Comment

Which Ifs Have Causal Answers

DONALD B. RUBIN*

If

If all the world were apple pie,
And all the sea were ink,
And all the trees were bread and cheese,
What should we have for drink?

—*The Real Mother Goose*

I congratulate my friend Paul Holland on his lucidly clear description of the basic perspective for causal inference referred to as Rubin's model. I have been advocating this general perspective for defining problems of causal inference since Rubin (1974), and with very little modification since Rubin (1978). The one point concerning the definition of causal effects that has continued to evolve in my thinking is the key role of the *stable-unit-treatment-value assumption* (SUTVA, as labeled in Rubin 1980) for deciding which questions are formulated well enough to have causal answers.

Under SUTVA, the model's representation of outcomes is adequate. More explicitly, consider the situation with N units indexed by $u = 1, \ldots, N$; T treatments indexed by $t = 1, \ldots, T$; and outcome variable Y, whose possible values are represented by Y_{tu} ($t = 1, \ldots, T$; $u = 1, \ldots, N$). SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive, and this holds for all $u = 1, \ldots, N$ and all $t = 1, \ldots, T$. SUTVA is violated when, for example, there exist unrepresented versions of treatments (Y_{tu} depends on which version of treatment t was received) or interference between units (Y_{tu} depends on whether unit u' received treatment t or t').

FISHER'S NULL HYPOTHESIS AS A SPECIAL CASE OF SUTVA

SUTVA is automatically satisfied under the Fisher (1935) null hypothesis of absolutely no treatment effects of any kind, H_F, since under H_F the treatment labels are absolutely irrelevant: the values of outcome Y for unit u are exactly the same for all treatments,

$$H_F\colon Y_{tu} = Y_{t'u} \quad \text{for all } u \text{ and all pairs } t, t'. \qquad (1)$$

Thus when Fisher's null hypothesis is tested, which is typically but not necessarily done only in randomized experiments using randomization tests, a particular case of SUTVA is always assumed. If H_F is rejected, all that can be said is that this representation using a very special case of SUTVA is inadequate.

For example, many common language uses of "cause" are essentially statements of a Fisher null hypothesis. Consider

The sun causes the planets to travel in their orbits, (2)

in which the implied treatments are "sun" and "no sun," the unit is the group of planets, and Y is an indicator for their current orbits; or

If John Doe had been born a female,
his life would have been different, (3)

in which the implied treatments are "born as male" and "born as female," John Doe is the only unit, and Y is an indicator for his life as a male. In both Statements (2) and (3), all that is being claimed causally is that Fisher's null hypothesis is to be rejected: no matter how the units would be actually exposed to the relatively vague other treatment ("no sun" and "born as female"), the outcome would not be identical to the outcome under the existing treatment. Neither statement carries with it a precise description of the other treatment (the precise manipulations that would constitute exposure to the other treatment) nor a precise description of an alternative hypothesis under which SUTVA is satisfied but H_F is not.

Thus in the context of statement (3), the claim is simply that if John Doe were born female instead of male, whether because of some hypothetical Y to X chromosome treatment at conception, or massive doses of hormones in utero that would lead to female morphology at birth, or an at-birth sex-change operation, or so forth, John Doe's life would have been different. I accept this as a meaningful causal statement. Since maleness is an attribute of John Doe, however, Holland might not consider Statement (3) to be a meaningful causal claim, and similarly with Statement (2).

In any case, more careful consideration of the implications of SUTVA is required whenever sizes of causal effects are of interest or null hypotheses regarding typical causal effects are to be evaluated, because then actual values under more than one treatment must be contemplated. My formulation of Neyman's null hypothesis of no average causal effect differs somewhat from Holland's because I believe that versions of treatments are implicit in Neyman's discussion yet are absent from Holland's description of it.

NEYMAN'S NULL HYPOTHESIS FORMULATED TO SATISFY SUTVA

Consider the case of two fertilizers A and B, N units, which are plots of land at the time of an experiment, and

* Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138.

Journal of the American Statistical Association
December 1986, Vol. 81, No. 396, Theory and Methods

the outcome Y, which is crop yield on the plots. Each fertilizer has m (m very large) versions $\{A_1, A_2, \ldots, A_m\}$ and $\{B_1, B_2, \ldots, B_m\}$ corresponding to different bags, where one bag is needed to fertilize a plot. The bags are known to vary somewhat in effectiveness, and thus SUTVA only holds exactly when all $2m$ versions of the fertilizers are represented as treatments by $2m$ outcomes (i.e., $t = A_1, \ldots, A_m, B_1, \ldots, B_m$). Using only two treatments, A and B, violates SUTVA because the value of Y for unit u under treatment A (or B) depends on which bag was used.

The causal question of primary interest concerns the typical yields of plots when exposed to fertilizer A relative to their yields when exposed to fertilizer B. A natural way to specify this question is to define the average A versus B differential yield for plot u as

$$\bar{Y}_{Au} - \bar{Y}_{Bu} = \frac{1}{m}\sum_{t=A_1}^{A_m} Y_{tu} - \frac{1}{m}\sum_{t=B_1}^{B_m} Y_{tu}, \qquad (4)$$

and then define the causal estimand as the average A versus B differential yield,

$$\bar{Y}_A - \bar{Y}_B = \frac{1}{N}\sum_{u=1}^{N} (\bar{Y}_{Au} - \bar{Y}_{Bu}). \qquad (5)$$

I believe that this formulation is implicit although certainly not explicit in Neyman (1935). It differs from Holland's interpretation of Neyman in that Holland uses the two-treatment formulation, which violates SUTVA because of "technical errors . . . due solely to the inaccuracy of experimental technique" (Neyman 1935, p. 110). Nonadditivity of treatment effects [$Y_{tu} - Y_{t'u}$ being a function of u as well as (t, t')] arose in Neyman because of "soil errors" due to "variation in fertility of the plots."

Accepting the causal estimand defined in (4) and (5), Neyman's null hypothesis, H_N, is that the average differential effect of fertilizer A versus fertilizer B is 0,

$$H_N\colon \bar{Y}_A - \bar{Y}_B = 0.$$

In contrast, the Fisher null hypothesis is given by (1), where t and $t' = A_1, \ldots, A_m, B_1, \ldots, B_m$.

In an ideally designed randomized experiment in which bags of each type of fertilizer are randomly chosen and randomly applied to plots, it is relatively straightforward to address H_N as well as H_F, although not necessarily using identical statistical tools. But in other cases, H_N is more difficult to address than H_F—simply suppose that fertilizers A and B were randomly assigned to plots, but the bags of A and the bags of B to be used on the plots were carefully selected by the manufacturer of A.

APPLYING SUTVA TO SEX DISCRIMINATION

Careful consideration of SUTVA is especially important for clarifying questions that cannot be addressed by randomized experiments and for deciding precisely in what sense such questions can have causal answers. As a specific example, consider the following statement:

If the females at firm f had been male, their starting salaries would have averaged 20% higher. (6)

I believe Holland would claim that Statement (6) is causally meaningless because "femaleness" is an attribute. I too believe that Statement (6) is causally meaningless, but for a possibly different reason: the statement, by itself, is too vague to have a clear formulation satisfying SUTVA and thus is too vague to admit a clear causal answer. What are the units, treatments, and outcomes such that SUTVA is satisfied? I am not at all sure how to define anything except Y, which clearly involves starting salary.

One range of possibilities for making (6) more precise is generated by considering the units to be the female employees at entry and the treatments to be "female," which is well defined since the units are females, and "male," which has many possible versions ranging from some hypothetical "at conception X to Y chromosome treatment" to replacing an "F" with an "M" on a job application form. Certainly these different versions of the treatment "male" could lead to vastly different outcomes, and so SUTVA is totally implausible without agreement on which version of the treatment "maleness" is under study or agreement on a way to average over some collection of such versions.

Another possibility, and one more closely tied to potential real-world manipulations, is to consider the firm to be the unit, multivariate Y to be the starting salaries of the female employees, and the treatments to be "current hiring practices" and "hiring practices as would take place under court supervision." Or perhaps the job slots in the firm are the units, Y is the starting salary in each job slot, and applicants are the treatments: type A treatments are the female applicants and type B treatments are the male applicants, using the notation used for Neyman's null hypothesis. For related discussion of this perspective, see Pratt and Schlaifer (1984), especially the rejoinder to the discussion by Rosenbaum and Rubin (1984).

In any case, the crucial point with Statement (6) is that we are not ready to estimate, test, or even logically discuss *causal effects* until units, treatments, and outcomes have been defined in such a way that SUTVA is plausible.

NO CAUSATION WITHOUT MANIPULATION?

Since statisticians who study causal effects usually do so for the purpose of drawing inferences about the effects of actual manipulations to which some group of units have been or might be exposed, the motto "no causation without manipulation" is a critical guideline for clear thinking in empirical studies for causal effects. Thinking about actual manipulations forces an initial definition of units and treatments and thereby increases the likelihood of a formulation in which SUTVA is plausible. Such clarity is essential, yet commonly absent, in policy-oriented studies in which decisions to implement real-world manipulations can result from the statistician's causal inferences.

ADDITIONAL REFERENCES

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Pratt, J. W., and Schlaifer, R. (1984), "On the Nature and Discovery of Structure" (with discussion), *Journal of the American Statistical Association*, 79, 9–33.