# Donald Rubin[*]

Andrew Gelman[†]

3 Oct 2017

Donald Rubin (1943–) is a statistician who has made major contributions in statistical modeling, computation, and the foundations of causal inference. He is best known, perhaps, for the EM algorithm (a mathematical framework for iterative optimization, which has been useful for mixture models, hierarchical regression, and many other problems for which closed-form solutions are unavailable); multiple imputation as a method for accounting for uncertainty in statistical analysis with missing data; a Bayesian formulation of instrumental-variables analysis in economics; propensity scores for controlling for multiple predictors in observational studies; and, especially, the potential-outcomes framework of causal inference.

Causal inference is central to social science. The effect of an intervention on an individual $i$ (which could be a person, a firm, a school, a country, or whatever particular entity is being affected by the treatment) is defined as the difference in the outcome $y_i$, comparing what would have happened under the intervention to what would have happened under the control. If these *potential outcomes* are labeled as $y_i^T$ and $y_i^C$, then the causal effect for that individual is $y_i^T - y_i^C$. But for any given individual $i$, we can never observe both potential outcomes $y_j^0 \ and \ y_j^1$, thus the causal effect is impossible to directly measure. This is commonly referred to as the *fundamental problem of causal inference*, and it is at the core of modern economics and policy analysis.

Resolutions to the fundamental problem of causal inference are called "identification strategies"; examples include linear regression, nonparametric regression, propensity score matching, instrumental variables, regression discontinuity, and difference in differences. Each of these has spawned a large literature in statistics, econometrics, and applied fields, and all are framed in response to the problem that it is not possible to observe both potential outcomes on a single individual.

From this perspective, what is amazing is that this entire framework of potential outcomes and counterfactuals for causal inference is all relatively recent, deriving from three papers by Rubin in the 1970s ("Estimating causal effects of treatments in randomized and nonrandomized studies," Journal of Educational Psychology, 1974; "Assignment to treatment group on the basis of a covariate," Journal of Educational Statistics, 1977; "Bayesian inference for causal effects: The role of randomization," Annals of Statistics, 1978). Although these ideas seem so natural today, it was a conceptual leap to consider $y^T$ and $y^C$ to be two separate variables, given that at most only one of them can be observed. Like all good ideas, this one has echoes in the past, and connections have been drawn to a long-forgotten paper from 1923 by mathematician Jerzy Neyman in the Polish Annals of Agricultural Sciences defining potential outcomes for randomized experiments, a 1943 Econometria paper by Trygve Haavelmo ("The statistical implications of a system of simultaneous equations"), and a 1951 paper by economist A. D. Roy ("Some thoughts on the distribution of earnings") presenting a model for a latent bivariate distribution of skills.

The econometrician Guido Imbens has written, "The potential outcome framework became popular in the econometrics literature on causality around 1990. See Heckman (1990, American Economic Review, Papers and Proceedings, "Varieties of Selection Bias," 313–318) and Manski (1990, American Economic Review, Papers and Proceedings, "Nonparametric Bounds on Treatment Effects," 319–323). The causality literature is actually one where there is a lot of cross-discipline

[*] For the Encyclopaedia of Social Research Methods, edited by Paul Atkinson, Sara Delamont, Melissa Hardy, and Malcolm Williams.

[†] Department of Statistics and Department of Political Science, Columbia University, New York

referencing, and in fact a lot of cross-discipline collaborations between statisticians, econometricians, political scientists and computer scientists."

The potential-outcome or counterfactual-based model of casual inference has led to conceptual, methodological, and applied breakthroughs in core areas of applied statistics.

The key conceptual advances come from the idea of a unit-level treatment effect, $y_i^T - y_i^C$, which, although it is unobservable, can be aggregated in various ways. So, instead of the treatment effect being thought of as a parameter ("$\beta$" in a regression model), it is an average of individual effects. From one direction, this leads to the "local average treatment effect" of Angrist and Imbens, the principal stratification idea of Frangakis and Rubin, and various other average treatment effects considered in the causal inference literature. Looked at another way, the fractalization of treatment effects allows one to determine what exactly can be identified from any study. A randomized experiment can estimate the average treatment effects among the individuals under study; if those individuals are themselves a random sample, then the average causal effect in the population is also identifiable. With an observational study, one can robustly estimate a local average treatment effect in the region of overlap between treatment and control groups, but inferences for averages outside this zone will be highly sensitive to model specification. The overarching theme here is that the counterfactual expression of causal estimands is inherently nonparametric and unbounds causal inference from the traditional regression modeling framework. The counterfactual approach thus fits in very well with modern agent-based foundations of micro- and macro-economics which are based on individual behavior.

The methodological advances have come in estimation and in identification. The first innovation was propensity score matching (Rosenbaum and Rubin, 1983, 1984) which allowed researchers to control for imbalance in observational studies, under certain assumptions. Later work by statistician Jennifer Hill, economist Susan Athey, and others has moved to nonparametric models, bringing in modern tools of machine learning and prediction to attack longstanding issues of model dependence in regression-based causal estimates. Advances in causal identification have come from deeper understanding of the relationships between information and inference in the causal setting. Important work here includes the Bayesian formulation of instrumental variables from Angrist, Imbens, and Rubin (1996) and recent work on regression discontinuity and difference in differences estimation by many different econometricians. Again, this all takes place within the potential-outcome framework and the estimation of local average treatment effects and treatment interactions.

On the applied side, social science has moved in the past forty years to a much greater concern with causality, and much greater rigor in causal measurement, what in economics is called "identification." Traditionally, in statistics, identification comes from the likelihood, that is, from the parametric statistical model. The counterfactual model of causal effects has shifted this: with causality defined nonparametrically in terms of latent data, there is a separation between (a) definition of the estimand, and (b) the properties of the estimator—a separation that has been fruitful both in the definition of causal summaries such as various conditional average treatment effects, and in the range of applications of these ideas. Organizations such as MIT's Poverty Action Lab and Yale's Innovations for Poverty Action have revolutionized development economics using randomized field experiments, and similar methods have spread within political science. Within micro-economics, identification strategies have been used not just for media-friendly "cute-o-nomics" but also in areas such as education research and the evaluation of labor and trade policies where randomized experiments are either impossible or impractical to do at scale. In psychology and medicine there are longstanding traditions of experimentation, but there the potential outcome framework has been useful in addressing real-world complexities such as dropout and noncompliance.

In addition to his aforementioned contributions to the theory and methods of causal inference, Rubin has made several other major advances in statistical methods which have been impactful in social science. The EM algorithm, presented in a 1977 paper by Dempster, Laird, and Rubin, is a general framework for maximum likelihood estimation with missing data and has been used in thousands of examples, most notably mixture models and latent variable models. Rubin's 1976 paper on inference and missing data introduced the concept of missingness at random, and later work by Rosenbaum and Rubin delineated the related concept of ignorability, thus making rigorous various previously unclear notions of when it was necessary to account for selection in data analysis. Rubin also, with Rod Little, wrote the standard textbook on statistics with missing data, and did all this in the context of active applications in education research, economics, and public health. Finally, and in addition to all of this foundational work, Rubin was a key contributor (with Lindley, Novick, Dempster, and a few others) in the Bayesian hierarchical-modeling or random-effects revolution in statistical analysis, which has had major impacts in education and sociology (for example, the study of school, neighborhood, and other "contextual" effects) and which is beginning to make its way into economics with the study of varying treatment effects in experiments and observational studies. Indeed, Rubin has had a major influence in social science just by virtue of being a coauthor of the leading textbook on Bayesian statistics. Rubin has also been influential within the field of Bayesian statistics through his work on posterior predictive checking, which generalizes the classical composite hypothesis testing problem to the scenario in which no pivotal quantity is available.

This article about Donald Rubin's contributions to social research should not be taken as a denigration of the work of many others in this area. Indeed, Rubin has throughout his career engaged in longstanding collaborations with the psychologist Robert Rosenthal, the statisticians Arthur Dempster and Roderick Little, the economist Guido Imbens, and many students and others. And, just keeping the focus on causal inference, important related work has been done by the economists James Heckman, Charles Manski, Joshua Angrist, and Guido Imbens, the epidemiologist Sander Greenland and biostatistician James Robins, the computer scientist Judea Pearl, the statistician Paul Rosenbaum, the psychometrician Kenneth Bollen, and many others. Rubin has been at the center of this revolution and has helped focus it on the interaction between statistical models and applied problems, with others working in more theoretical directions or in specific application areas.

Finally, one mark of Rubin's influence in applied statistics are the terms that he and his colleagues introduced, including "missing at random," "ignorability," "propensity scores," "potential outcomes," and "Bayesian data analysis." All these terms are now standard in statistics and represent a particular attitude toward statistical modeling and inference.