." Revision to appear in Biostatistics.

Holland, P. (1986). 'Statistics and Causal Inference.' Journal of the American Statistical Association, 81, pp. 945-970.

Holland, P.W. (1988a). "Causal Inference, Path Analysis, and Recursive Structural Equation Models.' Sociological Methodology, pp. 449484.

Holland, P.W. (198&b). Comment on "Employment Discrimination and Statistical Science' by A.P. Dempster. Statistical Science, 3, pp. 186-188.

Holland, P.W. (1989). 'It's Very Clear." Comment on 'Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Man over Training' by I Heckman and V. Hotz. Journal of the American Statistical Association, 84, pp. 875-877.

Holland, P.W. and Rubin, D.B. (1983). "On Lord's Paradox.' Principles of Modern Psychological Measurement: A Festschrift for Frederic M.

Lord, Wainer and Messick (eds.), pp. 3-25. Hillsdale, NJ: Earlbaum.

Horowitz, J.L.and Manski, C.F. (1999). 'Nonparametric Analysis of Randomized Experiments with Missing Covariates and Outcome Data.' Journal of the American Statistical Association.

Horvitz, D.G. and Thompson, D.J. (1952). 'A Generalization of Sampling Without replacement From a Finite Population.' Journal of the American Statistical Association, 47, pp. 663-685.

Imbens, G. and Rubin, D.B. (1997). 'Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." The Annals of Statistics, 25, 1, pp. 305-327.

Kadane, J. B. and Seidenfeld, T . (1990). "Randomization in a Bayesian Perspective." Journal of Statistical Planning and Inference, 25, pp. 329-346.

LaLonde, R. (1986). "Evaluating the Econometric Evaluations of Training Programs." American Economic Review, 76, pp. 604-620.

Manski, C.F. and Nagin, D.S. (1998). "Bounding Disagreements About Treatment Effects: A Study of Sentencing and Recidivism." Sociological Methodology, 28, pp. 99-137.

Neyman, J. (1923). "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9." Translated in Statistical Science, 5, pp. 465-480, 1990.

Neyman, J. (1934). "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection.' Journal of the Roval Statistical Society, Series A, 97, pp. 558-606.

Pratt, J.W. and Schlaifer, R. (1984). "On the Nature and Discovery of Structure." Journal of the American Statistical Association, 79, pp. 9-33. Pratt, J.W. and Schlaifer, R. (1988). "On the Interpretation and Observation of Laws." Journal of Econometrics, 39, pp. 23-52.

Reinisch, J., Sanders, S., Mortensen, E. and Rubin, D. B. (1995). "In Utero Exposure to Phenobarbital and Intelligence Deficits in Adult Men.' The Journal of the American Medical Association, 274, 19, pp. 1518-1525.

Robins, J.M. (1987). "A Graphical Approach to the Identification and Estimation of causal Parameters in Mortality Studies With Sustained Exposure Periods." Journal of Chronic Diseases (Suppl. 2), 40, pp. 139S161S.

Robins, J.M. (1989). "The Control of Confounding by Intermediate Variables." Statistics in Medicine, 8, pp. 679-701.

Robins, J.M. (1997)."Causal Inference From Complex Longitudinal Data.' Latent Va6able Modeling and Applications to Causality, Lecture Notes in Statistics, M. Berkane (ed.), 120, pp. 69-117. New York: SpringerVerlag.

Rosenbaum, P.R. (1984). 'From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment.' Journal of the American Statistical Association, 79, pp. 41-48.

Rosenbaum, P.R. (1987). "The Role of a Second Control Group in an Observational Study (with discussion)." Statistical Science, 2, pp. 292316.

Rosenbaum, P.R. (1995). Observational Studies. New York: Springer-Verlag.

Rosenbaum, P. and Rubin, D.B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects.' Biometrika, 70, pp. 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.' Journal of the American Statistical Association, 79, pp. 516-524.

Rosenbaum, P. R. and Rubin, D.B. (1985). "Constructing a Control Group Using Multivariate Matched Sampling Incorporating the Propensity Score.' The American Statistician, 39, pp. 33-38.

Rubin, D.B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psycboloizv, 66, 5, pp. 688-701. Rubin, D.B. (1976). "Inference and Missing Data.' Biometrika, 63, 3, pp. 581-592. With discussion and reply.

Rubin, D.B. (1977). "Assignment to Treatment Group on the Basis of a Covariate." Journal of Educational Statistics, 2, 1, pp. 1-26.

Printer's correction note 3, P. 384.

Rubin, D.B. (1978). "Bayesian Inference for Causal Effects: The Role of Randomization.' The Annals of Statistics, 7, 1, pp. 34-58.
Rubin, D. B. (4a4). "William G. Cochran's Contributions to the Design, Analysis, and Evaluation of Observational Studies." W.G. Cochran's Impact on Statistics, Rao and Sedransk (eds.). New York: Wiley, pp. 37-69.