BUSS975 Causal Inference in Financial Research

Ji-Woong Chung
chung_jiwoong@korea.ac.kr
Korea University Business School

Chapter 1

Introduction

1.1 Descriptive, Predictive, and Causal Questions

When conducting empirical research in finance, it is useful to distinguish three types of questions we might ask about data:

- **Descriptive questions:** These aim to describe how things are or were, focusing on statistical properties and relationships. For example, "Has industry concentration increased over time?" or "What is the typical cost of financial distress for firms?" Such questions summarize data without making guesses about new or unseen outcomes.

- **Predictive questions:** These aim to predict an unknown outcome without necessarily changing anything in the system. For instance, "What will the GDP be next quarter?" or "Can we predict a website visitor's income from their browsing behavior?" Here we use historical patterns to guess future or otherwise unobserved values, but we are not explicitly asking about causal effects.

- **Causal questions:** These ask what happens to Y when we change X, all else being equal. In other words, they concern cause-and-effect relationships. For example, "Do increases in financial disclosure requirements cause changes in firm value?" or "What is the effect of raising the minimum wage on employment levels?" Causal questions

involve imagining some intervention or change in X and asking how it would alter Y in a ceteris paribus world.

Only the third type – causal questions – requires us to grapple with the idea of what would happen under different scenarios for the same entity. Descriptive and predictive analyses can often be done with straightforward statistics or machine learning, but causal analysis demands something more: we must infer an outcome under a hypothetical change, which is fundamentally a missing data problem (since we don't actually see the hypothetical scenario directly). As we'll see, this is where concepts like counterfactuals, potential outcomes, and careful research design come into play.

1.2 What is Causality? A Counterfactual Notion

Causality, in the narrow sense we use in this course, means estimating the effect of a change in one variable (X) on another variable (Y), holding everything else equal. The Latin phrase ceteris paribus ("other things being equal") is often used to emphasize this. Intuitively, to say "X causes Y" we want to know: if we could somehow change X while keeping all other relevant factors unchanged, how much would Y change as a result? If we cannot hold other factors equal, then any observed association might be due to those other changing factors rather than a true causal effect of X on Y.

A formal way to think about causality is through counterfactuals. Consider an individual firm or person. We imagine two scenarios: one where the "treatment" X happens, and one where it does not. If we could compare that entity's outcome Y in these two scenarios, the difference would be the causal effect of X on Y for that entity. For example, suppose we want to know the effect of obtaining an MBA on a person's earnings. Ideally, we'd take one person and observe their earnings in two alternate universes: one in which they earned an MBA, and one in which they did not. The difference in those earnings would be the individual causal effect of the MBA for that person. The catch is: we can never observe both universes at once. In reality, each person either gets an MBA or not – we see only one outcome for each person. The other outcome (the one that would have happened under the alternative scenario) is the counterfactual, which is fundamentally unobservable. This is often called the Fundamental Problem of Causal Inference: for any given

unit, we cannot observe the counterfactual outcome that corresponds to the alternative treatment state. We only get to see one side of the "what if."

Remark 1.1 (Fundamental Problem of Causal Inference)**.** *In the words of statistician Paul Holland, "the fundamental problem of causal inference is that it is impossible to observe the value of* $Y_i(t)$ *and* $Y_i(t')$ *on the same unit and same moment in time". We cannot rewind time and apply a different treatment to the same unit to see what would happen. This is why causal inference is challenging – we must infer those missing pieces (the counterfactuals) indirectly*

Because we can't directly see counterfactual outcomes, a core strategy in causal inference is to find ways to approximate the all-else-equal comparison. That often involves making assumptions or using clever study designs to ensure that, aside from the treatment of interest, the groups we compare are as similar as possible.

1.3 Correlation vs. Causation

It's crucial to distinguish correlation from causation. A saying you might have heard is "Correlation does not imply causation." Two variables X and Y can be correlated (statistically associated) without X truly causing changes in Y. There are several reasons this can happen:

- **Reverse causality:** Maybe Y actually causes X, rather than the other way around. A tongue-in-cheek example: carrying an umbrella is correlated with rain, but obviously carrying umbrellas does not cause rain – it's the rain that causes people to carry umbrellas. Similarly, in finance, a rise in stock prices might be correlated with CEO optimism, but it could be that rising stock prices make CEOs optimistic (rather than optimism raising stock prices). Time order and logic help clarify which way causality could run.

- **Confounding:** X and Y might be correlated because some third factor influences both. Consider a classic nonsensical example: in some data, ice cream sales strongly correlate with shark attack incidents. Does buying ice cream cause shark attacks? Of course not. The confounder here is the weather: hot summer weather leads to more ice

cream consumption and more people swimming in the ocean (hence more shark encounters). In a financial context, suppose we find that companies with more fire insurance have higher profits. It might be that well-managed (and thus more profitable) firms are also the ones savvy enough to buy insurance – the underlying management quality drives both insurance purchase and profits, confounding the naive relationship.

- **Spurious correlations:** Sometimes two variables move together purely by chance or due to broad trends, with no direct or indirect causal link. Especially when many variables are tested, you'll eventually find some weird correlations. For example, one humorous analysis found a 94% correlation between the per-capita consumption of American cheese and the stock price of a large asset management firm (BlackRock) over a certain period. This doesn't mean eating cheese influences stock prices (or vice versa) – it's a spurious correlation. In the stock market realm, analysts have noted bizarre correlations like the "Super Bowl indicator" (whereby stock market performance correlates with which conference wins the Super Bowl) or instances where completely unrelated stocks move together for some time. These are coincidences or driven by unseen common factors, not true causation.

- **Selection bias:** Sometimes what looks like a relationship can be due to who or what is observed in each condition. For example, imagine observing that patients who receive a certain medical treatment tend to have worse health outcomes than those who don't. Does the treatment harm patients? Possibly not – it could be that the sickest patients are the ones receiving the treatment (a doctor gives the treatment to those in dire need), whereas healthier patients didn't need it. A "perfect doctor" who always gives the right treatment might paradoxically appear to have worse patient survival rates, because we only see her treating the most serious cases. In finance, we might observe that firms who undertake defensive mergers have poorer stock performance afterward than those that don't – not necessarily because the mergers caused poor performance, but perhaps because only firms already facing decline choose to merge (the underlying trouble causes both the merger decision and the performance drop). Selection bias can thus mask or even reverse the true causal effect in raw data.

The key lesson is that coming first or being associated is not enough to establish causality. Roosters crow before sunrise, but they do not cause the sun to rise. A finance example: Suppose we find that companies announcing stock buybacks see their stock price go up on average. It might be tempting to say "buybacks cause price increases." But what if companies tend to announce buybacks when they feel undervalued or when business is strong? The observed price rise could partly reflect underlying good news rather than a causal impact of the buyback itself. To claim causation, we must rule out or control for alternative explanations and isolate the effect of the variable of interest.

To isolate causation, we ideally want to compare identical worlds where only X differs. In practice, we approximate this by finding or creating comparison groups that are as similar as possible except for X. Randomized experiments, natural experiments, and various econometric techniques are all about making the treated vs. control groups comparable so that any remaining differences in Y can be attributed to the difference in X (rather than confounders). We will delve into these methods later, but first, let's formalize the causal inference framework more concretely.

1.4 The Potential Outcomes Framework

One powerful way to formalize causal ideas is the potential outcomes framework, also known as the Neyman-Rubin causal model (after Jerzy Neyman and Donald Rubin). This framework introduces the idea of labeling outcomes by the treatment status under which they are realized.

Consider a binary treatment for simplicity (extension to multiple treatments is possible but we'll start with two states like "treated" vs "not treated"). Let D_i be an indicator for whether unit i (which could be an individual, a firm, etc.) receives the treatment. In our examples, $D_i = 1$ might indicate "went to college" (or "got the treatment"), and $D_i = 0$ indicates "did not go to college" (or "no treatment"). For each unit i, we define two potential outcomes:

- $Y_i(1)$ = outcome for i if treated, $D_i = 1$
- $Y_i(0)$ = outcome for i if untreated, $D_i = 0$

These $Y_i(1)$ and $Y_i(0)$ represent the two parallel universe outcomes we discussed earlier. Only one of them will actually materialize for unit i in reality,

depending on whether i is treated or not. The other is the counterfactual outcome. We sometimes call $Y_i(1)$ and $Y_i(0)$ unit-level potential outcomes. They formalize the "what-if": $Y_i(0)$ answers "what would i's outcome be under no treatment?" and $Y_i(1)$ answers "what would it be under treatment?"

By observing data, we see either $Y_i(1)$ or $Y_i(0)$ for each i, never both. The observed outcome Y_i can be written in terms of these potential outcomes and the treatment indicator:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0).$$

This expression means: if $D_i = 1$, then $Y_i = 1 \cdot Y_i(1) + 0 \cdot Y_i(0) = Y_i(1)$ (we observe the treated potential outcome); if $D_i = 0$, then $Y_i = 0 \cdot Y_i(1) + 1 \cdot Y_i(0) = Y_i(0)$ (we observe the untreated outcome). This relies on an assumption called consistency, which essentially says the observed outcome corresponds to the appropriate potential outcome for the treatment actually received – a reasonable assumption as long as we have a well-defined treatment condition.

Because we never observe both $Y_i(1)$ and $Y_i(0)$ together for the same i, the individual-level causal effect $Y_i(1) - Y_i(0)$ is inherently unobservable for any given unit. This is just restating the fundamental problem: we lack the counterfactual for each individual. However, we can talk about average effects across a population, and those can sometimes be estimated with the right assumptions.

1.4.1 Defining Causal Effect Measures

Since individual causal effects $Y_i(1) - Y_i(0)$ are not directly observable, researchers usually focus on average causal effects over populations or subpopulations. Here are some common effect measures in the potential outcomes framework:

- **Individual Treatment Effect (ITE):** For unit i, $\tau_i = Y_i(1) - Y_i(0)$. This is the individual-level causal effect. As noted, τ_i is not observable for any single i because we cannot see both $Y_i(1)$ and $Y_i(0)$ for the same unit.

- **Average Treatment Effect (ATE):** $\tau_{\text{ATE}} = E[\, Y_i(1) - Y_i(0) \,]$. This is the expected causal effect of the treatment for a randomly chosen unit in the population. It answers: on average, how much does the outcome change due to the treatment? The ATE includes everyone,

regardless of whether they actually receive the treatment or not. It is often the target parameter for policy questions like "if we implement this policy for everyone, what would be the average effect?"

- **Average Treatment Effect on the Treated (ATT):** $\tau_{\text{ATT}} = E[\,Y_i(1) - Y_i(0) \mid D_i = 1\,]$. This is the average causal effect among those who actually received the treatment. In other words, it's the average benefit (or harm) that the treated group got from being treated, relative to what they would have experienced without treatment. For example, if we look at people who went to college ($D = 1$), τ_{ATT} is how much college increased their earnings on average compared to if those same people hadn't gone. A key thing to note is that one part of this quantity is observable: $E[Y_i(1) \mid D_i = 1]$ is just the average outcome we observe for treated units. But $E[Y_i(0) \mid D_i = 1]$ is counterfactual (what the treated would have gotten on average had they not been treated).

- **Average Treatment Effect on the Untreated (ATU):** $\tau_{\text{ATU}} = E[\,Y_i(1) - Y_i(0) \mid D_i = 0\,]$. This is the average effect for those who did not receive the treatment (had they received it). For instance, how much would the currently non-college individuals benefit on average if they did go to college? Again, one part of this (the average $Y_i(0)$ for untreated) is observable, but $E[Y_i(1) \mid D_i = 0]$ is a counterfactual mean.

- **Conditional Average Treatment Effect (CATE):** $\tau(x) = E[\,Y_i(1) - Y_i(0) \mid X_i = x\,]$. This is the average treatment effect for a specific subgroup defined by some characteristics $X_i = x$. For example, we might ask: "What is the average effect of college on earnings for women, as opposed to men?" That would be a CATE by gender. CATEs allow the treatment effect to potentially vary with observed covariates.

Each of these measures can be the parameter of interest depending on the research question. For instance, if policymakers are considering a broad policy applied to everyone, the ATE is often the relevant quantity. If we want to evaluate a program that only affects those who participate, we might care about the ATT (the effect on those who actually get the treatment). If we're curious about bringing a treatment to a group that currently doesn't have it (e.g., extending a financial product to people who haven't adopted it yet),

ATU could be of interest. If we are focusing on a particular segment (say, small firms vs. large firms), we might estimate a CATE for that segment.

Example (Choosing the right effect measure): Imagine a policy that encourages all high school graduates to attend college. If we consider implementing it universally, the ATE (the effect for a randomly selected student) would be most relevant. Now imagine a more targeted policy: suppose we only want to encourage students from schools with historically low college enrollment rates to go to college. In that case, we are interested in the effect on those who otherwise might not attend. That is closer to an ATU for the currently untreated population (those who typically wouldn't go without encouragement). In contrast, if we were evaluating the return on college for those who do attend (say, to justify student loan programs for current college students), we might focus on the ATT. The point is, the causal question dictates which parameter is most relevant. Regardless of which effect we look at, the central challenge remains: how do we estimate these causal effect parameters using the data we have? We need to relate these theoretical quantities ($E[Y(1) - Y(0)]$, etc.) to things we can actually observe. This leads us to the idea of identification and estimands.

1.5 Target Parameters vs. Estimands

A **parameter** is a quantity that describes some aspect of the truth in the population. In causal inference, our target parameter might be a causal effect like the ATE or ATT – something defined in terms of the potential outcome distributions. For example, the true ATE, denoted τ^*, could be written as:

$$\tau^* = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)].$$

This is what we want to know: say, the difference in average earnings if everyone goes to college vs. if no one goes to college. It's a fixed number (for a given population and well-defined treatment), not a random variable – though we don't know its value, it exists out there as a fact about the population.

The problem is that τ^* involves those counterfactual expectations $E[Y_i(1)]$ and $E[Y_i(0)]$. We cannot directly observe those because any given individual contributes to either the $Y_i(1)$ group or the $Y_i(0)$ group, not both. However, we can observe things like $E[Y_i \mid D = 1]$ = the average outcome among

those who actually received the treatment, and $E[Y_i \mid D = 0]$ = the average outcome among those who did not. These are features of the observed data distribution. We call a quantity like[1]

$$\tau = E[Y \mid D = 1] - E[Y \mid D = 0]$$

An **estimand** (or sometimes just an identified contrast in the data). An estimand is a function of the observable distribution that we intend to use to estimate the target parameter. In our case, τ is the difference in average outcomes between treated and untreated units in the population. If the treated and untreated groups were comparable in all relevant aspects, τ would equal the true causal effect τ^*. But in general, τ could differ from τ^* due to selection bias or confounding.

Let's consider a simple hypothetical example to illustrate this distinction between the causal effect and the naive observed difference: Suppose we have a small population of 10 individuals, and we're interested in the effect of college ($D = 1$) on earnings Y. The table below lists each individual's potential earnings with college $Y_i(1)$, potential earnings without college $Y_i(0)$, and their actual college attendance D_i. (This example is contrived for illustration.)

Individual i	$Y_i(1)$ (college)	$Y_i(0)$ (no college)	D_i (college?)
1	8	3	1
2	6	2	1
3	5	3	1
4	8	2	1
5	7	3	1
6	4	4	0
7	8	6	0
8	6	2	0
9	8	2	0
10	9	3	0
Population Mean:	**6.9**	**3.0**	–

Table 1.1: Hypothetical potential outcomes and treatment assignment for 10 individuals.

[1]To simplify notation, I will drop the subscript i whenever the meaning is clear from the context.

In this hypothetical population, five individuals went to college ($D = 1$ for $i = 1$–5) and five did not ($D = 0$ for $i = 6$–10). The true average causal effect here can be computed by comparing the two potential outcome columns:

- $E[Y(1)]$ = average of $Y_i(1)$ for $i = 1$ to $10 = 6.9$ (as shown in the table).
- $E[Y(0)]$ = average of $Y_i(0)$ for $i = 1$ to $10 = 3.0$.

So the true ATE in this toy example is $6.9 - 3.0 = 3.9$. On average, college attendance raises earnings by 3.9 (in whatever units these are).

Now, what would a naive observer compute from the observable data? They would compare the average earnings of those who went to college vs those who didn't:

- $E[Y \mid D = 1]$ = average Y among the college group. Those who went to college (IDs 1–5) have observed earnings $Y = Y(1)$ (since they took the treatment). The mean for them is $(8 + 6 + 5 + 8 + 7)/5 = 6.8$.
- $E[Y \mid D = 0]$ = average Y among the no-college group. Those who didn't go (IDs 6–10) have $Y = Y(0)$. The mean for them is $(4 + 6 + 2 + 2 + 3)/5 = 3.4$.

The observed difference is $6.8 - 3.4 = 3.4$. This $\tau = 3.4$ is our estimand (difference in group means). Compare that to the true effect 3.9. They're not the same – our estimand is biased by 0.5 in this scenario. Why the discrepancy? Because in this setup, the individuals who chose to go to college happened to be those with somewhat lower $Y(0)$ outcomes on average (notice the $Y(0)$ for IDs 1–5 average to 2.6, whereas for IDs 6–10 it's 3.4). In other words, the college-goers in this toy example tended to come from backgrounds or have characteristics that gave them lower earnings potential had they not gone compared to those who didn't go. This is a form of selection bias. Here it caused the naive observed gap (3.4) to understate the true effect (3.9) – but it could easily go the other way or even flip the sign in other cases.

We call $\tau = E[Y \mid D = 1] - E[Y \mid D = 0]$ the observational difference in means (an example of an estimand), and $\tau^* = E[Y(1) - Y(0)]$ the causal effect (target parameter). The gap between them is precisely due to the fact that $D = 1$ and $D = 0$ groups differ in ways other than the treatment.

Formally, we can write:

$$\underbrace{E[Y \mid D=1] - E[Y \mid D=0]}_{\text{observed difference}}$$
$$= \underbrace{E[Y(1) - Y(0) \mid D=1]}_{\text{ATT}} + \underbrace{(E[Y(0) \mid D=1] - E[Y(0) \mid D=0])}_{\text{selection bias}}$$

The first term on the right is the ATT (average treatment effect on the treated in our population). The second term is the selection bias term – the difference in average $Y(0)$ between those who would choose treatment and those who wouldn't. If the treated group already had different expected outcomes even in the absence of treatment, then simply comparing their observed outcomes to others will mix up the true effect with this pre-existing difference.

In our numeric example: $E[Y(1) - Y(0) \mid D = 1]$ for IDs 1–5 was about 4.2 (the ATT), and $E[Y(0) \mid D = 1] - E[Y(0) \mid D = 0] = 2.6 - 3.4 = -0.8$. Plugging in: $4.2 + (-0.8) = 3.4$, which matches the observed gap. The selection bias here is negative (meaning the treated had lower $Y(0)$ potential), causing the observed gap to be smaller than the true effect.

The goal of causal inference is to find ways to eliminate this selection bias (or more generally, any difference between the estimand and the target parameter), so that we can interpret an observed data contrast as a causal effect. In other words, we want conditions under which our estimand equals the target parameter.

When we succeed, we say the parameter is identified by that estimand. Identification is a property of the data-generating process and our assumptions – it means we can conceptually express the causal effect using only observable quantities. After establishing identification, we still have the task of estimation (using sample data to approximate the estimand). We tackle identification first, because if you cannot even write the causal effect in terms of observable distributions, no amount of data or fancy statistics will recover it.

1.6 Identification: Assumptions for Causal Inference

Identification refers to the link between the causal parameter (e.g. ATE) and some observed data distribution. To identify a causal effect, we need to

make assumptions that bridge the unobserved world of $(Y(0), Y(1))$ and the observed world of (Y, D). The classic assumptions that allow identification of treatment effects in the potential outcomes framework are:

1. **SUTVA (Stable Unit Treatment Value Assumption)** – which has two parts:

2. **No interference between units:** The outcome for unit i is not affected by the treatment status of other units. In other words, $Y_i(D)$ depends only on i's own treatment D, not on who else was treated. (So if Alice receives the treatment, it doesn't change Bob's potential outcomes.) This rules out spillovers or contagion effects across units for the scope of the analysis.

3. **No hidden variations of treatment:** There is effectively one version of the treatment and one version of control. Receiving the treatment $D = 1$ has a well-defined effect – there aren't multiple distinct ways or intensities of being "treated" that could lead to different outcomes. This also implies a consistency: if we say $Y_i(1)$ is the outcome under treatment, whenever i is observed with $D_i = 1$ we assume that outcome is indeed $Y_i(1)$ (and not some other variant).

In simpler terms, SUTVA means each unit's potential outcomes are stable in that they don't fluctuate based on others' treatments or different forms of the treatment. Formally, we can state: for any units i and j, and treatment indicators D, D' for j, $Y_i(D)$ is the same regardless of D' (no interference), and $D_i = 1$ unambiguously corresponds to the outcome $Y_i(1)$ (no multiple versions). SUTVA is usually considered a foundational assumption to even define $Y(1), Y(0)$ properly. In many finance applications, no interference often holds by design, but not always (e.g., one firm's treatment might not directly affect another in some studies, though caution: in market settings interference can occur, such as spillover of policies across firms or macroeconomic effects). No hidden variation means we have clearly defined what the "treatment" entails (e.g., a specific policy change, not a vague bundle of different interventions).

1. **Ignorability** (a.k.a. Exogeneity or Unconfoundedness) – this is the crucial assumption that who gets treated is "as good as random," at least conditional on some observed variables. Formally, ignorability

means the treatment D_i is independent of the potential outcomes once we account for some covariates X_i. In notation:

$$\{Y(0), Y(1)\} \perp\!\!\!\perp D \mid X,$$

for all i, and additionally we require positivity (overlap): $0 < P(D_i = 1 \mid X_i = x) < 1$ for all x in the support of X_i.

The independence part says that, within strata defined by X_i, the distribution of potential outcomes does not depend on whether $D = 1$ or 0. Intuitively, conditional on X, the treated and untreated units are comparable – there are no unobserved factors that systematically tilt outcomes between treated and control. The positivity part just ensures that for every combination of covariates, there is a nonzero chance of seeing both treatment and control; if some group never gets the treatment (or always gets it), then we can't compare outcomes in that subgroup – you can't learn the effect there because there's no variation in D to use.

If X_i includes all confounding variables, this assumption is often termed strong ignorability or conditional unconfoundedness. If we don't need any X (i.e., D is outright random unconditionally), then D is independent of $Y(0), Y(1)$ and we call it ignorability without caveats. In practice, X_i would be things like pre-treatment characteristics (e.g., in a finance study, we might condition on firm size, industry, prior performance, etc. – basically any observable that could affect both the likelihood of treatment and the outcome). Ignorability means that after controlling for those, there's no hidden bias; all variables that influenced both the treatment selection and the outcome have been accounted for. In an ideal randomized experiment, ignorability holds by design (random assignment makes D independent of all potential outcomes). In observational studies, ignorability is an assumption – a strong one – that we hope holds, often aided by control variables or strategies like matching.

In plain language, ignorability means we can ignore how individuals ended up in the treated vs. control group when analyzing outcomes. The potential outcomes are effectively exchangeable between treated and untreated groups once we condition on X. It's as if the treatment was random (within cells of X). If this holds, then any outcome differences we observe (conditional on X) can be attributed to the treatment, not to systematic differences in who got treated.

Given these assumptions, we can establish a very important result:

Theorem 1.1 (Identification of the ATE under SUTVA and Ignorability)**.** *Under SUTVA and conditional ignorability, the average treatment effect is identified by the difference in observed outcome means between treated and untreated groups (conditional on X). In particular, if $Y(0), Y(1) \perp\!\!\!\perp D \mid X$, then:*

$$E[Y_i(1) - Y_i(0)] = E[Y_i \mid D_i = 1, X] - E[Y_i \mid D_i = 0, X],$$

with the right-hand side understood as adjusted for X (i.e. an average over X values if needed).

Proof. Under SUTVA, $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ as noted earlier. Under ignorability, $E[Y(0) \mid D = 1, X = x] = E[Y(0) \mid D = 0, X = x] = E[Y(0) \mid X = x]$ because the distribution of $Y(0)$ doesn't depend on D once X is fixed (and similarly for $Y(1)$). By consistency (part of SUTVA), for units with $D = 1$, $Y = Y(1)$; for $D = 0$, $Y = Y(0)$. So consider the conditional expectation of Y given D and X:

- For the treated ($D = 1$): $E[Y \mid D = 1, X = x] = E[Y(1) \mid D = 1, X = x]$. By ignorability, this equals $E[Y(1) \mid X = x]$. By consistency, that is $E[Y_i(1) \mid X = x]$.
- For the untreated ($D = 0$): $E[Y \mid D = 0, X = x] = E[Y(0) \mid D = 0, X = x] = E[Y(0) \mid X = x]$.

Thus within each stratum $X = x$, the difference in observed means is $E[Y \mid D = 1, X = x] - E[Y \mid D = 0, X = x] = E[Y(1) \mid X = x] - E[Y(0) \mid X = x]$, which is the true conditional ATE for that stratum. If we then average over the distribution of X, we get $E_X E[Y(1) \mid X] - E[Y(0) \mid X] = E[Y(1)] - E[Y(0)] = \text{ATE}$. In the special case that no X is needed (completely random D), the reasoning is even simpler: $E(Y \mid D = 1) = E(Y(1))$ and $E(Y \mid D = 0) = E(Y(0))$, so their difference gives $E[Y(1) - Y(0)]$. □

Proof. By consistency, $Y = Y(1)$ if $D = 1$ and $Y = Y(0)$ if $D = 0$. Ignorability implies $E[Y(1) \mid X, D] = E[Y(1) \mid X]$ and similarly for $Y(0)$. Hence, within strata of X, differences in observed means equal true causal effects. Averaging over X yields the ATE. □

The above result is basically saying: if treatment selection is effectively random (conditional on observables), the selection bias vanishes. The treated and control groups can serve as proper counterfactuals for each other. In our earlier table example, ignorability was violated (since who went to college

depended on unobserved factors that also affected $Y(0)$), hence the observed difference diverged from the true effect. But if, hypothetically, students had been randomly assigned to college or not, then those who went and those who didn't would have, in expectation, the same distribution of ability, family background, etc. – all the things that affect earnings besides college. In that randomized scenario, we would expect the observed earnings gap to reflect only the causal effect of college. Indeed, randomization is the gold standard because it guarantees (in expectation) that $Y(0), Y(1)$ are independent of D.

To emphasize, positivity/overlap is required for the identification to be meaningful. If a certain type of unit always gets treated, we cannot learn their untreated outcome from data (since we never observe any untreated like them). In randomized experiments, overlap is usually by design (unless the experiment is stratified with some strata always assigned to one condition). In observational studies, one must be cautious that for all relevant subgroups, there are some treated and some untreated; otherwise, we have an extrapolation problem.

A quick corollary of the identification result: under the same assumptions, not only is ATE identified, but also ATT and ATU can be identified and in fact they equal each other and the ATE. If D is independent of potential outcomes, then the treated group is essentially a random sample of the population in terms of outcomes, and similarly for untreated. Thus $E[Y(0) \mid D = 1] = E[Y(0) \mid D = 0] = E[Y(0)]$, and $E[Y(1) \mid D = 1] = E[Y(1) \mid D = 0] = E[Y(1)]$. It follows that:

- ATT $= E[Y(1) - Y(0) \mid D = 1] = E[Y(1) \mid D = 1] - E[Y(0) \mid D = 1] = E[Y(1)] - E[Y(0) \mid D = 1]$. But $E[Y(0) \mid D = 1] = E[Y(0)]$ by independence, so ATT $= E[Y(1)] - E[Y(0)] =$ ATE.

- Similarly, ATU $= E[Y(1) - Y(0) \mid D = 0]$ will also equal $E[Y(1)] - E[Y(0)]$ under independence, giving ATU = ATE.

In other words, if treatment is unconfounded (ignorable), the distinction between ATE, ATT, and ATU disappears – the treatment effect is the same for everyone on average because selection into treatment isn't related to outcome potential. This is often approximately true in experiments (since treatment is random). In observational studies, however, ATT and ATE can differ if, say, those who self-select into treatment have different effects than those who don't (heterogeneous treatment effects coupled with selective uptake).

To recap: we have established that, under SUTVA and ignorability, the causal effect of interest can be expressed in terms of observable quantities. This step – expressing $E[Y(1) - Y(0)]$ as something like $E[Y \mid D = 1] - E[Y \mid D = 0]$ – is the identification step. The assumptions we invoked are strong, but without some assumptions, causal inference is impossible (by the fundamental problem). The art and science of causal research is often about finding situations where these assumptions hold plausibly, or using designs that make the assumptions more credible.

1.7 Statistical Inference: From Estimand to Estimate

Identification tells us what quantity in the population we need to look at (e.g. a difference in means). The next challenge is that we don't have the whole population data – we typically have a sample of data. We need to use the sample to estimate the estimand, and then infer the parameter. This is where familiar statistical inference tools come in (like law of large numbers, central limit theorem, confidence intervals, etc.).

Suppose we have a random sample of N units from the population. A natural estimator for the ATE (given it's identified by difference in means) is the difference in sample averages between treated and control units:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i:D_i=1} Y_i - \frac{1}{N_0} \sum_{i:D_i=0} Y_i.$$

where N_1 is the number of treated units in the sample and N_0 is the number of controls. This $\hat{\tau}$ is just the observed difference in outcomes in our sample. It is a random variable (since it depends on the particular sample we drew), whereas the true τ in the population is fixed (but unknown). We consider $\hat{\tau}$ an estimator of τ.

Under standard statistical assumptions (like that our sample is i.i.d. from the population), $\hat{\tau}$ will be a good estimator of τ in large samples. By the Law of Large Numbers (LLN), as $N \to \infty$, $\frac{1}{N_1} \sum_{D_i=1} Y_i$ will converge to $E[Y \mid D = 1]$ and $\frac{1}{N_0} \sum_{D_i=0} Y_i$ will converge to $E[Y \mid D = 0]$. So $\hat{\tau}$ will converge to $\tau = E[Y \mid D = 1] - E[Y \mid D = 0]$. If our identification assumptions hold, that τ equals the causal effect of interest, then $\hat{\tau}$ is a consistent estimator

of that causal effect. In short, with enough data, the sample difference in means will be arbitrarily close to the population difference in means.

The Central Limit Theorem (CLT) further tells us that $\hat{\tau}$ will be approximately normally distributed around the true τ for large N, with a variance that we can estimate. This lets us construct standard errors and confidence intervals for our estimate of the causal effect. For instance, we might say "our estimated ATE is 3.4, with a 95% confidence interval of [1.2, 5.6]," acknowledging sampling uncertainty.

This two-step perspective – first identification (defining which population quantity equals the causal effect under assumptions), then estimation (using sample data to approximate that quantity) – is extremely helpful. It separates the conceptual question "Can this causal effect be learned from this kind of data, in principle?" from the practical question "How do we actually compute it and how precise will our answer be?"

Joshua Angrist and Jörn-Steffen Pischke often phrase it as: First solve the identification problem (a modeling task – deciding what assumptions let you interpret something as causal), then solve the statistical inference problem (an estimation task – using data to get numbers with uncertainty)

In the context of our course, initially we assume we have an identification strategy in place (so we know which estimand equals the causal effect). We might then worry about how to actually estimate it (taking into account issues like sampling variation, finite samples, etc.). Both steps are important for credible research.

So far in this chapter, we've mostly discussed the first part: concepts of causality and identification. We treated comparisons as if we had the entire population. But of course, in practice you'll be working with samples of data, running regressions, etc., which come with noise. Rest assured, classical statistical techniques (estimation, hypothesis testing) will be our tools there – and later in the text we will also discuss how to compute correct standard errors in various scenarios, since issues like clustering, heteroskedasticity, etc., often arise in real financial data.

1.8 Randomized Experiments: The Gold Standard

If you take one thing from the identification discussion, it should be: random assignment of treatment eliminates selection bias. In a randomized controlled trial (RCT), units are assigned to treatment or control purely by chance (like flipping a coin). This ensures, in expectation, that the treated units are no different from control units in any systematic way except that one group got the treatment and the other didn't. Formally, randomization guarantees ignorability: $D \perp\!\!\!\perp Y(0), Y(1)$ (often even without needing any X) because the coin flip doesn't care about your potential outcomes. Thus $E[Y(1) \mid D = 1] = E[Y(1) \mid D = 0]$ and similarly for $Y(0)$. By our theorem, $E[Y \mid D = 1] - E[Y \mid D = 0]$ directly estimates the ATE. No fancy adjustments needed – just comparing means is valid.

Thought experiment: Imagine we could randomize who goes to college. Take a large group of high school graduates and randomly send some to college and some straight to work, irrespective of their personal preferences or backgrounds. If this were possible (and ethical), after a number of years we could compare the average earnings of the two groups. Because of randomization, any differences in innate ability, family background, motivation, etc., should wash out between the groups. Thus, any earnings difference could be attributed to the college education itself. Randomization would have solved our selection problem by making the two groups comparable. Indeed, randomization "balances" both observable and unobservable factors on average.

Randomized experiments are considered the gold standard for causal inference. In fields like medicine or psychology, performing experiments is common (clinical trials, A/B tests, etc.). In finance and economics, pure experiments are rarer, but not unheard of. For example, governments or researchers might run field experiments on microfinance loan offers, or regulators might pilot a new rule in a randomly chosen subset of firms. However, in many cases, experiments in finance are difficult, costly, or unethical. We usually cannot randomly assign interest rates, randomly force some firms to adopt a new accounting standard and not others, or randomly decide which banks get bailed out, etc. Ethical and practical constraints mean we often deal with observational data – data where the "treatments" (policy changes, corporate decisions, economic shocks) were not under our control.

Even when we can't conduct a true experiment, the ideal of randomization

guides our approach. We look for ways to approximate the experiment. This leads to what Angrist and Pischke (2010) call "identification strategies": essentially, a research design intended to solve the causal inference problem. An identification strategy typically involves either:

- **Finding or exploiting a natural experiment:** situations where something like random assignment happened by accident or by design of someone else. For example, a regulatory change might affect some firms but not others in a way that is arguably unrelated to their characteristics (as if randomly assigned). Or geographic boundary differences: perhaps one region got a policy and a neighboring region didn't, and it's plausibly arbitrary. These scenarios can sometimes be analyzed like experiments (this is the idea behind techniques like difference-in-differences and some natural experiment studies).

- **Using instrumental variables (IV):** finding a variable (instrument) that influences the treatment but is independent of the outcome except through that treatment. An IV can mimic random assignment by pushing some units to take the treatment and others not, in a way unrelated to their outcome potential. We'll devote a chapter to this.

- **Regression discontinuity designs (RDD):** when treatment assignment is determined by a rule or cutoff (like only firms above a certain size get regulated, those below do not), we can sometimes treat units near the cutoff as quasi-randomly split. RDD leverages that idea to estimate causal effects at the margin.

- **Controlled regression and matching methods:** using statistical control for observed covariates X (as in multiple regression or propensity score matching) to make treated and control groups comparable. Essentially, these adjust for differences in X hoping ignorability holds conditional on X. This is a classical econometric approach (think of including all the controls in a regression to reduce omitted variable bias).

- **Panel data methods:** if we observe the same units over time, we can difference out some sources of bias. Fixed-effects models, for example, control for time-invariant differences between units. Event studies around policy changes can help as well.

Each of these strategies has its own assumptions and conditions for validity, and they aim to emulate the ideal experiment in different ways. We will explore each in subsequent chapters. For instance, an instrumental variable provides a source of variation in D that is as-good-as random, a regression discontinuity exploits local randomization at a cutoff, matching tries to replicate a randomized block design by pairing similar units, and so on.

Terminology: We often say a study has a credible identification strategy if the authors can persuasively argue that, given their design, the assumptions (like ignorability) hold and thus their estimand equals the causal effect. Reviewers will ask "What is your identification strategy?", which is essentially "How are you getting at causation and not just correlation? What quasi-experimental variation or assumptions are you relying on?". A clear identification strategy is at the heart of modern empirical finance research, especially in what Angrist & Pischke dubbed the "Credibility Revolution" in econometrics.

In our course, after this conceptual introduction, we will delve into specific identification strategies commonly used in financial research:

- We will review linear regression and how it relates to causal inference (with a focus on what happens if you include controls, etc., and what regression coefficients mean causally under assumptions).

- We'll discuss panel data techniques (difference-in-differences, fixed effects, etc.) which are frequently used when we have time-series cross-sectional data on firms or countries.

- We'll explore instrumental variables (IV) for scenarios with endogeneity concerns (like reverse causality or omitted confounders).

- We'll examine regression discontinuity (RD) designs where applicable in finance (e.g., certain financing thresholds, rating cutoffs, etc., that create discontinuities).

- We'll cover matching methods and propensity score techniques as ways to preprocess observational data.

- Additionally, practical issues like standard errors (clustering by firm or time, dealing with heteroskedasticity, etc.) are important because getting a correct estimate is not enough; we also need correct uncertainty

quantification, especially with panel data where observations are not independent.

By the end, you should have a toolkit of methods to tackle causal questions in finance, and a clear understanding of the assumptions each method requires. We emphasize both theory and practice: the theoretical foundations (like the potential outcomes framework and assumptions we introduced here) ensure you understand what you're estimating and when you can interpret it causally, while the practical examples (from real financial research settings) ensure you can actually apply these methods to data and interpret results.

To conclude this introductory chapter, let's circle back to the big picture in a conversational tone:

Causal inference is like detective work. You, the analyst, are trying to figure out the effect of a "suspect" (the treatment) on an "outcome". But you arrive at the scene after the fact – you see the outcome and whether the treatment happened, but you can't directly see the alternate scenario. So you gather evidence (data), control for alibis (confounders), maybe find an informant (instrument) who randomly gave away some treatments, or look at natural accidents (policy changes) that acted like experiments. You do all this to reconstruct the counterfactual story: what would have happened without the treatment? Only then can you finger the treatment as the cause (or exonerate it).

In financial research, this process is both challenging and exciting. Financial markets and firms are complex, and purely random assignments are rare. But with creativity and rigor, we can leverage theory, institutional details, and statistical tools to make credible causal claims. As we proceed, remember: always ask "What am I comparing, and is that comparison apples-to-apples for causality?" Keep the mantra "all else equal" in your head. If you satisfy that, you're doing causal inference; if not, you're likely still in the realm of descriptive or predictive analysis. Moving forward, each chapter will build on these ideas, examining specific methods to achieve that ceteris paribus comparison in various contexts. By the end of this journey, you'll not only grasp the foundations laid out here, but also be able to implement and critically evaluate causal inference techniques in the wild world of financial data.