Chapter 1

Donald B. Rubin

Yau Mathematical Sciences Center, Tsinghua University

September 2, 2021

Introduction

Three essential concepts in the book and lectures

1. *Potential outcomes*, each corresponding to one of the levels of a *treatment* or *manipulation*, following the *dictum* "no causation without manipulation" (Rubin, 1975, p. 238).
   - Each of these potential outcomes is *a priori* observable, in the sense that it could be observed if the unit were to receive the corresponding treatment level.
   - But at most one potential outcome can be observed.
2. The necessity, when drawing causal inferences, to observe *multiple units*, and the utility of the related *stability* assumption, which is used to exploit the presence of multiple units.
3. The central role of the *assignment mechanism*.

Potential outcomes

In everyday life, causal language is widely used in an informal way. One might say:

"my headache went away because I took an aspirin," or

"she got a good job last year because she went to college," or

"she has long hair because she is a girl."

As such, these observations generally involve informal statistical analyses, drawing conclusions from associations between measurements of different quantities that vary from individual to individual, commonly called *variables* or *random variables*.

Potential outcomes

Nevertheless, statistical theory has been relatively silent on questions of causality. Many, especially older, textbooks avoid any mention of the term other than in settings of randomized experiments.

Some mention it mainly to stress that correlation or association is not the same as causation, and some even caution their readers to avoid using causal language in statistics.

Nevertheless, for many users of statistical methods, causal statements are exactly what they seek.

Potential outcomes

The fundamental notion underlying the book is that causality is tied to an *action* (or manipulation, treatment, or intervention), applied to a *unit*.

A unit here can be a physical object, a firm, an individual person, or collection of objects or persons, such as a classroom or a market, at a particular point in time.

Thus, the same physical object or person at a different time is a different unit.

From this perspective, a causal statement presumes that, although a unit was (at a particular point in time) subject to, or exposed to, a particular action, treatment, or regime, the same unit could have been exposed to an alternative action, treatment, or regime (at the same point in time).

Potential outcomes

For instance, when deciding to take an aspirin to relieve your headache, you could also have chosen not to take the aspirin, or you could have chosen to take an alternative medicine.

The book primarily considers settings with two actions, although many of the extensions to multi-valued treatments are conceptually straightforward.

Often one of these actions corresponds to a more active treatment (*e.g.*, taking an aspirin) in contrast to a more passive action (*e.g.*, not taking the aspirin).

In such cases we sometimes refer to the first action as the *active treatment* as opposed to the *control treatment*, but these are merely labels.

Given a unit and a set of actions, we associate each action/unit pair with a *potential outcome*.

Potential outcomes

We refer to these outcomes as potential outcomes because only one will ultimately be realized and therefore possibly observed: the potential outcome corresponding to the action actually taken. *Ex post*, the other potential outcomes cannot be observed.

The causal effect of one action or treatment relative to another involves the comparison of these potential outcomes, one realized (and perhaps, though not necessarily, observed), and the others not realized and therefore not observable.

Any treatment must occur temporally before the observation of any associated potential outcome.

Potential outcomes

Although the above argument may appear obvious, its force is revealed by its ability to clarify otherwise murky concepts, as can be demonstrated by considering the three examples of informal "because" statements.

1. At the time that I took the aspirin, I could have followed the alternate course of not taking an aspirin.
2. At the time she went to college she could have decided not to go to college.
3. "she has long hair because she is a girl."

Potential outcomes

In (1) different outcome might have resulted, and the "because" statement is causal in the perspective taken in this book.

In (2) the implied causal statement compares the quality of the job she actually had, say 10 years after her decision, to the quality of the job she would have had, had she not gone to college. The alternative treatment is somewhat murky: had she not enrolled in college, would she have enrolled in the military, or would she have joined an artist's colony? As a result, the potential outcome under the alternative action is not as well defined as in the first example.

The implicit treatment in (3) is being a girl, and the implicit alternative is being a boy, but there is no action articulated that would have made her a boy and allowed us to observe the alternate potential outcome of hair length as a boy.

Potential outcomes

As stated, however, there is no clear action described that would have allowed us to observe the unit exposed to the alternative treatment. Hence, in our approach, this "because" statement is ill–defined as a causal statement.

It may seem restrictive to exclude from consideration causal questions regarding e.g. gender and ethnic discrimination.

However, the reason to do so in this framework is that without further explication of the intervention being considered, the causal question is not well defined.

Potential outcomes

In practice, the distinction between well-defined and poorly-defined causal statements is one of degree.

The important point is, however, that causal statements become more clearly defined by more precisely articulating the intervention that would have made the alternative potential outcome the realized one.

Definition of Causal Effects

Consider one single unit, I, at a particular point in time, contemplating whether or not to take an aspirin for my headache.

If I take the aspirin, my headache may be gone or it may remain, say, an hour later; we denote this outcome, which can be either "Headache" or "No Headache," by $Y(\text{Aspirin})$.

Similarly, if I do not take the aspirin, my headache may remain an hour later, or it may not; we denote this potential outcome by $Y(\text{No Aspirin})$, which also can be either "Headache," or "No Headache."

There are therefore two potential outcomes, $Y(\text{Aspirin})$ and $Y(\text{No Aspirin})$.

The causal effect of aspirin involves the comparison of these two potential outcomes.

Definition of Causal Effects

Each potential outcome can take two values , and that is why the unit-level causal effect involves one of four possibilities

1. Headache gone only with aspirin:
   $Y(\text{Aspirin}) = \text{No Headache}$, $Y(\text{No Aspirin}) = \text{Headache}$;
2. No effect of aspirin, with a headache in both cases: $Y(\text{Aspirin}) = \text{Headache}$, $Y(\text{No Aspirin}) = \text{Headache}$;
3. No effect of aspirin, with the headache gone in both cases:
   $Y(\text{Aspirin}) = \text{No Headache}$, $Y(\text{No Aspirin}) = \text{No Headache}$;
4. Headache gone only without aspirin:
   $Y(\text{Aspirin}) = \text{Headache}$, $Y(\text{No Aspirin}) = \text{No Headache}$.

There is a zero causal effect of taking aspirin in the second and third possibilities. In the other two cases the aspirin has a non-zero causal effect.

Definition of Causal Effects

There are two important aspects of this definition of a causal effect:

1. the definition of the causal effect depends on the potential outcomes.
2. the causal effect is the comparison of potential outcomes, for the same unit, at the same moment in time post-treatment.

Thus, the causal effect is *not* defined in terms of comparisons of outcomes at different times, as in a before-and-after comparison.

"The fundamental problem of causal inference" (Holland, 1986, p.947, paraphrasing Rubin, 1975, p.38) is therefore the problem that at most one of the potential outcomes can be realized and thus observed.

Definition of Causal Effects

In general, therefore, even though the unit–level causal effect (the comparison of the two potential outcomes) may be well defined, by *definition* we cannot learn its value from just the single realized potential outcome. Table 1.2 illustrates this concept for the aspirin example, assuming the action taken was that you took the aspirin.

Table 1.2: Example of Potential Outcomes, Causal Effect, Actual Treatment and Observed Outcome with One Unit

Unit	Unknown Potential Outcomes Y(Aspirin)	Y(No Aspirin)	Unknown Causal Effect Y (A)- Y(NA)	Known Actual Treatment	Known Observed Outcome
You	No Headache	Headache	Improv. due to Asp.	Aspirin	No Headache

Definition of Causal Effects

For the *estimation* of causal effects we need to compare *observed* outcomes, that is, observed realizations of potential outcomes, and because there is only one realized potential outcome per unit, we will need to consider multiple units.

For example, a before-and-after comparison of the same physical object involves distinct units in this framework, and also the comparison of two different physical objects at the same time involves distinct units.

Such comparisons are critical for *estimating* causal effects, but they *do not* define causal effects.

For estimation it is critical to know about, or make assumptions about, the reason why certain potential outcomes were realized and not others. That is, we will need to think about the *assignment mechanism*.

Learning About Causal Effects: Multiple Units

Because with a single unit, we can at most observe a single potential outcome, we must rely on multiple units to make causal inferences from observed data.

There is sometimes a tendency to view the same physical object at different times as the same unit. We view this as a fundamental mistake. "myself at different times" are not the same unit in our approach to causality.

It is often reasonable to assume that time makes little difference for inanimate objects but this assumption is typically less reasonable with human subjects, and it is never correct to confuse assumptions (e.g., about similarities between different units), with definitions (e.g., of a unit, or of a causal effect).

Learning About Causal Effects: Multiple Units

As an alternative to observing the same physical object repeatedly, one might observe different physical objects at approximately the same time.

It is more obvious here that "you" and "I" at the same point in time are different units.

Your headache status after taking an aspirin can obviously differ from what my headache status would have been had I taken an aspirin. I may be more or less sensitive to aspirin, or I may have started with a more or less severe headache.

This type of comparison, often involving many different individuals, is widely used in informal assessments of causal effects, but it is also the basis for many formal studies of causal effects in the social and bio-medical sciences.

Learning About Causal Effects: Multiple Units

By itself, however, the presence of multiple units does not solve the problem of causal inference.

Consider the aspirin example with two units—You and I—and two possible treatments for each unit—aspirin or no aspirin. For simplicity, assume that the two available aspirin tablets are equally effective. There are now a total of four treatment levels and four potential outcomes for each of us.

For "I" these four potential outcomes are the state of my headache (i) if neither of us takes an aspirin, (ii) if I take an aspirin and you do not, (iii) if you take an aspirin and I do not, and (iv) if both of us take an aspirin. "You," of course, have your corresponding set of four potential outcomes.

Learning About Causal Effects: Multiple Units

We can still only *observe* at most one of these four potential outcomes for each unit, namely the one realized corresponding to whether you and I took, or did not take, an aspirin.

In this situation, there are six different comparisons defining causal effects for each of us, depending on which two of the four potential outcomes for each unit are conceptually compared

For example, we can compare the status of my headache if we both take aspirin with the status of my headache if neither of us take an aspirin, or we can compare the status of my headache if only you take an aspirin to the status of my headache my headache if we both do.

Learning About Causal Effects: Multiple Units

We typically make the assumption that whether you take an aspirin does not affect my headache status, and it is important to understand the force of such an assumption.

One should not lose sight of the fact that it is an assumption, often a strong and controversial one, not a fact, and therefore may be false.

Consider a setting where I take aspirin, and I will have a headache if you do not take an aspirin, whereas I will not have a headache if you do take an aspirin: we are in the same room, and unless you take an aspirin to ease your own headache, your incessant complaining will maintain my headache!

Such interactions or "spillover" effects are an important feature of many educational programs, and often motivate changing the unit of analysis from individual children to schools or other groups of individuals, which also changes the treatment being studied.

The Stable Unit Treatment Value Assumption (SUTVA)

The *Stable Unit Treatment Value Assumption* (Rubin, 1980) incorporates both the idea that units do not interfere with one another, and the concept that for each unit there is only a single version of each treatment level (ruling out, in this case, that a particular individual could take aspirin tablets of varying efficacy):

Assumption

(SUTVA)
The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

These two elements of the stability assumption enables us to exploit the presence of multiple units for estimating causal effects.

The Stable Unit Treatment Value Assumption (SUTVA)

SUTVA can be referred to as *exclusion restrictions*: assumptions that rely on external, substantive, information to rule out the existence of a causal effect of a particular treatment relative to an alternative.

For instance, in the aspirin example, we exclude the possibility that your taking or not taking aspirin has any effect on my headache in order to learn about the causal effect.

Similarly, we exclude the possibility that the aspirin tablets are of different strengths.

Note, that these restriction, and other discussed later, are not directly informed by observations—they are assumptions. That is, they rely on previously acquired knowledge of the subject matter for their justification.

Causal inference is generally impossible without such assumptions, and thus it is critical to be explicit about their content and their justifications

SUTVA: No Interference

Researchers have long been aware of the importance of this concept. For example, when studying the effect of different types of fertilizers in agricultural experiments on plot yields, traditionally researchers have taken care to separate plots using "guard rows," unfertilized strips of land between fertilized areas.

These guard rows make SUTVA more credible; without them we might suspect that the fertilizer applied to one plot affected the yields in contiguous plots.

In our headache example, in order to address the no-interference assumption, one has to argue, on the basis of a prior knowledge of medicine and physiology, that someone else taking an aspirin in a different location cannot have an effect on my headache.

SUTVA: No Interference

There exist settings, moreover, in which the no-interference part of SUTVA will not hold.

A classic example of interference between units arises in settings with immunizations against infectious diseases. The causal effect of your immunization versus no immunization will surely depend on the immunication of others.

Other examples is with regard to supply and demand effect in a market.

In an extreme example, the effect on your future earnings of going to a graduate program in statistics would surely be very different if everybody your age also went to a graduate program in statistics.

SUTVA: No Interference

Economists refer to this concept as a *general equilibrium* effect, in contrast to a *partial equilibrium* effect, which is the effect on your earnings of a statistics graduate degree under the *ceteris paribus* assumption that "everything else" stayed equal.

In all the above cases, sometimes a more restrictive form of SUTVA can be considered by defining the unit to be the community within which individuals interact, e.g., schools in educational settings, or specifically limiting the number of units assigned to a particular treatment.

SUTVA: No Hidden Variations of Treatments

Fundamentally, the second component of SUTVA is again an exclusion restriction.

The requirement is that the label of the aspirin tablet, or the nature of the administration of the treatment, cannot alter the potential outcome for any unit.

This assumption does *not* require that all forms of each level of the treatment are identical across all units, but only that unit i exposed to treatment level w specifies a well-defined potential outcome, $Y_i(w)$, for all i and w.

Basically, the assumption in that $Y_i(w)$ is a well-defined function of i and w: given i and w, there is a specified value of $Y_i(w)$ in its range.

SUTVA: No Hidden Variations of Treatments

One strategy to make SUTVA more plausible relies on re-defining the represented treatment levels to comprise a larger set of treatments, e.g., $\mathrm{Aspirin}-$(=weak), $\mathrm{Aspirin}+$(=strong) and no-aspirin instead of only $\mathrm{Aspirin}$ and no-aspirin. Then we re-define the treatment as taking a randomly selected level (either none, or $\mathrm{Aspirin}-$ or $\mathrm{Aspirin}+$).

A second strategy involves coarsening the outcome; for example, SUTVA may be more plausible if the outcome is defined as dead or alive rather than for a detailed measurement of health status.

The point is that SUTVA implies that the potential outcomes for each unit and each treatment are well-defined functions (possibly with stochastic images) of the unit index and the treatment.

Alternatives to SUTVA

Throughout most of the book, SUTVA is maintained.

In some cases, however, specific information may suggest that alternative assumptions are more appropriate.

For example, in some early AIDS drug trial settings, where patients where (i) mixing placebos and active drugs and (ii) "testing" if given a sugar-pill or not. In (ii), if found given a sugar-pill then most likely $Y \neq Y(0)$.

Given this kind of knowledge, it is clearly no longer appropriate to assert the no-interference element of SUTVA.

With specific information to model how treatments are received across patients in a study, one can make alternative assumptions that allow some inference.

Alternatives to SUTVA

For example, SUTVA may be more appropriate using subgroups of people as units in such AIDS drug trials.

Similarly, in educational settings, SUTVA may be more plausible with classrooms or schools as the units of analysis than with students as the units of analysis.

In many economic examples, interactions between units are often modeled through assumptions on market structure, again avoiding the no-interference element of SUTVA.

Consequently, SUTVA is only one candidate exclusion restriction for modeling the potentially complex interactions between units and the entire set of treatment levels in a particular experiment.

In many settings, however, it appears that SUTVA is the leading choice.

SUTVA with two units: you and I

Table 1.3: Example of Potential Outcomes and Causal Effects under SUTVA with Two Units

Unit	Unknown			Known	
	Potential Outcomes		Causal Effect	Actual	Observed
	Y(Aspirin)	Y(No Aspirin)	Y (A)- Y(NA)	Treatment W_i	Outcome Y_i^{obs}
You	No Headache	Headache	Improv. due to Asp.	Aspirin	No Headache
Me	No Headache	No Headache	None	No Aspirin	No Headache

Potential Outcomes with Two Units and Interference Between Units (Part (b) of SUTVA Does Not Hold)

Potential Outcomes and Values in Example

You take: I take:	Asp Asp	Not Asp	Asp Not	Not Not
Unit				
1=you 2=me	Y_1([Asp,Asp])=0 Y_2([Asp,Asp])=0	Y_1([Not,Asp])=100 Y_2([Not,Asp])=0	Y_1([Asp,Not])=50 Y_2([Asp,Not])=100	Y_1([Not,Not])=100 Y_2([Not, Not])=100

The Assignment Mechanism – An Introduction

The previous notation extends readily to many units, where we index the units in the population of size N by i, taking on values $1, \ldots, N$, and let the treatment indicator W_i take on the values 0 (the control treatment, e.g., no aspirin) and 1 (the active treatment, e.g., aspirin).

For unit i, with $i \in \{1, \ldots, N\}$, let Y_i^{obs} denote this realized (and possibly observed) outcome:

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

For each unit we also have one missing potential outcome, for unit i denoted by Y_i^{mis}:

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

The Assignment Mechanism – An Introduction

This information alone, still, does not allow us to infer the causal effect of taking an aspirin on headaches.

Suppose, in the two-person headache example, that the person who chose not to take the aspirin did so because he only had a minor headache. Suppose then that an hour later both headaches have faded.

When comparing these two observed potential outcomes, we might conclude that the aspirin had no effect, whereas in fact it may have been the cause of easing the more serious headache.

The key piece of information that we lack is *how* each individual came to receive the treatment level actually received: in our language of causation, the *assignment mechanism*.

The Assignment Mechanism – An Introduction

Because causal effects are defined by comparing potential outcomes, they are well defined irrespective of the actions actually taken.

But, because we observe at most half of all potential outcomes, and none of the unit-level causal effects, there is an inferential problem. In this sense causal inference is a *missing data problem*.

A key role is therefore played by the missing data mechanism, or the assignment mechanism. How is it determined which units get which treatments, or equivalently, which potential outcomes are realized and which are not?

This mechanism is, in fact, so crucial to the problem of causal inference that Parts II through V of the book and these lectures are organized by varying assumptions concerning this mechanism.

The Assignment Mechanism – An Introduction

An example involves four patients, and two medical procedures labeled 0 (Drug) and 1 (Surgery).

Table 1.4: Medical Example with Two Treatments, Four Units and SUTVA: Surgery (S) and Drug Treatment (D)

Unit	Potential Outcomes $Y_i(0)$	$Y_i(1)$	Causal Effect $Y_i(1) - Y_i(0)$
Patient #1	1	7	6
Patient #2	6	5	-1
Patient #3	1	5	4
Patient #4	8	7	-1
Average	4	6	2

The table displays each patient's potential outcomes, in terms of years of post-treatment survival, under each treatment

The Assignment Mechanism – An Introduction

Suppose that the doctor, through expertise, knows enough about these potential outcomes, and so assigns each patient to the treatment that is more beneficial to that patient.

In this scenario, Patients 1 and 3 will receive surgery, and Patients 2 and 4 will receive the drug treatment.

The *observed* treatments and outcomes will then be as displayed in Table 1.5

The Assignment Mechanism – An Introduction

Table 1.5: Ideal Medical Practice: Patients Assigned to the Individually Optimal Treatment; Example from Table 1.4

Unit i	Treatment W_i	Observed Outcome Y_i^{obs}
Patient #1	1	7
Patient #2	0	6
Patient #3	1	5
Patient #4	0	8

Here the average observed outcome with surgery is one year less than the average observed outcome with the drug treatment.

The Assignment Mechanism – An Introduction

Thus, a casual observer might be led to believe that, on average, the drug treatment is superior to surgery.

Based on this example, we can see that we cannot simply look at the observed values of potential outcomes under different treatments, that is, $\{Y_i^{\text{obs}} | i : \text{ s.t. } W_i = 0\}$ and $\{Y_i^{\text{obs}} | i : \text{ s.t. } W_i = 1\}$, and reach valid causal conclusions irrespective of the assignment mechanism.

In order to draw valid causal inferences, we must consider why some units received one treatment rather than another.

Attributes, Pretreatment Variables, or Covariates

Often the presence of unit-specific background attributes (i.e. pre-treatment variables, or covariates) can assist in making these predictions: these are denoted here by the K-component row vector X_i for unit i.

In our headache example, X_i could include the intensity of the headache before making the decision to take aspirin or not for unit i.

In the job training example, X_i could include age, previous educational achievement, family and socio-economic status, or pre–training earnings.

As these examples illustrate, sometimes a covariate (e.g., pre-training earnings) differs from the potential outcome (post-training earnings) solely in the timing of measurement, in which case the covariates can be highly predictive of the potential outcomes.

Attributes, Pretreatment Variables, or Covariates

Often the presence of unit-specific background attributes (i.e. pre-treatment variables, or covariates) can assist in making these predictions: denoted here by the K-component row vector X_i for unit i.

In our headache example, X_i could include the intensity of the headache before making the decision to take aspirin or not.

In the job training example, X_i could include age, previous educational achievement, family and socio-economic status, or pre–training earnings.

As these examples illustrate, sometimes a covariate (e.g., pre-training earnings) differs from the potential outcome (post-training earnings) solely in the timing of measurement, in which case the covariates can be highly predictive of the potential outcomes.

Attributes, Pretreatment Variables, or Covariates

Often the presence of unit-specific background attributes (i.e. pre-treatment variables or covariates) can assist in making these predictions: denoted here by the K-component row vector X_i for unit i.

In our headache example, X_i could include the intensity of the headache before making the decision to take aspirin or not.

In the job training example, X_i may include age, previous educational achievement, family and socio-economic status, or pre–training earnings.

As these examples illustrate, sometimes a covariate (e.g., pre-training earnings) differs from the potential outcome (post-training earnings) solely in the timing of measurement, in which case the covariates can be highly predictive of the potential outcomes.

Attributes, Pretreatment Variables, or Covariates

The key characteristic of these covariates is that they are *a priori* known to be unaffected by the treatment assignment, i.e. "pre-treatment" variables.

The information available in these covariates can be used in three ways.

1. covariates commonly serve to make estimates more precise by explaining some of the variation in outcomes.
2. for substantive reasons, the researcher may be interested in the typical (e.g., average) causal effect of the treatment on subgroups (as defined by a covariate) in the population of interest.
3. their effect on the assignment mechanism is generally important.

Attributes, Pretreatment Variables, or Covariates

With regard to (3) consider the evaluation of the effects of a job training program.

Young unemployed individuals may be more interested in training programs aimed at acquiring new skills. As a result, those taking the active treatment may differ in the values of their background characteristics from those taking the control treatment.

At the same time, these characteristics may be associated with the potential outcomes.

As a result, assumptions about the assignment mechanism, and in particular its possible freedom from dependence on potential outcomes are typically more plausible within subpopulations that are homogenous with respect to some covariates, *i.e.*, conditionally given the covariates, than unconditionally.

Potential Outcomes and Lord's Paradox

> *"A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded."(Lord, 1967, p. 304)*

The results of the hypothetical study includes the finding that for the males and females the average weight is identical at the end of the school year to what it was at the beginning.

The only difference is that the females started and ended the year lighter on average than the males.

The paradox is generated by considering the contradictory conclusions of two statisticians asked to comment on the data.

Potential Outcomes and Lord's Paradox

From Lord's description of the problem, the object of interest, i.e. the *estimand*, is the difference between the causal effect of the university diet on males and the causal effect of the university diet on females. That is, a "differential" causal effect.

Statistician 1 concludes that

> *"...as far as these data are concerned, there is no evidence of any interesting effect of diet (or of anything else) on student weight. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change. "(Lord, 1967, p. 305)*

Statistician 2 looks at the data in a more "sophisticated" way by conditioning on the initial weight in September.

Potential Outcomes and Lord's Paradox

He notices that conditional on their initial weight in September the males tended to gain weight on average and that the females tended to lose weight on average (i.e. covariance adjustment or regression adjustment described in Chapter 7).

He also notices that this result is true no matter what the value of initial weight he focuses on.

His conclusion, therefore, is that after "controlling for" initial weight, the diet has a differential positive effect on males relative to females because for males and females with the same initial weight, on average the males gain more than the females.

Potential Outcomes and Lord's Paradox

Who's right? Statistician 1 or Statistician 2?

Notice the focus of both statisticians on before-after or gain scores and recall that such gain scores are not causal effects because they do not compare potential outcomes at the same time post-treatment; rather, they compare changes over time.

If both statisticians confined their comments to *describing* the data, both would be correct, but for causal inference, both are wrong because these data cannot support any conclusions about the causal effect of the diet without making some very strong, and arguably implausible, assumptions.

Potential Outcomes and Lord's Paradox

The units are obviously the students, and the time of application of active treatment (the university diet) is clearly September and the time of the recording of the outcome Y is clearly June.

Notice that Lord's statement of the problem uses the already criticized notation with a treatment indicator and the observed variable, Y_i^{obs}, rather than the potential outcome notation being advocated.

The potential outcomes are June weight under the university diet $Y_i(1)$ and under the "control" diet $Y_i(0)$. The covariates are sex of students, male vs. female, and September weight, both of which are pre-treatment.

But the assignment mechanism has assigned everyone to the new treatment!

Potential Outcomes and Lord's Paradox

Thus $Y_i(0)$ is never observed. Hence, there is no purely empirical basis on which to compare the effects, either raw or differential, of the university diet with the control diet.

By making the problem complicated with the introduction of the covariates "male/female" and "initial weight", Lord has created partial confusion.

The "paradox" is resolved through the explicit use of potential outcomes. Either answer could be correct for causal inference depending on what we are willing to assume about the (never-observed) potential outcome under the control diet.

Causal Estimands

We start with a population of units, indexed by $i = 1, \ldots, N$, which is our focus.

Each unit in this population can be exposed to one of a set of treatments. In the most general case, let $\mathbb{W}_i$ denote the set of treatments to which unit i can be exposed. In most cases, this set will be identical for all units.

The set $\mathbb{W}_i$ consists of possible treatments for each unit, (*e.g.*, taking or not taking a drug),

$$\mathbb{W}_i = \Big\{0, 1\Big\}, \text{ for all } i = 1, \ldots, N$$

Causal Estimands

For each unit i, and for each treatment in the set of treatments, $\mathbb{W} = \{0, 1\}$, there are corresponding potential outcome, $Y_i(0)$ and $Y_i(1)$.

Comparisons of $Y_i(1)$ and $Y_i(0)$ are *unit-level causal effects*. Often these are simple differences,

$$Y_i(1) - Y_i(0), \qquad \text{or ratios } Y_i(1)/Y_i(0),$$

but in general the comparisons can take different forms.

A leading example of summarizing these unit-level causal effects – what in general is refereed to as a *causal estimand* — is the average treatment effect estimand:

$$\tau_{\text{fs}} = \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(1) - Y_i(0) \Big),$$

where the subscript "fs" indicates that we average over the finite sample.

Causal Estimands

We can generalize this example in a number of ways.

First (i), we can average over subpopulations rather than over the full population.

Second (ii), we should focus on more general functions of potential outcomes.

Regarding (i), we can consider subpopulations defined by *covariates* or by *treatment status*.

As an example with covariates, consider estimating an average effect of a new drug only for females:

$$\tau_{\text{fs}}(f) = \frac{1}{N(f)} \sum_{i:X_i=f} \Big(Y_i(1) - Y_i(0) \Big).$$

Causal Estimands

Here $X_i \in \{f, m\}$ is an indicator for being female, and $N(f) = \sum_{i=1}^{N} \mathbf{1}_{X_i=f}$ is the number of females in the finite population, where $\mathbf{1}_A$ is the indicator function for the event A, equal to 1 if A is true and zero otherwise.

As an example of an effect conditional on treatment status the average effect of the treatment for those who were exposed is highly policy relevant estimand

$$\tau_{\text{fs,t}} = \frac{1}{N_t} \sum_{i:W_i=1} \Big(Y_i(1) - Y_i(0) \Big),$$

where N_t is the number of units exposed to the active treatment.

An example of (ii) is an interest in the median (over the entire population or over a subpopulation) of $Y_i(1)$ versus the median of $Y_i(0)$.

Causal Estimands

One may also be interested in the median of the difference $Y_i(1) - Y_i(0)$, which generally differs from the difference in medians.

In all cases with $\mathbb{W}_i = \{0, 1\}$, we can write the causal estimand as a row-exchangeable function of all potential outcomes for all units, all treatment assignments, and pretreatment variables:

$$\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{W}).$$

In this expression $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ are the N-component column vectors of potential outcomes with ith elements equal to $Y_i(0)$ and $Y_i(1)$, $\mathbf{W}$ is the N-component column vector of treatment assignments, with ith element equal to W_i, and $\mathbf{X}$ is the $N \times K$ matrix of covariates with ith row equal to X_i.

Causal Estimands

Not all such functions necessarily have a causal interpretation, but the converse is true:

all the causal estimands considered in this book can be written in this form, and

all such estimands are comparisons of $Y_i(0)$ and $Y_i(1)$ for all units in a common set of units whose definition, as the previous examples illustrate, may depend on $\mathbf{Y}(0)$, $\mathbf{Y}(1)$, $\mathbf{X}$, and $\mathbf{W}$.

Samples, Populations, and Super-populations

The focus is on the finite set of units for which we observe covariates, treatments, and realized outcomes is the set of units we are interested in, and we will refer to this as the *population*.

Uncertainty about causal estimands in this setting arises from the randomness of the assignment mechanism.

For some parts it we assume the set of units as drawn randomly from a larger population.

In that case we typically take the population that the units were drawn from as infinite (i.e. a super-population).

Uncertainty about causal estimands in this setting arises from the random sampling and from the assignment mechanism.