Experimental Workshop: Lecture 2

Covariates, Block Randomization, Cluster Design and Power

Raymond Duch

May 3, 2022

University of Oxford

Lecture 2 Road Map

- Randomization Inference
- Covariates
- Block Randomization
- Cluster Design
- Power

Randomization Inference

Observed Outcomes Local Budget

	Budget share if village head is male	Budget share if village head is female
Village 1	?	15
Village 2	15	?
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	?
Village 7	?	30

Potential Outcomes Local Budget

	Budget share if village head is male	Budget share if village head is female	Treatment Effect
Village 1	10	15	5
Village 2	15	15	0
Village 3	20	30	10
Village 4	20	15	-5
Village 5	10	20	10
Village 6	15	15	0
Village 7	15	30	15
Average	15	30	15

2 ways of thinking about statistical uncertainty

- Sampling-based inference (Neyman):
  - Experimental subjects are a random draw from some "super-population"
  - Realized ATE has a sampling distribution with reference to that superpopulation
  - Different (random) experimental samples $\rightarrow$ different ATE's from draw to draw
  - Uncertainty arises from random sampling of subjects: How are ATE's distributed in the population?
  - Sampling distribution under the H0 is typically $X \sim \mathcal{N}(\mu, \sigma^2)$

Neyman's plan for inference

1. Define the estimand
2. Find unbiased estimator of the estimand
3. Calculate true sampling variance of the estimator
4. Find unbiased estimator of true sampling variance of estimator
5. Assume approximate normality to obtain p-value and confidence interval
6. Where $H0 : E[Y_i(1)] - E[Y_i(0)] = 0$

2 ways of thinking about statistical uncertainty

- Randomization-based inference (Fisher):
  - Treatment assignments are a random draw from the set of all possible assignment combinations $\rightarrow$ finite sample
  - Realized ATE has a distribution over those possible random assignments
  - Different ways of assigning subjects to treatment $\rightarrow$ different ATE's from allocation to allocation
  - Uncertainty arises from random assignment and missing potential outcomes

This has implications for other methods: use the ones that are *directly justified* by randomization $\rightarrow$ design instead of analysis for covariate adjustment; diff-in-group means estimator; reduce reliance on auxiliary modelling assumptions

Table 1: *

The goal of randomization inference is to derive a *sampling distribution* of estimated ATEs. In our application, generated when two of the seven villages listed in Table 2 are assigned to treatment

Estimated ATE	Frequency with which an estimate occurs
-1	2
0	2
0.5	1
1	2
1.5	2
2.5	1
6.5	1
7.5	3
8.5	3
9	1
9.5	1
10	1
16	1
Average 5	
Total	21

Potential Outcomes Local Budget

2 of 21 possible worlds:

World 1:

	Budget share if village head is male	Budget share if village head is female	Treatment Effect
Village 1		15	
Village 2		15	
Village 3	20		
Village 4	20		
Village 5	10		
Village 6	15		
Village 7	15		
Average	16	15	-1

Potential Outcomes Local Budget

World 2:

	Budget share if village head is male	Budget share if village head is female	Treatment Effect
Village 1	10		
Village 2	15		
Village 3	20		
Village 4	20		
Village 5	10		
Village 6		15	
Village 7		30	
Average	15	22.5	7.5

We can calculate the variation of these estimates:

Sum of squared deviations

$$
\begin{aligned}
&= (-1-5)^2 + (-1-5)^2 + (0-5)^2 + (0-5)^2 + (0.5-5)^2 + (1-5)^2 + (1-5)^2 \\
&+ (1.5-5)^2 + (1.5-5)^2 + (2.5-5)^2 + (6.5-5)^2 + (7.5-5)^2 + (7.5-5)^2 \\
&+ (7.5-5)^2 + (8.5-5)^2 + (8.5-5)^2 + (8.5-5)^2 + (9-5)^2 + (9.5-5)^2 \\
&+ (10-5)^2 + (16-5)^2 = 445
\end{aligned}
$$

$$
\textit{Square root of the average squared deviation} = \sqrt{\frac{1}{21}(445)} = 4.60
$$

Neyman variance estimator

Neyman quantifies the variance of our difference-in-means estimator with the Neyman variance estimator. Formally,

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left\{\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1)\right\}}$$

In our application,

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{6}\left\{\frac{(2)(14.29)}{5} + \frac{(5)(42.86)}{2} + (2)(7.14)\right\}} = 4.60$$

You can see that the covariance of the two potential outcomes is fundamentally unobservable, so we assume constant treatment effects, and the sample analog reduces to

$$\widehat{SE} = \sqrt{\frac{\widehat{Var}(Y_i(0))}{N-m} + \frac{\widehat{Var}(Y_1(1))}{m}}$$

Formal Randomization Inference

- Now, randomization inference is different. We only ever observe one particular realization of the randomized treatment assignment
- Yet, given m, N and a binary treatment, there is a set of all possible randomization realizations such that $\Omega = \frac{N!}{m!(N-m)!}$
- For the Abadie and Cattaneo (2018) example, we have $\Omega = \frac{8!}{4!(8-4)!} = 70$, and we are interested in the distribution of $\hat{\tau}(\omega)$, i.e. for each possible realization of the randomized assignment $\omega \in \Omega$, as in the following table

Abadie and Cattaneo 2018

Table 1 Randomization distribution of a difference in means

Panel A: Sample and sample statistic									
Y_i	12	4	6	10	6	0	1	1	
W_i	1	1	1	1	0	0	0	0	$\widehat{\tau} = 6$
Panel B: Randomization distribution									$\widehat{\tau}(\omega)$
$\omega = 1$	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

Covariates

Pre-treatment and post-treatment

- Q. Why would you want to collect the same information twice, pre-treatment and post-treatment? Do you gain anything?
- A. Yes, precision!
- Instead of having a single outcome measure Y_i, redefine as *change* from pre-test to post-test
- We compare 2 quantities:
  - $(Y_i - X_i)$ for $d_i = 1$
  - $(Y_i - X_i)$ for $d_i = 0$
  - *difference-in-differences* estimator

Pre-treatment and post-treatment

Is this estimator unbiased?

$$
\begin{aligned}
E(\widehat{ATE}) &= E[Y_i - X_i | D_i = 1] - E[Y_i - X_i | D_i = 0] \\
&= E[Y_i | D_i = 1] - E[X_i | D_i = 1] - E[Y_i | D_i = 0] - E[X_i | D_i = \\
&= E[Y_i(1)] - E[Y_i(0)]
\end{aligned}
$$

In general, difference-in-means and difference-in-differences generate unbiased estimates – but what if we also care about sampling variability of this estimator?

Pre-treatment and post-treatment

In general, $SE(\widehat{ATE}') < SE(\widehat{ATE})$ if either holds

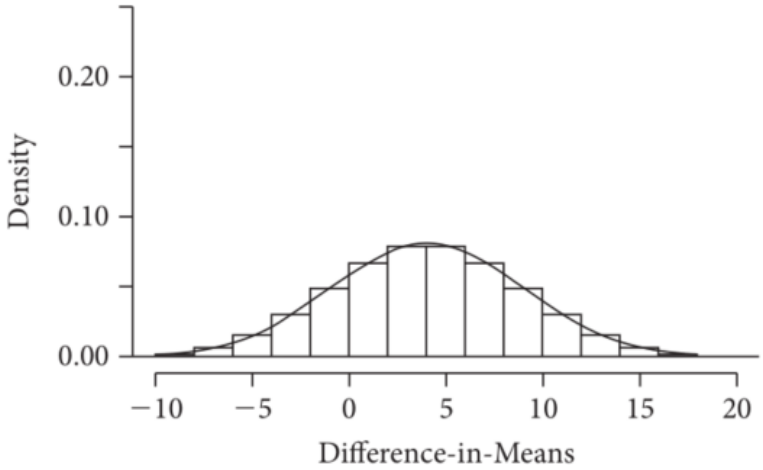
$$Cov(Y_i(0), X_i) + Cov(Y_i(1), X_i) > Var(X_i)$$

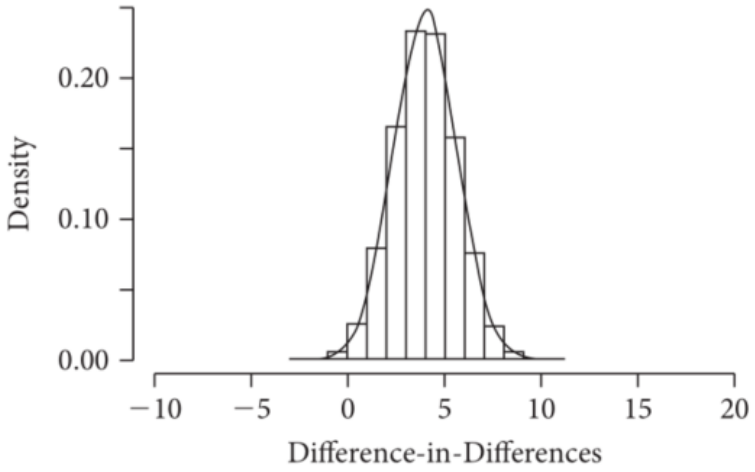
$$\frac{Cov(Y_i(0), X_i)}{Var(X_i)} + \frac{Cov(Y_i(1), X_i)}{Var(X_i)} > 1.$$

That is, *when a covariate X_i strongly predicts potential outcomes*

Pre-treatment and post-treatment

Sampling distribution of two estimators: difference-in-means and difference-in-differences





Block Random Assignment

Why Block Random Assignment: Practical Concerns

- Program requirements may restrict number of subjects allowed to receive treatment
- E.g. summer reading program concerned about students with low levels of preparedness: 60% of the admitted students must pass basic skills test
- If 50 students are admitted, randomly select 20 from the applicants that failed and 30 from those who passed
- Fairness concerns require each treatment of demographic groups
- Resource constraints mean you are only able to sample a certain number of subjects from certain groups

Why Block Random Assignment: Statistical Concerns

- Reduces sampling variability
- Subjects in blocks likely to have similar potential outcomes (those who fail and those who pass)
- Especially effective in small samples
- Ensures the ability to do subgroup analysis, e.g. women and men
- Complete random assignment may lead to imbalance

Potential Outcomes

Village	Block	$Y_i(0)$	$Y_i(1)$
1	A	0	0
2	A	1	0
3	A	2	1
4	A	4	2
5	A	4	0
6	A	6	0
7	A	6	2
8	A	9	3
9	B	14	12
10	B	15	9
11	B	16	8
12	B	16	15
13	B	17	5
14	B	18	17
⋮	⋮	⋮	⋮

Schedule of potential outcomes for public works projects when audited ($Y(1)$) and not audited ($Y(0)$)

Village	Block	All subjects $Y(0)$	All subjects $Y(1)$	Block A subjects $Y(0)$	Block A subjects $Y(1)$	Block B subjects $Y(0)$	Block B subjects $Y(1)$
1	A	0	0	0	0		
2	A	1	0	1	0		
3	A	2	1	2	1		
4	A	4	2	4	2		
5	A	4	0	4	0		
6	A	6	0	6	0		
7	A	6	2	6	2		
8	A	9	3	9	3		
9	B	14	12			14	12
10	B	15	9			15	9
11	B	16	8			16	8
12	B	16	15			16	15
13	B	17	5			17	5
14	B	18	17			18	17
	Mean	9.14	5.29	4.00	1.00	16.0	11.0
	Variance	40.41	32.49	7.75	1.25	1.67	17.0
	$Cov(Y(0), Y(1))$	31.03		2.13		1.00	

	All subjects		Block A		Block B	
	Y_i^c	Y_i^t	Y_i^c	Y_i^t	Y_i^c	Y_i^t
Mean	9.14	5.29	4.00	1.00	16.00	11.00
Variance	40.41	32.49	7.75	1.25	1.67	17.00
Covariance	31.03		2.13		1.00	

Estimating *ATE* with Block Random Assignment

$$ATE = \sum_{j=1}^{J} \frac{N_j}{N} ATE_j$$

- Where J is the number of blocks and $\frac{N_j}{N}$ is the share of all subjects in block j
- Weighted average of the block-specific ATEs

Observed Outcomes

Village	Block	$Y_i(0)$	$Y_i(1)$
1	A	0	?
2	A	1	?
3	A	?	1
4	A	4	?
5	A	4	?
6	A	6	?
7	A	6	?
8	A	?	3
9	B	14	?
10	B	?	9
11	B	16	?
12	B	16	?
13	B	17	?
14	B	?	17
⋮	⋮	⋮	⋮

Estimating ATE with Block Random Assignment

$$\begin{aligned}\widehat{ATE} &= (\widehat{ATE}_1)\left(\frac{N_1}{N}\right) + (\widehat{ATE}_2)\left(\frac{N_2}{N}\right) \\ &= (-1.5)\left(\frac{8}{14}\right) + (-2.75)\left(\frac{6}{14}\right) \\ &= -2.04\end{aligned}$$

Standard Error of the Estimated *ATE*

$$\widehat{SE}(\widehat{ATE}) = \sqrt{\widehat{SE}_1^2 \left(\frac{N_1}{N}\right)^2 + \widehat{SE}_2^2 \left(\frac{N_2}{N}\right)^2}$$

where for each of the two blocks:

$$\widehat{SE} = \sqrt{\frac{\widehat{Var}(Y_i^c)}{N - m} + \frac{\widehat{Var}(Y_i^t)}{m}}$$

SE with Random Block Design

$SE(\widehat{ATE})$ with complete random assignment

$$= \sqrt{\frac{1}{k-1}\left\{\frac{mVar(\bar{Y}_i^c)}{N-m} + \frac{(N-m)Var(\bar{Y}_j^t)}{m} + 2Cov(Y_j^c, Y_j^t)\right\}}$$

$$= \sqrt{\frac{1}{13}\left\{\frac{4(40.41)}{10} + \frac{(10)(32.49)}{4} + 2(31.03)\right\}}$$

$$= 3.50$$

$SE(\widehat{ATE})$ with block random assignment

$$= \sqrt{SE_1^2\left(\frac{N_1}{N}\right)^2 + SE_2^2\left(\frac{N_2}{N}\right)^2}$$

$$= \sqrt{(1.23)^2\left(\frac{8}{14}\right)^2 + (2.71)^2\left(\frac{6}{14}\right)^2}$$

$$= 1.36$$

Regression Estimation in Block Design

$$Y_i = \beta_1 \alpha_{\text{BR}} D_i + \beta_2 J_2 + \beta_3 J_3 + \cdots + \beta_j J_j + \mu_i$$

where $J_2, J_3, \ldots, J_j$ are dummy variables indicating each block.

- This regression estimator is valid if the treatment probability $p_j = \frac{m_j}{N_j}$ is the same in all blocks.
  - Regression weights each block specific $\hat{\text{ATE}}$ by $(\frac{N_j}{N}) p_j (1 - p_j)$
- If p_j varies across blocks, regression can lead to bias since treatment assignment is correlated with block characteristics.
- Need to use weighted regression with unit weights:

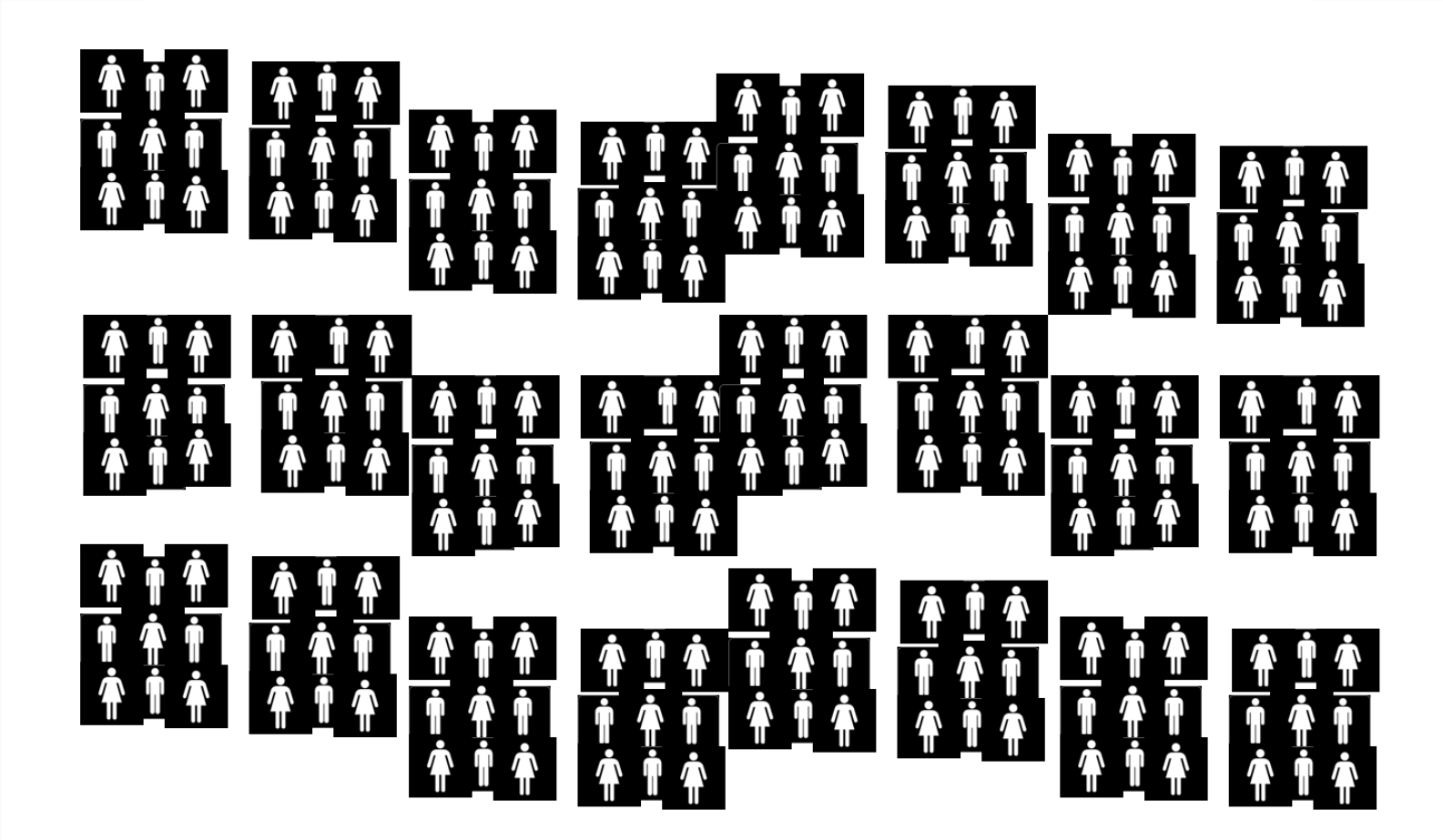
$$w_{ij} = (\frac{1}{p_{ij}}) D_i + (\frac{1}{1 - p_{ij}})(1 - D_i) \qquad (1)$$

Cluster Design

Unit of Random Assignment?

- Options:
  - Individual
  - Clusters or groups
- What is the level of random assignment?
- Considerations
  - What is the level of treatment?
  - What is the unit of analysis?

Random assignment at the level of the individual?

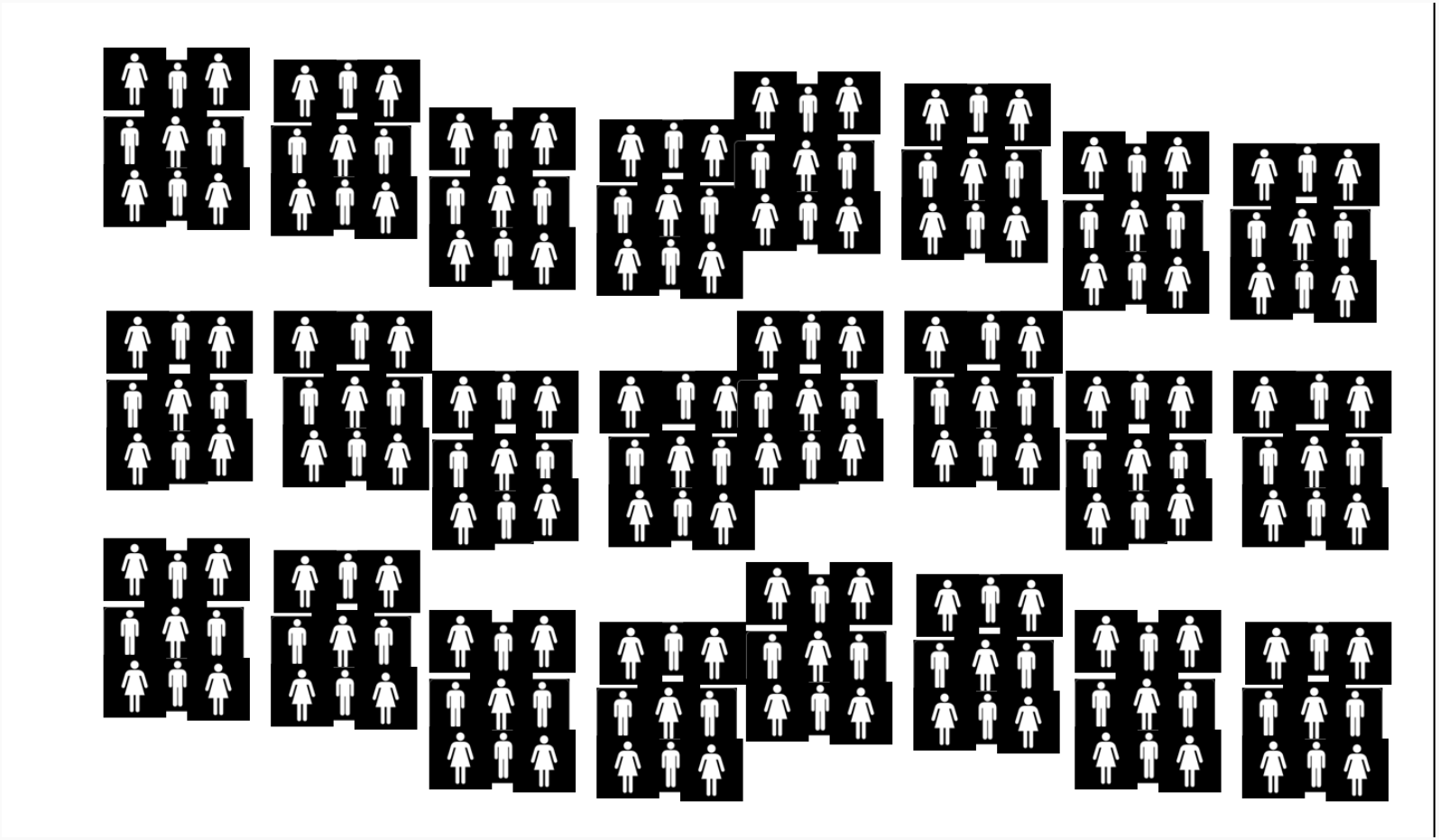


Random assignment at the level of the individual?

Cluster random assignment

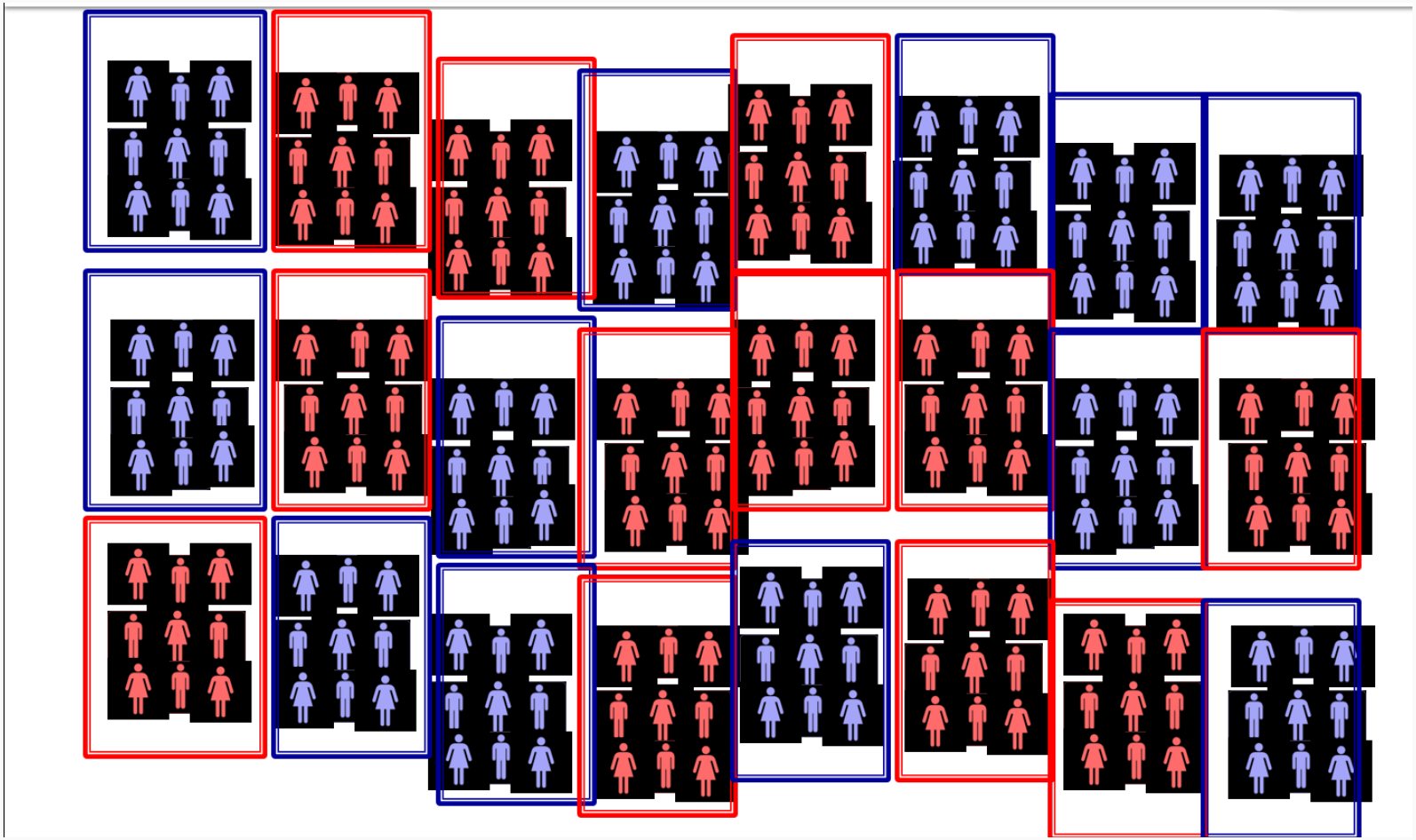
- Cluster randomized experiments allocate treatments to groups
- But measure outcomes at the level of the individuals that compose the groups
- restricts the number of ways that the treatment and control groups can be composed, relative to randomization at the individual level
- leads to underestimating the variance in our estimator

Random assignment at the level of the cluster or group?



Random assignment at the level of the cluster or group?

Random assignment at the level of the cluster or group?



Education: Level of the class.

Random assignment at the level of the cluster or group?

Education: Level of the school.

Why Clustering Matters: Example

- same sample size and same participants could contain very different amounts of information depending on whether units are clustered
- 10 people: 5 assigned to treatment and 5 to control
- Version 1: treatment is assigned to individuals
- Version 2: 5 individuals with black hair and the 5 individuals with some other color of hair are assigned to treatment as a group
- 252 combinations versus 2 combinations

```
treatment_effect      <- 1
# Define the individual ids (i)
person                <- 1:10
# Define the cluster indicator (j)
hair_color            <- c(rep("black",5),rep("brown
    ",5))
# Define the control outcome (Y0)
outcome_if_untreated <- rnorm(n = 10)
# Define the treatment outcome (Y1)
outcome_if_treated    <- outcome_if_untreated +
    treatment_effect

# Version 1 - Not cluster randomized
# Generate all possible non-clustered assignments
    of treatment (Z)
non_clustered_assignments <- combn(x = unique(
    person),m = 5)
```

```r
# Estimate the treatment effect
treatment_effects_V1 <-
      apply(
            X = non_clustered_assignments,
            MARGIN = 2,
            FUN = function(assignment) {
                  treated_outcomes   <- outcome_if_
                      treated[person %in% assignment]
                  untreated_outcomes <- outcome_if_
                      untreated[!person %in%
                      assignment]
                  mean(treated_outcomes) - mean(
                      untreated_outcomes)
            })
# Estimate the true standard error
standard_error_V1 <- sd(treatment_effects_V1)
# Plot the histogram of all possible estimates of
    the treatment effect
hist(treatment_effects_V1, xlim = c(-1,2.5), breaks =
    20)
```

```
### Cluster

# Version 2 - Cluster randomized
# Generate all possible assignments of treatment
    when clustering by hair color (Z)
clustered_assignments      <- combn(x = unique(hair_
    color),m = 1)
# Estimate the treatment effect
treatment_effects_V2 <-
  sapply(
    X = clustered_assignments,
    FUN = function(assignment) {
      treated_outcomes    <- outcome_if_treated[
          person %in% person[hair_color==assignment
          ]]
      untreated_outcomes <- outcome_if_untreated[
          person %in% person[!hair_color==
          assignment]]
```

```
# Estimate the true standard error
standard_error_V2 <- sd(treatment_effects_V2)
# Plot the histogram of all possible estimates of
    the treatment effect
hist(treatment_effects_V2, xlim = c(-1,2.5), breaks =
     20)
```

Individual variation within and between clusters

- Two cluster randomized studies with $J = 10$ villages and $n_j = 100$ people per village may have different information about the treatment effect on individuals
- Version 1: differences between villages are much greater than the differences in outcomes within them
    - all individuals in any village acted exactly the same
    - different villages showed different outcomes
    - we have 10 pieces of information:
    - all info about causal effects would be at the village level
- Version 2: if individuals within a village acted independently of each other, then we have $10 * 100 = 1000$ pieces of information.

Intracluster Correlation Coefficient

- indicates extent to which highly dependent clusters provide less information than the highly independent clusters

$$\text{ICC} = \frac{\text{Variance between clusters in y}}{\text{Total variance in y}} = +\frac{\sigma_j^2}{\sigma_j^2 + \sigma_i^2} \tag{2}$$

where:

- y is the outcome variable
- j clusters
- i units
- σ_j^2 is variation in outcomes defined at the cluster level
- σ_i^2 is variation between units within the population

Robust Clustered Standard Errors

$$\text{Var}\beta_{ols} = \frac{\sigma^2}{\sum_c \sum_i (T_{ic} - T)^2} \tag{3}$$

$$\text{Var}\beta_{Clustered} = \frac{\sigma^2(1 + (\overline{n_c} - 1)\rho))}{\sum_c \sum_i (T_{ic} - T)^2} \tag{4}$$

where:

- T is a treatment variables
- ρ is the ICC
- c are clusters
- n is number of units in cluster

Power Analysis

Statistical Power

- What is the power of a statistical test? H_0: null hypothesis
- Apply estimator to test some alternative H_A
- Type I error: False positive
  - If the null is true, how likely does the estimated effect (or greater) occur by chance?
  - Our tolerance for these errors is set by α
  - When $\alpha = 0.05$, 95% of the CIs we construct from repeated sampling will contain the true parameter

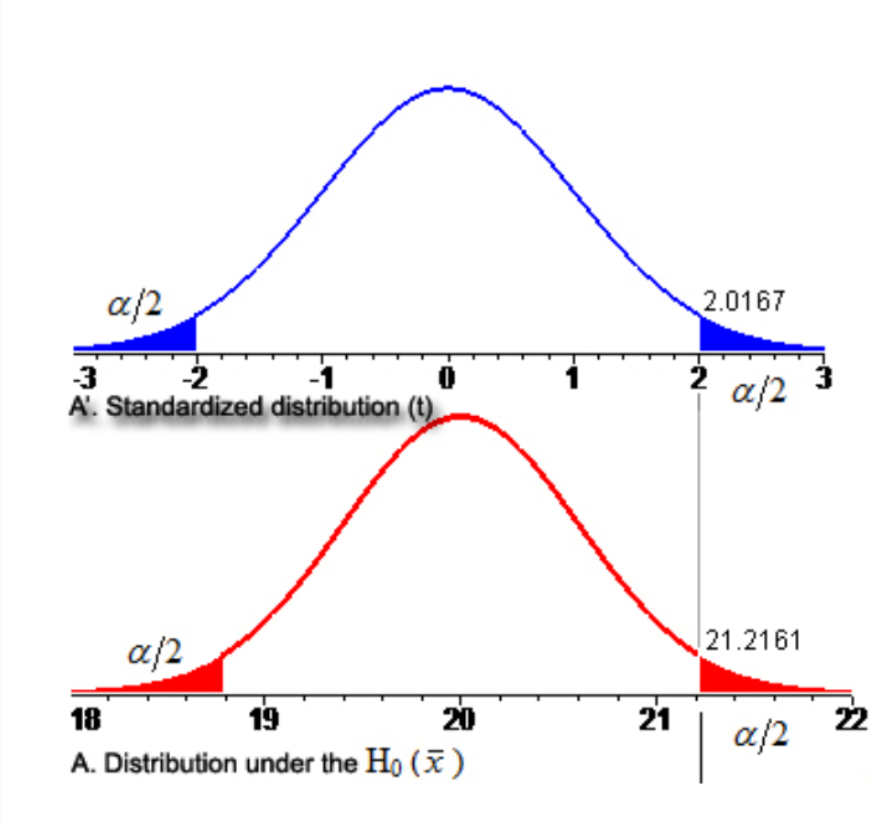
Statistical Power

- Type II error: False negative
  - If the null is not true, how often can we reject the null successfully?
  - Probability or rate of Type II error, β
- Power of a test: probability that the test rejects H_0, $1 - \beta$

Basic Inference Revisited

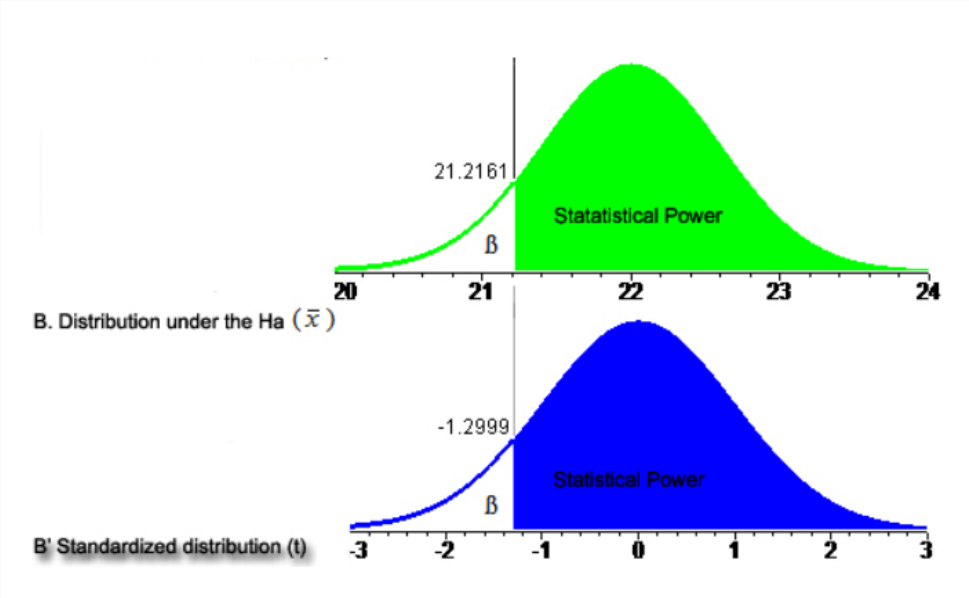
- What is the effect of losing Medicaid on infant mortality?
- $H_0 = 20$ deaths per 1,000 live births (assumed known without uncertainty here)
- True effect is an increase of 2 deaths per 1,000 live births
- Standard deviation in population is 4, we have N=44 observations; sampling distribution yields a standard error of 0.60
- $\hat{x}$ is our estimate of the new infant mortality rate
- Let's say we get an estimate right at the true estimate, $\hat{x} = 22$
- How unlikely is it we get this estimate, if the null is actually true?

Sampling Distribution Under Null



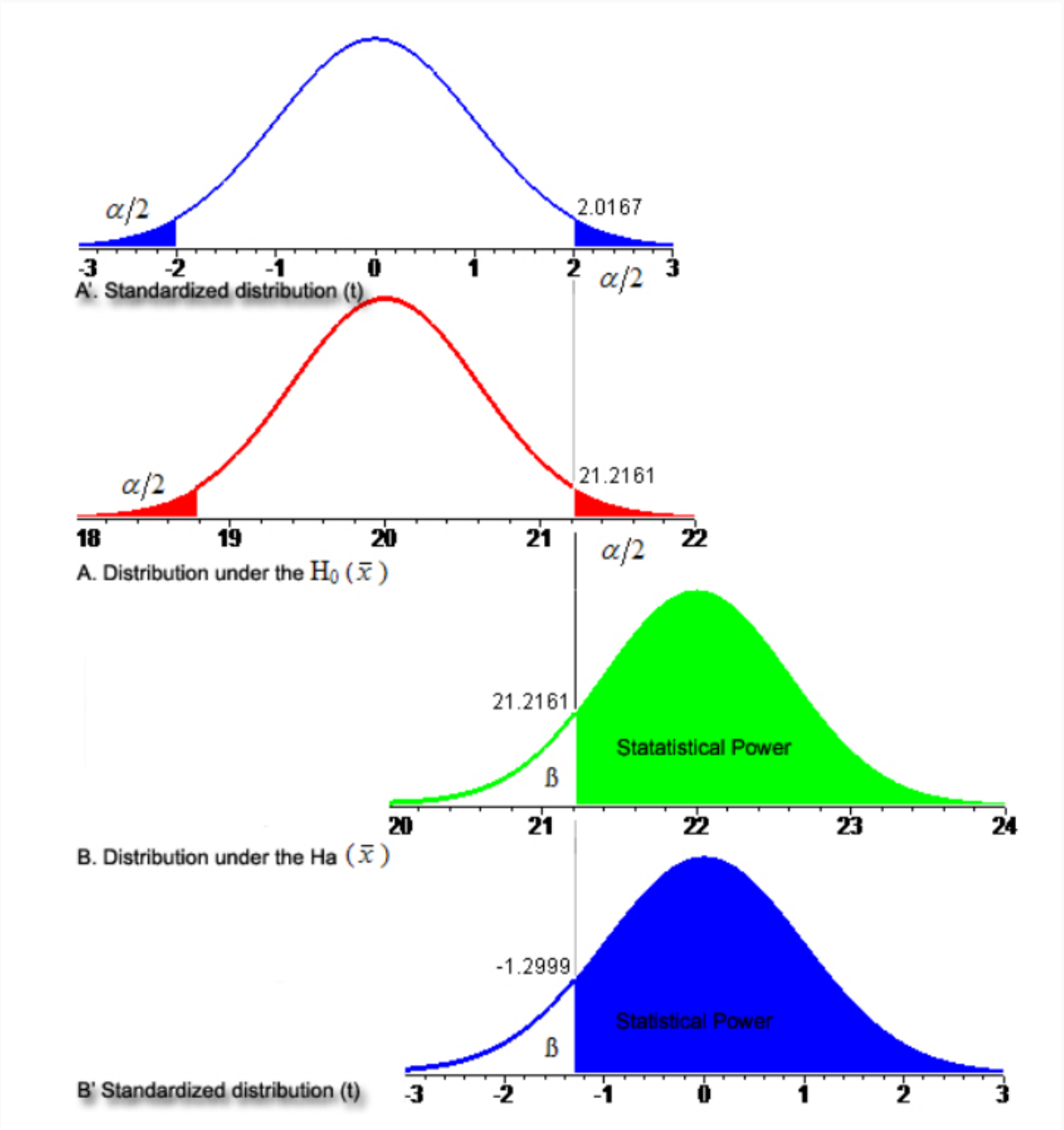
- Say for our test $\alpha = 0.05$
- Can rescale via Z-transformation
- What does this graphic mean?
- For $\hat{x} = 22$,
- t-stat=3.32, $p < 0.01$

Sampling Distribution of $\hat{x}$



- Interpret this graphic
- $1 - \beta$ is fraction of estimates that reject null hypothesis
- Power of the test
- What $x_t rue$ yields $1 - \beta = 0.5$?
- What parameters are needed?

The Relationship Between α and β



Sample Size Increases Power

- Of primary interest because it can be manipulated
- Law of large numbers: for independent data, statistical precision of estimates increases with the square root of the sample size, $\sqrt{n}$
- Test statistics often have the form $T = \hat{\theta}/\sqrt{\hat{V}(\hat{\theta})}$
- Example: Mean of normal distribution θ, data $y = (y_1, ..., y_n)$, iid

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} y_i = \bar{y}$$

$$\hat{V}(\hat{\theta}) = V(y)/n \text{ and } \sqrt{\hat{V}(\hat{\theta})} = s_y/\sqrt{n}$$

$$T = \bar{y}/(s_y/\sqrt{n})$$

- This logic extends to two-sample case (e.g., treated vs control in an experiment), regression, logistic regression, etc.

Reverse Engineer T to Determine Sample Size

- How much sample do I need to give myself a "reasonable" chance of rejecting H_0, given expectations as to the magnitude of the "effect"
- Example:

  A proportion $\theta \in [0, 1]$ estimated as $\hat{\theta}$

  Variance is $\theta(1-\theta)/n,$ maxes at 0.5

  A 95% CI at $\theta = 0.5$ is $0.5 \pm 2\sqrt{0.25/n}$

  Width of that interval is $W = 4\sqrt{0.25/n} \rightarrow n = 4/W^2$

- Typical use: how big must a poll be to get reasonable MOE?
- For researchers, how big must a poll be to detect a campaign effect?
  - Answer depends on beliefs about likely magnitude of campaign effects

Calculating Power (β)

$$\beta = \Phi(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\alpha} - \Phi^{-1}(1 - \frac{\alpha}{2}))$$

where:

- β= Power [0,1]
- Φ = CDF of normal and Φ^{-1} is its inverse
- μ_t is average outcome treatment – assume 65
- μ_c is average outcome treatment – assume 60
- treatment effect $\mu_t - \mu_c = 5$
- need an assumption for standard deviation of the outcome, σ – say $\sigma = 20$
- assume $\alpha = 0.05$ and N=500

Cohen's D

Cohen's D Definition

$$\text{Cohen's d} = \frac{(M_1 - M_2)}{\text{Pooled SD}}$$

where

$$\text{Pooled SD} = \sqrt{\frac{(sd_1^2 + sd_2^2)}{2}}$$

assume that group 1 as rnorm(n, 1,2)

assuming equal variance (t distribution assumption)

$$\text{Pooled SD} = \sqrt{\frac{(2^2 + 2^2)}{2}} = 2$$

Cohen's D: Estimating

$$\text{Cohen's d} = \frac{(1-0)}{2} = .5$$

Solving for the Pooled Standard Deviation

$$0.5 = \frac{(1-0)}{(\text{pooled SD})}$$

R code: solve(0.5,1) cohens d of .5

R result: [1] 2

Simulations in R

- Simulated t-test and power
- Simulating necessary N for t-test power
- Simulating necessary N for bi-variate regression

Assignment 2

- Generate similar power simulations for a multi-variate regression
- Illustrate the effect that covariates can have on Power