Causality, Potential Outcomes, and the Estimation of Treatment Effects in Randomized Studies

2020 AEA Continuing Education Program
Mastering Mostly Harmless Econometrics

Alberto Abadie
MIT

Contents

1. Causality, counterfactuals and potential outcomes
2. Randomized experiments, Fisher's exact test
3. Threats to internal and external validity in randomized experiments
4. Appendix: Experimental design

Purpose, scope, and examples

The goal of **policy/program evaluation** is to assess the causal effect of policy interventions. Examples:

- Job training programs on earnings and employment
- Class size on test scores
- Minimum wage on employment
- Tax-deferred saving programs on savings accumulation

More generally, we may be interested in the effect of interventions that are not public policies. Examples:

- Interest rate on credit card usage
- Incentive schemes on employee productivity

Causality with potential outcomes

Treatment

D_i: Indicator of treatment intake for *unit i*

$$D_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise.} \end{cases}$$

Outcome

Y_i: Observed outcome variable of interest for unit i

Potential Outcomes

Y_{0i} and Y_{1i}: Potential outcomes for unit i

$$\begin{array}{ll} Y_{1i}: & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_{0i}: & \text{Potential outcome for unit } i \text{ without treatment} \end{array}$$

Causality with potential outcomes

Treatment Effect

The treatment effect or causal effect of the treatment on the outcome for unit i is the difference between its two potential outcomes:

$$Y_{1i} - Y_{0i}$$

Observed Outcomes

Observed outcomes are realized as

$$Y_i = Y_{1i} D_i + Y_{0i}(1 - D_i) \quad \text{or} \quad Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

Fundamental Problem of Causal Inference

Cannot observe both potential outcomes (Y_{1i}, Y_{0i})

Stable unit treatment value assumption (SUTVA)

Assumption

Observed outcomes are realized as

$$Y_i = Y_{1i} D_i + Y_{0i}(1 - D_i)$$

- Implies that potential outcomes for unit i are unaffected by the treatment of unit j
- Rules out interference across units
- Example: Effect of flu vaccine on hospitalization
- This assumption may be problematic, so we should choose the units of analysis to minimize interference across units.

Quantities of interest (estimands)

ATE

Average treatment effect is:

$$\alpha_{ATE} = E[Y_1 - Y_0]$$

ATET

Average treatment effect on the treated is:

$$\alpha_{ATET} = E[Y_1 - Y_0 | D = 1]$$

Average treatment effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$Y_{1i} - Y_{0i}$
1	3	0	3	1	3
2	1	1	1	1	0
3	1	0	0	0	1
4	1	1	1	0	0
$E[Y_1]$	1.5				
$E[Y_0]$		0.5			
$E[Y_1 - Y_0]$				1	

$$\alpha_{ATE} = E[Y_1 - Y_0] = 3 \cdot (1/4) + 0 \cdot (1/4) + 1 \cdot (1/4) + 0 \cdot (1/4) = 1$$

Average treatment effect on the treated (ATET)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$Y_{1i} - Y_{0i}$
1	3	0	3	1	3
2	1	1	1	1	0
3	1	0	0	0	1
4	1	1	1	0	0
$E[Y_1 \vert D = 1]$	2				
$E[Y_0 \vert D = 1]$		0.5			
$E[Y_1 - Y_0 \vert D = 1]$					1.5

$$\alpha_{ATET} = E[Y_1 - Y_0|D = 1] = 3 \cdot (1/2) + 0 \cdot (1/2) = 1.5$$

Selection bias

Problem

Comparisons of earnings for the treated and the untreated do not usually give the right answer:

$$\begin{aligned} E[Y|D = 1] - E[Y|D = 0] &= E[Y_1|D = 1] - E[Y_0|D = 0] \\ &= \underbrace{E[Y_1 - Y_0|D = 1]}_{ATET} + \underbrace{\{E[Y_0|D = 1] - E[Y_0|D = 0]\}}_{BIAS} \end{aligned}$$

- Selection into treatment often depends on potential outcomes
- Bias term may be positive or negative depending on the setting

Selection bias

Problem

Comparisons of earnings for the treated and the untreated do not usually give the right answer:

$$
\begin{aligned}
E[Y|D=1] - E[Y|D=0] &= E[Y_1|D=1] - E[Y_0|D=0] \\
&= \underbrace{E[Y_1 - Y_0|D=1]}_{ATET} + \underbrace{\{E[Y_0|D=1] - E[Y_0|D=0]\}}_{BIAS}
\end{aligned}
$$

Example: Job training program for disadvantaged

- participants are self-selected from a subpopulation of individuals in difficult labor situations
- post-training period earnings would be lower for participants than for nonparticipants in the absence of the program ($E[Y_0|D=1] - E[Y_0|D=0] < 0$)

Training program for the disadvantaged in the U.S.

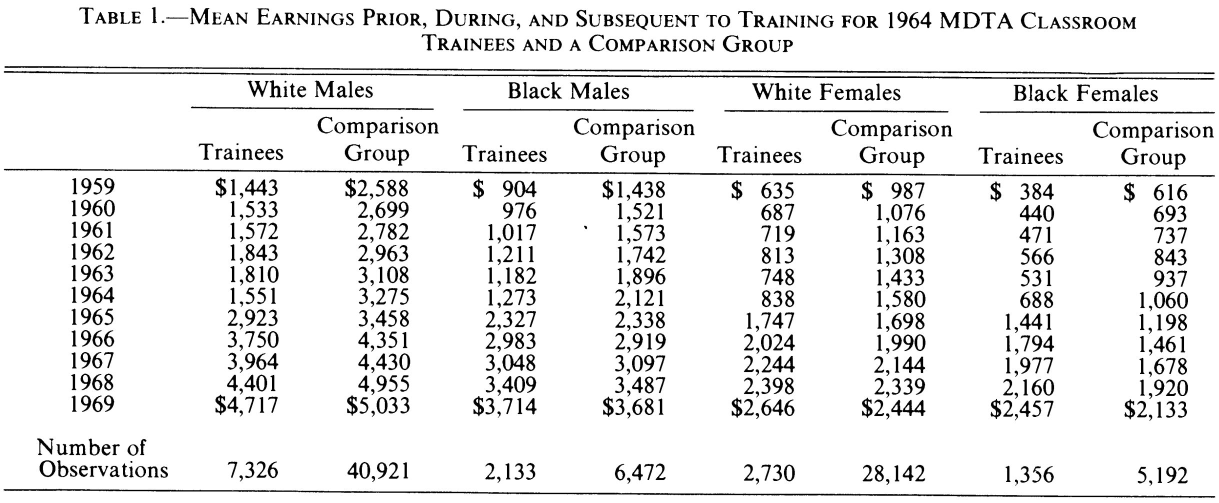


TABLE 1.—MEAN EARNINGS PRIOR, DURING, AND SUBSEQUENT TO TRAINING FOR 1964 MDTA CLASSROOM TRAINEES AND A COMPARISON GROUP

	White Males		Black Males		White Females		Black Females	
	Trainees	Comparison Group	Trainees	Comparison Group	Trainees	Comparison Group	Trainees	Comparison Group
1959	$1,443	$2,588	$ 904	$1,438	$ 635	$ 987	$ 384	$ 616
1960	1,533	2,699	976	1,521	687	1,076	440	693
1961	1,572	2,782	1,017	1,573	719	1,163	471	737
1962	1,843	2,963	1,211	1,742	813	1,308	566	843
1963	1,810	3,108	1,182	1,896	748	1,433	531	937
1964	1,551	3,275	1,273	2,121	838	1,580	688	1,060
1965	2,923	3,458	2,327	2,338	1,747	1,698	1,441	1,198
1966	3,750	4,351	2,983	2,919	2,024	1,990	1,794	1,461
1967	3,964	4,430	3,048	3,097	2,244	2,144	1,977	1,678
1968	4,401	4,955	3,409	3,487	2,398	2,339	2,160	1,920
1969	$4,717	$5,033	$3,714	$3,681	$2,646	$2,444	$2,457	$2,133
Number of Observations	7,326	40,921	2,133	6,472	2,730	28,142	1,356	5,192

Assignment mechanism

Assignment mechanism

Assignment mechanism is the procedure that determines which units are selected for treatment intake. Examples include:

- random assignment
- selection on observables
- selection on unobservables

Typically, treatment effects models attain identification by restricting the assignment mechanism in some way.

Key ideas

- Causality is defined by potential outcomes, not by realized (observed) outcomes
- Observed association is neither necessary nor sufficient for causation
- Estimation of causal effects of a treatment (usually) starts with studying the assignment mechanism

Selection bias

Recall the selection problem when comparing the mean outcomes for the treated and the untreated:

$$\underbrace{E[Y|D=1]-E[Y|D=0]}_{\text{Difference in Means}} = E[Y_1|D=1]-E[Y_0|D=0]$$
$$= \underbrace{E[Y_1-Y_0|D=1]}_{\text{ATET}} + \underbrace{\{E[Y_0|D=1]-E[Y_0|D=0]\}}_{\text{BIAS}}$$

- Random assignment of units to the treatment forces the selection bias to be zero
- The treatment and control group will tend to be similar along all characteristics (including Y_0)

Identification in randomized experiments

Randomization implies:

$$(Y_1, Y_0) \;\text{ independent of } D, \quad \text{or} \quad (Y_1, Y_0) \perp\!\!\!\perp D.$$

We have that $E[Y_0|D=1]=E[Y_0|D=0]$ and therefore

$$\alpha_{ATET} = E[Y_1 - Y_0|D=1] = E[Y|D=1]-E[Y|D=0]$$

Also, we have that

$$\alpha_{ATE} = E[Y_1-Y_0] = E[Y_1-Y_0|D=1] = E[Y|D=1]-E[Y|D=0]$$

As a result,

$$\underbrace{E[Y|D=1]-E[Y|D=0]}_{\text{Difference in Means}} = \alpha_{ATE} = \alpha_{ATET}$$

Identification in randomized experiments

The identification result extends beyond average treatment effects. Let $Q_\theta(Y)$ be the θ-th quantile of the distribution of Y:

$$\Pr(Y \leq Q_\theta(Y)) = \theta.$$

Given random assignment, $Y_0 \perp\!\!\!\perp D$. Therefore,

$$Y_0 \ \sim \ Y_0|D = 0 \ \sim \ Y|D = 0$$

where "$\sim$" means "has the same distribution as". Similarly,

$$Y_1 \ \sim \ Y|D = 1.$$

So effect of the treatment at any quantile, $Q_\theta(Y_1) - Q_\theta(Y_0)$ is identified.

- Randomization identifies the entire marginal distributions of Y_0 and Y_1
- Does not identify the quantiles of the effect: $Q_\theta(Y_1 - Y_0)$ (the difference of quantiles is not the quantile of the difference)

Estimation in randomized experiments

Consider a randomized trial with N individuals. Suppose that the estimand of interest is ATE:

$$\alpha_{ATE} = E[Y_1 - Y_0] = E[Y|D = 1] - E[Y|D = 0].$$

Using the **analogy principle**, we construct an estimator:

$$\widehat{\alpha} = \bar{Y}_1 - \bar{Y}_0,$$

where

$$\bar{Y}_1 = \frac{\sum Y_i \cdot D_i}{\sum D_i} = \frac{1}{N_1} \sum_{D_i=1} Y_i\,;$$

$$\bar{Y}_0 = \frac{\sum Y_i \cdot (1 - D_i)}{\sum (1 - D_i)} = \frac{1}{N_0} \sum_{D_i=0} Y_i$$

with $N_1 = \sum_i D_i$ and $N_0 = N - N_1$.
$\widehat{\alpha}$ is an unbiased and consistent estimator of α_{ATE}.

Testing in large samples: Two-sample t-test

Notice that:

$$\frac{\widehat{\alpha} - \alpha_{ATE}}{\sqrt{\dfrac{\widehat{\sigma}_1^2}{N_1} + \dfrac{\widehat{\sigma}_0^2}{N_0}}} \xrightarrow{d} N(0,1),$$

where

$$\widehat{\sigma}_1^2 = \frac{1}{N_1 - 1} \sum_{D_i=1} (Y_i - \bar{Y}_1)^2,$$

and $\widehat{\sigma}_0^2$ is analogously defined. In particular, let

$$t = \frac{\widehat{\alpha}}{\sqrt{\dfrac{\widehat{\sigma}_1^2}{N_1} + \dfrac{\widehat{\sigma}_0^2}{N_0}}}.$$

We reject the null hypothesis H$_0$: $\alpha_{ATE} = 0$ against the alternative H$_1$: $\alpha_{ATE} \neq 0$ at the 5% significance level if $|t| > 1.96$.

Testing in small samples: Fisher's exact test

- Test of differences in means with large N:

$$H_0 : E[Y_1] = E[Y_0], \quad H_1 : E[Y_1] \neq E[Y_0]$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null)}$$

- Let Ω be the set of all possible randomization realizations.
- We only observe the outcomes, Y_i, for one realization of the experiment. We calculate $\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$.
- Under the sharp null hypothesis we can calculate the value that the difference of means would have taken under any other realization, $\hat{\alpha}(\omega)$, for $\omega \in \Omega$.

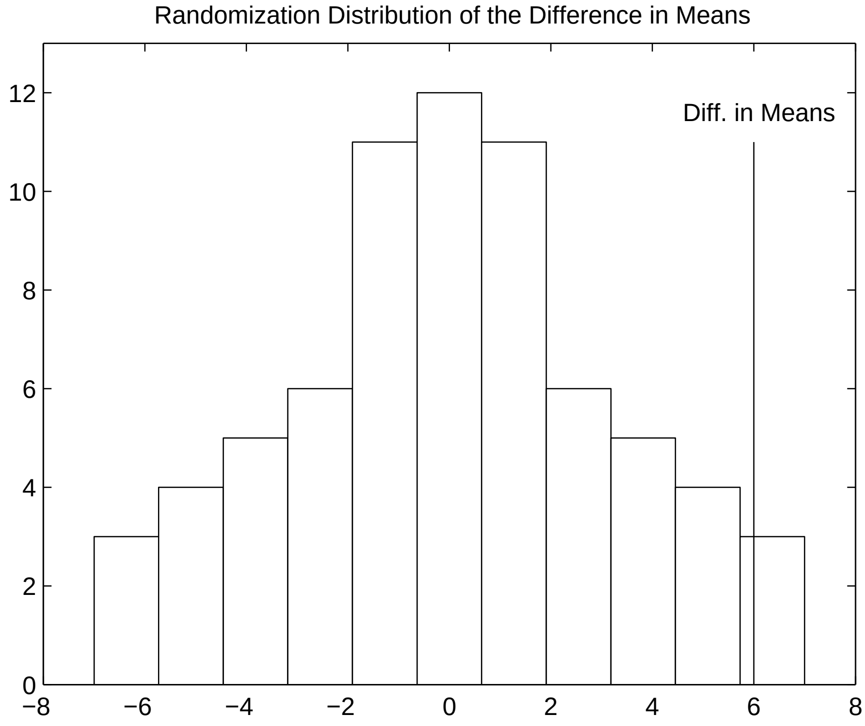
Testing in small samples: Fisher's exact test

Suppose that we assign 4 individuals out of 8 to the treatment:

Y_i	12	4	6	10	6	0	1	1	
D_i	1	1	1	1	0	0	0	0	$\hat{\alpha} = 6$
									$\hat{\alpha}(\omega)$
$\omega = 1$	1	1	1	1	0	0	0	0	6
$\omega = 2$	1	1	1	0	1	0	0	0	4
$\omega = 3$	1	1	1	0	0	1	0	0	1
$\omega = 4$	1	1	1	0	0	0	1	0	1.5
				. . .					
$\omega = 70$	0	0	0	0	1	1	1	1	-6

- The randomization distribution of $\widehat{\alpha}$ (under the sharp null hypothesis) is $\Pr(\widehat{\alpha} \leq z) = \frac{1}{70} \sum_{\omega \in \Omega} 1\{\widehat{\alpha}(\omega) \leq z\}$
- Now, find $\bar{z} = \inf\{z : P(|\widehat{\alpha}| > z) \leq 0.05\}$
- Reject the null hypothesis, H_0: $Y_{1i} - Y_{0i} = 0$ for all i, against the alternative hypothesis, H_1: $Y_{1i} - Y_{0i} \neq 0$ for some i, at the 5% significance level if $|\widehat{\alpha}| > \bar{z}$

Testing in small samples: Fisher's exact test



$\Pr(|\hat{\alpha}(\omega)| \geq 6) = 0.0857$

Covariate balance

- Randomization balances observed but also unobserved characteristics between treatment and control group
- Can check random assignment using so called “balance tests” (e.g., t-tests) to see if distributions of the observed covariates, X, are the same in the treatment and control groups
- X are pre-treatment variables that are measured prior to treatment assignment (i.e., at “baseline”)

Threats to the validity of randomized experiments

- Internal validity: can we estimate treatment effect for our particular sample?
  - Fails when there are differences between treated and controls (other than the treatment itself) that affect the outcome and that we cannot control for
- External validity: can we extrapolate our estimates to other populations?
  - Fails when the treatment effect is different outside the evaluation environment

Most common threats to internal validity

- Failure of randomization
- Non-compliance with experimental protocol
- Attrition

Most common threats to external validity

- Non-representative sample
- Non-representative program
    - The treatment differs in actual implementations
    - Scale effects
    - Actual implementations are not randomized (nor full scale)
- Hawthorne effects

Appendix: Experimental Design

Experimental design: Relative sample sizes for fixed N

Suppose that you have N experimental subjects and you have to decide how many will be in the treatment group and how many in the control group. We know that:

$$\bar{Y}_1 - \bar{Y}_0 \sim \left(\mu_1 - \mu_0, \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}\right).$$

We want to choose N_1 and N_0, subject to $N_1 + N_0 = N$, to minimize the variance of the estimator of the average treatment effect.

The variance of $\bar{Y}_1 - \bar{Y}_0$ is:

$$\text{var}(\bar{Y}_1 - \bar{Y}_0) = \frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}$$

where $p = N_1/N$ is the proportion of treated in the sample.

Experimental design: Relative sample sizes for fixed N

Find the value p^* that minimizes $\text{var}(\bar{Y}_1 - \bar{Y}_0)$:

$$-\frac{\sigma_1^2}{p^{*2}N} + \frac{\sigma_0^2}{(1-p^*)^2 N} = 0.$$

Therefore:

$$\frac{1-p^*}{p^*} = \frac{\sigma_0}{\sigma_1},$$

and

$$p^* = \frac{\sigma_1}{\sigma_1 + \sigma_0} = \frac{1}{1 + \sigma_0/\sigma_1}.$$

A "rule of thumb" for the case $\sigma_1 \approx \sigma_0$ is $p* = 0.5$

For practical reasons it is sometimes better to choose unequal sample sizes (even if $\sigma_1 \approx \sigma_0$)

Experimental design: Power calculations to choose N

- Recall that for a statistical test:
  - Type I error: Rejecting the null if the null is true.
  - Type II error: Not rejecting the null if the null is false.
- Size of a test is the probability of type I error, usually 0.05.
- Power of a test is one minus the probability of type II error, i.e. the probability of rejecting the null if the null is false.
- Statistical power increases with the sample size.
- But when is a sample "large enough"?
- We want to find N such that we will be able to detect an average treatment effect of size α or larger with high probability.

Experimental design: Power calculations to choose N

Assume a particular value, α, for $\mu_1 - \mu_0$.
Let $\widehat{\alpha} = \bar{Y}_1 - \bar{Y}_0$ and

$$\text{s.e.}(\widehat{\alpha}) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}.$$

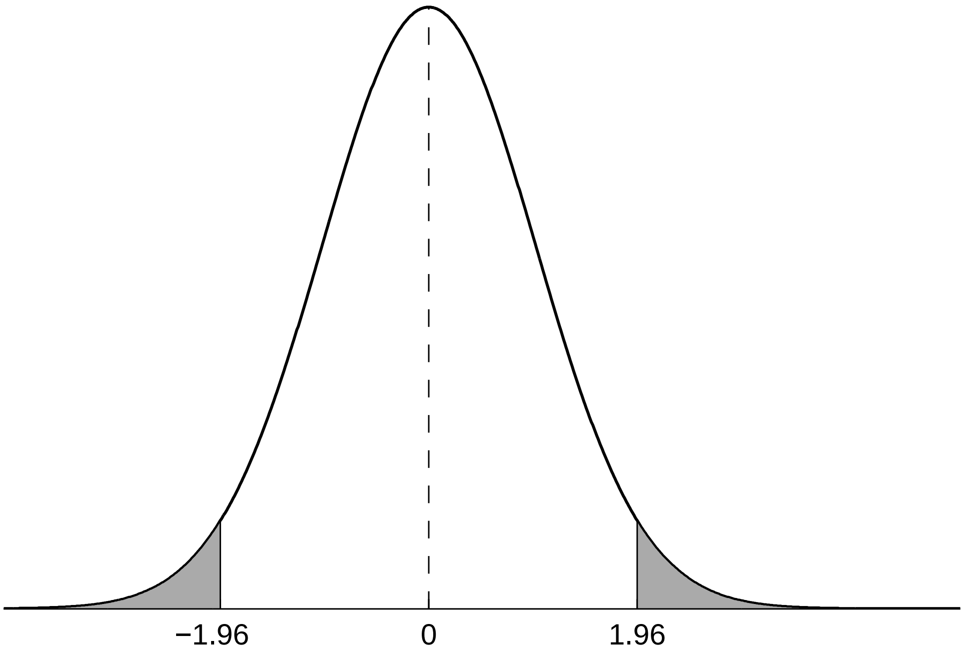
For a large enough sample, we can approximate:

$$\frac{\widehat{\alpha} - \alpha}{\text{s.e.}(\widehat{\alpha})} \sim N(0, 1).$$

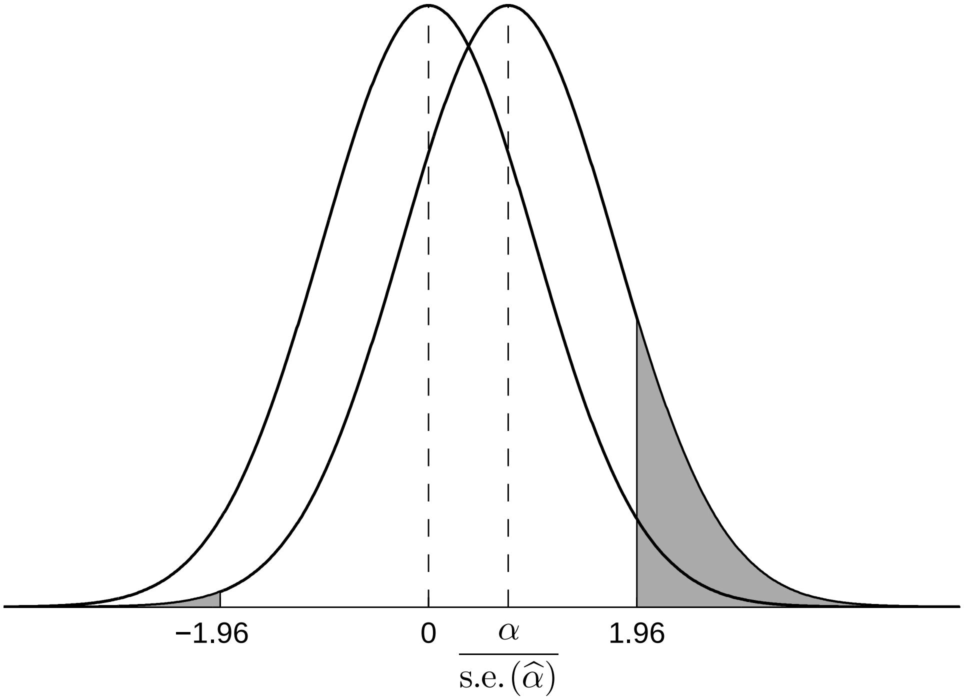
Therefore, the t-statistic for a test of significance is:

$$t = \frac{\widehat{\alpha}}{\text{s.e.}(\widehat{\alpha})} \sim N\left(\frac{\alpha}{\text{s.e.}(\widehat{\alpha})}, 1\right).$$

Probability of rejection if $\mu_1 - \mu_0 = 0$



Probability of rejection if $\mu_1 - \mu_0 = \alpha$



Experimental design: Power calculations to choose N

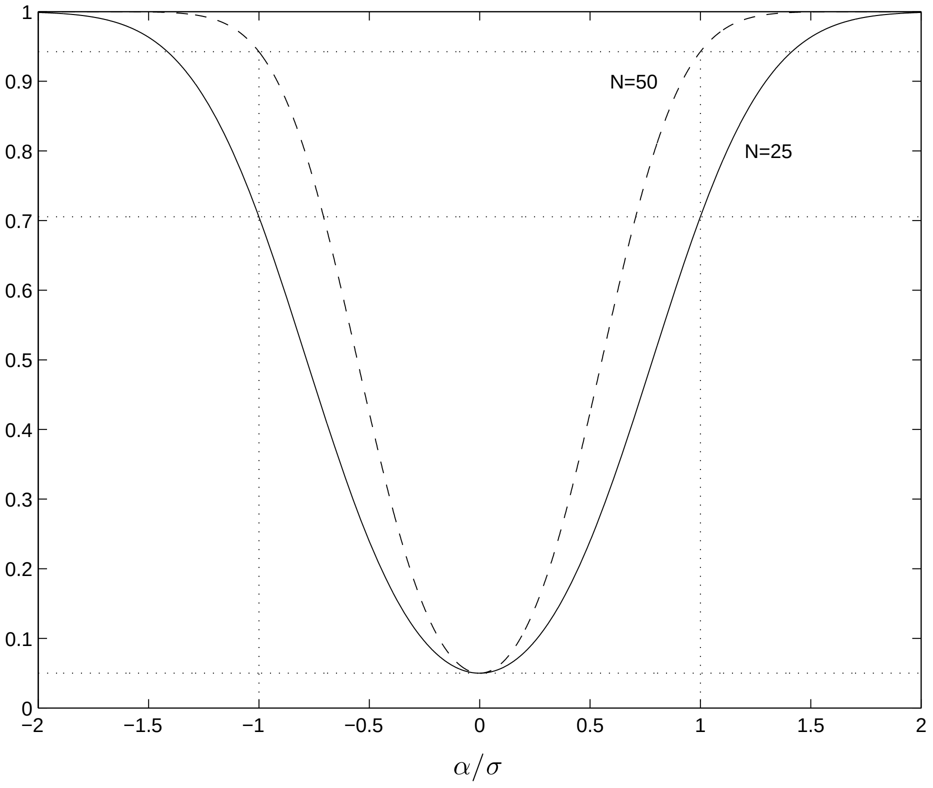
The probability of rejecting the null $\mu_1 - \mu_0 = 0$ is:

$$\begin{aligned}
\Pr(|t| > 1.96) &= \Pr(t < -1.96) + \Pr(t > 1.96) \\
&= \Pr\left(t - \frac{\alpha}{\text{s.e.}(\widehat{\alpha})} < -1.96 - \frac{\alpha}{\text{s.e.}(\widehat{\alpha})}\right) \\
&+ \Pr\left(t - \frac{\alpha}{\text{s.e.}(\widehat{\alpha})} > 1.96 - \frac{\alpha}{\text{s.e.}(\widehat{\alpha})}\right) \\
&= \Phi\left(-1.96 - \frac{\alpha}{\text{s.e.}(\widehat{\alpha})}\right) + \left(1 - \Phi\left(1.96 - \frac{\alpha}{\text{s.e.}(\widehat{\alpha})}\right)\right)
\end{aligned}$$

Suppose that $p = 1/2$ and $\sigma_1^2 = \sigma_0^2 = \sigma^2$. Then,

$$\begin{aligned}
\text{s.e.}(\widehat{\alpha}) &= \sqrt{\frac{\sigma^2}{N/2} + \frac{\sigma^2}{N/2}} \\
&= \frac{2\sigma}{\sqrt{N}}.
\end{aligned}$$

Power functions with $p = 1/2$ and $\sigma_1^2 = \sigma_0^2$



General formula for the power function ($p \neq 1/2$, $\sigma_0^2 \neq \sigma_1^2$)

$$\begin{aligned}
&\Pr\left(\text{reject } \mu_1 - \mu_0 = 0 | \mu_1 - \mu_0 = \alpha\right) \\
&\quad = \Phi\left(-1.96 - \alpha \middle/ \sqrt{\frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}}\right) \\
&\qquad + \left(1 - \Phi\left(1.96 - \alpha \middle/ \sqrt{\frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}}\right)\right).
\end{aligned}$$

To choose N we need to specify:

1. α: minimum detectable magnitude of treatment effect
2. Power value (usually 0.80 or higher)
3. σ_1^2 and σ_0^2 (usually $\sigma_1^2 = \sigma_0^2$) (e.g., using previous measures)
4. p: proportion of observations in the treatment group If $\sigma_1 = \sigma_0$, then the power is maximized by $p = 0.5$