Essential Concepts for Causal Inference in Randomized Experiments and Observational Studies: a remarkable history

Donald B. Rubin
Harvard University

Prologue to causal inference

- My Introduction
  - Physics – John Wheeler 1961
  - Psychology & consciousness – Julian Jaynes 1964
  - Experimental design – William Cochran 1968
- Clear Separation Between
  - Science = object of inference
  - What is done to learn about the science
    - Intervene to measure aspects at a point in time
- Same notation/representation of science no matter how we try to learn about or measure it
- Missing data always exist
  - Cannot go back in time
  - Heisenberg uncertainty principle & observer effect

Fisher (1925) – Actually randomize

- Thereby create balance on all pre-treatment variables in expectation
- Recondite advice = rerandomize if unbalanced for observed ones
  - Theory and application finally being pursued – computing possible
- Hypothetically rerandomize for assessment – Fiducial distribution
  - Stochastic proof by contradiction of sharp null hypothesis
  - Special case of posterior predictive p-value (Rubin, AoS 1984)
- Implicit definition of non-null causal effects from Fisher, 1918):

  If we say, "This boy has grown tall because he has been well fed," we are not merely tracing out the cause and effect in the individual instance; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter .
- But NO explicit notation for formalizing non-null causal effects, despite tremendous geometric insights under sharp null hypothesis based on symmetry arguments

Neyman (1923) – Potential outcomes

- Define estimands in randomized experiments as functions of potential outcomes for N units
  - Y(0) = Array of potential outcomes under treatment 0
  - Y(1) = Array of potential outcomes under treatment 1
  - Etc., implicitly assuming SUTVA: $Y_i(w)$ function of (i ,w)
  - Cannot observe both on any one unit – "Heisenberg"
- Evaluate operating characteristics of procedures (e.g., estimators) over randomization distribution
- Role of non-additive unit-level causal effects
- But later denied depth of understanding in 1923

Comments on these insights

- 20th Century insights, like those in quantum mechanics
  - Estimands are defined in terms of measurable quantities, which are not simultaneously measurable, even theoretically
- Was Fisher really the first? Peirce, late 19th century?
- RCTs quickly dominated agriculture and animal breeding – (Fisher DOE 1935) -- throughout Commonwealth and US
  - More applied work (e.g., Kempthorne, Cochran & G Cox, Box, D Cox)
  - Supporting mathematical work (e.g., ISI Mahalanobis, Bose, Nair, Rao)
- Subsequently RCTs entered western industry
  - Post WW II: (e.g., Deming Medal in Japan, 1951, for QC)
- But insights limited to RCTs with non-conscious units

Transportation of insights to RCTs with conscious units

- Medicine
  - UK in 1946 MRC & Hill -- strep
  - Salk vaccine RCT in US – 1954
  - US FDA and pharma – Paul Meier 1950s
  - Overzealous adherence to ITT to estimate the effect of assignment rather than the effect of assignment to and receipt of drug – not sage
- But no use of Fisher/Neyman insights or notation in non-RCTs
  - E.g., 1964 US Surgeon General on cigarette smoking & lung cancer
  - Epi, econ, social science all used regression with potential outcomes replaced by observed outcome with an indicator w_i for each unit
  - But $Y_i^{obs} = W_i\, Y_i(1) + (1-W_i)\, Y_i\,(0)$ violates the principle of separating the science from what we do to learn about the science

RCM (1974, 1975, 1976, 1977, 1978)

- Potential outcomes define causal estimands in all situations, not just in randomized experiments
  - Neyman disagreed ("too speculative" sans randomization)
- Assignment mechanism *needed* for causal inference:

  Probability for treatment indicator W given science

  Pr(W|X,Y(0),Y(1)), general dependence on Y(0) and Y(1)

  Uncounfounded = Pr(W|X), as in RCTs

  Ignorable = Pr(W|X,Y_{obs}), as in sequential RCTs

  Y_{obs} = observed values of Y(0) and Y(1)
- Bayes = model the science (X,Y(0),Y(1)) in addition to the assignment mechanism, which creates missing and observed potential outcomes; artistic touch is needed

Fundamental problem facing causal inference (Rubin, 1975) It is a missing data problem

- For each *i*, only $Y_i(1)$ or $Y_i(0)$ can be observed

Units	Y(1)	Y(0)	W
1	✓	?	1
.	✓	?	1
.	✓	?	1
.	?	✓	0
.	?	✓	0
N	?	✓	0

- Random assignment of active versus control representative sample of $Y_i(1)$ will be compared to representative sample of $Y_i(0)$ – obvious?
- Have to model the assignment mechanism to draw inferences about the missing potential outcomes

"Mistake" to regress $Y_{obs,i}$ on W_i and X_i

- Lose potential outcomes and key Fisher/Neyman concepts when using observed value notation

  $$Y_{obs,i} = W_i Y_i(1) + (1 - W_i) Y_i(0)$$

- Mixes assignment mechanism and science!
- Suppresses key insights– no missing data!
  - What to estimate? Must be parameters??!
- Standard in biostatistics, economics, epidemiology, everywhere!!
- Even great statisticians and epidemiologists (e.g., Fisher, Cochran, Cornfield, …) confused themselves with the observed value notation

Bad influence on observational studies – no design phase, confused analyses

- Association versus causation -- muddled
- E.g., Case – non-case (control) studies
  - These use sampling mechanisms that are confounded because they, by definition, depend on observed potential outcomes
- Try to "fix" with regression but no principles

"We now consider the distinction between the kinds of inferences that can be supported by observational studies, whether prospective or retrospective, and those that can be supported by experimental studies. That there is a distinction seems undeniable, but its exact nature is elusive." Cornfield, 1959, Principles of Research

Conclusions

- Retain key insights from the past
- Eschew confusion from the past
- Realize RCT ideas are extremely recent
- Update statistical methods, for both design and analysis, to take advantage of modern computing environments
- Encourage mathematical precision, especially in notation and in logical flow
- Precision can have critical consequences for challenging applications, e.g., placebo effects