Emory University

QTM 220 Regression Analysis

Learning Notes

Jiuru Lyu

March 4, 2024

Contents

1 Statistical Inference

1.1 Descriptive Statistics and Binary Covariates

Definition 1.1.1 (Location). The *location* of the data is where it is. It is about approximating the data by a constant.

$$Y_i \approx \mu, \quad \text{for } i = 1, \dots, n$$

Example 1.1.2

Different ways to summarize location: mean, median

Definition 1.1.3 (Spread). The *spread* of the data is how far it tends to be from is location.

Definition 1.1.4 (Residuals). Spread summarizes the size of the *residuals* left over after constant approximation. We use $\widehat{\varepsilon}$ to denote residuals.

$$\varepsilon_i \coloneqq Y_i - \widehat{\mu}.$$

Definition 1.1.5 (Median Absolute Deviation and Standard Deviation).

- The *median absolute deviation (MAD)* is the median size of residuals.
- The *standard deviation (sd)* is the square root of the mean squared size of residuals.

  Remark 1.1 *The standard deviation is a sort of average in which big residuals count more than smaller ones.*

Definition 1.1.6 (Distribution). We use *histograms* to summarize the *distribution* of the data.

Remark 1.2 *Distribution of the data tells us more information than location and spread, but less than dot plot.* For example, in this context, dot plot also include the identities of the individuals in addition to the number of people having salary in the range.

Definition 1.1.7 (Binary Data). *Binary data* only have two options, and we usually denote those two options as 1's and 0's.

Corollary 1.8 : Hence, when drawing a dot plot, everyone falls into either of the two lines representing 1 and 0.

***Theorem 1.1.9** Location of Binary Data*

The median is whichever outcome is the most common, and the mean is the proportion of 1's in the data.

Remark 1.3 *Hence, a histogram tells us no more information than $\widehat{\mu}$.*

Theorem 1.1.10 *Spread of Binary Data*

- Median absolute deviation will always be 0 in a binary case.
- The standard deviation is the square root of the mean squared distance from the mean, and
$$\text{sd} = \sqrt{\widehat{\mu}(1-\widehat{\mu})}.$$

Proof 1. The claim concerning MAD is trivial. *Hint: there's only two possible values in the data, so median and MAD should always be the same.*

Now, let's consider the claim on standard deviation.

$$\begin{aligned}
\text{sd}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\widehat{\mu})^2 \\
&= \frac{1}{n}\sum_{y:\{0,1\}}\sum_{i:Y_i=y}(Y_i-\widehat{\mu})^2 \\
&= \frac{1}{n}\big\{N_1\big(1-\widehat{\mu}^2\big)+(n-N_1)\big(0-\widehat{\mu}^2\big)\big\} && [N_1 = \text{number of } 1\text{'s}] \\
&= \frac{1}{n}\big\{N_1\big(1-2\widehat{\mu}+\widehat{\mu}^2\big)+(n-N_1)\widehat{\mu}^2\big\} \\
&= \frac{1}{n}\big\{N_1-2N_1\widehat{\mu}+n\widehat{\mu}^2\big\} \\
&= \frac{1}{n}\big\{n\widehat{\mu}-2n\widehat{\mu}\cdot\widehat{\mu}+n\widehat{\mu}^2\big\} && [N_1 = n\widehat{\mu}] \\
&= \frac{1}{n}\big\{n\widehat{\mu}-n\widehat{\mu}^2\big\} \\
&= \widehat{\mu}-\widehat{\mu}^2 = \widehat{\mu}(1-\widehat{\mu}).
\end{aligned}$$

Therefore, we know
$$\text{sd} = \sqrt{\widehat{\mu}(1-\widehat{\mu})}.$$

■

Remark 1.4 *In binary data, knowing the mean $\equiv$ knowing everything else.*

1.2 Population Inference for a Proportion

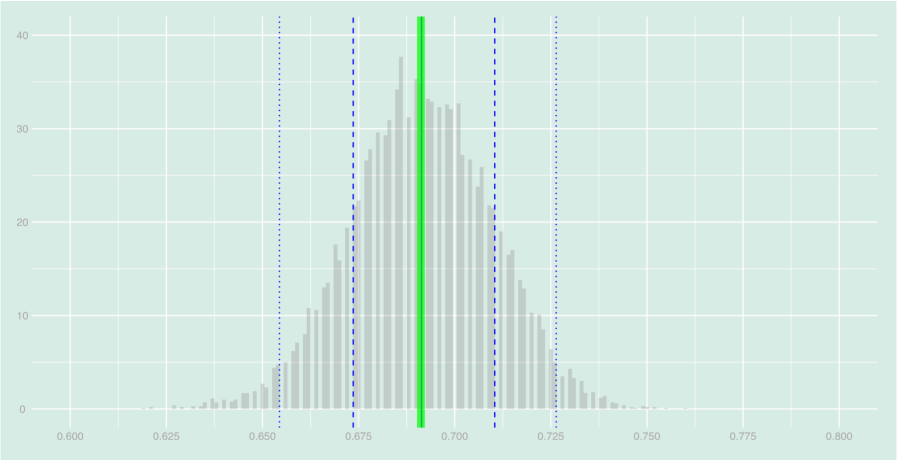
Definition 1.2.1 (Sampling Distribution). The *sampling distribution* is the distribution of estimates we'd get if we **replicated** our experiment over and over.

- Think of lots of people rolling the dice and reporting what they got.

- We consider this because it actually tells us something: it gives us an **interval** we can expect the proportion is in, and a statement about how much **confidence** we should have about it.

Example 1.2.2 Connecting Sample and Population

For each call i, we randomly select a voter with an id we'll call J_i. And we record as the call's outcome the turnout of the voter: $Y_i = y_{J_i}$. We can run this simulation using `R`.



- The *mean of the sampling distribution* is the solid blue line.
- The middle 2/3 of the sampling distribution lies between the dashed blue lines.
- The middle 95% of the sampling distribution lies between the dotted blue lines.
- Also, the population proportion is drawn as a wide green line.
- The question is "Could we predict how close we can get from the sampling before the election happened?" – Yes!
  - We will use an **interval estimate**: a *range of values* the population proportion is likely to be in.
  - The **width** of this interval speaks to the "how close" question.
  - The **coverage probability** (the probability we are right) qualifies this answer.
    * Our **point estimate** of the population proportion is the sample proportion $\overline{Y_n}$, where n is the size of the sample.
    * Now, we will try with some size of the interval. Say, x. Then, we are interested in the range of data $\overline{Y_n} \pm \frac{x}{2}$ (since the interval can be two-tailed).
    * Repeat the sampling process multiple times, say M times, and we notice that out of t times our interval "touches" the population proportion.

        * Then, we can define the coverage probability as follows:

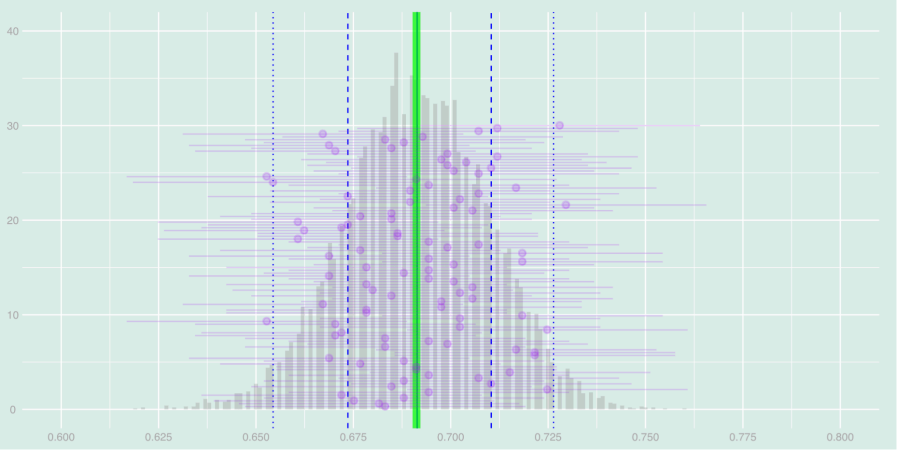
$$\text{coverage probability} = \frac{t}{M} = \mathbf{P}\Big(\overline{Y}_n \in \overline{y}_N \pm \frac{x}{2}\Big),$$

          where $\overline{Y}_n$ is our point estimate, $\overline{y}_N$ is the population proportion, and x is the width of the interval.

- Most of the time, we would like a 95% coverage probability, which means we will need to use a wider interval.

- Therefore, what we want to do is to choose a coverage probability and calculate the right width. An interval estimate like this (to ensure a given coverage) is called a **confidence interval**.

- The following figure shows a 95% coverage probability:



    – Our sample proportion 0.68 is close to the population proportion 0.69. Did we get luck? *No! In a million runs, almost all are within 0.05.*
    – Could we have predicted how close we would get before seeing the 0.69? *Yes! We can use a calibrated interval estimate – a Confidence Interval.*

- However, notice that this approach is not perfect: we cannot calibrate intervals like this in real life.

    – When we run our pool, we get a single point estimate $\overline{Y}_n$ based on our sample.
    – We don't know the sampling distribution of this point estimate until the election day.
    – However, what we actually do is almost the same: we will use an estimate of the sampling distribution in place of the thing itself.

1.3 Calibrating Interval Estimates

Theorem 1.3.1

An interval estimate covers the population proportion $\iff$ the corresponding point estimate is between the population proportion's arms.

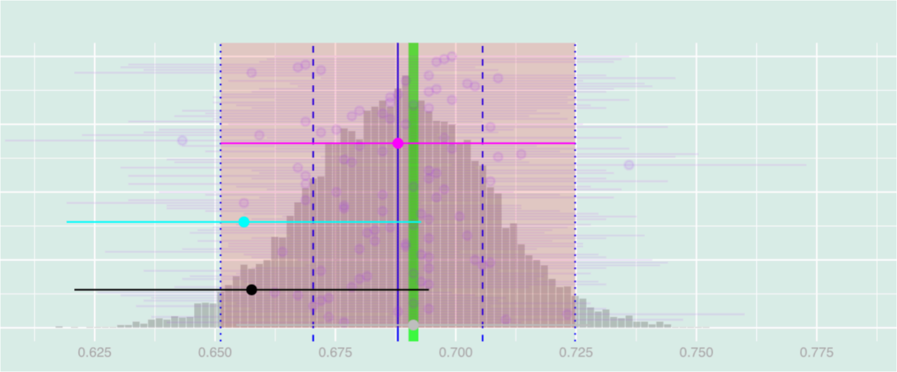
Remark 1.5 *With this Theorem, in Example 1.2.2, instead of looking at every interval and its arms to calculate coverage, we can draw arms of the same width around the population proportion.*

*Equivalently, we can calculate the **mass of the histogram** between the population proportion's arms.*

*However, even with this Theorem, the problem still exists: unless we've seen the population, we cannot run simulations. Hence, in reality we will do calibration using an **estimate of the sampling distribution**.*

Example 1.3.2

Here, we use our sample to estimate the sampling distribution. Compared with the actual population mean, the sample mean is a bit lower. Will this impact our estimation?



Solution 1.

It will not because we are not putting arms on draws from the estimated sampling distribution. We are putting arms on our point estimate, which is a draws from the actual sampling distribution. All that matters is the **width** of the estimated sampling distribution, and not the center. It turns out that the width calculated from the estimated sampling distribution and the population distribution should be the same (or close to the same). □

***Theorem 1.3.3** Binomial Distribution Estimation*

For a binary data type, we collected $Y_1, \cdots, Y_n$ as our sample. The distribution of the sample should follow a *Binomial Distribution*:

$$Y_i \sim \text{Binomial}(n, p).$$

Remark 1.6 *In theory, p should be calculated from the population. However, in the case of estimation, we will use our sample to estimate p. In the binary case, $p = \overline{Y}$.*

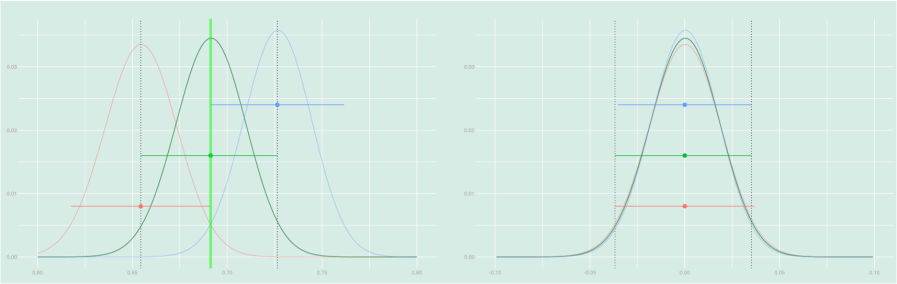
Binomial Distribution

```r
dbinom(x, n, p)
# x = number of heads, n = number of flips, p = probability of heads
```

Binomial Sample

```r
# To draw samples from estimated sampling distribution
samples <- rbinom(num, n, p)
# num = total number of draws,
# n = number of elements per drawing,
# p = probability of head
```

Example 1.3.4 How does the estimation work



The Binomial distribution is *continuous* as a function of p, so when p cahgnes little, the distribution changes little. That is to say that if we are not far off in the proportion, the estimated and actual sampling distributions are similar. The relevant difference (after centering) is even smaller because the way the binomial changes is mostly location.

This can be thought of a sort of "confidence interval" for our estimate of the sampling distribution. 95% of the time, we will get an estimate somewhere between the red and blue ones. As a result, a width of our interval estimate somewhere between the red and blue widths.

Example 1.3.5 The Bootstrap Interpretation

Let's revisit our distribution:

$$\text{Binomial}(n, p)/n.$$

This sample distribution indicates the proportion of 1's if we poll a sample of n object among whom the proportion of 1's is exactly $\overline{Y} = p$. This means that we can get a draw from our estimated sampling distribution by running a "poll" of the objects in our sample: rolling a n-sided die n times, calling up the corresponding object in our sample, and counting up the 1's we observe.

Boostraping Sample

```r
bootstrap.samples = array(dim=10000)
for (rr in 1:10000) {
    Y.boot = Y[sample(1:n, n, replace=TRUE] # Boostraping
    bootstrap.samples[rr] = sum(Y.boot)/n
}
```

Remark 1.7 (Bootstrap Sampling) *The usual way of sampling is to use a sample to approximate the population. Since one sample can only generate one estimate, we need many samples. However, we cannot do this in reality. Instead, we only have one sample, so we will use bootstrapping. In that way, we resample the sample multiple times to general different estimates. Using the resampling distribution, we can eventually approximate the population.*

Definition 1.3.6 (Normal Distribution). The *normal distribution* is a function of two parameters: its mean and its standard deviation:

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

Theorem 1.3.7 *Central Limit Theorem*

The sampling distribution and the bootstrap sampling distribution is approximately normal, and this is a consequence of the Central Limit Theorem.

1.4 Normal Approximation and Power Analysis

Previously, we have being talking about the context of estimating a **difference in proportions**. The difference does not have a sampling distribution with a simple parametric form (e.g., the binomial distribution), but we could use the bootstrapping to approximate its sampling distribution.

Theorem 1.4.1 *Approximation using a Normal Distribution*

A normal approximation $f_{\mu,\sigma}(x)$ for $\mu = p$ and $\sigma = \sqrt{\frac{p(1-p)}{n}}$ with $p = \overline{y}$ has a corresponding estimate using the sample:

$$\mu = \widehat{p}, \quad \widehat{\sigma} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \quad \text{with} \widehat{p} = \overline{Y}.$$

Theorem 1.4.2 *The Width of Normal Distributions*

Data Proportion	$\mu \pm \lambda\sigma$
68.3%	$\mu \pm 1\sigma$
95.4%	$\mu \pm 2\sigma$
99.7%	$\mu \pm 3\sigma$
95%	$\mu \pm 1.96\sigma$

Example 1.4.3 95% **Proportion**

By the CLT, we know $\overline{Y} \sim N(\mu, \sigma)$ for $\mu = \overline{y}$ and $\sigma = \sqrt{\overline{y}(1-\overline{y})/n}$. So, in roughly 95% of the polls, our estimate $\overline{Y}$ will be within 1.96 standard deviations of the population proportion $\overline{y}$. Converly, the population proportion $\overline{y}$ will be within 1.96 standard deviations of our estimate $\overline{Y}$. Or, mathematically,

$$\mu \in \overline{Y} \pm 1.96\sigma \quad \text{for } \sigma = \sqrt{\overline{y}(1-\overline{y})/n} \text{ in roughly 95\% of polls.} \tag{1}$$

However, Eq. (1) is not usable because we do not know the true $\overline{y}$. Therefore, we turn everything into estimate:

$$\mu \in \overline{Y} \pm 1.96\widehat{\sigma} \quad \text{for } \widehat{\sigma} = \sqrt{\overline{Y}(1-\overline{Y})/n} \text{ in roughly 95\% of polls.} \tag{95\% C.I.}$$

With Eq. (95% C.I.), we can avoid spending time doing bootstrap. That is, without bootstrapping, we can construct a 95% confidence interval from the sample mean.

Definition 1.4.4 (Power Analysis). Given the confidence interval (or, margin of error), we can find out the sample size needed.

Theorem 1.4.5

Suppose we want the estimate $\widehat{\mu}$ with a 95% confidence interval has a margin of error of d. Then, the smallest sample size to achieve so will be $n = \dfrac{1.96^2}{4d^2}$.

Proof 1. Suppose we want the estimate $\widehat{\mu}$ with a 95% confidence interval has a margin of error of d, then we know $\mu \in \overline{Y} \pm d = \overline{Y} \pm 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}$. Then, we know

$$d = 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}. \tag{2}$$

Here, though we are using $\overline{Y}$ to estimate $\widehat{p}$, it will not be accurate as different samples will have different $\overline{Y}$'s. So, we have to figure out another way to estimate $\widehat{p}$. Before we going to estimating $\widehat{p}$, let solve for n from Eq. (2) first:

$$\sqrt{\widehat{p}(1-\widehat{p})/n} = \frac{d}{1.96} \implies \widehat{p}(1-\widehat{p})/n = \left(\frac{d}{1.96}\right)^2 \implies n = \frac{1.96^2\widehat{p}(1-\widehat{p})}{d^2}.$$

Now, let's consider $f(\widehat{p}) = \widehat{p}(1-\widehat{p})$, a simple quadratic function, which will be maximized when $\widehat{p} = \dfrac{1}{2}$. Since we are conservative in our estimation, we want to find the "smallest" n that ensures the margin of error, so we take $\max f(\widehat{p}) = \dfrac{1}{4}$, and thus

$$n = \frac{1.96^2}{4d^2}.$$

■

2 Probability: Expectation and Variance

2.1 Probability Review

Theorem 2.1.1 *Properties of Expectations*

Linearity of Expectation Suppose Y, Z are random variables and $a, b \in \mathbb{R}$. Then

$$\mathbf{E}(aY + bZ) = \mathbf{E}(aY) + \mathbf{E}(bZ) = a\mathbf{E}(Y) + b\mathbf{E}(Z).$$

Multiplication Rules Suppose Y, Z are independent ($\perp\!\!\!\perp$) random variables, then

$$\mathbf{E}(Y \cdot Z) = \mathbf{E}(Y) \cdot \mathbf{E}(Z).$$

Remark 2.1 *Without special notice, we assume random variables are independent in this course.*

Theorem 2.1.2 *Variacne Decomposition*

If $Y \perp\!\!\!\perp Z$, then $\mathbf{Var}(Y + Z) = \mathbf{Var}(Y) + \mathbf{Var}(Z)$.

Proof 1. Notice that

$$\begin{aligned}
\mathbf{Var}(Y+Z) &= \mathbf{E}[(Y+Z)^2] - \mathbf{E}(Y+Z)^2 \\
&= \mathbf{E}(Y^2 + Z^2 + 2YZ) - [\mathbf{E}(Y) + \mathbf{E}(Z)]^2 \\
&= \mathbf{E}(Y^2) + \mathbf{E}(Z^2) + 2\mathbf{E}(YZ) - \mathbf{E}(Y)^2 - \mathbf{E}(Z)^2 - 2\mathbf{E}(Y)\mathbf{E}(Z) && \text{[Linearity]} \\
&= (\mathbf{E}(Y^2) - \mathbf{E}(Y)^2) + (\mathbf{E}(Z^2) - \mathbf{E}(Z)^2) + 2\mathbf{E}(YZ) - 2\mathbf{E}(YZ) && \text{[Independence]} \\
&= \mathbf{Var}(Y) + \mathbf{Var}(Z).
\end{aligned}$$

■

Theorem 2.1.3 *Binomial Expection and Variance*

If $Y \sim \text{Binomial}(n, p)$, then

$$\mathbf{E}(\overline{Y}) = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i)\right] = p$$

and

$$\mathbf{Var}(\overline{Y}) = \frac{p(1-p)}{n}.$$

The proof of these two quantities are omitted.

Definition 2.1.4 (Conditional Expectation). The *conditional expectation* is $\mathbf{E}[Y \mid X = x]$, namely the expected value of Y given that $X = x$.

Theorem 2.1.5 *Properties of Conditional Expectation*

Law of Iterated Expectations for any random variables X and Y, we have

$$\mathbf{E}(Y) = \mathbf{E}\{\mathbf{E}(Y \mid X)\}.$$

Irrelevance of Independent Conditioning Variables When $Z \perp\!\!\!\perp X$ and Y, we have

$$\mathbf{E}(Y \mid X, Z) = \mathbf{E}(Y \mid X).$$

In other words, if Z is unrelated to X and Y, holding it constant does not affect the relationship between X and Y.

2.2 Working with Expectations

Before we going into any details, let's review and introduce some notations.

Notation 2.1

- *The mean of our population:* $\mu = \frac{1}{m}\sum_{j=1}^{m} y_j$.
- *The mean of sample:* $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- *We use* y *to denote a number from the list, and* Y *to represent a random variable.*
- *To represent the subpopulation, we use* $X_i = x$.
- *The mean of the subpopulation* $X_i = x$ *is* $\mu(x) = \frac{1}{m_x}\sum_{j:x_j=x} y_j$ *where* $m_x = \sum_{j:x_j=x} 1$.
- *The mean of the subsample* $X_i = x$ *is* $\widehat{\mu}(x) = \frac{1}{N_x}\sum_{X_i=x} Y_i$*, where* $N_x = \sum_{i:X_i=x} 1$.

Definition 2.2.1 (Expectations). The *expected value* of a random variable Y_i is the **probability-weighted average** of the values it can take on.

$$\mathbf{E}[Y_i] = \sum_y \mathbf{P}(Y_i = y) \times y = \sum_y \left(\sum_{j:y_j=y} \frac{1}{m}\right) \times y = \frac{1}{m}\sum_y \sum_{j:y_j=y} y = \frac{1}{m}\sum_{j=1}^{m} y_j.$$

Definition 2.2.2 (Independence). We say random variables or *independent* ($\perp\!\!\!\perp$) if knowing the value of one does not tell us anything about the value of the other.

Corollary 2.3 : If Y_1 and Y_2 are independent,

$$\mathbf{P}(Y_1 = y_1 \text{ and } Y_2 = y_2) = \mathbf{P}(Y_1 = y_1)\mathbf{P}(Y_2 = y_2).$$

Remark 2.2 *When we draw a sample $Y_1, \dots, Y_n$, we claim $Y_1, \dots, Y_n$ are independent and identically distributed $(i.i.d.)$.*

Theorem 2.2.4 *The Law of Iterated Expectations*

For any random variables X and Y,

$$\mathbf{E}(Y) = \mathbf{E}\{\mathbf{E}(Y \mid X)\}.$$

Theorem 2.2.5 *Irrelevance of Independent Conditional Variables*

If X' is unrelated to X and Y, holding it constant does not affect the relationship between them. Suppose X' is independent of X and Y, then

$$\mathbf{E}(Y \mid X, X') = \mathbf{E}(Y \mid X).$$

Theorem 2.2.6 *Linearity of Expectations*

For random variables Y and Z, and numbers a and b,

$$\mathbf{E}(aY + bZ) = \mathbf{E}(aY) + \mathbf{E}(bZ) = a\mathbf{E}(Y) + b\mathbf{E}(Z).$$

Proof 1.

$$\begin{aligned}
\mathbf{E}(aY + bZ) &= \sum_y \sum_z (ay + bz)\mathbf{P}(Y = y, Z = z) && \text{def of expectation} \\
&= \sum_y \sum_z ay\mathbf{P}(Y = y, Z = z) + \sum_z \sum_y bz\mathbf{P}(Y = y, Z = z) \\
&= \sum_y ay \sum_z \mathbf{P}(Y = y, Z = z) + \sum_z bz \sum_y \mathbf{P}(Y = y, Z = z) \\
&= \sum_y ay\mathbf{P}(Y = y) + \sum_z bz \sum_y \mathbf{P}(Z = z) \\
&= a \sum_y y\mathbf{P}(Y = y) + \sum_z z \sum_y \mathbf{P}(Z = z) \\
&= a\mathbf{E}(Y) + b\mathbf{E}(Z)
\end{aligned}$$

■

Theorem 2.2.7 *Linearity of Conditional Expectations*

For random variables X, Y, Z, and some functions a and b of X,

$$\begin{aligned}\mathbf{E}\{a(X)Y + b(X)Z \mid X\} &= \mathbf{E}\{a(X)Y \mid X\} + \mathbf{E}\{b(X)Z \mid X\} \\ &= a(X)\mathbf{E}(Y \mid X) + b(X)\mathbf{E}(Z \mid X)\end{aligned}$$

Corollary 2.8 : When $Y \perp\!\!\!\perp Z$, then $\mathbf{E}[YZ] = \mathbf{E}[Y]\mathbf{E}[Z]$.

Proof 2.

$$\begin{aligned}\text{LHS} &= \mathbf{E}[YZ] \\ &= \mathbf{E}\{\mathbf{E}[YZ \mid Z]\} && \text{Law of Iterated Expectation} \\ &= \mathbf{E}[Z \cdot \mathbf{E}[Y \mid Z]] \\ &= \mathbf{E}\left[Z \cdot \underbrace{\mathbf{E}[Y]}_{\text{constant}}\right] && \text{Indepence} \\ &= \mathbf{E}[Y]\mathbf{E}[Z] = \text{RHS} && \text{Linearity}\end{aligned}$$

■

Theorem 2.2.9 *Indicator Trick*

When X is binary, we have

$$\mathbf{E}[X\mu(X)] = \mathbf{E}[X\mu(1)] = \mu(1)\mathbf{E}[X].$$

Similarly,

$$\mathbf{E}[(1-X)\mu(X)] = \mathbf{E}[(1-X)\mu(0)] = \mu(0)\mathbf{E}(1-X).$$

Definition 2.2.10 (Bias). The *bias* of an estimator is the difference between its expected value and the value of the thing it's estimating. When an estimator attains a bias of 0, we call it an *unbiased* estimator.

Claim 2.2.11

The sample mean is an unbiased estimator of the population mean: $\mathbf{E}[\widehat{\mu}] = \mu$.

Proof 3.

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[Y_i](\text{by linearity}) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\times n\times\mu = \mu.$$

■

Claim 2.2.12

The subsample mean is unbiased for the subpopulation mean: $\mathbf{E}[\widehat{\mu}(1)] = \mu(1)$.

Proof 4. Note that by definition, $\sum_{i=1}^{n} X_i Y_i = \begin{cases} Y_i & \text{if } X_i = 0 \\ 0 & \text{if } X_i = 0 \end{cases}$, so, $\widehat{\mu}(1) = \dfrac{\displaystyle\sum_{i=1}^{n} X_i Y_i}{\displaystyle\sum_{i=1}^{n} X_i}$. Then,

$$
\begin{aligned}
\mathbf{E}[\widehat{\mu}(1)] &= \mathbf{E}\left[\frac{\displaystyle\sum_{i=1}^{n} X_i Y_i}{\displaystyle\sum_{i=1}^{n} X_i}\right] \\
&= \mathbf{E}\left\{\mathbf{E}\left[\frac{\sum X_i Y_i}{\underbrace{\sum X_i}_{\text{constant}}}\right| X_1, \ldots, X_n\right\} && \text{Law of Iterated Expectation} \\
&= \mathbf{E}\left\{\frac{\mathbf{E}[\sum X_i Y_i \mid X_1, \ldots, X_n]}{\sum X_i}\right\} && \text{Linearity} \\
&= \mathbf{E}\left\{\frac{\sum \left(X_i \mathbf{E}[Y_i \mid X_1, \ldots, X_n]\right)}{\sum X_i}\right\} && \text{Linearity} \\
&= \mathbf{E}\left\{\frac{\sum \left(X_i \overbrace{\mathbf{E}[Y_i \mid X_i]}^{=\mu(X_i)}\right)}{\sum X_i}\right\} && \text{Irrelavent Expectation: } Y_i \leftrightsquigarrow X_i \text{ only} \\
&= \mathbf{E}\left\{\frac{\sum (X_i \mu(X_i))}{\sum X_i}\right\} \\
&= \mathbf{E}\left\{\frac{\sum (X_i \mu(1))}{\sum X_i}\right\} && \text{Indicator Trick} \\
&= \mathbf{E}\left\{\frac{\mu(1) \sum X_i}{\sum X_i}\right\} && \text{Linearity} \\
&= \mathbf{E}(\mu(1)) = \mu(1) \implies \text{unbiasedness}
\end{aligned}
$$

■

Claim 2.2.13

The difference in subsample means is unbiased for the difference in subpopulation means: $\mathbf{E}[\widehat{\mu}(1) - \widehat{\mu}(0)] = \mu(1) - \mu(0)$.

Proof 5. This follows from the linearity of expectations and unbiasedness of the subsample means.

$$\mathbf{E}[\widehat{\mu}(1) - \widehat{\mu}(0)] = \mathbf{E}[\widehat{\mu}(1)] - \mathbf{E}[\widehat{\mu}(0)] = \mu(1) - \mu(0).$$

■

2.3 Working with Variance

Definition 2.3.1 (Spread and Variance). To summarize the *spread* of a distribution, we talk about variance. The *variance* is an expectation: it is the **expected squared difference** between a random variable and its expectation:

$$\mathbf{Var}(Y) = \mathbf{E}\left[\{Y - \mathbf{E}(Y)\}^2\right].$$

Definition 2.3.2 (Standard Deviation). Standard deviation is the square root of the variance:

$$\mathrm{sd} = \sqrt{\mathbf{Var}(Y)}.$$

Theorem 2.3.3 *Variance as Excess*

We can think of variance in the idea of spread (the mean square of a centered version of Y), but variance can also be defined as the excess: the average amount Y^2 exceeds the square of its mean. So, variance can also be defined as

$$\mathbf{Var}(Y) = \mathbf{E}\big(Y^2\big) - \{\mathbf{E}(Y)\}^2.$$

Proof 1.

$$\begin{aligned}
\mathbf{E}\left[\{Y - \mathbf{E}(Y)\}^2\right] &= \mathbf{E}\left[Y^2 + \mathbf{E}(Y)^2 - 2Y\mathbf{E}(Y)\right] && \text{Complete Square} \\
&= \mathbf{E}\big(Y^2\big) + \mathbf{E}\left(\underbrace{\mathbf{E}(Y)^2}_{\text{constant}}\right) - 2\mathbf{E}\left(Y\ \underbrace{\mathbf{E}(Y)}_{\text{constant}}\right) && \text{Linearity} \\
&= \mathbf{E}\big(Y^2\big) + \mathbf{E}(Y)^2 - 2\mathbf{E}(Y)\mathbf{E}(Y) && \text{Linearity} \\
&= \mathbf{E}\big(Y^2\big) + \mathbf{E}(Y)^2 - 2\mathbf{E}(Y)^2 \\
&= \mathbf{E}\big(Y^2\big) - \mathbf{E}(Y)^2.
\end{aligned}$$

■

Definition 2.3.4 (Conditional Variance). The *conditional variance* describes the spread of one random variable within groups. The *conditional variance function* is defined as

$$\sigma^2(x) := \mathbf{Var}(Y \mid X = x) = \mathbf{E}\big[\{Y - \mathbf{E}(Y \mid X)\}^2 \mid X = x\big]$$

- This is the variance of Y, within the subpopulation $X = x$.
- It is the conditional expectation function of random variable $\{Y - \mathbf{E}(Y \mid X)\}^2$.
- We call it *the conditional variance of* Y *given* $X = x$.

Therefore, the *conditional variance* is defined by

$$\sigma^2(X) := \mathbf{Var}(Y \mid X) = \mathbf{E}\big[\{Y - \mathbf{E}(Y \mid X)\}^2 \mid X\big].$$

- This is the variance of Y within a random subpopulation of people.
  - It is the conditional variance function evaluated at the random variable X.
  - And the conditional expectation of the random variable $\{Y - \mathbf{E}(Y \mid X)\}^2$.
- We call it *the conditional variance of* Y *given* X.

Theorem 2.3.5 *Law of Total Variance*

The law of total variance relates conditional and unconditional variance:

$$\mathbf{Var}[Y] = \mathbf{E}\{\mathbf{Var}(Y \mid X)\} + \mathbf{Var}\{\mathbf{E}(Y \mid X)\}.$$

This is a useful way to decompose the variance of a random variable.

Proof 2.

$$\begin{aligned}
\text{RHS} &= \mathbf{E}\big[\mathbf{E}\big(Y^2 \mid X\big) - \{\mathbf{E}(Y \mid X)\}^2\big] + \mathbf{E}\big(\{\mathbf{E}(Y \mid X)\}^2\big) - [\mathbf{E}(\mathbf{E}(Y \mid X))]^2 \\
&= \mathbf{E}\big(\mathbf{E}\big(Y^2 \mid X\big)\big) - \mathbf{E}\big(\mathbf{E}(Y \mid X)^2\big) + \mathbf{E}(Y \mid X)^2 - [\mathbf{E}(\mathbf{E}(Y \mid X))]^2 \\
&= \mathbf{E}\big(\mathbf{E}\big(Y^2 \mid X\big)\big) - \mathbf{E}(Y \mid X)^2 + \mathbf{E}(Y \mid X)^2 - [\mathbf{E}(\mathbf{E}(Y \mid X))]^2 \\
&= \mathbf{E}(Y^2) - \mathbf{E}(Y)^2 \qquad \text{[Law of Iterated Expectation]} \\
&= \mathbf{Var}[Y] = \text{LHS}
\end{aligned}$$

■

Theorem 2.3.6 *The Variance of the Mean*

Suppose $\overline{Y}$ is the sample mean, n is the sample size, and σ is the population standard deviation, then $\mathbf{Var}\big(\overline{Y}\big) = \dfrac{\sigma^2}{n}$.

Proof 3.

$$\begin{aligned}
\mathbf{Var}\left(\overline{Y}\right) &= \mathbf{E}\left[\left\{\frac{1}{n}\sum_i Y_i - \mathbf{E}\left(\frac{1}{n}\sum_i Y_i\right)\right\}^2\right] && \text{definition}\\
&= \mathbf{E}\left[\left\{\frac{1}{n}\sum_i (Y_i - \mathbf{E}(Y_i))\right\}^2\right] && \text{Linearity}\\
&= \mathbf{E}\left[\left\{\frac{1}{n}\sum_i Z_i\right\}^2\right] && \text{define } Z_i \equiv Y_i - \mathbf{E}(Y_i)\\
&= \mathbf{E}\left[\frac{1}{n^2}\sum_i\sum_j Z_i Z_j\right] && \text{Complete Square}
\end{aligned}$$

Lemma *If* $Z_i \perp\!\!\!\perp Z_j$*, then* $\mathbf{E}[Z_iZ_j] = \mathbf{E}[Z_i]\mathbf{E}[Z_j] = 0$.

Proof.

$$\begin{aligned}
\mathbf{E}[Z_iZ_j] &= \mathbf{E}[Z_i]\mathbf{E}[Z_j] && \textit{Independence}\\
&= \mathbf{E}[Y_i - \mathbf{E}(Y_i)]\mathbf{E}[Y_j - \mathbf{E}(Y_j)] && \textit{Definition of } Z_i\\
&= \left(\mathbf{E}[Y_i] - \mathbf{E}\left[\overbrace{\mathbf{E}(Y_i)}^{\textit{constant}}\right]\right)\left(\mathbf{E}[Y_j] - \mathbf{E}\left[\overbrace{\mathbf{E}(Y_j)}^{\textit{constant}}\right]\right)\\
&= (\mathbf{E}(Y_i) - \mathbf{E}(Y_i))(\mathbf{E}(Y_j) - \mathbf{E}(Y_j))\\
&= 0.
\end{aligned}$$

The proof is thereby completed. □

Further note that

$$\begin{aligned}
\mathbf{Var}\left(\overline{Y}\right) &= \mathbf{E}\left[\frac{1}{n^2}\sum_i\sum_j Z_iZ_j\right]\\
&= \frac{1}{n^2}\sum_i\sum_j \mathbf{E}[Z_iZ_j] && \text{Linearity}\\
&= \frac{1}{n^2}\sum_i\left(\sum_{j\neq i}\mathbf{E}\left[Z_i^2\right] + \mathbf{E}[Z_iZ_j]\right) && \text{Take out terms with } i = j\\
&= \frac{1}{n^2}\sum_i\left(\sum_{j\neq i}\mathbf{E}\left[Z_i^2\right] + \mathbf{E}[Z_i]\mathbf{E}[Z_j]\right)
\end{aligned}$$

By the Lemma, we know $\mathbf{E}[Z_i]\mathbf{E}[Z_j] = 0$ as $Z_i \perp\!\!\!\perp Z_j$ whenever $i \neq j$. Therefore,

$$\begin{aligned}
\mathbf{Var}\left(\overline{Y}\right) &= \frac{1}{n^2}\sum_i \left(\sum_{j\neq i}\mathbf{E}\left[Z_i^2\right] + 0\right) && \text{By Lemma}\\
&= \frac{1}{n^2}\sum_i \left(\sum_{j\neq i}\mathbf{E}\left[Z_i^2\right]\right) && \text{Not related to } j\\
&= \frac{1}{n^2}\sum_i \mathbf{E}\left[Z_i^2\right]\\
&= \frac{1}{n^2}\cdot n \cdot \mathbf{E}\left[(Y_i - \mathbf{E}[Y_i])^2\right] && \text{Definition of } Z_i\\
&= \frac{1}{n}\mathbf{Var}(Y_i) && \text{Definition of Variance}\\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

■

Claim 2.3.7

The Variance of our point estimate is the **expected value** of the subpopulation variance divided by the number of poeple in the sub sample. That is,

$$\mathbf{Var}[\widehat{\mu}(1)] = \mathbf{E}\left[\frac{\sigma^2(1)}{N_1}\right] \quad \text{for} \quad N_1 = \sum_{i=1}^{n} X_i.$$

Proof 4. Recall that $\widehat{\mu}(1)$ is conditionally unbiased. That is, $\mathbf{E}(\widehat{\mu}(1) \mid X_1, \ldots, X_n) = \mu(1)$. So, by the law of total variance,

$$\begin{aligned}
\mathbf{Var}[\widehat{\mu}(1)] &= \mathbf{E}[\mathbf{Var}\{\widehat{\mu}(1) \mid X_1, \ldots, X_n\}] + \mathbf{Var}[\mathbf{E}\{\widehat{\mu}(1) \mid X_1, \ldots, X_n\}]\\
&= \mathbf{E}[\mathbf{Var}\{\widehat{\mu}(1) \mid X_1, \ldots, X_n\}] + \mathbf{Var}\left[\underbrace{\mu(1)}_{\text{constant}}\right]\\
&= \mathbf{E}[\mathbf{Var}\{\widehat{\mu}(1) \mid X_1, \ldots, X_n\}] && \text{variance of constants=0}\\
&= \mathbf{E}\left[\mathbf{E}\left\{(\widehat{\mu}(1) - \mu(1))^2 \mid X_1, \ldots, X_n\right\}\right] && \text{definition of variance}\\
&= \mathbf{E}\left[\mathbf{E}\left\{\left(\frac{\sum X_i Y_i}{\sum X_i} - \frac{\mu(1)\sum X_i}{\sum X_i}\right)^2 \mid X_1, \ldots, X_n\right\}\right]\\
&= \mathbf{E}\left[\mathbf{E}\left\{\left(\frac{\sum X_i (Y_i - \mu(1))^2}{\sum X_i}\right)^2 \mid X_1, \ldots, X_n\right\}\right]\\
&= \mathbf{E}\left[\frac{\mathbf{E}\left\{\left[\sum X_i (Y_i - \mu(1)\right]^2\right\} \mid X_1, \ldots, X_n}{(\sum X_i)^2}\right]
\end{aligned}$$

Define $Z_i = X_i\{Y_i - \mu(1)\}$, then

$$\left(\sum_i Z_i\right)^2 = \left(\sum_i Z_i\right)\left(\sum_j Z_j\right) = \sum_i \sum_j Z_i Z_j.$$

Then,

$$\begin{aligned}
\mathbf{Var}[\widehat{\mu}(1)] &= \mathbf{E}\left[\frac{\mathbf{E}\{[\sum X_i(Y_i - \mu(1)]^2\} \mid X_1, \ldots, X_n}{(\sum X_i)^2}\right] \\
&= \mathbf{E}\left[\frac{\mathbf{E}\left\{\left(\sum_i Z_i\right)^2 \mid X_1, \ldots, X_n\right\}}{(\sum X_i)^2}\right] \\
&= \mathbf{E}\left[\frac{\mathbf{E}\left\{\sum_i \sum_j Z_i Z_j \mid X_1, \ldots, X_n\right\}}{(\sum X_i)^2}\right] \\
&= \mathbf{E}\left[\frac{\sum_i \sum_i \mathbf{E}(Z_i Z_j \mid X_1, \ldots, X_n)}{(\sum X_i)^2}\right] && \text{Linearity} \\
&= \mathbf{E}\left[\frac{\sum_i \sum_j \mathbf{E}(Z_i Z_j \mid X_i, X_j)}{(\sum X_i)^2}\right] && \text{Condition on Irrelevance}
\end{aligned}$$

Lemma *Suppose* $Z_j = Y_j - \mu(X_j)$, *then* $\mathbf{E}(Z_j \mid X_j) = 0$.

Proof.

$$\begin{aligned}
\mathbf{E}(Z_j \mid X_j) &= \mathbf{E}[X_j(Y_j - \mu(X_j))] \\
&= \mathbf{E}\{\mathbf{E}[X_j(Y_j - \mu(1) \mid X_j]\} \\
&= \mathbf{E}\{X_j \mathbf{E}\{[Y_j \mid X_j] - \mu(X_j\}\} \\
&= \mathbf{E}\{X_j \cdot \{\mu(X_j) - \mu(X_j\}\} = 0
\end{aligned}$$

The lemma is therefore proved as desired. □

Claim *Suppose* $Z_j = Y_j - \mu(X_j)$, *then* $\mathbf{E}\{Z_i Z_j \mid X_i, X_j\} = 0$.

Proof.

$$\begin{aligned}
\mathbf{E}\{Z_iZ_j \mid X_i, X_j\} &= \mathbf{E}\{\mathbf{E}\{Z_iZ_j \mid Z_i, X_i, X_j\}\} && \textit{Law of Iterated Expectation}\\
&= \mathbf{E}\{Z_i\mathbf{E}\{Z_j \mid X_i, X_j\}\} && \textit{Constant}\\
&= \mathbf{E}\{Z_i\mathbf{E}\{Z_j \mid X_j\}\} && \textit{Independence}\\
&= \mathbf{E}\{Z_i \cdot 0\} = 0 && \textit{By Lemma}
\end{aligned}$$

Hence, the proof is completed. □

Therefore,

$$\begin{aligned}
\mathbf{Var}(\widehat{\mu}(1)) &= \mathbf{E}\left[\frac{\sum_i \sum_j \mathbf{E}(Z_iZ_j \mid X_i, X_j)}{\left(\sum X_i\right)^2}\right] && Z_i = Y_i - \mu(X_i)\\
&= \mathbf{E}\left[\frac{\sum \mathbf{E}(Z_i^2 \mid X_i)}{\left(\sum X_i\right)^2}\right] && \text{By Claim, Corss-Terms=0}
\end{aligned}$$

Claim *Suppose* $Z_i = X_i(Y_i - \mu(X_i))$*, then* $\mathbf{E}(Z_i^2 \mid X_i) = X_i\sigma^2(1)$*.*

Proof.

$$\begin{aligned}
\mathbf{E}\big(Z_i^2 \mid X_i\big) &= \mathbf{E}\big[X_i\{Y_i - \mu(X_i)\}^2 \mid X_i\big]\\
&= X_i\mathbf{E}\big[(Y_i - \mu(X_i))^2 \mid X_i\big]\\
&= X_i\mathbf{Var}(Y_i \mid X_i)\\
&= X_i\sigma^2(X_i)\\
&= X_i\sigma^2(1).
\end{aligned}$$

The claim is therefore proved. □

Then, by the Claim and Linearity of Expectation, we have

$$\begin{aligned}
\mathbf{Var}(\widehat{\mu}(1)) = \mathbf{E}\left[\frac{\sum \mathbf{E}(Z_i^2 \mid X_i)}{\left(\sum X_i\right)^2}\right] &= \mathbf{E}\left[\frac{\sum X_i \overbrace{\sigma^2(1)}^{\text{constant}}}{\left(\sum X_i\right)^2}\right]\\
&= \mathbf{E}\left[\frac{\sigma^2(1)\sum X_i}{\left(\sum X_i\right)^2}\right]\\
&= \mathbf{E}\left[\frac{\sigma^2(1)}{\sum X_i}\right].
\end{aligned}$$

Recall that $\sum X_i = N_1$, substitute, by the Linearity of Expectation, we have

$$\mathbf{Var}(\widehat{\mu}(1)) = \mathbf{E}\left[\frac{\overbrace{\sigma^2(1)}^{\text{constant}}}{N_1}\right] = \sigma^2(1) \cdot \mathbf{E}\left[\frac{1}{N_1}\right].$$

■

Claim 2.3.8

The variance of the difference in subsample means is the **sum** of the variances of the subsample means themselves.

$$\mathbf{Var}[\widehat{\mu}(1) - \widehat{\mu}(0)] = \mathbf{E}\left[\frac{\sigma^2(1)}{N_1} + \frac{\sigma^2(0)}{N_0}\right] \quad \text{for } N_x = \sum_{i=1}^{n} \mathbb{1}_x(X_i), \text{ where } \mathbb{1}_x = \begin{cases} 1 & \text{if } X_i = x \\ 0 & \text{o/w} \end{cases}$$

Remark 2.3 *This claim is crucial for power calculations and analysis.*

3 Causal Inference

3.1 Introduction to Randomized Experiments

Definition 3.1.1 (Causal Inference). The formal approach to *casual inference* reframes the problem of causal inference in terms of counterfactuals which are treated as missing data.

Notation 3.1 *Potential Outcomes and Treatment Effects*

- *Observable Stuff*
  - Y *is an outcome variable*
  - D *is a treatment variable*
- *Conceptual Stuff*
  - *Potential Outcomes*
    - $Y_i(1)$ *is the outcome individual* i *would have attained if they'd received treatment* $D_i = 1$.
    - $Y_i(0)$ *is the outcome individual* i *would have attained if they'd not received treatment* $D_i = 0$.
  - *Individual-level Treatment Effect: It is the* difference *in what would happen with the two treatments. We call this* τ_i. *Naturally, we can never observe this.*

Example 3.1.2

	In	Counterfactual	World	In	Reality
i	$Y_i(1)$	$Y_i(0)$	τ_i	D_i	Y_i
Individual #	Vote with Letter	Vote w/o Letter	Effect	Received Letter	Vote
1	1	0	1	1	1
2	1	1	0	1	1
3	0	0	0	0	0
4	0	1	-1	0	1
$\vdots$					

In the table above, black are what is observable (happening in the reality), and orange are what is not observable (happening in the counterfactual world).

Remark 3.1 *A sample without counterfactual information can tell us about averages of quantities involving counterfactual.*

Theorem 3.1.3 *Randomized Experiments*

Fundamental Problem of Causal Inference: We cannot observe the individuals' outcome under treatment and control at the same time.

Solution: Randomized experiments. This means that we can think of the treated and untreated subsamples drawn from our population by a "coin-flipping" process. If we ignore the stuff downstream of treatment assignment, we can therefore consider people in those two subsamples have the same distribution.

To estimate the treatment effect,

$$\text{ATE} = \mathbf{E}[\tau] = \mathbf{E}[Y(1) - Y(0)] = \mathbf{E}[Y(1)] - \mathbf{E}[Y(0)].$$

However, here, we have to consider the subpopulations. If we take sample large enough and perform the randomization well enough, the following is true:

$$\mathbf{E}[Y_i(1) \mid D_i = 1] \approx \mathbf{E}[Y_i(1)] \quad \text{and} \quad \mathbf{E}[Y_i(0) \mid D_i = 0] \approx \mathbf{E}[Y_i(0)].$$

Therefore, the average treatment effect can be estimated as

$$\widehat{\text{ATE}} = \mathbf{E}[\widehat{\tau}] = \mathbf{E}[Y_i(1) \mid D_i = 1] - \mathbf{E}[Y_i(0) \mid D_i = 0].$$

Remark 3.2 *The key of randomization is to make the treated subgroup and untreated subgroup representative of the population treated outcome and untreated outcome.*

Theorem 3.1.4 *Expectation and Variance of Treatment Effect*

In a randomized experiments, let's use $\widehat{\mu}(1)$ to denote the sample mean of the treated group and $\widehat{\mu}(0)$ to denote the sample mean of the untreated group. Then,

$$\mathbf{E}[\widehat{\tau}] = \mathbf{E}[\widehat{\mu}(1) - \widehat{\mu}(0)] = \mathbf{E}[\widehat{\mu}(1)] - \mathbf{E}[\widehat{\mu}(0)].$$

Further, suppose $\sigma^2(0)$ is the sample variance of the treated group and $\sigma^2(1)$ the sample variance of the untreated group, then

$$\mathbf{Var}[\widehat{\tau}] = \mathbf{E}\left[\frac{\sigma^2(1)}{N_1} + \frac{\sigma^2(0)}{N_0}\right],$$

where $N_1 = \sum_{i=1}^{n} X_i$, $N_0 = \sum_{i=1}^{n} (1 - X_i)$, and $n = N_1 + N_2$.

Theorem 3.1.5 *Condition for Variance of Treatment Effect to be Minimized*

When $N_1 = N_0 = \dfrac{n}{2}$, we attain the minimum for the variance of treatment effect.