Causality: Rubin (1974)

2nd Lecture - Hedibert Lopes

Insper - Institute of Education and Research

October 6th, 2015

Donald Bruce Rubin

- Born Washington DC in 12/22/1943, but moved to Evanston, IL, in 1944
- High school: Physics and Math
- Princeton 1961-1965: Major in Psychology
- Harvard 1965-1970
  - MSc in Computer Science 1966
  - PhD Statistics in 1970 (Bill Cohran)
- Department of Statistics Harvard: 1970-1971
- Educational Testing Service: 1971-1981
  - Princeton, Harvard, UC Berkeley, UC Austin, UW Madison
- University of Chicago: 1981-1983
- Harvard: 1985-now

Source: Li and Mealli (2014) A Conversation with Donald B. Rubin. *Statistical Science*, 29(3), 439-457.

Ascendants - Mathematics Genealogy Project[1]

- Galileo Galilei, Università di Pisa 1585
- Vincenzo Viviani, Università di Pisa 1642
- M.A. University of Cambridge
  - Isaac Barrow (1652), Isaac Newton (1668)
  - Roger Cotes (1706), Robert Smith (1715), Walter Taylor (1723), Stephen Whisson (1742), Thomas Postlethwaite (1756), Thomas Jones (1782),
  - Adam Sedgwick (1811), William Hopkins (1830), Sir Francis Galton (1847), Karl Pearson (1879)
- John Wishart, ?
- William Cochran, M.A. University of Glasgow 1931
- **Don Rubin**, Ph.D. Harvard University 1971

[1] http://genealogy.math.ndsu.nodak.edu/id.php?id=47145

Rubin: 48 students and 166 descendants

1980s

Rosenbaum, Lax, Heitjan, Li, Schenker, Raghunathan, Weld, Thomas, Zaslavsky.

1990s

Gelman, Meng, Belin, Schafer, D'Agostino, Liu, Larsen, Li, Wu, Thurston, Du, Roseman, Zanutto, Frangakis, Reiter

2000s

Hill, Shen, Zhang, Cook, Stuart, Jin, Yu, Greiner, Morgan

2010s

Chretien, Licht, Gutman, Lock, Pattanayak, Loong, Espinosa, Kolar, Liublinska, Sabbaghi, Watson, Andric, Ding, Feller, Lee

Citation and h-index

Scholar	Citations[2]		h-index	
	All	$\geq$2010	All	$\geq$2010
Donald Rubin	178	86	113	81
James Heckman	119	54	141	101
Judea Pearl	65	25	81	52
James Robins	36	19	98	68
Josh Angrist	35	21	64	54
Guido Inbems	29	19	58	49
Phil Dawid	13	5	47	29
Daniel Kahneman	227	104	118	92
Amos Tversky	193	77	163	131

[2] In thousands.

Additional literature

Rubin (1974)
Estimating causal effects of treatments in randomized and nonrandomized studies.
Journal of Educational Psychology, 56, 688-701.

Rubin (1977)
Assignment to treatment group on the basis of a covariate.
Journal of Educational Statistics, 2, 1-26.

Rubin (1978)
Bayesian inference for causal effects: the role of randomization.
Annals of Statistics, 6, 34-58.

Rubin (1990)
Formal mode of statistical inference for causal effects.
Journal of Statistical Planning and Inference, 25, 279-292.

Rubin (2004)
Teaching statistical inference for causal effects in experiments and observational studies.
Journal of Educational and Behavioral Statistics, 29, 343-367.

Rubin (2005)
Causal inference using potential outcomes: design, modeling, decisions.
Journal of the American Statistical Association, 100, 322-331.

Rubin (2008)
For objective causal inference, design trumps analysis.
Annals of Applied Statistics, 2, 808-840.

The ETS decade: missing data, EM and causal inference

In 1976, Rubin published the paper *Inference and Missing Data* in Biometrika that lays the foundation for modern analysis of missing data.

In 1977, with Arthur Dempster and Nan Laird, Rubin published the EM paper *Maximum Likelihood from Incomplete Data via the EM Algorithm* in JRSS-B.

In 1974, 1977, 1978, Durbin published a series of papers that lay the foundation for the *Rubin Causal Model*.

Rubin on his 1974 paper[3]

I wrote this in some form when I was still at Harvard, more as notes for an introductory statistics course for psychologists.

Someone suggested that I spruce it up a bit and submit it for publication.
I did, but then couldn't get it published anywhere!
Every place that I submitted the piece, rejected it.
1) "every baby statistics student knows this",
2) "its completely wrong"!,

Sometimes the comments were even insulting, especially so because I was submitting statistical work from my position at ETS rather than a respected university statistics department.

I asked around ETS and someone suggested Journal of Educational Statistics.

[3] Rubin (2014) Converting rejections into positive stimuli. In Lin et al., eds, *Past, Present, and Future of Statistical Science*, Chapter 49, pages 593-603. Chapman and Hall/CRC.

Solving real problems

I always worried about solving real problems.

I didn't read the literature to uncover a hot topic to write about.

I always liked math, but I never regarded much of mathematical statistics as real math – much of it is just so tedious. Can you keep track of these epsilons?

Were you aware of Neyman's work before?

I wasn't aware of his work defining potential outcomes until 1990 when his Ph.D. thesis was translated into English, although I attributed much of the perspective to him because of his work on surveys in Neyman (1934) and onward (see Rubin, 1990a, followed by Rubin, 1990b).

- Neyman (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *JRSS-B*, 97, 558-625.
- Neyman (1990) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5, 465-472. Translated from the Polish and edited by Dąbrowska and Speed.
- Rubin (1990a). Formal modes of statistical inference for causal effects. *JSPI*, 25, 279-292.
- Rubin (1990b). Comment on "Neyman (1923) and causal inference in experiments and observational studies." *Statistical Science*, 5, 472-480.

Rubin Causal Model

Paul Holland coined the term **Rubin Causal Model (RCM)** referring to the potential outcome framework to causal inference (Holland, 1986).

Neyman is pristinely associated with the development of potential outcomes in randomized experiments, no doubt about that.

But in the 1974 paper, I made the potential outcomes approach for defining causal effects front and center, not only in randomized experiments, but also in observational studies, which apparently had never been done before.

Potential outcome: device to facilitate causal inference

Question: Do you believe potential outcomes exist in people as fixed quantities, or is the notion that potential outcomes are a device to facilitate causal inference?

Answer: Definitely the latter. Among other things, a person's potential outcomes could change over time, and how do we know the people who were studied in the past are still exchangeable with people today? But there are lots of devices like that in science.

Effects of a cause or Cases of an effect?

Question: In the RCM, cause/intervention should always be defined before you start the analysis. In other words, the RCM is a framework to investigate the effects of a cause, but not the causes of an effect. Some criticize this as a major limitation. Do you regard this as a limitation? Do you think it is ever possible to draw inference on the causes of effects from data, or is it, per se, an interesting question worth further investigation?

Answer: I regard "the cause" of an event topic as more of a cocktail conversation topic than a scientific inquiry, because it leads to an essentially infinite regress. You can't talk sensibly about the cause of an event; you can talk about "but for that cause (and there can be many but fors), what would have happened?" All these questions can be addressed hypothetically. But the cause? The notion is meaningless to me.

Rubin's (1974) abstract

A discussion of matching, randomization, random sampling, and other methods of controlling extraneous variation is presented.

The objective is to specify the benefits of randomization in estimating causal effects of treatments.

The basic conclusion is the

randomization should be employed whenever possible

but that the use of

carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases.

Randomization: cost, ethics and time

Often the only immediately available data are observational (nonrandomized) and either

a) *the* **cost** *of performing the equivalent randomized experiment to test all treatments is prohibitive(e.g., 100 reading programs under study);*

b) *there are* **ethical reasons** *why the treatments cannot be randomly assigned (e.g., estimating the effects of heroin addiction on intellectual functioning); or*

c) *estimates based on results of experiments would be* **delayed many years** *(e.g., effect of childhood intake of cholesterol on longevity).*

In cases such as these, it seems more reasonable to try to estimate the effects of the treatments from nonrandomized studies than to ignore these data and dream of the ideal experiment or make "armchair" decisions without the benefit of data analysis.

Very simple example

Context: $2N$ units (e.g., subjects), half having been exposed to an experimental (E) treatment (e.g., a compensatory reading program) and the other half having been exposed to a control (C) treatment (e.g., a regular reading program).

Randomized experiment: Treatments E and C were assigned to the 2N units randomly, that is, using some mechanism that assured each unit was equally likely to be exposed to E as to C.

Otherwise, the study is called a nonrandomized study, a quasi-experiment, or an observational study.

Objective and central question

Objective: determine for some population of units (e.g., underprivileged sixth-grade children) the "typical" causal effect of the E versus C treatment on a dependent Variable Y, where Y could be dichotomous (e.g., success-failure) or more continuous (e.g., score on a given reading test).

Central question: What are the benefits of randomization in determining the causal effect of the E versus C treatment on Y?

Causal effect of the E versus C treatment

Intuitively, the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between

> *what would have happened at time t_2 if the unit had been exposed to E initiated at time t_1*

and

> *what would have happened at time t_2 if the unit had been exposed to C initiated at time t_1.*

Example:

- If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone.
- Because an hour ago I took two aspirins instead of just a class of water, my headache is now gone.

Trial = (unit, t_1, t_2)

t_1: time of initiation of a treatment (unit exposed to E **or** C).
t_2: time of measurement of a dependent variable, Y.

$y(E)$: the value of Y measured at t_2 on the unit, given that the unit received the experimental Treatment E initiated at t_1;

$y(C)$: the value of Y measured at t_2 on the unit, given that the unit received the control Treatment C initiated at t_1;

Then

$$y(E) - y(C)$$

is the causal effect of the E versus C treatment on Y for that trial, that is, for that particular unit and the times t_1, t_2.

Example

unit = a particular child

experimental treatment (E) = enriched reading program

control treatment (C) = regular reading program.

y: score on a reading test 10 days after treatment was initiated.

$y(E) = 38$

$y(C) = 34$

Then, the causal effect of the reading test for the trial of the enriched program versus the regular program is $38 - 34 = 4$ more items correct.

Fundamental problem of causal inference

> *The problem of measuring $y(E) - y(C)$ is that we can never observe both $y(E)$ and $y(C)$ since we cannot return to time t_1 to give the other treatment.*

"Typical" causal effect[4]

M trials for which we want the "typical" causal effect. Thus,

$$y_j(E) - y_j(C)$$

is the causal effect of the E versus C treatment for the j^{th} trial.

An obvious definition of the "typical" causal effect of the E versus C treatment for the M trials is the average (mean) causal effect for the M trials:

$$\frac{1}{M}\sum_{j=1}^{M}[y_j(E) - y_j(C)]$$

[4]Other possible definitions of the typical causal effects are the median causal effect or the mid mean causal effect.

Two-trial study

The typical causal effect for the two trials is

$$\frac{1}{2}[y_1(E) - y_1(C) + y_2(E) - y_2(C)]. \quad (1)$$

The estimate of this quantity from the study is either

$$y_1(E) - y_2(C) \quad \text{or} \quad y_2(E) - y_1(C).$$

Neither equation is necessarily close to (1) or to the causal effect for either unit[5]

$$y_1(E) - y_1(C) \quad \text{or} \quad y_2(E) - y_2(C).$$

If the Treaments E and C were randomly assigned to units, then

$$\frac{1}{2}[y_1(E) - y_2(C)] + \frac{1}{2}[y_2(E) - y_1(C)]$$

is an "unbiased" estimate of the desired typical causal effect.

[5] Unless there is matching.

Randomization vs matching

If the two units react identically in their trials, then randomization is absolutely irrelevant.

Having closely "matched" trials increases the closeness of the calculated experimental minus control difference to the typical causal effect for the two trials, while random assignment of treatments does *not* improve that estimate.

Although two-trial studies are almost unheard of in the behavioral sciences, they are not uncommon in the physical sciences.

In the physical sciences . . . there are models that successfully assign most variability to specific causes than in the social sciences where often important causal variables have not been identified.

The $2N$ trial study

Suppose there are $2N$ trials ($N > 1$) in the study, half with N units having received the E treatment and the other half with N units having received the C treatment.

Typical causal effect of the E versus C treatment on Y for the 2N trials:

$$\tau = \frac{1}{2N} \sum_{j=1}^{2N} [y_j(E) - y_j(C)].$$

Let S_E (S_C) denote the set of indices of the E trials (C trials). Then,

$$\bar{y}_d = \frac{1}{N} \sum_{j \in S_E} y_j(E) - \frac{1}{N} \sum_{j \in S_C} y_j(C)$$

is the difference between the average observed Y in the E trials and the average observed Y in the C trials.

How close is the estimate $\bar{y}_d$ to typical causal effect τ?

Whether treatments are randomly assigned or not, no matter how carefully matched the trials, and no matter how large N, a skeptical observer could always eventually find some variable that systematically differs in the E trials and C trials (e.g., length of longest hair on the child) and claim that $\bar{y}_d$ estimates the effect of this variable rather than τ, the causal effect of E versus C treatment.

Within the experiment there can be no refutation of this claim; only a logical argument explaining that the variable cannot causally affect the dependent variable or additional data outside the study can be used to counter it.

What are the benefits of randomization...

...besides the intuitive ones that follow from making all systematic sources of bias into random ones?

Randomization provides a mechanism to derive probabilistic properties of estimates without making further assumptions.

- The average E-C difference is an "unbiased" estimate of τ, the typical effect for the 2N trials.
- Precise probabilistic statements can be made indicating how unusual the observed E-C difference, $\bar{y}_d$, would be under specific hypothesized causal effects.

Randomization set[6]

Completely randomized experiment
Treatments randomly assigned to trials with no restriction

$$r = \binom{2N}{N}$$

Randomized block experiment
Treatments randomly assigned within matched pairs

$$r = 2^N$$

[6] William Cochran & Gertrude Cox (1957) *Experimental Designs* (2nd Edition).

Typical causal effect for the 2N trials

The contribution of the j^{th} trial to the average E-C difference is

$$\frac{y_j(E)}{N} \quad \text{and} \quad -\frac{y_j(C)}{N}$$

in two halves of the r allocations in the randomization set.

The expected contribution of the j^{th} trial to the average E-C difference is

$$\frac{1}{2N}[y_j(E) - y_j(C)].$$

The typical causal effect for the 2N trials is

$$\tau = \frac{1}{2N}\sum_{j=1}^{2N}[y_j(E) - y_j(C)]$$

Probabilistic statements from the randomization set

The unbiasedness of the E-C difference for τ . . . hardly solves the problem of estimating the typical causal effect.

Randomization provides a mechanism for making precise probabilistic statements indicating how unusual the observed E-C difference, $\bar{y}_d$, would be under specific hypotheses.

Example: If hypothesized values of individual causal effects are $\tilde{\tau}_j,\ j = 1, \ldots, 2N$, then

$$\tilde{\tau} = \frac{1}{2N} \sum_{j=1}^{2N} \tilde{\tau}_j$$

is the hypothesized typical causal effect for the 2N trials.

Framework

1. From $\tilde{\tau}_j, y_j(E), j \in S_E, y_j(C), j \in S_C$, one can compute hypothesized values $\tilde{y}_j(C)$ and $\tilde{y}_j(E)$ for all 2N trials:
   - For $j \in S_E$
     - $\tilde{y}_j(E) = y_j(E)$
     - $\tilde{y}_j(C) = y_j(E) - \tilde{\tau}_j$
   - For $j \in S_C$
     - $\tilde{y}_j(E) = y_j(C) + \tilde{\tau}_j$
     - $\tilde{y}_j(C) = y_j(E)$
2. Then one can compute hypothesized average E-C difference for each of the r allocations of the 2N trials in the randomization set.
3. Finally, one can compute a significance level.

Under the hypothesis that the causal effects are given by the $\tilde{\tau}_j$, for $j = 1, \ldots, 2N$, the probability that we would observe an average E-C difference that is as far or farther form $\tilde{\tau}$ than the one we have observed is m/r where m is the number of allocations in the randomization set that yield E-C differences that are as far or farther from $\tilde{\tau}$ than $\bar{y}_d$.

More on potential outcomes

Rubin (2004)[7] says:

> *Using potential outcomes to define causal effects in general is now relatively well accepted in many fields.*

In economics, he points out to

> *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 (1984a), respectively.*

[7] Rubin (2004) Teaching Statistical Inference for Causal Effects in Experiments and Observational Studies. *Journal of Educational and Behavioral Statistics*, 29, 343-367.

Heckman (1979) Sample selection bias as a specification error. *Econometrica*, 47,153-161.

Heckman and Hotz (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *JASA*, 84, 862-874.

Holland (1989) It's very clear. *JASA*, 84, 875-877.

Heckman (1989) Causal inference and nonrandom samples. *Journal of Educational and Behavioral Statistics*, 14, 159-168.

Pratt and Schlaifer (1984) On the nature and discovery of structure. *JASA*, 79, 9-21.

Rosenbaum and Rubin (1984) Comment: Estimating the effects caused by treatments. *JASA*, 79, 26-28.

Pratt and Schlaifer (1988) On the interpretation and observation of laws. *Journal of Econometrics*, 39, 23-52.

Remaining lectures

1. Holland (1986) Statistics and causal inference
   October 13th - André Yoshizumi, IME/USP

2. Pearl (1995) Causal diagrams for empirical research
   October 20th - Paloma Uribe, IME/USP

3. Angrist-Imbens-Rubin (1996) Identification of causal effects using IV
   November 3rd - Sergio Firpo, EESP/FGV

4. Dawid (2000) Causal inference without counterfactuals
   November 10th - Julio Trecenti, IME/USP

5. Vansteelandt-Goetghebeur (2003) Causal inference with generalized structural mean models
   November 24th - Manasses Nóbrega, UFABC

6. Heckman-Pinto (2015) Causal analysis Haavelmo
   December 1st - Hedibert Lopes - INSPER