Chapter 10: Random Variability

Contents

Part I focused on causal inference in settings where we conceptualized study populations as effectively infinite, allowing us to ignore random variability and focus solely on systematic bias from confounding, selection, and measurement. Part II now introduces **random variability** and the use of **statistical models** for causal inference. This chapter bridges identification (Part I) and estimation (Part II), explaining why we need models and how to quantify uncertainty.

This chapter is based on Hernán and Robins (2020, chap. 10, pp. 131-142). It marks the transition from Part I (Causal inference without models) to Part II (Causal inference with models).

1 10.1 Identification Versus Estimation (pp. 131-134)

Up to now, we have focused on **identification**: determining whether causal effects can be computed from observed data under certain assumptions. Now we turn to **estimation**: using finite data to approximate those causal effects.

1.1 Key Concepts in Statistical Estimation

Estimand: The population parameter of interest (e.g., $Pr[Y = 1|A = a]$ in the super-population).

Estimator: A rule for computing the estimand from sample data.

Estimate: The numerical value obtained by applying the estimator to a particular sample (a point estimate).

> **Example 1.1** (Sample Proportion as an Estimator)**. Estimand**: Super-population risk $Pr[Y = 1|A = 1]$
> **Estimator**: Sample proportion $\widehat{Pr}[Y = 1|A = 1]$
> **Estimate**: From our 20-person study, $\widehat{Pr}[Y = 1|A = 1] = 7/13 \approx 0.54$

Identification vs. Estimation:

- **Identification** (Part I): Can we express the causal effect in terms of observable quantities?
  - Requires: Exchangeability, positivity, consistency
  - Answer: "Yes, the causal risk difference equals the associational risk difference under these conditions"
- **Estimation** (Part II): How do we estimate the causal effect from finite data?
  - Requires: Statistical methods to handle random variability
  - Answer: "Use sample statistics with appropriate confidence intervals"

1.2 Consistency of Estimators

An estimator is **consistent** if estimates get arbitrarily close to the true parameter as sample size increases.

$$Pr\left[|\hat{\theta}_n - \theta| > \epsilon\right] \to 0 \text{ as } n \to \infty, \text{ for all } \epsilon > 0$$

The sample proportion $\widehat{Pr}[Y = 1|A = a]$ is a consistent estimator of $Pr[Y = 1|A = a]$.

1.3 Confidence Intervals

A **95% confidence interval** quantifies uncertainty due to random sampling.

Construction (Wald interval):

1. Compute point estimate $\hat{p}$
2. Estimate standard error: $\widehat{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
3. Compute interval: $\hat{p} \pm 1.96 \times \widehat{SE}$

Example: For $\hat{p} = 7/13 \approx 0.54$ with $n = 13$:

- $\widehat{SE} = \sqrt{(7/13)(6/13)/13} = 0.138$
- 95% CI: $0.54 \pm 1.96(0.138) = (0.27, 0.81)$

Interpretation of 95% Confidence Intervals:

Correct interpretation (frequentist): In repeated sampling, 95% of such intervals will contain the true parameter.

Incorrect interpretation: "There is a 95% probability that the true parameter is in this interval."

The parameter is fixed (though unknown). The interval either contains it (probability 1) or doesn't (probability 0).

Calibrated vs. valid intervals:

- **Calibrated**: Contains the true parameter in exactly 95% of samples
- **Conservative**: Contains the true parameter in >95% of samples

- **Valid**: Calibrated or conservative (at least 95% coverage)

Most commonly used intervals (like Wald intervals) are **large-sample valid**: guaranteed to be valid only in large samples.

2 10.2 Estimation of Causal Effects (pp. 134-136)

In randomized experiments with random sampling, standard statistical methods can be used to estimate causal effects and compute confidence intervals.

2.1 Setting

Suppose:

1. Study population is a random sample from a super-population
2. Treatment is randomly assigned in the super-population (or in the sample)
3. All individuals adhere to assigned treatment
4. Exchangeability holds: $Pr[Y^a = 1] = Pr[Y = 1|A = a]$

2.2 Causal Inference

Because of exchangeability, the causal risk difference equals the associational risk difference in the super-population:

$$Pr[Y^{a=1} = 1] - Pr[Y^{a=0} = 1] = Pr[Y = 1|A = 1] - Pr[Y = 1|A = 0]$$

Estimators:

- Causal risk difference: $\widehat{Pr}[Y = 1|A = 1] - \widehat{Pr}[Y = 1|A = 0]$
- Causal risk ratio: $\widehat{Pr}[Y = 1|A = 1]/\widehat{Pr}[Y = 1|A = 0]$

Standard statistical methods provide confidence intervals for these causal effects.

Two perspectives on randomization:

1. **Super-population randomization**: Entire super-population is randomized, then we sample
   - Exchangeability holds in super-population
   - Sample estimates have random sampling variability
2. **Sample randomization**: We sample first, then randomize only the sample
   - Exchangeability holds on average across randomizations
   - Can lead to imbalance in small samples

Mathematically equivalent: Both approaches lead to the same statistical inference procedures.

Practical implication: We typically assume a super-population exists and use standard statistical methods.

2.3 Observational Studies

In observational studies, similar methods apply after adjusting for confounding:

- Compute standardized or IP weighted estimates
- Calculate confidence intervals using appropriate standard errors
- Account for adjustment when computing standard errors

Key difference from randomized experiments:

- **Randomized**: Association = Causation (under assumptions)
- **Observational**: Need to adjust for confounding first, then association = causation (under additional assumptions)

Both require handling random variability through confidence intervals.

3 10.3 The Myth of the Super-Population (pp. 136-139)

The concept of a "super-population" is a useful fiction that allows us to apply statistical methods, but it raises important questions about the sources of randomness.

3.1 Two Sources of Randomness

1. **Sampling variability**: Random selection of individuals from a super-population
2. **Nondeterministic counterfactuals**: Intrinsic randomness in outcomes even with fixed treatment and covariates

3.2 When Are Binomial Confidence Intervals Valid?

The standard binomial confidence interval for $p = Pr[Y = 1|A = a]$ is valid in two scenarios:

Scenario 1: Random sampling from a super-population

- Study participants are randomly sampled from a large super-population
- Each individual i has a fixed but unknown Y_i^a (deterministic counterfactuals)
- Randomness comes solely from which individuals are sampled

Scenario 2: Nondeterministic counterfactuals

- Study participants are the entire population of interest (no sampling)
- Each individual has a probability p_i of outcome, not a fixed Y_i^a
- Randomness comes from the probabilistic nature of outcomes

The super-population is often a fiction:

In many studies, we don't actually randomly sample from a larger population:

- Clinical trials: Enroll volunteers who meet criteria (not random sampling)
- Observational studies: Use convenient samples (hospitals, registries)
- Rare diseases: Study all available patients (no super-population)

Why use super-population framework anyway?

Even without actual random sampling, the super-population framework provides a convenient way to:

1. Quantify uncertainty via confidence intervals
2. Use standard statistical methods
3. Think about generalizability

Alternative: Randomization-based inference

Some authors prefer inference based solely on the randomization of treatment within the observed sample, without invoking a super-population. This approach has different technical requirements.

3.3 Practical Implications

Most applied researchers use confidence intervals computed under the super-population framework, even when:

- No random sampling was performed
- The super-population is not well-defined
- Generalization to a specific population is unclear

This practice is justified by the convenience and familiarity of standard methods, though it requires careful interpretation.

4 10.4 The Conditionality Principle (pp. 139-140)

When should we condition on variables when computing causal effects and their standard errors?

4.1 The Principle

Conditionality principle: If a variable L is independent of treatment A and outcome Y under the intervention, we may condition on L when computing estimates and standard errors without affecting validity.

4.2 Applications

Example 1: Stratified randomization

If treatment is randomized within strata of sex L:

- Can estimate effects unconditional on L (marginal effects)
- Can estimate effects conditional on L (stratum-specific effects)
- Both are valid; choice depends on scientific question

Example 2: Baseline covariates in randomized trials

Even when baseline covariates L are balanced across treatment groups:

- Conditioning on L may improve precision (narrower confidence intervals)
- Does not introduce bias if L is a pre-treatment variable
- Called "covariate adjustment" or "regression adjustment"

Why condition on baseline variables?

Precision gain:

- If L predicts outcome Y, adjusting for L reduces unexplained variability
- Leads to smaller standard errors and narrower confidence intervals
- Can increase statistical power

No bias introduced:

- If L is measured before treatment assignment
- And randomization ensures $L \perp A$
- Then adjusting for L doesn't create confounding

Practical recommendation:

In randomized trials, pre-specify baseline covariates to adjust for. This improves efficiency without cherry-picking.

5 10.5 The Curse of Dimensionality (pp. 140-142)

As the number of confounders or effect modifiers increases, nonparametric estimation becomes increasingly difficult. This is the **curse of dimensionality**.

5.1 The Problem

Suppose we need to adjust for 10 binary confounders:

- Number of possible covariate patterns: $2^{10} = 1024$
- Need enough observations in each pattern to estimate effects
- With limited data, many cells will be sparse or empty

Consequences:

1. **Positivity violations**: Some covariate patterns have no treated or untreated individuals
2. **Unstable estimates**: Small cell counts lead to large standard errors
3. **Impractical stratification**: Cannot stratify on so many variables simultaneously

5.2 The Solution: Parametric Models

Parametric models make assumptions about the functional form relating variables:

- Logistic regression: $\text{logit}Pr[Y = 1|A, L] = \beta_0 + \beta_1 A + \beta_2 L$
- Linear regression: $\text{E}[Y|A, L] = \beta_0 + \beta_1 A + \beta_2 L$

Advantages:

- Can "borrow strength" across covariate patterns
- Requires fewer parameters than nonparametric estimation
- Provides smooth estimates even with sparse data

Disadvantages:

- **Model misspecification**: If functional form is wrong, estimates are biased
- Trade-off between bias (from model assumptions) and variance (from limited data)

Why models are necessary:

With finite data and many covariates, we must make modeling assumptions. The question is not whether to use models, but which models to use and how to assess their adequacy.

Nonparametric vs. parametric:

- **Nonparametric**: No assumptions about functional form, but requires large data relative to dimension of covariates
- **Semiparametric**: Weaker assumptions than parametric, stronger than nonparametric (e.g., requires only correct propensity score model)
- **Parametric**: Strong assumptions about functional form, works with smaller data

Modern approach:

- Use flexible, data-adaptive methods (machine learning)
- Cross-validation to assess model fit
- Doubly robust methods that combine multiple models
- Sensitivity analyses to assess robustness to model assumptions

The curse applies to all adjustment methods:

- Stratification: Need cells for all covariate combinations
- Standardization: Need to estimate $\text{E}[Y|A, L]$ for all L values

- IP weighting: Need to estimate $Pr[A|L]$ for all L values
- Matching: Need to find matches for all covariate patterns

5.3 Why This Matters for Part II

The remainder of Part II describes methods that use parametric and semiparametric models to:

1. Adjust for many confounders efficiently
2. Handle high-dimensional covariate spaces
3. Estimate causal effects with adequate precision

Understanding the curse of dimensionality motivates the need for these modeling approaches.

6 Summary

This chapter introduced **random variability** and bridged Part I (identification) and Part II (estimation).

Key concepts:

1. **Identification vs. estimation**:
   - Identification: Can we express causal effects in terms of observables?
   - Estimation: How do we estimate from finite data?
2. **Statistical concepts**:
   - Estimands, estimators, estimates
   - Consistency: Estimates approach truth as $n \to \infty$
   - Confidence intervals: Quantify random sampling uncertainty
3. **The super-population**:
   - Convenient fiction for statistical inference
   - Two sources of randomness: sampling and nondeterministic counterfactuals
   - Justifies use of standard confidence intervals
4. **The conditionality principle**:
   - Can condition on variables independent of treatment/outcome
   - Conditioning on baseline variables can improve precision
5. **The curse of dimensionality**:
   - High-dimensional covariates require parametric/semiparametric models
   - Trade-off between bias (model misspecification) and variance (limited data)
   - Motivates modeling approaches in Part II

Transition to Part II:

Part I established **what** we want to estimate (causal effects) and **when** they can be identified (under exchangeability, positivity, consistency).

Part II addresses **how** to estimate causal effects from finite data using statistical models.

Looking ahead:

- **Chapter 11**: Standardization and IP weighting with parametric models
- **Chapter 12**: The parametric g-formula
- **Chapter 13**: Propensity scores and marginal structural models
- **Chapter 14**: Instrumental variable estimation
- **Chapter 15**: Outcome regression and g-estimation
- **Chapter 16**: Structural nested models

All of these methods address the curse of dimensionality by making modeling assumptions, with different trade-offs between flexibility and robustness.

Key message:

In practice, we almost always need models. The question is not whether to make assumptions, but which assumptions to make and how to assess their plausibility.

7 References

Hernán, Miguel A, and James M Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC. https://miguelhernan.org/whatifbook.