Stats 201A: Midterm Study Guide

Melody Y. Huang

Note: The intent of this was to create a Sparknotes, condensed version of topics we have gone over in lecture. Having typed it all up, it seems as if this isn't very concise or condensed. These are adapted from the lecture notes in class.

1. Characterizing an Estimator

We are interested in an *estimand* that is characterized by the population. Therefore, we construct an *estimator* that can serve to estimate an estimand. Estimators have various qualities that we care about, such as biasedness and precision. We care about how the estimator behaves in the context of our sample, as well as how the estimator behaves as we increase our sample size. The steps to characterize an estimator are summarized below:

1. Define an estimand of interest (θ).
2. Find an estimator ($\hat{\theta}$) to estimate our estimand. We usually can use the plug-in principle to use the sample analog of the estimand as an estimator.
3. Finite Sample Properties:
    - Unbiasedness: $\mathbb{E}(\hat{\theta})$
    - Variance of the estimator (precision): $\mathbb{V}(\hat{\theta})$
    - Estimated Variance of the estimator: $\hat{\mathbb{V}}(\hat{\theta})$
    - Finite Sampilng Distribution
4. Asymptotic Properties:
    - Consistency of Estimator
        - Note: To show $\hat{\theta}$ is consistent, we must show: (1) $\mathbb{E}(\hat{\theta}) \xrightarrow{p} \theta$, (2) $\mathbb{V}(\hat{\theta}) \xrightarrow{p} 0$ (we may also show $\mathbb{V}(\hat{\theta}) \xrightarrow{d} 0$, and appeal to the fact that by converging in distribution to a constant, we are also converging in probability.
    - Consistent estimator for the variance
    - Limiting Distribution

2. Sampling Design

Our goal in this portion is to understand different sampling designs. We care about sampling because in the real world, it is generally infeasible to survey the entire population (due to a variety of reasons, i.e., costs, non-response, etc.). Therefore, we are stuck with sampling from the population in hopes of using the information we gather from the sample to extrapolate larger information about the population.

More formally, the primary goal within sampling is to estimate summary characteristic of population by using only the information we have from the sample. We want to also assess how accurate (or precise) these estimates are. Different sampling designs will yield estimates that are more or less precise than others.

2.1. Notation

We begin by introducing a lot of notation.

We denote the population as U. Technically speaking, U is our sampling frame. $\mathcal{F}$ is all possible combinations of subsets that can exist from the universe U (i.e., $\mathcal{F}$ is the power set of U). A represents all the subsets of F that fit within our sampling design (i.e., if our sampling design restricts the number of things in our sample to be of a fixed size n, A would represent only the subsets of F that contain n elements). a represents one set from A. $A_{(i)}$ is defined as the set of all A's that contain the unit i.

Summarizing all of this:

$$a \in A \subseteq \mathcal{F}$$

In other words: a is the actual sample we have. A are the other permutations of the sample that we could have gotten under our sampling design.[1] We condition later on $\mathcal{F}$ because we are characterizing estimators in the context of these potential subsets and samples in our universe.

2.2. Design Linear Estimators

Definition 2.1 *Sampling Design*
Sampling design is the procedure by which the sample of units is selected from the population.

[1] It should be noted that in reality, A is really a set of indices that we can link back to U, but the implication is essentially the same.

In most of the sampling designs we discuss, we usually select units to be in our sample by assigning a probability to each unit of being chosen for the sample. This is known as the **inclusion probability**.

Definition 2.2 *Inclusion Probability*
The inclusion probability for a unit i is:

$$\pi_i = P(i \in A) = \sum_{a \in A_{(i)}} p(a)$$

Additionally, by introducing an indicator variable I_i that denotes whether the i-th unit is included in our population, we can think of the inclusion probability as:

$$\pi_i = \mathbb{E}(I_i),$$

where

$$I_i = \begin{cases} 1, & \text{if unit } i \text{ is in the sample} \\ 0, & \text{else} \end{cases}$$

We note that $\pi_{ii} = \pi_i$, and $\pi_{ik} = P(\text{Both unit } i \text{ and } k \text{ are included in sample}) = \mathbb{E}(I_i I_k)$.[2]

Recall that our goal in sampling is to be able to estimate some population estimand using an estimator. For now, we keep the estimand of interest relatively general, and denote it as θ.

Definition 2.3 *Linear Estimator*

$$\hat{\theta} = \sum_{i \in A} w_i y_i$$

We can think of this as a weighted combination of all the y_i's in a sample. The weights w_i must not be a function of y_i. Extending the concept of a linear estimator into the design-based framework, we can re-formulate the linear estimator to be summed not just over the A, but the entire population U. This is known as the **design linear estimator**.

Definition 2.4 *Design Linear Estimator*

$$\hat{\theta} = \sum_{i \in U} I_i w_i y_i$$

[2]This should look vaguely familiar as the example from Lecture 1, in which we introduced the indicator Z_i in explaining design-based inference.

The weights in the design linear estimators are always *pre-determined and fixed*, with respect to the sampling design. The great thing about design linear estimators is that so long as w_i and y_i are both fixed, we have a representation of the expectation, variance, and design-unbiased estimator for the variance for $\hat{\theta}$. This is more formally stated as two theorems (we refer to them in class as **Theorem 1** and **Theorem 2**).

Theorem 2.1 *Design Linear Estimators*
Assume N total units in a finite population. Let y_i represent a real-valued element for unit i from within the population, and let w_i be fixed. Then, given a design linear estimator of the following form:

$$\hat{\theta} = \sum_{i \in U} I_i w_i y_i,$$

the expectation of $\hat{\theta}$ will be:

$$\mathbb{E}(\hat{\theta} \mid \mathcal{F}) = \sum_{i=1}^{N} w_i \pi_i y_i$$

and the variance of $\hat{\theta}$ is:

$$\mathbb{V}(\hat{\theta} \mid \mathcal{F}) = \sum_{i=1}^{N} \sum_{k=1}^{N} (\pi_{ik} - \pi_i \pi_k) w_i y_i w_k y_k$$

Theorem 2.2 *Variance Estimator (for Design Linear Estimators)*
Once again, assume N total units in a finite population. Consistent with before, let y_i represent a real-valued element for unit i from within the population, and let w_i be fixed. Now, assuming that all joint inclusion probabilities are non-zero (i.e., $\pi_{ik} > 0 \; \forall \; i, k \in U$), then:

$$\hat{\mathbb{V}}(\hat{\theta} \mid \mathcal{F}) = \sum_{i,j \in A} \sum \pi_{ik}^{-1} (\pi_{ik} - \pi_i \pi_k) w_i y_i w_k y_k$$

Furthermore, $\hat{\mathbb{V}}(\hat{\theta} \mid \mathcal{F})$ will be a design-unbiased estimator for $\mathbb{V}(\hat{\theta} \mid \mathcal{F})$.

Different estimators can be derived using this framework by changing how we formulate the weights. The main family of estimators we care about are known as the **Horvitz-Thompson Estimator**.

2.3. Horvitz-Thompson Estimator

The Horvitz-Thompson Estimator defines the weights in our design linear estimator as being the inverse of inclusion probability:

$$w_i = \frac{1}{\pi_i}$$

Up until now, we have kept the estimand of interest very general (specified as θ). We will now introduce two common estimands that are of interest:

- Finite Population Total:

$$T_y = \sum_{i \in U} y_i = \sum_{i=1}^{N} y_i$$

- Finite Population Mean:

$$\bar{y}_N = \frac{1}{N} T_y = \frac{1}{N} \sum_{i=1}^{N} y_i$$

The Horvitz-Thompson Estimator can be used to derive an *unbiased estimate for both the population total and mean*:

- Estimator for Finite Population Total:

$$\hat{T}_y = \sum_{i \in U} I_i \pi_i^{-1} y_i = \sum_{i \in A} \pi_i^{-1} y_i$$

$$\hat{\mathbb{V}}(\hat{T}_y) = \sum_{i,j \in A} \sum \pi_{ik}^{-1} (\pi_{ik} - \pi_i \pi_k) \pi_i^{-1} y_i \pi_k^{-1} y_k$$

- Estimator for Finite Population Mean:

$$\bar{y}_{HT} = \frac{1}{N} \hat{T}_y = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} y_i$$

$$\hat{\mathbb{V}}(\bar{y}_{HT}) = \frac{1}{N^2} \sum_{i,j \in A} \sum \pi_{ik}^{-1} (\pi_{ik} - \pi_i \pi_k) \pi_i^{-1} y_i \pi_k^{-1} y_k$$

This will serve as the basis for all of our estimators that are derived from various sampling schemes. The basic estimator specification will remain the same, but we will change the inclusion probabilities (π_i) to fit with the sampling design.

Usually when using the estimator for the variance of the estimator, it is more optimal to reformulate the specification above such that we denote two case: one in which we look strictly at just the variance if unit i, and then the covariance between i and the other units. More formally:

$$\hat{\mathbb{V}}(\hat{T}_y) = \sum_{i \in A} \pi_i^{-2} (1 - \pi_i) y_i^2 + \sum_{i \neq k} \sum \pi_{ik}^{-1} (\pi_{ik} - \pi_i \pi_k) \pi_i^{-1} y_i \pi_k^{-1} y_k$$

Cleaning this up further:

$$\hat{\mathbb{V}}(\hat{T}_y) = \sum_{i \in A} \pi_i^{-2} (1 - \pi_i) y_i^2 + \sum_{i \neq k} \sum \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik} \pi_i \pi_k} y_i y_k$$

Some potential drawbacks of the Horvitz-Thompson estimator are that it can be very high variance[3] (i.e., Basu's elephant), and that it will be location variant.

Alternative Estimator: Hajek Estimator[4]
An alternative estimator to the Horvitz-Thompson estimator is the Hajek estimator. The Hajek estimator weights events that are more probable, such that the estimator is less susceptible to outliers or "weird" events that could be drawn. The Hajek estimator is more precise than the Horvitz-Thompson estimator; however, it will be biased. It takes on the following form:

$$\hat{T}_{Hajek} = \left(\sum_{i \in A} \pi_i^{-1} \right)^{-1} \sum_{i \in A} \pi_i^{-1} y_i$$

2.4. Sampling Schemes

1. Simple Random Sampling
    - Assume random draws (sampling without replacement)
    - $\binom{N}{n}$ total possible samples
    - $\pi_i = \frac{n}{N}$ for all i
    - $\pi_{ik} = \frac{n}{N} \cdot \frac{n-1}{N-1}$
    - Estimated Population Total:
    $$\hat{T}_y = N \cdot \frac{1}{n} \sum_{i \in A} y_i$$
        - Variance of Population Total:
        $$\mathbb{V}(\hat{T}_y \mid \mathcal{F}) = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \left[\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_N)^2 \right]$$

---

[3] Horvitz and Thompson showed that there exists an optimal weight assignment (or, optimal way to set up our inclusion probabilities) such that the high variance problem could be mitigated:

$$\frac{1}{n-1} \sum_{k=1}^{N} \pi_{ik} = \pi_i = \frac{n y_i}{\sum_{i=1}^{N} y_i}$$

As it turns out, this is profoundly unhelpful since we don't know what $\sum_{i=1}^{N} y_i$ is, and if we did know it, we wouldn't have to sample and use an estimator to begin with. It does, however, imply that we should attempt to find a good proxy for y and try to sample proportionally in accordance with this auxilary variable.

[4] We never really mention Hajek after this, and all the sampling designs are discussed in the context of the HT estimator framework, but I've included this section as a fun fact in case it ever shows up as a Jeopardy question!

- Estimated Variance of Population Total:

$$\hat{\mathbb{V}}(\hat{T}_y \mid \mathcal{F}) = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \left[\frac{1}{n-1} \sum_{i \in A} (y_i - \bar{y}_n)^2\right]$$

2. Poisson Sampling (a.k.a. p-coin flipping)

- Series of N independent Bernoulli trials
- Each unit has a probability of π_i (π_i can be different across units)
- $\pi_{ik} = \pi_i \pi_k$ - Independence across trials!
- Estimated Population Total:

$$\hat{T}_y = \sum_{i \in A} \pi_i^{-1} y_i$$

  - Variance of Population Total:

$$\mathbb{V}(\hat{T}_y) = \sum_{i=1}^{N} \pi_i^{-1} (1 - \pi_i) y_i^2$$

  - Estimated Variance of Population Total:

$$\hat{\mathbb{V}}(\hat{T}_y) = \sum_{i=1}^{N} \pi_i^{-2} (1 - \pi_i) y_i^2$$

3. Stratified Sampling

- We assume H total strata
- The h-th strata contains N_h element
- Each element is in only one strata
- We then have a sampling scheme that we apply to each strata.
- Estimate of population mean (assuming SRS of strata):

$$\bar{y} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h,$$

where $\bar{y}_h$ is the mean within each strata (i.e., $\bar{y}_h = \frac{1}{n_h} \sum_{i \in A_h} y_{h,i}$)

  - The variance of the estimated population mean is:

$$\mathbb{V}(\bar{y} \mid \mathcal{F}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

  - The estimated variance of the estimated population mean is:

$$\hat{\mathbb{V}}(\bar{y} \mid \mathcal{F}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

4. Cluster Sampling

2.5. Non-Sampling Error

We generally expect some degree of error to arise from extrapolating information from a sample to a population. As we increase the sample size to approach that of the population, we expect to mitigate most of this error. However, it turns out there exists errors that arise from non-sampling issues. These errors, aptly called non-sampling errors, will persist even as we increase $n \to N$. The problem we discuss mostly will be non-response error.

We care about non-response because there may exist biases that arise from persistent non-response. This can be problematic, because then we cannot estimate things accurately for the population.

Definition 2.5 *Non-Response*
Non-response is defined as when a unit does not respond.
We use an indicator variable R_i to represent this.

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds} \\ 0 & \text{if unit } i \text{ does not respond} \end{cases}$$

We also define something known as **stable outcomes**, which weights all the outcomes Y_i by whether or not the unit responded. Any unit which does not respond is given a value of -99.

Definition 2.6 *Stable Outcomes*

$$Y_i^* = Y_i R_i + (-99)(1 - R_i)$$

We call Y_i^* the censored version of Y_i.

There are different types of non-response errors. We care more about some than others. We can categorize missing data into three types:

1. Missing completely at random
   In the case that non-response is missing completely at random, there exists no relationship between the outcome Y_i and the response R_i. In other words, $Y_i \perp\!\!\!\perp R_i$. This is usually not too much of a concern. (We can simply `na.omit` the missing values!)

2. Missing at random
   In this case, there may exist some relationships between Y_i and R_i. However, we can explain this relationship by using some auxiliary variables X_i such that by conditioning on X_i, we get a conditionally independent relationship between Y_i and R_i. We denote this as: $Y_i \perp\!\!\!\perp R_i \mid X_i$. Conditioning on some auxiliary variables X_i is equivalent to re-weighting the data we have on some external variables. This is akin to **post-stratification**.

3. Not missing at random
   In this case, the probability of response depends on the outcome, which will cause bias. This is bad.

We can deal with the issue of non-response (assuming missing at random) by re-weighting things. There are different ways to re-weight. The methods covered are: propensity score weighting and calibration (raking), and post-stratification. I attempt to provide some detail about each one below.

Balancing Score (Response Propensity Score)

One way to re-weight the data is to use the propensity score.

Definition 2.7 *Balancing Score*
We define the balancing score as the conditional probability that a unit will respond:

$$p(x) = P(R_i = 1 \mid X_i = x).$$

This is also known as the propensity score. Conditioning on the propensity score allows for conditional independence.[5] *More specifically:*

$$R_i \perp\!\!\!\perp X_i \mid p(X_i)$$

$$Y_i \perp\!\!\!\perp R_i \mid p(X_i)$$

We care about the fact that $Y_i \perp\!\!\!\perp R_i \mid p(X_i)$ because this allows us to use the outcome for responders for non-responders without inducing bias. Conditioning on the propensity score is equivalent to conditioning on all the covariates.[6]

Re-weighting the data using the balancing score (also referred to as *inverse propensity score weighting*), the expected outcome takes on the following form:

$$\mathbb{E}(Y_i) = \mathbb{E}\left(\frac{Y_i^* R_i}{p(X_i)}\right)$$

[5] Independence and ignorability are often used somewhat interchangeably. Another term for this is exogeneity, for all the ex-economists out there (compare this to endogeneity, in which problematic things occur).

[6] In the ideal world, we would not need to do this because with randomization, we hope that the condition of $Y_i \perp\!\!\!\perp R_i$ will already be met. However, in reality, people who respond to the survey may often share specific characteristics that would not be necessary representative of the population. Biases may arise from this. Since we cannot claim total ignorability, we need conditional ignorability by conditioning on potential covariates that may characterize the probability of responding.

Derivation:
Like all derivations, we begin by doing some fun algebra. In particular, we substitute in Y_i^*

$$\begin{aligned}\frac{Y_i^* R_i}{p(X_i)} &= \frac{(Y_i R_i + (-99)(1 - R_i)) \cdot R_i}{p(X_i)} \\ &= \frac{Y_i R_i^2 + (-99)(R_i - R_i^2)}{p(X_i)} \\ &\text{Note that } R_i^2 = R_i\text{:} \\ &= \frac{Y_i R_i}{p(X_i)}\end{aligned}$$

Now we can take the conditional expectation:

$$\begin{aligned}\mathbb{E}\left(\frac{Y_i^* R_i}{p(X_i)} \;\middle|\; X_i = x\right) &= \mathbb{E}(\left(\frac{Y_i R_i}{p(X_i)}\middle| X_i = x\right) \\ &= \frac{\mathbb{E}(Y_i R_i |\, X_i = x)}{p(X_i)}\end{aligned}$$

We have assumed conditional independence between Y_i and R_i, so:

$$= \frac{\mathbb{E}(Y_i|X_i = x)\mathbb{E}(R_i|X_i = x)}{p(X_i)}$$

Plugging in the definition of the propensity score:

$$\begin{aligned}&= \frac{\mathbb{E}(Y_i|X_i = x)\mathbb{E}(R_i|X_i = x)}{P(R_i = 1 \mid X_i = x)} \\ &= \mathbb{E}(Y_i|X_i = x)\end{aligned}$$

Therefore, we can use the Law of Iterated Expectation to find:

$$\mathbb{E}\left(\frac{Y_i^* R_i}{p(X_i)}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{Y_i^* R_i}{p(X_i)} \;\middle|\; X_i = x\right)\right) = \mathbb{E}(\mathbb{E}(Y_i|X_i = x)) = \mathbb{E}(Y_i)$$

□

In reality, we do not have $p(X_i)$ and will need to estimate it somehow (i.e., using a logistic regression). We denote the estimate of $p(X_i)$ (conditional probability of response) as $\hat{p}(X_i)$. Therefore, to estimate $\mathbb{E}\left(\frac{Y_i^* R_i}{p(X_i)}\right)$, we can use the following:

$$\hat{\mathbb{E}}_{IPW}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i^* R_i}{\hat{p}(X_i)}$$

Notice that if we write it as such:

$$\hat{\mathbb{E}}_{IPW}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{\hat{p}(X_i)} \cdot Y_i^*,$$

this looks a lot like the Horvitz-Thompson Estimator!
There exists some problems with this, in that the probability of response could be very low, which could lead to a massive blow up, as the denominator approaches zero. Therefore, we introduce something known as the **stabilized estimator**:

$$\hat{\mathbb{E}}_{SIPW}(Y_i) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^* R_i}{\hat{p}(X_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\hat{p}(X_i)}}$$

This is like the Hajek estimator, in that it will be biased, but more precise than the IPW estimator. A drawback of propensity score weighting is that we have to be able to accurately estimate $\hat{p}(X_i)$, which can be a very non-trivial. We can use another method known as calibration that does not require a functional form for the probability of response.

Post-Stratification

Given the population proportions, we can re-weight our sample after the fact to match all the population moments of some auxiliary variables. This means we have access to the population moments of this exogenous information (i.e., from a census). This is done by calculating the proportion of our particular strata and then weighting by this proportion. Sometimes, we can run into issues with the lack of data. For example, assume we want to re-weight our data on the variables race, income, education, and gender. While we could likely find a middle class, college educated white male in our survey, it may be difficult to find data for every iteration of the various categories present. The nice thing, however, with post-stratification is that we are guaranteed a balance on even margins, provided no missing data.

Calibration

Calibration allows us to weight such that *population moments are met exactly*. Mathematically, it will try to minimize a distance function:

$$\sum_k D(w_{k,final}, w_{k,start}),$$

subject to certain constraints that are usually population moments. Intuitively, this means that we are trying to find the closest weights to the weights we have within our sample (the starting weights - $w_{k,start}$) that allow us to match the population moments (such as population mean). Different distance functions will correspond to different weighting schematics. For example, the maximum entropy distance function corresponds to raking.[7] Raking is nice because we are less likely to have a missing data problem, but not all the margins are guaranteed to be balanced perfectly.

[7] The slides also mention GREG, which corresponds to a regression estimator. That is likely beyond the scope of what is on the midterm.

3. Causality

Our goal is to measure the causal effect[8] of a particular *unit*. For example, assume you have a headache from studying for your midterm. We want to see if, after you take aspirin, the aspirin mitigates your headache so you may return back to productively studying. In order to discuss causal effects, we must introduce something known as the **potential outcomes framework**.

3.1. The Potential Outcomes Framework

Within the potential outcomes framework, we assume many different units i exist. We can assign a unit i to treatment, or we can assign the unit to control. This is denoted by an indicator variable T_i. Each unit has its own potential outcomes: $Y_i(T_i)$. Our goal is to find the treatment effect of a particular unit i, i.e., the effect that being in treatment had.

More formally:

Definition 3.1 *Treatment Assignment*

$$T_i = \begin{cases} 1, & \textit{unit } i \textit{ is assigned to treatment} \\ 0, & \textit{else} \end{cases}$$

Definition 3.2 *Potential Outcomes*

$$Y_i(T_i) = \begin{cases} Y_i(1), & \textit{potential outcome for unit } i\textit{, if unit } i \textit{ is assigned to treatment} \\ Y_i(0), & \textit{potential outcome for unit } i\textit{, if unit } i \textit{ is assigned to control} \end{cases}$$

The potential outcomes for each unit i is *fixed*. The only random component within our design is the treatment assignment.

Definition 3.3 *Treatment (Causal) Effect*

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

The above is the unit level treatment effect. Other estimands that are often of interest to us within a study are:

- Sample Average Treatment Effect (SATE):

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

[8]Rubin uses the term 'causal effect'. We use the term 'treatment effect' in class.

- Population Average Treatment Effect (PATE):

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

- Population Average Treatment Effect for the Treated (PATT):

$$\mathbb{E}(Y_i(1) - Y_i(0)|T_i = 1)$$

- Population Conditional Average Treatment Effect (CATT):

$$\mathbb{E}(Y_i(1) - Y_i(0)|X_i = x)$$

Notice that by definition, all of the estimands that represent the treatment effect compare the potential outcomes for the same unit, at the same moment in time, following treatment assignment. In reality, we can never observe both $Y_i(0)$ and $Y_i(1)$, because a unit cannot both be simultaneously in the treatment and control group, such that both potential outcomes are realized. Therefore, we can only see something known as the **observed outcome**.

Definition 3.4 *Observed Outcome*

$$Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$$

This brings to light the fundamental problem within causal inference: we want to see the specific effect of treatment on unit i by comparing what would have occurred if unit i had been under treatment, versus what would have occurred if unit i had received no treatment. However, by putting unit i in either treatment or control, we can no longer observe the other potential outcome! Therefore, we have a missing data problem. In other words, we can only compare observed outcomes, since there exists only one realized potential outcome per unit.

To bypass the fundamental problem of missing data, we can try to predict (or impute) the missing potential outcomes at an individual level. Alternatively, we can observe different units i that are assigned to treatment or control within the same time period and see if we can estimate the treatment effects. In order to perform any form of identification, the assignment mechanism must be *random*. In the next portion, we will talk about the necessary assumptions that need to be met within our designed experiment in order to estimate treatment effects.

Long story short: we cannot measure the causal effect of a treatment due to the missing data problem. Therefore, we must randomize our treatment assignments to *estimate* the treatment effects, such that the outcome of one unit is independent of the treatment assignments of other units. This primary assumption is known as SUTVA.

3.2. Stable Unit Treatment Value Assumption

The Stable Unit Treatment Value Assumption (SUTVA) is that the outcome of a single unit does not depend on the assignments of other units. This can be written as three explicit assumptions that must be met in order for us to make claims about causal effects.

The assumptions are:

1. No simultaneity (outcome cannot cause treatment)
2. No interference (a unit's outcome does not depend on whether other units are assigned treatments)
3. Same version of treatment is being applied

SUTVA is easily violated in the context of spillover effects, as well as scenarios in which there may be variation in which treatments are administered. In these cases, we often try to re-define the treatment or our framework. For example, we may change how we define a unit. Whereas before, a unit i may be an individual person, we may opt to define a unit as an entire neighborhood.

3.3. Assignment Mechanism

The assignment mechanism is formally defined as the process by which units receive treatment. There exist three basic restrictions on how our assignment mechanism works:

1. Individual Assignment: No dependence of a unit's assignment probability on the values of the covariates and potential outcomes for other units
2. Probabilistic Assignment: Non-zero probability for all units to be assigned treatment
3. Unconfounded Assignment: No dependence of assignment mechanism on potential outcomes

There are different types of assignment mechanisms, depending on the experiment design and question we are trying to answer. In particular:

1. Classical Randomized Experiments:[9]
   In a classical randomized experiment, the researcher controls who receives treatment at a unit level. These experiments have *high internal validity*, since assignment probabilities are fully controlled and determined by the researcher. However, they may have *lower external*

[9]This is what our class is primarily focused on.

validity due to the fact that the results within an experiment can be difficult to generalize to the wider population.

2. Regular Assignment Mechanism
   A regular assignment mechanism is where the actual assignment mechanism is unknown to the researcher, but the process by which units are assigned to treatment or control still meets the necessary assumptions. This is common within observational studies. It is usually up to the researcher to prove that the required assumptions are safe to make within this context. In these studies, findings may have *low internal validity*, but *high external validity*.

3. Irregular Assignment Mechanism
   With an irregular assignment mechanism, there are issues with confoundment. There may be differences within the treatment group, due to the fact that receipt of treatment may vary amongst units. In order to perform casual inference here, we must make additional assumptions.

3.4. Design-Based Inference with Randomization

By randomization, we are able to justify causal inference, despite of the missing data problem. This is because we assume that, since there exists randomization, there exist no relationship between the treatment and underlying characteristics of each unit. In other words, we assume:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i$$

This would allow us to perform *identification by randomization*. This is often times referred to as *ignorability*.

When we cannot assume that the potential outcomes and the treatments are independent, we may find covariates X_i (i.e., some pre-treatment characteristics) that serve as a backdoor criterion[10] such that:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$$

This would render $Y_i(\cdot)$ and T_i conditionally independent, and we can then perform *identification by conditional independence* (this is referred to as *conditional ignorability*).

[10] For all the homies out there who, like me, don't know what backdoor criterion is: assume we are trying to figure out the causal effect of a on y (i.e. $a \to y$) However, there's a confounding variable x, such that $x \to a$, and $x \to y$. This makes it hard to tease out the actual effect of a on y since there is a backdoor path that exists ($x \to a \to y$). Therefore, in order to tease out the effect of a on y, we have to first condition on x to "block" the pathway formed between x and a, and x and y. (This makes a lot more sense when looking at a DAG, but since I'm too lazy to draw one, words will have to suffice.) **TL;DR:** backdoor criterion is a Fancy Way of saying that we've conditioned on all potential confounding variables, such that there exists no confounding effects, so we can perform causal inference.

For the purposes of this course, we focus primarily on identification by randomization. In identification by randomization, we follow the following framework. We begin by fixing the number of units that must be assigned to treatment (call this n_1). There are n units in total. This means that the probability of treatment is:

$$P(T_i = 1) = \frac{n_1}{n}$$

Our estimands of interest are the SATE or PATE. The estimator that is used for SATE and PATE is known as the *difference in means estimator*.

Definition 3.5 *Difference-in-Means Estimator*

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)Y_i$$

The difference-in-means estimator is an unbiased estimator of SATE. It is also an unbiased estimator of PATE, given certain assumptions. We will discuss properties of the difference-in-means estimator below.

Properties of the Difference-in-Means Estimator

Property 1. The difference-in-means estimator is an unbiased estimator of SATE

Property 2. Under certain assumptions, the difference-in-means estimator is an unbiased estimator of PATE.

Property 3. To estimate the variance of the difference-in-means estimator for SATE, we must use the Neyman Estimator as an upper bound for variance.

Property 4. The Neyman Estimator will be an unbiased estimator for the variance of the difference-in-means estimator for PATE.

Each property will be expanded upon in more detail.

Property 1. Unbiased Estimator of SATE

Above, we stated that the difference-in-means estimator $\hat{\tau}$ is an unbiased estimator of SATE. To show this, we can take the expectation of $\hat{\tau}$. Let's define Ω as the set of potential outcomes in our sample of n units (formally: $\Omega = \{Y_i(0), Y_i(1)\}_{i=1}^{n}$):

$$\mathbb{E}(\hat{\tau} \mid \Omega) = \mathbb{E}\Bigg(\underbrace{\frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i}_{(1)} - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)Y_i \;\Bigg|\; \Omega \Bigg)$$

Recall that the observed outcome is defined as $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. We can substitute this in for Y_i. Let's begin by looking at just the first term (denoted as (1)):

$$\begin{aligned}\frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i &= \frac{1}{n_1}\sum_{i=1}^{n} T_i(T_i Y_i(1) + (1 - T_i)Y_i(0)) \\ &= \frac{1}{n_1}\sum_{i=1}^{n}(T_i^2 Y_i(1) + T_i(1 - T_i)Y_i(0)) \\ &= \frac{1}{n_1}\sum_{i=1}^{n}(T_i^2 Y_i(1) + (T_i - T_i^2)Y_i(0))\end{aligned}$$

Notice that $T_i^2 = T_i$ (Since $T_i = 1$ or 0, so the squared of this value will remain the same). We can rewrite the above as:

$$\begin{aligned}&= \frac{1}{n_1}\sum_{i=1}^{n}(T_i Y_i(1) + (T_i - T_i)Y_i(0)) \\ &= \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i(1)\end{aligned}$$

Now if we take the expectation across this sum, the only random component to this equation is T_i, since the potential outcomes $Y_i(\cdot)$ are fixed, as is n_1. Therefore:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i \;\middle|\; \Omega\right] &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}(T_i \mid \Omega)Y_i(1) \\ &= \frac{1}{n_1}\sum_{i=1}^{n}\frac{n_1}{n}Y_i(1) \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i(1)\end{aligned}$$

A similar substitution can be made on the second term. This gives us:

$$\begin{aligned}\mathbb{E}(\hat{\tau} \mid \Omega) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\left(1 - \frac{n_1}{n}\right)Y_i(0) \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) \\ &= \text{SATE}\end{aligned}$$

Property 2. Unbiased Estimator of PATE

The difference-in-means estimator can also be used as an unbiased estimator for PATE. However,

in order to do so, we must assume that all units are randomly sampled from an infinite population. Recall that SATE is conditioned on the set Ω. Therefore, we can think Ω as one sub-group that serves as one realization of infinitely many other samples that could have been drawn. Using Law of Iterated Expectation, we find:

$$\mathbb{E}(\hat{\tau}) = \mathbb{E}(\mathbb{E}(\hat{\tau} \mid \Omega)) = \mathbb{E}(\text{SATE})$$

This is where things get a little weird. Before, we said that SATE and PATE were both estimands of interest. They technically are. However, SATE is also an estimator for PATE (much like how the sample mean is an estimator for the population mean). We can rewrite SATE in a way such that this makes more sense. In particular, we can introduce an indicator variable Z_i. Z_i takes on a value of 1 when the i-th unit is in our sample, and 0 otherwise (this is consistent with before, when we first introduced design based estimators). By introducing Z_i, we may rewrite the summation to now be from $i = 1, ..., N$,[11] instead of $i = 1, ..., n$. Additionally, like before, the probability that $Z_i = 1$ is simply the number of units in the sample divided by the total number of units in the population.

In math speak:

$$\begin{aligned}
\text{SATE} &= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0) \\
&= \frac{1}{n}\sum_{i=1}^{N} Z_i Y_i(1) - Z_i Y_i(0) \\
&= \frac{1}{n}\sum_{i=1}^{N} Z_i(Y_i(1) - Y_i(0))
\end{aligned}$$

Taking the expectation of SATE:

$$\begin{aligned}
\mathbb{E}(\text{SATE}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{N} Z_i(Y_i(1) - Y_i(0))\right) \\
&= \frac{1}{n}\sum_{i=1}^{N} \mathbb{E}(Z_i)(Y_i(1) - Y_i(0)) \\
&= \frac{1}{n}\sum_{i=1}^{N} \frac{n}{N}(Y_i(1) - Y_i(0)) \\
&= \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0)
\end{aligned}$$

[11] We are going to ignore the fact that we've just assumed there exists an infinite population, in which case $N \to \infty$; instead, let's reframe infinity to be something akin to a Very Large Number, like a million in accordance with Wu, 2018.

As such, SATE is an unbiased estimator for PATE.

Therefore, after this fun digression, we can return back to the expectation of $\hat{\tau}$ to show that:

$$\mathbb{E}(\hat{\tau}) = \mathbb{E}(\mathbb{E}(\hat{\tau} \mid \Omega)) = \mathbb{E}(\text{SATE}) = \text{PATE}$$

Property 3. Estimating the Variance of $\hat{\tau}$ for SATE

What about the variance of $\hat{\tau}$?

We once again begin by conditioning on just our sample's potential outcomes.

$$var(\hat{\tau} \mid \Omega) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_{01}^2}{n}$$

Above, S_t^2 is the equivalent of the sample variance for $Y_i(t)$. More formally:

$$S_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(t) - \overline{Y(t)})^2$$

(This is equivalent to if you took your entire column of $Y_i(1)$'s and typed in `var(y1)` into R).

S_{01}^2 is the sample variance of τ_i (the treatment effect)[12]:

$$S_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \text{SATE})^2$$

We can rewrite this as:

$$S_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(1) - Y_i(0) - \text{SATE})^2$$

This is mildly problematic, because in order to estimate the variance of the estimator, we would need to somehow estimate the sample variance of the treatment effect. This would require us to know both potential outcomes, which we don't know, or we wouldn't be down this rabbit hole of inference through randomization to begin with!

Since we have no means of ever really measuring S_{01}^2 without entering the alternate universe in which the treatment and control groups were flipped, the next best option is to bound the variance. We can do this relatively trivially by noting the following:

$$var(\hat{\tau} \mid \Omega) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_{01}^2}{n} \leq \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$$

[12]Brief note on notation: in the lecture notes, we use τ to denote SATE. This is mildly confusing, because there also exists PATE, which could arguably also be represented by τ. In formulating the sample variance, the Rubin book denotes the term as τ_{fs} To bypass introducing more symbols, I simply write in SATE.

Neyman was the first to observe this[13], so we call this bound the **Neyman estimator**:

$$\hat{\mathbb{V}}_{Neyman}(\hat{\tau} \mid \Omega) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$$

The Neyman estimator can serve as an estimator for the variance of SATE. If the treatment effect is constant for all units i (i.e., $\tau_i = c$, $\forall\, i$, which would imply that SATE is also equal to c, thereby making the entire S_{01}^2 term go to zero), then the Neyman estimator will be unbiased (in the lecture notes, this is referred to explicitly as the *constant treatment assumption*). However, this constant treatment effect assumption is usually not met. This is actually okay. While the Neyman estimator may not be unbiased under non-constant treatment effect scenarios, it serves as a conservative estimate for what the actual variance is. This means that while we may have overcoverage, we can assume at least nominal coverage.

Property 4. Estimating the Variance of $\hat{\tau}$ for PATE

For the variance of the estimator when looking at PATE, we have the following:

$$\mathbb{V}(\hat{\tau}) = \mathbb{V}(\mathbb{E}(\hat{\tau} \mid \Omega)) + \mathbb{E}(\mathbb{V}(\hat{\tau} \mid \Omega)) \approx \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$$

Therefore, for PATE, there exists no need to know both potential outcomes! We can rejoice, because this implies that we can use the Neyman estimator as an unbiased estimate for the variance:

$$\hat{\mathbb{V}}(\hat{\tau}) = \hat{\mathbb{V}}_{Neyman}(\hat{\tau})$$

3.5. Randomization Inference

Up until this point, we have talked a lot about estimating the average treatment effects within both samples and populations. However, estimating an average treatment effect of zero does not necessarily imply a lack of treatment effect across all units. A very simple example could be that one sub-group within our sample has a positive treatment effect, while another sub-group has a negative treatment effect. Therefore, this could lead to an averaged zero treatment effect.

We want to examine the question whether or not a treatment effect is zero for every unit. To do this, we must introduce something known as **randomization inference**, also referred to as **permutation (or Fisherian) inference**.

To begin, we define something known as the Sharp Null Hypothesis.

Definition 3.6 *Sharp Null Hypothesis*
The Sharp Null Hypothesis assumes that the estimated treatment effect is zero for all units i. This

[13] I'm not sure if that's actually true, but I'd imagine so, since this is named after him. "Did you ever think about what a coincidence it is that Lou Gehrig had Lou Gehrigs disease?" (Kresin, 2018).

can be extended such that we may choose the estimated treatment effect to be any constant c under the sharp null.

The Sharp Null Hypothesis allows us to bypass the missing data problem, because under the sharp null, we can easily calculate the missing potential outcomes that are unobserved by using the observed outcomes. In other words, because we are assuming there is no treatment effect, $Y_i(0) = Y_i(1) = Y_i$.

Under the randomized inference framework, we calculate all the possible treatment assignments. This is possible because within the classical randomized experiment setting, we (the researcher) have full control over the assignment mechanism, and therefore know the randomization scheme that is being used to assign units to treatment or control. For some test statistic t, we can generate a distribution of all possible values of t under our assignment mechanism.[14] (It is important to reiterate that we can create such a distribution because all the potential outcomes are known, since we assume the only thing that changes under the Sharp Null is the label of whether or not a unit is in treatment or control; the outcome never changes under the Sharp Null.) As such, by generating such a distribution, we can directly calculate the probability of observing our actual test statistic under the sharp null.

The probability of observing our test statistic, or a more extreme value, under the sharp null is what is formally known as **Fisher's exact p-values**.

Definition 3.7 *Fisher's Exact P-Values*
Let Ω be defined as the set of all possible randomizations under our assignment mechanism.
For a one-sided test, Fisher's exact p-values are defined as:

$$P = \frac{1}{|\Omega|} \sum_{i}^{|\Omega|} \mathbb{I}(t_i \geq t_{obs})$$

For a two-sided test, the p-value would take on the following form:

$$P = \frac{1}{|\Omega|} \sum_{i}^{|\Omega|} \mathbb{I}(|t_i| \geq |t_{obs}|))$$

Essentially, we simply count up the total number of test statistics that, under all the various permutations of assignments, could be more extreme than the observed test statistic, and divide it by the total number of different permutations of assignments that exist in our design. If the probability of getting a test statistic like the one we observed is very low under the sharp null, then we have reason to believe that there may not necessarily be a zero treatment effect for all units, and, as such, can opt to reject the Sharp Null.

[14] t is never defined as anything explicit, but Rubin makes a note that the test statistic we have in question is going to be the function of the assignment variable T_i, the observed outcome Y_i, and potentially some covariates.

4. Appendix

4.1. Probability Theory: the Very Abridged Version

I have not provided any sort of intuition as to what is going on here. This is mostly going to be focused on derivations.

Convergence in Probability

$$\lim_{n\to\infty} P(|T_{(n)} - c| \geq \varepsilon) = 0$$

Alternatively, we can frame this as:

$$\lim_{n\to\infty} P(|T_{(n)} - c| \leq \varepsilon) = 1$$

(This alternative representation can be easily shown by taking the complement.)

Continuous Mapping Theorem (CMT)

Let $S_{(1)}, S_{(2)}, S_{(3)}, ...$ that converges in probability to some value a.
Then for some function g, $g(S_{(n)}) \xrightarrow{p} g(a)$.

Chebyshev's Inequality

$$P(|X - E(X)| \geq \varepsilon\sigma^2) \leq \frac{1}{\varepsilon^2}$$

Chebyshev's Inequality (for the Sample Mean):

$$P(|\bar{X} - E(X)| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Derivation trick: Define a new $\varepsilon' = \frac{\varepsilon}{\hat{\sigma}}$. Recall that $\hat{\sigma} = \frac{\sigma^2}{n}$. Therefore:

$$P(|\bar{X} - E(X)| \geq \varepsilon) = P(|\bar{X} - E(X)| \geq \varepsilon'\hat{\sigma})$$

By the original formulation of Chebyshev's Inequality, we know that this must be less than $\frac{1}{\varepsilon'^2}$:

$$P(|\bar{X} - E(X)| \geq \varepsilon'\hat{\sigma}) \leq \frac{1}{\varepsilon'^2}$$

$$P(|\bar{X} - E(X)| \geq \varepsilon'\hat{\sigma}) \leq \frac{\hat{\sigma}^2}{\varepsilon^2}$$

$$P(|\bar{X} - E(X)| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Weak Law of Large Numbers

$$\bar{X}_n \xrightarrow{p} \mu$$

As we increase the sample size such that n gets infinitely large, then our sample mean will converge in probability to the population mean. We can use Chebyshev's Inequality to derive this:

$$0 \leq P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon}$$

The lower bound comes from the fact that probability must be greater than or equal to zero. Taking the limit as n goes to infinity:

$$0 \leq \lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon}$$

$$0 \leq \lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) \leq 0$$

By Squeeze Theorem:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

Slutsky's Theorem
Let $S_{(1)}, S_{(2)}, S_{(3)}, ...$ and $T_{(1)}, T_{(2)}, T_{(3)}, ...$ be two sequences of random variables (i.i.d.) Assume the following:

$$S_{(n)} \xrightarrow{p} c$$

$$T_{(n)} \xrightarrow{d} T$$

Then the following is true:

1. $S_{(n)} + T_{(n)} \xrightarrow{d} c + T$
2. $S_{(n)} \cdot T_{(n)} \xrightarrow{d} cT$
3. $\frac{T_{(n)}}{S_{(n)}} \xrightarrow{d} \frac{T}{c}$

Convergence in Distribution
Let $T_{(1)}, T_{(2)}, T_{(3)}, ...$ be a bunch of random variables, with cumulative density functions $F_{(1)}, F_{(2)}, F_{(3)}, ...$ Let T be a random variable with a cdf F. Then we say $T_{(n)} \xrightarrow{d} T$ if $\forall\ t \in \mathbb{R}$:

$$\lim_{n\to\infty} F_{(n)}(t) = F(t)$$

$F(t)$ is known the *limiting (or asymptotic) distribution*.
Note: To show this, we must show that $P(X_n < t) \to P(X < t)\ \forall\ t \in \mathbb{R}$.

Central Limit Theorem
The standardized sample mean converges in distribution to a standard normal distribution. More formally, define the standardized sample mean as:

$$Z = \frac{\bar{X} - \mathbb{E}(\bar{X})}{\hat{\sigma}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

Then, Central Limit Theorem states that $Z \xrightarrow{d} N(0,1)$. Using Slutsky's Theorem, we can show that this also means:

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Derivation:
We know from CLT that $Z \xrightarrow{d} N(0,1)$. Denote $Z' = N(0,1)$. Then, $Z \xrightarrow{d} Z'$. Using Slutsky's Theorem:

$$\sigma Z \xrightarrow{d} \sigma Z' = N(0, \sigma)$$

Within the context of this class, we care about Central Limit Theorem because it implies that **the sampling distribution of the sample mean will be normally distributed, centered at the population mean, with variance $\frac{\sigma^2}{n}$.**

4.2. Algebra Hints

Useful algebraic bits that show up here and there:

$$\begin{aligned}
\frac{n-1}{N-1} - \frac{n}{N} &= \frac{N(n-1)}{N(N-1)} - \frac{n(N-1)}{N(N-1)} \\
&= \frac{nN - N - nN + n}{N(N-1)} \\
&= \frac{-(N-n)}{N(N-1)} \\
&\text{By dividing everything by } N\text{:} \\
&= \frac{-(1 - \frac{n}{N})}{N-1} \\
&= -\frac{1}{N-1} \cdot \left(1 - \frac{n}{N}\right)
\end{aligned}$$

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 &= \sum_{i=1}^{n} y_i^2 - 2\bar{y}_n \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{y}_n^2 \\
&= \sum_{i=1}^{n} y_i^2 - 2\bar{y}_n(n\bar{y}_n) + n\bar{y}_n^2 \\
&= \sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2 \\
&= \sum_{i=1}^{n} y_i^2 - n \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 \\
&= \sum_{i=1}^{n} y_i^2 - n \cdot \frac{1}{n^2}\left(\sum_{i=1}^{n} y_i\right)^2 \\
&= \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2
\end{aligned}
$$