CHAPTER 2

A Brief History of the Potential Outcomes Approach to Causal Inference

2.1 INTRODUCTION

The approach to causal inference outlined in the first chapter has important antecedents in the literature. In this chapter we review some of these antecedents to put the potential outcomes approach in perspective. The two most important early developments, in quick succession in the 1920s, are the introduction of potential outcomes in randomized experiments by Neyman (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).

Once introduced, the basic idea that causal effects are the comparisons of potential outcomes may seem so obvious that one might expect it to be a long-established tenet of scientific thought. Yet, although the seeds of the idea can be traced back at least to the eighteenth century, the formal notation for potential outcomes was not introduced until 1923 by Neyman. Even then, however, the concept of potential outcomes was used exclusively in the context of randomized experiments, not in observational studies. 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 for observational studies. It is only more recently, starting in the early seventies with the work of Donald Rubin (1974), that the language and reasoning of potential outcomes was put front and center in observational study settings, and it took another quarter century before it found widespread acceptance as a natural way to define and assess causal effects, irrespective of the setting.

Moreover, before the twentieth century there appears to have been only limited awareness of the concept of the assignment mechanism. Although by the 1930s randomized experiments were firmly established in some areas of scientific investigation, notably in agricultural experiments, there was no formal statement for a general assignment mechanism and, moreover, not even formal arguments in favor of randomization until Fisher (1925).

https://doi.org/10.1017/CBO9781139025751.003 Published online by Cambridge University Press

2.2 POTENTIAL OUTCOMES AND THE ASSIGNMENT MECHANISM BEFORE NEYMAN

Before the twentieth century we can find seeds of the potential outcomes definition of causal effects among both experimenters and philosophers. For example, one can see some idea of potential outcomes, although as yet unlabeled as such, in discussions by the philosopher and economist Mill (1973, p. 327), who offers:

> 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.

Applying the potential outcomes notation to this quotation, 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.

Similarly, in the early twentieth century, the father of much of modern statistics, Fisher (1918, p. 214), argued:

> 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 again we see a, somewhat implicit, reference to two potential outcomes, Y(well fed) $=$ tall and Y(not well fed) $=$ 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. Curiously, 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, the well-known American epidemiologist 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. 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.

https://doi.org/10.1017/CBO9781139025751.003 Published online by Cambridge University Press

2.3 NEYMAN'S (1923) POTENTIAL OUTCOME NOTATION IN RANDOMIZED EXPERIMENTS

Neyman (in the translated 1990 version) 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. The "best estimate" (Neyman's term) of the yield of the ith variety in the field is the average potential outcomes for that variety over all m plots,

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

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. 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), and 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.

Now, still using Neyman's notation, 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 formalism made three contributions: (*i*) explicit notation for potential outcomes, (*ii*) implicit consideration of something like the stability assumption, and (*iii*) 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 in his introduction to the translation of Neyman (1923): "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. See, for example, Pitman (1937), Welch (1937), McCarthy (1939), Anscombe (1948), Kempthorne (1952, 1955), Brillinger, Jones, and Tukey (1978), Hedges and Lehman (1970, sec. 9.4), and dozens of other places, often

https://doi.org/10.1017/CBO9781139025751.003 Published online by Cambridge University Press

assuming additivity as in Cox (1956, 1958), and even in introductory texts (Freedman, Pisani, and Purves, 1978, pp. 456–458). 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)

2.4 EARLIER HINTS FOR PHYSICAL RANDOMIZING

The notion of the central role of randomization, even if not actual randomized experiments, seems to have been "in the air" in the 1920s 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). Somewhat remarkably, however, an American psychologist and philosopher, Charles Sanders Peirce, appears to have proposed physical randomization decades earlier, although not as a basis for inference, as in Fisher (1925). Specifically, Peirce and Jastrow (1885, reprinted in Stigler, 1980, pp. 75–83) used physical randomization to create sequences of binary treatment conditions (heavier versus lighter weights) in a repeated-measures psychological experiment. 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, pp. 79–80).[1] 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).[2] But we can find no suggestion for the physical randomizing of treatments to units as a basis for inference under Fisher (1925).

2.5 FISHER'S (1925) PROPOSAL TO RANDOMIZE TREATMENTS TO UNITS

An interesting aspect of Neyman's analysis was that, as just mentioned, although he developed his notation to treat data as if they arose from what was later called a completely randomly assigned experiment, he did not take the further step of proposing the necessity of physical randomization for credibly assessing causal effects. It was instead Ronald Fisher, in 1925, who first grasped this. Although the distinction may seem trivial in hindsight, Neyman did not see it as such:

[1] Thanks to Stephen Stigler for noting this, possibly first, use of randomization in formal experiments, in correspondence with the second author.

[2] Thanks to Keith O'Rourke and Stephen Stigler for pointing this out.

https://doi.org/10.1017/CBO9781139025751.003 Published online by Cambridge University Press

> 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, p. 45)

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, p. 109)

Thus, independent of Neyman's work, Fisher (1925) proposed the physical randomization of units and furthermore developed a distinct method of inference for this special class of assignment mechanisms, that is, randomized experiments. The random assignments can be made, for instance, by choosing balls from an urn, as described by Neyman (1923). 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.

2.6 THE OBSERVED OUTCOME NOTATION IN OBSERVATIONAL STUDIES FOR CAUSAL EFFECTS

Despite the almost immediate acceptance of randomized experiments, Fisher's p-values, and Neyman's notation for potential outcomes in agricultural work and mathematical statistics by 1930 within such experiments, these same elements were not used for causal inference in observational studies. Among social scientists, who were using almost exclusively observational data, the work on randomized experiments by Fisher, Neyman, and others, received little or no attention, and researchers continued building models for observed outcomes rather than thinking in terms of potential outcomes. Even among statisticians involved in the analysis of both randomized and non-randomized data for causal effects, the ideas and mathematical language used for causal inference in the setting of randomized experiments were completely excluded from causal inference in the non-randomized settings. The approach in the latter continued to involve building 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, a tradition that appears to originate with Yule (1897).

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. In particular, the pair of the potential outcomes

https://doi.org/10.1017/CBO9781139025751.003 Published online by Cambridge University Press

$(Y_i(1), Y_i(0))$, which in our approach is fundamental for defining causal effects, was replaced by the observed value of Y for unit i, introduced in Section 1.7.

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

The observed outcome Y_i^{obs} was then typically regressed, using ordinary least squares methods, 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 $W_i = 0$. Somewhat remarkably, under very specific conditions, this approach works as outlined in Chapter 7. But in broad generality it does not. 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.

In fact, for the half century following Neyman (1923), statisticians who wrote with great clarity and insight on randomized experiments using the potential outcomes notation did not use it when discussing non-randomized studies for causal effects. For example, contrast the discussion in Cochran and Cox (1956) on experiments with that in Cochran (1965) on observational studies, and the discussion in Cox (1958) on randomized experiments with that in Cox and McCullagh (1982) on Lord's paradox (which we discussed using the potential outcome framework in Chapter 1).

2.7 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. 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 function ("any imaginable price π") and the actual price p.

https://doi.org/10.1017/CBO9781139025751.003 Published online by Cambridge University Press

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, reprinted in Hendry and Morgan, 1994, p. 456)

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.

There are two interesting aspects of the Haavelmo work and the link with potential outcomes. First, it appears that Haavelmo was directly influenced by Neyman (see Hendry and Morgan, 1994, p. 67) 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 of how to do econometrics" (Haavelmo, 1989). Second, the close connection between the Tinbergen and Haavelmo work and potential outcomes disappeared in later work. In the work by Koopmans and others associated with the Cowles Commission (e.g., the papers in Koopmans, 1950, and Hood and Koopmans, 1953), statistical models are formulated for observed outcomes in terms of observed explanatory variables. No distinction is made between variables that Cox describes as "treatments ... potentially causal" and "intrinsic properties of the [units] under study" (Cox, 1992, p. 296) that are characteristics or attributes of the units. 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.

2.8 POTENTIAL OUTCOMES AND THE ASSIGNMENT MECHANISM IN OBSERVATIONAL STUDIES: RUBIN (1974)

Rubin (1974, 1975, 1978) makes two key contributions. First, Rubin (1974) 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. Second, he 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 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 particular unit. (Rubin, 1974, p. 639)

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

https://doi.org/10.1017/CBO9781139025751.003 Published online by Cambridge University Press

assignment mechanism in general mathematical terms as possibly depending on the potential outcomes. Our formal consideration of the assignment mechanism begins in Chapter 3.

NOTES

When one of us (Rubin) was visiting the Department of Statistics at Berkeley in the mid-1970s, where Neyman was Professor Emeritus, he asked Neyman why no one ever used the potential outcomes notation from randomized experiments to define causal effects more generally. This meeting was fifteen years before the (re-)publication of Neyman (1923, 1990). Somewhat remarkably in hindsight, at this meeting, Neyman never mentioned that he invented the notation; his reply to the question as to why it was not used outside experiments was to the effect that defining causal effects in non-randomized settings was too speculative, and in such settings, statisticians should stick with statements concerning descriptions and associations (see Rubin, 2010, p. 42). This fits in with the Neyman quote given in Section 2.5: "without randomization, an experiment has little value irrespective of the subsequent treatment" (Reid, 1982, p. 45). The term "assignment mechanism," and its formal definition, including possible dependence on the potential outcomes, was introduced in Rubin (1975).

For discussions on the intention-to-treat principle, see Davies (1954), Fisher et al. (1990), Meier (1992), Cook and DeMets (2008), Wu and Hamada (2009), Altman (1991), Sheiner and Rubin (1995), and Lui (2011).

https://doi.org/10.1017/CBO9781139025751.003 Published online by Cambridge University Press