Chapter 2. A Brief History of the Potential Outcomes Approach to Causal Inference

Donald B. Rubin

Yau Mathematical Sciences Center, Tsinghua University

September 2, 2021

Introduction

The approach to causal inference outlined in the first chapter has important antecedents in the literature.

Potential outcomes in Randomized Controlled Trial – RCT (Neyman, 1923, translated and reprinted in Neyman, 1990), and the introduction of randomization as the "reasoned basis" for inference by Fisher (Fisher 1935, p. 14).

The same statisticians, analyzing both experimental and observational data with the goal of inferring causal effects, would regularly use the notation of potential outcomes in experimental studies, but switch to a notation purely in terms of realized and observed outcomes when discussing observational studies.

Introduction

Before the twentieth century there appears to have been only limited awareness of the concept of the assignment mechanism.

Notably in agricultural experiments, there was no formal statement for a general assignment mechanism, and moreover, apparently not even formal arguments in favor of randomization until Fisher (1925).

The language and reasoning of potential outcomes together with the focus on the assignment mechanism was put front and center in observational study settings in Donald Rubin (1974).

It took another quarter century before it found widespread acceptance as a natural way to define and assess causal effects, irrespective of the setting, experimental or observational.

Potential Outcomes and the Assignment Mechanism Before Neyman

Ideas of potential outcomes, although as yet unlabeled as such, e.g. (Mill, 1973, page 327):

> *"If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the source of his death."*

Mill appears to be considering the two potential outcomes, Y(eat dish) and Y(not eat dish) for the same person.

In this case the observed outcome, Y(eat dish), is "death", and Mill appears to posit that if the alternative potential outcome, Y(not eat dish), is "not death", then one could infer that eating the dish was the source (cause) of the death.

Potential Outcomes and the Assignment Mechanism Before Neyman

Similarly, (Fisher, 1918, p. 214):

> *"If we say, 'This boy has grown tall because he has been well fed,' ... we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter. "*

Here, implicitly, $Y(\text{well fed}) = \text{tall}$ and $Y(\text{not well fed}) = \text{shorter}$, associated with a single unit, a boy.

Despite the insights we may perceive in these quotations, their authors may or may not have intended their words to mean as we choose to interpret them.

For instance, in his argument, Mill goes on to require "constant conjunction" in order to assign causality; that is, for the dish to be the cause of death, this outcome must occur every time it is consumed, by this person, or perhaps by any person.

Potential Outcomes and the Assignment Mechanism Before Neyman

An early tobacco industry argument used a similar notion of causality: not everyone who smokes two or more packs of cigarettes a day gets lung cancer; therefore is smoking does not cause lung cancer.

Jerome Cornfield, who studied smoking and lung cancer also struggled with this: "If cigarettes are carcinogenic, why don't all smokers get lung cancer?" (Cornfield, 1959, p.242) without the benefits of the potential outcomes framework. See also Rubin (2012).

No matter how interpreted, however, we have found no early writer who formally pursued these intuitive insights about potential outcomes defining causal effects; in particular, until Neyman did so in 1923, no one developed a formal notation for the idea of potential outcomes.

Potential Outcomes and the Assignment Mechanism Before Neyman

Nor did anyone discuss the importance of the assignment mechanism, which is necessary for the evaluation of causal effects.

The first such formal mathematical use of the idea of potential outcomes was introduced by Jerzey Neyman (1923), and then only in the context of an urn model for assigning treatments to plots.

The general formal definition of causal effects in terms of potential outcomes, as well as the formal definition of the assignment mechanism, was still another half century away.

Potential Outcomes and the Assignment Mechanism Before Neyman

An early tobacco industry argument used a similar notion of causality: not everyone who smokes two or more packs of cigarettes a day gets lung cancer, therefore smoking does not cause lung cancer.

Jerome Cornfield, who studied smoking and lung cancer also struggled with this: "If cigarettes are carcinogenic, why don't all smokers get lung cancer?" (Cornfield, 1959, p.242) without the benefits of the potential outcomes framework. See also Rubin (2012).

Until Neyman (1923), no one developed a formal notation for the idea of potential outcomes, even in RCTs.

Potential Outcomes and the Assignment Mechanism Before Neyman

Nor did anyone discuss the importance of the assignment mechanism, which is necessary for the evaluation of causal effects.

The first formal mathematical use of the idea of potential outcomes was in Neyman (1923), and then only in the context of an urn model for assigning treatments to plots.

The general formal definition of causal effects in terms of potential outcomes, as well as the formal definition of the assignment mechanism, was still another half century away.

Neyman's (1923) Potential Outcome Notation in Randomized Experiments

Neyman begins with a description of a field experiment with m plots on which v varieties, might be applied.

Neyman introduces what he calls "potential yield" U_{ik}, where i indexes the variety, $i = 1, \ldots, v$, and k indexes the plot, $k = 1, \ldots, m$.

The potential yields are not equal to the actual or observed yield because i indexes all varieties and k indexes all plots, and each plot is exposed to only one variety.

Throughout, the collection of potential outcomes,

$$\mathbf{U} = \{U_{ik} : i = 1, \ldots, v;\ k = 1, \ldots, m\}$$

is considered *a priori* fixed but unknown.

Neyman's (1923) Potential Outcome Notation in Randomized Experiments

The "best estimate" [Neyman's term] of the yield of the ith variety in the field is then

$$a_i = \frac{1}{N} \sum_{k=1}^{m} U_{ik} \ .$$

Neyman calls a_i the "best estimate" because of his concern with the definition of "true yield," something that he struggled with again in Neyman (1935).

As we define potential outcomes, they are the "true" values under SUTVA, not estimates of them.

Neyman then goes on to describe an urn model for determining which variety each plot receives; this model is stochastically identical to the completely randomized experiment with $n = m/v$ plots exposed to each variety.

Neyman's (1923) Potential Outcome Notation in Randomized Experiments

He notes the lack of independence between assignments for different plots implied by this restricted sampling of treatments without replacement (i.e., if plot k receives variety i, then plot l is less likely to receive variety i).

He goes on to note that certain formulas for this situation that have been justified on the basis of independence (i.e., treating the U_{ik} as independent normal random variables given some parameters) need more careful consideration.

Neyman calls a_i the "best estimate" because of his concern with the definition of "true yield".

As we define potential outcomes, they are the "true" values under SUTVA, not estimates of them.

Neyman's (1923) Potential Outcome Notation in Randomized Experiments

The urn model for determining which variety each plot receives is stochastically identical to the completely randomized experiment with $n = m/v$ plots exposed to each variety.

Due to the sampling of treatments without replacement (i.e., if plot k receives variety i, then plot l is less likely to receive variety i) he goes on to note that certain formulas for this situation that have been justified on the basis of independence (i.e., treating the U_{ik} as independent normal random variables given some parameters) need more careful consideration.

Neyman's (1923) Potential Outcome Notation in Randomized Experiments

Let x_i be the sample average of the n plots actually exposed to the i^{th} variety, as opposed to a_i, the average of the potential outcomes over all m plots.

Neyman shows that the expectation of $x_i - x_j$, that is, the average value of $x_i - x_j$ over all assignments that are possible under his urn drawings, is $a_i - a_j$.

Thus, the standard estimate of the effect of variety i versus variety j, the difference in observed means, $x_i - x_j$, is "unbiased" (over repeated randomizations on the m plots) for the causal estimand, $a_i - a_j$, the average effect of variety i versus variety j across all m plots.

Neyman's (1923) Potential Outcome Notation in Randomized Experiments

Neyman's formalism made three contributions:

1. explicit notation for potential outcomes
2. implicit consideration of something like the stability assumption
3. implicit consideration of a model for the assignment of treatments to units that corresponds to the completely randomized experiment.

But as Speed (1990, p. 464) writes "Implicit is not explicit; randomization as a physical act, and later as a basis for analysis, was yet to be introduced by Fisher."

Nevertheless, the explicit provision of mathematical notation for potential outcomes was a great advance, and after Fisher's introduction of randomized experiments in 1925, Neyman's notation quickly became standard for defining average causal effects in randomized experiments.

Neyman's (1923) Potential Outcome Notation in Randomized Experiments

Neyman himself, in hindsight, felt that the mathematical model was an advance:

> *"Neyman has always depreciated the statistical works which he produced in Bydogszcz [which is where Neyman (1923) was done], saying that if there is any merit in them, it is not in the few formulas giving various mathematical expectations but in the construction of a probabilistic model of agricultural trials which, at that time, was a novelty."[Reid, 1982, p. 45].*

Earlier Hints for Physical Randomizing

Experiments is one of the most central concepts in scientific methodology. Experiments used as tool to confirm or disprove theories became important early in physics.

For example, Albert Einstein wrote in a letter to JS Switzer in 1953 that "Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance)." Einstein Archive 61-381.

This is no coincidence, as in physics, precise theoretical predictions can be made and the conditions for keeping (or controlling) disturbing (external) elements constant are relatively good. In areas such as medicine and the social sciences, it is difficult to carry out experiments with the same credibility as in physics.

Earlier Hints for Physical Randomizing

A first problem with experiments on humans is the great natural variation that exists between humans in exposed to either treatment conditions (for example with regard to human capital and genes) and behaviors (which can be shaped by the social environment), which means that what gives an effect to one individual does not have to do it for another.

A second problem is the human tendency to see patterns where there are no patterns (so-called apophenia), which makes it difficult to separate randomly occurring atypical results from the systematic results induced by varying the cause.

Thus, in order to make progress with experiments involving humans, or on biological material, one must take into account this natural variation and be able to assess when an effect exists that differs from the natural variation in outcome. This is where the classic randomized controlled experiment or RCT, comes into play.

Earlier Hints for Physical Randomizing

The notion of the central role of randomization for inference seems to have been "in the air" in the 1920's before it was explicitly introduced by Fisher. For example, "Student" (Gossett, 1923, pp. 281-282) writes:

"If now the plots had been randomly placed ... ", and Fisher and MacKenzie (1923, p. 473) write "Furthermore, if all the plots were undifferentiated, as if the numbers had been mixed up and written down in random order" (See Rubin, 1990, p. 477).

The psychologist and philosopher, Charles Sanders Peirce, appears to have proposed physical randomization already in 1885. Peirce and Jastrow (1885, reprinted in Stigler, 1980, pp. 75-83) used physical randomization to create sequences of blinded binary treatment conditions (heavier versus lighter weights) in a repeated-measures psychological experiment.

Earlier Hints for Physical Randomizing

The purpose of the randomization was to create sequences such that "any possible psychological guessing of what changes the operator [experimenter] was likely to select was avoided." (Stigler, p. 79-80)

Peirce also appears to have anticipated, in the late nineteenth century, Neyman's concept of unbiased estimation when using simple random samples and appears to have even thought of randomization as a physical process to be implemented in practice (Peirce, 1931).

But we can find no suggestion for the physical randomizing of treatments to units as a basis for inference before Fisher (1925)

Fisher's (1925) proposal to randomize treatments to units

An interesting aspect of Neyman's analysis was that he did propose the necessity of physical randomization for credibly assessing causal effects as in Fisher (1925). Although the distinction may seem trivial in hindsight, Neyman did not see it as such:

> *"On one occasion, when someone perceived him as anticipating the English statistician R.A. Fisher in the use of randomization, he objected strenuously: '... I treated* theoretically *an unrestrictedly randomized agricultural experiment and the randomization was considered a prerequisite to probabilistic treatment of the results. This is not the same as the recognition that without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher, and I consider it as one of the most valuable of Fisher's achievements.' (Reid, 1982, page 45)*

Fisher's (1925) proposal to randomize treatments to units

Also,

> *"Owing to the work of R.A. Fisher, "Student" and their followers, it is hardly possible to add anything essential to the present knowledge concerning local experiments ... One of the most important achievements of the English School is their method of planning field experiments known as the method of Randomized Blocks and Latin Squares' (Neyman, 1935, page 109)."*

Thus, independent of Neyman's work, Fisher (1925) proposed the physical randomization of units, and furthermore developed a distinct method of inference based for this special class of assignment mechanisms, that is, randomized experiments.

Fisher's "significance levels" (i.e., p-values), in the current text introduced and discussed in Chapter 5, remain the accepted rigorous standard for the analysis of randomized clinical trials at the start of the twenty-first century, and validate so-called *intent-to-treat* analyses, as discussed in Chapters 5 and 23.

The Observed Outcome Notation in Observational Studies for Causal Effects

Among social scientists, who were using almost exclusively observational data, the work on randomized experiments by Fisher, Neyman and others, received little if any attention.

The tradition was to build statistical models relating the observed value of the outcome variable to covariates and indicator variables for treatment levels, with the causal effects defined in terms of the parameters of these models.

This approach estimated associations, for example, correlations, between observed variables, and then attempted, using various external arguments about temporal ordering of the variables, to infer causation, that is to assess which of these associations might be reflecting a causal mechanism.

The Observed Outcome Notation in Observational Studies for Causal Effects

In particular, the pair of the potential outcomes $(Y_i(1), Y_i(0))$ were replaced by the observed value of Y for unit i.

$$Y_i^{\text{obs}} = Y_i(W_i) = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

In the analysis Y_i^{obs} was then typically regressed as in Yule (1897), on covariates X_i and the indicator for treatment exposure, W_i.

The regression coefficient of W_i in this regression was then interpreted as estimating the causal effect of $W_i = 1$ versus $W_i = 0$.

Under very specific conditions, this approach works (see Chapter 7). This tradition dominated economics, sociology, psychology, education, and other social sciences, as well as the biomedical sciences, such as epidemiology, for most of a century.

Early Uses of Potential Outcomes in Observational Studies in Social Sciences

Although the potential outcome notation did not find widespread adoption in observational studies until recently, in some specific settings researchers used frameworks for causal inference that are similar.

One of the most interesting examples is the use of potential outcomes in the analysis of demand and supply functions specifically, and the analysis of simultaneous equations models in economics in general.

In the 1930s and 1940s economists Tinbergen (1930) and Haavelmo (1944) formulated causal questions in such settings in terms that now appear very modern.

Early Uses of Potential Outcomes in Observational Studies in Social Sciences

Tinbergen writes:

> *"Let π be any imaginable price; and call total demand at this price $n(\pi)$, and total supply $a(\pi)$. Then the actual price p is determined by the equation $a(p) = n(p)$, so that the actual quantity demanded, or supplied, obeys the condition $u = a(p) = n(p)$, where u is this actual quantity. ... The problem of determining demand and supply curves ... may generally be put as follows: Given p and u as functions of time, what are the functions $n(\pi)$ and $a(\pi)$?"(Tinbergen, 1930, translated in Hendry and Morgan, 1994, p. 233).*

This quotation clearly describes the potential outcomes and the specific assignment mechanism corresponding to market clearing, closely following the treatment of such questions in economic theory. Note the clear distinction in notation between the price as an argument in the demand and supply functon ("any imaginable price π") and the actual price p.

Early Uses of Potential Outcomes in Observational Studies in Social Sciences

Similarly, Haavelmo (1934) writes:

> *"If the group of all consumers in society were repeatedly furnished with the total income, or purchasing power r per year, they would, on average or 'normally' spend a total amount $\bar{u}$ for consumption per year, equal to $\bar{u} = \alpha r + \beta$."(Haavelmo, 1943, p. 3)*

Although more ambiguous than the Tinbergen quote, this certainly suggests that Haavelmo viewed laws, or structural equations, in terms of potential outcomes that could have been observed by arranging an experiment.

Early Uses of Potential Outcomes in Observational Studies in Social Sciences

It appears that Haavelmo was directly influenced by Neyman and in fact studied with him for a couple of months at Berkeley:

"I then had the privilege of studying with the world famous statistician Jerzey Neyman for a couple of months in California. ... When I met him for that second talk I had lost most of my illusions regarding my understanding how to do econometrics." (Haavelmo, 1989)

Interestingly, the close connection between the Tinbergen and Haavelmo work and potential outcomes disappeared in later work.

This observed outcome framework for analyzing causal questions dominated economics and other social sciences, and continues to dominate the textbooks in econometrics, with few exceptions until very recently.

Potential Outcomes and the Assignment Mechanism in Observational Studies: Rubin (1974)

Rubin (1974, 1975, 1978) makes two key contributions. He

1. puts the potential outcomes center stage in the analysis of causal effects, irrespective of whether the study is an experimental one or an observational one
2. discusses the assignment mechanism in terms of the potential outcomes.

Rubin starts by *defining* the causal effect at the unit level in terms of the pair of potential outcomes:

> *"... define the causal effect of the E versus C treatment on Y for a particular trial (i.e., a particular unit ...) as follows: Let $y(E)$ be the value of Y measured at t_2 on the unit, given that the unit received the experimental Treatment E initiated at t_1; Let $y(C)$ be the value of Y measured at t_2 on the unit given that the unit recieved 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 particular unit "(Rubin, 1974, p. 639)*

Potential Outcomes and the Assignment Mechanism in Observational Studies: Rubin (1974)

This definition fits perfectly with Neyman's framework for analyzing randomized experiments, but shows that the definition has nothing to do with the assignment mechanism: it applies equally to observational studies as well as to randomized experiments.

Rubin (1975, 1978) then discusses the benefits of randomization in terms of eliminating systematic differences between treated and control units, and formulates the assignment mechanism in general mathematical terms as possibly depending on the potential outcomes (see Chapter 3).