

Causal Inference

Archived study guides

These archived materials are no longer updated. This archive was exported from a saved copy of these guides.

Export date: 2026-10-05T19:26:30.244Z

Contents

Unit 1 – Probability and Statistics Fundamentals - 5 guides

Unit 2 – Potential outcomes framework - 4 guides

Unit 3 – Randomized experiments - 5 guides

Unit 4 – Observational Studies & Confounding - 4 guides

Unit 5 – Matching and propensity scores - 4 guides

Unit 6 – Instrumental variables - 4 guides

Unit 7 – Regression discontinuity designs - 3 guides

Unit 8 – Difference-in-Differences & Panel Data Methods - 4 guides

Unit 9 – Causal Graphs & Structural Models - 4 guides

Unit 10 – Causal Discovery & Learning Algorithms - 4 guides

Unit 11 – Social Science & Policy Evaluation Applications - 4 guides

Unit 12 – Causal Inference: Limitations and Frontiers - 4 guides

Unit 1 – Probability and Statistics

Fundamentals

1.1 Probability theory

Probability theory provides the mathematical foundation for causal inference. It gives you the tools to quantify uncertainty, model relationships between variables, and reason about cause and effect. This guide covers the core probability concepts you'll need: axioms, distributions, independence, Bayes' theorem, expectation and variance, common distributions, limit theorems, and how probability connects to causal reasoning.

Basics of probability

Probability quantifies how likely an event is to occur, on a scale from 0 (impossible) to 1 (certain). In causal inference, you'll use probability constantly to express uncertainty about whether one thing actually caused another, and to assess how strong your evidence is for a causal claim.

Probability axioms

These are the foundational rules that all probabilities must follow:

- Non-negativity: The probability of any event is at least zero: $P(A) \geq 0$
- Normalization: The probabilities across the entire sample space sum to 1: $P(S) = 1$
- Additivity: For mutually exclusive events A and B (they can't both happen), $P(A \cup B) = P(A) + P(B)$
- Complement rule: An event and its complement (everything that isn't that event) always sum to 1: $P(A) + P(A') = 1$

The complement rule is derived from the first three axioms, but it's used so frequently that it's worth memorizing on its own.

Sample spaces and events

The sample space (S) is the set of all possible outcomes of a random experiment. For a coin toss, $S = \{H, T\}$. For rolling a die, $S = \{1, 2, 3, 4, 5, 6\}$.

An event is any subset of the sample space. Events can be:

- Simple: a single outcome (rolling a 3)
- Compound: a combination of outcomes (rolling an even number, which is $\{2, 4, 6\}$)
- Mutually exclusive: two events that cannot occur at the same time (rolling a 1 and rolling a 6 on a single roll)

Conditional probability

Conditional probability $P(A|B)$ is the probability of event A occurring given that event B has already occurred. It's calculated as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where $P(A \cap B)$ is the probability that both A and B occur, and $P(B) > 0$.

This formula captures a simple idea: once you know B happened, you restrict your attention to only those outcomes where B is true, then ask how often A also occurs within that restricted set. Conditional probability is central to causal inference because it lets you update your beliefs as new information arrives and reason about how variables depend on each other.

Probability distributions

A probability distribution is a function that assigns probabilities to the possible values of a random variable. Distributions are how you formally model uncertainty and variability in data.

Discrete probability distributions

Discrete random variables take on a countable number of values (e.g., the number of defective items in a batch, or the number of heads in 10 coin flips).

A probability mass function (PMF) assigns a probability to each possible value. For any valid PMF, all probabilities are non-negative and they sum to 1.

Common discrete distributions include the Bernoulli, binomial, and Poisson (covered in detail below).

Continuous probability distributions

Continuous random variables can take any value within a range (e.g., height, weight, temperature).

A probability density function (PDF) describes the relative likelihood of different values. Unlike a PMF, the PDF value at a single point is not a probability. Instead, you calculate probabilities by integrating the PDF over an interval:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Common continuous distributions include the normal, exponential, and uniform.

Joint probability distributions

A joint probability distribution describes the probabilities of two or more random variables occurring together, denoted $P(X, Y)$ for random variables X and Y.

Joint distributions let you model how variables relate to each other. From a joint distribution, you can derive both marginal and conditional distributions.

Marginal probability distributions

A marginal distribution gives the probability distribution of a single variable, ignoring the others. You obtain it by summing out (discrete case) or integrating out (continuous case) the other variables from the joint distribution.

For example, if you have $P(X, Y)$, the marginal distribution of X is:

$$P(X = x) = \sum_y P(X = x, Y = y)$$

This is useful when you want to focus on one variable's behavior without worrying about the others.

Independence and dependence

These concepts describe whether knowing about one event tells you anything about another. Getting this right is critical for causal inference, because confusing independence with dependence (or vice versa) leads to incorrect causal conclusions.

Independent events

Events A and B are independent if the occurrence of one doesn't change the probability of the other. Formally:

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B)$$

An equivalent condition: $P(A \cap B) = P(A) \times P(B)$

Example: Flipping a fair coin twice. The outcome of the second flip is independent of the first. Knowing you got heads on flip 1 tells you nothing about flip 2.

Dependent events

Events A and B are dependent if the occurrence of one does change the probability of the other:

$$P(A|B) \neq P(A)$$

Example: Drawing cards from a deck without replacement. If you draw an ace first, the probability of drawing another ace on the second draw drops from $\frac{4}{52}$ to $\frac{3}{51}$.

Conditional independence

Events A and B are conditionally independent given event C if:

$$P(A|B, C) = P(A|C) \quad \text{and} \quad P(B|A, C) = P(B|C)$$

This means that once you know C, learning about A gives you no additional information about B (and vice versa). Two variables can be dependent overall but become independent once you condition on a third variable.

Conditional independence is especially important in causal inference. It's the key concept behind identifying confounders: if you condition on the right set of variables, the treatment and outcome become independent of unmeasured confounding.

Bayes' theorem

Bayes' theorem tells you how to update your beliefs about an event after observing new evidence. It's one of the most important results in probability for causal inference.

Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Each piece has a name:

- $P(A)$: the prior (your belief about A before seeing evidence B)
- $P(B|A)$: the likelihood (how probable the evidence is if A is true)
- $P(B)$: the marginal likelihood or evidence (total probability of observing B)

- $P(A|B)$: the posterior (your updated belief about A after seeing B)

Example: Suppose a disease affects 1% of the population ($P(D) = 0.01$). A test has a 95% true positive rate ($P(+|D) = 0.95$) and a 5% false positive rate ($P(+|\neg D) = 0.05$). If someone tests positive, the probability they actually have the disease is:

$$P(D|+) = \frac{0.95 \times 0.01}{(0.95 \times 0.01) + (0.05 \times 0.99)} = \frac{0.0095}{0.059} \approx 0.161$$

Only about 16%. This result surprises many people, but it makes sense: when the disease is rare, most positive tests come from the large pool of healthy people.

Prior vs posterior probabilities

The prior probability $P(A)$ reflects what you know before observing data. The posterior probability $P(A|B)$ reflects what you know after. Bayes' rule is the bridge between them.

The prior can come from previous studies, domain knowledge, or population-level data. As you observe more evidence, the posterior gets increasingly driven by the data and less by the prior.

Bayesian inference

Bayesian inference applies Bayes' theorem systematically: you specify a prior distribution over parameters of interest, observe data, and compute a posterior distribution.

This approach naturally incorporates prior knowledge into your analysis and provides a full distribution over parameter values (not just a point estimate). In causal inference, Bayesian methods are used for estimating causal effects, handling missing data, and conducting sensitivity analyses to see how results change under different assumptions.

Expectation and variance

These two quantities summarize the center and spread of a probability distribution. You'll use them constantly when defining and estimating causal effects.

Expected value

The expected value (mean) of a random variable X , written $E(X)$, is its long-run average if you could repeat the experiment infinitely many times.

- Discrete: $E(X) = \sum_x x \cdot P(X = x)$

- Continuous: $E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$

For example, the expected value of a fair six-sided die is $E(X) = \frac{1+2+3+4+5+6}{6} = 3.5$. You'll never actually roll a 3.5, but it's the average over many rolls.

Variance and standard deviation

Variance measures how spread out a distribution is around its mean:

$$\text{Var}(X) = E[(X - E(X))^2]$$

A useful computational shortcut: $\text{Var}(X) = E[X^2] - (E[X])^2$

Standard deviation $\sigma = \sqrt{\text{Var}(X)}$ is in the same units as X , which makes it easier to interpret than variance. A small standard deviation means values cluster tightly around the mean; a large one means

they're spread out.

Covariance and correlation

Covariance measures how two random variables move together:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

- Positive covariance: X and Y tend to increase together
- Negative covariance: when X increases, Y tends to decrease
- Zero covariance: no linear relationship (but there could still be a nonlinear one)

Correlation standardizes covariance to a $[-1, 1]$ scale:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

A correlation of $+1$ or -1 means a perfect linear relationship. A correlation of 0 means no linear relationship. A critical point for causal inference: correlation does not imply causation. Two variables can be highly correlated because of a shared confounder, not because one causes the other.

Common probability distributions

Knowing these distributions lets you choose the right model for your data. Each one captures a different type of random process.

Bernoulli and binomial distributions

The Bernoulli distribution models a single trial with two outcomes: success (probability p) or failure (probability $1 - p$).

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

The binomial distribution extends this to n independent trials, counting the total number of successes:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}$$

For example, if a drug has a 70% success rate ($p = 0.7$) and you treat 10 patients ($n = 10$), the binomial distribution tells you the probability of exactly k patients improving.

Poisson distribution

The Poisson distribution models the count of events occurring in a fixed interval of time or space, given a constant average rate λ :

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

It's commonly used for rare events: the number of hospital admissions per hour, the number of typos per page, or the number of mutations per gene. Its mean and variance are both equal to λ .

Normal distribution

The normal (Gaussian) distribution is the symmetric, bell-shaped distribution defined by its mean μ and standard deviation σ :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}$$

About 68% of values fall within $\pm 1\sigma$ of the mean, 95% within $\pm 2\sigma$, and 99.7% within $\pm 3\sigma$. The normal distribution shows up everywhere, partly because of the Central Limit Theorem (see below): averages of many independent random variables tend toward a normal distribution regardless of the original distribution.

Exponential distribution

The exponential distribution models the time between events in a Poisson process:

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

Its mean is $\frac{1}{\lambda}$. A distinctive property is memorylessness: the probability of waiting another t minutes is the same regardless of how long you've already waited. This makes it useful for modeling waiting times (time between customer arrivals, time until equipment failure) when there's no "aging" effect.

Limit theorems

Limit theorems describe what happens to sample statistics as you collect more and more data. They justify many of the statistical methods used in causal inference.

Law of large numbers

The law of large numbers (LLN) says that the sample average of i.i.d. (independent and identically distributed) random variables converges to the true expected value as the sample size grows.

- Weak LLN: The sample mean converges in probability to $E(X)$. Roughly, the probability that the sample mean is far from $E(X)$ shrinks toward zero.
- Strong LLN: The sample mean converges almost surely to $E(X)$. This is a stronger guarantee that says the sample mean will eventually stay close to $E(X)$ with probability 1.

The practical takeaway: larger samples give you more reliable estimates of population quantities.

Central limit theorem

The Central Limit Theorem (CLT) states that the standardized sum (or average) of a large number of i.i.d. random variables approaches a standard normal distribution, regardless of the original distribution of those variables.

Formally, if $X_1, X_2, \dots, X_n$ are i.i.d. with mean μ and variance σ^2 , then:

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1) \quad \text{as } n \rightarrow \infty$$

This is why normal-based methods (confidence intervals, z-tests, t-tests) work even when your data isn't normally distributed, as long as your sample size is large enough. In practice, $n \geq 30$ is a common rule of thumb, though the required size depends on how skewed the underlying distribution is.

Convergence in probability vs distribution

These are two different ways a sequence of random variables can "settle down" as n grows:

- Convergence in probability: $X_n \xrightarrow{P} X$ means that for any $\epsilon > 0$, $P(|X_n - X| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$. The random variables get arbitrarily close to X with high probability.

- Convergence in distribution: $X_n \xrightarrow{d} X$ means the CDF of X_n approaches the CDF of X at all continuity points. The shape of the distribution stabilizes, but individual realizations might not get close to X .

Convergence in probability is the stronger condition. If X_n converges in probability to X , it also converges in distribution, but not vice versa. The LLN involves convergence in probability; the CLT involves convergence in distribution.

Probability in causal inference

This is where probability theory connects directly to the causal questions you'll spend the rest of the course on. The key idea: causal effects are defined using potential outcomes (also called counterfactuals), and probability is the language for expressing uncertainty about those outcomes.

Probability of causation

The probability of causation (PC) asks: given that the cause was present and the effect occurred, what's the probability the effect wouldn't have occurred without the cause?

$$PC = P(Y_0 = 0 \mid Y_1 = 1)$$

Here Y_1 is the outcome when the cause is present, and Y_0 is the counterfactual outcome when the cause is absent. PC quantifies how responsible a specific cause is for an observed effect. This concept appears in legal and medical contexts (e.g., "Did this exposure cause the plaintiff's illness?").

Probability of necessity and sufficiency

These refine the idea of causation into two distinct questions:

- Probability of necessity (PN): Among cases where the effect occurred, how likely is it that removing the cause would have prevented the effect?

$$PN = P(Y_0 = 0 \mid Y = 1)$$

- Probability of sufficiency (PS): Among cases where the effect did not occur, how likely is it that introducing the cause would have produced the effect?

$$PS = P(Y_1 = 1 \mid Y = 0)$$

High PN means the cause is necessary for the effect (without it, the effect likely wouldn't happen). High PS means the cause is sufficient (with it, the effect likely would happen). A cause can be necessary without being sufficient, or vice versa.

Probability and counterfactuals

Counterfactuals are hypothetical scenarios: "What would have happened if things had been different?" In causal inference, you define causal effects by comparing potential outcomes under different conditions.

The average causal effect (ACE) is defined as:

$$ACE = E[Y_1] - E[Y_0]$$

where Y_1 is the potential outcome under treatment and Y_0 is the potential outcome under control. The fundamental problem is that you can never observe both Y_1 and Y_0 for the same individual. Probability provides the framework for estimating these quantities from data, using assumptions about independence, conditional independence, and distributional properties to bridge the gap between what you observe and what you want to know.

1.2 Random variables and distributions

Random variables and distributions provide the mathematical language for modeling uncertainty in causal inference. They let you formally describe treatment assignments, potential outcomes, and observed data, which are the building blocks for estimating causal effects. This topic covers the types of random variables, their probability distributions and properties, common named distributions, and how to transform random variables.

Types of random variables

A random variable assigns a numerical value to each outcome of a random process. In causal inference, random variables represent things like whether a subject received treatment, what outcome they experienced, or what covariates they have. The type of random variable you're working with determines which mathematical tools you'll use.

Discrete vs continuous

Discrete random variables take on a countable set of values, each with a positive probability. Think of counts: the number of patients who recover, the number of defective items in a batch.

Continuous random variables take on uncountably many values within some interval of real numbers. Think of measurements: blood pressure, income, time until an event.

This distinction matters because discrete variables use sums (and probability mass functions), while continuous variables use integrals (and probability density functions). Choosing the wrong framework leads to incorrect calculations.

Bernoulli random variables

Bernoulli random variables are the simplest type: they model a single binary outcome. The variable equals 1 ("success") with probability p and 0 ("failure") with probability $1 - p$.

In causal inference, Bernoulli variables show up constantly. Treatment assignment in a randomized experiment is often Bernoulli: each subject either gets the treatment (1) or doesn't (0).

- Expected value: $E[X] = p$
- Variance: $Var(X) = p(1 - p)$

Notice the variance is maximized when $p = 0.5$ and shrinks toward zero as p approaches 0 or 1.

Binomial random variables

If you repeat a Bernoulli trial n independent times and count the total successes, you get a binomial random variable. For example, if 10 patients each independently have a 0.3 probability of recovery, the total number who recover follows a Binomial(10, 0.3) distribution.

- PMF: $P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}$
- Expected value: $E[X] = np$
- Variance: $Var(X) = np(1 - p)$

The binomial is just the sum of n independent Bernoulli variables, so its mean and variance follow directly from linearity of expectation and independence.

Poisson random variables

Poisson random variables model the count of events occurring in a fixed interval of time or space, when events happen independently at a constant average rate. Examples: the number of hospital admissions per day, or the number of mutations in a stretch of DNA.

The single parameter λ represents the average rate of events per interval.

- PMF: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Expected value: $E[X] = \lambda$
- Variance: $Var(X) = \lambda$

A useful fact: the mean equals the variance. If you see count data where the variance is much larger than the mean, a Poisson model may not be appropriate (this is called overdispersion).

Gaussian random variables

The Gaussian (normal) distribution is the most widely used continuous distribution. It's symmetric and bell-shaped, fully characterized by its mean μ and standard deviation σ .

- PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Expected value: $E[X] = \mu$
- Variance: $Var(X) = \sigma^2$

The central limit theorem explains why the Gaussian is so important: the sum (or average) of many independent random variables tends toward a Gaussian distribution regardless of the original distribution, as long as certain regularity conditions hold. This is why sample means are approximately normal in large samples, which underpins most of the hypothesis testing and confidence interval construction you'll encounter in causal inference.

Probability distributions

Probability distributions formalize how likely each possible value of a random variable is. In causal inference, they model uncertainty in treatment assignments, potential outcomes, and observed data. The three main ways to describe a distribution are through mass functions, density functions, and cumulative distribution functions.

Probability mass functions

A probability mass function (PMF) applies to discrete random variables. It gives the probability that the variable takes each specific value: $P(X = x)$.

Two requirements for a valid PMF: - $P(X = x) \geq 0$ for all x - $\sum_x P(X = x) = 1$

You can read probabilities directly off a PMF. For instance, if X is Binomial(3, 0.5), then $P(X = 2) = \binom{3}{2} (0.5)^2 (0.5)^1 = 0.375$.

Probability density functions

A probability density function (PDF) applies to continuous random variables. Unlike a PMF, the PDF value at a single point is not a probability. Instead, you integrate the PDF over an interval to get the probability of falling in that interval:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Two requirements for a valid PDF: - $f(x) \geq 0$ for all x - $\int_{-\infty}^{\infty} f(x) dx = 1$

A common mistake is interpreting $f(x)$ as a probability. It's a density, so it can actually exceed 1 at some points, as long as the total area under the curve equals 1.

Cumulative distribution functions

The cumulative distribution function (CDF) works for both discrete and continuous variables. It gives the probability that the variable is less than or equal to a given value:

$$F(x) = P(X \leq x)$$

Key properties: - Non-decreasing: if $a < b$, then $F(a) \leq F(b)$ - $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$ - For continuous variables, the CDF is a smooth curve; for discrete variables, it's a step function

The CDF is especially handy for computing interval probabilities: $P(a < X \leq b) = F(b) - F(a)$.

Joint distributions

Joint distributions describe the simultaneous behavior of two or more random variables. For discrete variables X and Y , the joint PMF is $P(X = x, Y = y)$. For continuous variables, the joint PDF is $f(x, y)$.

Joint distributions are essential in causal inference because you're almost always dealing with multiple variables at once (treatment, outcome, covariates). From a joint distribution, you can derive conditional and marginal distributions, which are the tools for reasoning about how variables relate to each other.

Conditional distributions

A conditional distribution describes the distribution of one variable given a known value of another. This is central to causal inference, where you often want to know the distribution of an outcome given a particular treatment.

- Discrete case: $P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$
- Continuous case: $f(y \mid x) = \frac{f(x, y)}{f(x)}$

The denominator must be nonzero (you can only condition on events that have positive probability or density). Conditional distributions are what connect observed associations to potential causal relationships, though moving from association to causation requires additional assumptions.

Marginal distributions

A marginal distribution gives the distribution of a single variable, ignoring (or "marginalizing out") the others.

- Discrete case: $P(X = x) = \sum_y P(X = x, Y = y)$

- Continuous case: $f(x) = \int_{-\infty}^{\infty} f(x, y) dy$

You recover the marginal by summing or integrating the joint distribution over all values of the other variable. In causal inference, comparing marginal distributions of outcomes across treatment groups is one way to assess average treatment effects.

Properties of distributions

These summary quantities let you describe and compare distributions without specifying the full PMF or PDF. In causal inference, they're used to quantify treatment effects, measure precision of estimates, and assess relationships between variables.

Expected value

The expected value (mean) measures the center of a distribution. It's the long-run average you'd observe if you could repeat the random process infinitely many times.

- Discrete: $E[X] = \sum_x x P(X = x)$
- Continuous: $E[X] = \int_{-\infty}^{\infty} x f(x) dx$

The most important property of expected value is linearity: for any random variables X and Y and constants a and b ,

$$E[aX + bY] = aE[X] + bE[Y]$$

This holds whether or not X and Y are independent. Linearity is used constantly in deriving estimators for causal effects.

Variance and standard deviation

Variance measures how spread out a distribution is around its mean:

$$\text{Var}(X) = E[(X - E[X])^2]$$

A useful computational shortcut: $\text{Var}(X) = E[X^2] - (E[X])^2$.

The standard deviation is $\sqrt{\text{Var}(X)}$, which has the same units as X and is often easier to interpret.

Unlike expectation, variance is not linear. For a constant a : $\text{Var}(aX) = a^2 \text{Var}(X)$. For independent variables: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. If they're not independent, you need the covariance term.

In causal inference, variance quantifies the precision of your treatment effect estimates and drives the width of confidence intervals.

Covariance and correlation

Covariance measures the linear association between two random variables:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Positive covariance means the variables tend to move together; negative means they tend to move in opposite directions; zero means no linear relationship (but there could still be a nonlinear one).

Correlation standardizes covariance to the range $[-1, 1]$:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

A correlation of ± 1 means a perfect linear relationship. In causal inference, covariance and correlation help identify potential confounders and assess relationships between variables, though correlation alone never establishes causation.

Moment generating functions

A moment generating function (MGF) uniquely characterizes a distribution (when it exists). For a random variable X :

$$M_X(t) = E[e^{tX}]$$

The name comes from the fact that you can extract moments by differentiation: - $E[X] = M'_X(0)$ (first derivative at $t = 0$) - $E[X^2] = M''_X(0)$ (second derivative at $t = 0$)

MGFs are particularly useful for proving that sums of independent random variables follow specific distributions, since the MGF of a sum equals the product of the individual MGFs.

Characteristic functions

A characteristic function (CF) serves a similar role to the MGF but always exists:

$$\phi_X(t) = E[e^{itX}]$$

where i is the imaginary unit. The MGF may not exist for distributions with heavy tails (like the Cauchy distribution), but the CF always does. CFs are used in more advanced theoretical work, such as proving convergence results for sequences of random variables.

Common distributions

These named distributions appear repeatedly across statistics and causal inference. Each one models a particular type of data-generating process.

Uniform distribution

The uniform distribution assigns equal probability to all values in an interval $[a, b]$.

- PDF: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and 0 otherwise
- Expected value: $\frac{a+b}{2}$
- Variance: $\frac{(b-a)^2}{12}$

In causal inference, uniform distributions naturally model completely random treatment assignment, where every unit has the same probability of being assigned to each group.

Exponential distribution

The exponential distribution models waiting times between events in a Poisson process.

- PDF: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$
- Expected value: $\frac{1}{\lambda}$
- Variance: $\frac{1}{\lambda^2}$

The exponential distribution has the memoryless property: $P(X > s + t \mid X > s) = P(X > t)$. The probability of waiting another t units doesn't depend on how long you've already waited. This is the only continuous distribution with this property.

Gamma distribution

The gamma distribution generalizes the exponential. While the exponential models the time until the first event, the gamma models the time until the k -th event in a Poisson process.

- PDF: $f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$ for $x \geq 0$, where $\Gamma(\cdot)$ is the gamma function
- Expected value: $\frac{k}{\lambda}$
- Variance: $\frac{k}{\lambda^2}$

When $k = 1$, the gamma reduces to the exponential. The gamma is useful for modeling positive, right-skewed continuous variables.

Beta distribution

The beta distribution is defined on $[0, 1]$, making it natural for modeling probabilities or proportions.

- PDF: $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$ for $0 \leq x \leq 1$, where $B(\cdot, \cdot)$ is the beta function
- Expected value: $\frac{\alpha}{\alpha + \beta}$
- Variance: $\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

The beta is extremely flexible: depending on α and β , it can be uniform ($\alpha = \beta = 1$), U-shaped, skewed left, skewed right, or symmetric and peaked. It's also the conjugate prior for Bernoulli and binomial likelihoods, which makes it central to Bayesian approaches in causal inference.

Chi-squared distribution

The chi-squared distribution with k degrees of freedom is the distribution of the sum of k independent squared standard normal variables: if $Z_1, \dots, Z_k \sim N(0, 1)$ independently, then

$$\sum_{i=1}^k Z_i^2 \sim \chi_k^2$$

- Expected value: k
- Variance: $2k$

Chi-squared distributions are used in hypothesis testing (goodness-of-fit tests, tests of independence) and in constructing confidence intervals for variances.

Student's t-distribution

The t-distribution with ν degrees of freedom arises when you estimate the mean of a normal population using the sample standard deviation instead of the true standard deviation.

- It's symmetric and bell-shaped like the normal, but has heavier tails, meaning extreme values are more likely
- As $\nu \rightarrow \infty$, the t-distribution converges to the standard normal
- At small sample sizes (say $\nu < 30$), the heavier tails matter and produce wider confidence intervals than you'd get using a normal distribution

The t-distribution is the workhorse for hypothesis tests and confidence intervals about means when the population variance is unknown, which is almost always the case in practice.

F-distribution

The F-distribution with d_1 and d_2 degrees of freedom is the distribution of the ratio of two independent chi-squared variables, each divided by their degrees of freedom.

- It's defined only for positive values and is right-skewed
- Used primarily in ANOVA (comparing means across multiple groups) and in F-tests for comparing variances

In causal inference, F-tests come up when testing whether treatment effects differ across multiple treatment arms simultaneously.

Transformations of random variables

Transformations create new random variables by applying functions to existing ones. In causal inference, you'll encounter transformations when computing propensity scores, inverse probability weights, log-transformed outcomes, and other derived quantities.

Linear transformations

A linear transformation takes the form $Y = aX + b$, where a and b are constants.

The effect on the distribution's summary statistics is straightforward: - $E[Y] = aE[X] + b$ -
 $Var(Y) = a^2 Var(X)$

Notice that adding a constant b shifts the mean but doesn't affect the variance. Multiplying by a scales both the mean and the standard deviation (by $|a|$), and scales the variance by a^2 .

A common example: standardizing a variable by computing $Z = \frac{X - \mu}{\sigma}$. This is a linear transformation that produces a variable with mean 0 and variance 1.

Functions of random variables

For a general (possibly nonlinear) transformation $Y = g(X)$, finding the distribution of Y requires more work. Two standard approaches:

1. CDF method: Find $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$, then differentiate to get the PDF.
2. Change of variables formula: If g is monotonic and differentiable, the PDF of Y is:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

The expected value of $Y = g(X)$ can be computed without first finding the distribution of Y , using the law of the unconscious statistician (LOTUS):

- Discrete: $E[g(X)] = \sum_x g(x) P(X = x)$
- Continuous: $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$

LOTUS is a practical shortcut you'll use frequently when computing expected values of transformed variables.

1.3 Sampling and estimation

Simple random sampling

Definition of simple random sampling

Simple random sampling (SRS) means selecting a subset of individuals from a population so that every individual has an equal probability of being chosen. You use random selection methods (random number generators, lottery systems) to ensure unbiased selection. This requires a complete list of all members of the population, known as a sampling frame.

Advantages of simple random sampling

- Eliminates selection bias by giving every member of the population an equal chance of inclusion
- Allows you to use standard statistical methods to analyze results and estimate population parameters
- Produces a representative sample, enabling generalizations about the larger population

Disadvantages of simple random sampling

- Can be time-consuming and expensive, especially for large or geographically dispersed populations
- Requires a complete and accurate sampling frame, which isn't always available
- May not adequately represent small subgroups if the overall sample size is too small

Stratified sampling

Definition of stratified sampling

Stratified sampling divides the population into distinct subgroups called strata based on known characteristics (age, gender, income level). You then draw an independent random sample from each stratum. This guarantees that each subgroup appears in the final sample.

Advantages of stratified sampling

- Guarantees representation of all important subgroups within the population
- Increases precision by reducing sampling variability within each stratum
- Allows direct comparison across subgroups

Disadvantages of stratified sampling

- Requires prior knowledge of the population's characteristics to define appropriate strata
- More complex and time-consuming than SRS, since you draw multiple separate samples
- Offers little advantage over SRS if variability within strata is similar to variability between strata

Proportional vs disproportional allocation

- Proportional allocation assigns sample sizes to each stratum in proportion to that stratum's share of the population. If women make up 60% of the population, 60% of your sample comes from the women stratum.

- Disproportional allocation assigns larger samples to strata with greater internal variability, regardless of their population share.
- Optimal allocation (Neyman allocation) is a specific form of disproportional allocation that minimizes the variance of the overall estimate for a fixed total sample size.

Cluster sampling

Definition of cluster sampling

Cluster sampling divides the population into naturally occurring groups called clusters (schools, city blocks, hospitals). You randomly select some clusters, then sample from within those clusters. This is especially useful when a complete list of individuals doesn't exist but clusters can be easily identified.

Advantages of cluster sampling

- Reduces cost and time by focusing data collection on selected clusters rather than scattered individuals
- Eliminates the need for a complete sampling frame of every individual
- Allows study of naturally occurring groups and within-group relationships

Disadvantages of cluster sampling

- Clusters may not be representative of the entire population, increasing sampling error
- Requires a larger total sample size than SRS to achieve the same precision
- The clustering effect (individuals within a cluster tend to be more similar to each other than to those in other clusters) reduces effective sample size

Single-stage vs multi-stage clustering

- Single-stage: you randomly select clusters and include all members within the chosen clusters.
- Multi-stage: you randomly select clusters, then randomly select individuals within each chosen cluster. This is more practical for large or geographically dispersed populations because you don't need to survey every person in a selected cluster.

Systematic sampling

Definition of systematic sampling

Systematic sampling selects every k th element from a population list, starting with a randomly chosen element between 1 and k . The sampling interval k is calculated by dividing the population size by the desired sample size. The list needs to be arranged in some order (alphabetical, chronological, etc.).

Advantages of systematic sampling

- Simple to implement: only the first element is chosen randomly, and the rest follow a fixed interval
- Ensures even coverage across the entire population list
- Can be more efficient than SRS for large populations

Disadvantages of systematic sampling

- Periodicity in the list can introduce bias. If the sampling interval happens to align with a repeating pattern in the data, your sample will be skewed.
- Requires a complete, ordered list of the population
- Estimating sampling error and constructing confidence intervals is more complex than with SRS

Sample size determination

Factors influencing sample size

- Desired precision: smaller margins of error require larger samples
- Population variability: more variable populations need larger samples
- Confidence level: higher confidence (99% vs. 95%) requires larger samples
- Available resources: budget, time, and personnel may constrain feasible sample size

Sample size for estimating means

The required sample size depends on the population standard deviation, desired margin of error, and confidence level:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

where n is the sample size, $z_{\alpha/2}$ is the critical value for the desired confidence level, σ is the population standard deviation, and E is the margin of error.

For example, to estimate a population mean within $E = 2$ units at 95% confidence with $\sigma = 10$:

$$n = \left(\frac{1.96 \times 10}{2} \right)^2 = 96.04, \text{ so you'd need at least 97 observations.}$$

Sample size for estimating proportions

$$n = \frac{z_{\alpha/2}^2 \cdot p(1-p)}{E^2}$$

where p is the estimated population proportion. If you have no prior estimate for p , use $p = 0.5$ because that maximizes $p(1-p)$ and gives the most conservative (largest) sample size.

Sample size for comparing means

This depends on the difference in means you want to detect, the population standard deviations, desired statistical power, and significance level. You'll need to specify the null and alternative hypotheses and whether the test is one-tailed or two-tailed.

Sample size for comparing proportions

Similarly, this depends on the difference in proportions to detect, the population proportions, desired power, and significance level. Again, you specify hypotheses and test direction.

Sampling error and bias

Definition of sampling error

Sampling error is the difference between a sample statistic and the corresponding population parameter. It arises from the random variation inherent in drawing a sample rather than measuring the entire population. You can reduce it by increasing sample size or using more efficient sampling methods (like stratified sampling).

Definition of sampling bias

Sampling bias is a systematic error that occurs when the sample is not representative of the population. Unlike sampling error, bias causes estimates to consistently deviate from the true value in one direction. Increasing sample size does not fix bias, because the problem is in the sampling method itself.

Types of sampling bias

- Selection bias: the sampling method favors certain members over others
- Voluntary response bias: people who choose to participate often have stronger opinions than those who don't
- Undercoverage bias: certain population members have a lower probability of being included (e.g., a phone survey that misses people without phones)
- Non-response bias: people who respond to a survey differ systematically from those who don't
- Measurement bias: the data collection process itself distorts responses
- Leading question bias: question wording pushes respondents toward a particular answer
- Social desirability bias: respondents answer in ways that make them look good rather than answering honestly

Reducing sampling error and bias

- Use probability sampling methods (SRS, stratified sampling) to ensure representativeness
- Increase sample size to reduce sampling error (but remember this won't fix bias)
- Design questionnaires carefully to minimize measurement bias
- Implement strategies to boost response rates (incentives, follow-up contacts) to reduce non-response bias
- Compare your sample's key characteristics to known population values and apply weighting techniques to adjust for discrepancies

Point estimation

Definition of point estimation

Point estimation uses sample data to calculate a single value as the best guess for a population parameter. Common point estimates include the sample mean (for the population mean), sample proportion (for the population proportion), and sample variance (for the population variance). A point estimate is concise but tells you nothing about the uncertainty around it.

Properties of good estimators

- Unbiasedness: the expected value of the estimator equals the true parameter. On average, across many samples, the estimator hits the right target.
- Efficiency: among all unbiased estimators, the efficient one has the smallest variance. It gives you the tightest clustering around the true value.
- Consistency: as sample size grows, the estimator converges in probability to the true parameter.
- Sufficiency: the estimator captures all the information in the sample that's relevant to the parameter.

Maximum likelihood estimation

Maximum likelihood estimation (MLE) finds the parameter values that make the observed data most probable.

1. Write down the likelihood function, which is the joint probability of observing your sample data given the parameter values.
2. Take the natural log to get the log-likelihood function (this simplifies the math).
3. Take the derivative of the log-likelihood with respect to the parameter, set it equal to zero, and solve.

MLE estimators are consistent and asymptotically efficient, which makes them widely used in practice.

Method of moments estimation

Method of moments (MoM) works by setting sample moments equal to their population counterparts and solving for the parameters.

1. Express population moments (mean, variance, etc.) in terms of the unknown parameters.
2. Replace population moments with the corresponding sample moments.
3. Solve the resulting system of equations for the parameter estimates.

MoM is generally less efficient than MLE but can be easier to compute, especially when the likelihood function is hard to work with.

Interval estimation

Definition of interval estimation

Interval estimation produces a range of plausible values for a population parameter rather than a single point estimate. These intervals, called confidence intervals, incorporate sample variability and a chosen confidence level to account for uncertainty.

Confidence intervals

A confidence interval is a range of values centered around the point estimate. It's constructed using three ingredients:

- The point estimate (e.g., sample mean)
- The standard error of the estimate
- A critical value from the appropriate distribution (z-distribution or t-distribution)

The confidence level (typically 90%, 95%, or 99%) specifies how often the procedure produces intervals that contain the true parameter.

Interpreting confidence intervals

This is one of the most commonly misunderstood concepts in statistics. A 95% confidence interval does not mean there's a 95% probability the true parameter lies within that specific interval. The true parameter is fixed; it's either in the interval or it isn't.

The correct interpretation: if you repeated the sampling process many times and built a 95% CI each time, about 95% of those intervals would contain the true parameter.

Factors affecting confidence interval width

- Sample size: larger samples produce narrower intervals
- Population variability: more variable populations produce wider intervals
- Confidence level: higher confidence (99% vs. 95%) produces wider intervals because you need a bigger net to be more certain
- Sampling method: more efficient methods like stratified sampling can produce narrower intervals than SRS for the same sample size

Estimating population means

Sample mean as an estimator

The sample mean $\bar{x}$ is an unbiased and consistent estimator of the population mean μ . It's calculated as:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Unbiasedness means $E(\bar{x}) = \mu$: if you could take every possible sample and average all the sample means, you'd get exactly μ .

Standard error of the mean

The standard error measures how much $\bar{x}$ varies from sample to sample:

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Notice the $\sqrt{n}$ in the denominator. To cut the standard error in half, you need to quadruple your sample size. When σ is unknown (which is almost always the case), substitute the sample standard deviation S .

Confidence intervals for population means

For a 95% confidence interval:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

With $z_{\alpha/2} = 1.96$ for 95% confidence. If σ is unknown and you use S , replace the z-value with the appropriate t-value.

t-distribution for small samples

When the sample size is small (typically $n < 30$) and σ is unknown, you use the t-distribution instead of the standard normal. The t-distribution has heavier tails, which produces wider confidence intervals. This accounts for the extra uncertainty introduced by estimating σ with S from a small sample.

The degrees of freedom are $df = n - 1$. As n grows large, the t-distribution converges to the standard normal distribution.

Estimating population proportions

Sample proportion as an estimator

The sample proportion $\hat{p}$ is an unbiased and consistent estimator of the population proportion P :

$$\hat{p} = \frac{X}{n}$$

where X is the number of "successes" (individuals with the characteristic of interest) and n is the sample size. Its expected value is $E(\hat{p}) = p$.

Standard error of the proportion

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

When P is unknown, substitute $\hat{p}$ from your sample.

Confidence intervals for population proportions

For a 95% confidence interval:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{\alpha/2} = 1.96$ for 95% confidence.

Normal approximation for large samples

The sampling distribution of $\hat{p}$ is technically binomial, but for large samples you can approximate it with a normal distribution. The standard rule of thumb is that both $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$ should hold.

This approximation comes from the Central Limit Theorem: as n increases, the distribution of $\hat{p}$ approaches a normal distribution regardless of the population's shape. If these conditions aren't met, you should use exact binomial methods instead.

1.4 Hypothesis testing

Hypothesis testing overview

Hypothesis testing is a statistical method for making decisions about population parameters using sample data. You set up two competing claims, then use probability to decide which one the data support. In causal inference, this is how you assess whether a treatment effect is real or just noise from random variation.

Null and alternative hypotheses

The null hypothesis (H_0) represents the default position: no effect, no difference, no relationship. The alternative hypothesis (H_a) is the claim you're actually trying to find evidence for.

The logic works like a courtroom. You assume H_0 is true (innocent until proven guilty), then ask whether the data are surprising enough to reject that assumption.

- H_0 : The mean weight of a population is 150 lbs; H_a : The mean weight is not 150 lbs
- H_0 : There is no association between smoking and lung cancer; H_a : There is an association between smoking and lung cancer

Notice that you never "accept" H_0 . You either reject it or fail to reject it. Failing to reject just means you didn't find enough evidence against it.

Significance level and p-values

The significance level (α) is the threshold you set before running the test for how much risk of a false alarm you'll tolerate. Common choices are 0.05 (5%) and 0.01 (1%).

The p-value is the probability of seeing data as extreme as (or more extreme than) what you observed, assuming H_0 is true. It is not the probability that H_0 is true.

- If $p < \alpha$, reject H_0 . The result is "statistically significant."
- If $p \geq \alpha$, fail to reject H_0 .
- A smaller p-value means the data are harder to explain under H_0 , giving stronger evidence against it.

One-tailed vs two-tailed tests

A two-tailed test checks for any difference in either direction ($H_a : \mu \neq \mu_0$). It splits α across both tails of the distribution.

A one-tailed test checks for a difference in a specific direction ($H_a : \mu > \mu_0$ or $H_a : \mu < \mu_0$). It puts all of α in one tail, making it easier to detect an effect in that direction but blind to effects in the other.

Use a one-tailed test only when you have a strong prior reason to expect the effect goes one way. If you're unsure about direction, go two-tailed.

Type I and Type II errors

These are the two ways a hypothesis test can go wrong:

	H_0 is actually true	H_0 is actually false
Reject H_0	Type I error (false positive)	Correct decision
Fail to reject H_0	Correct decision	Type II error (false negative)

- The probability of a Type I error equals α (your significance level).
- The probability of a Type II error is denoted β .
- There's a direct trade-off: lowering α (being stricter) reduces false positives but increases the chance of missing real effects.

Power of a test

Power is the probability of correctly rejecting H_0 when it's actually false. It equals $1 - \beta$.

A test with low power might fail to detect a real treatment effect, which is a serious problem in causal inference. Four main factors influence power:

- Sample size: Larger samples give more power.
- Effect size: Bigger true effects are easier to detect.
- Significance level: A more lenient α (e.g., 0.05 vs 0.01) increases power.
- Variability in the data: Less noise means more power.

Researchers often conduct a power analysis before collecting data to determine the sample size needed to detect a meaningful effect.

Common hypothesis tests

Different tests suit different data types and research questions. The choice depends on whether your data are numerical or categorical, whether you know the population standard deviation, and how many groups you're comparing.

Z-test for population mean

Use the z-test when the population standard deviation (σ) is known and the sample size is large ($n > 30$) or the population is normally distributed.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- $\bar{x}$ = sample mean
- μ_0 = hypothesized population mean
- σ = population standard deviation
- n = sample size

In practice, you rarely know σ , so the z-test is less common than the t-test. It assumes observations are independent and identically distributed.

T-test for sample mean

Use the t-test when σ is unknown (which is most of the time). It uses the sample standard deviation s instead and accounts for the extra uncertainty with heavier tails in the t-distribution.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The t-test comes in several forms: - One-sample: Compare a sample mean to a known value. - Two-sample (independent): Compare means of two separate groups (e.g., treatment vs. control). - Paired: Compare means from the same subjects measured twice (e.g., before and after treatment).

Assumptions: observations are independent, and the population is approximately normal (though with larger samples, the Central Limit Theorem makes this less critical).

Chi-square test for independence

The chi-square test checks whether two categorical variables are independent. You build a contingency table of observed frequencies and compare them to the frequencies you'd expect if the variables were unrelated.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O = observed frequency in each cell
- E = expected frequency in each cell (calculated as row total $\times$ column total / grand total)

For example, you might test whether gender and voting preference are independent. A large χ^2 value means the observed pattern deviates substantially from what independence would predict.

Assumptions: expected frequencies should be at least 5 in each cell, and observations must be independent.

F-test for equality of variances

The F-test compares the variances of two populations to check whether they're equal.

$$F = \frac{s_1^2}{s_2^2}$$

where s_1^2 and s_2^2 are the sample variances. If the populations have equal variances, this ratio should be close to 1.

The F-distribution also underlies ANOVA (analysis of variance), which extends the comparison to three or more group means. Assumptions include normality and independence of observations.

Steps in hypothesis testing

Here's the systematic procedure, from start to finish:

1. Formulating hypotheses

State H_0 and H_a clearly based on your research question. They must be mutually exclusive (both can't be true) and exhaustive (one of them must be true). Think carefully about what it would mean to reject H_0 in the context of your study.

2. Selecting the appropriate test

Choose a test based on: - Data type: Numerical $\rightarrow$ z-test or t-test. Categorical $\rightarrow$ chi-square. - What you know: Is σ known? $\rightarrow$ z-test. Unknown? $\rightarrow$ t-test. - Number of groups: Two groups $\rightarrow$ t-test. Three or more $\rightarrow$ ANOVA. - Direction of hypothesis: One-tailed or two-tailed.

Check whether the test's assumptions (normality, independence, minimum expected counts) are reasonably met.

3. Calculating the test statistic

Plug your sample data into the formula for the test you selected. The test statistic measures how far your sample result is from what H_0 predicts, in standardized units.

4. Determining the critical value

Look up the critical value for your chosen α and degrees of freedom using a statistical table or software. The critical value marks the boundary of the rejection region.

- For a two-tailed test at $\alpha = 0.05$, you split the rejection region: 2.5% in each tail.
- For a one-tailed test at $\alpha = 0.05$, the full 5% goes in one tail.

5. Making the decision

Compare your test statistic to the critical value, or compare your p-value to α : - Test statistic in the rejection region (or $p < \alpha$) → reject H_0 . - Test statistic outside the rejection region (or $p \geq \alpha$) → fail to reject H_0 .

Always interpret the result in context. "We reject H_0 at the 5% level" means something specific about your research question, not just a number comparison.

Interpreting results

Getting a p-value is only half the job. You also need to understand what the result means and what it doesn't mean.

Confidence intervals

A confidence interval gives a range of plausible values for the population parameter. A 95% confidence interval means that if you repeated the study many times, about 95% of the intervals you'd construct would contain the true parameter.

Confidence intervals complement hypothesis tests by showing both the direction and precision of an estimate. A narrow interval means your estimate is precise; a wide one signals more uncertainty. If a 95% CI for a treatment effect doesn't include zero, that's equivalent to rejecting H_0 at $\alpha = 0.05$.

Effect size and practical significance

Statistical significance tells you whether an effect exists. Effect size tells you how big it is. Common measures include:

- Cohen's d: Standardized difference between two means. Values of 0.2, 0.5, and 0.8 are often called small, medium, and large.
- Pearson's r: Correlation coefficient, ranging from -1 to 1.
- Odds ratio: Used in categorical data to compare the odds of an outcome between groups.

Practical significance asks whether the effect matters in the real world. With a large enough sample, you can get a statistically significant p-value for a tiny, meaningless difference. For example, a drug that lowers blood pressure by 0.5 mmHg might be statistically significant with 100,000 participants but clinically irrelevant.

Limitations of hypothesis testing

Hypothesis testing has real weaknesses you should be aware of:

- It does not prove causation on its own. A significant result shows an association, but confounding and reverse causation could still explain it.

- It's sensitive to sample size. Very large samples can make trivially small effects "significant," while small samples may miss important effects.
- Violations of assumptions (non-normality, dependence between observations) can produce misleading results.
- Multiple testing inflates the Type I error rate. If you run 20 tests at $\alpha = 0.05$, you'd expect about 1 false positive by chance alone. Corrections like Bonferroni adjustment help address this.

Never treat a p-value as the final word. Consider study design, data quality, effect size, and theoretical plausibility alongside your test results.

Applications in causal inference

Hypothesis testing connects directly to the core question of causal inference: did the treatment actually cause the observed outcome, or could the difference be due to chance?

Testing for significant treatment effects

In a randomized controlled trial (RCT), you randomly assign subjects to treatment and control groups, then test whether the difference in outcomes is statistically significant. Randomization is what justifies the causal interpretation: if the groups are comparable at baseline, a significant difference in outcomes can be attributed to the treatment.

For example, testing whether a new drug reduces blood pressure compared to a placebo. If the two-sample t-test yields $p = 0.003$, you'd reject H_0 and conclude the drug has a significant effect (at the 5% level).

Comparing groups in experiments

When experiments involve more than two groups, you need tests that handle multiple comparisons:

- ANOVA tests whether at least one group mean differs from the others.
- Post-hoc tests (like Tukey's HSD) identify which specific pairs of groups differ.

For example, comparing three teaching methods on exam scores. ANOVA might show a significant overall difference, and post-hoc tests would reveal that Method A outperforms Method C but not Method B.

Assessing validity of causal claims

In observational studies and quasi-experiments (where randomization isn't possible), hypothesis testing can evaluate the strength of evidence for a causal relationship. But the results are more ambiguous because you can't rule out confounders the way randomization does.

A significant association between two variables is necessary but not sufficient for a causal claim. You also need to address confounding, selection bias, and reverse causation through study design or statistical adjustment.

Hypothesis testing vs estimation approaches

Hypothesis testing gives you a binary answer: reject or don't reject. Estimation approaches give you richer information:

- Confidence intervals show the range and precision of the estimated effect.

- Bayesian methods incorporate prior knowledge and produce probability distributions over parameter values rather than a single yes/no decision.

In causal inference, the best practice is to use both. Report whether the effect is significant and how large it is with a confidence interval. A treatment effect of 5 points (95% CI: 2 to 8, $p = 0.002$) tells you far more than just " $p < 0.05$."

1.5 Regression analysis

Regression analysis is a powerful tool for understanding relationships between variables and estimating causal effects. It allows researchers to model how changes in one variable affect another, while controlling for other factors. This technique is essential for causal inference in observational studies.

From simple linear regression to more complex methods like instrumental variables and panel data analysis, regression offers a range of approaches for different research questions. Understanding the assumptions, limitations, and interpretation of regression models is crucial for drawing valid conclusions about cause-and-effect relationships in various fields of study.

Basics of regression analysis

- Regression analysis is a statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables
- It allows researchers to estimate the effect of an independent variable on the dependent variable, while controlling for potential confounding factors
- Regression models are widely used in causal inference to estimate causal effects and test hypotheses about cause-and-effect relationships

Simple linear regression

Relationship between variables

- Simple linear regression examines the linear relationship between two continuous variables - one dependent variable and one independent variable
- The goal is to find the best-fitting straight line that describes how the dependent variable changes as a function of the independent variable
- The strength and direction of the relationship is quantified by the slope of the regression line and the correlation coefficient

Ordinary least squares method

- The most common method for estimating the parameters of a simple linear regression model is ordinary least squares (OLS)
- OLS minimizes the sum of squared residuals, which are the differences between the observed values of the dependent variable and the values predicted by the regression line
- The resulting OLS estimates are unbiased and efficient under certain assumptions, such as linearity, homoscedasticity, and independence of errors

Assumptions and limitations

- Simple linear regression relies on several key assumptions, including:

- Linearity: The relationship between the dependent and independent variables is linear
- Independence: The observations are independent of each other
- Homoscedasticity: The variance of the errors is constant across all levels of the independent variable
- Normality: The errors are normally distributed
- Violations of these assumptions can lead to biased or inefficient estimates and invalid inference
- Simple linear regression is limited to modeling the relationship between two variables and cannot account for the influence of other potential confounding factors

Multiple linear regression

Controlling for confounding variables

- Multiple linear regression extends simple linear regression by including multiple independent variables in the model
- By controlling for additional variables, multiple regression can account for potential confounding factors that may bias the estimated relationship between the main independent variable and the dependent variable
- For example, when estimating the effect of education on income, multiple regression can control for factors such as age, gender, and work experience to isolate the causal effect of education

Interpreting regression coefficients

- In multiple regression, each independent variable has its own regression coefficient, which represents the change in the dependent variable associated with a one-unit increase in that independent variable, holding all other variables constant
- The interpretation of regression coefficients depends on the scale and coding of the variables
- Standardized coefficients can be used to compare the relative importance of different independent variables in the model

Model selection techniques

- With multiple independent variables, it is important to select a parsimonious model that includes only the most relevant predictors
- Model selection techniques, such as stepwise regression, best subset selection, and regularization methods (like lasso and ridge regression), can help identify the optimal set of variables to include in the model
- These techniques balance the trade-off between model complexity and goodness-of-fit, aiming to prevent overfitting and improve the model's generalizability

Logistic regression

Binary outcome variables

- Logistic regression is a type of regression analysis used when the dependent variable is binary or dichotomous (e.g., success/failure, yes/no)

- Unlike linear regression, which models the expected value of a continuous dependent variable, logistic regression models the probability of an event occurring
- The logistic function, also known as the sigmoid function, is used to transform the linear combination of independent variables into a probability between 0 and 1

Odds ratios and log odds

- In logistic regression, the coefficients are typically interpreted in terms of odds ratios or log odds
- The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group
- A coefficient in logistic regression represents the change in the log odds of the dependent variable for a one-unit increase in the independent variable, holding other variables constant
- An odds ratio greater than 1 indicates an increased likelihood of the event, while an odds ratio less than 1 indicates a decreased likelihood

Interpreting logistic coefficients

- To interpret the coefficients in a logistic regression model, it is common to exponentiate them to obtain odds ratios
- For example, if the coefficient for a binary independent variable is 0.7, the odds ratio would be $\exp(0.7) \approx 2.01$, meaning that the odds of the event occurring are about twice as high for the group represented by that variable compared to the reference group
- It is important to note that odds ratios do not directly translate to probabilities, but they provide a measure of the strength and direction of the association between the independent variable and the binary outcome

Regression discontinuity designs

Sharp vs fuzzy designs

- Regression discontinuity (RD) designs are quasi-experimental methods used to estimate causal effects when treatment assignment is determined by a cutoff value of a continuous variable
- In a sharp RD design, treatment assignment is deterministic - all units above the cutoff receive treatment, while all units below do not
- In a fuzzy RD design, treatment assignment is probabilistic - the probability of receiving treatment changes discontinuously at the cutoff, but some units may not comply with their assigned status

Identifying causal effects

- RD designs rely on the assumption that units just above and below the cutoff are similar in all respects except for their treatment status
- By comparing outcomes for units near the cutoff, researchers can estimate the local average treatment effect (LATE) at the cutoff
- The LATE represents the causal effect of the treatment for units whose treatment status is affected by the cutoff

Validity and robustness checks

- To ensure the validity of RD estimates, researchers must check for manipulation of the running variable (the variable determining treatment assignment) and test for discontinuities in other covariates at the cutoff
- Robustness checks, such as testing different bandwidth selections and functional forms, can help assess the sensitivity of the results to modeling choices
- Placebo tests, which involve testing for discontinuities at false cutoff values, can provide evidence for the credibility of the RD design

Instrumental variable regression

Endogeneity and omitted variables

- Endogeneity occurs when an independent variable is correlated with the error term in a regression model, leading to biased estimates of causal effects
- Omitted variable bias is a common source of endogeneity, which arises when a relevant variable is not included in the model and is correlated with both the dependent and independent variables
- Instrumental variable (IV) regression is a method used to address endogeneity by using an external variable (the instrument) to isolate the exogenous variation in the independent variable of interest

Two-stage least squares

- Two-stage least squares (2SLS) is a common estimation procedure for IV regression
- In the first stage, the endogenous independent variable is regressed on the instrument and other exogenous variables to obtain predicted values
- In the second stage, the dependent variable is regressed on the predicted values of the endogenous variable from the first stage and the other exogenous variables
- The resulting 2SLS estimates are consistent and asymptotically normal under certain assumptions

Validity of instrumental variables

- For an instrument to be valid, it must satisfy two key assumptions:
- Relevance: The instrument must be correlated with the endogenous independent variable
- Exogeneity: The instrument must be uncorrelated with the error term in the second-stage regression
- The strength of the instrument can be assessed using the F-statistic from the first-stage regression, with a common rule of thumb being that an F-statistic greater than 10 indicates a strong instrument
- Overidentification tests can be used to assess the exogeneity of the instruments when there are more instruments than endogenous variables

Regression with panel data

Fixed effects vs random effects

- Panel data, also known as longitudinal data, consists of repeated observations on the same units (e.g., individuals, firms, or countries) over time
- Fixed effects models control for unobserved, time-invariant heterogeneity by including unit-specific intercepts, which capture all stable characteristics of the units
- Random effects models treat the unit-specific effects as random variables, assuming they are uncorrelated with the independent variables
- The choice between fixed and random effects depends on the assumptions about the correlation between the unit-specific effects and the independent variables, as well as the goals of the analysis

Difference-in-differences estimation

- Difference-in-differences (DiD) is a method used to estimate causal effects by comparing the changes in outcomes over time between a treatment group and a control group
- DiD relies on the assumption that, in the absence of treatment, the treatment and control groups would have followed parallel trends in the outcome variable
- The DiD estimator is obtained by subtracting the change in the outcome for the control group from the change in the outcome for the treatment group, effectively removing any common time trends and time-invariant differences between the groups

Dealing with serial correlation

- Panel data often exhibits serial correlation, meaning that observations within a unit are correlated over time
- Ignoring serial correlation can lead to underestimated standard errors and over-rejection of null hypotheses
- Cluster-robust standard errors, which allow for arbitrary correlation within units over time, are a common solution to this problem
- Other approaches include modeling the serial correlation directly using autoregressive or moving average error structures

Quantile regression

Estimating effects at different quantiles

- Quantile regression is a method that allows researchers to estimate the effect of independent variables on different quantiles (percentiles) of the dependent variable's distribution
- Unlike ordinary least squares regression, which focuses on the conditional mean, quantile regression provides a more comprehensive picture of the relationship between variables
- Quantile regression is particularly useful when the effect of the independent variables is expected to vary across the distribution of the dependent variable

Comparing quantile coefficients

- Quantile regression coefficients can be interpreted as the change in the specific quantile of the dependent variable associated with a one-unit increase in the independent variable, holding other variables constant
- Comparing coefficients across different quantiles allows researchers to assess how the effect of the independent variable changes along the distribution of the dependent variable
- Tests for the equality of coefficients across quantiles can be used to formally examine whether the effect is constant or varies significantly

Robustness to outliers

- Quantile regression is more robust to outliers and heavy-tailed distributions compared to mean regression methods
- By focusing on specific quantiles rather than the mean, quantile regression is less sensitive to extreme values in the dependent variable
- This property makes quantile regression a valuable tool for analyzing data with non-normal distributions or in the presence of influential observations

Nonparametric regression

Kernel regression and local averaging

- Nonparametric regression methods do not assume a specific functional form for the relationship between the dependent and independent variables
- Kernel regression is a nonparametric method that estimates the conditional expectation of the dependent variable at each value of the independent variable by averaging the observations in a local neighborhood
- The size of the local neighborhood is determined by the bandwidth, which controls the trade-off between bias and variance in the estimates

Spline regression models

- Spline regression is another nonparametric approach that fits piecewise polynomial functions to the data, allowing for a flexible relationship between the variables
- The most common types of splines are cubic splines, which ensure continuity and smoothness at the knots (points where the polynomial pieces connect)
- The number and location of the knots can be chosen using data-driven methods, such as cross-validation, to optimize the fit of the model

Advantages over parametric models

- Nonparametric regression methods offer several advantages over parametric models:
- Flexibility: They can capture complex, nonlinear relationships between variables without imposing a specific functional form

- **Robustness:** They are less sensitive to model misspecification and can provide reliable estimates even when the true relationship is unknown
- **Adaptivity:** They can automatically adapt to the local structure of the data, allowing for different levels of smoothness in different regions of the predictor space
- However, nonparametric methods may require larger sample sizes and can be more computationally intensive than parametric models

Evaluating regression models

Goodness-of-fit measures

- Goodness-of-fit measures quantify how well a regression model fits the observed data
- The coefficient of determination (R-squared) measures the proportion of variance in the dependent variable that is explained by the independent variables
- Adjusted R-squared accounts for the number of independent variables in the model and penalizes the addition of irrelevant variables
- The root mean squared error (RMSE) and mean absolute error (MAE) measure the average magnitude of the residuals, with lower values indicating better fit

Residual diagnostics and plots

- Residual diagnostics involve examining the residuals (the differences between the observed and predicted values) to assess the validity of the regression assumptions
- Residual plots, such as residuals vs. fitted values and residuals vs. each independent variable, can reveal patterns that indicate violations of linearity, homoscedasticity, or independence
- Normal probability plots (Q-Q plots) can be used to check the normality assumption of the residuals
- Identifying and addressing violations of assumptions is crucial for ensuring the validity and reliability of the regression results

Cross-validation and out-of-sample testing

- Cross-validation is a technique used to assess the performance and generalizability of a regression model
- The data is split into k subsets (folds), and the model is trained on $k-1$ folds and tested on the remaining fold, with this process repeated k times
- The average performance across the k folds provides an estimate of the model's out-of-sample performance
- Out-of-sample testing involves evaluating the model's performance on a separate dataset that was not used in the model training process
- This approach helps to assess the model's ability to generalize to new, unseen data and can help prevent overfitting

Causal inference with regression

Establishing causal relationships

- Regression analysis can be a powerful tool for estimating causal effects, but it is important to recognize that regression alone does not guarantee a causal interpretation
- To establish a causal relationship, researchers must consider the potential sources of bias and confounding that may affect the estimates
- Randomized controlled trials (RCTs) are considered the gold standard for causal inference, as they ensure that treatment assignment is independent of potential outcomes

Potential outcomes framework

- The potential outcomes framework, also known as the Rubin causal model, provides a formal language for defining and estimating causal effects
- Each unit is assumed to have potential outcomes under different treatment conditions, and the causal effect is defined as the difference between these potential outcomes
- In observational studies, researchers observe only one potential outcome for each unit, leading to the fundamental problem of causal inference
- Statistical methods, such as propensity score matching, inverse probability weighting, and instrumental variable estimation, can be used to estimate causal effects under certain assumptions

Challenges and limitations

- Despite the usefulness of regression analysis for causal inference, there are several challenges and limitations to keep in mind:
- Unobserved confounding: Regression estimates may be biased if there are unobserved variables that affect both the treatment and the outcome
- Model misspecification: If the functional form of the relationship between the variables is incorrectly specified, the estimates may be biased and inconsistent
- Extrapolation: Regression estimates are most reliable within the range of the observed data and may not generalize well to units with covariate values outside this range
- Sampling bias: If the sample is not representative of the population of interest, the estimates may not be externally valid
- Careful design, data collection, and sensitivity analyses can help address these challenges and strengthen the credibility of causal claims based on regression analysis

Unit 2 – Potential outcomes framework

2.1 Counterfactuals and potential outcomes

Definition of counterfactuals

A counterfactual is the outcome that would have happened if circumstances had been different. In causal inference, this means asking: what would have happened to this person if they had received a different treatment?

This idea is the foundation of causal reasoning. You can't measure a treatment's effect without having something to compare it to, and that comparison is the counterfactual.

Counterfactuals vs observed outcomes

- Observed outcomes are what actually happened and got recorded. If a patient took the drug, their observed outcome is their health after taking it.
- Counterfactual outcomes are what would have happened under a different condition. For that same patient, the counterfactual is their health if they had not taken the drug.
- The causal effect for that individual is the difference between these two quantities. The catch: you can only ever observe one of them.

Counterfactuals in causal inference

Counterfactuals let you define causal effects precisely. Rather than vaguely saying "the treatment helped," you can specify exactly what comparison you're making: the outcome under treatment versus the outcome under control, for the same unit.

Even though you can never directly observe both outcomes for a single person, the counterfactual framing gives you a clear target. All the estimation methods you'll encounter (randomization, matching, etc.) are strategies for getting at this fundamentally unobservable comparison.

Potential outcomes framework

The potential outcomes framework takes the intuition behind counterfactuals and makes it mathematically precise. It gives you notation, definitions, and assumptions that let you move from "what if" reasoning to actual estimation of causal effects.

This framework is used across statistics, economics, epidemiology, and political science. It's sometimes called the Rubin causal model (more on that below).

Defining potential outcomes

For each individual i , you define a potential outcome for every possible treatment level. With a simple binary treatment (treated vs. control):

- $Y_i(1)$ = the outcome individual i would have under treatment
- $Y_i(0)$ = the outcome individual i would have under control

These potential outcomes exist conceptually for every person, regardless of which treatment they actually receive. The notation $Y_i(t)$ generalizes this to any treatment level t .

Potential outcomes vs observed outcomes

The observed outcome for individual i is just one of their potential outcomes, determined by which treatment they actually got. If individual i receives treatment ($T_i = 1$), you observe $Y_i(1)$ but not $Y_i(0)$. If they're in the control group ($T_i = 0$), you observe $Y_i(0)$ but not $Y_i(1)$.

You can write this compactly as:

$$Y_i^{obs} = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$$

The unobserved potential outcome is the counterfactual. This "missing data" structure is what makes causal inference hard.

Assumptions of the potential outcomes framework

The framework relies on several assumptions to connect potential outcomes to what you can actually estimate:

1. SUTVA (Stable Unit Treatment Value Assumption): Each person's potential outcomes depend only on their own treatment assignment, not on anyone else's. There's also only one version of each treatment level. (More detail in the SUTVA section below.)
2. Consistency: The observed outcome for someone who received treatment t equals their potential outcome $Y_i(t)$. In other words, the potential outcome notation actually maps onto reality.
3. No interference: One person's treatment doesn't affect another person's outcomes. (This is often folded into SUTVA.)

Individual-level causal effects

The most intuitive notion of a causal effect is at the individual level: how much did this specific treatment change this specific person's outcome?

Definition of individual-level causal effects

The individual causal effect for person i is:

$$\tau_i = Y_i(1) - Y_i(0)$$

This is simply the difference between what happens to person i under treatment and what happens under control. Notice that τ_i can vary from person to person. A drug might help some patients a lot and others not at all.

Fundamental problem of causal inference

Here's the core difficulty: you can never observe both $Y_i(1)$ and $Y_i(0)$ for the same person at the same time. Once someone takes the drug, you can't rewind time and see what would have happened without it.

This is called the fundamental problem of causal inference (a term coined by Paul Holland). For every individual, at least one potential outcome is always missing. Individual-level causal effects are therefore never directly observable.

Estimating individual-level causal effects

Because of the fundamental problem, you can't pin down τ_i for any specific person. But you can estimate average effects across groups of people. The key insight: if you randomly assign treatment, the treatment and control groups will be comparable on average, so the difference in group means estimates the average causal effect. This shifts the goal from individual effects to population-level summaries like the ATE.

Average treatment effect (ATE)

Since individual causal effects are unobservable, the average treatment effect (ATE) is the most common estimand in causal inference. It tells you the expected causal effect of the treatment across an entire population.

Definition of ATE

The ATE is defined as:

$$ATE = E[Y_i(1) - Y_i(0)]$$

For a finite population of N individuals:

$$ATE = \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0))$$

This is the average of all individual-level causal effects. It answers: if you could apply the treatment to everyone and also withhold it from everyone, what would the average difference in outcomes be?

ATE vs individual-level causal effects

The ATE is a summary. It smooths over individual variation. Two treatments could have the same ATE but very different distributions of individual effects. One treatment might help everyone a little; another might help half the population a lot and hurt the other half.

The practical advantage of the ATE is that it's estimable. You don't need to observe both potential outcomes for any single person. You just need groups that are comparable on average.

Estimating ATE from potential outcomes

- Randomized experiments are the most straightforward approach. Random assignment makes treatment independent of potential outcomes, so the difference in group means is an unbiased estimator of the ATE.
- Observational studies require additional assumptions and adjustments. Methods like propensity score weighting, matching, and regression can be used to control for confounders and approximate what randomization would have achieved.

Stable unit treatment value assumption (SUTVA)

SUTVA is one of the most important (and most easily violated) assumptions in the potential outcomes framework. Without it, potential outcomes aren't even well-defined.

Definition of SUTVA

SUTVA has two components:

1. No interference: Person i 's potential outcomes depend only on their own treatment, not on whether person j was treated. Your outcome under treatment is the same regardless of how many other people around you are also treated.
2. No hidden variations of treatment: There's only one version of "treatment" and one version of "control." If different people receive slightly different forms of the treatment, the potential outcomes become ambiguous.

Together, these ensure that $Y_i(t)$ is a single, well-defined quantity for each person and treatment level.

Importance of SUTVA in causal inference

When SUTVA holds, you can cleanly link observed outcomes to potential outcomes and estimate causal effects. If it fails, the potential outcomes themselves become ill-defined or depend on the full vector of everyone's treatment assignments, which makes estimation far more complex.

Violations of SUTVA

SUTVA violations are common in practice. Recognizing them matters for study design and interpretation.

- Interference (spillover effects): A vaccinated person reduces disease transmission to their unvaccinated neighbor, so the neighbor's outcome depends on the vaccinated person's treatment. Similarly, in an education study, a tutored student might share what they learned with classmates.
- Treatment variation: If "the treatment" is a job training program but different sites run it differently, then there isn't a single $Y_i(1)$. The potential outcome depends on which version of the program the person receives.

When you suspect SUTVA violations, you may need to redefine the unit of analysis (e.g., treat a whole classroom as one unit) or explicitly model the interference structure.

Rubin causal model

The Rubin causal model (RCM), developed by Donald Rubin, is the formal framework that organizes potential outcomes, treatment assignment mechanisms, and assumptions into a coherent system for causal inference.

Overview of Rubin causal model

The RCM defines causal effects as comparisons of potential outcomes and places the treatment assignment mechanism at the center of the analysis. The core question is always: how was treatment assigned, and does that assignment process allow you to identify causal effects?

Potential outcomes in Rubin causal model

In the RCM, potential outcomes are treated as fixed (non-random) attributes of each individual. What's random is the treatment assignment. The observed outcome is then determined by which treatment the person happened to receive. This perspective makes it clear that causal inference is fundamentally a missing data problem.

Assumptions of Rubin causal model

The RCM requires three assumptions for identification of causal effects:

1. SUTVA: No interference between units and a single version of each treatment.

2. Ignorability (unconfoundedness): Conditional on observed covariates X , treatment assignment is independent of potential outcomes: $T_i \perp (Y_i(1), Y_i(0)) \mid X_i$. In a randomized experiment, this holds by design. In observational studies, it's an assumption you need to argue for.
3. Positivity (overlap): Every individual has a nonzero probability of receiving each treatment level: $0 < P(T_i = 1 \mid X_i) < 1$. If some subgroup always or never receives treatment, you can't estimate causal effects for that subgroup.

Neyman-Rubin causal model

The Neyman-Rubin causal model integrates Jerzy Neyman's randomization-based inference with Rubin's potential outcomes framework. It's especially relevant for analyzing randomized experiments.

Comparison to Rubin causal model

The Neyman-Rubin model shares the same potential outcomes setup but emphasizes randomization as the basis for inference rather than relying on modeling assumptions. Where the broader Rubin framework accommodates observational studies through ignorability assumptions, the Neyman-Rubin approach draws its inferential power directly from the known randomization procedure.

It also focuses on finite population inference, meaning it targets the ATE for the specific group of individuals in the study, not a hypothetical superpopulation.

Randomization in Neyman-Rubin causal model

Randomization does the heavy lifting here. When you randomly assign treatment:

- Treatment is guaranteed to be independent of potential outcomes (ignorability holds by design)
- You don't need to model the outcome or control for confounders
- The randomization distribution itself provides the basis for hypothesis tests and confidence intervals

This is why randomized experiments are considered the gold standard for causal inference.

Estimating causal effects in Neyman-Rubin model

The standard estimator is the difference in sample means:

$$\hat{ATE} = \bar{Y}_1 - \bar{Y}_0$$

where $\bar{Y}_1$ is the average outcome in the treatment group and $\bar{Y}_0$ is the average outcome in the control group. Under random assignment, this is an unbiased estimator of the ATE.

To quantify uncertainty, you can use:

- Randomization inference (permutation tests): Assess what the difference in means would look like under the null hypothesis by considering all possible random assignments
- Neyman's conservative variance estimator: Provides standard errors that account for sampling variability, though they tend to be slightly conservative because individual treatment effects are unobservable

Applications of counterfactuals

The potential outcomes framework isn't just theoretical. It structures how researchers across many fields design studies and evaluate interventions.

Counterfactuals in policy evaluation

Policy evaluation asks: did this policy cause the observed change in outcomes? The counterfactual is what would have happened without the policy. For example, to evaluate a minimum wage increase on employment, you'd compare employment levels after the increase to what employment would have been without it. Methods like difference-in-differences and regression discontinuity are designed to construct credible counterfactuals in these settings.

Counterfactuals in medical research

Randomized controlled trials (RCTs) are the standard in medicine precisely because they create clean counterfactual comparisons. Patients are randomly assigned to treatment or placebo, so the placebo group's outcomes serve as the counterfactual for the treatment group. When RCTs aren't feasible (due to ethical or practical constraints), observational methods grounded in the potential outcomes framework are used instead.

Counterfactuals in social sciences

Social scientists use counterfactual reasoning to evaluate programs like job training, early childhood education, or crime prevention strategies. The challenge is that random assignment is often difficult in these contexts, so researchers rely on quasi-experimental designs (instrumental variables, natural experiments, matching) to approximate the counterfactual comparison that a randomized experiment would provide.

2.2 Average treatment effect (ATE)

Average Treatment Effect (ATE) is a key concept in causal inference that measures the average difference in outcomes between treatment and control groups. It provides an estimate of the causal effect of an intervention on an outcome, averaged across all individuals in a population.

Estimating ATE requires meeting certain assumptions, including SUTVA, positivity, and unconfoundedness. Various methods can be used to estimate ATE, such as randomized experiments, observational studies, and propensity score methods. Interpreting ATE estimates requires careful consideration of study design, assumptions, and limitations.

Definition of ATE

- The Average Treatment Effect (ATE) is a key concept in causal inference that measures the average difference in outcomes between treatment and control groups in a population
- ATE provides an estimate of the causal effect of a treatment or intervention on an outcome of interest, averaged across all individuals in a population
- Estimating ATE is crucial for making informed decisions about the effectiveness of treatments, policies, or interventions

Potential outcomes framework

- The potential outcomes framework is a conceptual framework for defining and estimating causal effects, including ATE
- Under this framework, each individual has two potential outcomes: one under treatment and one under control

- The individual-level treatment effect is the difference between the potential outcomes, but only one potential outcome is observed for each individual

Treatment vs control groups

- In the context of ATE, the population is divided into two groups: the treatment group, which receives the intervention or treatment of interest, and the control group, which does not receive the treatment
- The treatment group is exposed to the intervention (medication, policy, or program), while the control group serves as a comparison or baseline
- ATE is estimated by comparing the average outcomes between the treatment and control groups, after accounting for any differences in baseline characteristics

Assumptions for estimating ATE

- To obtain unbiased estimates of ATE, certain assumptions must be met, ensuring that the treatment and control groups are comparable and that the treatment effect is well-defined
- Violations of these assumptions can lead to biased estimates and incorrect conclusions about the causal effect of the treatment
- Carefully assessing and addressing these assumptions is crucial for valid causal inference and accurate estimation of ATE

Stable Unit Treatment Value Assumption (SUTVA)

- SUTVA assumes that the potential outcomes for each individual are unaffected by the treatment assignment of other individuals (no interference)
- It also assumes that there are no different versions of the treatment (consistency), meaning that the treatment is well-defined and applied uniformly across individuals
- Violations of SUTVA can occur when there are spillover effects (one individual's treatment affects another's outcome) or when there are multiple versions of the treatment

Positivity assumption

- The positivity assumption, also known as the overlap assumption, requires that each individual has a non-zero probability of being assigned to either the treatment or control group
- This assumption ensures that there are comparable individuals in both groups, allowing for the estimation of the treatment effect across the entire population
- Violations of positivity can occur when there are certain subgroups of individuals who have zero probability of receiving the treatment (lack of overlap in covariate distributions)

Unconfoundedness assumption

- The unconfoundedness assumption, also known as the ignorability assumption, states that treatment assignment is independent of the potential outcomes, given a set of observed covariates
- This assumption implies that all confounding factors that influence both treatment assignment and the outcome are measured and accounted for in the analysis
- Violations of unconfoundedness can occur when there are unmeasured confounders that affect both treatment assignment and the outcome, leading to biased estimates of ATE

Methods for estimating ATE

- Various methods can be used to estimate ATE, depending on the study design, data availability, and the plausibility of the assumptions
- These methods aim to create comparable treatment and control groups, either through randomization or by adjusting for confounding factors using statistical techniques
- The choice of method depends on the specific context and the researcher's ability to meet the necessary assumptions for valid causal inference

Randomized experiments

- Randomized controlled trials (RCTs) are considered the gold standard for estimating ATE, as they ensure that treatment assignment is independent of potential outcomes
- In an RCT, individuals are randomly assigned to either the treatment or control group, creating balanced groups that are comparable in terms of both observed and unobserved factors
- The ATE can be estimated by simply comparing the average outcomes between the treatment and control groups, as randomization ensures unconfoundedness

Observational studies

- In observational studies, treatment assignment is not randomized, and individuals self-select into treatment or control groups based on their characteristics
- To estimate ATE in observational studies, researchers must account for confounding factors that may influence both treatment assignment and outcomes
- Common methods for adjusting for confounding in observational studies include regression adjustment, matching, and inverse probability weighting

Propensity score methods

- Propensity score methods are a class of techniques used to estimate ATE in observational studies by balancing treatment and control groups based on their likelihood of receiving the treatment
- The propensity score is the probability of receiving the treatment given a set of observed covariates, estimated using a logistic regression model
- Methods such as propensity score matching, stratification, and weighting can be used to create balanced groups and estimate ATE, assuming unconfoundedness and positivity hold

Interpreting ATE estimates

- Once ATE is estimated, it is important to interpret the results correctly and understand their implications for the population of interest
- ATE estimates should be interpreted in the context of the study design, the assumptions made, and the limitations of the data and methods used
- Careful interpretation of ATE estimates is crucial for making informed decisions and drawing valid conclusions about the causal effect of the treatment

Population-level effects

- ATE estimates represent the average causal effect of the treatment on the outcome across the entire population of interest
- They provide a summary measure of the overall impact of the treatment, which can be useful for policy decisions and resource allocation
- However, population-level ATE estimates may mask important heterogeneity in treatment effects across subgroups

Subgroup analysis

- Subgroup analysis involves estimating ATE for specific subpopulations defined by characteristics such as age, gender, or baseline risk factors
- Examining treatment effects in subgroups can reveal important heterogeneity in the impact of the treatment and identify subpopulations that may benefit more or less from the intervention
- However, subgroup analyses should be prespecified and interpreted with caution, as multiple testing and post-hoc analyses can lead to false-positive findings

Limitations and caveats

- ATE estimates are subject to various limitations and caveats that should be considered when interpreting the results
- Estimates may be sensitive to violations of the underlying assumptions (SUTVA, positivity, unconfoundedness), leading to biased results
- The generalizability of ATE estimates may be limited by the characteristics of the study population and the specific definition of the treatment and outcome variables
- Confidence intervals and statistical significance should be reported alongside ATE estimates to communicate the uncertainty associated with the estimates

Relationship to other causal effects

- ATE is one of several causal effect measures used in the field of causal inference, each with its own interpretation and assumptions
- Understanding the relationships and differences between ATE and other causal effects is important for selecting the appropriate measure for a given research question and interpreting the results
- Comparing ATE to other causal effects can provide a more comprehensive understanding of the impact of a treatment or intervention

Comparison to average treatment effect on the treated (ATT)

- The average treatment effect on the treated (ATT) measures the average causal effect of the treatment among those who actually received the treatment
- ATT is often of interest when evaluating the effectiveness of a program or policy on those who participated in it
- Unlike ATE, which estimates the effect across the entire population, ATT focuses specifically on the treated subpopulation and may be more relevant for decision-making in some contexts

Comparison to local average treatment effect (LATE)

- The local average treatment effect (LATE) measures the average causal effect of the treatment among the subpopulation of "compliers," who would receive the treatment if assigned to the treatment group and not receive it if assigned to the control group
- LATE is often estimated using instrumental variable methods, which rely on the existence of a variable that affects treatment assignment but not the outcome directly
- LATE may be more relevant than ATE when there is non-compliance or when the treatment effect is expected to vary across different subpopulations

Applications of ATE

- ATE has numerous applications across various fields, including economics, public health, social sciences, and policy evaluation
- Estimating ATE is crucial for understanding the impact of interventions, programs, and policies on outcomes of interest
- Examples of applications include evaluating the effectiveness of medical treatments, assessing the impact of educational programs, and informing policy decisions

Policy evaluation

- ATE is often used to evaluate the impact of policies or programs implemented by governments or organizations
- By comparing outcomes between those exposed to the policy and those not exposed, policymakers can assess the effectiveness of the intervention and make informed decisions about continuing, modifying, or discontinuing the policy
- Examples include evaluating the impact of a job training program on employment rates or the effect of a tax credit on household income

Program impact assessment

- ATE can be used to assess the impact of social programs or interventions designed to improve outcomes for specific populations
- Estimating ATE allows program administrators to determine whether the program is achieving its intended goals and to identify areas for improvement
- Examples include evaluating the impact of a health education program on disease prevention or the effect of a mentoring program on academic achievement

Decision-making in healthcare and social sciences

- ATE is widely used in healthcare and social sciences to inform decision-making about treatments, interventions, and resource allocation
- In healthcare, ATE estimates can help clinicians and patients make informed decisions about the most effective treatments for a given condition
- In social sciences, ATE can be used to evaluate the impact of interventions on outcomes such as mental health, social well-being, or economic stability

Challenges in estimating ATE

- Estimating ATE in real-world settings can be challenging due to various factors that can introduce bias or violate the assumptions required for valid causal inference
- Researchers must be aware of these challenges and take steps to address them through appropriate study design, data collection, and statistical methods
- Failure to address these challenges can lead to biased estimates and incorrect conclusions about the causal effect of the treatment

Selection bias

- Selection bias occurs when the treatment and control groups differ systematically in ways that are related to the outcome, even after adjusting for observed confounders
- This can happen when individuals self-select into treatment or when treatment assignment is influenced by factors that are not fully captured in the data
- Selection bias can lead to biased estimates of ATE, as the observed differences in outcomes may be due to preexisting differences between the groups rather than the causal effect of the treatment

Unmeasured confounding

- Unmeasured confounding occurs when there are factors that influence both treatment assignment and the outcome but are not measured or accounted for in the analysis
- This violates the unconfoundedness assumption and can lead to biased estimates of ATE, as the observed differences in outcomes may be due to the unmeasured confounders rather than the treatment
- Addressing unmeasured confounding requires careful study design, the use of sensitivity analyses, or the application of advanced statistical methods such as instrumental variables or difference-in-differences

Violations of assumptions

- Violations of the key assumptions required for estimating ATE (SUTVA, positivity, and unconfoundedness) can lead to biased and unreliable estimates
- SUTVA violations can occur when there are spillover effects or different versions of the treatment, while positivity violations can occur when there are subgroups with zero probability of receiving the treatment
- Unconfoundedness violations can occur when there are unmeasured confounders that affect both treatment assignment and the outcome
- Researchers must carefully assess the plausibility of these assumptions in their specific context and take steps to address any violations through appropriate study design and statistical methods

Advanced topics in ATE estimation

- As the field of causal inference continues to evolve, researchers are developing advanced methods for estimating ATE that can address some of the challenges and limitations of traditional approaches

- These advanced topics aim to improve the robustness, efficiency, and flexibility of ATE estimation, particularly in the presence of complex data structures or violations of assumptions
- Familiarity with these advanced topics can help researchers select the most appropriate methods for their specific research questions and data contexts

Doubly robust estimation

- Doubly robust estimation combines two models: a model for the outcome and a model for the treatment assignment mechanism (propensity score)
- The doubly robust estimator is consistent if either the outcome model or the propensity score model is correctly specified, providing a safeguard against model misspecification
- Doubly robust methods can improve the efficiency and robustness of ATE estimates, particularly in the presence of high-dimensional confounders or complex data structures

Machine learning approaches

- Machine learning techniques, such as decision trees, random forests, and neural networks, can be used to estimate ATE by flexibly modeling the relationship between confounders, treatment, and outcomes
- These approaches can automatically detect and account for complex interactions and nonlinearities in the data, which may be difficult to specify in traditional parametric models
- Machine learning methods can be particularly useful when dealing with high-dimensional data or when the functional form of the relationships between variables is unknown

Sensitivity analysis for unobserved confounding

- Sensitivity analysis is a technique used to assess the robustness of ATE estimates to potential unmeasured confounding
- By simulating the presence of an unmeasured confounder and examining how much the confounder would need to influence treatment assignment and outcomes to alter the conclusions, researchers can gauge the sensitivity of their results
- Sensitivity analysis can help determine the plausibility of the unconfoundedness assumption and provide a range of possible ATE estimates under different scenarios of unmeasured confounding
- Examples of sensitivity analysis methods include the Rosenbaum bounds approach and the use of instrumental variables to assess the potential impact of unmeasured confounders

2.3 Conditional average treatment effect (CATE)

Conditional average treatment effect (CATE)

The conditional average treatment effect (CATE) estimates how a treatment effect differs across subgroups defined by observed characteristics. While the ATE gives you one number summarizing the treatment effect for an entire population, CATE answers a more targeted question: what is the treatment effect for individuals with these specific characteristics?

This matters because treatments rarely work the same way for everyone. A drug might help older patients more than younger ones. A marketing campaign might convert urban customers but not rural ones. CATE gives you the tools to detect and quantify that kind of variation.

Definition of CATE

CATE is the expected difference in potential outcomes for individuals who share a particular set of covariate values. Formally:

$$\tau(x) = E[Y(1) - Y(0) \mid X = x]$$

where $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control, and X represents observed covariates (age, gender, medical history, etc.).

The function $\tau(x)$ maps each covariate profile x to a treatment effect. When $\tau(x)$ is the same for all x , the treatment effect is homogeneous and CATE collapses to the ATE. When $\tau(x)$ varies, you have heterogeneous treatment effects.

CATE vs. ATE

The ATE averages over the entire population:

$$ATE = E[Y(1) - Y(0)]$$

You can think of the ATE as the average of CATE across the covariate distribution: $ATE = E[\tau(x)]$. So the ATE is a single summary, while CATE is a function that tells you how the effect changes with x .

This distinction has practical consequences. If a treatment has an ATE of zero, that could mean it helps no one, or it could mean it helps some subgroups and hurts others in ways that cancel out. Only CATE can distinguish between these two very different scenarios.

Heterogeneous treatment effects

Heterogeneous treatment effects means the treatment effect varies across individuals or subgroups. Recognizing this heterogeneity is what motivates CATE estimation in the first place.

- A blood pressure drug might reduce systolic pressure by 15 mmHg in patients over 60 but only 3 mmHg in patients under 30.
- A job training program might raise earnings substantially for workers without a college degree but have little effect on those who already have one.

Identifying which subgroups benefit most (or are harmed) allows for better resource allocation and more personalized decision-making.

Estimating CATE

Estimating CATE is harder than estimating the ATE because you need enough data within each subgroup to get reliable effect estimates. Two broad families of methods are commonly used.

Regression methods for CATE

The simplest approach uses regression with interaction terms between the treatment indicator and covariates.

For example, with a single covariate x :

$$Y = \beta_0 + \beta_1 T + \beta_2 x + \beta_3 (T \times x) + \varepsilon$$

Here $\beta_1 + \beta_3 x$ gives the estimated treatment effect at covariate value x . The interaction term β_3 captures how the effect changes with x .

This extends to generalized linear models and multilevel models. The main limitation is that you need to specify which interactions to include, and linear models may miss complex nonlinear patterns in $\tau(x)$.

Machine learning approaches to CATE

ML methods can discover heterogeneity patterns without requiring you to pre-specify interaction terms. Common approaches include:

- Causal forests (a variant of random forests designed specifically for treatment effect estimation)
- Bayesian additive regression trees (BART)
- Meta-learners (S-learner, T-learner, X-learner), which use standard ML models as building blocks for CATE estimation

These methods handle high-dimensional covariates and complex interactions well, but they require careful tuning and validation to avoid overfitting, especially when sample sizes are modest.

Challenges in estimating CATE

- Limited subgroup sample sizes. Splitting data by covariates reduces the effective sample size for each estimate, increasing variance.
- High-dimensional covariates. With many potential effect modifiers, the risk of finding spurious heterogeneity grows.
- Unmeasured confounders. If confounding differs across subgroups, CATE estimates can be biased even when the overall ATE estimate is not.
- Model selection. Choosing between methods (and validating results) is harder for CATE than for ATE, since you never observe both potential outcomes for the same individual. Cross-validation strategies need to be adapted for causal quantities.

Applications of CATE

Personalized medicine

CATE estimation is central to precision medicine. By estimating treatment effects conditional on a patient's genetic profile, comorbidities, and demographics, clinicians can recommend the treatment most likely to benefit that specific patient rather than relying on population-average evidence.

Targeted marketing campaigns

Marketers use CATE to estimate the uplift of a campaign for different customer segments. Rather than sending a promotion to everyone, you target the subgroup where the estimated CATE (incremental conversion probability) is highest, improving return on investment.

Policy evaluation and optimization

Policymakers can use CATE to determine which populations benefit most from an intervention. For example, estimating CATE for a job training program across education levels and regions can guide where to expand the program and where to reallocate resources.

Assumptions and limitations

CATE estimation inherits the core identification assumptions of the potential outcomes framework, and violations are especially consequential because they can distort the pattern of heterogeneity, not just the overall level.

Unconfoundedness assumption

Also called conditional ignorability, this assumption states:

$$Y(1), Y(0) \perp\!\!\!\perp T \mid X$$

In words: once you condition on the observed covariates X , treatment assignment T is independent of the potential outcomes. This means there are no unmeasured confounders lurking behind the relationship between treatment and outcome.

If this assumption fails, your CATE estimates will be biased. Worse, the bias can vary across subgroups, potentially making you think a treatment helps one group when it actually doesn't.

Overlap assumption

Also called positivity, this requires:

$$0 < P(T = 1 \mid X = x) < 1 \quad \text{for all } x$$

Every covariate profile must have some probability of receiving treatment and some probability of receiving control. If certain subgroups almost always (or never) receive treatment, you're essentially extrapolating rather than estimating, and CATE estimates in those regions become unreliable.

Limitations of CATE estimates

- Finite sample bias. With limited data per subgroup, estimates can be noisy or systematically off.
- Model misspecification. Parametric models may impose the wrong functional form on $\tau(X)$; flexible models may overfit.
- Validation difficulty. You can't directly validate individual-level treatment effects since the fundamental problem of causal inference (you only observe one potential outcome per unit) still applies.
- Always interpret CATE estimates alongside sensitivity analyses and cross-validation results.

Advanced topics in CATE

Efficient estimation of CATE

Efficient estimators aim to minimize mean squared error by balancing bias and variance. Key approaches include:

- Doubly robust (DR) estimation combines an outcome model and a propensity score model. If either model is correctly specified, the estimator remains consistent.
- Targeted maximum likelihood estimation (TMLE) uses a targeted bias-reduction step on top of an initial outcome model.
- Efficient influence function-based estimators achieve the semiparametric efficiency bound under certain regularity conditions.

These methods are more robust to partial model misspecification than approaches that rely on a single model.

Nonparametric estimation of CATE

Nonparametric methods avoid strong assumptions about the functional form of $\tau(x)$:

- Kernel regression and local linear regression estimate the treatment effect locally around each point in covariate space.
- Spline-based methods fit flexible smooth functions.

These approaches can capture complex nonlinear heterogeneity but typically require larger sample sizes and can be computationally expensive in high dimensions.

CATE with instrumental variables

When unconfoundedness fails, instrumental variables (IVs) offer an alternative identification strategy. An instrument Z must affect treatment assignment but have no direct effect on the outcome except through treatment.

- Two-stage least squares (2SLS) can be extended with interactions to estimate heterogeneous effects.
- The local average treatment effect (LATE) identifies the effect for compliers, and extensions estimate how LATE varies with covariates.
- Marginal treatment effect (MTE) methods trace out how the treatment effect varies with individuals' unobserved propensity to select into treatment, providing a richer picture of heterogeneity.

CATE in longitudinal settings

When treatments, covariates, and outcomes evolve over time, standard cross-sectional CATE methods can break down due to time-dependent confounding (where past treatment affects future covariates, which in turn affect future treatment).

Methods designed for this setting include:

- Marginal structural models (MSMs) with inverse probability of treatment weighting
- Structural nested mean models (SNMMs) that directly model the causal effect of treatment at each time point
- G-computation, which models the full sequence of conditional distributions to estimate counterfactual outcomes

Each of these accounts for the dynamic interplay between treatment and confounders over time, though they come with their own assumptions and data requirements.

2.4 Assumptions: Stable Unit Treatment Value Assumption (SUTVA) and consistency

Definition of SUTVA

The Stable Unit Treatment Value Assumption (SUTVA) is what makes causal effects well-defined in the potential outcomes framework. Without it, you can't cleanly write down what $Y_i(1)$ and $Y_i(0)$ even mean, because the potential outcomes could shift depending on what happens to other units or how the treatment gets delivered.

SUTVA has two distinct components:

1. No interference between units — one unit's treatment assignment doesn't affect another unit's outcomes.
2. No hidden variations in treatment — there's only one version of each treatment level, so "treated" means the same thing for every treated unit.

Both components must hold for the potential outcomes to be stable, well-defined quantities that you can actually estimate.

Potential Outcomes Framework Connection

SUTVA is baked into the notation of the potential outcomes framework itself. When you write $Y_i(t)$ for unit i under treatment t , you're implicitly assuming that this value is a fixed quantity. SUTVA is what justifies that assumption. If other units' treatments could change $Y_i(t)$, or if treatment t could mean different things for different units, then $Y_i(t)$ isn't a single number — it's a moving target.

No Interference Between Units

This component states that unit i 's potential outcomes depend only on unit i 's own treatment assignment, not on the treatments assigned to units $j, k, l, \dots$

Formally: $Y_i(t_1, t_2, \dots, t_n) = Y_i(t_i)$ for all possible treatment vectors. The potential outcome for unit i is a function of t_i alone.

This rules out spillover effects — situations where treating one unit changes outcomes for others. Common violations include:

- Peer effects: A job training program changes labor market conditions for non-participants by altering labor supply.
- Spatial spillovers: An environmental policy in one region affects pollution levels in neighboring regions.
- Contagion: Vaccinating some people in a community reduces disease exposure for unvaccinated people (herd immunity).

No Hidden Variations in Treatment

This component states that there is only one "version" of each treatment level. If unit i and unit j are both assigned $T = 1$, they receive the same treatment in all relevant respects.

Violations arise when the treatment label masks meaningful differences in what units actually receive:

- Dosage differences: Two patients both "receive the drug" but at different doses.

- Quality of delivery: Two classrooms both "receive the new curriculum" but with teachers of vastly different skill levels.
- Adherence variation: Two participants are both "assigned to exercise" but one exercises 5 days a week and the other exercises 1 day.

When this component fails, $Y_i(1)$ is ambiguous because "treatment = 1" doesn't correspond to a single, well-defined intervention.

Consistency Assumption

Consistency bridges the gap between the theoretical potential outcomes and the data you actually observe. It says: if unit i receives treatment t , then the outcome you observe for unit i equals the potential outcome under that treatment.

Formal Definition

If $T_i = t$, then $Y_i = Y_i(t)$

where Y_i is the observed outcome, $Y_i(t)$ is the potential outcome under treatment t , and T_i is the treatment actually received.

This might seem obvious, but it's doing real work. Consistency says the potential outcome $Y_i(t)$ is the same outcome you'd observe in the real world when unit i gets treatment t . If the treatment is vaguely defined or has multiple versions, this link breaks down.

Connection to SUTVA

Consistency and SUTVA are tightly linked. SUTVA's "no hidden treatment variations" component is essentially what makes consistency hold. If there are multiple versions of treatment t , then $Y_i(t)$ isn't a single value, and the equation $Y_i = Y_i(t)$ when $T_i = t$ becomes meaningless — which version of $Y_i(t)$ are you referring to?

Some textbooks fold consistency into SUTVA; others treat them as separate assumptions. Either way, the logic is the same: you need well-defined treatments and no interference for the observed data to tell you about potential outcomes.

Consistency vs. Exchangeability

These are different assumptions that address different problems:

- Consistency concerns the relationship between observed outcomes and potential outcomes. It asks: does the data I see correspond to the potential outcomes I've defined?
- Exchangeability concerns the relationship between treatment assignment and potential outcomes. It asks: are treated and untreated groups comparable, so that differences in observed outcomes reflect causal effects rather than confounding?

You need both for valid causal inference, but they can fail independently. A study could have perfect consistency (well-defined treatments, no interference) but terrible exchangeability (treatment assignment driven by confounders). Or it could have exchangeability through randomization but poor consistency because the treatment varies in implementation.

SUTVA Violations and Their Consequences

What Goes Wrong

When SUTVA or consistency fails, the core machinery of causal inference breaks down in specific ways:

- Biased effect estimates: If interference exists, the observed outcome for a control unit may already be influenced by treated units nearby. The treatment effect estimate gets contaminated.
- Ambiguous estimands: If treatment has multiple versions, the "average treatment effect" is an average over an undefined mixture of different interventions. You don't know what you're estimating.
- Misleading policy conclusions: If spillover effects inflate (or deflate) your estimate, scaling up the intervention could produce very different results than your study predicted.

Concrete Examples

Interference example: You randomize villages to receive a microfinance program. But treated villages trade with control villages, and the economic activity spills over. Control village outcomes improve partly because of the treatment in neighboring villages. Your estimated effect underestimates the true direct effect of the program.

Treatment variation example: You study the effect of "surgery vs. no surgery" for a condition, but surgical technique varies widely across hospitals. The causal effect you estimate is a blend of many different surgical procedures, making it hard to know what specific intervention you're recommending.

Strategies for Addressing Violations

Redefining Treatment or Units

Sometimes you can fix a SUTVA violation by changing what counts as a "unit" or a "treatment":

- If interference happens within households, analyze at the household level instead of the individual level.
- If interference happens within classrooms, randomize at the classroom or school level (cluster randomization).
- If treatment varies in intensity, replace a binary treatment indicator with a more granular measure (e.g., hours of tutoring rather than tutored/not tutored).

The goal is to redefine things so that the redefined treatment is truly uniform and the redefined units don't interfere with each other.

Modeling Interference Explicitly

When interference is unavoidable, you can build it into the model rather than assuming it away. This requires specifying the structure of interference:

- Partial interference: Units interact within known clusters (e.g., classrooms, villages) but not across clusters. You can define potential outcomes as functions of both a unit's own treatment and the proportion of treated units in its cluster.
- Network models: If you know the social network, you can model outcomes as depending on a unit's own treatment and the treatments of its network neighbors.

These approaches require stronger assumptions about how interference works, but they're more honest than pretending it doesn't exist.

Sensitivity Analysis

When you can't rule out violations or model them directly, sensitivity analysis asks: how bad would the violation need to be to change my conclusions?

You vary assumptions about the magnitude of interference or treatment variation and see how the estimated effects shift. If your conclusions hold across a wide range of plausible violations, that's reassuring. If a small amount of interference could flip your result, that's a red flag.

SUTVA and Consistency in Practice

Assessing Plausibility

Before running any analysis, think carefully about whether SUTVA and consistency are reasonable for your setting:

- Could units interact? Consider physical proximity, social networks, shared markets, or shared resources.
- Is the treatment truly uniform? Consider whether implementation quality, dosage, timing, or adherence varies across units.
- Is the treatment well-defined enough? If you can't precisely describe what "receiving treatment" means, consistency is likely shaky.

These assessments should draw on subject-matter knowledge, not just statistical diagnostics.

Role Across Causal Methods

SUTVA and consistency aren't specific to one method. They're required (at least implicitly) by randomized experiments, propensity score methods, instrumental variables, regression discontinuity, and difference-in-differences. No amount of clever identification strategy can rescue you if the potential outcomes themselves are ill-defined.

Limitations and Trade-offs

Addressing violations always involves trade-offs:

- Redefining units to a coarser level (e.g., clusters) reduces interference concerns but also reduces your sample size and the precision of your estimates.
- Modeling interference explicitly gives you richer results but requires assumptions about the interference structure that may be hard to verify.
- Sensitivity analysis doesn't fix the problem; it just tells you how much the problem matters.

Being transparent about which assumptions you're relying on, and how plausible they are, is more valuable than pretending the assumptions are automatically satisfied.

Unit 3 – Randomized experiments

3.1 Completely randomized designs

Completely randomized designs are a cornerstone of experimental research. They involve randomly assigning units to treatment groups, ensuring each unit has an equal chance of receiving any treatment. This randomization helps balance observed and unobserved covariates across groups, allowing for unbiased estimation of treatment effects.

These designs offer several benefits, including elimination of confounding and balanced covariates on average. By randomly allocating units, researchers can attribute observed differences in outcomes to the treatment itself. This approach strengthens internal validity and enables causal inference, making it a powerful tool in various fields, from clinical trials to social science experiments.

Definition of completely randomized designs

- Completely randomized designs are experimental designs where units are randomly assigned to treatment and control groups
- Involves a single randomization step to allocate units to different treatment conditions
- Ensures that each unit has an equal probability of being assigned to any of the treatment groups
- Provides a basis for making causal inferences about the effect of the treatment on the outcome of interest

Benefits of randomization

- Randomization is a key feature of completely randomized designs that helps to ensure the internal validity of the study
- Allows for the unbiased estimation of treatment effects by balancing both observed and unobserved covariates across treatment groups on average

Elimination of confounding

- Randomization helps to eliminate confounding by ensuring that treatment assignment is independent of potential outcomes and covariates
- Balances the distribution of potential confounders across treatment groups, reducing the risk of bias in treatment effect estimates
- Enables researchers to attribute observed differences in outcomes between treatment groups to the causal effect of the treatment itself

Balance of covariates on average

- Randomization ensures that, on average, the distribution of covariates is similar across treatment groups
- Balancing of covariates helps to isolate the effect of the treatment from the influence of other factors
- Increases the comparability of treatment groups and strengthens the internal validity of the study
- Allows for the estimation of treatment effects without the need for statistical adjustment for covariates

Mechanics of randomization

- The process of randomization involves assigning units to treatment groups using a random mechanism
- Ensures that each unit has an equal chance of being assigned to any of the treatment conditions
- Can be performed using various methods, such as simple randomization or stratified randomization

Randomization methods

- Simple randomization: Each unit is independently assigned to a treatment group with equal probability
- Stratified randomization: Units are first stratified based on important covariates, and then randomization is performed within each stratum
- Matched-pair randomization: Units are paired based on similarity in covariates, and then one unit from each pair is randomly assigned to each treatment group
- Blocked randomization: Units are divided into blocks of a fixed size, and randomization is performed within each block to ensure balance

Stratified vs unstratified randomization

- Unstratified randomization (simple randomization) assigns units to treatment groups without considering any covariates
- Stratified randomization involves first dividing units into strata based on important covariates and then performing randomization within each stratum
- Stratified randomization can improve the balance of covariates and increase the precision of treatment effect estimates, especially when the sample size is small
- Stratification is particularly useful when there are known prognostic factors that could influence the outcome of interest

Estimating causal effects

- Completely randomized designs allow for the unbiased estimation of causal effects by comparing outcomes between treatment groups
- The average treatment effect (ATE) is a common measure of the causal effect in completely randomized designs
- The difference in means estimator is a simple and unbiased estimator of the ATE

Average treatment effect (ATE)

- The ATE is the expected difference in outcomes between the treatment and control groups
- Represents the average causal effect of the treatment on the outcome of interest across the entire population
- Can be estimated by comparing the mean outcomes of the treatment and control groups in a completely randomized design
- Provides a summary measure of the overall effectiveness of the treatment

Difference in means estimator

- The difference in means estimator is a simple and unbiased estimator of the ATE in completely randomized designs
- Calculated by taking the difference between the sample means of the outcomes in the treatment and control groups
- Assumes that the treatment assignment is independent of potential outcomes and that the sample is representative of the population
- Provides a straightforward way to estimate the causal effect of the treatment on the outcome of interest

Hypothesis testing and p-values

- Hypothesis testing is used to assess the statistical significance of the estimated treatment effect
- Involves specifying a null hypothesis (usually no treatment effect) and an alternative hypothesis (presence of a treatment effect)
- P-values are used to quantify the strength of evidence against the null hypothesis
- A small p-value (typically < 0.05) indicates strong evidence against the null hypothesis and supports the presence of a treatment effect
- Hypothesis testing and p-values help to determine whether the observed difference in outcomes between treatment groups is likely due to chance or a true treatment effect

Statistical inference

- Statistical inference is the process of drawing conclusions about population parameters based on sample data
- In completely randomized designs, statistical inference is used to make statements about the true average treatment effect in the population
- Involves estimating the sampling distribution of the estimator, calculating standard errors, and constructing confidence intervals

Sampling distributions

- The sampling distribution is the probability distribution of an estimator (e.g., difference in means) over repeated random samples from the same population
- Describes the variability and central tendency of the estimator under the null hypothesis of no treatment effect
- The sampling distribution is used to calculate standard errors and construct confidence intervals for the true average treatment effect
- In completely randomized designs, the sampling distribution of the difference in means estimator is approximately normal for large sample sizes (due to the Central Limit Theorem)

Standard errors and confidence intervals

- The standard error is a measure of the variability of an estimator (e.g., difference in means) across different samples
- Calculated as the standard deviation of the sampling distribution of the estimator
- Used to construct confidence intervals for the true average treatment effect
- A 95% confidence interval is a range of values that has a 95% probability of containing the true average treatment effect
- Confidence intervals provide a measure of the precision and uncertainty associated with the estimated treatment effect

Randomization inference

- Randomization inference is a non-parametric method for testing the null hypothesis of no treatment effect in completely randomized designs
- Involves comparing the observed test statistic (e.g., difference in means) to the distribution of test statistics under all possible random assignments of units to treatment groups
- Provides exact p-values and confidence intervals without relying on distributional assumptions
- Particularly useful when the sample size is small or the assumptions of parametric tests are violated
- Preserves the validity of the inference by taking into account the actual randomization process used in the study

Efficiency of completely randomized designs

- The efficiency of a completely randomized design refers to its ability to precisely estimate the average treatment effect
- Efficiency is influenced by various factors, such as sample size, variability of the outcome, and the presence of covariates
- Completely randomized designs can be compared to other experimental designs in terms of their efficiency

Comparison to other designs

- Completely randomized designs are often less efficient than designs that make use of covariates, such as stratified or blocked designs
- Stratified designs can improve efficiency by reducing the variability of the treatment effect estimator within each stratum
- Blocked designs can increase efficiency by removing the variability associated with the blocking factors from the error term
- However, completely randomized designs are simpler to implement and require fewer assumptions about the relationship between covariates and the outcome

Factors affecting efficiency

- Sample size: Larger sample sizes generally lead to more precise estimates of the treatment effect and increased efficiency
- Variability of the outcome: Outcomes with high variability require larger sample sizes to achieve the same level of precision as outcomes with low variability
- Presence of covariates: Adjusting for prognostic covariates can increase the precision of the treatment effect estimator and improve efficiency
- Allocation ratio: The allocation ratio between treatment and control groups can affect efficiency, with equal allocation typically being the most efficient
- Heterogeneity of treatment effects: If the treatment effect varies across subgroups, the overall efficiency of the design may be reduced

Limitations of completely randomized designs

- While completely randomized designs have several advantages, they also have some limitations that researchers should be aware of
- These limitations can affect the internal and external validity of the study and should be considered when interpreting the results

Lack of covariate balance in small samples

- In small samples, completely randomized designs may not achieve adequate balance of covariates between treatment groups by chance alone
- Imbalance of important prognostic factors can lead to biased estimates of the treatment effect and reduced efficiency
- Stratified or blocked randomization can be used to improve covariate balance in small samples
- Alternatively, statistical adjustment methods (e.g., regression) can be used to account for covariate imbalances in the analysis

Ethical considerations

- In some cases, completely randomized designs may not be ethically appropriate, particularly when there is a strong prior belief that one treatment is superior to another
- Withholding a potentially beneficial treatment from some participants through randomization may be considered unethical
- Alternative designs, such as adaptive or preference-based designs, may be more appropriate in these situations
- Researchers should carefully consider the ethical implications of their study design and ensure that the benefits outweigh the risks for all participants

Examples of completely randomized designs

- Completely randomized designs are widely used in various fields, including medicine, psychology, and social sciences

- Some common examples include clinical trials and A/B testing in web design

Clinical trials

- Clinical trials often use completely randomized designs to evaluate the efficacy and safety of new medical interventions (drugs, medical devices)
- Patients are randomly assigned to receive either the new treatment or a control (placebo or standard treatment)
- Randomization helps to ensure that treatment groups are comparable and that observed differences in outcomes can be attributed to the treatment effect
- Example: A randomized controlled trial comparing the effectiveness of a new blood pressure medication to a placebo

A/B testing in web design

- A/B testing is a form of completely randomized design used to compare two versions of a web page or application
- Users are randomly assigned to either the "A" version (control) or the "B" version (treatment) of the web page
- Metrics such as click-through rates, conversion rates, or user engagement are compared between the two versions
- Randomization ensures that any differences in user behavior can be attributed to the design changes made in the "B" version
- Example: An e-commerce website randomly assigns visitors to either a standard product page or a redesigned page with new features to assess the impact on sales

Variations on completely randomized designs

- While the basic completely randomized design is widely used, there are several variations that can be employed to address specific research questions or design considerations
- These variations include blocked randomized designs and factorial designs

Blocked randomized designs

- In blocked randomized designs, units are first divided into blocks based on one or more covariates and then randomized within each block
- Blocking helps to ensure balance of the covariates across treatment groups and can increase the precision of the treatment effect estimator
- Blocks are typically chosen based on factors that are known or suspected to influence the outcome of interest
- Example: In an agricultural experiment, fields may be blocked by soil type before randomly assigning different fertilizer treatments within each block

Factorial designs

- Factorial designs allow researchers to investigate the effects of two or more factors simultaneously

- Each factor has two or more levels, and treatments are formed by combining the levels of the factors
- Completely randomized factorial designs involve randomly assigning units to each combination of factor levels
- Factorial designs enable the estimation of main effects (the effect of each factor averaging over the levels of the other factors) and interaction effects (the extent to which the effect of one factor depends on the level of another factor)
- Example: A 2x2 factorial design comparing the effects of two different drugs (drug A and drug B) and their combination (drug A + drug B) on patient outcomes

Analyzing completely randomized designs

- The analysis of completely randomized designs typically involves comparing the outcomes between treatment groups
- Several statistical methods can be used to estimate treatment effects and assess their significance, including regression analysis and analysis of variance (ANOVA)

Regression analysis

- Regression analysis is a flexible method for estimating treatment effects in completely randomized designs
- Involves fitting a regression model with the outcome as the dependent variable and treatment indicators as independent variables
- Can easily accommodate covariates and assess their influence on the treatment effect
- Provides estimates of the average treatment effect, standard errors, and confidence intervals
- Example: Using linear regression to estimate the effect of a new teaching method on student test scores, controlling for baseline performance and demographic factors

Analysis of variance (ANOVA)

- ANOVA is a common method for analyzing completely randomized designs with a single factor (treatment)
- Partitions the total variability in the outcome into two components: variability between treatment groups and variability within treatment groups
- Tests the null hypothesis of no difference in means between treatment groups using an F-test
- Provides estimates of the average treatment effect and can be extended to include blocking factors (e.g., blocked ANOVA)
- Example: Using one-way ANOVA to compare the mean weight loss achieved by participants in three different diet groups (low-carb, low-fat, and control) in a randomized trial

3.2 Stratified and blocked designs

Stratified and blocked designs are powerful tools in causal inference. They help control confounding variables and reduce variability, leading to more precise treatment effect estimates. These designs divide populations into subgroups based on key characteristics, ensuring balanced representation and increased power to detect effects.

By combining stratification and blocking, researchers can control for both confounding variables and nuisance factors. This approach improves balance, reduces variability, and increases power. However, these designs have limitations, including feasibility constraints, generalizability concerns, and challenges with small subgroups.

Stratified designs

- Stratified designs involve dividing the population into subgroups (strata) based on key characteristics and then randomly assigning treatments within each stratum
- Stratification aims to improve the precision of treatment effect estimates by reducing variability within strata and ensuring balanced representation of important subgroups
- Stratified designs are commonly used in causal inference to control for confounding variables and increase the power to detect treatment effects

Benefits of stratification

- Reduces variability within strata, leading to more precise treatment effect estimates
- Ensures balanced representation of important subgroups (age groups, gender) in the treatment and control groups
- Allows for the estimation of treatment effects within specific subgroups of interest
- Increases the power to detect treatment effects by reducing noise from confounding variables

Stratification vs randomization

- Stratification is a form of restricted randomization where randomization occurs within each stratum separately
- While randomization alone can balance treatment groups on average, stratification guarantees balance on the selected stratification variables
- Stratification can be more effective than simple randomization in reducing bias and improving precision, especially when the stratification variables are strongly related to the outcome

Selecting stratification variables

- Stratification variables should be chosen based on their expected relationship with the outcome and potential for confounding
- Common stratification variables include demographic characteristics (age, gender, race), baseline risk factors, or geographic location
- The number of strata should be carefully considered to avoid creating too many small strata, which can reduce statistical power
- It is important to select stratification variables that are easily measurable and available for all participants before randomization

Blocked designs

- Blocked designs involve dividing the experimental units into homogeneous subgroups (blocks) based on extraneous factors that may influence the outcome

- Treatments are then randomly assigned within each block, ensuring that each treatment appears an equal number of times within each block
- Blocked designs are commonly used in causal inference to control for nuisance factors and reduce variability in the treatment effect estimates

Benefits of blocking

- Reduces variability within blocks, leading to more precise treatment effect estimates
- Controls for extraneous factors (batch effects, time effects) that may influence the outcome but are not of primary interest
- Ensures balanced allocation of treatments within each block, reducing the impact of nuisance factors on the treatment effect estimates
- Increases the power to detect treatment effects by removing the variability associated with the blocking factors

Blocking vs stratification

- While both blocking and stratification involve dividing the population into subgroups, they differ in their purpose and implementation
- Stratification is used to ensure balance on key baseline characteristics and estimate subgroup-specific treatment effects, while blocking is used to control for nuisance factors and reduce variability
- Stratification variables are typically related to the outcome and used in the analysis, while blocking factors are not necessarily related to the outcome and may not be included in the analysis
- Stratification is more commonly used in observational studies, while blocking is more common in experimental studies

Selecting blocking variables

- Blocking variables should be chosen based on their potential to introduce variability in the outcome, but are not of primary interest in the study
- Common blocking variables include batch effects (different production runs), time effects (different days or seasons), or geographic location (different study sites)
- The number of blocks should be carefully considered to ensure sufficient replication within each block while maintaining a manageable study design
- Blocking variables should be easily measurable and available before randomization, but do not need to be included in the final analysis

Analyzing stratified designs

- The analysis of stratified designs involves estimating treatment effects within each stratum and then combining the estimates to obtain an overall treatment effect
- Stratified designs allow for the assessment of treatment effect heterogeneity across strata and the identification of subgroups that may benefit more or less from the treatment
- The analysis of stratified designs must account for the stratification variables to obtain valid and precise treatment effect estimates

Estimating treatment effects

- Treatment effects are estimated within each stratum by comparing the outcomes between the treatment and control groups
- Common estimators include the difference in means, risk difference, risk ratio, or odds ratio, depending on the type of outcome variable
- The within-stratum treatment effect estimates are then combined using a weighted average, where the weights are typically proportional to the stratum sizes or the inverse of the stratum-specific variances

Pooled vs stratified analysis

- A pooled analysis ignores the stratification and estimates the treatment effect across all strata combined, while a stratified analysis estimates the treatment effects within each stratum and then combines them
- A stratified analysis is generally more efficient than a pooled analysis, as it accounts for the reduced variability within strata and can provide more precise treatment effect estimates
- However, a pooled analysis may be preferred when the treatment effect is expected to be homogeneous across strata or when some strata have very small sample sizes

Assessing balance within strata

- It is important to assess the balance of baseline characteristics between the treatment and control groups within each stratum
- Imbalances within strata can lead to biased treatment effect estimates and reduce the effectiveness of stratification
- Common methods for assessing balance include comparing means or proportions of baseline variables between treatment groups within each stratum, or using standardized differences or hypothesis tests
- If substantial imbalances are found within strata, it may be necessary to adjust for these variables in the analysis or consider alternative stratification variables

Analyzing blocked designs

- The analysis of blocked designs involves estimating treatment effects within each block and then combining the estimates to obtain an overall treatment effect
- Blocked designs allow for the control of nuisance factors and the reduction of variability in the treatment effect estimates
- The analysis of blocked designs must account for the blocking structure to obtain valid and precise treatment effect estimates

Estimating treatment effects

- Treatment effects are estimated within each block by comparing the outcomes between the treatment and control groups
- Common estimators include the difference in means, risk difference, risk ratio, or odds ratio, depending on the type of outcome variable

- The within-block treatment effect estimates are then combined using a weighted average, where the weights are typically proportional to the block sizes or the inverse of the block-specific variances

Accounting for block effects

- The analysis of blocked designs should include block effects to account for the variability between blocks and improve the precision of the treatment effect estimates
- Block effects can be modeled as fixed effects (assuming the blocks are a fixed set of categories) or random effects (assuming the blocks are a random sample from a larger population)
- Including block effects in the analysis can help to reduce the residual variability and increase the power to detect treatment effects
- If block effects are not accounted for, the treatment effect estimates may be biased and less precise

Assessing balance within blocks

- It is important to assess the balance of baseline characteristics between the treatment and control groups within each block
- Imbalances within blocks can lead to biased treatment effect estimates and reduce the effectiveness of blocking
- Common methods for assessing balance include comparing means or proportions of baseline variables between treatment groups within each block, or using standardized differences or hypothesis tests
- If substantial imbalances are found within blocks, it may be necessary to adjust for these variables in the analysis or consider alternative blocking variables

Combining stratification and blocking

- Stratification and blocking can be combined in the same study design to control for both confounding variables and nuisance factors simultaneously
- A combined stratified and blocked design involves first stratifying the population based on key baseline characteristics and then applying blocking within each stratum
- Combining stratification and blocking can provide the benefits of both techniques, including improved balance, reduced variability, and increased power to detect treatment effects

Benefits of combined approach

- Controls for both confounding variables (through stratification) and nuisance factors (through blocking), leading to more precise and unbiased treatment effect estimates
- Ensures balance on key baseline characteristics across treatment groups within each stratum and block
- Reduces variability within strata and blocks, increasing the power to detect treatment effects
- Allows for the estimation of treatment effects within specific subgroups of interest while controlling for extraneous factors

Design considerations

- The choice of stratification and blocking variables should be based on their expected relationships with the outcome and their potential for confounding or introducing variability
- The number of strata and blocks should be carefully considered to ensure sufficient sample sizes within each combination of stratum and block
- The order of stratification and blocking (i.e., stratifying first and then blocking within strata, or vice versa) may depend on the relative importance of the stratification and blocking variables
- The randomization procedure should be clearly defined and implemented within each combination of stratum and block

Analysis of combined designs

- The analysis of combined stratified and blocked designs should account for both the stratification and blocking variables to obtain valid and precise treatment effect estimates
- Treatment effects can be estimated within each combination of stratum and block, and then combined using weighted averages across strata and blocks
- The analysis should include both stratum and block effects to control for the variability associated with these factors
- Assessing balance and treatment effect heterogeneity across strata and blocks can provide insights into the generalizability and robustness of the findings

Limitations of stratified and blocked designs

- While stratified and blocked designs offer several benefits for causal inference, they also have some limitations that should be considered when planning a study
- These limitations can impact the feasibility, generalizability, and analysis of the study, and may require careful planning and adaptation to address

Feasibility constraints

- Stratified and blocked designs require a larger sample size compared to simple randomization, as each stratum and block needs a sufficient number of participants to ensure balance and precision
- Recruiting participants and implementing the randomization procedure within each stratum and block can be logistically challenging and resource-intensive
- In some cases, the desired stratification or blocking variables may not be available or measurable before randomization, limiting the applicability of these designs

Generalizability concerns

- Stratified and blocked designs are designed to control for specific variables and factors, which may limit the generalizability of the findings to other populations or settings
- If the selected stratification or blocking variables are not relevant or representative of the target population, the treatment effect estimates may not be applicable beyond the study sample
- Over-stratification or using too many blocking variables can lead to small sample sizes within each stratum or block, reducing the power and generalizability of the study

Handling small strata or blocks

- When the sample size is limited, some strata or blocks may end up with very few participants, leading to imprecise treatment effect estimates and reduced power
- Small strata or blocks can also be more sensitive to imbalances in baseline characteristics, as even minor differences can have a large impact on the treatment effect estimates
- In such cases, it may be necessary to combine small strata or blocks, use alternative stratification or blocking variables, or consider other design approaches (covariate adjustment, matching)
- The analysis of designs with small strata or blocks may require special techniques (exact tests, Bayesian methods) to obtain valid and reliable results

3.3 Factorial designs

Factorial designs let you study the effects of multiple independent variables (factors) at the same time within a single experiment. Instead of running separate experiments for each factor, you cross all factor levels together, which reveals not only how each factor affects the outcome on its own but also how factors combine to produce effects that neither would produce alone. This makes factorial designs one of the most efficient and informative structures available for randomized experiments.

Factorial design overview

A factorial design is an experiment in which every level of every factor is combined with every level of every other factor. Subjects are then randomly assigned to these combinations. The result is that you can estimate main effects (the average impact of each factor) and interaction effects (whether the impact of one factor changes depending on the level of another).

Factors and levels

A factor is an independent variable that the experimenter manipulates. Each factor has two or more levels, which are the specific values or categories it can take.

For example, in a 2×2 factorial design studying weight loss, you might have:

- Factor A: Diet with levels low-fat and high-fat
- Factor B: Exercise with levels sedentary and active

The notation " 2×2 " tells you there are two factors, each with two levels.

Treatment combinations

A treatment combination is one specific pairing of factor levels. You get the total number of combinations by multiplying the number of levels across all factors.

- A 2×2 design has $2 \times 2 = 4$ treatment combinations
- A 2×3 design has $2 \times 3 = 6$ treatment combinations
- A $2 \times 2 \times 2$ design has $2 \times 2 \times 2 = 8$ treatment combinations

In the diet-and-exercise example, the four treatment combinations are: low-fat/sedentary, low-fat/active, high-fat/sedentary, and high-fat/active.

Balanced vs unbalanced designs

- A balanced factorial design assigns an equal number of subjects to each treatment combination. This is preferred because it gives equal precision for all effect estimates and simplifies the analysis.
- An unbalanced design has unequal cell sizes, which can happen due to dropout or practical constraints. Unbalanced designs require more complex analysis (e.g., Type III sums of squares) and can make interaction effects harder to estimate cleanly.

Benefits of factorial designs

Efficiency vs single-factor experiments

Factorial designs get more information per subject than running separate one-factor-at-a-time experiments. Consider a 2×2 design with 20 subjects per cell (80 total). Each main effect estimate uses all 80 subjects (40 per level of each factor), so you effectively get two experiments' worth of information from one sample.

Running two separate single-factor experiments with the same precision would also require 80 total subjects, but you'd learn nothing about how the factors interact.

Interaction effects detection

An interaction effect occurs when the effect of one factor depends on the level of another factor. Factorial designs are the only way to detect these.

For example, suppose a study finds that medication reduces depression symptoms by 10 points on average, and therapy reduces them by 8 points. If the combination reduces symptoms by 25 points rather than the expected 18, there's a positive interaction: the treatments are synergistic. You'd miss this entirely if you studied medication and therapy in separate experiments.

Cost and time savings

By combining factors into a single experiment, you avoid duplicating recruitment, setup, and measurement procedures across multiple studies. This matters especially in settings where running experiments is expensive or time-sensitive, such as clinical trials or policy evaluations.

Assumptions of factorial designs

The standard analysis of factorial designs (ANOVA-based) relies on several assumptions. Violations can bias your estimates or inflate your Type I error rate.

Independence of observations

Each observation should be independent of every other observation. If subjects can influence each other's outcomes, this assumption breaks down. A common violation occurs in cluster settings: students in the same classroom, patients in the same clinic, etc. When clustering is present, you need to account for it (e.g., cluster-robust standard errors or multilevel models).

Normality of residuals

The residuals (observed minus predicted values) should be approximately normally distributed. You can check this with Q-Q plots or the Shapiro-Wilk test. Moderate departures from normality are usually tolerable with large samples due to the central limit theorem, but heavily skewed or heavy-tailed distributions may call for data transformations or nonparametric alternatives.

Homogeneity of variances

The variance of residuals should be roughly equal across all treatment combinations. This is called homoscedasticity. You can check it with Levene's test or by plotting residuals against predicted values. If variances differ substantially across cells, consider using Welch-type corrections or robust standard errors.

Designing factorial experiments

Choosing factors and levels

1. Select factors that have a plausible causal relationship with the outcome and are directly relevant to your research question.
2. Choose levels that are meaningfully distinct from each other and representative of the range you care about.
3. Consider practical feasibility and ethics. Can you actually manipulate each factor at the chosen levels without harming participants?

For instance, in an agricultural experiment on plant growth, you might choose temperature levels of 15°C, 25°C, and 35°C to span a realistic range, rather than picking two values that are barely different.

Determining sample size

Use power analysis before collecting data to figure out how many subjects you need per cell. The required sample size depends on:

- The effect size you want to detect (smaller effects need larger samples)
- Your chosen significance level (typically $\alpha = 0.05$)
- Your desired power (typically 80% or higher)
- The number of treatment combinations (more cells means more subjects overall)

For a 2×2 design aiming to detect a medium-sized interaction effect (Cohen's $f = 0.25$) with 80% power at $\alpha = 0.05$, a power analysis might indicate you need roughly 100 subjects per cell (400 total). Software like G*Power can compute these numbers for you.

Randomization and blocking

Randomization is what makes factorial designs valid for causal inference. Randomly assigning subjects to treatment combinations ensures that, in expectation, the groups are comparable on both observed and unobserved characteristics.

Blocking adds precision on top of randomization. You group subjects by a known source of variability (the blocking variable), then randomize treatment assignments within each block. For example, if prior academic performance predicts the outcome, you could create blocks of high-, medium-, and low-performing students, then randomly assign treatments within each block. This reduces residual variance and increases power.

Analyzing factorial designs

ANOVA for factorial designs

Analysis of variance (ANOVA) is the standard tool. It partitions the total sum of squares in the outcome into components:

- Main effect of Factor A
- Main effect of Factor B
- Interaction effect A×B
- Residual (error)

Each component is tested with an F-test, which compares the variance explained by that component to the residual variance. The ANOVA table reports degrees of freedom, sums of squares, mean squares, F-values, and p-values for each source.

Main effects vs interaction effects

A main effect is the average effect of one factor, collapsing across all levels of the other factor(s). A significant main effect of diet means that, on average across exercise conditions, one diet produces different outcomes than the other.

An interaction effect means the effect of one factor changes depending on the level of another. If diet matters a lot for sedentary people but barely matters for active people, that's an interaction.

When a significant interaction is present, interpreting main effects in isolation can be misleading. The main effect of diet might look small on average, but that average hides the fact that diet has a large effect for one subgroup and no effect for another. Always check for interactions before drawing conclusions from main effects alone.

Multiple comparisons and post-hoc tests

When a main effect or interaction is significant, you often want to know which specific groups differ. Post-hoc tests do this while controlling for the inflated Type I error rate that comes from making multiple comparisons.

Common procedures include:

- Tukey's HSD: Controls the family-wise error rate across all pairwise comparisons. Good default choice.
- Bonferroni correction: Divides α by the number of comparisons. Conservative but simple.
- Scheffé's method: Most conservative; useful when you're testing complex contrasts, not just pairwise differences.

For example, if a significant diet×exercise interaction is found, post-hoc tests might reveal that the low-fat diet leads to significantly more weight loss than the high-fat diet among sedentary individuals, but not among active individuals.

Effect size and power

Statistical significance alone doesn't tell you whether an effect is practically meaningful. Always report effect sizes alongside p-values.

- Partial eta-squared (η_p^2): The proportion of variance in the outcome explained by a given effect, after removing variance due to other effects. Values of 0.01, 0.06, and 0.14 are often considered small, medium, and large.
- Cohen's f : Related to η_p^2 by

$$f = \sqrt{\frac{\eta_p^2}{1 - \eta_p^2}}$$

. Values of 0.10, 0.25, and 0.40 correspond to small, medium, and large effects.

A study with a very large sample might find a statistically significant effect that is too small to matter in practice. Conversely, a small study might miss a real and important effect. Reporting both significance and effect size gives the full picture.

Interpreting factorial results

Significance of main effects

A significant main effect tells you that the average outcome differs across levels of that factor, collapsing over the other factor(s). Check the direction (which level produces higher outcomes) and the magnitude (how large is the difference relative to the variability in the data).

Remember: if there's also a significant interaction involving that factor, the main effect is an average that may not describe any particular subgroup well.

Significance of interaction effects

A significant interaction means the story changes depending on which combination of factor levels you're looking at. To understand the interaction, examine the simple effects: the effect of one factor at each fixed level of the other factor.

For example, if study method and subject difficulty interact in their effect on test scores, you'd look at the effect of study method separately for easy subjects and hard subjects. The interaction tells you these simple effects are not the same.

Graphical representations of effects

Interaction plots are the most useful visual tool. Plot the mean outcome on the y-axis, levels of one factor on the x-axis, and use separate lines for each level of the other factor.

- Parallel lines indicate no interaction: the effect of one factor is the same regardless of the other factor's level.
- Non-parallel lines suggest an interaction. The more the lines diverge or cross, the stronger the interaction.

Bar graphs with error bars (showing confidence intervals) for each cell mean are also helpful for comparing specific treatment combinations.

Factorial design variations

Two-way vs three-way designs

A two-way design has two factors and can estimate two main effects and one two-way interaction. A three-way design has three factors and can estimate three main effects, three two-way interactions, and one three-way interaction.

Three-way (and higher) designs provide richer information but come with costs: more treatment combinations, larger required sample sizes, and harder-to-interpret higher-order interactions. Use them when you have strong theoretical reasons to expect a three-way interaction and sufficient resources to power the design.

Within-subjects vs between-subjects factors

- Between-subjects: Each participant experiences only one treatment combination. Simple to implement, but individual differences add noise.
- Within-subjects (repeated measures): Each participant experiences all levels of a factor. This controls for individual differences and increases power, but introduces risks of carryover effects (earlier conditions affecting later ones) and order effects.

Counterbalancing (varying the order of conditions across participants) helps mitigate order effects in within-subjects designs.

Mixed factorial designs

A mixed design combines at least one between-subjects factor and at least one within-subjects factor. This is common when one factor can't be repeated (e.g., a drug vs. placebo, assigned once) but another naturally involves repeated measurement (e.g., time points).

For example, a study might assign participants to caffeine or placebo (between-subjects) and test them in both the morning and afternoon (within-subjects). The mixed design lets you estimate the main effects of caffeine and time of day, plus their interaction.

Limitations of factorial designs

Large number of treatment combinations

The number of cells grows multiplicatively. A 2^4 design has 16 cells; a 3^3 design has 27. Each cell needs enough subjects for adequate power, so the total sample size can become impractical quickly.

Fractional factorial designs address this by running only a strategically chosen subset of all possible treatment combinations. You sacrifice the ability to estimate some higher-order interactions, but you can still estimate main effects and lower-order interactions with far fewer subjects.

Difficulty interpreting higher-order interactions

A two-way interaction is relatively straightforward: the effect of A depends on B. A three-way interaction means the two-way interaction between A and B itself changes across levels of C. Four-way interactions are even harder to parse.

Higher-order interactions are also more likely to be spurious, especially when you're testing many effects simultaneously. Treat them with caution and look for replication before building theory around them.

Confounding and lurking variables

Randomization protects against confounding in expectation, but it doesn't guarantee balance in any single experiment, especially with small samples. Lurking variables are unmeasured factors that could influence the outcome and co-vary with your treatment assignments by chance.

To guard against this:

- Use randomization (the most important step)
- Block on known prognostic variables
- Check for baseline balance across treatment groups
- Consider covariate adjustment in the analysis

Factorial designs in practice

Real-world examples and case studies

- Psychology: Bandura, Ross, and Ross (1961) used a factorial design to study children's aggression after observing aggressive vs. non-aggressive models who were either rewarded or punished. The interaction revealed that model reward moderated the effect of observed aggression on children's behavior.
- Marketing: A company might test price (low, medium, high) crossed with ad type (emotional, informational) to see which combination maximizes purchase intent. The interaction could show that emotional ads work best at high price points.
- Healthcare: The STAR*D trial and similar studies use factorial-like structures to evaluate combinations of medications and psychotherapy, testing whether treatments are additive or synergistic.

Reporting factorial design results

Follow established reporting guidelines (APA Style, CONSORT for trials). A complete report should include:

1. A clear description of all factors, their levels, and the resulting treatment combinations
2. Sample size per cell and overall, with information on how the sample size was determined
3. The randomization procedure and any blocking variables
4. Full ANOVA results: F-values, degrees of freedom, p-values, and effect sizes for every main effect and interaction
5. Post-hoc comparisons where relevant, with the correction method specified
6. Graphical displays (interaction plots, bar graphs with confidence intervals) to make the pattern of results accessible to readers

3.4 Cluster randomized designs

Cluster randomized trials involve randomizing groups of individuals rather than individuals themselves. This approach is useful in public health and social science research when individual randomization isn't feasible or when studying interventions that operate at a group level.

These trials require special considerations in design, analysis, and interpretation. Key aspects include accounting for intracluster correlation, handling variability in cluster sizes, and using appropriate statistical methods like generalized estimating equations or mixed effects models to analyze the data.

Cluster randomized trials

- Involve randomizing intact social units or clusters of individuals (schools, communities, hospitals) to different intervention groups rather than randomizing individuals
- Commonly used in public health, education, and social science research when individual randomization is not feasible or desirable
- Require special considerations in design, analysis, and interpretation compared to individually randomized trials

Rationale for cluster randomization

- Avoids contamination between intervention and control groups that may occur when individuals within the same cluster are assigned to different groups
- Allows for the evaluation of interventions that operate at a group level or cannot be delivered to individuals (community-wide health promotion campaigns)
- May be more feasible and cost-effective than individual randomization in certain settings (school-based interventions)
- Enables the assessment of both direct and indirect effects of interventions on outcomes

Design of cluster randomized trials

Parallel vs stepped wedge designs

- Parallel designs involve randomizing clusters to different intervention groups at the start of the trial and following them over time
- Stepped wedge designs involve sequentially rolling out the intervention to clusters over time, with the order of rollout determined by randomization
- Stepped wedge designs may be preferred when it is unethical to withhold the intervention from some clusters or when logistical constraints prevent simultaneous implementation across all clusters

Matching and stratification

- Matching involves pairing clusters based on important baseline characteristics (size, socioeconomic status) and randomizing within pairs to ensure balance between intervention groups
- Stratification involves dividing clusters into strata based on important characteristics and randomizing within each stratum to ensure balance
- Both matching and stratification can improve the efficiency and validity of cluster randomized trials by reducing the potential for confounding and increasing statistical power

Randomization procedures

- Simple randomization involves randomly assigning clusters to intervention groups with equal probability

- Restricted randomization techniques (block randomization, minimization) can be used to ensure balance between intervention groups on important baseline characteristics
- Randomization should be performed by an independent party using a valid random allocation sequence to minimize the risk of selection bias

Sample size and power calculations

Intraclass correlation coefficient

- Measures the degree of similarity among individuals within the same cluster compared to individuals in different clusters
- Higher values indicate greater clustering of outcomes within clusters and reduced statistical power compared to individual randomization
- Must be accounted for in sample size calculations to ensure adequate power to detect intervention effects

Variability in cluster sizes

- Unequal cluster sizes can reduce statistical power and efficiency compared to equal cluster sizes
- Sample size calculations should account for the expected variability in cluster sizes and the potential for attrition or missing data at the cluster level
- Analysis methods that account for variable cluster sizes (weighted regression, mixed effects models) may be necessary to obtain valid inferences

Analysis of cluster randomized trials

Generalized estimating equations

- Extension of generalized linear models that accounts for the correlation of outcomes within clusters
- Provides population-averaged estimates of intervention effects while adjusting for clustering
- Requires specification of a working correlation structure (exchangeable, autoregressive) to model the dependence of outcomes within clusters

Mixed effects models

- Incorporate both fixed effects (intervention group) and random effects (cluster-specific intercepts and slopes) to account for the hierarchical structure of the data
- Provide cluster-specific estimates of intervention effects and allow for the estimation of between-cluster variability
- May be more efficient than generalized estimating equations when the number of clusters is small or the cluster sizes are variable

Small-sample corrections

- Cluster-level analyses that treat each cluster as a single observation can lead to biased estimates and inflated Type I error rates when the number of clusters is small (<30 per group)

- Small-sample corrections (Kenward-Roger, Satterthwaite) adjust the degrees of freedom for hypothesis tests to account for the uncertainty in estimating the covariance parameters
- Permutation tests that randomly reassign clusters to intervention groups can provide valid inferences when parametric assumptions are violated or the number of clusters is very small

Challenges in cluster randomized trials

Selection bias and imbalance

- Cluster randomization may not always result in balance between intervention groups on important baseline characteristics, especially when the number of clusters is small
- Imbalance can lead to confounding and biased estimates of intervention effects if not properly accounted for in the analysis
- Strategies to minimize imbalance include matching, stratification, and covariate adjustment in the analysis

Contamination and interference

- Contamination occurs when individuals in the control group are exposed to the intervention, diluting the observed intervention effect
- Interference occurs when the outcome of an individual depends on the treatment status of other individuals in the same or neighboring clusters
- Both contamination and interference can bias the estimates of intervention effects and reduce statistical power
- Strategies to minimize contamination and interference include using geographically dispersed clusters, blinding participants and investigators, and measuring the extent of contamination and interference

Attrition and missing data

- Attrition at the cluster level (entire clusters dropping out) or the individual level (individuals within clusters dropping out) can lead to biased estimates of intervention effects if the missingness is related to the intervention or the outcome
- Strategies to minimize attrition include engaging with cluster leaders, providing incentives for participation, and using methods to follow up with individuals who miss assessments
- Analysis methods that account for missing data (multiple imputation, inverse probability weighting) may be necessary to obtain valid inferences

Reporting and interpretation

CONSORT guidelines for cluster trials

- Extension of the CONSORT (Consolidated Standards of Reporting Trials) statement for individually randomized trials
- Provide a checklist of essential items to include when reporting a cluster randomized trial (rationale for clustering, randomization procedures, sample size calculations)

- Aim to improve the transparency, completeness, and accuracy of reporting and facilitate the critical appraisal and interpretation of cluster randomized trials

Generalizability and external validity

- Cluster randomized trials may have limited generalizability if the included clusters are not representative of the broader population of interest
- External validity may be compromised if the intervention is not feasible or acceptable in real-world settings outside of the trial context
- Assessing the generalizability and external validity of cluster randomized trials requires careful consideration of the characteristics of the included clusters and the context in which the intervention was delivered
- Replication of findings in different populations and settings can enhance the external validity and inform the scale-up and implementation of effective interventions

3.5 Analysis of randomized experiments

Fundamentals of Randomized Experiments

Randomized experiments are the gold standard for causal inference because they let you isolate the effect of a treatment from everything else. By randomly assigning units to treatment and control groups, you make those groups comparable on average across every characteristic, whether you measured it or not.

A well-designed experiment requires four things: a clearly defined population of interest, a precisely specified treatment, an adequate sample size, and a valid randomization procedure for assigning units to conditions.

Benefits of Randomization

Why does randomization matter so much? Because it balances both observed and unobserved confounders between groups on average. No observational method can do that.

- Creates comparable groups, enabling unbiased estimation of causal effects
- Provides a basis for statistical inference: you can quantify the uncertainty around your treatment effect estimates using known probability distributions
- Strengthens the credibility of findings by minimizing bias and confounding, which is why randomized experiments carry more weight than observational studies in most policy and scientific debates

Estimating Causal Effects

Average Treatment Effect (ATE)

The average treatment effect (ATE) is the most common estimand in a randomized experiment. It captures the average difference in outcomes if everyone were treated versus if no one were treated:

$$ATE = E[Y(1)] - E[Y(0)]$$

Here, $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control. You never observe both for the same unit (the fundamental problem of causal inference), but randomization lets you estimate the ATE by comparing group averages.

You can estimate the ATE using a simple difference in sample means or through regression analysis, both covered below.

Intention-to-Treat (ITT) Analysis

Intention-to-treat (ITT) analysis estimates the effect of being assigned to treatment, regardless of whether units actually received or complied with it.

- Preserves the benefits of randomization because you analyze everyone according to their original assignment
- Yields a conservative estimate of the treatment effect, since noncompliers dilute the measured impact
- Reflects real-world effectiveness: in practice, not everyone follows through on an assigned treatment, so the ITT tells you what to expect from a policy of offering the treatment

Validity in Randomized Experiments

Internal vs. External Validity

These two concepts address different questions about your experiment's conclusions.

Internal validity asks: Did the treatment actually cause the observed outcome difference within this study? Randomization directly strengthens internal validity by eliminating confounding.

External validity asks: Do these findings generalize to other populations, settings, or times? A perfectly randomized experiment can still have limited external validity if the sample is unrepresentative or the study conditions are artificial. For example, a drug trial conducted only on young, healthy volunteers may not generalize to elderly patients with comorbidities.

Analyzing Completely Randomized Designs

Difference in Means

In a completely randomized design, units are assigned to treatment or control without any stratification or matching. The simplest estimator for the ATE is the difference in sample means:

$$\hat{ATE} = \bar{Y}_1 - \bar{Y}_0$$

This estimator is unbiased under successful randomization. You construct confidence intervals and hypothesis tests using the standard error of the difference, which depends on the variance within each group and the sample sizes.

Regression Analysis

Regression offers a more flexible way to estimate the ATE. The basic model is:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

Here T_i is a binary treatment indicator (1 = treated, 0 = control), and β_1 is the estimated ATE. Without covariates, this gives you exactly the same answer as the difference in means.

Why bother with regression, then? Two reasons:

- Precision gains. Including pre-treatment covariates (e.g., baseline scores) reduces residual variance, which shrinks your standard errors and tightens confidence intervals.

- Heterogeneous effects. Adding interaction terms between T_i and covariates lets you estimate how the treatment effect varies across subgroups.

Analyzing Stratified and Matched Designs

Conditional Average Treatment Effect

Stratified and matched designs group units into strata or pairs based on observed covariates before randomization occurs within each group. This guarantees balance on those covariates.

The conditional average treatment effect (CATE) is the average treatment effect within a specific stratum. You estimate it by computing the difference in means (or running a regression) within each stratum separately.

To get the overall ATE, take a weighted average of the CATEs across strata, where the weights are the proportion of units in each stratum.

Regression with Strata Indicators

You can handle stratification in a single regression by including strata fixed effects:

$$Y_i = \beta_0 + \beta_1 T_i + \sum_{s=1}^{S-1} \gamma_s I_{is} + \varepsilon_i$$

The I_{is} terms are indicator variables for each stratum s , and β_1 gives you the pooled treatment effect across strata (assuming a constant effect). If you want to allow the treatment effect to differ by stratum, interact T_i with the strata indicators.

Noncompliance in Experiments

One-Sided vs. Two-Sided Noncompliance

Noncompliance happens when units don't follow their assigned treatment condition.

- One-sided noncompliance: Some units assigned to treatment don't take it, but no one in the control group receives the treatment. Common in trials where the treatment is optional (e.g., a training program people can skip).
- Two-sided noncompliance: Units in both groups deviate. Some assigned to treatment don't take it, and some assigned to control obtain the treatment on their own.

Noncompliance is a problem because if the people who comply differ systematically from those who don't, a naive comparison of actual treatment recipients vs. non-recipients is confounded.

Complier Average Causal Effect (CACE)

The CACE (also called the local average treatment effect, or LATE) targets the treatment effect specifically among compliers: units who would take the treatment if assigned to it and would not take it if assigned to control.

$$CACE = E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0]$$

Here $D(1)$ and $D(0)$ are potential treatment-receipt indicators under assignment to treatment and control, respectively. Estimating the CACE requires two key assumptions beyond random assignment:

- Exclusion restriction: Assignment affects outcomes only through its effect on actual treatment receipt.

- Monotonicity: No unit does the opposite of their assignment in both directions (i.e., no "defiers" who would take treatment only when assigned to control).

Instrumental Variables Estimation

Instrumental variables (IV) estimation uses the randomized assignment as an instrument to recover the CACE. The estimator is straightforward:

$$CACE = \frac{\hat{ITT}_Y}{\hat{ITT}_D}$$

The numerator is the ITT effect on the outcome (reduced-form effect), and the denominator is the ITT effect on treatment receipt (the first-stage effect, i.e., the compliance rate difference between groups).

This works because dividing by the compliance rate "scales up" the diluted ITT to reflect the effect among those who actually responded to the assignment. The required assumptions are: random assignment, exclusion restriction, monotonicity, and SUTVA (stable unit treatment value assumption, meaning one unit's outcome isn't affected by another's assignment).

Dealing with Missing Data

Inverse Probability Weighting

Inverse probability weighting (IPW) handles missing outcome data by upweighting observed units that resemble the missing ones.

1. Estimate the probability of being observed for each unit, typically using logistic regression with treatment assignment and baseline covariates as predictors.
2. Assign each observed unit a weight equal to the inverse of its estimated probability of being observed.
3. Compute the weighted difference in means (or weighted regression) using these weights.

The intuition: if a certain type of unit is underrepresented among the observed data because similar units tend to drop out, IPW gives those remaining units more influence to compensate. This produces unbiased estimates as long as the model for the observation probability is correctly specified.

Multiple Imputation

Multiple imputation (MI) takes a different approach: instead of reweighting, it fills in the missing values.

1. Fit a model to the observed data and draw multiple plausible values for each missing observation from the model's posterior predictive distribution.
2. Create several complete datasets (typically 5-20), each with different imputed values.
3. Analyze each dataset separately using your standard method.
4. Combine the results across datasets using Rubin's rules, which pool the point estimates and adjust the standard errors to reflect both within-imputation and between-imputation variability.

MI assumes data are missing at random (MAR): the probability of missingness depends only on observed variables, not on the missing values themselves. This is a weaker assumption than "missing completely at random" but still requires careful thought about what drives attrition.

Subgroup Analysis and Heterogeneity

Interaction Terms in Regression

To test whether a treatment works differently for different types of people, add an interaction term to your regression:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 (T_i \times X_i) + \varepsilon_i$$

Here X_i is a subgroup indicator (e.g., male vs. female), and β_3 captures the difference in treatment effects between subgroups. A statistically significant β_3 suggests treatment effect heterogeneity: the treatment doesn't affect everyone equally.

The treatment effect for the baseline subgroup ($X_i = 0$) is β_1 , and for the other subgroup ($X_i = 1$) it's $\beta_1 + \beta_3$.

Dangers of Post-Hoc Subgroup Analysis

Subgroup analyses that weren't planned before data collection are risky:

- Multiple comparisons problem. Testing many subgroups inflates the chance of finding a "significant" difference by chance. With 20 subgroups tested at $\alpha = 0.05$, you'd expect about one false positive even if the treatment effect is truly constant.
- Overfitting to noise. Post-hoc subgroups are often suggested by patterns in the data, which means you're partly "discovering" random variation.
- Lower credibility. Reviewers and policymakers rightly treat post-hoc findings as exploratory, not confirmatory.

Pre-specified subgroup analyses (registered before data collection) are far more credible. Post-hoc findings should be replicated in independent studies before being taken seriously.

Statistical Power and Sample Size

Minimum Detectable Effect Size

Statistical power is the probability that your study will detect a true treatment effect when one exists. The minimum detectable effect size (MDES) is the smallest effect your study can reliably detect at a given significance level and power.

The MDES depends on:

- Sample size (larger $n \rightarrow$ smaller MDES)
- Outcome variance (less noisy outcomes $\rightarrow$ smaller MDES)
- Significance level (α , the Type I error rate)
- Desired power (typically 0.80 or higher; $1 - \beta$, where β is the Type II error rate)

Factors Influencing Power

Several design choices affect your power:

- Increase sample size. The most direct way to boost power.

- Reduce outcome variance. Including pre-treatment covariates in your analysis absorbs some outcome variability, effectively increasing power without adding participants.
- Use stratified or matched designs. These reduce the variance of the treatment effect estimator compared to a completely randomized design, improving power.
- Accept a higher significance level. Moving from $\alpha = 0.01$ to $\alpha = 0.05$ increases power but also increases the false positive rate, so this is a tradeoff.

Practical Considerations

Ethics and Informed Consent

Randomized experiments must follow core ethical principles:

- Respect for persons: Participants must give informed consent, meaning they understand the procedures, risks, and benefits and voluntarily agree to participate.
- Beneficence: The study should maximize potential benefits and minimize harm.
- Justice: The burdens and benefits of research should be distributed fairly.

Equipoise is a prerequisite for ethical randomization: there must be genuine uncertainty about which treatment is better. If strong evidence already favors one arm, it's unethical to randomize people away from it. Additional scrutiny applies when working with vulnerable populations or when participants receive limited direct benefit.

Generalizability of Results

Even a well-executed experiment may not generalize beyond its specific context. Factors that limit external validity include:

- Narrow inclusion/exclusion criteria that produce an unrepresentative sample
- Artificial study settings that differ from real-world implementation
- Specific populations or geographic contexts that may not reflect the broader target

Replication across different settings, populations, and implementation conditions is the strongest way to establish that a treatment effect is robust and generalizable. When designing an experiment, think carefully about who your target population is and whether your study sample and conditions reflect it.

Unit 4 – Observational Studies & Confounding

4.1 Confounding variables

Confounding variables can distort the true relationship between exposure and outcome in causal studies. They can create false associations or mask real ones, leading to biased estimates and incorrect conclusions. Understanding confounders is crucial for accurate causal inference.

Researchers use various methods to control for confounding, including randomization, stratification, matching, and regression adjustment. Identifying potential confounders requires causal diagrams and subject matter expertise. Assessing residual confounding and addressing challenges like unobserved confounding are essential for robust causal analyses.

Definition of confounding variables

- Confounding variables are extraneous factors that influence both the exposure and outcome variables in a study, leading to a distortion of the true causal relationship between the exposure and outcome
- Confounders can create a spurious association between the exposure and outcome, making it appear as if there is a causal relationship when there is none, or they can mask a true causal relationship
- The presence of confounding variables can lead to biased estimates of the causal effect and incorrect conclusions about the relationship between the exposure and outcome

Impact of confounding variables

Bias in causal estimates

- Confounding variables can introduce bias in the estimated causal effect of the exposure on the outcome
- Positive confounding occurs when the confounder is positively associated with both the exposure and outcome, leading to an overestimation of the causal effect
- Negative confounding occurs when the confounder is positively associated with the exposure but negatively associated with the outcome (or vice versa), leading to an underestimation of the causal effect

Incorrect conclusions

- The presence of confounding variables can lead to incorrect conclusions about the causal relationship between the exposure and outcome
- Researchers may conclude that there is a causal relationship when there is none (false positive) or fail to detect a true causal relationship (false negative)
- Incorrect conclusions can have significant implications for public health policies, clinical practice, and future research directions

Types of confounding variables

Measured vs unmeasured confounders

- Measured confounders are variables that are observed and recorded in the study data
- These confounders can be controlled for in the analysis using various statistical methods (stratification, regression adjustment)
- Unmeasured confounders are variables that are not observed or recorded in the study data
- These confounders cannot be directly controlled for in the analysis and can lead to residual confounding

Time-varying vs time-invariant confounders

- Time-varying confounders are variables that change over time and can affect the exposure and outcome at different points in the study
- Examples of time-varying confounders include age, socioeconomic status, and health behaviors (smoking status)
- Time-invariant confounders are variables that remain constant throughout the study period
- Examples of time-invariant confounders include gender, race, and genetic factors

Examples of confounding variables

Confounding by indication

- Confounding by indication occurs when the indication for treatment (exposure) is also associated with the outcome
- This type of confounding is common in observational studies of medical interventions
- Example: In a study examining the effectiveness of a new medication for treating depression, the severity of depression may be a confounder because patients with more severe depression are more likely to receive the medication and also more likely to have poor outcomes

Socioeconomic status as confounder

- Socioeconomic status (SES) is a common confounder in many epidemiological studies
- SES can influence both the exposure (access to healthcare, health behaviors) and outcome (disease risk, mortality)
- Example: In a study investigating the association between air pollution and respiratory health, SES may be a confounder because individuals with lower SES may live in areas with higher air pollution and also have poorer respiratory health due to other factors (smoking, occupational exposures)

Identifying potential confounders

Causal diagrams

- Causal diagrams, also known as directed acyclic graphs (DAGs), are graphical representations of the assumed causal relationships between variables in a study

- DAGs can help identify potential confounders by visualizing the paths between the exposure, outcome, and other variables
- By examining the DAG, researchers can determine which variables need to be controlled for to minimize confounding

Subject matter expertise

- Subject matter expertise is crucial in identifying potential confounders based on prior knowledge of the research topic
- Researchers should consult with experts in the field to identify variables that may be associated with both the exposure and outcome
- Literature reviews can also provide insights into potential confounders that have been identified in previous studies

Controlling for confounding variables

Randomization

- Randomization is a powerful tool for controlling confounding in experimental studies
- By randomly assigning participants to exposure groups, randomization ensures that potential confounders are balanced across the groups
- Randomization minimizes the risk of confounding by distributing both measured and unmeasured confounders equally between the exposure groups

Stratification

- Stratification involves dividing the study population into subgroups (strata) based on the levels of the confounding variable
- The association between the exposure and outcome is then examined within each stratum separately
- Stratification allows for the assessment of the exposure-outcome relationship while holding the confounder constant within each stratum

Matching

- Matching involves pairing exposed and unexposed individuals based on their similarity in terms of the confounding variable(s)
- By matching on the confounder, researchers can create groups that are balanced with respect to the confounder
- Matching can be done using various methods, such as individual matching, frequency matching, or propensity score matching

Regression adjustment

- Regression adjustment involves including the confounding variable(s) as covariates in a regression model

- By adjusting for the confounders in the regression model, researchers can estimate the association between the exposure and outcome while controlling for the effect of the confounders
- Regression adjustment allows for the simultaneous control of multiple confounders and can handle both continuous and categorical confounders

Propensity score methods

- Propensity score methods involve estimating the probability of exposure given the observed confounders (propensity score) and using this score to balance the exposure groups
- Propensity score matching pairs exposed and unexposed individuals with similar propensity scores
- Propensity score stratification divides the study population into strata based on the propensity score and estimates the exposure-outcome association within each stratum
- Propensity score weighting assigns weights to individuals based on their propensity scores to create a pseudo-population in which the exposure is independent of the confounders

Assessing residual confounding

Sensitivity analysis techniques

- Sensitivity analyses are used to assess the robustness of the study results to potential unmeasured confounding
- One approach is to simulate the presence of an unmeasured confounder and examine how it would affect the estimated causal effect
- Another approach is to vary the strength of the association between the unmeasured confounder and the exposure and outcome to determine the threshold at which the study conclusions would change

Quantitative bias analysis

- Quantitative bias analysis involves estimating the magnitude and direction of bias due to unmeasured confounding
- Researchers can use external data sources or expert knowledge to estimate the prevalence of the unmeasured confounder and its associations with the exposure and outcome
- By incorporating these estimates into the analysis, researchers can quantify the potential impact of unmeasured confounding on the study results

Challenges with confounding variables

Unobserved confounding

- Unobserved confounding occurs when there are unmeasured confounders that are not accounted for in the analysis
- Unobserved confounding can lead to biased estimates of the causal effect and incorrect conclusions
- Addressing unobserved confounding requires the use of sensitivity analyses or the incorporation of external data sources

Collider bias

- Collider bias occurs when conditioning on a variable that is affected by both the exposure and outcome (collider) introduces a spurious association between the exposure and outcome
- Conditioning on a collider can create an association between the exposure and outcome even in the absence of a true causal relationship
- Identifying and avoiding collider bias requires careful consideration of the causal structure and the use of appropriate analytical methods

Overadjustment bias

- Overadjustment bias occurs when a variable that is on the causal pathway between the exposure and outcome is inappropriately controlled for in the analysis
- Adjusting for a mediator can attenuate or eliminate the true causal effect of the exposure on the outcome
- Identifying and avoiding overadjustment bias requires a clear understanding of the causal relationships between the variables and the use of appropriate causal inference methods

4.2 Selection bias

Selection bias is a critical issue in causal inference that can skew study results and lead to incorrect conclusions. It occurs when the sample used doesn't accurately represent the target population, potentially distorting estimates and limiting generalizability.

Understanding the sources, consequences, and methods for detecting and addressing selection bias is crucial for researchers. By recognizing and mitigating this bias, we can improve the validity of causal inferences and ensure more reliable and meaningful research outcomes.

Sources of selection bias

- Selection bias arises when the sample used in a study is not representative of the target population, leading to distorted estimates and conclusions
- Different sources of selection bias can occur at various stages of the research process, from data collection to analysis, impacting the validity of causal inferences

Sampling bias in data collection

- Occurs when the sampling method systematically excludes certain subgroups of the population (non-probability sampling)
- Can result from convenience sampling, where participants are selected based on accessibility rather than representativeness (students on a college campus)
- Oversampling or undersampling specific segments of the population introduces bias (oversampling urban residents in a national survey)
- Inadequate sampling frame that fails to capture the entire target population (outdated voter registration lists)

Non-response bias in surveys

- Arises when individuals who respond to a survey differ systematically from those who do not respond
- Non-respondents may have different characteristics, opinions, or behaviors compared to respondents (healthier individuals more likely to participate in health surveys)
- Can be influenced by factors such as survey mode, incentives, and follow-up procedures (online surveys may exclude those without internet access)
- Leads to biased estimates if the non-response is related to the outcome of interest (political polls with low response rates)

Volunteer bias in studies

- Happens when participants self-select into a study based on their own interests, motivations, or characteristics
- Volunteers may differ from non-volunteers in ways that affect the study outcomes (health-conscious individuals more likely to enroll in a nutrition study)
- Can limit the generalizability of findings to the broader population (results from a study with highly motivated volunteers may not apply to the general public)
- Particularly problematic in observational studies and clinical trials that rely on voluntary participation

Survivorship bias in analysis

- Occurs when the analysis focuses only on the "survivors" or successful cases, ignoring those that dropped out or failed
- Can lead to overestimating the effectiveness of interventions or the prevalence of positive outcomes (analyzing the performance of successful companies while ignoring bankrupt ones)
- Fails to account for the missing data or attrition that may be related to the outcome of interest (studying long-term effects of a drug only among patients who tolerated it well)
- Requires careful consideration of the reasons for dropout or failure and their potential impact on the results

Consequences of selection bias

Biased estimates of effects

- Selection bias can lead to overestimation or underestimation of the true causal effect, depending on the direction and magnitude of the bias
- Biased estimates can occur when the selection process is related to both the exposure and the outcome (self-selection into a treatment group based on perceived benefits)
- Can distort the magnitude and even the direction of the observed association (a study with healthy volunteer bias may underestimate the effect of a risk factor on disease)

Incorrect causal conclusions

- Selection bias can lead to erroneous conclusions about the presence, absence, or strength of causal relationships
- Biased samples may create spurious associations or mask true causal effects (concluding that a treatment is effective when the observed benefit is due to selection bias)
- Can undermine the internal validity of a study, making it difficult to establish causal claims with confidence

Limits to generalizability

- Selection bias can restrict the external validity or generalizability of study findings to the target population
- Results based on a biased sample may not be applicable to the broader population of interest (findings from a study of college students may not generalize to the general adult population)
- Limits the ability to make valid inferences or predictions beyond the specific study context
- Requires careful consideration of the representativeness of the sample and the factors that may influence selection

Detecting selection bias

Comparing sample vs population

- Assessing the representativeness of the sample by comparing its characteristics to those of the target population
- Examining key demographic, socioeconomic, or clinical variables to identify systematic differences (comparing age, gender, or income distribution)
- Using external data sources or census information to benchmark the sample against the population
- Helps identify potential sources of selection bias and gauge the extent of the problem

Assessing missing data patterns

- Analyzing the patterns and mechanisms of missing data to detect potential selection bias
- Examining the relationship between missingness and key variables of interest (missing income data may be related to socioeconomic status)
- Using statistical tests or graphical methods to assess the randomness or non-randomness of missing data (Little's MCAR test, missing data patterns plot)
- Provides insights into the nature and potential impact of selection bias due to missing data

Sensitivity analysis techniques

- Conducting sensitivity analyses to assess the robustness of findings to different assumptions about selection bias
- Varying the assumptions about the missing data mechanism or the selection process to examine the impact on the results (assuming different scenarios for the values of missing data)

- Using methods such as propensity score matching or inverse probability weighting to adjust for selection bias under different assumptions
- Helps quantify the potential impact of selection bias and the sensitivity of conclusions to alternative assumptions

Addressing selection bias

Randomization in study design

- Using random assignment to allocate participants to treatment and control groups, ensuring that the groups are balanced on both observed and unobserved characteristics
- Minimizes selection bias by eliminating systematic differences between the groups at baseline
- Particularly effective in experimental studies, such as randomized controlled trials (RCTs)
- Requires careful implementation and monitoring to ensure the integrity of the randomization process

Weighting methods for adjustment

- Applying statistical weights to the sample data to make it more representative of the target population
- Using techniques such as inverse probability weighting (IPW) or propensity score weighting to adjust for selection bias based on observed characteristics
- Assigning higher weights to underrepresented groups and lower weights to overrepresented groups to balance the sample
- Requires accurate measurement of the relevant characteristics and appropriate specification of the weighting models

Imputation for missing data

- Using statistical methods to fill in missing values based on the observed data and assumptions about the missing data mechanism
- Applying techniques such as multiple imputation or maximum likelihood estimation to handle missing data in a principled manner
- Preserving the variability and uncertainty associated with the missing values, rather than relying on a single imputation
- Requires careful consideration of the missing data mechanism and the appropriateness of the imputation model

Bounds on effect estimates

- Calculating bounds or ranges for the causal effect estimates to account for potential selection bias
- Using methods such as Manski bounds or sensitivity analysis to determine the range of plausible effect sizes under different assumptions about the selection process
- Providing a more conservative and robust assessment of the causal relationship, acknowledging the uncertainty introduced by selection bias

- Helps convey the sensitivity of the findings to different scenarios and the limits of causal inference in the presence of selection bias

Selection bias vs confounding

Differences in causal structures

- Selection bias arises from the non-random selection of individuals into the sample or study groups, while confounding occurs when a third variable influences both the exposure and the outcome
- Selection bias is related to the sampling or selection process, whereas confounding is related to the causal relationships between variables
- Selection bias can occur even in the absence of confounding, and vice versa (a perfectly representative sample can still be subject to confounding)

Implications for bias direction

- The direction of bias introduced by selection bias depends on the specific nature of the selection process and its relationship to the exposure and outcome
- Selection bias can lead to overestimation or underestimation of the causal effect, depending on how the selection process is related to the variables of interest
- Confounding typically leads to bias in a specific direction, determined by the direction of the associations between the confounder, exposure, and outcome (positive confounding or negative confounding)

Strategies for distinguishing

- Assessing the plausibility and potential impact of selection bias requires careful consideration of the study design, sampling methods, and data collection processes
- Examining the causal structure and identifying potential confounders can help distinguish between selection bias and confounding
- Using directed acyclic graphs (DAGs) to visually represent the causal relationships and identify potential sources of bias
- Applying appropriate statistical methods to address selection bias (weighting, imputation) and confounding (adjustment, stratification, matching) based on the underlying causal structure and assumptions
- Conducting sensitivity analyses to assess the robustness of findings to different assumptions about selection bias and confounding

4.3 Controlling for confounding: Stratification and regression adjustment

Controlling for confounding is crucial in causal inference to obtain unbiased estimates of causal effects. Confounding variables can distort the true relationship between exposure and outcome, making it difficult to determine direct causal effects.

Two key methods for controlling confounding are stratification and regression adjustment. Stratification divides the study population into subgroups based on confounding variables, while regression adjustment includes confounders as covariates in statistical models. Both aim to isolate the direct causal effect of exposure on outcome.

Confounding variables

- Confounding variables are extraneous factors that influence both the exposure and outcome variables, potentially leading to biased estimates of the causal effect
- Confounding can distort the true relationship between the exposure and outcome, making it difficult to determine the direct causal effect
- Confounding variables can be measured or unmeasured, and they can arise from various sources such as demographic factors, lifestyle factors, or environmental factors

Definition of confounding

- Confounding occurs when a third variable is associated with both the exposure and outcome variables, creating a spurious relationship between them
- Confounding variables can mask the true causal effect or create the appearance of a causal effect when none exists
- Confounding can lead to overestimation or underestimation of the causal effect, depending on the direction and strength of the associations

Impact on causal inference

- Confounding can severely bias the estimates of causal effects, leading to incorrect conclusions about the relationship between the exposure and outcome
- Failing to account for confounding can result in erroneous policy decisions or ineffective interventions based on flawed causal inferences
- Confounding can also reduce the precision of effect estimates and decrease the power to detect true causal relationships

Examples of confounding

- Age is a common confounder in many epidemiological studies, as it is often associated with both the exposure (e.g., smoking) and the outcome (e.g., lung cancer)
- Socioeconomic status can confound the relationship between education and health outcomes, as it is related to both variables and may explain some of the observed association
- In a study examining the effect of alcohol consumption on liver disease, smoking could be a confounder if it is associated with both alcohol intake and liver disease risk

Controlling for confounding

- Controlling for confounding is crucial in causal inference to obtain unbiased estimates of the causal effect
- Researchers employ various methods to control for confounding, such as stratification, matching, regression adjustment, and propensity score methods
- Controlling for confounding helps isolate the direct causal effect of the exposure on the outcome, independent of other factors

Importance of controlling confounding

- Controlling for confounding is essential to ensure the internal validity of a study and to draw accurate conclusions about causal relationships
- Failing to control for confounding can lead to biased effect estimates and incorrect inferences about the effectiveness of interventions or the impact of exposures
- Adequate control of confounding is necessary to inform evidence-based decision-making in fields such as medicine, public health, and social policy

Methods for controlling confounding

- Researchers can control for confounding through study design, such as randomization in experiments or restriction and matching in observational studies
- Statistical methods for controlling confounding include stratification, regression adjustment, propensity score matching, and inverse probability weighting
- The choice of method depends on factors such as the nature of the confounding variables, the sample size, and the assumptions underlying each method

Stratification

- Stratification is a method for controlling confounding by dividing the study population into subgroups (strata) based on the levels of the confounding variable
- Within each stratum, the confounding variable is held constant, allowing for the estimation of the causal effect within each subgroup
- The stratum-specific effect estimates can then be combined using a weighted average to obtain an overall causal effect estimate adjusted for the confounder

Definition of stratification

- Stratification involves dividing the study population into mutually exclusive and exhaustive subgroups based on the levels of the confounding variable
- Each stratum represents a homogeneous subpopulation with respect to the confounder, eliminating its influence on the exposure-outcome relationship within the stratum
- Stratification allows for the estimation of the causal effect within each stratum, which can be combined to obtain an overall adjusted effect estimate

Stratification process

1. Identify the key confounding variable(s) to be controlled for
2. Divide the study population into strata based on the levels of the confounding variable (e.g., age groups, gender categories)
3. Estimate the causal effect within each stratum using appropriate measures (e.g., risk difference, risk ratio, odds ratio)
4. Combine the stratum-specific effect estimates using a weighted average (e.g., Mantel-Haenszel method) to obtain an overall adjusted effect estimate

5. Assess the homogeneity of the stratum-specific effect estimates to ensure that the confounding variable has been adequately controlled for

Stratification vs matching

- Stratification and matching are both methods for controlling confounding, but they differ in their approach and implementation
- Matching involves pairing exposed and unexposed individuals based on their similarity in terms of the confounding variable(s), creating a balanced subset of the original population
- Stratification, on the other hand, divides the entire study population into subgroups based on the levels of the confounding variable, without necessarily creating matched pairs
- Matching is typically used in the design phase of a study, while stratification can be applied in both the design and analysis phases

Limitations of stratification

- Stratification can lead to small sample sizes within each stratum, especially when controlling for multiple confounding variables simultaneously
- Small sample sizes within strata can result in imprecise effect estimates and wide confidence intervals, reducing the power to detect true causal effects
- Stratification may not be feasible when dealing with continuous confounding variables or a large number of confounding variables, as the number of strata grows exponentially
- Stratification does not control for unmeasured or unknown confounding variables, which can still bias the effect estimates

Regression adjustment

- Regression adjustment is a statistical method for controlling confounding by including the confounding variables as covariates in a regression model
- The regression model estimates the association between the exposure and outcome while adjusting for the effects of the confounding variables
- Regression adjustment allows for the estimation of the causal effect of the exposure on the outcome, holding the confounding variables constant

Definition of regression adjustment

- Regression adjustment involves fitting a regression model with the outcome variable as the dependent variable and the exposure and confounding variables as independent variables
- The regression coefficient for the exposure variable represents the adjusted causal effect, controlling for the confounding variables included in the model
- Common types of regression models used for causal inference include linear regression for continuous outcomes, logistic regression for binary outcomes, and Cox proportional hazards regression for time-to-event outcomes

Regression models for causal inference

- Linear regression: Used when the outcome variable is continuous and assumes a linear relationship between the exposure and outcome, adjusting for confounding variables
- Logistic regression: Used when the outcome variable is binary (e.g., disease status) and estimates the adjusted odds ratio or risk ratio for the exposure, controlling for confounding variables
- Cox proportional hazards regression: Used when the outcome variable is time-to-event data (e.g., survival time) and estimates the adjusted hazard ratio for the exposure, controlling for confounding variables and accounting for censoring

Interpreting regression coefficients

- The regression coefficient for the exposure variable represents the adjusted causal effect, holding the confounding variables constant
- In linear regression, the coefficient represents the change in the outcome variable associated with a one-unit increase in the exposure variable, adjusted for the confounding variables
- In logistic regression, the exponentiated coefficient (odds ratio or risk ratio) represents the adjusted relative change in the odds or risk of the outcome associated with the exposure, controlling for the confounding variables
- In Cox regression, the exponentiated coefficient (hazard ratio) represents the adjusted relative change in the hazard of the outcome associated with the exposure, controlling for the confounding variables

Assumptions of regression adjustment

- Linearity: The relationship between the exposure and outcome is linear (for linear regression) or can be transformed to be linear (e.g., using log-transformation)
- No multicollinearity: The exposure and confounding variables are not highly correlated with each other, which can lead to unstable coefficient estimates
- Homoscedasticity: The variance of the residuals is constant across all levels of the exposure and confounding variables (for linear regression)
- No omitted variable bias: All relevant confounding variables are included in the regression model, and there are no unmeasured or unknown confounders

Limitations of regression adjustment

- Regression adjustment relies on the correct specification of the functional form of the relationships between the variables (e.g., linearity assumption)
- Misspecification of the regression model can lead to biased effect estimates and incorrect inferences about the causal relationship
- Regression adjustment may not adequately control for confounding if there are unmeasured or unknown confounding variables not included in the model
- Regression adjustment can be sensitive to outliers or influential observations, which can distort the effect estimates

Stratification vs regression adjustment

- Stratification and regression adjustment are both methods for controlling confounding in observational studies, but they differ in their approach and underlying assumptions
- Stratification divides the study population into subgroups based on the levels of the confounding variable, while regression adjustment includes the confounding variables as covariates in a regression model
- The choice between stratification and regression adjustment depends on factors such as the nature of the confounding variables, the sample size, and the desired interpretation of the effect estimates

Similarities between methods

- Both stratification and regression adjustment aim to control for confounding and estimate the causal effect of the exposure on the outcome
- Both methods rely on the assumption of no unmeasured confounding, meaning that all relevant confounding variables are accounted for in the analysis
- Both methods can be used to estimate adjusted effect measures, such as risk differences, risk ratios, odds ratios, or hazard ratios

Differences between methods

- Stratification is a non-parametric method that does not assume a specific functional form for the relationships between variables, while regression adjustment relies on parametric assumptions (e.g., linearity)
- Stratification can handle categorical confounding variables more easily than continuous ones, while regression adjustment can accommodate both categorical and continuous confounding variables
- Stratification may lead to small sample sizes within strata, reducing the precision of effect estimates, while regression adjustment can handle larger numbers of confounding variables without necessarily reducing sample size
- Stratification provides stratum-specific effect estimates that can be combined into an overall adjusted estimate, while regression adjustment directly estimates the adjusted effect from the regression coefficient

Choosing the appropriate method

- Stratification may be preferred when the confounding variables are categorical and there are few levels, ensuring adequate sample sizes within each stratum
- Regression adjustment may be preferred when the confounding variables are continuous or there are many confounding variables to control for simultaneously
- Regression adjustment may be more appropriate when the goal is to estimate the adjusted effect while controlling for multiple confounding variables and potential effect modifiers
- A combination of stratification and regression adjustment can be used, such as stratifying on key categorical confounders and then applying regression adjustment within each stratum to control for additional continuous confounders

Assessing confounding control

- After applying methods to control for confounding, it is important to assess the adequacy of the confounding control and the robustness of the causal effect estimates
- Researchers can use various techniques to evaluate the balance of confounding variables between exposed and unexposed groups after adjustment and to assess the sensitivity of the results to potential unmeasured confounding
- Reporting the results of confounding control assessments is crucial for transparency and allowing readers to interpret the findings in light of the strengths and limitations of the methods used

Evaluating balance after adjustment

- Balance assessment involves comparing the distribution of confounding variables between exposed and unexposed groups after applying the confounding control method (e.g., stratification or regression adjustment)
- For categorical confounding variables, researchers can compare the proportions or frequencies of each level between the exposed and unexposed groups within each stratum or after regression adjustment
- For continuous confounding variables, researchers can compare the means, medians, or distributions of the variables between the exposed and unexposed groups after adjustment
- Adequate balance is achieved when the distributions of the confounding variables are similar between the exposed and unexposed groups after adjustment, indicating that the confounding has been effectively controlled for

Sensitivity analysis for unmeasured confounding

- Sensitivity analysis assesses the robustness of the causal effect estimates to potential unmeasured confounding variables that were not accounted for in the primary analysis
- One approach is to simulate the presence of an unmeasured confounder with varying strengths of association with the exposure and outcome and examine how the effect estimates change under different scenarios
- Another approach is to calculate the E-value, which represents the minimum strength of association that an unmeasured confounder would need to have with both the exposure and outcome to explain away the observed effect
- Sensitivity analyses help determine the conditions under which the causal conclusions would be altered by unmeasured confounding and provide a measure of the robustness of the findings

Reporting results after confounding control

- When reporting the results of a study that used methods to control for confounding, it is important to provide clear and transparent information about the methods used and the assumptions made
- Researchers should report the specific confounding control method(s) applied (e.g., stratification, regression adjustment), the variables used for adjustment, and any balance assessments conducted
- The adjusted effect estimates should be presented with their corresponding confidence intervals and p-values, along with the crude (unadjusted) estimates for comparison

- Any sensitivity analyses performed to assess the robustness of the findings to unmeasured confounding should be described, and the results should be interpreted in light of these analyses
- Limitations of the confounding control methods and the potential for residual confounding should be acknowledged and discussed in the context of the study's findings and conclusions

4.4 Sensitivity analysis

Sensitivity analysis is a crucial tool in causal inference, helping researchers assess how robust their findings are to potential violations of key assumptions. It involves systematically varying the strength of unmeasured confounders or simulating different scenarios to quantify uncertainty in causal estimates.

By conducting sensitivity analyses, researchers can transparently communicate the range of plausible causal effects under different assumptions. This approach allows for a more nuanced understanding of causal relationships, highlighting the importance of carefully considering untestable assumptions in observational studies.

Sensitivity analysis overview

Definition of sensitivity analysis

- Sensitivity analysis assesses the robustness of causal estimates to potential unmeasured confounding or violations of key assumptions
- Involves systematically varying the strength of unmeasured confounders or simulating potential outcomes under different scenarios
- Helps quantify the degree to which causal conclusions might change under alternative assumptions

Purpose in causal inference

- Causal inference relies on strong assumptions (exchangeability, positivity, consistency) that are often untestable in observational studies
- Sensitivity analysis provides a way to assess the plausibility of these assumptions and the robustness of causal estimates to potential violations
- Allows researchers to transparently communicate the uncertainty surrounding causal conclusions and identify the most critical assumptions

Methods for conducting sensitivity analysis

Varying unmeasured confounder strength

- One approach to sensitivity analysis is to posit the existence of an unmeasured confounder and vary its strength of association with treatment and outcome
- This can be done by simulating the confounder's distribution and specifying a range of plausible values for its effect on treatment and outcome
- The causal estimate is then re-calculated under each scenario to assess how much it changes as the unmeasured confounder strength varies

Simulating potential outcomes

- Another approach is to directly simulate the potential outcomes under different treatment assignments, based on the observed data and assumptions about the causal mechanism
- This allows for assessing the sensitivity of causal estimates to violations of the consistency assumption or the presence of interference between units
- The distribution of simulated potential outcomes can be used to construct bounds on the true causal effect

Bounds on treatment effects

- Sensitivity analysis can also be used to derive bounds on the treatment effect, rather than point estimates
- This involves specifying a range of plausible values for the sensitivity parameters (unmeasured confounder strength, potential outcome distributions) and calculating the corresponding range of possible treatment effects
- Bounds provide a more transparent way to communicate the uncertainty in causal estimates and the range of conclusions that are consistent with the data and assumptions

Interpreting sensitivity analysis results

Robustness of causal estimates

- The primary goal of sensitivity analysis is to assess the robustness of causal estimates to potential violations of key assumptions
- If the causal estimate remains relatively stable across a wide range of sensitivity parameters, this suggests that the conclusion is robust and less sensitive to unmeasured confounding or other violations
- Conversely, if the estimate changes substantially or even reverses sign under plausible scenarios, this indicates that the conclusion is more sensitive and uncertain

Threshold for effect reversal

- One key metric in sensitivity analysis is the threshold for effect reversal, which is the minimum strength of unmeasured confounding required to explain away the observed treatment effect
- This threshold can be expressed in terms of the association between the unmeasured confounder and treatment/outcome, or the proportion of variation in treatment/outcome that would need to be explained by the confounder
- A high threshold for reversal suggests that the causal conclusion is more robust, while a low threshold indicates greater sensitivity to unmeasured confounding

Communicating uncertainty

- Sensitivity analysis provides a way to transparently communicate the uncertainty surrounding causal conclusions, rather than presenting point estimates as definitive
- Results should be presented in terms of the range of plausible causal effects under different assumptions, along with the threshold for effect reversal

- This allows readers to assess the credibility of the causal claim and understand the degree to which it depends on untestable assumptions

Sensitivity analysis in practice

Selecting key assumptions to test

- In practice, sensitivity analysis requires carefully selecting the most critical assumptions to test, based on substantive knowledge and the potential for violations
- This may involve focusing on the exchangeability assumption and potential unmeasured confounders, or the consistency assumption and possible interference between units
- The choice of sensitivity parameters should be guided by prior literature, expert opinion, and the specific causal question and study design

Incorporating domain knowledge

- Sensitivity analysis is most informative when it incorporates domain knowledge to specify plausible ranges for sensitivity parameters
- This may involve using prior studies or theoretical arguments to bound the strength of unmeasured confounding, or using substantive expertise to rule out certain causal mechanisms
- Incorporating domain knowledge helps to ensure that sensitivity analyses are grounded in reality and not just mathematical exercises

Limitations of sensitivity analysis

- While sensitivity analysis is a valuable tool for assessing the robustness of causal conclusions, it has some important limitations
- Sensitivity analysis cannot prove that a causal estimate is unbiased, only that it is robust to certain violations of assumptions
- The range of scenarios considered in sensitivity analysis is necessarily limited, and there may be other sources of bias or uncertainty that are not captured
- Sensitivity analysis should be seen as a complement to, not a substitute for, careful study design and data collection to minimize potential confounding and other biases

Sensitivity analysis vs other approaches

Comparison to instrumental variables

- Instrumental variable (IV) methods provide an alternative approach to estimating causal effects in the presence of unmeasured confounding
- IV methods rely on finding a variable that affects treatment but not outcome except through treatment, and using this to estimate the causal effect
- Sensitivity analysis can be used to assess the robustness of IV estimates to violations of the exclusion restriction or other assumptions

Relation to bounding methods

- Bounding methods, such as the no-assumptions bounds or the Manski bounds, provide a way to estimate the range of possible causal effects consistent with the observed data alone
- These methods do not rely on untestable assumptions, but often yield very wide bounds that may not be informative
- Sensitivity analysis can be seen as a way to narrow these bounds by incorporating additional assumptions, while still acknowledging the uncertainty in these assumptions

Sensitivity analysis with multiple confounders

- Most sensitivity analysis methods focus on a single unmeasured confounder, but in reality there may be multiple sources of confounding
- Extending sensitivity analysis to multiple confounders requires specifying the joint distribution of the confounders and their effects on treatment and outcome
- This can quickly become intractable as the number of confounders grows, requiring simplifying assumptions or alternative approaches such as bias formulas or Bayesian methods

Advanced topics in sensitivity analysis

Sensitivity parameters for complex designs

- Sensitivity analysis for complex study designs, such as matched or stratified designs, requires specifying sensitivity parameters that capture the relevant features of the design
- For example, in a matched design, the sensitivity parameter may be the difference in unmeasured confounder prevalence between matched pairs, rather than the overall confounder-outcome association
- Sensitivity analysis for time-varying treatments or outcomes may require specifying parameters for the time-varying confounding or selection bias

Bayesian sensitivity analysis

- Bayesian methods provide a natural framework for incorporating uncertainty about sensitivity parameters into causal estimates
- Prior distributions can be specified for the sensitivity parameters, reflecting the range of plausible values based on domain knowledge
- The posterior distribution of the causal effect can then be calculated, marginalizing over the sensitivity parameters to reflect the overall uncertainty

Sensitivity analysis with missing data

- Missing data is a common problem in observational studies, and can be a source of bias if the missingness is related to treatment, outcome, or confounders
- Sensitivity analysis can be used to assess the robustness of causal estimates to different assumptions about the missing data mechanism
- This may involve specifying a range of plausible values for the missing data, or using pattern mixture models to estimate causal effects under different missingness scenarios

Unit 5 – Matching and propensity scores

5.1 Propensity score

Definition of propensity scores

The propensity score is the probability that a subject receives treatment, conditional on their observed baseline characteristics. In notation, for subject i with covariates X_i :

$$e(X_i) = P(T_i = 1 \mid X_i)$$

In a randomized experiment, every subject has a known probability of being assigned to treatment. Observational studies lack this design, so the propensity score estimates that probability from the data. The core idea comes from Rosenbaum and Rubin (1983): if two subjects have the same propensity score, their observed covariates are, on average, balanced across treatment groups. This means you can condition on a single scalar (the propensity score) instead of trying to match or adjust on many covariates simultaneously.

Propensity scores only balance on observed covariates. They cannot fix confounding from variables you didn't measure.

Estimating propensity scores

Logistic regression for propensity score estimation

The most common approach is logistic regression with treatment assignment as the outcome and observed covariates as predictors:

$$\log\left(\frac{e(X_i)}{1 - e(X_i)}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$$

The predicted probabilities from this model are your estimated propensity scores. Logistic regression is straightforward to implement and lets you inspect how each covariate relates to treatment assignment through the estimated coefficients.

One thing to keep in mind: the goal here is prediction of treatment assignment, not inference about individual coefficients. You're not testing whether a covariate "significantly" predicts treatment. You just want the best possible estimate of each subject's treatment probability.

Machine learning methods for propensity score estimation

Methods like random forests, gradient boosting (e.g., GBM, XGBoost), and other flexible algorithms can also estimate propensity scores. These approaches:

- Capture non-linear relationships and interactions automatically
- Can handle high-dimensional covariate spaces
- May outperform logistic regression when the true relationship between covariates and treatment is complex

The tradeoff is reduced interpretability and the risk of overfitting, which can produce extreme propensity scores near 0 or 1. Regardless of the estimation method, what ultimately matters is whether the resulting scores achieve good covariate balance.

Assessing propensity score balance

Estimating propensity scores is not the end goal. You need to verify that conditioning on them actually balances covariates between groups.

Visual assessment of propensity score balance

- Plot the distribution of estimated propensity scores for treated and control groups using histograms, density plots, or box plots
- Look for overlap (common support): both groups should have propensity scores across a similar range. Regions where one group has scores but the other doesn't indicate positivity violations
- After applying your propensity score method (matching, weighting, etc.), re-plot covariate distributions to confirm they've become more similar

Statistical assessment of propensity score balance

The standardized mean difference (SMD) is the preferred balance metric. For a covariate X :

$$SMD = \frac{\bar{X}_{treated} - \bar{X}_{control}}{\sqrt{(s_{treated}^2 + s_{control}^2)/2}}$$

Calculate SMDs before and after propensity score adjustment to quantify improvement. A common rule of thumb is that SMDs below 0.1 indicate adequate balance, though this threshold isn't absolute.

Avoid relying on p-values from t-tests or chi-square tests for balance checking. These tests are sensitive to sample size: with large samples, trivial imbalances appear "significant," while with small samples, real imbalances can be missed. SMDs are sample-size-independent and more informative.

Propensity score methods

Once you have estimated propensity scores, there are four main ways to use them.

Matching on propensity scores

Match each treated subject to one or more control subjects with similar propensity scores. Common approaches:

- Nearest neighbor matching: pair each treated subject with the control subject whose propensity score is closest
- Caliper matching: same as nearest neighbor, but only allow matches within a specified distance (e.g., 0.2 standard deviations of the logit propensity score)
- Optimal matching: minimize the total distance across all matched pairs simultaneously

After matching, you estimate the treatment effect by comparing outcomes between matched treated and control subjects. Matching can discard unmatched subjects, which reduces sample size but typically targets the ATT (average treatment effect on the treated).

Stratification on propensity scores

Divide subjects into strata (often 5 to 10) based on propensity score quantiles. Within each stratum, treated and control subjects should have roughly similar covariate distributions. You then:

1. Estimate the treatment effect within each stratum

2. Combine stratum-specific estimates into an overall effect (typically a weighted average, with weights proportional to stratum size)

Stratification retains all subjects, but balance within strata is only approximate. Using more strata improves balance but reduces the sample size per stratum.

Inverse probability weighting (IPW)

IPW reweights subjects to create a pseudo-population where treatment is independent of observed covariates. The weights are:

- Treated subjects: $W_i = \frac{1}{e(X_i)}$
- Control subjects: $W_i = \frac{1}{1 - e(X_i)}$

These weights target the ATE (average treatment effect). If you want the ATT instead, use weights of 1 for treated subjects and $e(X_i)/(1 - e(X_i))$ for controls.

IPW retains all subjects and is computationally simple, but it's sensitive to extreme propensity scores. A subject with $e(X_i) = 0.01$ gets a weight of 100, which can make estimates unstable. Stabilized weights (multiplying by the marginal probability of treatment) and weight trimming help address this.

Propensity score as a covariate

Include the estimated propensity score directly as a covariate in a regression model for the outcome. This is the simplest approach, but it relies heavily on correctly specifying the outcome model and tends to be less effective at reducing bias than matching or weighting. It's generally considered the weakest of the four methods.

Advantages vs. disadvantages of propensity score methods

Advantages: - Reduce a high-dimensional covariate set to a single balancing score - Allow transparent, pre-outcome assessment of covariate balance (you can check balance before ever looking at the outcome) - Separate the "design" phase (achieving balance) from the "analysis" phase (estimating the effect), reducing reliance on outcome model assumptions

Disadvantages: - Only balance on observed covariates; unmeasured confounding remains a threat - Require correct specification of the propensity score model - Can reduce effective sample size (matching may discard subjects; extreme weights in IPW inflate variance)

Propensity scores vs. regression adjustment

Regression adjustment models the outcome directly as a function of treatment and covariates. Propensity score methods instead model the treatment assignment mechanism. Key differences:

- Propensity scores let you assess balance before analyzing outcomes. With regression, you're adjusting and estimating simultaneously, so you can't easily tell whether the adjustment "worked."
- Propensity score methods can be more robust when the outcome model is misspecified, since they don't require you to get the outcome-covariate relationship right.
- Regression adjustment can be more statistically efficient when the outcome model is correctly specified, because it uses information about the outcome-covariate relationship.
- The two approaches can be combined ("doubly robust" estimation): if either the propensity score model or the outcome model is correct, the treatment effect estimate is consistent.

Propensity scores with multiple treatments

When there are more than two treatment levels (e.g., three different drugs), you estimate a generalized propensity score: the probability of receiving each specific treatment level given observed covariates. For treatment level t :

$$e_t(X_i) = P(T_i = t \mid X_i)$$

These can be estimated using multinomial logistic regression or a series of binary models. Generalized propensity scores support matching, stratification, or weighting across multiple treatment groups and enable pairwise treatment comparisons or dose-response estimation.

Propensity scores with time-varying treatments

When treatment changes over time (e.g., a patient starts and stops medication across multiple visits), standard propensity scores don't suffice. Instead, you estimate time-varying propensity scores at each time point, conditional on the covariate and treatment history up to that point.

These are most commonly used with inverse probability of treatment weighting (IPTW) to handle time-varying confounding, especially when past treatment affects future confounders. This connects to marginal structural models, where the weights create a pseudo-population free of time-varying confounding.

Propensity score methods in practice

Selecting variables for propensity score estimation

Variable selection matters a lot. The guiding principles:

- Include all variables that are confounders (associated with both treatment and outcome)
- Include variables strongly associated with the outcome, even if weakly related to treatment. These improve precision without introducing bias.
- Exclude variables that are consequences of treatment (post-treatment variables), as conditioning on them can introduce bias
- Exclude instrumental variables (strongly predict treatment but have no direct effect on outcome), as including them can increase variance without reducing bias
- Use causal diagrams (DAGs) and subject matter knowledge to guide these decisions

Specifying the propensity score model

1. Start with a simple model using main effects of all selected covariates
2. Estimate propensity scores and check covariate balance
3. If balance is inadequate, add interaction terms, quadratic terms, or other non-linear transformations
4. Re-check balance after each modification
5. Use cross-validation if employing machine learning methods to avoid overfitting

The iterative nature of this process is a feature, not a bug. You keep refining until balance is satisfactory.

Checking propensity score overlap

Overlap (also called positivity or common support) means that for every combination of covariate values, there's a non-zero probability of receiving either treatment or control. In practice:

- Plot propensity score distributions for both groups and look for regions where they overlap
- If some subjects have propensity scores very near 0 or 1, they have near-certain treatment assignment, making causal comparisons unreliable in that region
- Consider trimming subjects with extreme scores (e.g., dropping those with scores below 0.05 or above 0.95) to improve overlap and reduce sensitivity to model misspecification

Sensitivity analysis for propensity score methods

Because propensity scores can't address unmeasured confounding, sensitivity analysis is critical:

- Compare results across different propensity score methods (matching, stratification, IPW) to see if conclusions are robust
- Vary the propensity score model specification (different covariates, functional forms) and check stability of estimates
- Use formal sensitivity analysis frameworks (e.g., Rosenbaum bounds) to quantify how strong unmeasured confounding would need to be to change your conclusions
- Report the range of estimates under different assumptions to honestly convey uncertainty

5.2 Matching methods

Matching methods are powerful tools for estimating causal effects in observational studies. They aim to create balanced treatment and control groups by pairing similar units based on observed characteristics, mimicking randomized experiments.

These techniques rely on key assumptions like SUTVA, unconfoundedness, and positivity. Various algorithms, from greedy to optimal matching, can be used. Assessing balance and estimating treatment effects are crucial steps in the matching process.

Overview of matching methods

- Matching methods are a set of techniques used in causal inference to estimate treatment effects by creating balanced treatment and control groups
- These methods aim to mimic the randomization process in observational studies by matching treated and control units on observed covariates
- Matching methods can be used to estimate various causal estimands, such as the average treatment effect (ATE) and the average treatment effect on the treated (ATT)

Assumptions for matching

Stable unit treatment value assumption (SUTVA)

- SUTVA assumes that the potential outcomes for each unit are unaffected by the treatment assignment of other units

- This assumption rules out interference between units and ensures that there are no hidden versions of the treatment
- Violations of SUTVA can lead to biased estimates of treatment effects

Unconfoundedness assumption

- The unconfoundedness assumption, also known as the ignorability assumption, states that treatment assignment is independent of potential outcomes given the observed covariates
- This assumption implies that there are no unobserved confounders that influence both treatment assignment and outcomes
- Meeting the unconfoundedness assumption is crucial for the validity of matching methods

Positivity assumption

- The positivity assumption requires that each unit has a non-zero probability of receiving each level of the treatment given their observed covariates
- This assumption ensures that there is sufficient overlap between the treatment and control groups in terms of their covariate distributions
- Violations of the positivity assumption can lead to extrapolation and unreliable estimates of treatment effects

Matching algorithms

Greedy matching

- Greedy matching is a simple and fast algorithm that matches each treated unit to the nearest control unit based on a distance metric (Mahalanobis distance)
- This algorithm proceeds sequentially, making the best match for each treated unit without considering the overall balance of the matched sample
- Greedy matching can be implemented with or without replacement of control units

Optimal matching

- Optimal matching aims to minimize the total distance between matched pairs while ensuring that each treated unit is matched to a control unit
- This algorithm uses optimization techniques, such as network flow algorithms, to find the globally optimal match
- Optimal matching can be computationally intensive for large datasets but often results in better balance than greedy matching

Genetic matching

- Genetic matching is an adaptive algorithm that uses a genetic search algorithm to find the optimal weight for each covariate in the distance metric
- This method iteratively improves the balance of the matched sample by modifying the distance metric based on the balance achieved in previous iterations

- Genetic matching can automatically select the most important covariates for achieving balance

Coarsened exact matching (CEM)

- CEM involves coarsening the continuous covariates into discrete categories and then performing exact matching on the coarsened data
- This method ensures that matched units have identical covariate values within each coarsened category
- CEM can be useful when there are many covariates or when exact matching on the original scale is infeasible

Assessing balance after matching

Standardized mean differences

- Standardized mean differences (SMDs) measure the difference in means between the treatment and control groups for each covariate, divided by the pooled standard deviation
- SMDs are commonly used to assess balance, with values close to zero indicating good balance
- A common rule of thumb is that SMDs should be less than 0.1 for the matched sample to be considered well-balanced

Variance ratios

- Variance ratios compare the variances of each covariate between the treatment and control groups
- Ideally, the variance ratios should be close to one, indicating that the spread of the covariate is similar in both groups
- Large deviations from one suggest imbalance and may require further adjustment or a different matching algorithm

Kolmogorov-Smirnov tests

- Kolmogorov-Smirnov (KS) tests compare the empirical cumulative distribution functions (ECDFs) of each covariate between the treatment and control groups
- A significant KS test suggests that the distributions of the covariate differ between the groups, indicating imbalance
- KS tests can detect differences in the shape of the distributions that may not be captured by SMDs or variance ratios

Graphical diagnostics

- Visual inspection of covariate balance can be done using various plots, such as histograms, density plots, or QQ plots
- These plots allow for a more intuitive assessment of balance and can help identify any remaining differences between the groups
- Examples of graphical diagnostics include side-by-side boxplots, overlaid density plots, and love plots (a grid of SMDs before and after matching)

Estimating treatment effects

Average treatment effect (ATE)

- The ATE is the expected difference in outcomes between the treatment and control groups across the entire population
- Estimating the ATE requires that the matched sample is representative of the overall population
- The ATE can be estimated by taking the difference in mean outcomes between the matched treatment and control groups

Average treatment effect on the treated (ATT)

- The ATT is the expected difference in outcomes between the treatment and control groups, focusing only on the population of treated units
- Estimating the ATT is often of interest when the treatment is not randomly assigned and there is selection into treatment based on observed covariates
- The ATT can be estimated by taking the difference in mean outcomes between the treated units and their matched controls

Subgroup analysis

- Subgroup analysis involves estimating treatment effects for specific subpopulations defined by one or more covariates (age groups)
- This analysis can help identify heterogeneous treatment effects and provide insights into which subgroups benefit most from the treatment
- When conducting subgroup analysis, it is important to pre-specify the subgroups of interest and adjust for multiple testing to avoid false positives

Advantages of matching methods

Intuitive approach

- Matching methods mimic the randomization process by creating balanced treatment and control groups, making the causal inference more intuitive and easier to communicate
- The matched sample can be thought of as a "virtual RCT," where the only remaining difference between the groups is the treatment itself
- This intuitive approach can be particularly appealing to non-technical audiences and decision-makers

Transparency in matched samples

- Matching methods provide a clear and transparent way to create balanced treatment and control groups
- The matched sample can be easily inspected and compared to the original sample to assess the quality of the matching and the resulting balance
- This transparency allows for greater scrutiny of the causal inference process and can increase confidence in the results

Flexibility in estimands

- Matching methods can be used to estimate various causal estimands, such as the ATE and ATT, depending on the research question and the population of interest
- The choice of estimand can be tailored to the specific context and the available data
- This flexibility allows researchers to answer a wide range of causal questions using the same matching framework

Limitations of matching methods

Curse of dimensionality

- As the number of covariates used for matching increases, the difficulty of finding good matches grows exponentially
- This phenomenon, known as the curse of dimensionality, can lead to poor balance and reduced sample size when matching on a large number of covariates
- Techniques such as coarsened exact matching (CEM) and propensity score matching can help mitigate this issue by reducing the dimensionality of the matching problem

Reduced sample size

- Matching methods often result in a reduced sample size, as some units may not have suitable matches and are therefore discarded
- This reduction in sample size can lead to decreased statistical power and wider confidence intervals for the estimated treatment effects
- The impact of reduced sample size should be considered when interpreting the results and planning the study design

Sensitivity to matching algorithm choice

- Different matching algorithms can produce different matched samples and, consequently, different estimates of treatment effects
- The choice of matching algorithm, distance metric, and tuning parameters can have a substantial impact on the results
- It is important to assess the sensitivity of the results to these choices and to justify the selected matching approach based on the research question and the characteristics of the data

Matching vs other methods

Matching vs regression adjustment

- Regression adjustment involves controlling for observed covariates in a regression model to estimate treatment effects
- Matching can be seen as a nonparametric preprocessing step that creates a balanced sample, which can then be analyzed using regression adjustment

- Matching followed by regression adjustment can help ensure that the regression model is not extrapolating beyond the observed data and can improve the robustness of the estimates

Matching vs propensity score methods

- Propensity score methods involve estimating the probability of treatment assignment given the observed covariates and using these probabilities for matching, stratification, or weighting
- Matching can be performed directly on the propensity scores, which reduces the dimensionality of the matching problem
- Propensity score methods can be more efficient than matching on the original covariates when there are many covariates, but they rely on the correct specification of the propensity score model

Matching vs inverse probability weighting

- Inverse probability weighting (IPW) involves weighting each unit by the inverse of its probability of receiving the observed treatment, given the covariates
- IPW can be used to estimate the ATE by creating a pseudo-population in which treatment assignment is independent of the covariates
- Matching can be more robust to model misspecification than IPW, but IPW can be more efficient when the propensity score model is correctly specified and there is good overlap between the treatment and control groups

Sensitivity analysis

Rosenbaum bounds

- Rosenbaum bounds are a method for assessing the sensitivity of the estimated treatment effects to unobserved confounding
- This approach calculates the magnitude of hidden bias that would be necessary to explain away the observed treatment effect
- A study is considered sensitive to unobserved confounding if a relatively small amount of hidden bias could alter the conclusions

Simulation-based sensitivity analysis

- Simulation-based sensitivity analysis involves simulating the impact of unobserved confounders on the estimated treatment effects
- This approach requires specifying the distribution and strength of the unobserved confounder and its relationship with the treatment and outcome
- By varying the characteristics of the simulated confounder, researchers can assess the robustness of the results to different scenarios of unobserved confounding

Software for matching

R packages for matching

- Several R packages are available for implementing matching methods, such as MatchIt, Matching, and optmatch

- These packages provide a wide range of matching algorithms, diagnostic tools, and functions for estimating treatment effects
- The MatchIt package is particularly popular due to its ease of use and compatibility with other causal inference packages, such as cobalt for balance assessment

Stata commands for matching

- Stata offers several commands for matching, including `teffects nnmatch` for nearest-neighbor matching and `teffects psmatch` for propensity score matching
- The `kmatch` command is a versatile tool that implements various matching algorithms, such as Mahalanobis distance matching and coarsened exact matching
- Stata's `tebalance` command provides a range of diagnostic tools for assessing balance before and after matching

Python libraries for matching

- Python has a growing ecosystem of libraries for causal inference, including matching methods
- The CausalInference library provides a simple interface for propensity score matching and balance assessment
- The causalml library offers a range of matching algorithms, including nearest-neighbor matching, optimal matching, and coarsened exact matching, along with tools for estimating treatment effects and conducting sensitivity analyses

5.3 Inverse probability weighting

Overview of inverse probability weighting

Inverse probability weighting (IPW) is a technique for estimating causal effects from observational data. The core problem it solves: in observational studies, the people who get treated differ systematically from those who don't, so a naive comparison of outcomes is confounded. IPW fixes this by reweighting observations so that the confounders are balanced across treatment groups, effectively creating a pseudo-population that mimics what you'd see in a randomized experiment.

The weight assigned to each observation is the inverse of the probability that the person received the treatment they actually got, given their covariates. If a treated person had a low probability of being treated (based on their characteristics), they get a large weight because they represent many similar people who were not treated. This upweighting of "unlikely" observations is what drives the balance.

Motivation for the weighting approach

In a randomized experiment, treatment is assigned independently of confounders, so you can compare outcomes directly. Observational data doesn't give you that luxury. Confounders influence who gets treated, which biases simple comparisons.

IPW tackles this by asking: given someone's covariates, how likely were they to receive the treatment they got? Then it weights each person inversely by that probability. The result is a weighted sample where the distribution of confounders looks roughly the same in the treated and untreated groups. You're not removing confounded individuals from the data; you're re-balancing the entire sample through weights.

Assumptions of IPW

Positivity assumption

Every individual, regardless of their covariate values, must have a non-zero probability of receiving each treatment level. Formally:

$$P(A = a \mid X = x) > 0 \quad \text{for all } a \text{ and } x$$

If some region of the covariate space has zero (or near-zero) probability of a particular treatment, the corresponding weights blow up. This is called a positivity violation, and it leads to extreme weights and unstable estimates. Practical near-violations (propensity scores very close to 0 or 1) cause similar problems even when strict positivity technically holds.

Exchangeability assumption

Also called the no unmeasured confounders assumption. It requires that, conditional on the observed covariates X , treatment assignment is independent of the potential outcomes:

$$Y(a) \perp A \mid X \quad \text{for all } a$$

This means every variable that jointly affects both treatment and outcome must be measured and included in the propensity score model. If an unmeasured confounder exists, the weights won't balance it, and your causal estimate will be biased. This assumption is untestable from the data alone, which is why subject-matter knowledge is so important when specifying the model.

Estimating weights

Propensity score models

The propensity score is the probability of receiving treatment given covariates:

$$e(X) = P(A = 1 \mid X)$$

For binary treatments, this is most commonly estimated with logistic regression, where treatment status is the outcome and the confounders are predictors. Variable selection matters a lot here. You should include true confounders (variables that affect both treatment and outcome). Omitting an important confounder introduces bias; including variables that only predict the outcome (not treatment) can actually improve efficiency, but including instruments (variables that affect treatment but not outcome) can increase variance.

Stabilized vs. unstabilized weights

Unstabilized weights for individual i :

$$w_i = \frac{A_i}{e(X_i)} + \frac{1 - A_i}{1 - e(X_i)}$$

A treated person gets weight $1/e(X_i)$, and an untreated person gets weight $1/(1 - e(X_i))$. These weights can be highly variable, especially when propensity scores are near 0 or 1.

Stabilized weights replace the numerator of 1 with the marginal probability of the treatment actually received:

$$sw_i = \frac{A_i \cdot P(A = 1)}{e(X_i)} + \frac{(1 - A_i) \cdot P(A = 0)}{1 - e(X_i)}$$

Stabilized weights have the same expectation for balancing confounders but lower variance. They also produce a pseudo-population closer in size to the original sample. In practice, stabilized weights are preferred because they yield more efficient and stable estimates.

Fitting outcome models

Weighted regression

Once you have the weights, you fit a regression model for the outcome with each observation weighted by its IPW. For continuous outcomes, use weighted least squares. For binary outcomes, use weighted logistic regression.

The treatment variable should be included as a predictor. If residual imbalance remains after weighting (checked via balance diagnostics), you may also include additional covariates in the outcome model, though the primary confounding adjustment comes from the weights.

Estimating causal effects

The coefficient on the treatment variable in the weighted model estimates the average treatment effect (ATE) in the pseudo-population. For a binary treatment, this is the difference in expected outcomes between treated and untreated groups after reweighting.

Standard errors from a naive weighted regression are typically too small because they ignore the uncertainty in estimating the weights. Use robust (sandwich) standard errors to get valid confidence intervals. Bootstrapping is another option that accounts for the full estimation process, including the propensity score step.

Assessing covariate balance

Standardized mean differences

After applying weights, you need to verify that confounders are actually balanced. The standardized mean difference (SMD) for each covariate compares the weighted means between treatment groups, scaled by the pooled standard deviation.

- SMD close to 0 indicates good balance.
- The conventional threshold is $\text{SMD} < 0.1$ for adequate balance.
- Check SMDs for every confounder, not just a few. A Love plot (SMDs before and after weighting, plotted side by side) is a useful visual tool.

Variance ratios

Balancing means isn't enough; you should also check that the distributions have similar spread. The variance ratio (VR) divides the weighted variance of a covariate in the treated group by its weighted variance in the untreated group.

- VR close to 1 indicates good balance.
- Values far from 1 suggest the weighting hasn't adequately balanced higher-order features of the distribution.

If substantial imbalances persist, revisit the propensity score model. Consider adding interaction terms, nonlinear terms, or using a more flexible estimation method (e.g., generalized boosted models).

Comparison to other methods

IPW vs. matching

Both IPW and matching use propensity scores to address confounding, but they do so differently. Matching pairs (or groups) treated and untreated individuals with similar propensity scores and discards unmatched units. IPW keeps all observations and reweights them instead.

- Matching is often more intuitive to explain and can be easier to diagnose visually.
- IPW retains the full sample, which can improve efficiency.
- Matching results can be sensitive to the choice of algorithm (nearest-neighbor, caliper width, with/without replacement).
- IPW results can be sensitive to extreme propensity scores.

IPW vs. stratification

Stratification divides the sample into subgroups (often quintiles) based on propensity scores and estimates treatment effects within each stratum. The overall ATE is a weighted average across strata.

- Stratification is more robust to propensity score model misspecification because it only requires approximate ranking, not exact probability estimates.
- However, stratification can leave residual confounding within strata if the strata are too coarse (e.g., using only 5 strata for a complex confounder structure).
- IPW generally provides more precise estimates when the propensity score model is well-specified.

Limitations of IPW

Sensitivity to model misspecification

IPW's validity depends on the propensity score model being correctly specified. If you omit a confounder or get the functional form wrong (e.g., assuming linearity when the true relationship is nonlinear), the weights won't properly balance the confounders, and your causal estimate will be biased.

Sensitivity analyses can help. Try varying the set of included confounders, adding interaction or polynomial terms, or using flexible machine learning methods for propensity score estimation. If results change substantially, that's a warning sign.

Instability with extreme weights

When propensity scores are near 0 or 1, the resulting weights become very large, and a small number of observations can dominate the analysis. This inflates variance and makes estimates unreliable.

Common remedies:

1. Weight truncation: cap weights at a chosen percentile (e.g., the 99th percentile) so no single observation has outsized influence.
2. Weight trimming: exclude observations with propensity scores below or above certain thresholds (e.g., drop anyone with $e(X) < 0.01$ or $e(X) > 0.99$).
3. Stabilized weights: as discussed above, these naturally reduce weight variability.

Both truncation and trimming introduce some bias because they effectively change the target population. Use them judiciously and report the sensitivity of your results to the chosen threshold.

Applications of IPW

Time-varying treatments

IPW extends naturally to settings where treatment changes over time. At each time point, you estimate the probability of the treatment actually received, conditional on the full covariate and treatment history up to that point. The weight for each individual is the product of the inverse probabilities across all time points.

This is where IPW really shines relative to standard regression, because time-varying confounders that are also affected by prior treatment create a problem that conventional adjustment can't handle without introducing bias.

Marginal structural models

Marginal structural models (MSMs) are the modeling framework most commonly paired with IPW for time-varying treatments. An MSM specifies the marginal mean of the potential outcome as a function of the treatment history, without conditioning on the time-varying confounders.

The key steps:

1. Estimate the probability of treatment at each time point, given covariate and treatment history.
2. Construct stabilized weights as the cumulative product of time-point-specific weights.
3. Fit a weighted outcome model (e.g., weighted GEE or weighted pooled logistic regression) with the treatment history as predictors.

MSMs also typically incorporate inverse probability of censoring weights to handle loss to follow-up, multiplying treatment weights by censoring weights.

Simulation studies of IPW

Evaluating bias reduction

Simulation studies let you test IPW under controlled conditions where the true causal effect is known. You generate data with a specified causal structure, apply IPW, and compare the estimated effect to the truth.

This approach is valuable for understanding how IPW performs when assumptions are met versus violated. For example, you can simulate data with an unmeasured confounder and quantify how much bias remains, or introduce near-positivity violations and observe the resulting instability.

Comparing to other estimators

Simulations also allow head-to-head comparisons of IPW against matching, stratification, g-computation, and doubly robust methods. The usual metrics are bias, variance, mean squared error (MSE), and confidence interval coverage.

A consistent finding across many simulation studies: IPW performs well when the propensity score model is correct but can be outperformed by doubly robust methods when there's a risk of model misspecification. No single method dominates in all scenarios, which is why understanding the tradeoffs matters.

Extensions of IPW

Doubly robust estimation

Doubly robust (DR) estimators combine a propensity score model with an outcome model. The key property: the estimate is consistent if either model is correctly specified (though not necessarily both). This gives you two chances to get it right.

A common DR estimator is the augmented IPW (AIPW) estimator, which starts with an IPW estimate and adds a correction term based on the outcome model. If the propensity score model is wrong but the outcome model is right, the correction term removes the bias, and vice versa. DR estimators also tend to be more efficient than IPW alone when both models are at least approximately correct.

Targeted maximum likelihood estimation

Targeted maximum likelihood estimation (TMLE) is another doubly robust approach with some additional advantages. The procedure works in three stages:

1. Fit an initial outcome model (e.g., using machine learning methods like Super Learner).
2. Update the initial model using a "clever covariate" derived from the propensity score. This targeting step optimizes the estimate specifically for the causal parameter of interest.
3. Average the targeted predictions to obtain the ATE estimate.

TMLE is doubly robust and, under regularity conditions, achieves the semiparametric efficiency bound. It also respects the bounds of the outcome (e.g., predicted probabilities stay between 0 and 1), which AIPW does not always guarantee. These properties make TMLE a strong default choice for causal inference in observational studies.

5.4 Doubly robust estimation

Doubly robust estimation overview

Doubly robust estimation combines outcome regression and inverse probability weighting (IPW) into a single estimator for causal effects in observational studies. The core appeal: your estimate remains consistent if either the outcome model or the propensity score model is correctly specified. You don't need both to be right, just one. This "two chances to get it right" property makes doubly robust methods a major improvement over relying on a single modeling strategy, where one wrong specification means biased results.

Potential outcome framework

Notation and setup

The potential outcomes framework underpins all doubly robust methods. For each individual, define:

- $Y(1)$: the outcome that would be observed under treatment ($A = 1$)
- $Y(0)$: the outcome that would be observed under control ($A = 0$)

Only one potential outcome is ever observed. The observed outcome links to the potential outcomes through:

$$Y = A \cdot Y(1) + (1 - A) \cdot Y(0)$$

The target estimand is the average treatment effect (ATE):

$$ATE = E[Y(1) - Y(0)]$$

Assumptions for causal inference

Three assumptions must hold for these methods to identify causal effects:

- Consistency: the observed outcome for an individual equals the potential outcome under the treatment they actually received. No hidden versions of treatment.
- Positivity: every individual has a non-zero probability of receiving each treatment level, given their covariates. If some subgroup always gets treated, you can't estimate what would have happened without treatment for that subgroup.
- Conditional exchangeability (no unmeasured confounding): treatment assignment is independent of potential outcomes given the measured covariates X . All confounders are captured in X .

Inverse probability weighting

Propensity score estimation

The propensity score is the probability of receiving treatment given observed covariates:

$$e(X) = P(A = 1 \mid X)$$

It's typically estimated with logistic regression, though machine learning methods (random forests, gradient boosting) are also used. The propensity score serves to balance the covariate distribution between treatment and control groups: within strata of similar propensity scores, treated and untreated individuals should look comparable on observed covariates.

Weighted estimators

The IPW estimator for the ATE is:

$$\hat{\psi}_{IPW} = \frac{1}{n} \sum_{i=1}^n \left(\frac{A_i Y_i}{e(X_i)} - \frac{(1 - A_i) Y_i}{1 - e(X_i)} \right)$$

Each individual's outcome is weighted by the inverse of their probability of receiving the treatment they actually got. This creates a pseudo-population where treatment assignment is independent of covariates, so a simple weighted comparison of outcomes estimates the ATE.

Limitations of IPW

- Entirely dependent on correct specification of the propensity score model. If that model is wrong, the weights are wrong, and the estimate is biased.
- Extreme weights arise when propensity scores are near 0 or 1, inflating the variance and making estimates unstable. A single observation with $e(X) = 0.01$ gets a weight of 100.
- IPW doesn't model the outcome at all, so it can't leverage information about the outcome-covariate relationship to improve precision.

Outcome regression

Regression adjustment

Outcome regression directly models the expected outcome as a function of treatment and covariates. You fit separate models for each treatment group:

$$\bullet \hat{\mu}_1(X) = \hat{E}[Y \mid A = 1, X]$$

$$\bullet \hat{\mu}_0(X) = \hat{E}[Y \mid A = 0, X]$$

Then for each individual, predict their outcome under both treatment conditions (regardless of which they actually received) and average the differences:

$$\hat{\psi}_{\text{OR}} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$$

Limitations of outcome regression

- Relies on correct specification of the outcome model. With high-dimensional or nonlinear data, getting this right is hard.
- Doesn't directly balance covariates between groups, so the estimate depends heavily on the model's ability to extrapolate.
- Extrapolation bias is a real concern: if treated individuals have covariate values rarely seen among controls, the outcome model is predicting outside its training data.

Doubly robust estimators

Combining IPW and outcome regression

Doubly robust estimators use both the propensity score and the outcome model in a single estimating equation. The key insight is that the outcome model "augments" the IPW estimator: when the propensity score model is slightly off, the outcome model compensates, and vice versa.

The result is an estimator that's consistent if either model is correctly specified. You still want both to be right (that gives you the best performance), but you only need one.

Semiparametric efficiency

When both models are correctly specified, doubly robust estimators achieve the semiparametric efficiency bound. This is the smallest possible asymptotic variance among all regular estimators for the ATE. In practical terms, this means narrower confidence intervals and more precise estimates than you'd get from IPW or outcome regression alone.

Robustness to model misspecification

The double robustness property is especially valuable in observational studies where the true data-generating process is unknown. You're hedging your bets: even if your propensity score model captures the treatment assignment mechanism but your outcome model is wrong (or the reverse), the estimator still converges to the truth. This is a strict improvement over methods that give you only one shot at correct specification.

Augmented inverse probability weighting

Efficient influence function

The efficient influence function (EIF) is the theoretical building block for constructing doubly robust estimators. It characterizes how much a single observation influences the estimator and determines the asymptotic variance.

For the ATE, the EIF is:

$$\frac{A(Y - \mu_1(X))}{e(X)} - \frac{(1-A)(Y - \mu_0(X))}{1-e(X)} + \mu_1(X) - \mu_0(X) - \psi$$

where $\mu_1(X)$ and $\mu_0(X)$ are the true outcome models and ψ is the true ATE. Notice the structure: the first two terms are IPW-style terms applied to the residuals (not the raw outcomes), and the last terms are the outcome regression estimate. When either model is correct, the "error" terms vanish in expectation.

AIPW estimator

The augmented inverse probability weighting (AIPW) estimator sets the sample average of the EIF to zero:

$$\hat{\psi}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^n \left(\frac{A_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \frac{(1-A_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)} + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right)$$

Here $\hat{\mu}_1(X)$, $\hat{\mu}_0(X)$ are estimated outcome models and $\hat{e}(X)$ is the estimated propensity score.

Think of it this way: start with the outcome regression estimate $\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$, then add a bias-correction term that uses IPW on the residuals. If the outcome model is perfect, the residuals are zero and the correction vanishes. If the outcome model is wrong but the propensity score is right, the correction fixes the bias.

Comparison to standard IPW

- AIPW is doubly robust; standard IPW is not. IPW fails entirely if the propensity score model is wrong.
- When the outcome model is well-specified, AIPW is more efficient than IPW because the augmentation term reduces variance.
- The augmentation term also stabilizes the estimator against extreme propensity scores, since it operates on residuals rather than raw outcomes. Large weights matter less when the residuals they multiply are small.

Targeted maximum likelihood estimation

TMLE algorithm

Targeted maximum likelihood estimation (TMLE) is a doubly robust procedure that iteratively updates initial model estimates to specifically target the parameter of interest (here, the ATE).

The steps are:

1. Fit initial models: Estimate the outcome model $\hat{\mu}(A, X)$ and the propensity score $\hat{e}(X)$.
2. Compute clever covariate: For each observation, calculate $H_i = \frac{A_i}{\hat{e}(X_i)} - \frac{1-A_i}{1-\hat{e}(X_i)}$. This covariate encodes the propensity score information.

3. Targeting step: Fit a regression of Y on H using the initial outcome predictions as an offset (on the appropriate link scale, e.g., logit for binary outcomes). This "tilts" the outcome model toward the ATE.

4. Update predictions: Use the fitted coefficient from step 3 to update the outcome model predictions.

5. Estimate the ATE: Compute the ATE as the average difference in updated predictions:
$$\hat{\psi}_{\text{TMLE}} = \frac{1}{n} \sum_i [\hat{\mu}_1^*(X_i) - \hat{\mu}_0^*(X_i)],$$
 where the starred quantities are the updated predictions.

Steps 2-4 can be iterated until convergence, though in practice one iteration is often sufficient.

Advantages over AIPW

- TMLE is a substitution estimator: the final ATE estimate is always a function of predicted outcomes, which means it respects the natural bounds of the outcome (e.g., probabilities stay between 0 and 1). AIPW can produce estimates outside the parameter space.
- TMLE is a more general framework, applicable to survival analysis, longitudinal data, and other complex settings.
- The targeting step can improve finite-sample stability, particularly when propensity scores are near the boundaries.

Limitations and considerations

- TMLE is more computationally intensive than AIPW due to the iterative updating.
- Performance depends on the choice of initial estimators and convergence criteria for the targeting step.
- Like all doubly robust methods, TMLE relies on the positivity assumption. Near-violations (propensity scores very close to 0 or 1) can still cause instability, though TMLE tends to handle this better than AIPW in practice.

Doubly robust estimation in practice

Model selection and validation

- Use subject matter knowledge to select covariates for both models. Include variables that are confounders (related to both treatment and outcome). Avoid conditioning on instruments or colliders.
- Data-driven approaches like regularization (LASSO, ridge) can help with variable selection, especially in high-dimensional settings.
- Cross-validation is important for assessing model fit. With doubly robust estimators, cross-fitting (sample splitting) is recommended when using flexible machine learning estimators to avoid overfitting bias.

Diagnostics and sensitivity analysis

- Check overlap: plot the propensity score distributions for treated and control groups. Poor overlap signals positivity violations and unreliable estimates.
- Examine the distribution of IPW weights. If a few observations carry enormous weight, consider trimming or truncating extreme propensity scores.
- Run sensitivity analyses: vary the functional form of the models, swap in different estimators (AIPW vs. TMLE), or use formal sensitivity analysis methods for unmeasured confounding (e.g., E-values).

Extensions and variations

- Doubly robust methods extend to multiple treatments, continuous treatments, and time-varying treatments (longitudinal settings).
- Machine learning can be incorporated into both the propensity score and outcome models to capture complex, nonlinear relationships. This is the basis of doubly robust machine learning (DML) and targeted learning frameworks.
- Adaptations exist for specific study designs such as matched cohorts, clustered data, and survey-weighted samples.

Comparison of doubly robust methods

AIPW vs. TMLE

Feature	AIPW	TMLE
Double robustness	Yes	Yes
Substitution estimator	No	Yes
Respects parameter bounds	Not guaranteed	Yes
Computational cost	Lower	Higher
Generality	ATE and similar	Broad (survival, longitudinal, etc.)

Both are asymptotically equivalent when both models are correctly specified, achieving the semiparametric efficiency bound.

Efficiency and robustness tradeoffs

- With both models correct, AIPW and TMLE perform identically in large samples.
- With one model misspecified, both remain consistent, but their finite-sample behavior can differ. TMLE's targeting step often provides better stability when propensity scores are extreme.
- Neither method can save you from misspecifying both models. The "doubly robust" label means two chances, not a guarantee.

Guidelines for method choice

- If your setting is straightforward (binary treatment, cross-sectional data, no extreme propensity scores), AIPW is simpler to implement and often sufficient.
- If you're working with bounded outcomes, complex data structures, or want to pair with machine learning (Super Learner), TMLE is the more flexible choice.
- When in doubt, run both. If AIPW and TMLE give similar answers, that's reassuring. If they diverge meaningfully, investigate why: it often points to model misspecification or positivity problems.

Unit 6 – Instrumental variables

6.1 Instrumental variables (IV) assumptions

Definition of Instrumental Variables

Instrumental variables (IVs) let you estimate causal effects when unmeasured confounding makes standard regression unreliable. The core idea: find a variable that nudges the treatment but has no other connection to the outcome. That variable becomes your instrument.

For an IV to work, four assumptions must hold: relevance, the exclusion restriction, exchangeability, and monotonicity. If any of these breaks down, your causal estimate can be biased or uninterpretable. The rest of this guide walks through each assumption, what happens when they fail, and how to think about choosing and interpreting IVs.

Relevance Assumption

Instrument Correlated with Treatment

The instrument must actually predict the treatment. If the instrument has no association with treatment assignment, it provides no useful variation to exploit, and the whole IV strategy falls apart.

The strength of this correlation directly affects the precision of your estimates. A weak instrument (only weakly correlated with treatment) inflates standard errors and can introduce finite-sample bias toward the confounded OLS estimate.

Strength of Instrument

You can assess instrument strength using the F-statistic from the first-stage regression (regressing the treatment on the instrument). The classic rule of thumb: an F-statistic above 10 suggests the instrument is strong enough, though more recent work by Stock and Yogo provides more nuanced thresholds.

The partial R^2 from that first-stage regression is another useful diagnostic. It tells you how much of the treatment variation the instrument explains after accounting for other covariates.

With weak instruments, you'll see:

- Large standard errors and wide confidence intervals
- IV estimates that are biased toward the OLS estimate in finite samples
- Unreliable inference (confidence intervals with poor coverage)

Exclusion Restriction

Instrument Affects Outcome Only Through Treatment

The exclusion restriction states that the instrument Z affects the outcome Y only through the treatment X . In DAG terms, there should be no direct arrow from Z to Y and no backdoor path from Z to Y that bypasses X .

This is the assumption that does the heaviest lifting in an IV analysis, and it is untestable with data alone. You have to argue for it on substantive grounds.

Violations of Exclusion Restriction

If the instrument directly affects the outcome, the IV estimate conflates the causal effect of treatment with the instrument's direct effect. Suppose Z is a physician's prescribing preference (used as an instrument for medication use). If that preference also correlates with other aspects of care the physician provides, the exclusion restriction is violated.

When the exclusion restriction fails, the IV estimate is biased, and the direction of bias depends on the sign and magnitude of the direct effect. Careful reasoning about the causal structure, not statistical tests, is the primary way to evaluate this assumption.

Exchangeability

Instrument Independent of Potential Outcomes

Exchangeability requires that the instrument Z is independent of the potential outcomes Y^a . Put differently, the instrument should be "as good as randomly assigned" with respect to the outcome. This rules out any common cause of Z and Y that doesn't go through X .

This ensures that different levels of the instrument correspond to groups that are comparable in their potential outcomes, so the variation the instrument creates is unconfounded.

Instrument Independent of Measured Confounders

One way to partially check this: examine whether measured covariates are balanced across levels of the instrument. If people with $Z = 1$ look systematically different from those with $Z = 0$ on observed characteristics, that's a red flag.

Balance checks on measured covariates are helpful but not definitive. Passing them is necessary but not sufficient for exchangeability.

Instrument Independent of Unmeasured Confounders

The harder part is that the instrument must also be independent of unmeasured confounders affecting both treatment and outcome. This is untestable by definition, since you can't check balance on variables you haven't measured.

The plausibility of this assumption rests entirely on your understanding of the data-generating process. Natural experiments (draft lotteries, geographic distance to a facility, Mendelian randomization) are popular IV strategies precisely because randomization or quasi-randomization makes this assumption more credible.

Monotonicity Assumption

Concept of Defiers

To understand monotonicity, you need the four compliance types defined by how individuals respond to the instrument:

- Compliers: take treatment when $Z = 1$, don't take it when $Z = 0$
- Always-takers: take treatment regardless of Z
- Never-takers: refuse treatment regardless of Z

- Defiers: do the opposite of what the instrument encourages (take treatment when $Z = 0$, refuse when $Z = 1$)

Always-takers and never-takers don't contribute to the IV estimate because their treatment status doesn't change with the instrument. Compliers are the group whose behavior the instrument actually shifts.

Absence of Defiers

Monotonicity assumes there are no defiers in the population. The instrument must push everyone in the same direction (or not at all). Nobody systematically does the opposite of what the instrument encourages.

Why does this matter? If defiers exist, their treatment effects get mixed in with the compliers' effects but with the wrong sign, making the IV estimate uninterpretable. With monotonicity satisfied, the IV estimate has a clean interpretation as the local average treatment effect (LATE) for compliers.

In practice, monotonicity is most plausible when the instrument operates through a simple, one-directional mechanism. For example, being randomly assigned an encouragement to exercise is unlikely to cause anyone to exercise less.

Consequences of Violated Assumptions

Bias in IV Estimates

Each assumption violation produces different problems:

- Weak instruments (relevance violation): bias toward the confounded OLS estimate, unreliable inference
- Exclusion restriction violation: bias that reflects the instrument's direct effect on the outcome
- Exchangeability violation: bias from confounding between the instrument and the outcome, similar in spirit to omitted variable bias in OLS
- Monotonicity violation: the LATE interpretation breaks down, and the estimate may not correspond to any well-defined causal quantity

The severity of bias depends on how badly the assumption is violated. Small violations may produce small bias, but you generally can't know the magnitude without strong assumptions.

Sensitivity Analyses for Violations

Since most IV assumptions are untestable, sensitivity analyses help you assess how fragile your conclusions are:

- For the exclusion restriction, you can ask: how large would a direct effect of Z on Y need to be to nullify or reverse the IV estimate?
- For exchangeability, you can simulate unmeasured confounders of varying strength and see how the estimate shifts.
- For instrument strength, you can report weak-instrument-robust confidence intervals (e.g., Anderson-Rubin confidence sets) that remain valid even with weak instruments.

These analyses don't prove your assumptions hold, but they show whether your conclusions depend on the assumptions holding exactly or whether they're robust to plausible violations.

Considerations for Choosing Instruments

Subject Matter Knowledge

Good instruments come from understanding the causal structure of your problem. You need a convincing story for why the proposed instrument satisfies all four assumptions. Common sources of instruments include:

- Policy changes or natural experiments (e.g., a lottery that determines eligibility)
- Geographic or institutional variation (e.g., distance to a treatment facility)
- Genetic variants (Mendelian randomization)

The best instruments have a clear, well-understood mechanism linking them to treatment and a compelling argument for why they don't affect the outcome through other channels.

Data-Driven Approaches

Statistical methods can help screen for candidate instruments by identifying variables that strongly predict treatment. Machine learning approaches can be useful here, particularly for constructing instruments from many weak predictors.

However, data-driven methods can only assess relevance. They cannot tell you whether the exclusion restriction or exchangeability holds. A variable that predicts treatment perfectly but violates the exclusion restriction is worse than useless. Always pair data-driven screening with substantive reasoning.

Interpretation of IV Estimates

Local Average Treatment Effect (LATE)

Under the four assumptions, the IV estimand is the LATE: the average treatment effect among compliers. This is not the average treatment effect for the whole population.

Compliers are the people whose treatment status actually changes in response to the instrument. If the instrument is a draft lottery for military service, compliers are those who serve when drafted but wouldn't volunteer otherwise. The LATE tells you the causal effect of service for that specific group.

Generalizability of IV Estimates

The LATE applies to compliers, and compliers may differ from always-takers, never-takers, and the general population. This limits external validity.

Key questions to ask:

- Who are the compliers? Can you characterize them using observed covariates?
- Are compliers likely to have larger or smaller treatment effects than the general population?
- Does the specific instrument define a complier population that's relevant to the policy question you care about?

Different instruments for the same treatment can identify different complier populations and therefore produce different LATEs. This isn't a contradiction; it reflects genuine heterogeneity in treatment effects across subgroups.

Extensions of IV Methods

Multiple Instruments

When multiple instruments are available for the same treatment, you can:

- Combine them to improve efficiency (tighter estimates)
- Run overidentification tests (e.g., Sargan/Hansen test) to check whether the instruments produce consistent estimates. If they don't, at least one instrument likely violates the exclusion restriction.

Using multiple instruments requires that each instrument individually satisfies the IV assumptions. It also raises questions about whether the LATEs identified by different instruments are the same, which depends on treatment effect homogeneity.

Nonlinear Models with IV

Standard IV (two-stage least squares, or 2SLS) assumes a linear relationship between treatment and outcome. Extensions exist for nonlinear settings:

- Binary outcomes can be handled with IV probit or bivariate probit models
- Nonlinear first stages can be accommodated, though care is needed to avoid the "forbidden regression" (plugging fitted values from a nonlinear first stage into a linear second stage)

These extensions require additional modeling assumptions (correct functional form, distributional assumptions) on top of the standard IV assumptions, so they demand extra scrutiny.

6.2 Two-stage least squares (2SLS)

Two-stage least squares (2SLS) is a powerful tool for estimating causal effects when dealing with endogeneity. It uses instrumental variables to isolate exogenous variation in the explanatory variable, allowing researchers to overcome biases from omitted variables or reverse causality.

The 2SLS method involves two stages: first, regressing the endogenous variable on instruments, then using predicted values to estimate the causal effect. This approach provides consistent estimates when OLS fails, but requires finding valid instruments that satisfy relevance and exclusion conditions.

Overview of 2SLS

- Two-stage least squares (2SLS) is an instrumental variable estimation method used to estimate causal effects in the presence of endogeneity
- 2SLS consists of two stages: the first stage regresses the endogenous variable on the instrumental variables, and the second stage uses the predicted values from the first stage to estimate the causal effect
- 2SLS is widely used in causal inference when randomized experiments are not feasible or ethical, and it relies on finding suitable instrumental variables that satisfy certain conditions

Motivation for 2SLS

Endogeneity problem

- Endogeneity occurs when an explanatory variable is correlated with the error term in a regression model, violating the assumption of exogeneity

- Endogeneity can arise due to omitted variables, measurement error, or simultaneous causality, leading to biased and inconsistent estimates of causal effects
- Examples of endogeneity include the relationship between education and earnings (ability bias) or the impact of police on crime rates (reverse causality)

Limitations of OLS

- Ordinary least squares (OLS) estimation assumes that the explanatory variables are exogenous and uncorrelated with the error term
- In the presence of endogeneity, OLS estimates are biased and inconsistent, leading to incorrect conclusions about causal relationships
- OLS cannot disentangle the causal effect of an endogenous variable from the confounding factors that affect both the explanatory variable and the outcome

Instrumental variables (IVs)

Definition of IVs

- An instrumental variable (IV) is a variable that is correlated with the endogenous explanatory variable but uncorrelated with the error term in the regression model
- IVs provide exogenous variation in the endogenous variable, allowing for the estimation of causal effects
- Examples of IVs include distance to college as an instrument for education, or rainfall as an instrument for agricultural output

Relevance condition

- The relevance condition requires that the IV is sufficiently correlated with the endogenous explanatory variable
- A strong correlation between the IV and the endogenous variable is necessary for the IV to be a valid instrument and to avoid the weak instruments problem
- The first-stage F-statistic is commonly used to assess the strength of the IV, with a rule of thumb suggesting an F-statistic greater than 10 for a strong instrument

Exclusion restriction

- The exclusion restriction assumes that the IV affects the outcome variable only through its effect on the endogenous explanatory variable
- This means that the IV should have no direct effect on the outcome, and any effect should be mediated entirely through the endogenous variable
- Violations of the exclusion restriction can lead to biased estimates and invalidate the IV approach

First stage of 2SLS

Regressing endogenous variable on IVs

- In the first stage of 2SLS, the endogenous explanatory variable is regressed on the IV(s) and any other exogenous control variables
- This regression estimates the relationship between the IV(s) and the endogenous variable, isolating the exogenous variation in the endogenous variable
- The predicted values from the first-stage regression are used in the second stage to estimate the causal effect

Assessing IV strength

- The strength of the IV is assessed using the first-stage F-statistic, which tests the joint significance of the IV(s) in predicting the endogenous variable
- A high F-statistic (greater than 10) indicates a strong IV, while a low F-statistic suggests a weak instrument problem
- Weak instruments can lead to biased and imprecise estimates in the second stage, and they may perform poorly in small samples

Weak instruments problem

- The weak instruments problem arises when the IV is only weakly correlated with the endogenous variable, leading to biased and imprecise estimates in the second stage
- Weak instruments can cause the 2SLS estimator to be biased towards the OLS estimator, and the bias may not disappear even in large samples
- Solutions to the weak instruments problem include using multiple IVs, estimating the reduced form directly, or employing alternative estimation methods (limited information maximum likelihood)

Second stage of 2SLS

Using first-stage predictions

- In the second stage of 2SLS, the predicted values of the endogenous variable from the first stage are used as a regressor in the main regression model
- By using the predicted values, which are based on the exogenous variation in the IV, the second stage isolates the causal effect of the endogenous variable on the outcome
- The second-stage regression includes the predicted endogenous variable and any other exogenous control variables

Estimating causal effect

- The coefficient on the predicted endogenous variable in the second stage represents the causal effect of the endogenous variable on the outcome
- This causal effect is interpreted as the local average treatment effect (LATE) for the subpopulation affected by the IV (compliers)

- Standard errors in the second stage need to be adjusted to account for the two-stage estimation process, typically using robust or clustered standard errors

Properties of 2SLS estimator

Consistency

- The 2SLS estimator is consistent, meaning that it converges in probability to the true causal effect as the sample size increases, under the assumptions of IV validity
- Consistency requires that the IV satisfies the relevance condition and the exclusion restriction, and that the sample size is large enough for the asymptotic properties to hold
- Consistency ensures that the 2SLS estimator provides an unbiased estimate of the causal effect in large samples

Asymptotic normality

- The 2SLS estimator is asymptotically normally distributed, which allows for the construction of confidence intervals and hypothesis testing
- Asymptotic normality holds under the assumptions of IV validity and the usual regularity conditions for the central limit theorem
- The asymptotic variance of the 2SLS estimator is larger than that of the OLS estimator, reflecting the uncertainty introduced by the two-stage estimation process

Efficiency vs OLS

- The 2SLS estimator is less efficient than the OLS estimator when the explanatory variables are exogenous, meaning that it has a larger variance
- The efficiency loss of 2SLS compared to OLS depends on the strength of the IV and the degree of endogeneity in the model
- However, when endogeneity is present, the OLS estimator is biased and inconsistent, making the 2SLS estimator a preferred choice despite its lower efficiency

Assumptions for 2SLS validity

Linear model specification

- 2SLS assumes that the relationship between the endogenous variable, the IV, and the outcome is linear in parameters
- Misspecification of the functional form can lead to biased estimates and invalidate the 2SLS approach
- Non-linear relationships can be accommodated by transforming variables or using non-linear 2SLS estimators

Homoskedasticity

- 2SLS assumes that the error term in the second-stage regression has constant variance (homoskedasticity)

- Violations of homoskedasticity (heteroskedasticity) can lead to inefficient estimates and incorrect standard errors
- Heteroskedasticity-robust standard errors can be used to account for non-constant variance in the error term

No multicollinearity

- 2SLS assumes that there is no perfect multicollinearity among the explanatory variables, including the IV and the endogenous variable
- Perfect multicollinearity can lead to non-identification of the model and inability to estimate the causal effect
- Near multicollinearity (high correlation among explanatory variables) can lead to imprecise estimates and large standard errors

Testing 2SLS assumptions

Overidentification tests

- Overidentification tests are used when there are more IVs than endogenous variables (overidentified model) to test the validity of the exclusion restriction
- The Sargan-Hansen test is commonly used, which tests the joint null hypothesis that all IVs are valid (uncorrelated with the error term)
- Rejecting the null hypothesis suggests that at least one of the IVs violates the exclusion restriction, and the model may need to be reconsidered

Hausman test for endogeneity

- The Hausman test is used to compare the OLS and 2SLS estimates and test for the presence of endogeneity in the model
- The test is based on the difference between the OLS and 2SLS estimates, which should be small if the explanatory variables are exogenous
- Rejecting the null hypothesis of the Hausman test indicates the presence of endogeneity, and the 2SLS estimator is preferred over OLS

Interpreting 2SLS results

Local average treatment effect (LATE)

- The 2SLS estimator identifies the local average treatment effect (LATE), which is the causal effect for the subpopulation affected by the IV (compliers)
- Compliers are individuals whose treatment status (endogenous variable) is influenced by the IV, while always-takers and never-takers are not affected by the IV
- The LATE may differ from the average treatment effect (ATE) for the entire population, depending on the heterogeneity of treatment effects

Generalizability of findings

- The generalizability of 2SLS findings depends on the similarity between the compliers and the target population of interest
- If the compliers are a non-representative subgroup, the LATE may not be informative about the causal effect for the entire population
- Assessing the external validity of 2SLS results requires careful consideration of the characteristics of the compliers and the context of the study

Limitations of 2SLS

Finding suitable instruments

- The main challenge in implementing 2SLS is finding suitable instruments that satisfy the relevance condition and the exclusion restriction
- Instruments that are weakly correlated with the endogenous variable or violate the exclusion restriction can lead to biased and inconsistent estimates
- Identifying credible IVs often requires deep institutional knowledge and theoretical justification, and the validity of IVs cannot be tested directly

Sensitivity to assumptions

- 2SLS estimates are sensitive to violations of the assumptions, such as non-linearity, heteroskedasticity, or invalid instruments
- Small violations of the exclusion restriction can lead to substantial bias in the 2SLS estimator, especially when the instruments are weak
- Sensitivity analysis and robustness checks are important to assess the stability of 2SLS results under different assumptions and specifications

Extensions of 2SLS

Multiple endogenous variables

- 2SLS can be extended to handle multiple endogenous variables by using multiple instruments and estimating the first stage for each endogenous variable separately
- The second stage includes the predicted values of all endogenous variables and estimates their causal effects simultaneously
- Identification in the presence of multiple endogenous variables requires at least as many instruments as endogenous variables (order condition) and sufficient variation in the instruments (rank condition)

Nonlinear models

- 2SLS can be adapted to estimate causal effects in nonlinear models, such as probit, logit, or Poisson regression
- The two-stage residual inclusion (2SRI) method is commonly used, where the residuals from the first stage are included as an additional regressor in the second stage

- Nonlinear 2SLS estimators have different properties than linear 2SLS and require careful interpretation of the marginal effects

Panel data settings

- 2SLS can be applied to panel data settings, where multiple observations are available for each individual or unit over time
- Panel data 2SLS estimators can control for unobserved time-invariant heterogeneity by including individual fixed effects or using first-differencing
- Instruments in panel data settings can be time-varying or time-invariant, and the relevance and exclusion conditions need to be satisfied in the presence of fixed effects or first-differencing

6.3 Local average treatment effect (LATE)

Local Average Treatment Effect (LATE) is a crucial concept in causal inference. It estimates the causal effect of a treatment for a specific subpopulation induced to take up the treatment by an instrumental variable. LATE is particularly useful when there's non-compliance or self-selection into treatment.

LATE differs from the Average Treatment Effect (ATE) by focusing on compliers rather than the entire population. It relies on key assumptions like relevance, exclusion restriction, and monotonicity. Understanding LATE is essential for researchers estimating causal effects in complex real-world scenarios.

Definition of LATE

- Local Average Treatment Effect (LATE) estimates the causal effect of a treatment for a specific subpopulation of individuals who are induced to take up the treatment by a change in an instrumental variable
- LATE is a key concept in causal inference that allows researchers to estimate causal effects in the presence of non-compliance or self-selection into treatment
- LATE is particularly relevant when there is heterogeneity in treatment effects and the subpopulation of interest is those who would change their treatment status based on the instrumental variable

LATE vs ATE

- LATE differs from the Average Treatment Effect (ATE) because it focuses on a specific subpopulation (compliers) rather than the entire population
- ATE estimates the average causal effect of a treatment across all individuals in a population, regardless of their compliance with treatment assignment
- LATE is more relevant when there is heterogeneity in treatment effects and the subpopulation of interest is those who would change their treatment status based on the instrumental variable

Assumptions for LATE

- LATE relies on several key assumptions, including relevance, exclusion restriction, and monotonicity
- Relevance assumes that the instrumental variable is correlated with the treatment variable
- Exclusion restriction assumes that the instrumental variable only affects the outcome through its effect on the treatment variable

- Monotonicity assumes that there are no defiers in the population (individuals who always do the opposite of their treatment assignment)

Interpretation of LATE

- LATE estimates the average causal effect of a treatment for the subpopulation of compliers
- Compliers are individuals who would take up the treatment if assigned to the treatment group and would not take up the treatment if assigned to the control group
- LATE can be interpreted as the causal effect of the treatment for the subpopulation of individuals who are induced to take up the treatment by a change in the instrumental variable

Instrumental variables

- Instrumental variables (IVs) are a key tool in causal inference for estimating causal effects in the presence of unmeasured confounding or self-selection into treatment
- An instrumental variable is a variable that affects the treatment variable but does not directly affect the outcome variable, except through its effect on the treatment
- IVs allow researchers to isolate the causal effect of a treatment by exploiting exogenous variation in treatment assignment

Relevance condition

- The relevance condition requires that the instrumental variable is correlated with the treatment variable
- This means that the instrumental variable must have a non-zero effect on the probability of receiving the treatment
- The strength of the correlation between the instrumental variable and the treatment variable determines the strength of the instrument

Exclusion restriction

- The exclusion restriction assumes that the instrumental variable only affects the outcome through its effect on the treatment variable
- This means that there should be no direct effect of the instrumental variable on the outcome, and no unmeasured confounding between the instrumental variable and the outcome
- Violations of the exclusion restriction can lead to biased estimates of the causal effect

Monotonicity assumption

- The monotonicity assumption requires that there are no defiers in the population
- Defiers are individuals who always do the opposite of their treatment assignment (take up treatment when assigned to control, and do not take up treatment when assigned to treatment)
- Monotonicity is necessary for the interpretation of LATE as a causal effect for compliers

Strength of instruments

- The strength of an instrumental variable is determined by the magnitude of its correlation with the treatment variable
- Weak instruments (those with low correlation) can lead to biased estimates and large standard errors
- Strong instruments (those with high correlation) provide more precise estimates of the causal effect
- The F-statistic from the first stage regression of the treatment on the instrument is often used to assess instrument strength ($F > 10$ is a common rule of thumb)

Compliers vs defiers

- Compliers and defiers are two key subpopulations in the context of LATE estimation
- The behavior of these subpopulations has important implications for the interpretation and validity of LATE estimates

Definition of compliers

- Compliers are individuals who would take up the treatment if assigned to the treatment group and would not take up the treatment if assigned to the control group
- In other words, compliers are those whose treatment status is determined by their assignment to treatment or control
- LATE estimates the causal effect of the treatment for the subpopulation of compliers

Definition of defiers

- Defiers are individuals who always do the opposite of their treatment assignment
- They take up the treatment when assigned to the control group and do not take up the treatment when assigned to the treatment group
- The presence of defiers violates the monotonicity assumption required for LATE estimation

Implications for LATE

- The interpretation of LATE as a causal effect for compliers relies on the absence of defiers in the population
- If defiers are present, LATE estimates a weighted average of the causal effects for compliers and defiers, which is difficult to interpret
- The presence of defiers can also lead to biased estimates of the causal effect
- Researchers often argue for the plausibility of the monotonicity assumption in their specific context to justify the use of LATE

Estimation of LATE

- Several methods can be used to estimate LATE, including the Wald estimator and two-stage least squares (2SLS)

- These methods rely on the instrumental variable to isolate the causal effect of the treatment for compliers

Wald estimator

- The Wald estimator is a simple method for estimating LATE
- It is calculated as the ratio of the difference in outcomes between the treatment and control groups (reduced form) to the difference in treatment uptake between the treatment and control groups (first stage)
- The Wald estimator provides a consistent estimate of LATE under the assumptions of relevance, exclusion restriction, and monotonicity

Two-stage least squares

- Two-stage least squares (2SLS) is a more general method for estimating LATE that can accommodate multiple instruments and covariates
- In the first stage, the treatment variable is regressed on the instrumental variable(s) and any covariates
- In the second stage, the outcome variable is regressed on the predicted values of the treatment from the first stage and any covariates
- The coefficient on the predicted treatment in the second stage provides an estimate of LATE

Confidence intervals for LATE

- Confidence intervals for LATE can be constructed using standard errors from the 2SLS regression
- These confidence intervals reflect the uncertainty in the LATE estimate due to sampling variability
- It is important to note that confidence intervals for LATE only capture sampling uncertainty and do not account for uncertainty in the validity of the IV assumptions
- Researchers should assess the plausibility of the IV assumptions and consider sensitivity analyses to evaluate the robustness of their results

Applications of LATE

- LATE has been widely applied in various fields, including economics, public health, and social sciences
- Researchers use LATE to estimate causal effects in settings where there is non-compliance or self-selection into treatment

Examples in economics

- Angrist and Krueger (1991) used quarter of birth as an instrumental variable to estimate the causal effect of education on earnings
- Card (1995) used geographic proximity to college as an instrumental variable to estimate the causal effect of education on earnings
- Duflo and Saez (2003) used randomized information sessions as an instrumental variable to estimate the causal effect of retirement savings plan participation on savings

Examples in public health

- Greenland (2000) used physician prescribing preferences as an instrumental variable to estimate the causal effect of a drug treatment on patient outcomes
- Brookhart et al. (2006) used physician prescribing preferences as an instrumental variable to estimate the causal effect of COX-2 inhibitors on gastrointestinal complications
- Hernán and Robins (2006) used randomization as an instrumental variable to estimate the causal effect of hormone replacement therapy on cardiovascular outcomes

Limitations of LATE

- LATE estimates the causal effect for a specific subpopulation (compliers) and may not generalize to the entire population
- The validity of LATE estimates depends on the plausibility of the IV assumptions (relevance, exclusion restriction, monotonicity), which may be difficult to justify in some contexts
- LATE estimates can be sensitive to violations of the IV assumptions, and researchers should conduct sensitivity analyses to assess the robustness of their results
- LATE estimates may have limited policy relevance if the subpopulation of compliers is not representative of the target population for an intervention

Comparison to other estimands

- LATE is one of several estimands that can be used to estimate causal effects in the presence of non-compliance or self-selection into treatment
- Other estimands include the Average Treatment Effect (ATE), Average Treatment Effect on the Treated (ATT), and Average Treatment Effect on the Untreated (ATU)

LATE vs ATT

- ATT estimates the average causal effect of a treatment for those who actually received the treatment
- LATE estimates the average causal effect of a treatment for those who would receive the treatment if assigned to the treatment group (compliers)
- ATT and LATE may differ if there is selection bias in who receives the treatment

LATE vs ATC

- ATC (Average Treatment effect on the Compliers) is another name for LATE
- Both ATC and LATE refer to the average causal effect of a treatment for the subpopulation of compliers
- The terms are often used interchangeably in the literature

When to use LATE

- LATE is appropriate when there is non-compliance or self-selection into treatment and the researcher has a valid instrumental variable

- LATE is particularly useful when there is heterogeneity in treatment effects and the subpopulation of interest is those who would change their treatment status based on the instrumental variable
- Researchers should carefully consider the plausibility of the IV assumptions and the policy relevance of the complier subpopulation when deciding whether to use LATE
- If the IV assumptions are not plausible or the complier subpopulation is not of interest, other estimands (such as ATE or ATT) may be more appropriate

6.4 Weak instruments and identification

Weak instruments can lead to biased and inconsistent estimates in instrumental variable analysis. This topic explores the challenges of weak instruments, methods for detecting them, and their consequences for causal inference.

Strategies for addressing weak instruments include careful instrument selection and combining multiple instruments. The notes also cover partial identification, bounds on parameters, and alternative methods for estimating causal effects when weak instruments are a concern.

Weak instruments overview

- Weak instruments are a common challenge in instrumental variable (IV) estimation, arising when the correlation between the instrument and the endogenous regressor is low
- Weak instruments can lead to biased and inconsistent estimates, as well as unreliable hypothesis tests and confidence intervals
- Identifying and addressing weak instruments is crucial for obtaining valid causal estimates in IV analysis

Bias in weak instruments

- Weak instruments can introduce bias in the IV estimator, even in large samples, due to the correlation between the instrument and the error term
- The bias of the IV estimator with weak instruments is approximately proportional to the degree of endogeneity and inversely proportional to the strength of the instrument
- In finite samples, the bias of the IV estimator with weak instruments can be substantial and may even exceed the bias of ordinary least squares (OLS) estimation

Inconsistency of estimators

- Weak instruments can lead to inconsistent IV estimators, meaning that the estimator does not converge to the true parameter value as the sample size increases
- The inconsistency of IV estimators with weak instruments arises because the correlation between the instrument and the endogenous regressor does not increase sufficiently with the sample size
- Inconsistent estimators can produce misleading results and invalid inferences, even in large samples

Detecting weak instruments

- Several diagnostic tools and tests have been developed to assess the strength of instruments and detect potential weak instrument problems

- These methods help researchers determine whether their instruments are sufficiently strong to provide reliable IV estimates
- Commonly used tests for weak instruments include the first-stage F-statistic, the Cragg-Donald statistic, and the Stock-Yogo critical values

First-stage F-statistic

- The first-stage F-statistic is a widely used diagnostic for assessing instrument strength in the context of a single endogenous regressor
- It is obtained by regressing the endogenous regressor on the instrument(s) and control variables, and testing the joint significance of the excluded instruments
- A rule of thumb suggests that an F-statistic greater than 10 indicates sufficiently strong instruments, while values below 10 suggest potential weak instrument problems

Cragg-Donald statistic

- The Cragg-Donald statistic is a generalization of the first-stage F-statistic for models with multiple endogenous regressors
- It tests the null hypothesis that the instruments are weak, based on the eigenvalues of a matrix involving the reduced-form coefficients and the covariance matrix of the instruments
- Higher values of the Cragg-Donald statistic indicate stronger instruments, while lower values suggest weak instruments

Stock-Yogo critical values

- Stock and Yogo (2005) developed critical values for the Cragg-Donald statistic that allow researchers to test for weak instruments based on the maximum acceptable bias or size distortion of the IV estimator
- These critical values provide thresholds for determining whether the instruments are sufficiently strong to limit the bias or size distortion to a desired level (e.g., 10% or 15%)
- Comparing the Cragg-Donald statistic to the appropriate Stock-Yogo critical value helps assess the strength of the instruments in the presence of multiple endogenous regressors

Consequences of weak instruments

- Weak instruments can have several adverse consequences for inference and interpretation in IV analysis
- These consequences include distorted hypothesis tests, invalid confidence intervals, and potentially misleading conclusions about causal relationships
- Understanding the implications of weak instruments is essential for interpreting IV results and drawing appropriate inferences

Hypothesis tests

- Weak instruments can distort the size and power of hypothesis tests in IV estimation
- When instruments are weak, the standard errors of the IV estimator may be severely underestimated, leading to over-rejection of the null hypothesis (Type I error)

- Hypothesis tests based on weak instruments may have poor power, meaning they may fail to reject the null hypothesis when it is false (Type II error)

Confidence intervals

- Weak instruments can lead to invalid and misleading confidence intervals for the IV estimator
- Confidence intervals constructed using standard asymptotic approximations may have incorrect coverage probabilities when instruments are weak
- In the presence of weak instruments, confidence intervals may be too narrow, failing to capture the true parameter value with the desired probability

Strategies for weak instruments

- Several strategies can be employed to mitigate the impact of weak instruments and improve the reliability of IV estimates
- These strategies involve careful selection of relevant instruments, combining multiple instruments, and using alternative estimation methods that are more robust to weak instruments
- Implementing appropriate strategies can help researchers obtain more reliable causal estimates in the presence of weak instruments

Selecting relevant instruments

- Choosing instruments that are strongly correlated with the endogenous regressor is crucial for mitigating weak instrument problems
- Researchers should carefully consider the theoretical and empirical justification for their chosen instruments, ensuring that they satisfy the relevance condition
- Using subject matter knowledge and prior research to identify relevant instruments can help strengthen the instrument-regressor relationship and improve the reliability of IV estimates

Combining multiple instruments

- When multiple instruments are available, combining them can help increase the overall strength of the instrument set
- Methods such as two-stage least squares (2SLS) and limited information maximum likelihood (LIML) can efficiently combine multiple instruments to estimate the causal effect of interest
- Combining instruments can provide more precise estimates and reduce the impact of weak instruments, especially when the instruments are jointly strong but individually weak

Weak instruments vs irrelevant instruments

- Weak instruments and irrelevant instruments are two distinct problems in IV estimation, with different implications for inference
- Weak instruments are correlated with the endogenous regressor but only weakly, while irrelevant instruments are uncorrelated with the endogenous regressor
- Irrelevant instruments violate the exclusion restriction and lead to inconsistent IV estimators, while weak instruments can lead to biased and inconsistent estimators in finite samples

Identification with weak instruments

- Weak instruments can pose challenges for identification in IV models, as the causal effect of interest may not be point-identified
- In the presence of weak instruments, the parameters of interest may only be partially identified, meaning that they can be bounded but not precisely estimated
- Partial identification and bounds on parameters can provide valuable information about the range of possible causal effects, even when point identification is not possible

Partial identification

- Partial identification refers to the situation where the causal effect of interest is not point-identified but can be bounded within a range of possible values
- With weak instruments, the IV estimator may not converge to a single point, but it can still provide information about the set of parameter values consistent with the data and assumptions
- Partial identification methods, such as set estimation and moment inequality approaches, can be used to construct bounds on the causal effect in the presence of weak instruments

Bounds on parameters

- When instruments are weak, researchers can derive bounds on the parameters of interest, such as the average treatment effect or the local average treatment effect
- These bounds represent the range of possible parameter values that are consistent with the observed data and the assumptions of the IV model
- Bounds on parameters can be informative for policy analysis and decision-making, even when point estimates are not available or reliable

Alternatives to instrumental variables

- When weak instruments are a concern, researchers may consider alternative methods for estimating causal effects that do not rely on instrumental variables
- These alternative methods exploit different sources of variation or make different assumptions to identify the causal effect of interest
- Some common alternatives to IV estimation include direct control for confounding, difference-in-differences, and regression discontinuity designs

Direct control for confounding

- Direct control for confounding involves including observed confounders as control variables in a regression model
- This approach assumes that all relevant confounders are observed and can be adequately controlled for in the analysis
- While direct control for confounding does not require instruments, it relies on the strong assumption of no unobserved confounding, which may not always be plausible

Difference-in-differences

- Difference-in-differences (DID) is a method that exploits variation in treatment timing across groups to estimate causal effects
- DID compares the change in outcomes before and after treatment for a treated group to the change in outcomes for an untreated control group
- DID relies on the assumption of parallel trends, which requires that the treated and control groups would have followed the same trend in outcomes in the absence of treatment

Regression discontinuity

- Regression discontinuity (RD) designs exploit a discontinuity in treatment assignment based on a continuous running variable
- In an RD design, units just above and below a threshold value of the running variable are compared, assuming that they are similar in all respects except for the treatment assignment
- RD designs can provide credible estimates of causal effects in the vicinity of the threshold, but they require a sufficient number of observations near the threshold and the appropriate specification of the functional form

Unit 7 – Regression discontinuity designs

7.1 Sharp and fuzzy regression discontinuity

Regression discontinuity designs are powerful tools for estimating causal effects in observational studies. They exploit discontinuities in treatment assignment based on a cutoff value of an assignment variable, allowing researchers to compare outcomes just above and below the threshold.

Sharp designs have deterministic treatment assignment at the cutoff, while fuzzy designs have probabilistic assignment. Key assumptions include continuity of potential outcomes and local randomization around the cutoff. Proper implementation involves careful bandwidth selection, validity checks, and consideration of limitations like external validity concerns.

Sharp vs fuzzy designs

- Sharp regression discontinuity (RD) designs have a deterministic treatment assignment based on a cutoff value of the assignment variable, while fuzzy RD designs have a probabilistic treatment assignment at the cutoff
- In sharp RD, the probability of treatment jumps from 0 to 1 at the cutoff, whereas in fuzzy RD, the probability of treatment increases discontinuously but not necessarily from 0 to 1
- Fuzzy RD designs are more common in practice due to imperfect compliance or measurement error in the assignment variable

Assignment variable types

- The assignment variable (running variable) in RD designs can be continuous (age, test score) or discrete (birth year, grade level)
- Continuous assignment variables allow for more precise estimation of the treatment effect at the cutoff
- Discrete assignment variables may lead to larger bandwidths and less precise estimates

Cutoff for treatment assignment

- The cutoff value of the assignment variable determines treatment assignment in RD designs
- The cutoff can be based on policy rules (eligibility thresholds for social programs) or arbitrary values (median test score)
- The choice of cutoff should be exogenous to potential outcomes to ensure the validity of the RD design

Identification assumptions

- RD designs rely on several key assumptions for identifying causal effects at the cutoff
- These assumptions are crucial for the internal validity of the RD estimates and should be carefully assessed in empirical applications

Continuity of potential outcomes

- The potential outcomes (treated and untreated) are assumed to be continuous functions of the assignment variable at the cutoff

- This assumption rules out any discontinuities in potential outcomes at the cutoff that are unrelated to the treatment
- Continuity can be assessed by testing for covariate balance and placebo discontinuities at non-cutoff points

Local randomization

- RD designs assume that treatment assignment is as good as randomly assigned in a small neighborhood around the cutoff
- This assumption holds if individuals have imprecise control over the assignment variable near the cutoff
- Local randomization can be tested by examining the density of the assignment variable and checking for manipulation around the cutoff

Exclusion restriction

- The assignment variable is assumed to affect outcomes only through its effect on treatment assignment at the cutoff
- This assumption rules out any direct effects of the assignment variable on outcomes that are not mediated by the treatment
- The exclusion restriction can be assessed by testing for discontinuities in outcomes at placebo cutoffs and examining the robustness of estimates to different bandwidth choices

Estimating causal effects

- RD designs allow for the estimation of causal effects of the treatment at the cutoff under the identification assumptions
- The causal effect is defined as the difference in potential outcomes between treated and untreated units at the cutoff

Local average treatment effect (LATE)

- The LATE is the causal effect of the treatment for compliers at the cutoff in fuzzy RD designs
- Compliers are units that receive the treatment if and only if they are above the cutoff
- The LATE can be estimated using instrumental variables (IV) methods, with the treatment indicator as the endogenous variable and the above-cutoff indicator as the instrument

Bandwidth selection

- The choice of bandwidth around the cutoff is crucial for the bias-variance trade-off in RD estimates
- Smaller bandwidths reduce bias by focusing on units closer to the cutoff but increase variance due to smaller sample sizes
- Optimal bandwidth selection methods (Imbens-Kalyanaraman, Calonico-Cattaneo-Titiunik) balance this trade-off and provide data-driven bandwidth choices

Parametric vs non-parametric approaches

- Parametric RD approaches use polynomial functions (linear, quadratic) to model the relationship between the assignment variable and outcomes on either side of the cutoff
- Non-parametric approaches use local linear regression or kernel methods to estimate the treatment effect at the cutoff
- Non-parametric methods are more flexible and rely on fewer functional form assumptions but may be more sensitive to bandwidth choice

Regression discontinuity plots

- RD plots are graphical representations of the relationship between the assignment variable and outcomes, with a focus on the discontinuity at the cutoff
- These plots provide a visual assessment of the main RD assumptions and the magnitude of the treatment effect

Binned scatter plots

- Binned scatter plots divide the assignment variable into discrete bins and plot the mean outcome within each bin against the bin midpoint
- These plots can reveal discontinuities in outcomes at the cutoff and the overall shape of the relationship between the assignment variable and outcomes
- The choice of bin width involves a bias-variance trade-off, with narrower bins providing more local estimates but higher variability

Polynomial fits

- Polynomial fits overlay parametric regression lines (linear, quadratic) on the RD plot to model the relationship between the assignment variable and outcomes on either side of the cutoff
- The order of the polynomial should be chosen to balance flexibility and overfitting, with higher-order polynomials providing a better fit but potentially leading to biased estimates

Confidence intervals

- Confidence intervals around the binned means or polynomial fits provide a measure of the uncertainty in the RD estimates
- These intervals can be used to assess the statistical significance of the treatment effect at the cutoff
- Robust confidence intervals (Calonico-Cattaneo-Titiunik) account for the bias-correction and clustering in the RD estimator

Validity checks

- RD designs rely on several key assumptions for identifying causal effects, and these assumptions should be carefully assessed in empirical applications
- Validity checks help to evaluate the credibility of the RD estimates and rule out alternative explanations for the observed discontinuities

Density test of running variable

- The density test examines the distribution of the assignment variable around the cutoff to check for manipulation or sorting
- A discontinuity in the density at the cutoff may indicate that individuals have precise control over the assignment variable, violating the local randomization assumption
- The McCrary density test and Cattaneo-Jansson-Ma test are commonly used to detect discontinuities in the density

Covariate balance tests

- Covariate balance tests assess whether predetermined characteristics of individuals are continuous at the cutoff
- Discontinuities in covariates at the cutoff may suggest that the RD assumptions are violated or that the treatment effect is confounded by other factors
- Balance tests can be conducted using parametric or non-parametric RD methods, with the covariates as the outcome variables

Placebo tests at non-discontinuity points

- Placebo tests involve testing for discontinuities in outcomes at cutoff values where there is no true treatment effect
- These tests help to assess whether the observed discontinuities at the actual cutoff are due to the treatment or other factors
- Placebo tests can be conducted at multiple points along the assignment variable to check for the robustness of the RD estimates

Applications of RD designs

- RD designs have been widely used in various fields, including economics, education, health, and public policy, to evaluate the causal effects of interventions and policies
- These applications demonstrate the versatility and relevance of RD methods for addressing important research questions

Education policy evaluations

- RD designs have been used to evaluate the effects of education policies, such as school entry age cutoffs (Angrist & Krueger, 1991), class size caps (Angrist & Lavy, 1999), and financial aid eligibility (Van der Klaauw, 2002)
- These studies exploit discontinuities in policy rules to estimate the causal effects of education inputs on student outcomes

Health insurance impacts

- RD designs have been applied to study the impacts of health insurance coverage on healthcare utilization and health outcomes

- Examples include the effects of Medicaid eligibility cutoffs (Card & Shore-Sheppard, 2004) and Medicare eligibility at age 65 (Card et al., 2008)
- These studies leverage discontinuities in insurance coverage rules to identify the causal effects of insurance on various outcomes

Effects of social programs

- RD methods have been used to evaluate the effects of social programs, such as welfare benefits (Lemieux & Milligan, 2008), unemployment insurance (Lalive, 2008), and housing assistance (Jacob & Ludwig, 2012)
- These studies exploit eligibility cutoffs or program rules to estimate the causal impacts of these programs on labor market outcomes, consumption, and well-being

Limitations and extensions

- While RD designs offer a powerful tool for causal inference, they also have some limitations and have been extended in various ways to address specific challenges
- These limitations and extensions are important to consider when applying RD methods in practice

External validity concerns

- RD estimates are local average treatment effects (LATE) that apply to units near the cutoff and may not generalize to units far from the cutoff or to other contexts
- The external validity of RD estimates depends on the similarity between the population of interest and the units around the cutoff
- Assessing the heterogeneity of treatment effects across different subgroups or using multiple cutoffs can help to address external validity concerns

Fuzzy RD with multiple cutoffs

- Fuzzy RD designs can be extended to settings with multiple cutoffs, where the probability of treatment changes discontinuously at each cutoff
- Multiple cutoffs can be used to estimate the treatment effect at different points along the assignment variable and to assess the heterogeneity of effects
- Estimation methods for fuzzy RD with multiple cutoffs include parametric (Angrist & Pischke, 2009) and non-parametric (Cattaneo et al., 2016) approaches

Combining RD with other methods

- RD designs can be combined with other methods, such as difference-in-differences (RD-DD) or regression kink designs (RKD), to strengthen causal inference
- RD-DD methods exploit both the discontinuity in treatment assignment and the variation in treatment timing to estimate the causal effect (Grembi et al., 2016)
- RKD methods exploit kinks in the relationship between the assignment variable and the treatment to identify the causal effect (Card et al., 2015)
- These combined methods can help to relax some of the assumptions of standard RD designs and improve the robustness of the estimates

7.2 Bandwidth selection and local polynomial regression

Bandwidth in local polynomial regression

Bandwidth selection controls how much data goes into each local fit in a local polynomial regression. Pick too narrow a window and your estimates jump around with noise; pick too wide and you blur over the very discontinuity you're trying to detect. In regression discontinuity (RD) designs, getting this right directly determines whether your treatment effect estimate is credible.

Importance of bandwidth selection

The bandwidth h defines the size of the neighborhood around each point where data contributes to the local polynomial fit. Observations within the bandwidth receive positive kernel weight; observations outside receive zero (or near-zero) weight.

This matters because:

- Too few observations in the window means high variance — your estimate is noisy and unstable across samples.
- Too many observations means you're averaging over a wide range of the running variable, which pulls in data far from the cutoff and introduces bias.
- In RD specifically, the treatment effect estimate is only valid at the cutoff, so you need enough data near the cutoff to be precise, but not so much that you're fitting the wrong functional form.

Bias-variance tradeoff

The core tension in bandwidth choice is between bias and variance, and the goal is to minimize mean squared error (MSE), which combines both:

$$MSE = Bias^2 + Variance$$

- Smaller h : Uses fewer observations near the target point. The local polynomial closely tracks the true function (low bias), but the estimate is noisy (high variance).
- Larger h : Uses more observations, stabilizing the estimate (low variance), but the polynomial must approximate the regression function over a wider range, which introduces systematic error (high bias).

The optimal bandwidth is the value of h that minimizes MSE. It shrinks as sample size grows — with more data, you can afford a narrower window and still keep variance under control.

Smaller vs. larger bandwidths

Think of this as a spectrum:

- Undersmoothing (small h): The fitted curve is wiggly and responsive to local features. Useful when the regression function has sharp turns or local structure, but sensitive to outliers and noise.
- Oversmoothing (large h): The fitted curve is smooth and stable but may miss genuine local patterns — including, in RD, the jump at the cutoff itself.

In practice, you'll often see researchers report results across a range of bandwidths to show that the treatment effect estimate isn't driven by one particular choice. This is a key part of RD robustness checks.

Methods for bandwidth selection

No single method dominates in all settings. The three main approaches each handle the bias-variance tradeoff differently.

Rule-of-thumb approach

This is the simplest option. It assumes the regression function is smooth and the errors are roughly normal, then derives a closed-form bandwidth:

$$h = c \cdot \sigma \cdot n^{-1/5}$$

where c is a constant (depending on the kernel), σ is the standard deviation of the residuals, and n is the sample size.

The $n^{-1/5}$ rate means the bandwidth shrinks slowly as you get more data. This method is fast to compute and gives a reasonable starting point, but it can perform poorly when the smoothness assumption is wrong — for instance, when the regression function has sharp local features near the cutoff.

Cross-validation techniques

Cross-validation lets the data choose the bandwidth without strong assumptions about the regression function.

1. Choose a set of candidate bandwidths.
2. For each candidate h , fit the local polynomial using all observations except one (leave-one-out) or except one fold (k -fold).
3. Predict the held-out observation(s) and compute the prediction error.
4. Repeat for every observation or fold.
5. Select the h that minimizes the average prediction error (typically MSE).

Leave-one-out cross-validation (LOOCV) uses every observation as its own validation set, which is thorough but computationally expensive for large datasets. K -fold cross-validation splits data into k groups and is faster but introduces some randomness from the fold assignment.

Cross-validation is data-driven and flexible, but it optimizes global fit rather than estimation at the cutoff specifically. For RD applications, this distinction matters.

Plug-in methods

Plug-in methods work directly with the asymptotic mean integrated squared error (AMISE) formula, which expresses the optimal bandwidth in terms of unknown quantities — typically the second derivative of the regression function and the noise variance.

1. Estimate the unknown quantities (e.g., fit a preliminary regression to estimate curvature).
2. Plug those estimates into the AMISE formula.
3. Solve for the bandwidth that minimizes the estimated AMISE.

Silverman's rule of thumb is a simple plug-in method. The direct plug-in (DPI) method is more sophisticated, iteratively refining the preliminary estimates. These methods are faster than cross-validation but only as good as the preliminary estimates they rely on. If the pilot estimate of curvature is poor, the

resulting bandwidth will be too.

For RD designs specifically, the most widely used approach is the method of Imbens and Kalyanaraman (2012) and its refinement by Calonico, Cattaneo, and Titiunik (CCT, 2014). CCT's procedure selects an MSE-optimal bandwidth and then applies a bias correction with robust confidence intervals. If you're doing applied RD work, the `rdrobust` package (available in R and Stata) implements this directly.

Local polynomial regression

Local polynomial regression fits a polynomial of order P in a neighborhood around each target point, weighting observations by a kernel function. Unlike global polynomial regression (which fits one polynomial to all the data), local polynomial regression adapts to the shape of the regression function at each location.

Advantages over kernel regression

Standard kernel regression (the Nadaraya-Watson estimator) fits a constant locally — it's essentially a weighted average. Local polynomial regression improves on this in several ways:

- Captures curvature: A local linear or quadratic fit approximates the slope and curvature of the true function, not just its level.
- Reduces boundary bias: Kernel regression is biased at the edges of the data because the kernel is asymmetric there. Local polynomials naturally adapt because the polynomial extrapolates into the sparse region.
- Better theoretical properties: Local polynomial estimators achieve faster convergence rates and lower minimax MSE compared to kernel estimators of the same order.

These advantages are especially relevant in RD, where you're estimating the regression function right at the cutoff — a boundary point by construction.

Fitting local polynomials

To estimate the regression function at a point X :

1. Assign weights to each observation i using a kernel function: $K\left(\frac{X_i - x}{h}\right)$, where h is the bandwidth.
2. Construct the local design matrix with polynomial terms: $1, (X_i - x), (X_i - x)^2, \dots, (X_i - x)^P$.
3. Solve the weighted least squares problem — minimize

$$\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) [Y_i - \beta_0 - \beta_1(X_i - x) - \dots - \beta_P(X_i - x)^P]^2$$

.

4. The estimate at X is $\hat{m}(x) = \hat{\beta}_0$ — the intercept of the local fit.

The resulting estimator can be written as a linear smoother:

$$\hat{m}(x) = \sum_{i=1}^n W_i(x) Y_i$$

where $W_i(x)$ are the equivalent kernel weights that come out of the weighted least squares solution.

Order of polynomial fit

The polynomial order P controls how flexible each local fit is:

- $p = 0$ (local constant): Equivalent to kernel regression. Simple but prone to boundary bias.
- $p = 1$ (local linear): The workhorse choice in RD designs. It handles boundary bias well and is relatively stable.
- $p = 2$ (local quadratic): Captures curvature better but introduces more variance, especially with small samples.

Higher-order polynomials can approximate more complex local shapes, but they also increase variance and can produce erratic behavior near the boundaries. In most applied RD work, local linear regression is the default because it strikes a good balance: it corrects for boundary bias at the cutoff without overfitting.

A useful rule: the polynomial order and the bandwidth interact. If you increase p , you typically need a larger bandwidth to keep variance manageable.

Boundary bias in local polynomial regression

In RD designs, the cutoff is literally a boundary — you estimate the regression function separately on each side, so every estimate at the cutoff is a one-sided boundary estimate. This makes boundary bias a first-order concern, not just a technical nuisance.

Causes of boundary bias

Boundary bias occurs because the kernel window extends beyond the range of available data on one side:

- At the cutoff in an RD design, observations exist only to the left (control) or only to the right (treated) of the point you're estimating.
- The kernel is effectively truncated, making the weighted average asymmetric.
- This asymmetry means the estimator systematically over- or under-estimates the true function value at the boundary.

The severity depends on the bandwidth (wider bandwidths worsen it), the kernel shape, and the curvature of the true regression function near the boundary.

Techniques for bias correction

Several strategies reduce boundary bias:

- Boundary kernels: Modified kernel functions (e.g., Epanechnikov boundary kernel, Gamma boundary kernel) that adjust their shape near boundaries to restore symmetry in the weighting. They work but require knowing where the boundaries are and can be tricky to implement.
- Local linear regression: Automatically adapts at boundaries because the linear fit extrapolates naturally. This is the most common practical solution in RD.
- Explicit bias correction: The CCT (2014) procedure estimates the leading bias term (which depends on the second derivative of the regression function) and subtracts it from the point estimate. The confidence intervals are then constructed to account for the remaining uncertainty from this correction.
- Resampling methods: Jackknife and bootstrap procedures can estimate and correct for boundary bias, though they add computational cost.

Boundary kernels vs. local linear regression

Both approaches address boundary bias, but they work differently:

- Boundary kernels fix the problem by reshaping the weights so the effective kernel is symmetric even at the boundary. They require explicit construction for each boundary region.
- Local linear regression fixes the problem implicitly — the linear term in the local fit absorbs the first-order bias that causes boundary distortion.

In practice, local linear regression is far more common in RD applications because it handles boundary bias as a built-in feature rather than requiring a separate correction. Boundary kernels are more common in density estimation and other nonparametric settings.

Applications in causal inference

Local polynomial regression is not just a curve-fitting tool — it's central to several causal inference strategies.

Estimating treatment effects

In observational studies, local polynomial regression can estimate treatment effects nonparametrically:

- Fit separate local polynomial regressions for treated and control groups as functions of covariates.
- The difference in fitted values at each covariate value estimates the conditional average treatment effect (CATE).
- Averaging over the appropriate population gives the ATE or ATT.

This approach avoids imposing a parametric functional form on the outcome-covariate relationship, reducing the risk of specification bias.

Regression discontinuity designs

This is where bandwidth selection and local polynomial regression come together most directly. In a sharp RD design:

1. Choose a bandwidth h around the cutoff c (using, e.g., the CCT optimal bandwidth selector).
2. Restrict the sample to observations with running variable values in $[c - h, c + h]$.
3. Fit a local polynomial (typically local linear) separately on each side of the cutoff.
4. The treatment effect estimate is the difference in the two fitted values at the cutoff:

$$\hat{\tau} = \hat{m}_+(c) - \hat{m}_-(c).$$

The bandwidth determines how much data informs the estimate. Robustness checks typically vary the bandwidth (e.g., half and double the optimal) and the polynomial order to confirm the result isn't sensitive to these choices.

Generalized propensity score estimation

For continuous or multi-valued treatments, the generalized propensity score (GPS) extends the standard propensity score framework. Local polynomial regression can estimate the GPS by modeling the conditional density of treatment given covariates.

Once estimated, the GPS is used to:

- Balance covariates across treatment levels.
- Estimate dose-response functions showing how outcomes vary with treatment intensity.

Local polynomial regression is well-suited here because the relationship between treatment assignment and covariates may be nonlinear, and a flexible estimator avoids the misspecification risk of parametric models.

7.3 Validity tests and sensitivity analysis

Validity tests and sensitivity analysis are crucial tools in causal inference. They help researchers assess the plausibility of key assumptions and quantify the robustness of causal estimates. By conducting these analyses, we can strengthen the credibility of our causal claims and identify potential sources of bias.

These methods work together to provide a comprehensive evaluation of causal models. Validity tests offer empirical evidence to support or question assumptions, while sensitivity analysis examines how estimates change under different scenarios. Combining both approaches allows for more robust and reliable causal conclusions.

Importance of validity tests

- Validity tests play a crucial role in causal inference by assessing the plausibility of key assumptions underlying causal models
- These tests help researchers gain confidence in the validity of their causal estimates and identify potential sources of bias or violations of assumptions
- By conducting validity tests, researchers can strengthen the credibility of their causal claims and ensure the robustness of their findings

Assessing causal model assumptions

- Validity tests allow researchers to empirically assess the assumptions required for valid causal inference, such as unconfoundedness, positivity, and consistency
- These assumptions are often untestable directly, but validity tests provide indirect evidence of their plausibility
- Examples of assumptions assessed by validity tests include the absence of unmeasured confounding (unconfoundedness), sufficient overlap in treatment probabilities (positivity), and well-defined interventions (consistency)

Detecting potential biases

- Validity tests can help detect potential biases that may invalidate causal estimates, such as selection bias, measurement error, or model misspecification
- By identifying these biases, researchers can take appropriate steps to address them, such as adjusting for confounding variables, improving measurement accuracy, or refining the causal model
- Detecting and mitigating potential biases is essential for obtaining reliable and unbiased causal estimates

Types of validity tests

- There are several types of validity tests commonly used in causal inference to assess different aspects of causal model assumptions
- These tests provide empirical evidence to support or question the validity of causal estimates

- The choice of validity tests depends on the specific assumptions being assessed and the nature of the data and causal question at hand

Tests for unconfoundedness

- Unconfoundedness assumes that treatment assignment is independent of potential outcomes given observed covariates
- Tests for unconfoundedness assess whether there are any unmeasured confounders that may bias causal estimates
- Examples of tests for unconfoundedness include placebo tests (checking for treatment effects on pre-treatment outcomes) and covariate balance tests (assessing balance of covariates between treatment groups)

Tests for positivity

- Positivity assumes that there is a non-zero probability of receiving each treatment level for all covariate values
- Tests for positivity assess whether there is sufficient overlap in treatment probabilities across covariate strata
- Examples of tests for positivity include checking for extreme propensity scores (near 0 or 1) and assessing common support of propensity score distributions between treatment groups

Tests for consistency

- Consistency assumes that the observed outcome under a given treatment is the same as the potential outcome under that treatment
- Tests for consistency assess whether the treatment is well-defined and consistently applied across units
- Examples of tests for consistency include checking for treatment effect heterogeneity across subgroups and assessing the stability of treatment effects over time or across different settings

Implementing validity tests

- Implementing validity tests involves selecting appropriate tests based on the causal assumptions being assessed and the available data
- Validity tests should be carefully designed and executed to provide meaningful and reliable results
- The results of validity tests should be interpreted cautiously, considering their limitations and potential alternative explanations

Placebo tests

- Placebo tests assess the presence of treatment effects on outcomes that should not be affected by the treatment (placebo outcomes)
- If significant placebo effects are found, it suggests the presence of unmeasured confounding or other biases
- Placebo tests can be conducted by selecting pre-treatment outcomes or outcomes that are known to be unrelated to the treatment

Negative control tests

- Negative control tests assess the presence of treatment effects on outcomes that are known to be unaffected by the treatment (negative control outcomes)
- If significant effects are found on negative control outcomes, it suggests the presence of unmeasured confounding or other biases
- Negative control tests can be conducted by selecting outcomes that are similar to the main outcome but are not causally affected by the treatment

Covariate balance tests

- Covariate balance tests assess the balance of observed covariates between treatment groups
- If significant imbalances are found, it suggests the presence of confounding or selection bias
- Covariate balance tests can be conducted using statistical tests (t-tests, chi-square tests) or standardized mean differences to compare covariate distributions between treatment groups

Sensitivity analysis overview

- Sensitivity analysis is a technique used to assess the robustness of causal estimates to potential violations of causal assumptions
- It involves systematically varying the assumptions or parameters of the causal model and examining the impact on the causal estimates
- Sensitivity analysis helps quantify the degree to which causal conclusions are sensitive to unverified assumptions or potential sources of bias

Purpose of sensitivity analysis

- The primary purpose of sensitivity analysis is to assess the robustness of causal estimates to potential violations of causal assumptions
- It helps researchers understand how sensitive their causal conclusions are to unverified assumptions or potential sources of bias
- Sensitivity analysis provides a way to quantify the degree of uncertainty or skepticism surrounding causal claims

Quantifying robustness of causal estimates

- Sensitivity analysis allows researchers to quantify the robustness of causal estimates by measuring the degree to which the estimates change under different assumptions or scenarios
- It involves specifying a range of plausible values for sensitivity parameters (e.g., the strength of unmeasured confounding) and examining the impact on the causal estimates
- By quantifying the robustness of causal estimates, researchers can assess the credibility and reliability of their causal conclusions

Types of sensitivity analysis

- There are several types of sensitivity analysis techniques used in causal inference, each with its own strengths and limitations
- The choice of sensitivity analysis method depends on the specific causal assumptions being assessed and the nature of the data and causal question at hand
- Different sensitivity analysis methods may yield different insights into the robustness of causal estimates

Rosenbaum bounds

- Rosenbaum bounds is a sensitivity analysis method that assesses the robustness of causal estimates to potential unmeasured confounding
- It quantifies the degree of unmeasured confounding that would be needed to invalidate the causal conclusions
- Rosenbaum bounds provide a worst-case scenario analysis by determining the maximum amount of hidden bias that can be present while still maintaining a significant treatment effect

Simulation-based approaches

- Simulation-based sensitivity analysis involves simulating data under different assumptions or scenarios and examining the impact on the causal estimates
- It allows researchers to assess the sensitivity of causal estimates to a wide range of potential biases or violations of assumptions
- Examples of simulation-based approaches include Monte Carlo simulations and Bayesian sensitivity analysis

Bias formulas

- Bias formulas are mathematical expressions that quantify the bias in causal estimates due to specific sources of bias, such as unmeasured confounding or selection bias
- These formulas allow researchers to assess the sensitivity of causal estimates to different levels of bias and to determine the conditions under which the causal conclusions remain valid
- Examples of bias formulas include the VanderWeele-Arah bias formula for unmeasured confounding and the Greenland bias formula for selection bias

Conducting sensitivity analysis

- Conducting sensitivity analysis involves specifying sensitivity parameters, running the analysis, and interpreting the results
- Sensitivity analysis should be carefully planned and executed to provide meaningful and reliable insights into the robustness of causal estimates
- The results of sensitivity analysis should be reported transparently, including the assumptions made, the range of sensitivity parameters considered, and the impact on the causal conclusions

Specifying sensitivity parameters

- Specifying sensitivity parameters involves defining the range of plausible values for the parameters that capture potential violations of causal assumptions
- These parameters may include the strength of unmeasured confounding, the degree of selection bias, or the level of measurement error
- The choice of sensitivity parameters should be based on subject matter knowledge, previous research, or reasonable worst-case scenarios

Interpreting sensitivity results

- Interpreting sensitivity results involves examining how the causal estimates change under different sensitivity parameters or scenarios
- If the causal estimates remain stable and statistically significant across a wide range of sensitivity parameters, it suggests that the causal conclusions are robust to potential violations of assumptions
- If the causal estimates are highly sensitive to small changes in the sensitivity parameters, it indicates that the causal conclusions are less reliable and should be interpreted with caution

Reporting sensitivity findings

- Reporting sensitivity findings involves presenting the results of the sensitivity analysis in a clear and transparent manner
- This includes describing the sensitivity parameters considered, the range of values explored, and the impact on the causal estimates and conclusions
- Sensitivity findings should be discussed in the context of the strengths and limitations of the study, and any implications for the interpretation and generalizability of the results should be addressed

Validity tests vs sensitivity analysis

- Validity tests and sensitivity analysis are complementary approaches used in causal inference to assess the robustness and credibility of causal estimates
- While validity tests focus on assessing the plausibility of causal assumptions, sensitivity analysis examines the robustness of causal estimates to potential violations of those assumptions
- Both validity tests and sensitivity analysis are important tools for strengthening the validity and reliability of causal claims

Complementary roles in causal inference

- Validity tests and sensitivity analysis play complementary roles in causal inference by addressing different aspects of the causal modeling process
- Validity tests provide empirical evidence to support or question the plausibility of causal assumptions, while sensitivity analysis quantifies the robustness of causal estimates to potential violations of those assumptions
- Together, validity tests and sensitivity analysis provide a comprehensive assessment of the validity and reliability of causal conclusions

Combining both for robust conclusions

- Combining validity tests and sensitivity analysis is a powerful approach for drawing robust causal conclusions
- By conducting both types of analyses, researchers can gain a more complete understanding of the strengths and limitations of their causal claims
- If both validity tests and sensitivity analysis provide consistent and supportive results, it strengthens the credibility and robustness of the causal conclusions
- If the results of validity tests and sensitivity analysis are inconsistent or raise concerns, it highlights the need for further investigation or cautious interpretation of the causal findings

Unit 8 – Difference-in-Differences & Panel Data Methods

8.1 Difference-in-differences (DiD) estimation

Definition of DiD

Difference-in-differences (DiD) is a quasi-experimental research design used to estimate the causal effect of a treatment or intervention. It works by comparing how outcomes change over time between a group that received the treatment and a group that didn't. By looking at changes rather than levels, DiD controls for both time-invariant differences between groups and group-invariant changes over time.

The core logic: if both groups were on similar trajectories before the intervention, then any divergence afterward can be attributed to the treatment. This is what makes the parallel trends assumption so central to the whole approach.

Setup for DiD

DiD requires data on both treatment and control groups, observed both before and after the intervention. This can come from either panel data (same units tracked over time) or repeated cross-sections (different samples from the same populations at different times).

- The treatment group is exposed to the intervention at a specific point in time
- The control group remains unexposed throughout the study period
- The outcome variable is measured for both groups at baseline (pre-intervention) and follow-up (post-intervention)

Parallel Trends Assumption

This is the assumption that makes or breaks a DiD design. It states that, in the absence of treatment, the average change in the outcome would have been the same for both groups. In other words, the control group's trajectory serves as the counterfactual for what would have happened to the treatment group had they not been treated.

You can never directly test this assumption (because you can't observe the counterfactual), but you can assess its plausibility by checking whether the two groups followed similar trends in the pre-intervention period. If they were already diverging before the treatment, your DiD estimate is likely biased.

DiD Estimation

The DiD estimator isolates the treatment effect by taking a "difference of differences": the before-after change in the treatment group, minus the before-after change in the control group. This double-differencing removes both fixed group differences and common time shocks.

Simple Two-Period DiD

In the simplest case, you have two groups and two time periods. The estimator is:

$$\hat{\tau}_{DiD} = (\bar{Y}_{Treatment, Post} - \bar{Y}_{Treatment, Pre}) - (\bar{Y}_{Control, Post} - \bar{Y}_{Control, Pre})$$

where $\bar{Y}$ is the average outcome for each group-period cell.

This same estimator falls out of a regression framework. You estimate:

$$Y_{it} = \beta_0 + \beta_1 \cdot \text{Treat}_i + \beta_2 \cdot \text{Post}_t + \beta_3 \cdot (\text{Treat}_i \times \text{Post}_t) + \varepsilon_{it}$$

- β_1 captures baseline differences between groups
- β_2 captures the common time trend
- β_3 is the DiD estimate of the treatment effect

Multi-Period DiD

DiD extends naturally to settings with multiple pre- and post-intervention periods. With more time periods, you can:

- Examine whether the treatment effect grows, shrinks, or stays constant over time
- Provide stronger evidence on pre-trends (more pre-periods to check)
- Estimate more flexible time trend specifications

The basic logic is the same: compare the trajectory of the treatment group to that of the control group, now averaged across multiple periods.

Two-Way Fixed Effects Model

The standard regression implementation of DiD with panel data is the two-way fixed effects (TWFE) model:

$$Y_{it} = \alpha_i + \gamma_t + \beta \cdot D_{it} + \varepsilon_{it}$$

- α_i = unit fixed effects (absorb all time-invariant differences across units)
- γ_t = time fixed effects (absorb all common shocks in each period)
- D_{it} = treatment indicator (equals 1 for treated units in post-treatment periods)
- β = the estimated treatment effect

TWFE can handle multiple treatment and control groups, as well as staggered adoption designs. However, recent econometrics research has shown that TWFE can produce misleading estimates under staggered adoption with heterogeneous treatment effects. This is covered further in the extensions section below.

Key Assumptions

The validity of DiD depends on several assumptions beyond parallel trends. Each one should be carefully evaluated in any application.

Common Trends Assumption

Already discussed above, but worth restating: the treatment and control groups must have been on parallel outcome trajectories prior to the intervention. Researchers typically assess this by:

1. Plotting the outcome variable for both groups across pre-treatment periods
2. Estimating an event study specification and checking that pre-treatment coefficients are close to zero and statistically insignificant

3. Running formal statistical tests for differential pre-trends

Passing these checks doesn't prove parallel trends hold, but failing them is a strong warning sign.

Stable Unit Treatment Value Assumption (SUTVA)

SUTVA requires that one unit's treatment status doesn't affect another unit's outcomes. There should be no spillover effects or interference between units.

Violations are common in practice. For example, if a minimum wage increase in one state causes firms to relocate to a neighboring control state, the control group's outcomes are contaminated. Possible responses include redefining the unit of analysis to a broader geographic level or using methods that explicitly model spillovers.

No Anticipation Effects

Units should not change their behavior in anticipation of future treatment. If treated firms start adjusting their hiring before a policy takes effect, the pre-treatment data is contaminated, and the DiD estimate will be biased.

You can check for anticipation effects by looking at whether outcomes start shifting in the treatment group during the periods just before the intervention. Event study designs are particularly useful for this.

Interpreting DiD Estimates

Average Treatment Effect on the Treated (ATT)

Under the parallel trends assumption, DiD identifies the average treatment effect on the treated (ATT): the average causal effect for units that actually received the treatment. This is distinct from the average treatment effect (ATE), which is the average effect across the entire population.

- If treatment effects are homogeneous (the same for everyone), $ATT = ATE$
- If treatment effects are heterogeneous, ATT and ATE can differ. DiD tells you about the treated group specifically, not necessarily what would happen if you treated everyone

Heterogeneous Treatment Effects

Treatment effects often vary across subgroups. You can explore this by interacting the treatment indicator with subgroup indicators (e.g., age, income, region). This reveals which populations benefit most from the intervention and provides a richer picture of the policy's distributional impacts.

Robustness Checks

DiD estimates should always be stress-tested. Here are the most common approaches.

Testing for Pre-Trends

This is the single most important robustness check. Estimate a model that includes leads of the treatment indicator (interactions between the treatment group dummy and pre-intervention time period dummies).

- If these lead coefficients are close to zero and insignificant, that's reassuring
- If they're significant, it suggests the groups were already diverging before the intervention, casting doubt on the parallel trends assumption

An event study design plots these coefficients period by period, giving you a visual diagnostic of both pre-trends and the dynamic treatment effect.

Placebo Tests

Placebo tests check whether your design picks up effects where none should exist:

- Placebo outcomes: Estimate DiD on an outcome that shouldn't be affected by the treatment. A significant effect suggests confounding.
- Placebo treatments: Randomly reassign the treatment to different units or time periods. The DiD estimate should be approximately zero.
- Placebo timing: Move the treatment date to a pre-intervention period and re-estimate. A significant effect suggests pre-existing trends.

Triple Differences (DDD)

Triple differences adds a third layer of differencing by comparing DiD estimates across two subgroups that are differentially affected by the treatment. For example, if a policy targets older workers but not younger workers, you can compute DiD separately for each age group and then take the difference.

DDD helps control for time-varying confounders that affect both treatment and control groups equally, as long as those confounders don't differentially impact the two subgroups. The tradeoff is that DDD requires additional assumptions and a more complex data structure.

Extensions of DiD

Staggered Adoption DiD

In many real-world settings, treatment rolls out at different times for different units (e.g., states adopting a policy in different years). The standard TWFE approach has traditionally been used here, but recent research (Goodman-Bacon 2021, Callaway and Sant'Anna 2021, Sun and Abraham 2021) has shown that TWFE can produce biased estimates when treatment effects vary across units or over time.

The problem: TWFE implicitly uses already-treated units as controls for later-treated units, and if treatment effects evolve over time, these comparisons are contaminated. Newer estimators address this by:

- Restricting comparisons to never-treated or not-yet-treated units as controls
- Estimating group-time specific treatment effects and then aggregating them
- Allowing for heterogeneous and dynamic treatment effects

Synthetic Control Methods vs. DiD

Synthetic control methods (SCM) construct a weighted combination of control units that best matches the treated unit's pre-intervention outcomes. The treatment effect is the gap between the treated unit and its synthetic counterpart after the intervention.

DiD works well with many treated and control units and assumes parallel trends. SCM is designed for settings with a single (or few) treated unit(s) and many potential controls, and it relaxes the parallel trends assumption by matching on pre-treatment outcomes directly.

SCM can be seen as a more flexible version of DiD, but it has its own limitations: it requires a good pre-treatment fit, and inference can be challenging.

Dynamic Treatment Effects

Treatment effects don't have to be constant over time. A dynamic DiD model interacts the treatment indicator with dummies for each post-treatment period, producing a separate treatment effect estimate for each period after the intervention.

This is closely related to the event study framework. The resulting plot shows:

- Pre-treatment coefficients (should be near zero if parallel trends hold)
- The treatment effect at each post-treatment period (showing whether effects grow, fade, or remain stable)

Limitations of DiD

Violations of Parallel Trends

If the treatment and control groups were on different trajectories before the intervention, DiD will attribute those pre-existing differences to the treatment. Common sources of violation include time-varying confounders that differentially affect the groups, or compositional changes in the groups over time.

Time-Varying Confounders

DiD eliminates time-invariant confounders by design, but it cannot handle confounders that change over time and affect the two groups differently. For instance, if a recession hits the treatment region harder than the control region at the same time as the policy change, DiD will conflate the recession's effect with the policy's effect.

Spillover Effects

When treatment affects untreated units (through economic linkages, social networks, or general equilibrium effects), SUTVA is violated and DiD estimates are biased. The direction of bias depends on whether spillovers push control group outcomes in the same or opposite direction as the treatment effect.

Applications of DiD

Policy Evaluation Examples

DiD is one of the most widely used methods in applied economics and policy evaluation. Classic examples include:

- Card and Krueger (1994): Estimated the effect of New Jersey's minimum wage increase on fast-food employment, using Pennsylvania as the control group. They found no significant negative employment effect, challenging the standard competitive model prediction.
- Finkelstein et al. (2012): Used the Oregon Health Insurance Experiment to study the effects of Medicaid expansion on health care utilization, financial strain, and health outcomes.
- Chay and Greenstone (2003): Evaluated the effects of the Clean Air Act Amendments on air quality and infant health.

DiD with Panel Data

Panel data (repeated observations on the same units) is the natural setting for DiD. It allows you to include unit fixed effects, examine pre-trends, and estimate dynamic treatment effects. One concern specific to panel data is attrition bias: if units drop out of the sample over time, and attrition is correlated with treatment, your estimates can be biased.

DiD with Repeated Cross-Sections

When you can't track the same units over time, you can still apply DiD using repeated cross-sections (different samples from the same populations at different times). This approach requires an additional assumption: the composition of the treatment and control groups must remain stable over time, or any compositional changes must be unrelated to the treatment.

This is common in settings where researchers use survey data to evaluate population-level policy impacts, such as the effect of state-level policy changes on health or labor market outcomes.

8.2 Parallel trends assumption

The parallel trends assumption is crucial for difference-in-differences estimation, a method used to measure causal effects in observational studies. It states that without treatment, the average outcomes for treatment and control groups would follow parallel paths over time.

This assumption allows researchers to attribute differences in outcomes between groups to the treatment itself. Violating this assumption can lead to biased estimates and incorrect conclusions about causal effects, making it essential to carefully assess and address potential violations.

Definition of parallel trends assumption

- Fundamental assumption in difference-in-differences (DiD) estimation, a popular method for estimating causal effects in observational studies
- Requires that in the absence of treatment, the average outcomes for the treatment and control groups would have followed parallel paths over time
- Implies that any differences in outcomes between the two groups after treatment can be attributed to the causal effect of the treatment itself

Importance in difference-in-differences estimation

- Crucial for the validity of DiD estimates as it allows for the isolation of the causal effect of interest
- Enables the control group to serve as a valid counterfactual for the treatment group, representing what would have happened to the treatment group in the absence of treatment
- Violation of this assumption can lead to biased estimates and incorrect conclusions about the causal effect of the treatment

Conditions for parallel trends assumption

Treatment vs control groups

- Requires that the treatment and control groups are similar in terms of their pre-treatment characteristics and trends

- Groups should be affected by the same external factors and shocks over time
- Differences between the groups should be stable and not vary systematically with the treatment

Pre-treatment trends

- Assumption is more plausible when the treatment and control groups exhibit similar trends in the outcome variable prior to the treatment
- Parallel pre-treatment trends suggest that the groups would have continued to follow similar paths in the absence of treatment
- Divergence in pre-treatment trends raises concerns about the validity of the assumption

Absence of time-varying confounding

- Requires that there are no unobserved factors that affect the outcome variable differently for the treatment and control groups over time
- Time-varying confounders can cause the groups to diverge even in the absence of treatment, violating the parallel trends assumption
- Presence of such confounders can lead to biased estimates of the treatment effect

Assessing the validity of parallel trends assumption

Visual inspection of pre-treatment trends

- Plotting the outcome variable for the treatment and control groups over time can provide a visual assessment of the parallel trends assumption
- Similar pre-treatment trends suggest that the assumption is plausible, while diverging trends raise concerns
- Visual inspection should be supplemented with statistical tests for more rigorous evaluation

Placebo tests

- Involve artificially assigning the treatment to a different time period or group where no effect is expected
- If the parallel trends assumption holds, placebo tests should yield estimates close to zero
- Significant placebo effects suggest that the assumption may be violated and the DiD estimates may be biased

Covariate balance tests

- Check whether the treatment and control groups are balanced in terms of observable characteristics before and after the treatment
- Significant differences in covariates over time can indicate the presence of time-varying confounding and violation of the parallel trends assumption
- Covariate balance tests help assess the comparability of the groups and the plausibility of the assumption

Consequences of violating parallel trends assumption

Biased causal effect estimates

- Violation of the parallel trends assumption can lead to biased estimates of the treatment effect
- If the treatment and control groups would have followed different paths even in the absence of treatment, the DiD estimate will capture not only the true treatment effect but also the difference in underlying trends
- Bias can be positive or negative, depending on the direction of the violation

Misinterpretation of results

- Biased estimates can lead to incorrect conclusions about the effectiveness of the treatment
- Policymakers and researchers may attribute changes in outcomes to the treatment when they are actually driven by other factors
- Misinterpretation of results can have serious consequences for policy decisions and future research directions

Strategies for addressing violations

Including covariates

- Adding relevant covariates to the DiD model can help control for observable time-varying confounders
- Covariates should be selected based on theoretical considerations and data availability
- Including covariates can reduce bias and improve the plausibility of the parallel trends assumption

Synthetic control methods

- Involve constructing a synthetic control group as a weighted combination of untreated units that closely resembles the treatment group in terms of pre-treatment characteristics and trends
- Synthetic control methods can help address violations of the parallel trends assumption by creating a more suitable comparison group
- Requires a sufficient number of untreated units and careful selection of weighting variables

Triple difference estimators

- Extend the DiD approach by adding a third difference, such as a comparison group that is unaffected by the treatment
- Triple difference estimators can help control for time-varying confounders that affect both the treatment and control groups
- Requires the identification of a suitable third comparison group and additional assumptions about the nature of the confounding

Limitations of parallel trends assumption

Unobserved time-varying confounders

- Parallel trends assumption can be violated by the presence of unobserved factors that affect the treatment and control groups differently over time
- Such confounders are often difficult to measure or control for in observational studies
- Unobserved time-varying confounders can lead to biased estimates even if the parallel trends assumption appears to hold based on observable characteristics

Anticipation effects

- Occur when individuals or entities change their behavior in anticipation of the treatment, even before it is implemented
- Anticipation effects can cause the treatment and control groups to diverge prior to the actual treatment, violating the parallel trends assumption
- Ignoring anticipation effects can lead to biased estimates of the treatment effect

Functional form misspecification

- Parallel trends assumption is often tested and assessed using linear models, such as linear regression
- If the true relationship between the outcome and time is non-linear, the assumption may appear to hold even if it is actually violated
- Misspecification of the functional form can lead to incorrect conclusions about the validity of the parallel trends assumption and the causal effect of the treatment

Alternatives to difference-in-differences

Regression discontinuity designs

- Exploit a discontinuity in treatment assignment based on a continuous variable (running variable)
- Compares outcomes for units just above and below a cutoff value of the running variable
- Relies on the assumption that units near the cutoff are similar in terms of unobservable characteristics

Instrumental variables

- Use an exogenous source of variation (instrument) that affects the treatment but not the outcome directly
- Instrument should be correlated with the treatment and uncorrelated with unobserved confounders
- Allows for the estimation of causal effects in the presence of unmeasured confounding

Matching methods

- Involve pairing treated units with similar untreated units based on observable characteristics
- Matching methods aim to create a balanced sample that mimics a randomized experiment

- Common matching techniques include propensity score matching and coarsened exact matching

8.3 Fixed effects models

Fixed effects models are a powerful tool for controlling unobserved confounding in panel data. They estimate causal effects by leveraging within-group variation over time, assuming time-invariant confounders are correlated with explanatory variables.

These models offer advantages like reducing omitted variable bias, but have limitations such as inability to estimate time-invariant effects. Understanding when to use fixed effects versus alternatives like random effects is crucial for accurate causal inference in panel data analysis.

Definition of fixed effects models

- Fixed effects models are a type of regression model used to control for unobserved confounding in panel data or repeated measures data
- Assume that the unobserved confounders are time-invariant and correlated with the explanatory variables
- Estimate the effect of the time-varying explanatory variables while controlling for the fixed effects

Assumptions in fixed effects models

- The unobserved confounders are time-invariant within each individual or group
- The explanatory variables are strictly exogenous conditional on the fixed effects
- The idiosyncratic errors are uncorrelated with the explanatory variables and the fixed effects
- The idiosyncratic errors are homoscedastic and serially uncorrelated

Advantages of fixed effects models

Control for unobserved confounders

- Fixed effects models can control for unobserved confounders that are constant over time within each individual or group
- This helps to reduce omitted variable bias and obtain more consistent estimates of the causal effect
- Examples of unobserved confounders include individual ability, motivation, or family background

Consistent estimates with panel data

- Fixed effects models can provide consistent estimates of the causal effect even if the unobserved confounders are correlated with the explanatory variables
- This is because the fixed effects absorb the time-invariant confounders and the within-group variation is used to estimate the effect
- Panel data with multiple observations per individual or group over time is required to estimate fixed effects models

Limitations of fixed effects models

Cannot estimate time-invariant effects

- Fixed effects models cannot estimate the effect of time-invariant variables because they are absorbed by the fixed effects
- This includes variables such as gender, race, or region that do not change over time within each individual or group
- If the effect of time-invariant variables is of interest, random effects models or other approaches may be more appropriate

Inefficient with low within-group variation

- Fixed effects models rely on the within-group variation in the explanatory variables to estimate the effect
- If there is little variation in the explanatory variables within each individual or group over time, the fixed effects estimates may be imprecise and inefficient
- This can lead to large standard errors and low statistical power to detect significant effects

Fixed effects vs random effects models

- Fixed effects models assume that the unobserved confounders are correlated with the explanatory variables, while random effects models assume they are uncorrelated
- Fixed effects models control for all time-invariant confounders, while random effects models only control for the included time-invariant variables
- Fixed effects models are consistent under weaker assumptions, but random effects models are more efficient if the assumptions hold
- The choice between fixed and random effects models depends on the nature of the unobserved confounders and the research question

Estimating fixed effects models

Within transformation

- The within transformation subtracts the individual or group means from each variable to remove the fixed effects
- The transformed variables are then used in a standard OLS regression to estimate the fixed effects model
- This approach is equivalent to including a dummy variable for each individual or group in the regression

First differences

- First differencing subtracts the lagged values from each variable to remove the fixed effects
- The differenced variables are then used in a standard OLS regression to estimate the fixed effects model

- This approach is equivalent to the within transformation if there are only two time periods per individual or group

Least squares dummy variable approach

- The least squares dummy variable (LSDV) approach includes a dummy variable for each individual or group in the regression
- This directly estimates the fixed effects as the coefficients on the dummy variables
- The LSDV approach is computationally equivalent to the within transformation, but may be less efficient with many individuals or groups

Interpreting fixed effects estimates

- The fixed effects estimate represents the average change in the outcome variable for a one-unit change in the explanatory variable, holding the fixed effects constant
- The interpretation is within-group: it compares the outcome for the same individual or group at different times with different values of the explanatory variable
- The fixed effects estimate does not generalize to the population level or to other individuals or groups not included in the sample

Extensions of fixed effects models

Fixed effects vector decomposition

- Fixed effects vector decomposition (FEVD) is a method to estimate the effects of time-invariant variables in a fixed effects model
- FEVD decomposes the fixed effects into a part explained by the time-invariant variables and an unexplained part
- The unexplained part is then used as an additional regressor in a standard fixed effects model to estimate the effects of both time-varying and time-invariant variables

Interacted fixed effects

- Interacted fixed effects models include interactions between the individual or group fixed effects and time fixed effects
- This allows the unobserved confounders to have different effects at different times, relaxing the assumption of time-invariant effects
- Interacted fixed effects models can control for unobserved confounders that vary over time in a limited way, such as individual-specific time trends

Applications of fixed effects models

Panel data examples

- Fixed effects models are commonly used with panel data on individuals, households, firms, or countries over time

- Examples include estimating the effect of education on earnings, the impact of minimum wage laws on employment, or the relationship between democracy and economic growth
- Panel data allows researchers to control for unobserved individual-specific confounders and estimate causal effects

Difference-in-differences designs

- Difference-in-differences (DiD) designs are a special case of fixed effects models used to estimate the effect of a policy or treatment
- DiD compares the change in outcomes before and after the policy for a treatment group to the change in outcomes for a control group
- The DiD estimate is equivalent to a fixed effects estimate with a binary treatment variable interacted with time fixed effects

Specification tests for fixed effects

Hausman test

- The Hausman test compares the fixed effects and random effects estimates to test for correlation between the unobserved confounders and the explanatory variables
- If the test rejects the null hypothesis of no correlation, fixed effects models are preferred over random effects models
- The Hausman test helps to choose between fixed and random effects models based on the assumptions about the unobserved confounders

Testing for serial correlation

- Serial correlation in the idiosyncratic errors can lead to biased standard errors and invalid inference in fixed effects models
- Tests for serial correlation, such as the Breusch-Godfrey test or the Arellano-Bond test, can be used to detect serial correlation in fixed effects models
- If serial correlation is present, robust standard errors or alternative estimators (such as first differences or generalized least squares) may be needed

Best practices for fixed effects models

Choosing between fixed and random effects

- The choice between fixed and random effects models should be based on the assumptions about the unobserved confounders and the research question
- If the unobserved confounders are likely to be correlated with the explanatory variables, fixed effects models are preferred
- If the unobserved confounders are likely to be uncorrelated with the explanatory variables and the goal is to estimate the effects of time-invariant variables, random effects models may be more appropriate

Reporting fixed effects results

- When reporting fixed effects results, it is important to clearly state the assumptions and limitations of the model
- The fixed effects estimates should be interpreted as within-group effects, not population-level effects
- The number of individuals or groups and the number of time periods should be reported to assess the precision and reliability of the estimates
- Specification tests and robustness checks should be reported to support the validity of the fixed effects model

8.4 Synthetic control methods

Synthetic control methods offer a powerful tool for estimating causal effects in settings with a single treated unit and limited control options. By constructing a weighted combination of control units, researchers can create a data-driven counterfactual to compare against the treated unit.

This approach allows for more accurate treatment effect estimates compared to traditional methods. It relies on key assumptions like similar pre-treatment trends and no spillover effects. Researchers must carefully select predictor variables, optimize weights, and conduct sensitivity analyses to ensure robust results.

Overview of synthetic control methods

- Synthetic control methods provide a data-driven approach to estimate the causal effect of an intervention or treatment on a single unit (state, country, firm) by constructing a weighted combination of control units that closely resembles the treated unit in the pre-treatment period
- Enables researchers to create a counterfactual scenario representing what would have happened to the treated unit in the absence of the intervention, allowing for a more accurate estimation of the treatment effect compared to traditional methods that rely on a single control unit or a simple average of multiple controls
- Particularly useful in settings where there is a single treated unit and a small number of potential control units, making it difficult to find a suitable comparison group using conventional methods such as difference-in-differences or matching

Key assumptions of synthetic control methods

- The treated unit and control units should have similar characteristics and trends in the pre-treatment period, ensuring that the synthetic control provides a valid counterfactual for the treated unit in the absence of the intervention
- There should be no spillover effects or interference between the treated unit and control units, meaning that the intervention in the treated unit does not affect the outcomes of the control units
- The intervention should be exogenous and not anticipated by the units, as anticipation effects could lead to changes in behavior or outcomes prior to the actual implementation of the treatment
- The relationship between the predictor variables and the outcome variable should be stable over time, implying that the model used to estimate the weights for the synthetic control is correctly specified and captures the relevant factors influencing the outcome

Constructing synthetic control units

Selecting predictor variables

- Choose variables that are strong predictors of the outcome variable and are not affected by the intervention to ensure that the synthetic control closely mimics the treated unit in the pre-treatment period
- Include variables that capture relevant characteristics of the units, such as demographic, economic, or geographic factors, depending on the context of the study
- Consider using lagged values of the outcome variable as predictors to account for pre-treatment trends and improve the fit of the synthetic control

Choosing weights for control units

- Assign weights to each control unit based on their similarity to the treated unit in terms of the selected predictor variables
- Weights should be non-negative and sum to one, ensuring that the synthetic control represents a convex combination of the control units
- Control units with weights close to zero contribute little to the synthetic control, while those with higher weights are more influential in constructing the counterfactual

Optimizing weights to minimize pre-treatment differences

- Use an optimization algorithm (constrained least squares) to find the set of weights that minimizes the difference between the treated unit and the synthetic control in the pre-treatment period
- The objective function typically involves minimizing the mean squared prediction error (MSPE) or a similar measure of goodness-of-fit between the treated unit and the synthetic control
- The optimization process ensures that the resulting synthetic control provides the best possible match for the treated unit based on the available data and chosen predictor variables

Estimating treatment effects with synthetic controls

Comparing treated unit to synthetic control

- After constructing the synthetic control, compare the post-treatment outcomes of the treated unit to those of the synthetic control to estimate the causal effect of the intervention
- The difference between the actual outcome of the treated unit and the counterfactual outcome represented by the synthetic control provides an estimate of the treatment effect at each post-treatment time period
- Visualize the results by plotting the outcomes of the treated unit and the synthetic control over time, with the intervention point clearly marked

Interpreting estimated treatment effects

- The estimated treatment effect represents the impact of the intervention on the treated unit, measured in terms of the outcome variable

- A positive treatment effect indicates that the intervention had a beneficial impact on the treated unit, while a negative effect suggests that the intervention was detrimental
- Interpret the magnitude of the treatment effect in the context of the specific study and the scale of the outcome variable, considering the practical significance of the results

Advantages vs traditional methods

- Synthetic control methods can provide more accurate estimates of treatment effects compared to traditional methods (difference-in-differences, matching) when there is a single treated unit and a limited number of control units
- By constructing a data-driven counterfactual that closely mimics the treated unit, synthetic controls reduce the risk of bias due to unobserved confounders or differences in pre-treatment trends
- Synthetic controls allow for the estimation of dynamic treatment effects over time, capturing the evolution of the impact of the intervention rather than providing a single average effect estimate

Inference and uncertainty in synthetic control estimates

Placebo tests for statistical significance

- Assess the statistical significance of the estimated treatment effects by conducting placebo tests, which involve applying the synthetic control method to control units as if they were treated
- Generate a distribution of placebo treatment effects by iteratively assigning the intervention to each control unit and estimating the effect using the remaining units as potential controls
- Compare the actual treatment effect to the distribution of placebo effects to determine the likelihood of observing an effect of the same magnitude or larger by chance, providing a p-value for the significance of the results

Sensitivity analysis for robustness

- Test the robustness of the synthetic control estimates by varying the set of predictor variables, the optimization method, or the pool of potential control units
- Assess the sensitivity of the results to the inclusion or exclusion of specific control units that may have a disproportionate influence on the synthetic control
- Examine the stability of the weights assigned to the control units and the goodness-of-fit measures across different specifications to ensure that the results are not driven by arbitrary modeling choices

Confidence intervals for treatment effects

- Construct confidence intervals around the estimated treatment effects to quantify the uncertainty associated with the synthetic control estimates
- Use methods such as bootstrapping (resampling) or permutation tests to generate a distribution of treatment effect estimates and derive the corresponding confidence intervals
- Interpret the confidence intervals in terms of the range of plausible values for the true treatment effect, considering the width of the intervals and their implications for the significance and precision of the results

Extensions and variations of synthetic control methods

Multiple treated units or time periods

- Extend the basic synthetic control framework to settings with multiple treated units or multiple treatment periods by constructing separate synthetic controls for each treated unit or time period
- Pool the estimates across treated units or time periods to obtain an average treatment effect, taking into account the variability and potential heterogeneity of the individual estimates
- Use meta-analytic techniques (random effects models) to combine the results and account for the uncertainty associated with each estimate

Synthetic control methods vs other matching techniques

- Compare the performance and assumptions of synthetic control methods to other matching techniques, such as propensity score matching or coarsened exact matching
- Assess the relative strengths and weaknesses of each approach in terms of reducing bias, handling multiple treated units, and accommodating time-varying confounders
- Consider the trade-offs between the data-driven nature of synthetic controls and the potential for model misspecification, as well as the interpretability and transparency of the matching process

Combining synthetic controls with difference-in-differences

- Integrate synthetic control methods with difference-in-differences (DID) estimation to further improve the causal inference by accounting for both time-invariant and time-varying confounders
- Construct a synthetic control for the treated unit and estimate the treatment effect using a DID approach, comparing the change in outcomes between the treated unit and its synthetic control before and after the intervention
- Exploit the strengths of both methods, with synthetic controls providing a more accurate counterfactual and DID controlling for unobserved time-invariant factors, to obtain more robust and reliable estimates of the treatment effect

Applications of synthetic control methods

Examples from economics and public policy

- Apply synthetic control methods to evaluate the impact of economic policies (minimum wage laws), public health interventions (smoking bans), or environmental regulations (carbon taxes) on various outcomes (employment, health, emissions)
- Use synthetic controls to assess the effectiveness of regional development programs (enterprise zones) or infrastructure projects (transportation networks) on economic growth and social welfare
- Analyze the consequences of political events (elections, regime changes) or social movements (protests, strikes) on political, economic, or social outcomes using synthetic control methods

Evaluating impact of interventions or shocks

- Employ synthetic control methods to estimate the causal effect of natural disasters (earthquakes, hurricanes) or public health crises (pandemics) on economic, health, or social outcomes

- Assess the impact of policy interventions (gun control laws, immigration policies) or institutional reforms (education system changes, judicial reforms) on relevant outcomes using synthetic controls
- Evaluate the effectiveness of international aid programs (development assistance) or trade agreements (free trade zones) on economic growth, poverty reduction, or other development indicators in recipient countries

Limitations and caveats of synthetic control approach

- Synthetic control methods rely on the availability of a sufficient number of suitable control units and relevant predictor variables to construct a valid counterfactual, which may be challenging in some settings
- The approach assumes that the relationship between the predictor variables and the outcome is stable over time, and violations of this assumption can lead to biased estimates of the treatment effect
- Synthetic controls do not account for unobserved time-varying confounders that may affect the treated unit and control units differently, potentially biasing the results
- The method provides an estimate of the treatment effect for a single treated unit, and the generalizability of the findings to other units or contexts may be limited, requiring careful interpretation and additional analyses to assess external validity

Unit 9 – Causal Graphs & Structural Models

9.1 Directed acyclic graphs (DAGs)

Definition of DAGs

A Directed Acyclic Graph (DAG) is a graphical model that represents causal relationships between variables. It consists of nodes (representing variables) and directed edges (arrows representing causal relationships). DAGs give you a visual way to lay out your causal assumptions so you can reason clearly about what causes what in a complex system.

Nodes and Edges

Nodes represent variables or factors in a causal system, such as age, treatment assignment, or health outcome. Directed edges are arrows that indicate the direction of causal influence from one node to another.

- An edge from node A to node B ($A \rightarrow B$) means A has a direct causal effect on B.
- If there's no edge between two nodes, you're assuming there's no direct causal relationship between them.

That second point is easy to overlook but really matters: every missing edge in a DAG is itself a causal claim.

Acyclic Property

The "acyclic" part of DAG means the graph contains no directed cycles. You can't start at a node, follow the arrows, and end up back where you started. In other words, a variable can never directly or indirectly cause itself.

This property is what gives DAGs a clear causal ordering. Without it, you'd run into paradoxes where A causes B causes A, and the whole framework for reasoning about "upstream" and "downstream" effects breaks down.

Directed Paths

A directed path is a sequence of nodes connected by arrows, all pointing in the same direction. If there's a directed path from A to B ($A \rightarrow \dots \rightarrow B$), that means A has a causal effect on B, either directly or through intermediary variables.

Tracing directed paths is how you map out the full chain of causal influence. It also helps you spot where confounding might enter the picture, since confounders create non-causal paths between variables you care about.

Markov Properties of DAGs

The Markov properties connect a DAG's structure to probabilistic independence relationships among the variables. These properties are what make DAGs useful for computation and inference, not just for drawing pictures of causal stories.

Factorization According to the DAG

The joint probability distribution of all variables in a DAG can be broken into a product of simpler conditional probabilities. Each variable's distribution depends only on its parents (the nodes with arrows pointing into it):

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \text{Pa}(X_i))$$

Here, $\text{Pa}(X_i)$ denotes the set of parents of node X_i .

This is a big deal practically. Instead of specifying one enormous joint distribution over all your variables, you only need to specify each variable's relationship with its direct causes. That's far more manageable.

Conditional Independence

DAGs encode conditional independence relationships directly in their structure. Two variables X and Y are conditionally independent given a set Z if:

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

Two key versions of this show up in DAG theory:

- Local Markov property: Each node is independent of its non-descendants, given its parents.
- Global Markov property: Any two sets of nodes that are d-separated (see below) by a third set are conditionally independent.

These properties let you read independence relationships straight off the graph, which is essential for identifying confounders and mediators.

D-Separation Criterion

D-separation (directed separation) is the graphical rule for determining whether two sets of nodes are conditionally independent given a third set. This is one of the most important tools in the entire DAG framework.

Two sets of nodes X and Y are d-separated by a conditioning set Z if every path between them is blocked. A path is blocked when it contains:

- A non-collider (a node where arrows don't both point in, such as a chain $A \rightarrow M \rightarrow B$ or a fork $A \leftarrow C \rightarrow B$) that is in Z , or
- A collider (a node with two arrows pointing into it, like $A \rightarrow C \leftarrow B$) where C is not in Z and no descendant of C is in Z .

If X and Y are d-separated by Z , then $X \perp\!\!\!\perp Y \mid Z$ in any distribution compatible with the DAG.

The collider rule trips people up the most: conditioning on a collider (or its descendant) opens a path that was previously blocked. This is the mechanism behind collider bias, sometimes called "explaining away."

Causal Interpretation of DAGs

A DAG on its own is just a mathematical object encoding independence relationships. To use it for causal inference, you need to interpret the directed edges as genuine causal influences. That interpretation rests on three key assumptions.

Causal Markov Assumption

This assumption states that each variable is independent of all its non-descendants, given its direct causes (parents) in the DAG. In plain terms, once you know a variable's direct causes, knowing about other "upstream" or unrelated variables tells you nothing extra.

This is what lets the DAG's factorization correspond to the actual causal structure. If this assumption fails, the graph doesn't faithfully represent how the system generates data.

Causal Sufficiency Assumption

Causal sufficiency requires that all common causes of the observed variables are included in the DAG. There can be no unmeasured confounders lurking outside the graph that simultaneously affect two or more variables you've drawn.

This is a strong assumption and often the hardest to defend in practice. If it's violated, you can get spurious associations between variables that look causal but aren't, leading to biased effect estimates.

Causal Faithfulness Assumption

Faithfulness says that the only conditional independencies in the data are those implied by the DAG structure. There are no "accidental" independencies caused by causal effects that happen to perfectly cancel each other out.

For example, if A affects Y through two paths with effects of +3 and -3 that exactly cancel, faithfulness would be violated because A and Y would appear independent even though causal paths exist.

Faithfulness rules out these knife-edge cancellations, which makes it possible to learn causal structure from observed independence patterns.

DAGs vs. Other Graphical Models

Comparison with Bayesian Networks

Bayesian networks share the same mathematical structure as DAGs: directed edges, acyclicity, and the factorization property. The key difference is in interpretation. Bayesian networks represent probabilistic dependencies and are used for tasks like prediction and probabilistic reasoning. They don't necessarily carry causal meaning.

A DAG used for causal inference adds the causal assumptions described above, which lets you reason about interventions (what happens if you set a variable to a value) rather than just observations (what you'd predict given observed data).

Comparison with Structural Equation Models

Structural equation models (SEMs) pair a set of equations describing functional relationships between variables with a graphical representation (a path diagram). You can think of an SEM as a parametric version of a DAG: the DAG captures the qualitative causal structure (which variables affect which), while the SEM's equations specify the quantitative relationships (by how much).

DAGs and SEMs are closely related. A DAG often serves as the starting point for building an SEM, and the SEM fills in the numerical details. DAGs are more flexible in that they don't require you to commit to specific functional forms.

Constructing DAGs from Data

In many settings, you don't start with a known causal structure. Causal discovery methods try to learn the DAG from observed data by detecting statistical patterns of dependence and independence.

Constraint-Based Methods

Constraint-based methods use conditional independence tests to infer the graph structure. The most well-known is the PC algorithm. The process works roughly like this:

1. Start with a fully connected undirected graph (an edge between every pair of nodes).
2. Test pairs of variables for marginal independence. Remove edges where independence holds.
3. Test pairs for conditional independence given one variable, then two variables, and so on. Remove edges accordingly.
4. Orient edges by identifying collider patterns (v-structures) and applying orientation rules to avoid creating new colliders or cycles.

Common independence tests include the chi-square test (for categorical data) and partial correlation tests (for continuous/Gaussian data). The resulting graph represents the causal structure consistent with the observed conditional independencies.

Score-Based Methods

Score-based methods evaluate candidate DAG structures using a scoring function that balances model fit against complexity. The Greedy Equivalence Search (GES) algorithm is a prominent example.

- Common scoring functions include the Bayesian Information Criterion (BIC) and the Bayesian Dirichlet equivalent (BDe) score.
- The algorithm searches through the space of possible DAGs (or equivalence classes of DAGs) and selects the structure with the best score.
- Because the space of possible DAGs grows super-exponentially with the number of variables, these methods use greedy search heuristics rather than exhaustive enumeration.

Hybrid Methods

Hybrid methods combine constraint-based and score-based approaches to get the best of both worlds. A well-known example is the Max-Min Hill-Climbing (MMHC) algorithm:

1. Use a constraint-based method to identify a restricted set of candidate parents for each node (this narrows the search space).
2. Apply a score-based search over this restricted space to find the best DAG.

Hybrid methods tend to be more robust and computationally efficient than using either approach alone, especially in higher-dimensional settings.

Applications of DAGs in Causal Inference

Identifying Causal Effects

DAGs help you determine whether a causal effect can be estimated from observational data and, if so, which variables to adjust for. By examining the paths between treatment and outcome, you can apply identification strategies:

- Back-door adjustment: If you can block all back-door paths (non-causal paths) from treatment to outcome by conditioning on a set of variables, you can identify the causal effect. The back-door criterion tells you exactly which sets work.
- Front-door adjustment: When back-door adjustment isn't possible (e.g., due to unmeasured confounders), you may still identify the effect by exploiting a mediator that satisfies the front-door criterion.

The DAG makes these strategies explicit and verifiable, rather than relying on intuition about "what to control for."

Dealing with Confounding

A confounder is a variable that causally influences both the treatment and the outcome, creating a spurious association between them. On a DAG, confounding shows up as an open back-door path from treatment to outcome.

DAGs provide a systematic way to identify the minimal sufficient adjustment set: the smallest set of variables you need to condition on to block all confounding paths. This is more reliable than the common but error-prone strategy of "controlling for everything you can measure," which can actually introduce bias if you condition on colliders or mediators.

Mediation Analysis

Mediation analysis investigates how a treatment affects an outcome by examining intermediate variables (mediators). On a DAG, you can trace:

- The direct effect: the path from treatment to outcome that doesn't go through the mediator.
- The indirect effect: the path from treatment to outcome that passes through the mediator.

The DAG guides you in selecting which variables to condition on for estimating these effects and warns you about situations where mediation analysis may not be valid (e.g., when there's a confounder of the mediator-outcome relationship that is itself affected by treatment).

Limitations and Extensions of DAGs

Cyclic Graphs

Standard DAGs assume no feedback loops, but many real-world systems involve reciprocal causation (e.g., supply and demand, or anxiety and insomnia reinforcing each other). Directed cyclic graphs (DCGs) and equilibrium-based feedback models extend the framework to handle cycles, though they require different assumptions and more complex analysis techniques.

Time-Varying Treatments

Standard DAGs represent variables at a single time point. When treatments change over time and past outcomes influence future treatment decisions, you need models that capture this dynamic structure. Marginal structural models and g-methods extend the DAG framework to handle time-varying treatments and time-varying confounding, using techniques like inverse probability of treatment weighting.

Latent Variables and Selection Bias

DAGs typically assume all relevant variables are observed, but unmeasured (latent) variables are common in practice. Latent variable models and specialized causal discovery algorithms (like the FCI algorithm) extend the framework to account for hidden common causes.

Selection bias arises when the sample you observe isn't representative of the target population, often because some process determined who ended up in your dataset. DAGs can represent this by adding a selection node and modeling the selection mechanism explicitly. Addressing these issues requires additional assumptions (such as missing at random) and specialized methods like inverse probability weighting or multiple imputation.

9.2 d-separation and backdoor criterion

Graphical models, like Directed Acyclic Graphs (DAGs), are powerful tools for visualizing causal relationships. They help us understand how variables interact and influence each other in complex systems. DAGs use nodes to represent variables and directed edges to show causal effects.

D-separation and the backdoor criterion are key concepts in causal inference using DAGs. They allow us to determine conditional independence relationships and identify sets of variables that, when adjusted for, enable unbiased estimation of causal effects from observational data.

Graphical models of causality

- Graphical models provide a visual representation of the causal relationships between variables in a system
- Enable reasoning about independence relationships and identifiability of causal effects from observational data
- Fundamental tool in causal inference for representing expert knowledge and assumptions about a domain

Directed acyclic graphs (DAGs)

- DAGs are a specific type of graphical model used to represent causal relationships
- Consist of nodes representing variables and directed edges representing direct causal effects
- Must not contain any directed cycles, meaning no variable can be a cause of itself (directly or indirectly)
- Encode independence assumptions and enable application of graphical criteria for causal inference

Nodes, edges, and paths

- Nodes in a DAG represent variables, which can be observed or unobserved (latent)

- Directed edges between nodes represent direct causal effects, with the arrow pointing from cause to effect
- Paths are sequences of edges connecting nodes, which can be either directed (causal) or undirected (associational)
- Directed paths follow the direction of the arrows and represent causal relationships
- Undirected paths do not necessarily follow the direction of the arrows and represent associational relationships

Causal Markov assumption

- The causal Markov assumption is a fundamental assumption in causal inference using DAGs
- States that a variable is independent of its non-descendants given its direct causes (parents) in the graph
- Implies that the joint distribution of the variables factorizes according to the DAG structure
- Allows for the derivation of testable implications and identification of causal effects from observational data

d-separation

- d-separation is a graphical criterion for determining conditional independence relationships in a DAG
- Provides a way to read off independence statements directly from the graph structure
- Fundamental concept for assessing identifiability of causal effects and selecting adjustment sets

Conditional independence in DAGs

- Two variables are conditionally independent given a set of variables Z if they are independent in every subpopulation defined by Z
- In a DAG, conditional independence relationships are implied by the graph structure and can be determined using d-separation
- If two variables are d-separated by a set Z , then they are conditionally independent given Z in any probability distribution that factorizes according to the DAG

Blocked vs unblocked paths

- A path between two variables in a DAG is blocked by a set of variables Z if it contains a collider that is not in Z or has a descendant in Z , or a non-collider that is in Z
- Colliders are variables with two incoming arrows on the path (common effects)
- Non-colliders are variables with at least one outgoing arrow on the path (common causes or mediators)
- If all paths between two variables are blocked by Z , then the variables are d-separated and conditionally independent given Z
- If there exists an unblocked path between two variables given Z , then they are d-connected and potentially dependent given Z

Three basic configurations

- The three basic configurations in a DAG that determine d-separation relationships are chains, forks, and colliders
- Chains: $X \rightarrow M \rightarrow Y$, where M is a non-collider. Conditioning on M blocks the path and makes X and Y independent
- Forks: $X \leftarrow M \rightarrow Y$, where M is a non-collider. Conditioning on M blocks the path and makes X and Y independent
- Colliders: $X \rightarrow M \leftarrow Y$, where M is a collider. Conditioning on M (or a descendant of M) unblocks the path and makes X and Y dependent

Formal definition and rules

- Two variables X and Y are d-separated by a set Z in a DAG G if and only if every path between X and Y is blocked by Z
- The rules for d-separation can be summarized as follows: 1. A path is blocked if it contains a non-collider that is in Z 2. A path is blocked if it contains a collider that is not in Z and has no descendant in Z 3. A path is unblocked if it contains a collider that is in Z or has a descendant in Z
- These rules can be applied systematically to determine all d-separation relationships in a DAG

Backdoor criterion

- The backdoor criterion is a graphical criterion for determining whether a set of variables is sufficient for adjusting for confounding in a causal effect estimation
- Provides a way to identify admissible adjustment sets from the DAG structure alone, without requiring knowledge of the underlying probability distribution
- Fundamental tool for causal effect identification and estimation in observational studies

Confounding and backdoor paths

- Confounding occurs when there is a common cause of the treatment and outcome variables, leading to spurious associations
- In a DAG, confounding is represented by the presence of backdoor paths between the treatment and outcome
- Backdoor paths are paths that start with an arrow into the treatment variable and end with an arrow into the outcome variable
- Adjusting for a sufficient set of variables that blocks all backdoor paths can eliminate confounding and enable unbiased estimation of the causal effect

Adjusting for confounders

- To adjust for confounding, we need to condition on a set of variables that blocks all backdoor paths between the treatment and outcome
- Conditioning on a variable can be done through stratification, matching, or regression adjustment

- The goal is to create subpopulations within which the treatment assignment is unconfounded, and then aggregate the estimates across subpopulations
- Proper adjustment for confounders requires careful selection of variables based on the causal DAG and the backdoor criterion

Formal definition of criterion

- A set of variables Z satisfies the backdoor criterion relative to a treatment variable X and an outcome variable Y in a DAG G if: 1. No variable in Z is a descendant of X 2. Z blocks all backdoor paths from X to Y
- If a set Z satisfies the backdoor criterion, then the causal effect of X on Y can be identified by adjusting for Z , i.e., $P(Y|do(X)) = \sum_z P(Y|X,Z)P(Z)$

Relationship to d-separation

- The backdoor criterion is closely related to the concept of d-separation in DAGs
- If a set Z satisfies the backdoor criterion relative to X and Y , then X and Y are d-separated by Z in the graph $G_{\setminus X}$, which is the graph obtained by removing all arrows emanating from X
- This connection allows for the use of d-separation algorithms to find admissible adjustment sets and assess the identifiability of causal effects

Applications of d-separation and backdoor

- d-separation and the backdoor criterion are fundamental tools in causal inference with wide-ranging applications
- They provide a principled way to reason about conditional independence relationships and identifiability of causal effects from observational data
- Enable the selection of appropriate adjustment sets for confounding control and the assessment of the feasibility of causal effect estimation

Determining identifiability of causal effects

- The backdoor criterion can be used to determine whether a causal effect is identifiable from observational data given a causal DAG
- If there exists a set of variables that satisfies the backdoor criterion, then the causal effect can be identified by adjusting for that set
- If no such set exists, then the causal effect may not be identifiable without additional assumptions or data (e.g., instrumental variables)

Selecting variables for adjustment

- d-separation and the backdoor criterion provide guidance on which variables should be included in an adjustment set for confounding control
- They help avoid the inclusion of colliders or mediators, which can introduce bias or block the effect of interest
- By systematically applying these criteria, researchers can select adjustment sets that are sufficient for eliminating confounding while avoiding unnecessary or harmful adjustments

Examples in practice

- Epidemiological studies often use DAGs and the backdoor criterion to identify potential confounders and select adjustment variables (e.g., in studies of the effect of smoking on lung cancer, adjusting for age and sex)
- In social sciences, researchers use these tools to assess the feasibility of estimating causal effects from observational data and guide the design of studies (e.g., in studies of the effect of education on earnings, adjusting for family background and ability)

Limitations and assumptions

- The application of d-separation and the backdoor criterion relies on the correctness of the assumed causal DAG
- If important variables are omitted or the causal relationships are misspecified, the resulting conclusions may be invalid
- These tools also assume that there is no unmeasured confounding, which may not always hold in practice
- Sensitivity analyses and triangulation of evidence from multiple sources can help assess the robustness of findings to potential violations of assumptions

Advanced topics and extensions

- The concepts of d-separation and the backdoor criterion have been extended and generalized to handle more complex causal structures and identification scenarios
- These advances have expanded the range of causal questions that can be addressed using graphical models and have deepened our understanding of the foundations of causal inference

Frontdoor criterion and adjustment

- The frontdoor criterion is another graphical criterion for identifying causal effects in the presence of unmeasured confounding
- It applies when there is no set of variables that satisfies the backdoor criterion, but there exists a mediator variable that is not affected by the unmeasured confounders
- Adjusting for the frontdoor variable can enable the identification of the causal effect by decomposing it into the product of two identifiable components

m-separation and MAGs

- Maximal ancestral graphs (MAGs) are a generalization of DAGs that allow for the representation of unmeasured confounding and selection bias
- m-separation is a generalization of d-separation that applies to MAGs and provides a graphical criterion for conditional independence in the presence of unmeasured variables
- These extensions enable the identification of causal effects in a wider range of settings and provide a foundation for causal discovery algorithms

Causal discovery algorithms

- Causal discovery algorithms aim to learn the causal structure of a system from observational data
- They use the principles of d-separation and m-separation to infer the presence or absence of causal relationships between variables
- Examples include the PC algorithm, which uses conditional independence tests to construct a DAG, and the FCI algorithm, which learns a MAG in the presence of unmeasured confounding
- These algorithms provide a data-driven approach to causal structure learning and can complement expert knowledge in the specification of causal models

Challenges with unmeasured confounding

- Unmeasured confounding is a major challenge in causal inference from observational data
- When important confounders are not measured or included in the analysis, the resulting estimates of causal effects may be biased
- Sensitivity analysis techniques, such as the E-value and bias formulas, can help assess the robustness of findings to potential unmeasured confounding
- Instrumental variable methods and other approaches that leverage natural experiments or quasi-random variation can also help address unmeasured confounding in some settings

9.3 Structural causal models (SCMs)

Structural causal models (SCMs)

Structural causal models (SCMs) combine graphs and equations to represent how variables influence each other in a system. They let you move beyond correlation to estimate causal effects, evaluate interventions, and reason about counterfactuals. An SCM encodes your assumptions about causal structure, then gives you the mathematical machinery to work with those assumptions rigorously.

Definition of SCMs

An SCM is a mathematical model describing the causal relationships between variables in a system. Formally, it consists of three pieces:

- A set of variables (both observed and unobserved)
- A directed acyclic graph (DAG) representing the causal structure, where edges indicate direct causal influences
- A set of structural equations specifying the functional relationship between each variable and its direct causes

Together, these components also imply a joint probability distribution over the variables. The DAG tells you the qualitative story (what causes what), and the structural equations tell you the quantitative story (how much, and in what way).

Endogenous vs. exogenous variables

SCMs distinguish between two types of variables:

- Endogenous variables are determined by other variables within the model. They have at least one incoming edge in the DAG.
- Exogenous variables are determined by factors outside the model. No other variable in the SCM causes them, so they appear as root nodes (no incoming edges) in the DAG.

Exogenous variables are typically assumed to be independently distributed. They represent the "inputs" to the system that the model takes as given rather than explaining.

Structural equations

Structural equations specify how each endogenous variable is generated from its direct causes plus an error term. The error term captures unobserved factors or randomness.

For example, if Y is caused by X :

$$Y = f(X, \varepsilon_Y)$$

Here f is some function, X is the direct cause, and ε_Y is the error term. A critical point: structural equations are not the same as regression equations. A regression equation describes a statistical association, while a structural equation claims a causal mechanism. Changing X in a structural equation changes Y ; changing X in a regression equation doesn't necessarily mean anything causal.

Directed acyclic graphs (DAGs)

DAGs are the graphical backbone of an SCM:

- Nodes represent variables
- Directed edges (arrows) represent direct causal influences
- The absence of an edge between two nodes means there is no direct causal relationship between them

The "acyclic" part means there are no directed cycles. You can't follow the arrows from a variable and loop back to it. This rules out feedback loops (though extensions exist for cyclic systems, covered later).

Causal Markov condition

The causal Markov condition is a core assumption linking the DAG to the probability distribution. It states:

A variable is independent of all its non-descendants, given its parents (direct causes) in the DAG.

This assumption allows you to factorize the joint distribution according to the DAG structure. Each variable's conditional distribution depends only on its parents.

For example, in the chain $X \rightarrow Y \rightarrow Z$, the causal Markov condition tells you that X and Z are independent given Y . Once you know Y , learning X gives you no additional information about Z .

Causal sufficiency

Causal sufficiency assumes that all common causes of the observed variables are included in the model. In other words, there are no unmeasured confounders affecting multiple observed variables.

This is a strong assumption. In practice, it often doesn't hold perfectly, and violations can lead to biased causal effect estimates. When you suspect unmeasured confounding, you'll need techniques like sensitivity analysis or instrumental variables (discussed below).

Causal faithfulness

Causal faithfulness assumes that the only independencies in the data are those implied by the DAG. There are no "accidental" independencies caused by parameters perfectly canceling each other out.

For example, if two causal paths from X to Y have effects that exactly cancel (one positive, one negative, same magnitude), X and Y would appear independent even though causal paths exist. Faithfulness rules this out. Such exact cancellations are considered rare in practice, but they can occur.

Representing interventions with SCMs

Interventions are actions that force variables to take specific values, overriding their natural causes. SCMs give you a precise way to represent and reason about interventions, which is what separates causal reasoning from purely statistical reasoning.

Interventional distributions

An interventional distribution is the probability distribution of variables after an intervention. It's written using the *do*-operator:

$$P(Y \mid do(X = x))$$

This reads as "the probability of Y when we set X to value x ." This is fundamentally different from the conditional distribution $P(Y \mid X = x)$, which is what you'd observe by filtering data. The *do*-operator represents actively manipulating X , not passively observing it.

Graph mutilation

To derive an interventional distribution, you modify the DAG through graph mutilation:

1. Identify the variable X being intervened on
2. Remove all incoming edges to X (since the intervention overrides X 's natural causes)
3. Set X to the intervention value
4. The resulting "mutilated graph" represents the causal structure under the intervention

The rest of the DAG stays the same. Variables downstream of X are still affected by X , but X itself is no longer influenced by its former parents.

Truncated factorization

Truncated factorization is the mathematical counterpart of graph mutilation. Under the original DAG, the joint distribution factorizes as:

$$P(V_1, V_2, \dots, V_n) = \prod_{i=1}^n P(V_i \mid PA_i)$$

where PA_i are the parents of V_i . When you intervene on $X = x$, you drop the factor for X from the product (since X is now fixed, not generated by its parents) and substitute $X = x$ everywhere else. The result is the interventional distribution over the remaining variables.

Identification of causal effects

Identification asks: can you compute a causal effect from observational data alone, given your assumed DAG? If yes, the effect is "identified." If not, you need additional data or assumptions. SCMs provide graphical criteria to answer this question.

Back-door criterion

The back-door criterion is the most commonly used identification tool. A set of variables Z satisfies the back-door criterion relative to (X, Y) if:

1. Z blocks all back-door paths from X to Y (paths that enter X through an incoming arrow)
2. No variable in Z is a descendant of X

If such a Z exists, you can identify the causal effect by adjusting for Z :

$$P(Y \mid do(X = x)) = \sum_z P(Y \mid X = x, Z = z) \cdot P(Z = z)$$

For example, in the DAG $X \leftarrow Z \rightarrow Y$, the confounder Z creates a back-door path. Conditioning on Z blocks it, satisfying the criterion.

Front-door criterion

The front-door criterion applies when no set of variables satisfies the back-door criterion (e.g., because the confounder is unmeasured). A set Z satisfies the front-door criterion relative to (X, Y) if:

1. Z intercepts all directed paths from X to Y
2. There are no unblocked back-door paths from X to Z
3. All back-door paths from Z to Y are blocked by X

A classic example: $X \rightarrow Z \rightarrow Y$ with an unmeasured confounder U affecting both X and Y . You can't adjust for U directly, but Z (the mediator) satisfies the front-door criterion, letting you identify the causal effect through a two-step adjustment.

Instrumental variables

Instrumental variables (IVs) provide another identification strategy when unmeasured confounders are present. A variable Z is a valid instrument for the effect of X on Y if:

1. Z is associated with X (relevance)
2. Z does not directly affect Y except through X (exclusion restriction)
3. Z is independent of all confounders of the X - Y relationship (independence)

A well-known example: estimating the effect of education on income. A person's quarter of birth affects years of education (through compulsory schooling laws) but has no direct effect on income, making it a candidate instrument.

Mediation analysis

Mediation analysis decomposes the total causal effect of X on Y into:

- Direct effect: the effect of X on Y not passing through a mediator M

- Indirect effect: the effect of X on Y that operates through M

For example, a drug (X) might lower blood pressure (Y) both directly and indirectly by reducing heart rate (M). SCMs let you define and estimate these effects precisely, though mediation analysis requires assumptions about the absence of unmeasured confounders for both the $X \rightarrow Y$ and $M \rightarrow Y$ relationships.

Counterfactuals in SCMs

Counterfactuals ask "what would have happened if things had been different?" SCMs handle counterfactuals by using the structural equations to simulate alternative scenarios while holding the exogenous variables (the background context) fixed.

Potential outcomes framework

The potential outcomes framework (also called the Rubin Causal Model) defines causal effects in terms of hypothetical outcomes. For a binary treatment:

- $Y(1)$: the outcome if the unit receives treatment
- $Y(0)$: the outcome if the unit does not receive treatment

The individual causal effect is $Y(1) - Y(0)$. The fundamental problem is that you only observe one of these for each unit. SCMs and potential outcomes are complementary frameworks: SCMs provide the structural machinery, while potential outcomes provide a clean notation for defining effects.

Counterfactual queries

Counterfactual queries ask about outcomes under hypothetical interventions. In an SCM, you answer them by:

1. Abduction: Use the observed data to infer the values of the exogenous variables (error terms)
2. Action: Modify the structural equations to reflect the hypothetical intervention
3. Prediction: Use the modified equations with the inferred exogenous values to compute the counterfactual outcome

For example, "What would this patient's blood pressure have been without the drug?" You first use the patient's actual data to pin down their individual error terms, then re-run the model with $X = 0$ (no drug) to predict the counterfactual blood pressure.

Note that this is different from the interventional query $P(Y \mid do(X = 0))$, which asks about a population-level effect. The counterfactual is specific to a particular individual with known characteristics.

Twin networks

Twin networks are a graphical tool for computing counterfactual quantities. They work by creating two copies of the SCM:

- The factual network represents what actually happened
- The counterfactual network represents the hypothetical scenario

The two networks share the same exogenous variables, which ties the individual's background characteristics together across both worlds. You can then read off counterfactual quantities by comparing outcomes between the two networks.

For example, in a twin network for a drug study, the factual side shows the patient taking the drug and the observed outcome, while the counterfactual side shows the same patient not taking the drug, with the same exogenous factors.

Learning SCMs from data

Learning an SCM from data involves two tasks: discovering the causal structure (the DAG) and estimating the parameters of the structural equations. Both are challenging due to limited data, latent variables, and the fact that multiple DAGs can produce the same observed distribution.

Causal structure learning

Causal structure learning aims to infer the DAG from observational data. Methods fall into three categories:

- Constraint-based methods (e.g., PC algorithm): use conditional independence tests to determine which edges belong in the DAG
- Score-based methods (e.g., GES algorithm): search over possible DAGs to find the one that best fits the data according to a scoring function
- Hybrid methods: combine both approaches

A key limitation: observational data alone can typically only identify the DAG up to its Markov equivalence class (a set of DAGs that encode the same conditional independencies). You may need additional assumptions or experimental data to distinguish between equivalent structures.

Constraint-based methods

Constraint-based methods rely on the causal Markov condition and faithfulness to infer the DAG. The PC algorithm is the most well-known:

1. Start with a fully connected undirected graph
2. For each pair of adjacent nodes, test whether they are conditionally independent given some subset of other variables
3. Remove edges where conditional independence is found
4. Orient edges using rules based on v-structures and acyclicity constraints

These methods are computationally efficient but sensitive to errors in independence tests, especially with small samples or violations of faithfulness.

Score-based methods

Score-based methods search for the DAG that optimizes a scoring function balancing data fit and model complexity. Common scores include the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC).

The Greedy Equivalence Search (GES) algorithm:

1. Start with an empty graph (no edges)
2. Forward phase: iteratively add edges that most improve the score
3. Backward phase: iteratively remove edges that improve the score
4. Return the highest-scoring DAG

Score-based methods are less sensitive to individual test errors than constraint-based methods, but the search space grows super-exponentially with the number of variables, making them computationally expensive for large systems.

Hybrid methods

Hybrid methods combine constraint-based and score-based approaches to get the best of both worlds. A typical strategy:

1. Use constraint-based methods to prune the search space (eliminate edges that are clearly absent)
2. Use score-based methods to search over the remaining candidate structures

The Max-Min Hill-Climbing (MMHC) algorithm is a prominent example. It uses the MMPC algorithm to identify candidate parent-child relationships, then applies hill-climbing search to find the best-scoring DAG within that restricted space. This achieves a balance between computational efficiency and robustness.

Parameter estimation

Once you have the DAG, you need to estimate the parameters of the structural equations. The two main approaches:

- Maximum likelihood estimation (MLE): find parameter values that maximize the probability of observing the data, given the DAG structure
- Bayesian methods: specify prior distributions over parameters and update them with the data to get posterior distributions

Both approaches require assumptions about the functional form of the structural equations (e.g., linear vs. nonlinear) and the distribution of error terms (e.g., Gaussian vs. non-Gaussian). Getting these assumptions wrong can lead to poor estimates.

Applications of SCMs

SCMs are used across many fields, including epidemiology, economics, social sciences, and AI. They provide a principled framework for several practical tasks.

Causal effect estimation

SCMs let you estimate causal effects from observational data by leveraging the assumed causal structure. For example, using electronic health records, you could estimate the causal effect of a medication on patient outcomes by identifying the right adjustment set from the DAG and applying the back-door adjustment formula.

Policy evaluation

By simulating interventions on an SCM, you can predict the effects of policies before implementing them. For instance, you could model the causal relationships between taxation, income inequality, and economic growth, then simulate different tax policies to compare their predicted outcomes.

Transportability

Transportability addresses whether causal effects estimated in one population generalize to another. SCMs provide formal tools for assessing this. For example, if a clinical trial estimates a drug's effect in one demographic group, transportability analysis can determine under what conditions that estimate applies to a different population with different demographics and comorbidities.

Causal discovery

SCMs also serve as the target for causal discovery: inferring the causal structure itself from observational data. This is valuable for generating hypotheses. For example, applying causal discovery algorithms to cohort study data might reveal previously unknown causal factors for a disease, guiding future experimental research.

Limitations and extensions of SCMs

SCMs are powerful but come with assumptions that don't always hold. Understanding these limitations helps you apply SCMs responsibly.

Latent confounding

Standard SCMs assume causal sufficiency (all common causes are measured). In practice, unmeasured confounders are common and can bias causal effect estimates. For example, in studying smoking and lung cancer, unmeasured genetic factors might influence both.

Extensions to handle this include:

- Latent variable models that explicitly represent unmeasured confounders
- Sensitivity analysis that quantifies how robust your conclusions are to potential unmeasured confounding
- Bounds on causal effects when point identification isn't possible

Cyclic causal models

Standard SCMs require acyclicity, but many real-world systems involve feedback loops. Job satisfaction affects job performance, which in turn affects satisfaction. Cyclic causal models extend the SCM framework to handle such cases, though they require different assumptions and estimation techniques (e.g., equilibrium conditions or dynamic models).

Time-varying treatments

Standard SCMs typically model treatments as fixed at a single point in time. Many real-world treatments change over time (e.g., medication dosages adjusted based on patient response). Extensions like marginal structural models and structural nested models handle time-varying treatments by modeling the causal effects of treatment sequences rather than single treatment assignments. These methods use techniques like inverse probability weighting to account for time-varying confounding.

9.4 Interventions and do-calculus

Defining interventions

An intervention means you physically set a variable to a particular value, rather than just watching what value it naturally takes. This is the core move that separates causal inference from standard statistics: instead of asking "What is Y when we observe $X = x$?" you ask "What would Y be if we forced X to equal x ?"

That distinction matters because observational associations can be driven by confounders, selection effects, or reverse causation. Interventions cut through all of that by breaking the variable free from its usual causes.

Interventional distributions

An interventional distribution, written $P(Y \mid \text{do}(X = x))$, describes the probability distribution of Y in a world where X has been set to x by external action. Compare this to the observational conditional $P(Y \mid X = x)$, which reflects the distribution of Y among units that happen to have $X = x$.

These two quantities can differ dramatically. For example, $P(\text{recovery} \mid \text{drug} = 1)$ in hospital records might be low because sicker patients receive the drug more often. But $P(\text{recovery} \mid \text{do}(\text{drug} = 1))$ captures what would happen if you randomly assigned the drug, removing that selection effect.

Causal vs observational

- A causal relationship means changing X directly changes Y . An observational association means X and Y tend to co-occur, but manipulating X might not affect Y at all.
- Confounders are the classic reason these diverge: ice cream sales and drowning rates are correlated (both driven by hot weather), but banning ice cream won't prevent drownings.
- The whole point of the do-operator is to formalize this distinction so you can reason precisely about when observational data can tell you about causal effects, and when it can't.

Ideal randomized experiments

Randomized experiments are the gold standard for estimating causal effects because randomization breaks the link between the treatment and any confounders.

- Randomly assigning subjects to treatment vs. control ensures that, in expectation, the groups are identical on all pre-treatment characteristics.
- This means $P(Y \mid \text{do}(X = x)) = P(Y \mid X = x)$ in the experimental data, so the observational conditional is the causal quantity.
- Classic examples: clinical trials for drug efficacy, A/B tests for website design changes.

The challenge is that randomized experiments aren't always feasible. That's where the rest of this unit comes in.

Representing interventions

To work with interventions formally, you need notation and graphical tools that make causal assumptions explicit and manipulable.

Intervention notation

The do-operator, written $do(X = x)$, represents an external intervention that forces X to take value x . The key expressions to know:

- $P(Y \mid do(X = x))$ is the interventional distribution of Y when X is set to x .
- $P(Y \mid X = x)$ is the ordinary observational conditional.
- The central question of causal identification: Can we express $P(Y \mid do(X = x))$ purely in terms of observational distributions?

Graphical representation

Causal graphs (specifically DAGs) encode your causal assumptions visually:

- Nodes represent variables.
- Directed edges (arrows) represent direct causal effects.
- The graph tells you which variables cause which, and which paths connect treatment to outcome.

When you intervene on X , you represent this by deleting all incoming edges to X in the graph. This reflects the fact that the intervention overrides whatever would normally determine X .

Manipulated graphs

A manipulated graph (also called a mutilated graph) is the DAG you get after deleting incoming edges to the intervened variable. Formally, if you perform $do(X = x)$, the manipulated graph $G_{\bar{X}}$ is the original graph with all arrows pointing into X removed.

Why this matters: the manipulated graph encodes the independence structure of the post-intervention world. You use it to read off which variables are independent of which under the intervention, and that's what drives identification criteria like the back-door and front-door criteria.

Identifiability under interventions

Identifiability asks: given the causal graph and observational data, can you compute the interventional distribution $P(Y \mid do(X = x))$? If yes, the causal effect is identifiable. If not, you need additional data or stronger assumptions.

Back-door criterion

The back-door criterion gives you a sufficient condition for identifying a causal effect by adjusting for confounders.

A set of variables Z satisfies the back-door criterion relative to (X, Y) if:

1. No variable in Z is a descendant of X .
2. Z blocks every path between X and Y that has an arrow into X (these are the "back-door paths").

When Z satisfies the criterion, the causal effect is given by the back-door adjustment formula:

$$P(Y \mid do(X = x)) = \sum_z P(Y \mid X = x, Z = z) P(Z = z)$$

Example: to estimate the causal effect of smoking on cancer, you might adjust for age and socioeconomic status if those are confounders with no descendant relationship to smoking.

Front-door criterion

The front-door criterion handles cases where there are unmeasured confounders between X and Y , but there exists a mediator M such that:

1. X blocks all back-door paths from X to M .
2. M captures the entire causal effect of X on Y (all directed paths from X to Y go through M).
3. There are no unblocked back-door paths from M to Y after conditioning on X .

When these conditions hold, the causal effect is:

$$P(Y \mid do(X = x)) = \sum_m P(M = m \mid X = x) \sum_{x'} P(Y \mid X = x', M = m) P(X = x')$$

This is powerful because it lets you identify causal effects even when you can't measure the confounder directly.

Instrumental variables

An instrumental variable (IV) Z affects the treatment X but has no direct effect on the outcome Y , and is independent of unmeasured confounders between X and Y .

- The classic example: using proximity to a college as an instrument for years of education when estimating education's effect on income. Living near a college nudges people toward more education, but doesn't directly affect income through other channels (that's the assumption, at least).
- IVs allow causal effect estimation under unmeasured confounding, but they typically identify a local average treatment effect (LATE) rather than the average treatment effect for the whole population.
- The key assumptions (relevance, exclusion restriction, independence) must hold for IV estimates to be valid, and they're not testable from data alone.

Intro to do-calculus

Do-calculus, developed by Judea Pearl, is a complete set of rules for transforming expressions involving the do-operator. Its purpose is to convert interventional quantities into expressions that involve only observational distributions, which you can estimate from data.

Formal definition

Do-calculus consists of three inference rules that let you add, remove, or exchange observations and interventions within probability expressions. Each rule has a graphical condition that must be satisfied in a specific manipulated version of the DAG. Together, these rules are complete: if a causal effect is identifiable from the graph, some sequence of these three rules will derive the identifying formula.

Differences from standard probability

Standard probability gives you tools like Bayes' rule and the chain rule for manipulating conditional distributions. But it has no way to distinguish $P(Y \mid X = x)$ from $P(Y \mid do(X = x))$, because standard probability doesn't encode causal direction.

Do-calculus adds this layer. It operates on expressions that mix observational conditioning and do-operators, and its rules are justified by the causal structure (the DAG), not just by probability axioms. This is what makes it possible to answer causal questions from observational data.

Three rules of do-calculus

Each rule applies when a specific conditional independence holds in a specific manipulated graph. Let X , Y , Z , and W be disjoint sets of variables in a DAG G :

Rule 1 (Insertion/deletion of observations):

$$P(Y \mid do(X), Z, W) = P(Y \mid do(X), W)$$

if $Y \perp Z \mid X, W$ in the graph $G_{\bar{X}}$ (the graph with incoming edges to X removed).

This says you can ignore an observed variable Z if it's independent of Y given X and W in the manipulated graph.

Rule 2 (Action/observation exchange):

$$P(Y \mid do(X), do(Z), W) = P(Y \mid do(X), Z, W)$$

if $Y \perp Z \mid X, W$ in the graph $G_{\bar{X}, \underline{Z}}$ (incoming edges to X removed, outgoing edges from Z removed).

This lets you replace an intervention on Z with an observation of Z under the right conditions.

Rule 3 (Insertion/deletion of actions):

$$P(Y \mid do(X), do(Z), W) = P(Y \mid do(X), W)$$

if $Y \perp Z \mid X, W$ in the graph $G_{\bar{X}, \overline{Z(S)}}$, where $Z(S)$ is the set of nodes in Z that are not ancestors of any node in W in $G_{\bar{X}}$.

This lets you drop an intervention on Z entirely when the graphical condition is met.

Applying do-calculus

The practical use of do-calculus is to take a target causal quantity like $P(Y \mid do(X))$ and, through repeated application of the three rules, reduce it to an expression involving only observational probabilities.

Identifying causal effects

The typical workflow:

1. Write down the target interventional distribution, e.g., $P(Y \mid do(X = x))$.
2. Examine the DAG to determine which rules apply.
3. Apply rules to progressively eliminate do-operators.
4. If you can reduce the expression to one with no do-operators, the effect is identified, and you have a formula you can estimate from data.

Both the back-door and front-door adjustment formulas can be derived as special cases of do-calculus. The power of do-calculus is that it also handles cases where neither criterion directly applies.

Proving non-identifiability

Sometimes no sequence of do-calculus rules can eliminate all do-operators. In that case, the causal effect is not identifiable from observational data alone given the assumed graph.

- This is a genuine impossibility result, not just a failure to find the right trick. Completeness of do-calculus guarantees that if the rules can't get you there, no other method based on the same assumptions can either.
- Non-identifiability tells you that you need either additional data (e.g., from experiments or new measured variables) or stronger assumptions to estimate the effect.

Worked examples

A concrete example helps clarify the process. Consider the classic front-door model:

- U is an unmeasured confounder affecting both X and Y .
- $X \rightarrow M \rightarrow Y$, with $U \rightarrow X$ and $U \rightarrow Y$.

To identify $P(Y \mid do(X))$:

1. Apply Rule 2 to replace $do(X)$ with conditioning on X for the $X \rightarrow M$ relationship (since U doesn't confound $X \rightarrow M$ once you account for the graph structure).
2. Apply Rule 2 and Rule 3 to handle the $M \rightarrow Y$ relationship, adjusting for X to block the back-door path through U .
3. Combine the pieces to arrive at the front-door formula.

Working through examples like this is the best way to build fluency with do-calculus. Try deriving the back-door adjustment formula from the three rules as practice.

Interventions with SCMs

Structural Causal Models (SCMs) give you a complete formal package: a DAG for the qualitative causal structure, plus a set of structural equations that specify the quantitative relationships.

Modeling interventions

In an SCM, each variable X_i has a structural equation:

$$X_i = f_i(\text{pa}_i, U_i)$$

where pa_i are the parents of X_i in the DAG and U_i is an exogenous noise term.

To model an intervention $do(X_j = x)$, you replace the equation for X_j with the constant $X_j = x$ and leave all other equations unchanged. This is the formal counterpart of deleting incoming edges in the graph.

Interventions can also be stochastic: instead of setting X_j to a fixed value, you draw it from some specified distribution, replacing the original equation with a new random mechanism.

Truncated factorization

In an SCM without interventions, the joint distribution factors as:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i \mid \text{pa}_i)$$

Under an intervention $do(X_j = x)$, the factor for X_j is removed (since X_j is now fixed), giving the truncated factorization:

$$P(x_1, \dots, x_n \mid do(X_j = x)) = \prod_{i \neq j} P(x_i \mid \text{pa}_i) \big|_{x_j = x}$$

This formula is the algebraic expression of graph mutilation. It directly connects the graphical operation (delete incoming edges) to a probability computation.

Causal effect estimation

With the truncated factorization in hand, you can compute any interventional quantity by summing or integrating over the non-intervened variables.

Common estimands:

- Average Treatment Effect (ATE): $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$
- Effect of Treatment on the Treated (ETT): $E[Y_{X=1} - Y_{X=0} \mid X = 1]$

SCMs also support counterfactual reasoning (e.g., "What would have happened to this specific individual under a different treatment?"), which goes beyond what do-calculus alone can handle. Counterfactuals require the full SCM, including the noise terms U_i , not just the graph.

Practical considerations

The formal machinery of do-calculus and SCMs is powerful, but applying it in practice requires careful attention to real-world constraints.

Feasibility of interventions

Many interventions that are well-defined mathematically are impossible or impractical to carry out:

- You can write $do(\text{gender} = \text{male})$, but you can't actually randomize gender.
- Cost, logistics, and time constraints may rule out large-scale experiments.
- When direct intervention isn't feasible, researchers turn to natural experiments (where some external event approximates random assignment) or quasi-experimental designs (difference-in-differences, regression discontinuity).

Ethical implications

Even when an intervention is feasible, it may not be ethical:

- Withholding a potentially beneficial treatment from a control group raises concerns.
- Interventions on vulnerable populations require extra scrutiny.
- Institutional review boards (IRBs) enforce standards around informed consent, minimizing harm, and equitable selection of subjects.

These constraints are a major reason why observational causal inference methods exist in the first place.

Limitations of do-calculus

Do-calculus is complete for the class of problems it addresses, but it rests on assumptions that may not hold:

- Causal Markov condition: Each variable is independent of its non-descendants given its parents. Violations can occur with feedback loops or aggregated variables.
- Faithfulness: All observed independencies are consequences of the graph structure, not coincidental cancellations. This can fail in finely tuned systems.
- Correct graph specification: Do-calculus takes the DAG as given. If your graph is wrong (missing edges, wrong directions), your conclusions will be wrong too.

When you're uncertain about these assumptions, complement do-calculus with sensitivity analysis (how much would an unmeasured confounder need to shift results?) and partial identification (deriving bounds on the causal effect rather than a point estimate).

Unit 10 – Causal Discovery & Learning Algorithms

10.1 Constraint-based algorithms

Constraint-based algorithms are a key tool in causal inference, using conditional independence tests to uncover causal relationships in data. These methods aim to recover the Markov equivalence class of the true causal graph, revealing the underlying structure of variables and their connections.

These algorithms rely on key assumptions like the Causal Markov condition and faithfulness. They work by first discovering the skeleton of the graph, then orienting edges to identify v-structures. Popular examples include the PC, FCI, and RFCI algorithms, each with unique strengths in handling different scenarios.

Constraint-based algorithms overview

- Constraint-based algorithms are a class of methods used in causal inference to learn the structure of causal graphs from observational data
- These algorithms rely on conditional independence tests to identify the presence or absence of edges between variables in a causal graph
- Constraint-based algorithms aim to recover the Markov equivalence class of the true causal graph, which represents a set of graphs that encode the same conditional independence relationships

Assumptions of constraint-based algorithms

Causal Markov condition

- States that a variable is independent of its non-descendants given its direct causes (parents) in the causal graph
- Implies that the joint probability distribution of the variables can be factorized according to the causal graph structure
- Allows for the identification of conditional independence relationships from the causal graph

Causal faithfulness condition

- Assumes that the conditional independence relationships in the joint probability distribution are exactly those encoded by the causal graph
- Ensures that there are no additional independence relationships beyond those implied by the causal graph structure
- Violations of faithfulness can occur due to deterministic relationships or perfect cancellations of effects

Causal sufficiency assumption

- Requires that all common causes of the observed variables are included in the dataset
- Implies that there are no unmeasured confounders influencing the observed variables
- Violation of causal sufficiency can lead to incorrect inferences about the causal structure

- For example, if an unmeasured confounder affects two observed variables, the algorithm may incorrectly infer a direct causal relationship between them

Structure learning with constraint-based algorithms

Independence tests in constraint-based algorithms

- Constraint-based algorithms rely on statistical tests to assess conditional independence between variables
- Common independence tests include: 1. Chi-square test for categorical variables 2. Partial correlation test for continuous variables 3. Kernel-based tests for non-linear relationships
- The choice of independence test depends on the nature of the data and the assumed functional relationships between variables

Skeleton discovery phase

- The first step in constraint-based algorithms is to learn the undirected graph structure (skeleton) by performing a series of conditional independence tests
- The algorithm starts with a complete undirected graph and iteratively removes edges between variables that are conditionally independent given a subset of other variables
- The size of the conditioning set is gradually increased to test for higher-order conditional independencies
- The resulting skeleton represents the undirected version of the causal graph, capturing the adjacencies between variables

V-structures and orientation rules

- After learning the skeleton, constraint-based algorithms employ orientation rules to determine the direction of edges and identify v-structures (colliders)
- V-structures are patterns where two variables are not directly connected but have a common effect (child) and are not conditionally independent given any subset of other variables
- Other orientation rules, such as the Meek rules, are applied to propagate edge orientations and ensure acyclicity of the resulting graph
- The output of this phase is a partially directed acyclic graph (PDAG) or a completed partially directed acyclic graph (CPDAG), representing the Markov equivalence class of the true causal graph

Popular constraint-based algorithms

PC algorithm

- Named after its authors, Peter Spirtes and Clark Glymour
- Employs a two-phase approach: skeleton discovery followed by edge orientation
- Starts with a complete undirected graph and iteratively removes edges based on conditional independence tests
- Applies orientation rules to determine the direction of edges and identify v-structures

- Outputs a CPDAG representing the Markov equivalence class of the true causal graph

FCI algorithm

- Fast Causal Inference (FCI) algorithm is an extension of the PC algorithm that relaxes the causal sufficiency assumption
- Allows for the presence of latent confounders and selection bias
- Employs additional orientation rules to identify potential latent common causes and distinguish between directed and bidirected edges
- Outputs a partial ancestral graph (PAG) that represents the equivalence class of causal graphs with latent variables

RFCI algorithm

- Really Fast Causal Inference (RFCI) algorithm is a computationally efficient variant of the FCI algorithm
- Avoids unnecessary conditional independence tests by exploiting the information gained from previous tests
- Achieves faster runtime compared to FCI while still allowing for latent confounders and selection bias
- Outputs a PAG similar to FCI, but with fewer conditional independence tests performed

Advantages vs disadvantages of constraint-based algorithms

Scalability and computational efficiency

- Constraint-based algorithms are generally more computationally efficient compared to score-based algorithms, especially for large datasets with many variables
- The number of conditional independence tests grows polynomially with the number of variables, making these algorithms scalable to high-dimensional settings
- However, the computational complexity can still be high for datasets with a large number of variables and high-order conditional independence tests

Robustness to latent confounders

- Some constraint-based algorithms, such as FCI and RFCI, can handle the presence of latent confounders and selection bias
- These algorithms can identify potential latent common causes and distinguish between directed and bidirected edges in the output graph
- This robustness to latent confounders is an advantage over score-based algorithms that typically assume causal sufficiency

Sensitivity to independence test parameters

- Constraint-based algorithms rely heavily on the accuracy of the conditional independence tests
- The choice of independence test, significance level, and sample size can impact the performance of these algorithms

- False positives or false negatives in the independence tests can lead to incorrect graph structures and causal conclusions
- Careful selection of independence test parameters and multiple testing correction methods is crucial for reliable results

Applications of constraint-based algorithms

Gene regulatory network inference

- Constraint-based algorithms are widely used to infer gene regulatory networks from gene expression data
- By treating genes as variables and their expression levels as observations, these algorithms can identify causal relationships between genes
- The inferred networks provide insights into the complex interactions and regulatory mechanisms underlying biological processes (cell differentiation, stress response)

Causal discovery in social sciences

- Constraint-based algorithms find applications in social sciences to uncover causal relationships between variables of interest
- For example, researchers can use these algorithms to study the causal factors influencing educational outcomes, socioeconomic status, or political preferences
- The identified causal structures can inform policy decisions and interventions aimed at addressing social issues (poverty, inequality)

Causal structure learning in healthcare

- Constraint-based algorithms are employed in healthcare to learn causal relationships between clinical variables, treatments, and patient outcomes
- By analyzing electronic health records or clinical trial data, these algorithms can identify potential causal pathways and inform treatment decisions
- The discovered causal structures can aid in personalized medicine, risk prediction, and the design of more effective interventions (drug development, disease management strategies)

10.2 Score-based algorithms

Score-based algorithms overview

Score-based algorithms estimate treatment effects by balancing observed covariates between treatment and control groups. They aim to mimic a randomized controlled trial by creating pseudo-populations where treatment assignment is independent of observed covariates. The three main approaches are propensity score matching, inverse probability weighting, and doubly robust estimation.

Each method uses a different strategy to handle confounding, but they all rely on the same core assumption: that you've measured all the relevant confounders. Understanding when and how to apply each one, and how to check whether it's working, is what this section covers.

Propensity score matching

Propensity score calculation

The propensity score is the probability of receiving treatment given the observed covariates:

$$e(X) = P(T = 1 \mid X)$$

You typically estimate this with logistic regression, where treatment assignment is the dependent variable and the covariates are predictors. The key idea is dimensionality reduction: instead of trying to match on dozens of covariates simultaneously, you collapse them into a single number (the propensity score) and match on that.

Once you have estimated propensity scores, you pair treated and control units that have similar scores. The goal is a matched sample where the distribution of observed covariates looks roughly the same across both groups.

Matching methods

- Nearest neighbor matching pairs each treated unit with the control unit whose propensity score is closest (1:1 matching). Simple and widely used, but can produce poor matches when the nearest neighbor is still far away.
- Caliper matching adds a constraint: matched pairs must have propensity scores within a specified maximum distance (the caliper width). This prevents bad matches at the cost of potentially discarding some treated units.
- Matching with replacement allows a single control unit to serve as the match for multiple treated units. This improves match quality, especially when good controls are scarce, but means some control units carry more weight in the analysis.
- Optimal matching minimizes the total distance across all matched pairs simultaneously, rather than greedily pairing one unit at a time. It tends to produce better overall balance than nearest neighbor matching.

Advantages vs limitations

Propensity score matching is intuitive and produces a matched dataset that's easy to analyze and interpret. It can effectively reduce confounding bias when the relevant covariates are measured.

The downsides are real, though. Matching often discards unmatched units, shrinking your sample and reducing statistical power. If there's limited overlap in propensity scores between groups (poor common support), you may lose a large fraction of your data. Most critically, matching cannot fix bias from unmeasured confounders. The entire approach rests on the no unmeasured confounding assumption.

Inverse probability weighting

Propensity score weighting

Inverse probability weighting (IPW) takes a different approach than matching. Instead of pairing units, it reweights the entire sample so that the observed covariate distributions are balanced across treatment groups.

Each unit receives a weight based on the inverse of its probability of receiving the treatment it actually got:

$$w_i = \frac{T_i}{e(X_i)} + \frac{1 - T_i}{1 - e(X_i)}$$

The intuition: units that received an unlikely treatment assignment (e.g., a treated unit with a low propensity score) get upweighted because they carry more information about what would have happened under the other condition. The result is a weighted pseudo-population where covariates are balanced.

Stabilized vs unstabilized weights

Unstabilized weights can become very large when propensity scores are close to 0 or 1. A single unit with a propensity score of 0.01 gets a weight of 100, which inflates variance and makes estimates unstable.

Stabilized weights address this by multiplying by the marginal probability of treatment:

$$sw_i = \frac{P(T_i)}{e(X_i)} \cdot T_i + \frac{1 - P(T_i)}{1 - e(X_i)} \cdot (1 - T_i)$$

Stabilized weights have a mean closer to 1 and lower variance, which translates to more efficient and stable estimates. In practice, you should almost always prefer stabilized weights.

Marginal structural models

Marginal structural models (MSMs) extend IPW to settings with time-varying treatments and time-varying confounders. This is the scenario where standard regression adjustment breaks down: if a confounder at time t is affected by treatment at time $t - 1$, conditioning on it introduces bias, but ignoring it also introduces bias.

MSMs solve this by modeling the counterfactual outcome as a function of the full treatment history, using IPW to adjust for the time-varying confounding at each time point. This makes them especially useful in longitudinal studies (e.g., estimating the effect of sustained medication use over time).

Doubly robust estimation

Outcome regression

Outcome regression directly models the relationship between covariates and the outcome, separately for treated and control groups. You fit a model (linear regression for continuous outcomes, logistic for binary), then use it to predict each unit's potential outcomes under both treatment conditions.

The treatment effect estimate comes from comparing these predicted potential outcomes. The weakness: if your outcome model is misspecified, your estimates will be biased.

Propensity score modeling

In the doubly robust framework, you also estimate propensity scores (typically via logistic regression), just as you would for IPW. These scores are used to construct weights or to adjust the outcome model for the treatment assignment mechanism.

Combining outcome and propensity models

The doubly robust estimator combines both models. The critical property is this: the estimator is consistent if either the outcome model or the propensity score model is correctly specified. You don't need both to be right, just one.

This "double protection" is why the method is attractive. In practice, of course, having both models approximately correct yields the best performance. Doubly robust estimators also tend to be more efficient than IPW alone when the outcome model is reasonably well-specified.

Assessing covariate balance

After applying any score-based method, you need to verify that it actually worked. Balance checking is not optional.

Standardized mean differences

The standardized mean difference (SMD) for a covariate is the difference in group means divided by the pooled standard deviation. SMDs are scale-free, so you can compare balance across covariates measured in different units.

Rules of thumb: - SMD < 0.1: good balance - SMD between 0.1 and 0.2: borderline - SMD > 0.2: meaningful imbalance that likely needs to be addressed

Always check SMDs after matching or weighting, not just before.

Variance ratios

Variance ratios compare the spread of a covariate across groups. A ratio near 1 means the variability is similar; ratios far from 1 indicate that one group has a much wider or narrower distribution. This matters most for continuous covariates, where two groups could have identical means but very different distributions.

Graphical diagnostics

Plots like side-by-side boxplots, overlaid density curves, or quantile-quantile plots let you visually inspect the full distribution of each covariate across groups. These are especially valuable for catching non-linear imbalances (e.g., different distributional shapes) that summary statistics like SMDs would miss.

Sensitivity analysis

Unobserved confounding

Every score-based method assumes no unmeasured confounding. Since this assumption is untestable, sensitivity analysis asks: how strong would an unobserved confounder need to be to change our conclusions?

If the answer is "implausibly strong," you can be more confident in your results. If even a modest confounder could flip the findings, the evidence is fragile.

Rosenbaum bounds

Rosenbaum bounds are the standard sensitivity analysis tool for matched studies. The approach works by introducing a parameter Γ that represents the degree of hidden bias. At $\Gamma = 1$, there's no hidden bias. As Γ increases, you're allowing for stronger unmeasured confounding.

You then ask: at what value of Γ does the treatment effect become statistically insignificant? A study that remains significant up to $\Gamma = 3$ is more robust than one that breaks down at $\Gamma = 1.2$.

Simulation-based approaches

An alternative is to directly simulate hypothetical unobserved confounders. You specify a plausible confounder's relationship with both treatment and outcome, then re-estimate the treatment effect including this simulated variable. By varying the confounder's strength and prevalence, you map out how sensitive your results are across different confounding scenarios.

This approach is more flexible than Rosenbaum bounds and can be applied to any score-based method, not just matched designs.

Extensions and variations

Matching with replacement

Allowing control units to be reused across multiple matches improves match quality when good controls are scarce. The trade-off is that your effective sample size shrinks (since some controls are counted multiple times), and you need to use variance estimators that account for the non-independence this creates.

Caliper matching

The caliper width controls the bias-variance trade-off. A narrow caliper ensures high-quality matches but may discard many units, reducing power and potentially limiting generalizability. A common starting point is a caliper of 0.2 standard deviations of the logit of the propensity score, though this should be tuned based on the data.

Coarsened exact matching

Coarsened exact matching (CEM) takes a fundamentally different approach. Instead of estimating propensity scores, you coarsen each covariate into bins (e.g., age groups instead of exact ages), then require exact matches on the coarsened values.

CEM guarantees balance on the coarsened covariates by construction, handles both continuous and categorical variables, and avoids model dependence in the matching step. The downside is that fine-grained coarsening with many covariates can leave few or no exact matches, especially in smaller samples.

Practical considerations

Sample size and overlap

Score-based methods need adequate sample size and sufficient overlap (common support) in propensity score distributions between groups. Check overlap by plotting the propensity score distributions for treated and control units. If there are regions where one group has scores but the other doesn't, units in those regions can't be reliably compared.

When overlap is poor, consider trimming the sample to the region of common support, or switching to methods like CEM that handle limited overlap differently.

Missing data handling

Missing covariate data complicates propensity score estimation. Your approach should depend on the missing data mechanism:

- Missing completely at random (MCAR): Complete case analysis is unbiased but wasteful.
- Missing at random (MAR): Multiple imputation is generally preferred. Estimate propensity scores within each imputed dataset, then combine results using Rubin's rules.
- Missing not at random (MNAR): No standard fix exists. Sensitivity analysis is essential.

Using missing data indicators as covariates is sometimes done in practice, but this approach can introduce bias and should be used cautiously.

Software implementations

Several well-maintained packages support these methods:

- R: MatchIt (matching), WeightIt (weighting), twang (generalized boosted models for propensity scores), CBPS (covariate balancing propensity scores)
- Stata: teffects (built-in treatment effects), psmatch2 (propensity score matching), ipw (inverse probability weighting)
- Python: DoWhy (causal inference framework), causalinference (propensity score methods)

Each package has different defaults and capabilities, so check the documentation for your specific use case rather than relying on default settings.

10.3 Hybrid algorithms

Hybrid algorithms combine machine learning's flexibility with the statistical rigor needed for valid causal inference. They let you use powerful predictive models (random forests, neural nets, ensemble learners) to estimate nuisance parameters while still getting unbiased causal effect estimates with proper confidence intervals. The three workhorses are targeted maximum likelihood estimation (TMLE), augmented inverse probability weighting (AIPW), and double machine learning (DML).

Hybrid algorithms overview

The core problem these methods solve: traditional causal inference estimators (like plain IPW or outcome regression) require you to correctly specify parametric models. Get the model wrong, and your estimates are biased. But if you just plug in a machine learning model instead, you lose the ability to do valid statistical inference because ML introduces its own biases (regularization bias, overfitting).

Hybrid algorithms resolve this tension through two key ideas:

- Double robustness: The estimator remains consistent if either the propensity score model or the outcome regression model is correctly specified (not necessarily both).
- Cross-fitting: Nuisance parameters are estimated on separate data folds from those used to construct the final estimator, preventing overfitting bias from contaminating inference.

All three methods build on the efficient influence function (EIF), which provides the theoretical backbone for achieving optimal efficiency and valid confidence intervals.

Targeted maximum likelihood estimation (TMLE)

TMLE procedure

TMLE starts with initial estimates of the outcome regression and propensity score, then updates those estimates in a targeted way so the final estimator solves the efficient influence function equation. This targeting step is what distinguishes TMLE from simply plugging ML predictions into a formula.

The procedure works as follows:

1. Estimate the outcome regression $\hat{m}(A, X) = \hat{E}[Y | A, X]$ using a flexible ML method (e.g., Super Learner, an ensemble of multiple algorithms).
2. Estimate the propensity score $\hat{e}(X) = \hat{P}(A = 1 | X)$ using another ML method.
3. Define a fluctuation submodel that tilts the initial outcome regression estimate along the direction of the EIF. This submodel is parameterized by a scalar ϵ and typically uses the "clever covariate" $H(A, X) = \frac{A}{\hat{e}(X)} - \frac{1-A}{1-\hat{e}(X)}$ as an offset in a logistic regression.
4. Fit the fluctuation parameter ϵ by maximizing the targeted likelihood (e.g., running a simple logistic regression of Y on $H(A, X)$ with $\hat{m}(A, X)$ as an offset).
5. Update the initial estimate: $\hat{m}^*(A, X) = \text{expit}(\text{logit}(\hat{m}(A, X)) + \hat{\epsilon} \cdot H(A, X))$.
6. Compute the plug-in estimate:

$$\hat{\psi}_{TMLE} = \frac{1}{n} \sum_{i=1}^n [\hat{m}^*(1, X_i) - \hat{m}^*(0, X_i)]$$

for the ATE.

The update step ensures the estimator solves the EIF estimating equation, which is what gives TMLE its double robustness and efficiency properties.

TMLE for causal effect estimation

TMLE can target different causal parameters by changing the EIF used in the fluctuation step:

- Average treatment effect (ATE): $\psi = E[Y(1) - Y(0)]$
- Average treatment effect on the treated (ATT): $\psi = E[Y(1) - Y(0) | A = 1]$
- Conditional average treatment effect (CATE): $\psi(x) = E[Y(1) - Y(0) | X = x]$

A common choice for the initial ML estimator is Super Learner, which builds a weighted ensemble of candidate algorithms (GLMs, random forests, gradient boosting, etc.) using cross-validated risk to select the optimal combination. This reduces the risk of choosing a single poorly-specified model.

TMLE vs traditional methods

Property	TMLE	IPW	Outcome Regression
Double robustness	Yes	No	No
Asymptotic efficiency	Yes (when both models correct)	No	No
ML-compatible	Yes (via targeting step)	Limited	Limited
Valid inference with ML nuisance estimates	Yes (with cross-fitting)	Generally no	Generally no

The key advantage: IPW fails if the propensity score is wrong, outcome regression fails if the outcome model is wrong, but TMLE only needs one of the two to be right. When both are correct, TMLE achieves the semiparametric efficiency bound.

Augmented inverse probability weighting (AIPW)

AIPW estimator

AIPW combines IPW and outcome regression into a single estimator that is doubly robust. For estimating the ATE, the AIPW estimator for $E[Y(1)]$ is:

$$\hat{\psi}_{AIPW} = \frac{1}{n} \sum_{i=1}^n \left(\frac{A_i Y_i}{\hat{e}(X_i)} - \frac{A_i - \hat{e}(X_i)}{\hat{e}(X_i)} \hat{m}(1, X_i) \right)$$

where $\hat{e}(X_i)$ is the estimated propensity score and $\hat{m}(1, X_i)$ is the estimated outcome under treatment.

The intuition: start with the IPW estimator, then add an augmentation term that corrects for errors in the propensity score model using the outcome regression. If the propensity score is correct, the augmentation term has expectation zero and doesn't hurt. If the propensity score is wrong but the outcome model is right, the augmentation term corrects the bias. This is where double robustness comes from.

An analogous expression is constructed for $E[Y(0)]$, and the ATE estimate is the difference.

AIPW for missing data problems

AIPW extends naturally to missing data under the missing at random (MAR) assumption. The logic is the same: replace the treatment indicator A with a missingness indicator R , the propensity score with the probability of being observed, and the outcome regression with the expected outcome given covariates.

- Estimate $P(R = 1 | X)$, the probability of observing the outcome.
- Estimate $E[Y | X, R = 1]$, the outcome regression among observed cases.
- The AIPW estimator reweights observed outcomes by the inverse probability of being observed and augments with the predicted outcome for everyone.

This gives consistent estimates as long as either the missingness model or the outcome model is correct.

AIPW vs IPW and outcome regression

- IPW is consistent only if the propensity score is correctly specified. It can also be highly variable when propensity scores are near 0 or 1 (positivity violations).
- Outcome regression is consistent only if the outcome model is correctly specified.

- AIPW is consistent if either model is correct, and it achieves the semiparametric efficiency bound when both are correct.

AIPW also tends to be more stable than IPW in practice because the augmentation term reduces variance, especially when propensity scores are extreme.

Efficient influence functions (EIF)

EIF definition and properties

The efficient influence function (EIF) is a central object in semiparametric efficiency theory. It characterizes the best possible estimator for a given target parameter in a nonparametric or semiparametric model.

The EIF has several defining properties:

- It is a mean-zero function of the observed data: $E[\text{EIF}(O; \psi_0, \eta_0)] = 0$, where ψ_0 is the true parameter and η_0 are the true nuisance parameters.
- It is the pathwise derivative of the target parameter functional, meaning it captures how the parameter changes under small perturbations of the data distribution.
- Its variance equals the semiparametric efficiency bound: no regular asymptotically linear (RAL) estimator can have smaller asymptotic variance than $\text{Var}(\text{EIF})/n$.

For the ATE, the EIF takes the form:

$$\text{EIF}(O; \psi, \eta) = \frac{A(Y - m(1, X))}{e(X)} - \frac{(1 - A)(Y - m(0, X))}{1 - e(X)} + m(1, X) - m(0, X) - \psi$$

This expression shows exactly how each component (propensity score, outcome regression, observed data) contributes to the efficient estimator.

EIF in TMLE and AIPW

Both TMLE and AIPW are constructed so that they solve the EIF estimating equation:

$$\frac{1}{n} \sum_{i=1}^n \widehat{\text{EIF}}(O_i; \hat{\psi}, \hat{\eta}) \approx 0$$

- In TMLE, the targeting/fluctuation step forces this equation to hold exactly. The clever covariate in the fluctuation submodel is derived directly from the EIF.
- In AIPW, the estimator is literally the sample mean of the EIF evaluated at estimated nuisance parameters: $\hat{\psi}_{\text{AIPW}} = \psi_{\text{init}} + \frac{1}{n} \sum_i \widehat{\text{EIF}}(O_i)$.

Because both methods solve the same estimating equation, they are asymptotically equivalent when both nuisance models are consistently estimated. They can differ in finite samples, though, because TMLE respects the bounds of the outcome (e.g., staying in $[0, 1]$ for binary outcomes) while AIPW does not.

EIF-based confidence intervals

Once you have an estimator $\hat{\psi}$ that solves the EIF equation, you can construct confidence intervals using the estimated EIF values as if they were i.i.d. observations:

$$\hat{\psi} \pm z_{\alpha/2} \sqrt{\frac{1}{n^2} \sum_{i=1}^n \widehat{\text{EIF}}(O_i; \hat{\psi}, \hat{\eta})^2}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

These confidence intervals have correct asymptotic coverage as long as:

- The estimator is $\sqrt{n}$ -consistent and asymptotically normal.
- The nuisance parameter estimators converge fast enough (faster than $n^{-1/4}$).
- The propensity score is bounded away from 0 and 1.

This is a major practical advantage: you get valid inference even though the nuisance parameters were estimated with black-box ML methods.

Double machine learning (DML)

DML framework

Double machine learning, introduced by Chernozhukov et al. (2018), provides a general framework for combining ML nuisance estimation with valid causal inference. The name "double" refers to two features: double robustness and the use of two stages of estimation.

The DML procedure:

1. Split the data into K folds (typically $K = 5$ or $K = 10$).
2. For each fold k : use the remaining $K - 1$ folds to train ML models for the nuisance parameters (propensity score $\hat{e}^{(-k)}$ and outcome regression $\hat{m}^{(-k)}$).
3. Predict on the held-out fold k using these trained models.
4. Compute the EIF-based score for each observation in fold k using the out-of-fold predictions.
5. Average the scores across all n observations to get the final DML estimate.

The critical insight: by always predicting on data that was not used for training, cross-fitting prevents overfitting bias from leaking into the causal estimate. Without cross-fitting, regularization bias from ML methods (which converge slower than $n^{-1/2}$) would distort the asymptotic distribution of the estimator.

DML for treatment effect estimation

For the ATE, the DML estimator takes the form:

$$\hat{\psi}_{DML} = \frac{1}{n} \sum_{i=1}^n \left[\hat{m}^{(-k_i)}(1, X_i) - \hat{m}^{(-k_i)}(0, X_i) + \frac{A_i(Y_i - \hat{m}^{(-k_i)}(1, X_i))}{\hat{e}^{(-k_i)}(X_i)} - \frac{(1 - A_i)(Y_i - \hat{m}^{(-k_i)}(0, X_i))}{1 - \hat{e}^{(-k_i)}(X_i)} \right]$$

where k_i denotes the fold that observation i belongs to, and the $(-k_i)$ superscript means "trained without fold k_i ."

DML estimators are $\sqrt{n}$ -consistent and asymptotically normal under a key rate condition: the product of the convergence rates of the two nuisance estimators must be faster than $n^{-1/2}$. Concretely, if both the propensity score and outcome regression converge at rate $n^{-1/4}$ or faster, the DML estimator achieves $\sqrt{n}$ -normality. This is a much weaker requirement than needing either model to converge at the parametric $n^{-1/2}$ rate.

DML vs traditional machine learning

Traditional ML optimizes for prediction accuracy (minimizing out-of-sample loss). DML optimizes for causal parameter estimation with valid inference. The differences matter:

- Prediction ML can be biased for causal effects due to regularization, confounding, and lack of double robustness. A well-tuned random forest predicts Y accurately but doesn't tell you the causal effect of A .

- DML uses ML only for nuisance estimation, then constructs the causal estimate through the EIF. The Neyman orthogonality of the EIF makes the final estimate insensitive to small errors in the nuisance models.
- Inference: DML provides asymptotically valid confidence intervals and p-values. Standard ML does not.

Cross-fitting technique

Cross-fitting procedure

Cross-fitting is the sample-splitting strategy that makes hybrid algorithms work with ML. It's similar to cross-validation but serves a different purpose: cross-validation selects models, while cross-fitting prevents overfitting bias in causal estimation.

Step-by-step:

1. Randomly partition the data into K equally-sized folds ($K = 5$ is common).
2. For fold $k = 1$: train the propensity score model and outcome regression model on folds 2 through K . Generate predicted values for observations in fold 1.
3. Repeat for each fold: train on all folds except k , predict on fold k .
4. Now every observation has out-of-fold predictions for both nuisance parameters.
5. Compute the EIF-based estimator using these out-of-fold predictions for all n observations.

The result: each observation's nuisance parameter estimates come from a model that never saw that observation during training.

Cross-fitting in TMLE and DML

In DML, cross-fitting is built into the core procedure. The EIF scores are computed using out-of-fold nuisance estimates, and the final estimator averages these scores.

In TMLE, cross-fitting is used for the initial estimation step. The outcome regression and propensity score are estimated via cross-fitting, and then the targeting/fluctuation step can be performed using these out-of-fold estimates. The final TMLE estimate averages the targeted predictions across folds.

Both approaches achieve the same goal: ensuring that the ML-induced bias in nuisance estimation is orthogonal to the causal parameter of interest, so it vanishes in the asymptotic distribution.

Cross-fitting benefits and trade-offs

Benefits: - Eliminates overfitting bias that would otherwise invalidate inference when using adaptive ML methods. - Allows you to use the full sample for both nuisance estimation and causal effect estimation (unlike a simple train/test split, which wastes data). - Enables the use of highly flexible ML methods (deep learners, boosted trees) without sacrificing inferential validity.

Trade-offs: - Computational cost scales linearly with K , since you train nuisance models K times. - Each training set is only $(K - 1)/K$ of the full sample, so nuisance estimates may be slightly less accurate than using all the data. - The choice of K involves a bias-variance trade-off: larger K means larger training sets per fold (less bias in nuisance estimation) but more computation. In practice, $K = 5$ or $K = 10$ works well for most sample sizes.

Hybrid algorithms performance

Efficiency and robustness

The two defining properties of hybrid algorithms are:

- Semiparametric efficiency: When both nuisance models are consistently estimated, the estimator achieves the smallest possible asymptotic variance among all RAL estimators. This variance equals $\text{Var}(\text{EIF})/n$.
- Double robustness: The estimator remains $\sqrt{n}$ -consistent if either the propensity score or the outcome regression (but not necessarily both) is consistently estimated.

These properties stem from the Neyman orthogonality of the EIF: the EIF score is locally insensitive to perturbations in the nuisance parameters around their true values. This is why small errors in ML-estimated nuisance parameters don't propagate into the causal estimate.

Finite sample properties

Asymptotic guarantees don't always translate directly to finite samples. Several factors affect real-world performance:

- Sample size: With small samples (say, $n < 500$), the $n^{-1/4}$ rate condition may not be effectively satisfied, and confidence interval coverage can be poor.
- Positivity violations: If some propensity scores are very close to 0 or 1, all three methods can become unstable. Trimming or truncating extreme propensity scores helps but introduces bias.
- ML model choice: The quality of the nuisance estimates matters. Poorly tuned ML models yield poor causal estimates even with double robustness, because double robustness protects against misspecification of one model, not both.
- Tuning and regularization: Overly aggressive regularization can slow nuisance convergence rates below the required $n^{-1/4}$ threshold.

Simulation studies and sensitivity analyses are essential for assessing performance in any specific application.

Asymptotic properties

Under regularity conditions, hybrid algorithms satisfy:

- $\sqrt{n}$ -consistency: $\hat{\psi} - \psi_0 = O_p(n^{-1/2})$
- Asymptotic normality: $\sqrt{n}(\hat{\psi} - \psi_0) \xrightarrow{d} N(0, \text{Var}(\text{EIF}))$
- Semiparametric efficiency: The asymptotic variance equals the efficiency bound.

These results require: - Nuisance estimators converge at rates such that the product of their errors is $o_p(n^{-1/2})$. The standard sufficient condition is that each converges at $o_p(n^{-1/4})$. - The propensity score is bounded: $0 < \delta \leq e(X) \leq 1 - \delta < 1$ for some $\delta > 0$ (positivity). - Standard regularity conditions (finite variance of the EIF, Donsker-type conditions or use of cross-fitting to avoid them).

Cross-fitting is particularly valuable because it relaxes the Donsker condition, which would otherwise restrict the complexity of the ML methods you can use.

Applications of hybrid algorithms

Observational studies

Observational studies are the primary use case for hybrid algorithms. Without randomization, treatment assignment depends on covariates, creating confounding. Hybrid algorithms address this by:

- Estimating the propensity score $P(A = 1 | X)$ to model treatment assignment.
- Estimating the outcome regression $E[Y | A, X]$ to model the outcome mechanism.
- Combining both through the EIF to get a doubly robust causal estimate.

Classic applications include estimating the effect of job training programs on earnings (as in the LaLonde dataset), assessing medical treatment effectiveness from electronic health records, and evaluating policy interventions using administrative data. In all these settings, the true data-generating process is unknown, making double robustness especially valuable.

Randomized trials with non-compliance

Even in randomized experiments, non-compliance complicates causal inference. If some participants assigned to treatment don't actually take it (or vice versa), the intention-to-treat (ITT) estimate dilutes the true treatment effect.

Hybrid algorithms can estimate the complier average causal effect (CACE), the effect among those who would comply with their assignment, by:

- Using randomized assignment Z as an instrumental variable.
- Estimating the compliance model and outcome model with ML.
- Applying TMLE or DML adapted for IV estimation.

This approach has been used in studies of school voucher programs, medication adherence trials, and behavioral interventions where dropout is common.

Longitudinal studies

Hybrid algorithms extend to longitudinal settings where treatments, confounders, and outcomes evolve over time. Time-varying confounding is the central challenge: a variable at time t can be both a confounder for future treatment and a mediator of past treatment.

Adapted methods include:

- Longitudinal TMLE (LTMLE): Sequentially targets the outcome regression at each time point, working backward from the final outcome.
- Longitudinal IPW / marginal structural models: Weight each person-time observation by the cumulative product of inverse probability weights across time points.
- Sequential DML: Applies the DML framework to sequential decision problems.

These methods handle time-varying confounding and informative censoring (when dropout depends on covariates or treatment history) under sequential versions of the no unmeasured confounders assumption.

10.4 Causal feature selection

Causal feature selection is a crucial technique in causal inference that identifies variables directly impacting outcomes. It distinguishes between predictive and causal features, helping build interpretable models for decision-making and understanding underlying mechanisms.

This approach improves model accuracy by focusing on relevant variables, reducing complexity, and enhancing interpretability. It addresses challenges like confounding, collider bias, and selection bias, using methods such as filter, wrapper, and embedded techniques to select the most impactful features.

Causal feature selection overview

- Causal feature selection focuses on identifying the most relevant variables for predicting the outcome of interest while considering the underlying causal relationships
- Aims to distinguish between predictive features and causal features, which have a direct causal effect on the outcome
- Plays a crucial role in building interpretable and reliable causal models for decision-making and understanding the underlying mechanisms

Importance of causal features

Identifying relevant variables

- Causal feature selection helps identify the variables that have a direct causal effect on the outcome, rather than just being correlated with it
- Distinguishes between confounders, mediators, and colliders, which can lead to biased estimates if not properly accounted for
- Enables researchers to focus on the most important variables for intervention and policy-making

Reducing model complexity

- By selecting only the causal features, the complexity of the model can be reduced, leading to improved computational efficiency and reduced overfitting
- Eliminates irrelevant or redundant features that may introduce noise and reduce the model's performance
- Facilitates the development of more parsimonious and interpretable models

Improving interpretability

- Causal feature selection enhances the interpretability of the model by focusing on the variables that have a direct causal effect on the outcome
- Allows for a clearer understanding of the underlying causal mechanisms and the potential impact of interventions
- Facilitates communication of the model's findings to stakeholders and decision-makers

Causal feature selection methods

Filter methods

- Filter methods assess the relevance of features based on their statistical properties, such as correlation or mutual information, without considering the specific model
- Examples include correlation-based feature selection (CFS) and minimum redundancy maximum relevance (mRMR)
- Filter methods are computationally efficient but may not capture the complex causal relationships between variables

Wrapper methods

- Wrapper methods evaluate feature subsets by training and testing a specific model, iteratively adding or removing features to optimize the model's performance
- Examples include recursive feature elimination (RFE) and genetic algorithms
- Wrapper methods can capture the interactions between features and the model but are computationally expensive and may overfit the data

Embedded methods

- Embedded methods incorporate feature selection as part of the model training process, simultaneously learning the model parameters and selecting the relevant features
- Examples include Lasso regression, which uses L1 regularization to shrink the coefficients of irrelevant features to zero
- Embedded methods provide a balance between computational efficiency and capturing the interactions between features and the model

Challenges in causal feature selection

Confounding variables

- Confounding variables are associated with both the exposure and the outcome, leading to biased estimates if not properly adjusted for
- Causal feature selection methods must account for confounding variables to identify the true causal effects
- Techniques such as propensity score matching or inverse probability weighting can be used to adjust for confounding

Collider bias

- Collider bias occurs when conditioning on a variable that is a common effect of both the exposure and the outcome, inducing a spurious association
- Causal feature selection methods must be cautious when selecting variables to avoid introducing collider bias
- Directed acyclic graphs (DAGs) can help identify potential colliders and guide the selection process

Selection bias

- Selection bias arises when the sample used for analysis is not representative of the target population, leading to biased estimates
- Causal feature selection methods must consider the potential impact of selection bias on the validity of the causal inferences
- Techniques such as inverse probability weighting or doubly robust estimation can be used to mitigate selection bias

Causal graphs for feature selection

Directed acyclic graphs (DAGs)

- DAGs are graphical representations of the causal relationships between variables, with arrows indicating the direction of causality
- DAGs can help identify confounders, mediators, and colliders, guiding the selection of appropriate adjustment sets for causal inference
- Examples of DAGs include the "back-door" and "front-door" criteria for identifying valid adjustment sets

Markov blankets

- A Markov blanket of a variable is the minimal set of variables that shields it from the rest of the variables in the graph
- The Markov blanket of the outcome variable contains all the variables needed for optimal prediction and causal inference
- Causal feature selection methods can leverage Markov blankets to identify the most relevant variables for the outcome

Backdoor criterion

- The backdoor criterion is a graphical criterion for identifying valid adjustment sets in a DAG
- A set of variables satisfies the backdoor criterion if it blocks all backdoor paths between the exposure and the outcome, while not including any descendants of the exposure
- The backdoor criterion can guide the selection of variables for causal inference and help avoid bias due to confounding

Causal discovery algorithms

PC algorithm

- The PC algorithm is a constraint-based causal discovery algorithm that learns the structure of a DAG from observational data
- It starts with a fully connected graph and iteratively removes edges based on conditional independence tests

- The PC algorithm can identify the Markov equivalence class of DAGs that are consistent with the observed data

FCI algorithm

- The Fast Causal Inference (FCI) algorithm is an extension of the PC algorithm that can handle latent confounders and selection bias
- FCI learns a partial ancestral graph (PAG) that represents the equivalence class of MAGs (maximal ancestral graphs) consistent with the observed data
- The FCI algorithm can identify the presence of latent confounders and guide the selection of adjustment sets for causal inference

GES algorithm

- The Greedy Equivalence Search (GES) algorithm is a score-based causal discovery algorithm that searches for the optimal DAG based on a scoring criterion (e.g., Bayesian Information Criterion)
- GES starts with an empty graph and iteratively adds and removes edges to maximize the score, while ensuring the graph remains acyclic
- The GES algorithm can identify the Markov equivalence class of DAGs that best fit the observed data

Evaluation of causal feature subsets

Causal effect estimation

- Causal effect estimation involves quantifying the effect of an exposure on an outcome, adjusting for potential confounders
- Techniques such as propensity score matching, inverse probability weighting, and doubly robust estimation can be used to estimate causal effects
- The performance of causal feature subsets can be evaluated based on the accuracy and precision of the estimated causal effects

Sensitivity analysis

- Sensitivity analysis assesses the robustness of the causal inferences to potential unmeasured confounders or violations of assumptions
- Techniques such as the E-value and the bias formula can be used to quantify the strength of unmeasured confounding needed to explain away the observed effect
- Causal feature subsets that are robust to potential unmeasured confounding are preferred for reliable causal inference

Cross-validation techniques

- Cross-validation techniques, such as k-fold cross-validation or leave-one-out cross-validation, can be used to assess the performance of causal feature subsets on unseen data
- The causal feature subsets are evaluated based on their ability to accurately estimate causal effects or predict outcomes in the validation set

- Cross-validation helps prevent overfitting and provides a more reliable assessment of the generalizability of the causal feature subsets

Applications of causal feature selection

Biomedical research

- Causal feature selection is crucial in biomedical research for identifying the key risk factors and potential targets for intervention
- Examples include identifying the causal genes for complex diseases, understanding the mechanisms of drug action, and developing personalized treatment strategies
- Causal feature selection can help prioritize the most promising biomarkers and guide the design of clinical trials

Social sciences

- Causal feature selection is important in social sciences for understanding the complex interplay of social, economic, and environmental factors on human behavior and outcomes
- Examples include identifying the causal determinants of educational attainment, understanding the impact of social policies on health disparities, and evaluating the effectiveness of interventions
- Causal feature selection can help disentangle the causal effects of various factors and guide evidence-based policy-making

Business decision-making

- Causal feature selection can inform business decision-making by identifying the key drivers of customer behavior, market trends, and organizational performance
- Examples include understanding the causal impact of marketing campaigns on sales, identifying the factors that influence employee retention, and optimizing supply chain processes
- Causal feature selection can help businesses make data-driven decisions and allocate resources more effectively

Future directions in causal feature selection

Integration with machine learning

- Integrating causal feature selection with machine learning algorithms can improve the interpretability and robustness of predictive models
- Techniques such as causal regularization and causal representation learning can incorporate causal knowledge into the learning process
- Combining causal inference with machine learning can lead to more accurate and reliable predictions, especially in the presence of distribution shifts or interventions

Handling high-dimensional data

- Causal feature selection methods need to be adapted to handle high-dimensional data, such as genomic or neuroimaging data, where the number of features far exceeds the sample size

- Techniques such as sparse causal discovery algorithms and regularization methods can be used to identify the most relevant features in high-dimensional settings
- Incorporating prior knowledge or domain expertise can help guide the selection process and reduce the search space

Addressing unmeasured confounding

- Unmeasured confounding remains a major challenge in causal feature selection, as it can lead to biased estimates and invalid inferences
- Techniques such as instrumental variables, sensitivity analysis, and causal discovery algorithms that can handle latent confounders (e.g., FCI) can help mitigate the impact of unmeasured confounding
- Developing new methods that can explicitly model and adjust for unmeasured confounding is an important area of research in causal feature selection

Unit 11 – Social Science & Policy

Evaluation Applications

11.1 Economics and labor markets

Labor markets are complex systems where employers and workers interact to determine wages and employment levels. This chapter explores key concepts like labor demand and supply, wage determination, and factors affecting labor market outcomes.

Understanding labor markets is crucial for analyzing economic policies and their impacts on workers and firms. We'll examine topics such as human capital investment, discrimination, unions, unemployment, and the effects of globalization on labor markets.

Demand for labor

- Demand for labor is a derived demand, meaning it depends on the demand for the goods and services that labor is used to produce
- Firms hire workers to produce output, and the demand for labor is determined by the marginal productivity of labor and the price of the output produced

Marginal productivity theory

- States that a profit-maximizing firm will hire workers up to the point where the marginal product of labor equals the wage rate
- The marginal product of labor (MPL) is the additional output produced by hiring one more worker, holding all other inputs constant
- For example, if a firm can produce 100 units of output with 10 workers and 105 units with 11 workers, the MPL of the 11th worker is 5 units
- The value of the marginal product of labor (VMPL) is the marginal product of labor multiplied by the price of the output
- This represents the additional revenue generated by hiring one more worker

Elasticity of labor demand

- Measures the responsiveness of the quantity of labor demanded to a change in the wage rate
- Calculated as the percentage change in quantity of labor demanded divided by the percentage change in the wage rate
- Factors affecting elasticity of labor demand include:
 - Ease of substituting other inputs (capital, technology) for labor
 - Labor's share of total costs (a higher share implies more elastic demand)
 - Elasticity of demand for the final product (more elastic product demand implies more elastic labor demand)

Supply of labor

- Labor supply refers to the number of workers willing and able to work at various wage rates
- Individuals face a tradeoff between work and leisure, and the supply of labor is determined by factors such as wage rates, non-labor income, and preferences

Labor-leisure tradeoff model

- Assumes that individuals allocate their time between work and leisure to maximize utility
- The budget constraint shows the combinations of income and leisure that are possible given the wage rate and total available time
- The indifference curves represent the individual's preferences for income and leisure
- The optimal choice of labor supply occurs where the budget constraint is tangent to the highest possible indifference curve

Factors affecting labor supply

- Wage rates: Higher wages increase the opportunity cost of leisure, leading to an increase in labor supply (substitution effect)
- However, higher wages also increase income, which may lead to a decrease in labor supply (income effect)
- Non-labor income: An increase in non-labor income (inheritance, spouse's income) shifts the budget constraint outward, leading to a decrease in labor supply
- Preferences: Individuals with a stronger preference for leisure will supply less labor at any given wage rate
- Demographics: Labor supply varies by age, gender, and education level
- For example, labor force participation rates are typically higher for prime-age workers (25-54) than for younger or older workers

Labor market equilibrium

- The labor market reaches equilibrium when the quantity of labor demanded equals the quantity of labor supplied at a given wage rate
- In a competitive labor market, the equilibrium wage and employment level are determined by the intersection of the labor demand and supply curves

Competitive labor markets

- Characterized by a large number of firms and workers, with no single firm or worker able to influence the wage rate
- Firms are wage-takers, meaning they must pay the market wage determined by supply and demand
- In the long run, workers earn a wage equal to their marginal product, and firms earn zero economic profits

Monopsony in labor markets

- A monopsony is a labor market with a single buyer of labor (or a few buyers with significant market power)
- Monopsonists face an upward-sloping labor supply curve, meaning they must pay higher wages to attract additional workers
- As a result, monopsonists hire fewer workers and pay lower wages compared to a competitive labor market
- Examples of monopsonistic labor markets include:
 - Company towns where a single firm dominates employment
 - Specialized occupations with few potential employers (professional sports, academia)

Wage determination

- Various factors beyond labor market competition can influence wage rates, leading to differences in pay across workers, occupations, and industries

Compensating wage differentials

- Differences in wages that compensate workers for job characteristics such as risk, unpleasantness, or location
- For example, workers in dangerous occupations (logging, mining) may earn higher wages to compensate for the increased risk of injury or death
- Compensating differentials can also exist for positive job attributes, such as flexible hours or generous benefits

Efficiency wages

- Wages set above the market-clearing level to increase worker productivity and reduce turnover
- Firms may pay efficiency wages to:
 - Attract higher-quality workers
 - Reduce shirking (slacking off) by increasing the cost of job loss
 - Foster loyalty and reduce turnover, saving on hiring and training costs
- Efficiency wage theory helps explain persistent unemployment, as wages remain above the level that would clear the labor market

Minimum wage laws

- Legal minimum wages set a floor on the wage rate, making it illegal to pay workers less than the specified amount
- Proponents argue that minimum wages raise living standards for low-wage workers and reduce poverty
- Critics argue that minimum wages can lead to job losses, particularly for low-skilled workers, as firms reduce employment in response to higher labor costs

- The impact of minimum wages on employment depends on factors such as:
- The level of the minimum wage relative to the market wage
- The elasticity of labor demand in affected industries
- The ability of firms to pass on higher labor costs to consumers through price increases

Human capital investments

- Human capital refers to the knowledge, skills, and abilities that workers acquire through education, training, and experience
- Investments in human capital can increase worker productivity and lead to higher wages over time

Education and training

- Formal schooling (primary, secondary, tertiary) provides general knowledge and skills that are valuable in the labor market
- Vocational training and apprenticeships provide job-specific skills
- On-the-job training, both formal and informal, allows workers to acquire and refine skills while employed

Returns to human capital

- The returns to education can be measured by comparing the wages of workers with different levels of schooling
- For example, the college wage premium is the difference in earnings between college graduates and high school graduates
- Returns to education vary by field of study, with higher returns for degrees in STEM fields (science, technology, engineering, math) and business
- The returns to human capital also include non-monetary benefits, such as improved health, increased job satisfaction, and better decision-making skills

Discrimination in labor markets

- Labor market discrimination occurs when workers with similar skills and abilities are treated differently based on characteristics such as race, gender, age, or disability
- Discrimination can lead to wage gaps and occupational segregation, with disadvantaged groups earning lower wages and being concentrated in lower-paying jobs

Taste-based vs statistical discrimination

- Taste-based discrimination (Becker, 1957) occurs when employers, co-workers, or customers have a preference for interacting with certain groups and are willing to pay a price to avoid interacting with other groups
- For example, an employer with a distaste for a particular group may be willing to hire workers from that group only at a wage discount

- Statistical discrimination (Phelps, 1972; Arrow, 1973) occurs when employers use group membership as a proxy for unobserved productivity-related characteristics
- For example, if an employer believes that workers from a particular group are less reliable on average, they may be less likely to hire from that group, even if an individual job candidate is highly qualified

Measuring wage gaps

- Wage gaps can be measured using regression analysis to control for differences in worker characteristics such as education, experience, and occupation
- The unexplained portion of the wage gap (the residual) is often attributed to discrimination, although it may also reflect unobserved differences in productivity
- Examples of wage gaps:
 - The gender wage gap: women earn less than men on average, even after controlling for observable characteristics
 - The racial wage gap: in the US, black and Hispanic workers earn less than white workers on average, even after controlling for observable characteristics

Unions and collective bargaining

- Unions are organizations that represent workers in negotiations with employers over wages, benefits, and working conditions
- Collective bargaining is the process by which unions and employers negotiate employment contracts that cover union members

Union membership and coverage

- Union membership refers to the percentage of workers who belong to a union
- Union coverage refers to the percentage of workers whose terms of employment are covered by a collective bargaining agreement, regardless of whether they are union members
- In the US, union membership and coverage have declined significantly since the 1950s, with private-sector unionization falling from around 35% to less than 7% in 2020

Effects on wages and employment

- Unions aim to increase wages and benefits for their members through collective bargaining
- The union wage premium is the difference in wages between union and non-union workers, controlling for observable characteristics
- Studies typically find a union wage premium of 10-20% in the US
- Critics argue that unions can lead to job losses and reduced competitiveness, as higher labor costs may cause firms to reduce employment or go out of business
- The impact of unions on employment depends on factors such as:
 - The competitiveness of the product market
 - The substitutability of union and non-union labor

- The bargaining power of the union relative to the firm

Unemployment and job search

- Unemployment occurs when individuals are actively seeking work but are unable to find a job
- The unemployment rate is the number of unemployed workers as a percentage of the labor force (employed + unemployed)

Types of unemployment

- Frictional unemployment: results from the time it takes for workers to search for and find new jobs
- Includes unemployment due to job switching, labor market entry, and temporary layoffs
- Structural unemployment: occurs when there is a mismatch between the skills and location of unemployed workers and the requirements of available jobs
- Can result from technological change, shifts in industry composition, or geographic immobility
- Cyclical unemployment: results from fluctuations in aggregate demand over the business cycle
- During recessions, falling demand leads to layoffs and reduced hiring, increasing unemployment

Job search models

- Job search theory models the behavior of unemployed workers as they search for employment
- Key elements of job search models include:
 - The distribution of wage offers
 - The cost of searching for a job (time, effort, foregone earnings)
 - The expected duration of unemployment
- Optimal search strategy: workers set a reservation wage (the lowest wage they will accept) and continue searching until they receive a wage offer above that threshold
- Factors affecting job search:
 - Unemployment benefits: higher benefits may increase the reservation wage and prolong unemployment duration
 - Labor market conditions: a tighter labor market (more job openings relative to job seekers) may reduce search times and increase the likelihood of finding a high-wage job

Impact of labor market policies

- Governments implement various policies aimed at improving labor market outcomes, such as increasing employment, reducing inequality, and supporting low-income workers

Earned income tax credit (EITC)

- A refundable tax credit for low- and moderate-income working individuals and families
- Designed to encourage work and reduce poverty by increasing the returns to employment for low-wage workers

- The credit amount increases with earned income up to a maximum, then plateaus and eventually phases out at higher income levels
- Research suggests that the EITC has increased labor force participation among single mothers and reduced poverty rates

Unemployment insurance (UI) benefits

- Provides temporary income support to workers who have lost their jobs through no fault of their own
- Aims to smooth consumption and provide a safety net during job search
- Eligibility typically requires a minimum level of prior earnings and employment, and benefits are usually a fraction of previous wages (50-60% in most US states)
- Critics argue that UI benefits can prolong unemployment by reducing job search intensity and increasing the reservation wage
- The optimal design of UI balances the benefits of consumption smoothing against the potential disincentive effects on job search

Globalization and labor markets

- Globalization, the increasing integration of economies through trade, investment, and migration, has significant impacts on labor markets

Offshoring and outsourcing

- Offshoring refers to the relocation of production or service activities to another country, while outsourcing refers to the contracting out of activities to external providers (domestic or foreign)
- Offshoring and outsourcing can affect labor markets by:
 - Displacing workers in industries exposed to import competition or job relocation
 - Changing the relative demand for different types of skills (e.g., increasing demand for high-skilled workers in design and marketing, while reducing demand for low-skilled production workers)
- The impact of offshoring and outsourcing on employment and wages depends on factors such as:
 - The magnitude and speed of job displacement
 - The ability of affected workers to transition to new industries or occupations
 - The overall competitiveness and productivity of the economy

Immigration and labor supply

- Immigration increases the supply of labor in the receiving country, potentially affecting wages and employment for native-born workers
- The impact of immigration on native wages and employment depends on:
 - The skill composition of immigrants relative to the native population
 - The degree of substitutability between immigrant and native labor
- The response of firms to the increased labor supply (e.g., expanding production, adopting new technologies)

- Studies on the impact of immigration on native wages and employment have found mixed results, with most estimates suggesting small negative effects on average, but larger effects on specific skill groups (e.g., high school dropouts)
- In the long run, immigration can also have positive effects on the economy through increased innovation, entrepreneurship, and labor market flexibility

11.2 Education and social programs

Education programs can significantly impact student outcomes, but isolating causal effects requires careful research design. Randomized controlled trials are the gold standard, while quasi-experimental methods offer alternatives when randomization isn't feasible.

Understanding mechanisms linking education to outcomes is crucial. Non-cognitive skills, peer effects, and social networks play important roles. Exploring heterogeneous treatment effects can reveal which interventions work best for specific student populations, guiding targeted policy decisions.

Causal effects of education programs

- Analyzing how educational interventions and policies impact student outcomes requires careful research designs to isolate causal effects
- Education programs can affect a wide range of outcomes including academic achievement, social-emotional development, and long-term life success
- Establishing causal links between education and outcomes is crucial for guiding evidence-based policy decisions and resource allocation in schools

Randomized controlled trials in education

- RCTs are considered the gold standard for causal inference, randomly assigning students or schools to treatment and control groups
- In education, RCTs can test the effectiveness of interventions like curricula, teaching practices, technology, or school resources
- Well-designed RCTs can provide unbiased estimates of the average treatment effect of an educational program or policy

Overcoming challenges of RCTs in schools

- Implementing RCTs in schools often faces logistical and ethical challenges around random assignment and consent
- Ensuring compliance with treatment assignment and minimizing attrition is important for preserving the benefits of randomization
- Conducting RCTs at a large scale across many schools and districts can enhance the generalizability of the findings

Interpreting education RCT results

- RCTs estimate the average treatment effect, but there may be important variation in impacts across student subgroups

- The choice of outcome measures in education RCTs matters for understanding the scope and magnitude of program impacts
- Statistical power, multiple hypothesis testing, and other issues need to be considered when interpreting the significance of education RCT results

Quasi-experimental designs for education research

- When RCTs are not feasible, quasi-experimental methods can be used to estimate causal effects by leveraging natural experiments or policy changes
- Quasi-experimental designs rely on finding a convincing control group that can serve as a counterfactual for what would have happened without the intervention
- Careful consideration of the identifying assumptions and potential threats to internal validity is essential for quasi-experimental education research

Difference-in-differences for policy changes

- Difference-in-differences (DD) methods can be used to study the impacts of education policies that are implemented at a specific point in time for some schools or regions
- DD compares the change in outcomes for the treated group before and after the policy to the change over the same period for a control group
- Parallel trends between the treatment and control groups in the pre-period is a key assumption for DD to provide causal estimates

Regression discontinuity in education

- Regression discontinuity (RD) designs can be used when there is a cutoff or threshold that determines assignment to an educational treatment
- RD estimates the causal effect by comparing outcomes for students just above and below the assignment cutoff, who should be similar except for treatment status
- RD has been used to study the effects of school entry age cutoffs, exam score thresholds for program eligibility, and school quality rankings

Propensity score matching for education programs

- Propensity score matching (PSM) attempts to create a convincing control group by matching treated students to untreated students with similar predicted probabilities of treatment
- PSM can adjust for selection bias due to observable characteristics, but cannot account for unmeasured confounders
- PSM has been used to estimate the effects of charter schools, gifted and talented programs, and financial aid on student outcomes

Mechanisms linking education to outcomes

- Beyond estimating the overall causal effect of education, it is important to understand the mechanisms or channels through which education influences life outcomes
- Potential mediating pathways between education and outcomes include the development of cognitive and non-cognitive skills, peer networks, and access to opportunities

- Identifying the key mechanisms can inform the design of more targeted and effective educational interventions

Education's impact on earnings and employment

- Education can increase human capital and productivity, leading to higher individual earnings and aggregate economic growth
- Additional years of schooling and higher degree attainment are associated with increased employment rates and job stability
- The returns to education in the labor market can vary across fields of study, institutions, and local economic conditions

Non-cognitive skill development from schooling

- In addition to academic knowledge, education can foster non-cognitive or socioemotional skills such as motivation, self-regulation, and social skills
- Non-cognitive skills are increasingly recognized as important determinants of success in school, work, and life
- Teaching practices, school climate, and extracurricular activities are some of the ways that schools can intentionally develop students' non-cognitive skills

Peer effects and social networks in schools

- Students' peers in school can influence their academic achievement, behavior, and attitudes through direct interactions, norm-setting, and information sharing
- Peer effects operate through multiple channels, including classroom disruptions, study partnerships, and role model inspiration
- The structure of social networks within and between schools can shape the flow of peer influences and the persistence of educational inequality

Heterogeneous treatment effects of education

- The impacts of education programs and policies may differ across student subgroups defined by demographic, socioeconomic, or academic characteristics
- Exploring heterogeneous treatment effects can identify programs that are particularly effective for disadvantaged or underserved student populations
- Understanding variation in educational impacts can guide the targeting and customization of interventions to the students who need them most

Variation by student characteristics

- Student characteristics such as race, ethnicity, gender, and family income may moderate the effects of educational interventions
- For example, the Tennessee STAR class size experiment found larger test score gains for minority and low-income students
- Analyzing impacts by English learner status, special education needs, and prior academic performance can also reveal important subgroup differences

Differences by school and teacher quality

- The effectiveness of education programs may depend on the quality of the schools and teachers implementing them
- Students in lower-performing schools or with less experienced teachers may have more room for improvement and thus larger treatment effects
- School contextual factors such as resources, leadership, and climate can influence the success of educational interventions

Cost-benefit analysis of education programs

- To inform education policy decisions, it is important to compare the costs and benefits of different programs and interventions
- Cost-benefit analysis quantifies the direct and indirect costs of implementing a program and estimates the monetary value of its impacts on outcomes
- Results of cost-benefit analysis can be expressed as net benefits, benefit-cost ratios, or returns on investment

Calculating costs and monetizing benefits

- The costs of education programs include personnel, facilities, materials, and training, as well as opportunity costs of student and teacher time
- Some benefits of education, such as higher earnings, can be readily monetized using data on labor market returns to schooling
- Other benefits, such as improved health or reduced crime, may require more assumptions and projections to translate into dollar terms

Comparing education policies and programs

- Cost-benefit analysis allows for comparative evaluation of alternative educational investments and interventions
- For example, early childhood education programs like Perry Preschool have been found to have higher benefit-cost ratios than adolescent interventions like Job Corps
- Comparing the cost-effectiveness of different programs per unit gain in student outcomes can guide resource allocation decisions

Generalizing education research findings

- To inform evidence-based policy and practice, the results of education studies need to generalize beyond the specific contexts and samples in which they were conducted
- External validity refers to the extent to which causal effects estimated in one setting would hold in other populations, places, or times
- Several strategies can be used to assess and enhance the generalizability of education research, including replication, meta-analysis, and theory-building

External validity concerns

- The students, schools, and social conditions in any given study may not be representative of the larger population of interest
- Educational interventions may not scale or sustain as successfully when implemented by different personnel or in different resource environments
- The timeliness of education research is important as the effectiveness of interventions may depend on the prevailing policy, technology, and cultural contexts

Replicating education studies in new contexts

- Directly replicating education studies in new settings can test the robustness and transferability of the original findings
- Systematic replication efforts, such as the What Works Clearinghouse, assess the quality and quantity of evidence for education programs across multiple studies
- Replications with consistent results can strengthen confidence in the generalizability of the causal effects, while divergent results can spur further investigation into the sources of variation

11.3 Epidemiology and public health

Epidemiology is the study of disease patterns in populations. It provides crucial tools for understanding health issues, from identifying risk factors to evaluating interventions. This field forms the backbone of public health, guiding policies and strategies to improve population health.

Causal inference in epidemiology aims to determine if observed associations are truly causal. This involves applying criteria like Bradford Hill's, using counterfactual thinking, and employing methods like directed acyclic graphs. These approaches help researchers draw meaningful conclusions from complex health data.

Epidemiology fundamentals

- Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems
- Fundamental concepts in epidemiology provide a foundation for understanding how diseases and health conditions are distributed in populations and the factors that influence their occurrence
- Epidemiological methods are used to investigate the causes of diseases, evaluate the effectiveness of interventions, and inform public health policies and decision-making

Study designs in epidemiology

- Cohort studies follow a group of individuals over time to assess the incidence of the outcome of interest (prospective)
- Case-control studies compare individuals with the outcome of interest to those without it, looking back in time to assess exposure (retrospective)
- Cross-sectional studies assess the prevalence of an outcome and exposure at a single point in time
- Ecological studies compare populations rather than individuals, often using aggregate data

Measures of disease frequency

- Prevalence is the proportion of a population that has a disease or condition at a specific point in time
- Incidence is the rate of new cases of a disease or condition in a population over a specified period
- Incidence proportion (cumulative incidence) is the proportion of a population at risk that develops the disease over a specified period
- Incidence rate (person-time incidence rate) is the number of new cases per population at risk per unit of time

Measures of association

- Relative risk (risk ratio) compares the risk of an outcome in the exposed group to the risk in the unexposed group
- Odds ratio compares the odds of exposure in cases to the odds of exposure in controls
- Attributable risk (risk difference) is the difference in risk between the exposed and unexposed groups
- Population attributable risk is the proportion of cases in a population that can be attributed to a specific exposure

Bias and confounding

- Selection bias occurs when the study population does not accurately represent the target population (non-random sampling, loss to follow-up)
- Information bias arises from inaccurate measurement or classification of exposure or outcome (recall bias, misclassification)
- Confounding occurs when a third variable is associated with both the exposure and the outcome, distorting the true relationship
- Strategies to address confounding include randomization, restriction, matching, stratification, and adjustment in statistical analysis

Causal inference in epidemiology

- Causal inference is the process of determining whether an observed association between an exposure and an outcome represents a causal relationship
- Establishing causality requires consideration of multiple criteria and the use of various methodological approaches
- Causal inference is crucial for guiding public health interventions and policies aimed at preventing disease and promoting health

Bradford Hill criteria

- Strength of association: A strong association between exposure and outcome supports causality
- Consistency: The association is observed repeatedly in different populations and settings
- Specificity: The exposure is associated with a specific outcome or group of outcomes
- Temporality: The exposure precedes the outcome in time

- Biological gradient (dose-response): Increasing levels of exposure are associated with increasing risk of the outcome
- Plausibility: The association is consistent with existing biological and medical knowledge
- Coherence: The association is compatible with the natural history and biology of the disease
- Experiment: Experimental evidence, such as randomized controlled trials, supports the causal relationship
- Analogy: Similar exposures have been shown to cause similar outcomes

Counterfactuals and potential outcomes

- Counterfactuals are hypothetical scenarios that describe what would have happened to an individual if they had received a different exposure or treatment
- Potential outcomes are the outcomes that would be observed under different exposure or treatment conditions
- The causal effect is the difference between the potential outcomes under different exposures or treatments
- The fundamental problem of causal inference is that only one potential outcome is observed for each individual

Directed acyclic graphs (DAGs)

- DAGs are graphical representations of the causal relationships between variables
- Nodes represent variables, and directed edges (arrows) represent causal relationships
- DAGs help identify confounding, mediation, and collider bias
- They guide the selection of variables for adjustment in statistical analyses to estimate causal effects

Mediation analysis

- Mediation analysis assesses the extent to which the effect of an exposure on an outcome is mediated through an intermediate variable
- The total effect of an exposure on an outcome can be decomposed into a direct effect and an indirect effect (mediated through the intermediate variable)
- Methods for mediation analysis include the Baron and Kenny approach, the difference method, and the product method
- Mediation analysis can provide insights into the mechanisms through which exposures affect outcomes

Study design and analysis

- The choice of study design depends on the research question, the available resources, and the ethical considerations
- Different study designs have their strengths and limitations in terms of causal inference, generalizability, and efficiency

- Appropriate statistical methods should be used to analyze the data collected from epidemiological studies

Randomized controlled trials

- Randomized controlled trials (RCTs) are considered the gold standard for assessing the efficacy of interventions
- Participants are randomly allocated to the intervention or control group, minimizing confounding and selection bias
- RCTs can establish a causal relationship between the intervention and the outcome
- Limitations of RCTs include high costs, ethical constraints, and limited generalizability to real-world settings

Observational studies

- Observational studies are non-experimental studies where the investigator does not control the assignment of exposures
- They include cohort studies, case-control studies, cross-sectional studies, and ecological studies
- Observational studies are useful for investigating exposures that cannot be randomized (smoking, environmental factors)
- They are susceptible to confounding and bias, requiring careful design and analysis to minimize these issues

Cohort vs case-control studies

- Cohort studies are better suited for investigating rare exposures and multiple outcomes
- Case-control studies are more efficient for studying rare outcomes and can investigate multiple exposures
- Cohort studies allow for the calculation of incidence rates and risk ratios, while case-control studies estimate odds ratios
- Case-control studies are more prone to selection and recall bias compared to cohort studies

Dealing with confounding variables

- Confounding can be addressed through study design (randomization, restriction, matching) or statistical analysis (stratification, adjustment)
- Stratification involves analyzing the association between exposure and outcome within homogeneous subgroups of the confounding variable
- Multivariable regression models can adjust for multiple confounding variables simultaneously
- Propensity score methods (matching, stratification, weighting) can balance the distribution of confounding variables between exposed and unexposed groups

Public health applications

- Epidemiological methods are applied to various public health issues to inform prevention, control, and policy decisions
- Public health applications of epidemiology include disease surveillance, outbreak investigation, screening and diagnostic tests, and vaccine effectiveness studies

Screening and diagnostic tests

- Screening tests are used to identify individuals with asymptomatic disease or risk factors in a population
- Diagnostic tests are used to confirm the presence of a disease in individuals with signs or symptoms
- The performance of screening and diagnostic tests is evaluated using measures such as sensitivity, specificity, predictive values, and likelihood ratios
- The choice of a screening or diagnostic test depends on the prevalence of the disease, the consequences of false-positive and false-negative results, and the available resources

Vaccines and herd immunity

- Vaccines are a primary prevention strategy that protect individuals from infectious diseases by inducing immunity
- Herd immunity refers to the indirect protection of unvaccinated individuals when a large proportion of the population is vaccinated
- The effectiveness of vaccines can be assessed through observational studies (cohort or case-control) or randomized controlled trials
- Vaccine safety is monitored through post-marketing surveillance systems (passive reporting, active surveillance)

Outbreak investigation steps

- Verify the diagnosis and confirm the outbreak
- Establish a case definition and identify cases
- Describe the outbreak in terms of person, place, and time
- Develop and test hypotheses about the source and mode of transmission
- Implement control and prevention measures
- Communicate findings and recommendations to stakeholders

Disease surveillance systems

- Disease surveillance involves the ongoing, systematic collection, analysis, and interpretation of health data for public health action
- Passive surveillance relies on health care providers, laboratories, or other sources to report cases to public health authorities

- Active surveillance involves public health officials proactively seeking out cases through regular contact with health care providers or other sources
- Syndromic surveillance uses data on symptoms or other indicators to detect outbreaks or unusual health events
- Surveillance data are used to monitor disease trends, detect outbreaks, evaluate control measures, and allocate resources

Ethics in epidemiology

- Ethical principles and guidelines are essential for conducting epidemiological research and public health practice
- Key ethical considerations in epidemiology include informed consent, privacy and confidentiality, balancing risks and benefits, and addressing conflicts of interest
- Special attention should be given to the protection of vulnerable populations, such as children, pregnant women, and marginalized groups

Informed consent and privacy

- Informed consent is the process by which participants voluntarily agree to take part in a study after being informed of the purpose, procedures, risks, and benefits
- Informed consent should be obtained from participants or their legally authorized representatives before enrolling in a study
- Privacy and confidentiality of participants' personal and health information should be protected through secure data storage and management
- Exceptions to informed consent may be granted in certain circumstances (minimal risk studies, public health emergencies)

Balancing risks and benefits

- Epidemiological studies should have a favorable balance of risks and benefits for participants and society
- Risks may include physical, psychological, social, or economic harms associated with participation in the study or the use of the findings
- Benefits may include direct benefits to participants (access to care or interventions) or indirect benefits to society (knowledge gained, public health improvements)
- The level of risk should be minimized and proportional to the anticipated benefits

Vulnerable population considerations

- Vulnerable populations are groups of individuals who may have a limited capacity to provide informed consent or are at increased risk of harm or exploitation
- Examples of vulnerable populations include children, pregnant women, prisoners, mentally ill individuals, and economically or educationally disadvantaged persons
- Special protections and considerations are required when involving vulnerable populations in epidemiological studies

- Researchers should ensure that the participation of vulnerable populations is equitable and that their rights and welfare are protected

Conflicts of interest

- Conflicts of interest occur when professional judgment or actions regarding a primary interest (validity of research, welfare of participants) may be unduly influenced by a secondary interest (financial gain, career advancement)
- Researchers should disclose any potential conflicts of interest to participants, research ethics committees, and publishers
- Institutions should have policies and procedures in place to manage and mitigate conflicts of interest
- Transparency and independence in the design, conduct, and reporting of epidemiological studies are essential to maintain public trust and credibility

11.4 Marketing and advertising

Marketing experiments are crucial for businesses to make data-driven decisions and optimize campaigns. By testing different strategies and measuring their impact, marketers can improve ROI and provide causal evidence for the effectiveness of their interventions.

However, challenges like selection bias and confounding variables can complicate causal inference in marketing. Randomized controlled trials and quasi-experimental designs help overcome these challenges, allowing marketers to estimate the true impact of their efforts and make informed decisions.

Importance of marketing experiments

- Marketing experiments enable businesses to make data-driven decisions by testing different strategies and measuring their impact on key metrics (conversion rates, customer lifetime value)
- Conducting rigorous experiments helps marketers optimize their campaigns, allocate resources efficiently, and improve overall marketing ROI
- Experiments provide causal evidence for the effectiveness of marketing interventions, allowing marketers to distinguish between correlation and causation

Challenges in marketing causal inference

Selection bias in marketing

- Selection bias occurs when the assignment of customers to different marketing treatments is not random, leading to systematic differences between treatment and control groups
- For example, if a company targets a promotional offer only to high-value customers, the observed differences in outcomes may be due to pre-existing differences between the groups rather than the offer itself
- Selection bias can lead to incorrect conclusions about the causal impact of marketing interventions, as the observed effects may be confounded by other factors

Confounding variables for advertising

- Confounding variables are factors that influence both the assignment of customers to advertising treatments and the outcomes of interest, making it difficult to isolate the causal effect of the advertisement
- Examples of confounding variables in advertising include seasonality (holiday periods may drive both ad exposure and purchases), customer demographics (age, income), and prior purchase history
- Failing to account for confounding variables can lead to biased estimates of advertising effectiveness and suboptimal budget allocation decisions

Randomized controlled trials (RCTs)

Benefits of RCTs for marketing

- RCTs are the gold standard for causal inference, as they ensure that treatment and control groups are balanced on both observed and unobserved characteristics
- By randomly assigning customers to different marketing interventions (ad creatives, pricing strategies), RCTs eliminate selection bias and confounding, allowing for unbiased estimates of causal effects
- RCTs enable marketers to test multiple hypotheses simultaneously, compare the effectiveness of different treatments, and make evidence-based decisions to optimize their campaigns

Limitations of RCTs in advertising

- Conducting RCTs can be expensive and time-consuming, especially for large-scale advertising campaigns across multiple channels and platforms
- RCTs may not always be feasible due to technical limitations (inability to randomize ad exposure on certain platforms) or ethical concerns (withholding potentially beneficial treatments from some customers)
- The external validity of RCT results may be limited, as the findings from a specific experiment may not generalize to different contexts, customer segments, or time periods

Quasi-experimental designs

Difference-in-differences for marketing

- Difference-in-differences (DID) is a quasi-experimental approach that compares the changes in outcomes over time between a treatment group and a control group
- DID can be used to evaluate the impact of marketing interventions (new product launches, pricing changes) by comparing the pre- and post-intervention trends between exposed and unexposed customer segments
- The key assumption of DID is that the treatment and control groups would have followed parallel trends in the absence of the intervention, which can be assessed by examining pre-intervention data

Regression discontinuity in advertising

- Regression discontinuity (RD) designs exploit sharp cutoffs in treatment assignment to estimate causal effects
- In advertising, RD can be applied when ad exposure is determined by a threshold on a continuous variable (age, income, or a targeting score)
- By comparing outcomes for customers just above and below the cutoff, RD provides a local estimate of the causal effect of the advertisement, assuming that customers near the threshold are similar on other characteristics

Propensity score matching for campaigns

- Propensity score matching (PSM) is a method for creating balanced treatment and control groups by matching customers based on their estimated probability of receiving the treatment (propensity score)
- In marketing campaigns, PSM can be used to match customers who were exposed to a campaign with similar unexposed customers, based on observable characteristics (demographics, past purchases)
- PSM helps reduce selection bias and confounding, but its effectiveness depends on the quality of the matching variables and the assumption that there are no unobserved confounders

Uplift modeling

Treatment vs control groups

- Uplift modeling aims to estimate the incremental impact of a marketing treatment on individual customers, rather than the average effect across the entire population
- Treatment groups are the customers who receive the marketing intervention (targeted advertisement, promotional offer), while control groups are the customers who do not receive the treatment
- By comparing the outcomes (purchases, revenue) between treatment and control groups, uplift models can identify the customers who are most likely to respond positively to the marketing intervention

Estimating incremental impact

- Incremental impact refers to the additional value generated by a marketing treatment, over and above what would have happened in the absence of the treatment
- Uplift models estimate incremental impact by predicting the difference in outcomes between the treatment and control scenarios for each individual customer
- Accurate estimation of incremental impact enables marketers to target their interventions to the customers with the highest expected uplift, maximizing the return on marketing investment

Measuring advertising effectiveness

Conversion rates and attribution

- Conversion rates measure the proportion of customers who take a desired action (purchase, sign-up) after being exposed to an advertisement

- Attribution models assign credit for conversions to different touchpoints in the customer journey, such as ad clicks, website visits, and email interactions
- Multi-touch attribution models (time decay, position-based) attempt to distribute credit across multiple touchpoints, while single-touch models (first-click, last-click) assign all credit to a single touchpoint

Long-term effects of advertising

- While conversion rates and attribution focus on short-term responses to advertising, it is important to consider the long-term effects of advertising on brand awareness, customer loyalty, and lifetime value
- Long-term effects can be measured through brand tracking studies, customer surveys, and cohort analyses that compare the behavior of customers exposed to advertising over an extended period
- Incorporating long-term metrics into advertising effectiveness measurement helps marketers optimize their campaigns for sustainable growth and profitability

Ethics of marketing experiments

Informed consent and data privacy

- Marketing experiments often involve collecting and analyzing customer data, raising concerns about informed consent and data privacy
- Marketers should provide clear information about what data is being collected, how it will be used, and obtain explicit consent from customers before conducting experiments
- Ensuring data privacy involves implementing appropriate security measures, anonymizing customer data, and complying with relevant regulations (GDPR, CCPA)

Fairness across customer segments

- Marketing experiments should be designed and conducted in a way that ensures fairness and non-discrimination across different customer segments
- Biased targeting or differential treatment based on sensitive attributes (race, gender, age) can lead to unethical and potentially illegal outcomes
- Marketers should regularly audit their experiments and algorithms for fairness, transparency, and accountability, and take steps to mitigate any identified biases or disparities

Unit 12 – Causal Inference: Limitations and Frontiers

12.1 Unmeasured confounding

Unmeasured confounding is a major hurdle in causal inference. It occurs when variables affecting both treatment and outcome are left out of the analysis, leading to biased estimates and incorrect conclusions about relationships between variables.

Several methods have been developed to detect and address unmeasured confounding, but it remains a difficult problem in observational studies. Understanding these methods is crucial for researchers aiming to draw accurate causal conclusions from non-experimental data.

Unmeasured confounding

- Unmeasured confounding is a major challenge in causal inference that occurs when there are variables that affect both the treatment and outcome but are not included in the analysis
- Failing to account for unmeasured confounding can lead to biased estimates of causal effects and incorrect conclusions about the relationship between variables
- Several methods have been developed to detect and address unmeasured confounding, but it remains a difficult problem in observational studies

Definition of unmeasured confounding

- Unmeasured confounding occurs when a variable that is not included in the analysis affects both the treatment and outcome variables
- This unmeasured variable is called a confounder because it confounds the relationship between the treatment and outcome
- Unmeasured confounders can be either observed variables that were not included in the analysis or unobserved variables that cannot be directly measured

Threats to causal inference

- Unmeasured confounding poses a serious threat to causal inference because it can lead to biased estimates of causal effects
- When unmeasured confounders are present, the observed association between the treatment and outcome may not reflect the true causal relationship
- Unmeasured confounding can cause both overestimation and underestimation of causal effects, depending on the direction and magnitude of the confounding

Bias from unmeasured confounding

- Bias from unmeasured confounding occurs when the unmeasured confounder is associated with both the treatment and outcome, leading to a spurious association
- The magnitude and direction of the bias depend on the strength of the associations between the unmeasured confounder, treatment, and outcome

- Positive confounding occurs when the unmeasured confounder is positively associated with both the treatment and outcome, leading to an overestimation of the causal effect
- Negative confounding occurs when the unmeasured confounder is negatively associated with either the treatment or outcome, leading to an underestimation of the causal effect

Detecting unmeasured confounding

- Detecting unmeasured confounding is challenging because the confounders are not directly observed
- One approach is to use sensitivity analyses to assess how robust the causal estimates are to potential unmeasured confounders
- Another approach is to use negative controls, which are variables that are not expected to be affected by the treatment but may be affected by unmeasured confounders
- Instrumental variables can also be used to detect unmeasured confounding by identifying variables that are associated with the treatment but not the outcome

Sensitivity analysis for unmeasured confounding

- Sensitivity analysis is a method for assessing how robust causal estimates are to potential unmeasured confounders
- The basic idea is to hypothesize the existence of an unmeasured confounder and then calculate how strong the associations between the confounder, treatment, and outcome would need to be to explain away the observed effect
- Sensitivity analyses can provide a range of plausible causal effects under different assumptions about the strength of unmeasured confounding
- However, sensitivity analyses cannot definitively prove the absence of unmeasured confounding

Bounds on causal effects

- In the presence of unmeasured confounding, it may not be possible to obtain a precise estimate of the causal effect
- However, it is sometimes possible to derive bounds on the causal effect, which provide a range of possible values consistent with the observed data
- Bounds can be derived using various methods, such as instrumental variables, sensitivity analyses, and linear programming
- While bounds may not provide a definitive answer, they can still be informative for decision-making and policy evaluation

Instrumental variables for unmeasured confounding

- Instrumental variables (IVs) are a method for addressing unmeasured confounding by identifying variables that are associated with the treatment but not the outcome
- A valid IV must satisfy three assumptions: (1) it must be associated with the treatment, (2) it must not be associated with the outcome except through its effect on the treatment, and (3) it must not be associated with any unmeasured confounders

- When these assumptions are met, the IV can be used to estimate the causal effect of the treatment on the outcome, even in the presence of unmeasured confounding
- Examples of IVs include random assignment in experiments, natural experiments, and genetic variants in Mendelian randomization studies

Negative controls for unmeasured confounding

- Negative controls are variables that are not expected to be affected by the treatment but may be affected by unmeasured confounders
- The idea is that if an association is observed between the treatment and a negative control outcome, it suggests the presence of unmeasured confounding
- Negative controls can be used to assess the plausibility of causal claims and to detect potential sources of bias
- Examples of negative controls include outcomes that occur before the treatment (pre-treatment outcomes) and outcomes that are not plausibly related to the treatment (placebo outcomes)

Difference-in-differences for unmeasured confounding

- Difference-in-differences (DID) is a method for addressing unmeasured confounding by comparing changes over time in a treatment group and a control group
- The basic idea is that if the treatment and control groups have similar trends before the treatment, any differences in trends after the treatment can be attributed to the causal effect of the treatment
- DID relies on the assumption that the unmeasured confounders affect both groups similarly over time (parallel trends assumption)
- DID can be used in observational studies when there is a natural experiment or policy change that affects one group but not the other

Panel data methods for unmeasured confounding

- Panel data methods are a class of methods for addressing unmeasured confounding by using repeated observations of the same units over time
- Fixed effects models control for all time-invariant unmeasured confounders by including unit-specific intercepts
- Random effects models allow for time-varying unmeasured confounders by including unit-specific random intercepts and slopes
- Panel data methods rely on the assumption that the unmeasured confounders are either time-invariant or have a specific distribution over time

Simulation studies of unmeasured confounding

- Simulation studies are a useful tool for investigating the impact of unmeasured confounding on causal estimates
- By generating data with known causal relationships and unmeasured confounders, researchers can assess the performance of different methods for addressing unmeasured confounding

- Simulation studies can help to identify the conditions under which different methods are likely to succeed or fail
- However, simulation studies are limited by the assumptions and parameters used to generate the data, and may not fully capture the complexity of real-world settings

Famous examples of unmeasured confounding

- One famous example of unmeasured confounding is the association between hormone replacement therapy (HRT) and cardiovascular disease in postmenopausal women
- Observational studies suggested that HRT reduced the risk of cardiovascular disease, but randomized trials later showed that HRT actually increased the risk
- The discrepancy was likely due to unmeasured confounding by factors such as socioeconomic status and health behaviors, which were associated with both HRT use and cardiovascular risk
- Another example is the association between smoking and lung cancer, which was initially dismissed as confounded by factors such as air pollution and occupational exposures

Strategies to minimize unmeasured confounding

- While it is impossible to completely eliminate unmeasured confounding, there are strategies that can help to minimize its impact
- One strategy is to carefully select the variables to include in the analysis, based on subject matter knowledge and prior research
- Another strategy is to use matching or weighting methods to balance the distribution of observed confounders between the treatment and control groups
- Sensitivity analyses can also be used to assess the robustness of the results to potential unmeasured confounders
- Finally, randomization can be used to ensure that unmeasured confounders are balanced between the treatment and control groups

Randomization vs observational studies

- Randomization is the gold standard for causal inference because it ensures that the treatment is independent of all potential confounders, both measured and unmeasured
- In a randomized experiment, the treatment is assigned randomly to units, so any differences in outcomes between the treatment and control groups can be attributed to the causal effect of the treatment
- Observational studies, on the other hand, rely on naturally occurring variation in the treatment and are more susceptible to unmeasured confounding
- While observational studies can be useful for generating hypotheses and studying rare outcomes or exposures, they cannot definitively establish causal relationships

Challenges of addressing unmeasured confounding

- Despite the many methods available for addressing unmeasured confounding, it remains a challenging problem in causal inference

- One challenge is that the assumptions required for different methods (e.g., instrumental variables, parallel trends) may not hold in practice
- Another challenge is that the results of different methods may not always agree, leading to uncertainty about the true causal effect
- Unmeasured confounding can also interact with other sources of bias, such as selection bias and measurement error, making it difficult to isolate its impact
- Finally, the choice of method for addressing unmeasured confounding often depends on the specific research question, data, and assumptions, requiring careful consideration and justification

12.2 External validity and generalizability

External validity is crucial in causal inference, determining whether study results apply to real-world situations. It's about generalizing findings beyond the specific study sample to broader populations or settings. This concept is key for making research meaningful and applicable.

Assessing external validity involves evaluating sample representativeness, experimental setting realism, and potential interactions between causal effects and sample or setting characteristics. Improving it requires strategies like examining heterogeneous treatment effects, using stratified sampling, and replicating studies in diverse contexts.

External validity

- Refers to the extent to which the results of a study can be generalized to other populations, settings, or contexts beyond the specific study sample and conditions
- A crucial consideration in causal inference as it determines whether the causal effects estimated in a study are applicable and relevant to real-world situations and decision-making
- Closely related to the concept of generalizability, which assesses whether the findings from a particular study hold true for a broader population or different circumstances

Generalizability of causal effects

- The degree to which the causal effects estimated in a study can be extended to a larger population or different settings
- Depends on factors such as the representativeness of the study sample, the similarity of the study conditions to real-world contexts, and the stability of the causal relationship across different subgroups or environments
- Example: A study on the effectiveness of a new educational intervention conducted in a specific school district (Los Angeles Unified School District) may have limited generalizability to other districts with different student demographics or resources

Target population vs study sample

- The target population refers to the entire group of individuals, units, or entities to which the researcher aims to generalize the study findings
- The study sample is the subset of the target population that is selected and included in the actual study

- Discrepancies between the characteristics of the study sample and the target population can limit the external validity of the findings
- Example: A study on the effects of a new drug conducted on a sample of healthy young adults (18-25 years old) may not be directly generalizable to the target population of all adults, including older individuals or those with pre-existing health conditions

Threats to external validity

- Factors that can limit the generalizability of study findings to other populations or settings
- Selection bias: When the study sample is not representative of the target population due to non-random selection or self-selection
- Interaction effects: When the causal effect varies depending on specific characteristics of the sample or setting, limiting generalizability
- Hawthorne effect: When participants' behavior is influenced by their awareness of being observed or studied, leading to results that may not generalize to real-world settings
- Example: A study on the effectiveness of a new teaching method conducted in a highly controlled laboratory setting may not generalize to real classrooms with more diverse student populations and less controlled environments

Assessing external validity

- Involves evaluating the extent to which the results of a study can be generalized beyond the specific study sample and setting
- Requires careful consideration of factors such as the representativeness of the sample, the realism of the experimental setting, and potential interactions between the causal effect and sample characteristics or settings
- Helps determine the applicability and relevance of study findings to real-world contexts and decision-making

Representativeness of sample

- The degree to which the study sample accurately reflects the characteristics and diversity of the target population
- Assessing representativeness involves comparing the demographic, socioeconomic, or other relevant characteristics of the sample to those of the target population
- Non-representative samples can limit the generalizability of findings, as the causal effects may not hold for subgroups that are underrepresented or excluded from the study
- Example: A study on the effectiveness of a new weight loss program conducted on a sample of primarily white, middle-class participants may not be representative of the broader population, which includes individuals from diverse racial, ethnic, and socioeconomic backgrounds

Realism of experimental setting

- The extent to which the study conditions and environment resemble real-world contexts
- Highly controlled or artificial settings may limit the external validity of findings, as the causal effects may not generalize to more naturalistic environments

- Assessing realism involves considering factors such as the presence of real-world constraints, the level of control over extraneous variables, and the similarity of the study tasks or interventions to real-life situations
- Example: A study on the effects of a new educational software conducted in a quiet, well-equipped computer lab may not generalize to real classrooms with more distractions, varying levels of technology access, and diverse student needs

Interaction of causal effect with sample characteristics

- The possibility that the causal effect may vary depending on specific characteristics of the study sample, such as age, gender, education level, or baseline risk factors
- Assessing interaction effects involves examining whether the causal relationship is consistent across different subgroups within the sample
- If the causal effect interacts with sample characteristics, the findings may not generalize to populations with different characteristics
- Example: A study on the effectiveness of a new blood pressure medication may find that the treatment effect is stronger in older adults compared to younger adults, limiting the generalizability of the findings to all age groups

Interaction of causal effect with settings

- The possibility that the causal effect may vary depending on the specific features of the study setting, such as geographic location, cultural context, or organizational characteristics
- Assessing interaction effects with settings involves examining whether the causal relationship is consistent across different environments or contexts
- If the causal effect interacts with setting characteristics, the findings may not generalize to other settings with different features
- Example: A study on the impact of a new management strategy conducted in a large, urban corporation may not generalize to smaller, rural businesses or non-profit organizations due to differences in organizational structure, resources, and culture

Improving external validity

- Strategies and approaches aimed at enhancing the generalizability of study findings to broader populations and settings
- Involves designing studies that are more representative of the target population, conducting research in realistic settings, and exploring potential heterogeneity in treatment effects
- Helps increase the applicability and relevance of study results to real-world contexts and decision-making

Heterogeneous treatment effects

- The recognition that the causal effect of an intervention or treatment may vary across different subgroups within a population
- Examining heterogeneous treatment effects involves conducting subgroup analyses or using statistical methods (such as interaction terms or multilevel modeling) to identify potential moderators of

the causal relationship

- Understanding heterogeneity in treatment effects can help identify subpopulations for whom the intervention is more or less effective, informing targeted interventions and policy decisions
- Example: A study on the effectiveness of a new job training program may find that the program has stronger effects for individuals with lower levels of education compared to those with higher levels of education, suggesting the need for tailored interventions based on educational background

Stratified sampling

- A sampling technique that involves dividing the target population into distinct subgroups (strata) based on key characteristics and selecting a separate sample from each stratum
- Stratified sampling helps ensure that the study sample is representative of the target population in terms of important characteristics, such as age, gender, or socioeconomic status
- By using stratified sampling, researchers can improve the external validity of their findings by capturing the diversity and heterogeneity of the target population
- Example: A study on the prevalence of a specific health condition may use stratified sampling to select participants from different age groups and geographic regions, ensuring that the sample reflects the age and regional distribution of the target population

Replication in diverse samples and settings

- Conducting multiple studies with different samples and settings to assess the consistency and generalizability of findings
- Replication studies help determine whether the causal effects observed in one study hold true in other contexts and populations
- If the findings are consistent across diverse samples and settings, it provides stronger evidence for the external validity of the causal relationship
- Example: A study on the effectiveness of a new teaching method may be replicated in different schools, grade levels, and student populations to assess whether the positive effects on learning outcomes generalize across various educational contexts

Extrapolation and transportability

- Extrapolation refers to the process of extending the causal effects estimated in a study to a target population or setting that differs from the study sample or conditions
- Transportability is a related concept that focuses on the application of causal effects from one population or setting to another, often in the context of policy or intervention implementation
- Both extrapolation and transportability require careful consideration of the assumptions and conditions necessary for valid generalization of causal effects

Formalizing extrapolation of causal effects

- Developing mathematical frameworks and statistical methods to quantify and assess the validity of extrapolating causal effects from a study sample to a target population
- Formalizing extrapolation involves specifying the assumptions and conditions under which the causal effects can be generalized, such as the absence of unmeasured confounders or the stability of the

causal mechanism across populations

- Example: Pearl and Bareinboim (2014) proposed a formal framework for extrapolation based on graphical models and the concept of "transportability," which specifies the conditions under which causal effects can be transferred from one population to another

Assumptions for valid extrapolation

- The assumptions that need to be met for extrapolation of causal effects to be valid and unbiased
- Conditional ignorability: The assumption that the treatment assignment is independent of the potential outcomes given the observed covariates, implying that there are no unmeasured confounders
- Positivity: The assumption that each unit in the target population has a non-zero probability of receiving each level of the treatment, given their observed covariates
- Stable unit treatment value assumption (SUTVA): The assumption that the potential outcomes for each unit are not affected by the treatment assignments of other units, and that there is a single version of the treatment
- Example: When extrapolating the effects of a job training program from a study sample to a larger population, researchers need to assume that the baseline characteristics of the study participants are similar to those of the target population (conditional ignorability), that individuals in the target population have access to the training program (positivity), and that the training program is delivered consistently across participants (SUTVA)

Transportability of causal effects across populations

- The application of causal effects estimated in one population (the source population) to another population (the target population) that may differ in terms of characteristics or context
- Transportability requires assessing the similarity between the source and target populations in terms of factors that may modify the causal effect, such as age, gender, or socioeconomic status
- If the relevant effect modifiers are measured and accounted for, and the causal structure remains stable across populations, the causal effects may be transportable
- Example: A policymaker may wish to apply the findings from a study on the effectiveness of a crime prevention program conducted in one city to another city with different demographic and socioeconomic characteristics, assuming that the relevant effect modifiers (such as poverty rates or educational attainment) are similar across the two cities

Strategies for transportability

- Methods and approaches for assessing and improving the transportability of causal effects across populations or settings
- Subgroup analysis: Examining the causal effects within different subgroups of the source population to identify potential effect modifiers and assess the consistency of the causal relationship across subpopulations
- Sensitivity analysis: Evaluating the robustness of the causal effects to potential violations of the transportability assumptions, such as the presence of unmeasured effect modifiers or differences in the distribution of covariates between the source and target populations

- Data fusion: Combining data from multiple sources (such as the study sample and external population-level data) to improve the representativeness of the sample and the generalizability of the findings
- Example: When transporting the effects of a health intervention from a clinical trial to a real-world population, researchers may use subgroup analysis to examine whether the treatment effect varies by age or comorbidities, conduct sensitivity analyses to assess the impact of potential unmeasured confounders, and incorporate population-level data on disease prevalence and risk factors to improve the external validity of the findings

Limits of external validity

- The inherent challenges and constraints in generalizing causal effects from a study sample to broader populations or settings
- Recognizing the limits of external validity is crucial for making appropriate and cautious inferences based on study findings and for guiding future research and policy decisions

Fundamental problem of extrapolation

- The inherent uncertainty in generalizing causal effects from a study sample to a target population that differs in terms of characteristics or context
- The fundamental problem arises because the true causal effect in the target population is unobserved and can only be estimated based on assumptions and statistical models
- Even when the assumptions for valid extrapolation are met, there may still be residual uncertainty due to potential differences between the study sample and the target population that are not captured by the observed covariates
- Example: When extrapolating the effects of a new medical treatment from a clinical trial to a real-world patient population, there may be unmeasured differences in patient characteristics, adherence to treatment, or healthcare access that limit the generalizability of the findings

Irreducible uncertainty in generalization

- The recognition that there will always be some level of uncertainty when generalizing causal effects from one population or setting to another, even when using rigorous methods for extrapolation or transportability
- This irreducible uncertainty stems from the fact that no study can perfectly capture all the relevant factors that may influence the causal relationship in different contexts
- Researchers and policymakers need to acknowledge and communicate this uncertainty when making decisions based on study findings and be prepared to update their conclusions as new evidence emerges
- Example: When applying the results of a study on the effectiveness of a poverty alleviation program from one country to another, policymakers should recognize that differences in cultural norms, institutional structures, or macroeconomic conditions may lead to some irreducible uncertainty in the generalizability of the findings

Cautious interpretation of findings

- The need for researchers and consumers of research to exercise caution when interpreting and applying study findings to new populations or settings
- Cautious interpretation involves considering the limitations of the study design, the potential for unmeasured confounding or effect modification, and the extent to which the study sample and conditions differ from the target population or setting
- Researchers should transparently report the assumptions and limitations of their analyses, and consumers of research should critically evaluate the external validity of the findings before making decisions based on the results
- Example: When interpreting the results of a study on the impact of a new educational intervention on student achievement, policymakers should consider factors such as the characteristics of the study sample, the specific features of the intervention, and the educational context in which the study was conducted before deciding to scale up the intervention to other schools or districts

Scope of causal claims

- The importance of clearly defining and communicating the scope of the causal claims that can be made based on a study's findings
- The scope of causal claims should be limited to the populations, settings, and conditions under which the study was conducted, and any extrapolation beyond these boundaries should be made with appropriate caveats and uncertainty
- Researchers should avoid overgeneralizing their findings and be explicit about the limitations and potential boundary conditions of their conclusions
- Example: A study on the effectiveness of a new therapy for depression conducted on a sample of adult patients in a clinical setting should limit its causal claims to similar populations and settings, rather than making broad statements about the therapy's effectiveness for all individuals with depression or in non-clinical contexts

12.3 Causal inference with complex data structures

Causal inference with complex data structures poses unique challenges. From panel data to networks and text, researchers must navigate issues like unobserved confounding, selection bias, and interdependencies to establish cause-and-effect relationships.

Advanced techniques like difference-in-differences, vector autoregressive models, and causal machine learning help address these challenges. Sensitivity analyses and robustness checks are crucial for assessing the validity of causal estimates in complex data environments.

Causal inference challenges

- Causal inference aims to establish cause-and-effect relationships between variables, but complex data structures pose significant challenges
- Limitations in data availability, unobserved confounding factors, and selection bias can lead to biased estimates and incorrect conclusions about causal effects

Data limitations for causal inference

- Insufficient sample size reduces statistical power to detect causal effects
- Missing data on key variables can introduce bias and limit the ability to control for confounding factors
- Measurement error in treatment or outcome variables attenuates estimated causal effects
- Lack of exogenous variation in treatment assignment hinders causal identification

Unobserved confounding variables

- Confounding occurs when a third variable influences both the treatment and outcome, leading to spurious associations
- Unobserved confounders cannot be directly controlled for in the analysis, potentially biasing causal effect estimates
- Examples of unobserved confounders include genetic factors, individual preferences, or environmental conditions
- Instrumental variables, fixed effects models, and sensitivity analyses can help address unobserved confounding

Selection bias in observational data

- Selection bias arises when the treatment and control groups differ systematically on observed or unobserved characteristics
- Non-random selection into treatment can confound the causal effect estimate
- Common sources of selection bias include self-selection (individuals choosing their treatment status) and sample attrition (non-random dropout from the study)
- Propensity score matching, inverse probability weighting, and Heckman correction methods can mitigate selection bias

Causal inference with panel data

- Panel data, also known as longitudinal data, consists of repeated observations on the same units (individuals, firms, or countries) over time
- Panel data allows for controlling time-invariant unobserved confounders and estimating dynamic causal effects

Fixed effects vs random effects models

- Fixed effects models control for all time-invariant unobserved confounders by focusing on within-unit variation over time
- Random effects models assume that unobserved confounders are uncorrelated with the treatment and allow for estimating effects of time-invariant variables
- The choice between fixed and random effects depends on the assumptions about the unobserved confounders and the research question
- Hausman tests can help assess the appropriateness of fixed versus random effects specifications

Difference-in-differences estimation

- Difference-in-differences (DiD) compares the change in outcomes for a treatment group to the change in outcomes for a control group, before and after a policy intervention
- DiD controls for time-invariant unobserved confounders and common time trends affecting both groups
- Parallel trends assumption: treatment and control groups would have followed the same trend in the absence of the intervention
- DiD can be extended to multiple time periods, multiple treatment groups, and staggered treatment adoption

Event study designs

- Event studies examine the dynamic causal effects of a treatment or event over time, relative to a baseline period
- Event studies allow for estimating lead and lag effects, testing for pre-trends, and assessing the persistence of treatment effects
- The key identifying assumption is that the timing of the event is exogenous with respect to the outcome variable
- Event studies can be implemented using a distributed lag model or a non-parametric approach with dummy variables for each time period

Causal inference with time series data

- Time series data consists of observations on a single unit (e.g., a country or a financial asset) at regular intervals over time
- Causal inference with time series data requires accounting for temporal dependence, trends, and potential feedback effects

Granger causality tests

- Granger causality tests whether past values of one variable help predict future values of another variable, beyond the information contained in the past values of the latter variable
- Granger causality does not imply true causality, as it may be driven by unobserved confounding or reverse causality
- Vector autoregressive (VAR) models are commonly used to implement Granger causality tests
- Limitations of Granger causality include its sensitivity to the choice of lag length and the potential for spurious results due to common trends or unobserved confounders

Vector autoregressive models

- VAR models capture the dynamic relationships among multiple time series variables, allowing for feedback effects and lagged dependencies
- Each variable in a VAR is modeled as a function of its own past values and the past values of the other variables in the system

- Impulse response functions derived from VAR models show the dynamic causal effects of a shock to one variable on the other variables over time
- VAR models can be extended to include exogenous variables (VARX), cointegration relationships (VECM), or time-varying parameters (TVP-VAR)

Cointegration and error correction models

- Cointegration refers to the existence of a long-run equilibrium relationship among non-stationary time series variables
- Error correction models (ECMs) capture both the short-run dynamics and the long-run equilibrium relationship among cointegrated variables
- ECMs can be used to estimate the speed of adjustment towards the long-run equilibrium and to test for Granger causality in the presence of cointegration
- Johansen tests and the Engle-Granger two-step procedure are common methods for testing cointegration and estimating ECMs

Causal inference with network data

- Network data consists of nodes (actors) and edges (relationships or interactions) among them
- Causal inference with network data requires accounting for network structure, interdependencies, and potential spillover effects

Network formation and homophily

- Network formation processes, such as preferential attachment and triadic closure, can lead to non-random network structures
- Homophily, the tendency for similar nodes to form connections, can confound the estimation of peer effects and social influence
- Stochastic actor-oriented models (SAOMs) and exponential random graph models (ERGMs) can jointly model network formation and behavior dynamics
- Separating homophily from influence requires exploiting exogenous variation in network structure or using dynamic network models

Peer effects and social influence

- Peer effects refer to the causal impact of an individual's peers' characteristics or behaviors on their own outcomes
- Social influence occurs when an individual's behavior is affected by the behavior of their peers
- Identifying peer effects is challenging due to reflection (simultaneity), correlated effects (common shocks), and endogenous peer group formation
- Instrumental variables, network fixed effects, and quasi-experimental designs (e.g., natural experiments or randomized experiments) can help identify peer effects

Identification strategies for network effects

- Network randomization: Randomly assigning individuals to different network positions or randomly rewiring network ties
- Partial population experiments: Randomly treating a subset of nodes and measuring spillover effects on untreated nodes
- Network instruments: Using network structure (e.g., friends of friends) or exogenous shocks to network ties as instruments for peer effects
- Structural models: Specifying a model of network formation and behavior and estimating the model parameters using maximum likelihood or Bayesian methods

Causal inference with text data

- Text data, such as documents, social media posts, or survey responses, can serve as treatments, outcomes, or confounders in causal inference
- Causal inference with text data requires quantifying text features, addressing high-dimensionality, and accounting for selection and measurement issues

Text as treatments, outcomes, or confounders

- Text as treatments: Exposure to specific content (e.g., news articles, advertisements) can affect individual attitudes or behaviors
- Text as outcomes: Causal effects can manifest in changes in the content, sentiment, or style of text produced by individuals or groups
- Text as confounders: Unobserved factors (e.g., political ideology, personality traits) can influence both the treatment and the text outcomes, leading to confounding

Topic modeling for causal inference

- Topic models, such as Latent Dirichlet Allocation (LDA), can uncover latent themes or topics in a corpus of text documents
- Topic proportions or topic assignments can be used as treatments, outcomes, or control variables in causal inference models
- Structural topic models (STMs) incorporate document-level metadata (e.g., author attributes, time periods) to estimate the effect of covariates on topic prevalence
- Dynamic topic models (DTMs) capture the evolution of topics over time and can be used for causal inference with time series text data

Sentiment analysis and causal effects

- Sentiment analysis aims to extract the emotional content or opinion polarity from text data
- Sentiment scores or classifications can be used as outcomes in causal inference models, e.g., to estimate the effect of a policy on public opinion
- Lexicon-based methods rely on pre-defined dictionaries of positive and negative words, while machine learning methods train classifiers on labeled data

- Challenges in sentiment analysis for causal inference include dealing with negation, sarcasm, and domain-specific language

Causal inference with spatial data

- Spatial data contains information about the geographic location of units (e.g., individuals, households, or regions)
- Causal inference with spatial data requires accounting for spatial dependence, heterogeneity, and potential spillover effects

Spatial dependence and spillover effects

- Spatial dependence occurs when the outcomes of nearby units are more similar than those of distant units, due to common factors or interactions
- Spillover effects arise when the treatment of one unit affects the outcomes of neighboring units
- Ignoring spatial dependence can lead to biased and inefficient estimates, while ignoring spillover effects can underestimate the total impact of a policy
- Spatial lag models, spatial error models, and spatial Durbin models can incorporate spatial dependence and spillover effects

Spatial regression discontinuity designs

- Spatial regression discontinuity (RD) designs exploit geographic boundaries (e.g., state borders, school districts) as a source of exogenous variation in treatment assignment
- Units just on either side of the boundary are assumed to be similar in unobserved characteristics, but are exposed to different treatments
- Spatial RD designs can estimate the local average treatment effect (LATE) at the boundary, under the assumption of no spatial spillovers across the boundary
- Challenges in spatial RD include selecting appropriate bandwidths, testing for spatial sorting, and accounting for potential spatial spillovers

Geographically weighted regression

- Geographically weighted regression (GWR) allows for spatial heterogeneity in the relationship between the treatment and the outcome
- GWR estimates local regression coefficients for each spatial unit, using a distance-based weighting scheme
- The local coefficients can reveal spatial patterns in the causal effect of the treatment, e.g., identifying regions where the treatment is more or less effective
- Challenges in GWR include choosing the appropriate spatial kernel and bandwidth, and interpreting the local coefficients as causal effects

Machine learning for causal inference

- Machine learning methods can improve the estimation of causal effects by flexibly modeling the relationship between the treatment, covariates, and outcomes

- Causal machine learning methods combine the strengths of machine learning (prediction) and causal inference (identification) to estimate heterogeneous treatment effects and optimize treatment assignment

Causal trees and forests

- Causal trees recursively partition the covariate space to identify subgroups with different treatment effects
- Causal forests average the predictions of multiple causal trees to improve the stability and accuracy of the estimated treatment effects
- Honest causal trees use a split sample approach to avoid overfitting and obtain unbiased estimates of the treatment effects
- Causal forests can be used for personalized treatment assignment, by assigning individuals to the treatment with the highest estimated effect given their covariates

Double machine learning

- Double machine learning (DML) is a general framework for estimating causal effects using machine learning methods while maintaining the validity of statistical inference
- DML involves estimating the propensity score and the outcome model using machine learning algorithms (e.g., lasso, random forests) and obtaining the causal effect estimate through a final regression step
- The key idea is to use orthogonalized estimating equations to remove the bias due to model selection and regularization in the nuisance parameter estimation
- DML can be applied to a variety of causal inference settings, including treatment effect estimation, instrumental variables, and mediation analysis

Deep learning for causal effect estimation

- Deep learning models, such as neural networks, can learn complex non-linear relationships between the treatment, covariates, and outcomes
- Causal effect estimation with deep learning requires careful design to ensure the identification assumptions are met and the estimates are interpretable
- Techniques such as Targeted Maximum Likelihood Estimation (TMLE) and Adversarial Balancing can be combined with deep learning to estimate causal effects
- Challenges in deep learning for causal inference include sensitivity to hyperparameters, potential for overfitting, and the need for large sample sizes

Sensitivity analysis and robustness checks

- Sensitivity analysis assesses the robustness of causal effect estimates to potential violations of the identification assumptions
- Robustness checks involve estimating the causal effect using alternative methods, data sources, or specifications to ensure the consistency of the results

Assessing sensitivity to unobserved confounding

- Unobserved confounding is a major threat to the validity of causal effect estimates in observational studies
- Sensitivity analysis methods quantify the degree of unobserved confounding necessary to explain away the estimated causal effect
- Examples include the Rosenbaum bounds approach for binary treatments and the VanderWeele and Ding bias formulas for continuous treatments
- Sensitivity parameters can be varied to create worst-case scenarios and assess the plausibility of the unobserved confounding

Placebo tests and falsification strategies

- Placebo tests involve estimating the causal effect on an outcome that should not be affected by the treatment, to check for potential confounding or selection bias
- Falsification tests involve estimating the causal effect of the treatment on pre-treatment outcomes or covariates, which should be zero if the identification assumptions hold
- Placebo and falsification tests can help assess the credibility of the causal identification strategy and detect potential violations of the assumptions
- Challenges in placebo and falsification tests include finding suitable placebo outcomes or pre-treatment variables and interpreting the results

Replication with alternative data or methods

- Replication involves re-estimating the causal effect using a different dataset, a subset of the original data, or an alternative identification strategy
- Successful replication increases the credibility of the original findings and helps rule out potential sources of bias or model misspecification
- Replication can also involve cross-validation techniques, such as k-fold or leave-one-out cross-validation, to assess the stability of the causal effect estimates
- Challenges in replication include the availability of suitable alternative datasets, differences in variable definitions or measurement, and the comparability of the identification strategies

12.4 Machine learning for causal inference

Machine learning methods are revolutionizing causal inference by handling complex data and relationships. These techniques estimate treatment effects, identify subgroups, and model outcomes using algorithms like decision trees, neural networks, and regularized estimators.

While powerful, machine learning in causal inference has limitations. Researchers must consider assumptions like positivity and overlap, interpret black-box models carefully, and address computational challenges. Proper evaluation and understanding of these methods are crucial for reliable causal estimates.

Machine learning methods

- Machine learning methods are increasingly being used in causal inference to estimate treatment effects and identify heterogeneous subgroups
- These methods leverage the predictive power of machine learning algorithms to model the relationship between covariates, treatments, and outcomes
- Machine learning approaches can handle high-dimensional data and complex non-linear relationships, making them well-suited for causal inference tasks

Supervised vs unsupervised learning

- Supervised learning involves training a model on labeled data, where the desired output (treatment effect) is known during training (e.g., using data from randomized controlled trials)
- Unsupervised learning aims to discover patterns and structures in the data without explicit labels (e.g., clustering similar individuals based on their covariates)
- In causal inference, supervised learning is more commonly used to estimate treatment effects, while unsupervised learning can be used for exploratory analysis and identifying subgroups

Classification vs regression models

- Classification models are used when the outcome of interest is categorical (e.g., binary treatment assignment)
- Examples of classification algorithms include logistic regression, decision trees, and support vector machines
- Regression models are used when the outcome is continuous (e.g., measuring the magnitude of the treatment effect)
- Examples of regression algorithms include linear regression, random forests, and gradient boosting machines

Estimating treatment effects

- Treatment effect estimation is a central goal in causal inference, which aims to quantify the causal impact of an intervention on an outcome
- Machine learning methods can be used to estimate various types of treatment effects, such as the average treatment effect (ATE) and conditional average treatment effect (CATE)
- These methods often involve modeling the potential outcomes under different treatment conditions and estimating the difference between them

Average treatment effect (ATE)

- The ATE is the expected difference in outcomes between the treated and control groups across the entire population
- It represents the overall causal effect of the treatment, assuming that the treatment is applied to everyone in the population
- Machine learning methods can be used to estimate the ATE by modeling the potential outcomes and taking the average difference

Conditional average treatment effect (CATE)

- The CATE is the expected difference in outcomes between the treated and control groups, conditional on specific covariates
- It allows for the identification of heterogeneous treatment effects, where the impact of the treatment may vary across different subgroups
- Machine learning methods can estimate the CATE by modeling the relationship between covariates, treatment, and outcomes, and predicting the treatment effect for each individual

Targeted maximum likelihood estimation (TMLE)

- TMLE is a double-robust estimation method that combines machine learning with targeted estimation to obtain unbiased and efficient estimates of the ATE or CATE
- It involves fitting an initial machine learning model for the outcome and propensity score, and then updating the estimates using a targeted maximum likelihood step
- TMLE is particularly useful when there is potential confounding and the propensity score model may be misspecified

Causal trees and forests

- Causal trees and forests are machine learning methods specifically designed for estimating heterogeneous treatment effects
- They extend traditional decision trees and random forests to the causal inference setting, allowing for the identification of subgroups with different treatment effects
- These methods can handle high-dimensional data and automatically detect interactions between covariates and the treatment

Recursive partitioning for heterogeneous effects

- Causal trees use recursive partitioning to divide the covariate space into subgroups with different treatment effects
- The splitting criteria are designed to maximize the difference in treatment effects between the resulting subgroups
- The tree structure allows for easy interpretation and identification of important covariates driving the heterogeneity

Honest estimation with causal trees

- Honest estimation is a technique used in causal trees to ensure unbiased estimates of the treatment effects
- It involves splitting the data into two parts: one for building the tree structure and identifying the subgroups, and another for estimating the treatment effects within each subgroup
- This approach prevents overfitting and ensures that the estimates are reliable and generalizable

Ensemble methods and causal forests

- Causal forests are an ensemble method that combines multiple causal trees to improve the stability and accuracy of the treatment effect estimates
- They work by bootstrapping the data and fitting a causal tree on each bootstrap sample, and then averaging the predictions across the trees
- Causal forests can provide more robust estimates of the CATE and capture complex interactions between covariates and the treatment

Neural networks for causal inference

- Neural networks are powerful machine learning models that can learn complex non-linear relationships between variables
- They have been adapted for causal inference tasks, such as estimating treatment effects and adjusting for confounding
- Neural networks can handle high-dimensional data and automatically learn relevant features for causal inference

Feed-forward neural networks

- Feed-forward neural networks are the simplest type of neural networks, where information flows in one direction from the input layer to the output layer
- They can be used to estimate treatment effects by modeling the potential outcomes as a function of the covariates and treatment
- The network architecture can be designed to incorporate propensity scores or other causal assumptions

Representation learning for confounding adjustment

- Representation learning involves learning a low-dimensional representation of the covariates that captures the relevant information for causal inference
- Neural networks can be used to learn these representations, which can then be used to adjust for confounding and estimate treatment effects
- Examples include using autoencoders or variational autoencoders to learn a balanced representation of the covariates

Deep instrumental variable estimation

- Instrumental variable (IV) estimation is a method for estimating causal effects in the presence of unmeasured confounding
- Deep IV estimation uses neural networks to learn the relationship between the instrument, treatment, and outcome
- The learned representations can be used to estimate the causal effect of the treatment, even when there are unmeasured confounders

Causal regularization techniques

- Regularization techniques are used in machine learning to prevent overfitting and improve the generalization performance of models
- In causal inference, regularization can be used to encourage the estimation of stable and interpretable treatment effects
- Causal regularization techniques incorporate causal assumptions or constraints into the regularization term of the machine learning objective

Targeted regularization for ATE estimation

- Targeted regularization is a technique that incorporates the propensity score into the regularization term to improve the estimation of the ATE
- It encourages the model to focus on the regions of the covariate space where the treatment assignment is balanced
- Examples include targeted maximum likelihood estimation with lasso regularization

Causal effect estimation with lasso

- Lasso (least absolute shrinkage and selection operator) is a regularization technique that encourages sparse solutions by shrinking some coefficients to zero
- In causal inference, lasso can be used to estimate treatment effects while performing variable selection
- It can help identify the most important covariates for estimating the causal effect and improve the interpretability of the model

Doubly robust regularized estimators

- Doubly robust estimators combine outcome modeling and propensity score weighting to estimate treatment effects
- Regularized doubly robust estimators incorporate regularization techniques, such as lasso or ridge regression, into the outcome and propensity score models
- These estimators can provide more stable and efficient estimates of the treatment effects, especially when the models are misspecified

Evaluating machine learning estimators

- Evaluating the performance of machine learning estimators in causal inference is crucial for assessing their reliability and validity
- Several methods can be used to assess the bias, variance, and robustness of the estimated treatment effects
- These evaluation techniques help ensure that the machine learning methods are providing accurate and trustworthy causal estimates

Assessing bias and variance

- Bias refers to the systematic deviation of the estimated treatment effects from the true values
- Variance refers to the variability of the estimates across different samples or datasets
- Techniques such as simulation studies and sensitivity analyses can be used to assess the bias and variance of machine learning estimators

Cross-validation strategies for causal inference

- Cross-validation is a common technique in machine learning for assessing the generalization performance of models
- In causal inference, cross-validation needs to be adapted to account for the specific challenges, such as the presence of confounding and the need for honest estimation
- Strategies such as nested cross-validation and doubly robust cross-validation can be used to evaluate machine learning estimators in causal inference settings

Sensitivity analysis and robustness checks

- Sensitivity analysis involves assessing the robustness of the estimated treatment effects to violations of the causal assumptions (e.g., unobserved confounding)
- It can help identify the range of plausible treatment effects under different assumptions and assess the sensitivity of the estimates to model misspecification
- Robustness checks, such as testing for covariate balance and assessing the overlap assumption, can further validate the reliability of the machine learning estimators

Limitations and considerations

- While machine learning methods offer powerful tools for causal inference, it is essential to be aware of their limitations and the assumptions required for valid causal estimates
- Several considerations need to be taken into account when applying machine learning methods in causal inference settings
- Understanding these limitations and considerations is crucial for the appropriate use and interpretation of machine learning estimators

Positivity and overlap assumptions

- The positivity assumption requires that there is a non-zero probability of receiving each treatment level for all covariate values
- The overlap assumption requires that there is sufficient overlap in the covariate distributions between the treated and control groups
- Violations of these assumptions can lead to biased and unstable estimates of the treatment effects, even with machine learning methods

Interpreting black-box models causally

- Many machine learning models, such as deep neural networks, are considered "black-box" models, meaning that their internal workings are not easily interpretable
- Interpreting these models causally can be challenging, as it may be difficult to disentangle the causal effects from the complex interactions learned by the model
- Techniques such as permutation feature importance and partial dependence plots can be used to gain some insight into the causal relationships captured by the model

Computational complexity and scalability

- Machine learning methods, particularly those involving complex models like neural networks, can be computationally intensive and may not scale well to large datasets
- The computational complexity can be a limiting factor when applying these methods to real-world causal inference problems
- Strategies such as dimensionality reduction, parallelization, and efficient algorithms can help mitigate these challenges and improve the scalability of machine learning methods in causal inference