CENTRE FOR
EXPERIMENTAL
SOCIAL
SCIENCES

Lecture 1: Causal Inference and Potential Outcomes

Raymond Duch

April 24, 2018

Director CESS Nuffield/Santiago

Road Map to Lecture 1

- Potential outcomes and causal inference
- Average Treatment Effects (ATE)
- Complier Average Causal Effect (CACE)
- Intention to Treat Effect (ITT)
- Power Calculations

Defining Treatment

- The variable d_i indicates whether the ith subject is treated
- In the typical case of binary treatments, $d_i = 1$ means the ith subject receives the treatment
- $d_i = 0$ means the ith subject does not receive the treatment
- It is assumed that d_i is observed for every subject

Potential Outcomes

- Y_i: the potential outcome for subject i
- $Y_i(d_i)$: the outcome for subject i, written as a function of the treatment i received; it is generally the case that we observe only one of the potential outcomes for each i
- For the binary-valued treatment, there are two "potential outcomes":
  - $Y_i(1)$, the potential outcome for i conditional on i being treated
  - $Y_i(0)$, the potential outcome for i conditional on i not being treated

Potential Outcome Schedule

- "Hypothetical"
- Comprehensive list of potential outcomes for all subjects
- Rows of this schedule are indexed by i, and the columns are indexed by d
- Potential outcomes for the fifth subject may be found in adjacent columns of the fifth row

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	20	5

Potential Outcome Subgroup

- Sometimes useful to refer to potential outcomes for a subset of the subjects
- Expressions of the form $Y_i(d)|X = x$ denote potential outcomes when the condition $X = x$ holds
- For example, $Y_i(0)|d_i = 1$ refers to the untreated potential outcome for a subject who actually receives the treatment

Individual Level Causal Effect

- For subject i, the effect of the treatment is conventionally defined as the difference between outcomes across the two potential outcomes:

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

- Alternatively:

$$Y_i = Y_i(0) + (Y_i(1) - Y_i(0))D_i$$

- Often referred to as the Rubin causal model; perhaps more appropriately, the Neyman-Holland-Rubin causal model
- **The Fundamental Problem of Causal Inference** only one of the two potential outcomes is realized, so that δ_i is typically non-operational

Relaized Potential Outcomes

- Use lower-case letters for realization of the potential quantities (again, typically only one of the two potential outcomes is realized)
    1. $y_i(1)$, the outcome observed for i conditional on $d_i = 1$ (i is treated)
    2. $y_i(0)$, the outcome observed for i conditional on $d_i = 0$ (i is not treated)

The Fundamental Problem of Causal Inference

Table 1: Table 2.1, p35 Morgan and Winship, *Counterfactuals and Causal Inference*

Group	$Y_i(1)$	$Y_i(0)$
Treatment ($D_i = 1$)	**Observable**	Counterfactual
Treatment ($D_i = 0$)	Counterfactual	**Observable**

Observed Outcomes

- The connection between the observed outcome and the underlying potential outcome is given by the equation $Y_i = d_i Y_i(1) + (1 - d_i) Y_i(0))$
- This equation indicates that the $Y_i(1)$ are observed for subjects who are treated, and the $Y_i(0)$ are observed for subjects who are not treated
- For any given subject, we observe either $Y_i(1)$ or $Y_i(0)$, not both

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

Average Treatment Effect

- Average Treatment Effect:

$$E(\delta) = E[Y(1)] - E[Y(0)] = E[Y(1) - Y(0)]$$

- where the expectation is over a population, and so no subscript i
- This is operational, in that we can compute sample estimates of E[Y(1)] and E[Y(0)]: e.g., the sample averages:

$$\hat{y}(1) = \frac{1}{n_1} \sum_{i:d_i=1} y_i(1) \text{ and } \frac{1}{n_0} \sum_{i:d_i=0} y_i(0)$$

- where n_1 and n_0 are the number of subjects in groups $d(1)$ and $d(0)$ respectively

Randomization Generates Unbiased Estimates of Average Treatment Effect

- Rubin (1974) calls this:

$$\begin{aligned}\hat{\delta} &= \hat{y}(1) - \hat{y}(0)\\ &= \hat{y}_d\end{aligned}$$

- Under certain circumstances, this is an unbiased estimate of the population average treatment effect δ
- Why? How?
- Nice, informal treatment in "Two Formal Benefits of Randomization"

Properties of Random Assignment?

- Under equal probability random assignment, the conditional ATE among the treated is the same as the conditional ATE among the control group, which is therefore the same as the ATE
- The expected $Y_i(0)$ in the treatment group is the same as the expected $Y_i(0)$ in the control group
- When random assignment is not used (i.e., observational research), the unbiasedness of the difference-in-means estimator becomes a matter of conjecture

Potential Outcomes: Core Assumptions

- We assume that each subject has two potential outcomes $Y_i(1)$ if treated and $Y_i(0)$ if not treated
- Each potential outcome depends **solely** on whether the subject **itself** receives the treatment
- Potential outcomes respond only to the treatment and not to some other feature of the experiment - such as assignment or measurement

Exclusion restriction

- Let $Y_i(z, d)$ be the potential outcome when $z_i = z$ and $d_i = d$ for $z \in (0, 1)$ and for $d \in (0, 1)$
- For example, if $z_i = 1$ and $d_i = 1$, the subject is assigned to the treatment group and receives the treatment
- Or $z_i = 1$ and $d_i = 0$ - subject is assigned treatment but does not receive treatment
- The exclusion restriction is that $Y_i(1, d) = Y_i(0, d)$ - subjects only respond to input from d_i
- The excludability assumption cannot be verified empirically because we never observe both and for the same subject

Classic Drug Experiment Example

- Treatment group receives a new drug
- Control group receives nothing
- Experiment confounds this treatment with receipt of a pill
- If patients respond to the pill rather than the pill's ingredients, excludability is violated
- Jeopardizes unbiasedness of the difference-in-means estimator

Non-interference

- Permits us to ignore the potential outcomes that would arise if subject i were affected by the treatment of other subjects
- Formally, we reduce the schedule of potential outcomes $Y_i(\mathbf{d})$, where $\mathbf{d}$ describes all of the treatments administered to all subjects, to a much simpler schedule $Y_i(d)$, where d refers to the treatment administered to subject i.
- Implies that so long as a subject's treatment status remains constant, that subject's outcome is unaffected by the particular way in which treatments happened to be assigned to other subjects

Non-interference violated

- Police patrols displace crime from treated to untreated areas
- Non-interference violated if your estimand is following:
    - Average potential outcome when a block is treated minus average potential outcome when no blocks are treated
- If police patrols displace crime from treated to untreated areas, observed outcomes in control will not be potential outcomes when no treatment administered anywhere
- Estimated ATE will tend to exaggerate the true ATE

Core assumptions violated?

- Blinding: Researchers are not blinded to experimental assignment when measuring outcomes
- Attrition: Some of the subjects in the treatment group become discouraged and drop out of the study
- Compensatory behavior: Upon noticing that some subjects are excluded from a poverty aid treatment because they were assigned to the control group, an aid organization endeavors to treat those who were assigned to the control group
- Multiple outcomes: A weight loss intervention randomly assigns students who come to a cafeteria for lunch to a treatment consisting of small dishes and portions; outcomes are measured in terms of total calories consumed at the cafeteria during lunchtime

Average Treatment Effect (ATE)

$$\text{ATE} = \frac{1}{N}(Y_i(D = 1) - Y_i(D = 0))$$

- i.e., treatment effects averaged over those *receiving treatment*

Chattopadhyay & Duflo 2004

- Randomized policy experiment in India
- 1990s, one-third of village council heads reserved for women
- women.csv contains subset of data from West Bengal
- Gram Panchayat (GP) = level of government
- Analysis?
    - Was randomization implemented properly?
    - Conjecture: more drinking facilities under women
    - Conjecture: no effect on irrigation

Name	Description
GP	An identifier for the Gram Panchayat (GP)
village	identifier for each village
reserved	binary variable indicating whether the GP was reserved for women leaders or not
female	binary variable indicating whether the GP had a female leader or not
irrigation	variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
water	drinking-water facilities in the village since the reserve policy started

Table 4.6: *The Variable Names and Descriptions of the Women as Policy Makers Data.*

```
>setwd("~/Dropbox/Experimental␣Methodology/
    DPIR␣2017/qss-master/PREDICTION")
>women <- read.csv("women.csv")
>##proportion of female politicians in
    reserved GP vs. unreserved GP
>mean(women$female[women$reserved] ==1)
 [1] 1
>mean(women$female[women$reserved == 0])
 [1] 0.07476636
```

```r
## drinking-water facilities
mean(women$water[women$reserved == 1]) -
    mean(women$water[women$reserved ==0])

## [1] 9.25223

## irrigation facilities
mean(women4irrigation[women$reserved == 1]) -
    mean(women$irrigation[women$reserved == 0])

## [1] -0.3693319
```

Intent-to-Treat Effect

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

- Where Z is an indicator for treatment assignment
- With 100% compliance ATE=ITT
- ITT captures the average effect of being assigned to the treatment group regardless of the proportion of the treatment group actually treated

Complier Average Causal Effect (CACE)

$$\text{CACE} = \frac{\text{ITT}}{\sigma}$$

- where σ is the share of those assigned to the treatment group receiving treatment
- CACE also referred to as Local Average Treatment Effect (LATE) and Treatment on Treated (TOT)
- ATE among Compliers

ITT, ATE: Potential Outcomes

Obs	$Y_i(0)$	$Y_i(1)$	$D_i(0)$	$D_i(1)$	Type
1	4	6	0	1	Complier
2	2	8	0	0	Never-Taker
3	1	5	0	1	Complier
4	5	7	0	1	Complier
5	6	10	0	1	Complier
6	2	10	0	0	Never-Taker
7	6	9	0	1	Complier
8	2	5	0	1	Complier
9	5	9	0	0	Never-Taker

Compare ATT, ATE, and CACE

- ATE does not consider noncompliance:

$$\text{ATE} = \frac{2+6+4+2+4+8+3+3+4}{9} = 4$$

- ITT accounts for the fact that never-takers will not receive the treatment:

$$\text{ITT} = \frac{2+0+4+2+4+0+3+3+0}{9} = 2$$

- CACE is based on the subset of Compliers:

$$\text{CACE} = \frac{2+4+2+4+3+3}{6} = 3$$

Personal Canvass & Voting

- Gerber and Green New Haven study APSR 2000
- Randomly assign voters different GOVT tactics
    - Personal canvassing contact?
    - Mail?
    - Telephone?
    - Control?

New Haven Voter Mobilization

Turnout Rate	Treatment Group	Control Group
Among those contacted	54.43 (395)	
Among those not contacted	36.48 (1050)	37.54 (5645)
Overall	41.38 (1445)	37.45 (5645)

- ITT $=$ 41.38 - 37.54 $=$ 3.84
- $\sigma = 395/1445 = 0.273$
- CACE$=$ITT$/\sigma = 3.84/0.273 = 14.1$

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

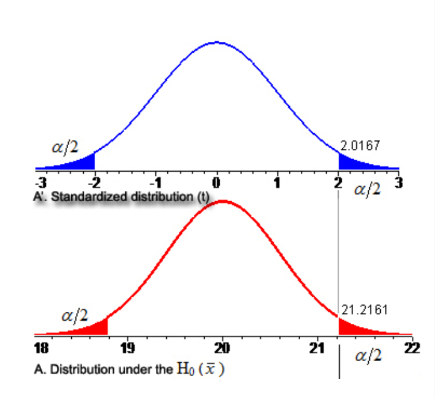
Statistical Power

- 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

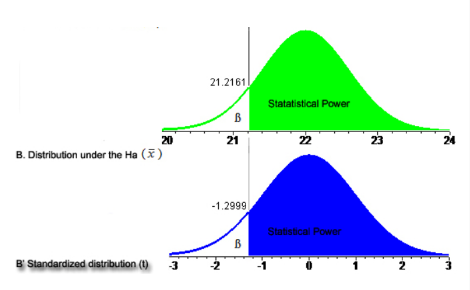
- What is the effect of losing Medicaid on infant mortality?
- $H_0 = 20$ deaths per 1,000 live births (assumed known without uncertainty here)
- True effect is an increase of 2 deaths per 1,000 live births
- Standard deviation in population is 4, we have N=44 observations; sampling distribution yields 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



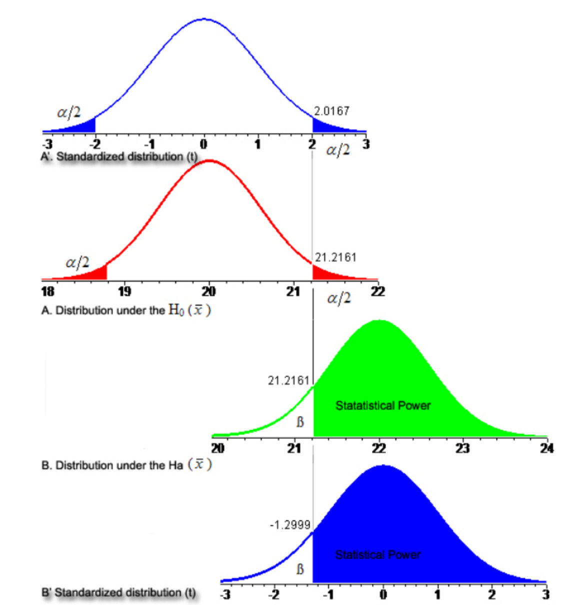
- Say for our test $\alpha = 0.05$
- Can rescale via Z-transformation
- What does this graphic mean?
- For $\hat{x} = 22$,
- t-stat=3.32, $p < 0.01$

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

Example 2: campaign effect

- In R, power.prop.test()
- Researcher thinks effects that move a proportion (i.e. vote support) from 50% to 52% are likely
- Would like to be able to detect effects of this size at conventional levels of statistical significance
- ($p = 0.05$; 95% confidence interval for the effect excludes zero), with power ($1 - \beta$) equal to 0.50
- $H_0 : \delta = \theta_1 - \theta_2 = 0$; $H_A : \delta \neq 0$ (two-sided alternative)

Power Estimate for 2 Point Effect

Two-sided alternative at conventional levels of significance

```
>power.prop.test(p1 = 0.5, p2 = 0.52, power = 0.5)
```

```
Two-sample comparison of proportions power calculation
n = 4799.903
p1 = 0.5
p2 = 0.52
sig.level = 0.05
power = 0.5
alternative = two.sided
NOTE: n is number in *each* group
```

Power Estimate for 2 Point Effect

One-sided alternative at conventional levels of significance

```
> power.prop.test(p1 = 0.5, p2 = 0.52,
   power = 0.5,
alternative= one.sided")
```

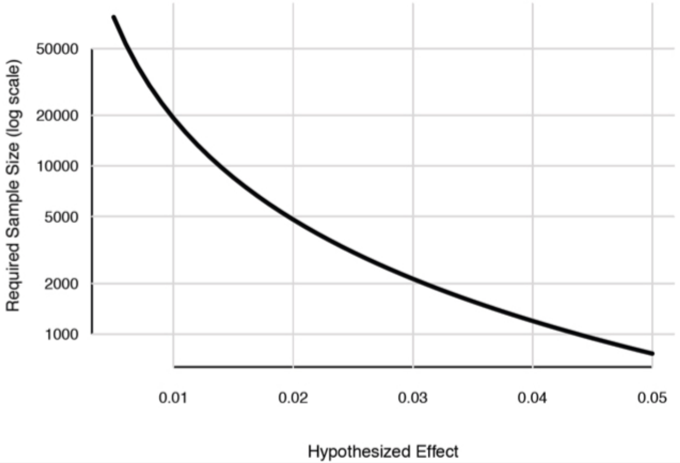
```
Two-sample comparison of proportions power calculation
n = 3380.577
p1 = 0.5
p2 = 0.52
sig.level = 0.05
power = 0.5
alternative = one.sided
NOTE: n is number in *each* group
```

Power Curves

```
> p> effects <- seq(0.005, 0.05, by =
    0.001)

> base <- 0.5
> m <- length(effects)
> n <- rep(NA, m)
> for (i in 1:m) {
n[i] <- power.prop.test(p1 = base, p2 =
    base + effects[i],
+ power = 0.5)$n
+})
```

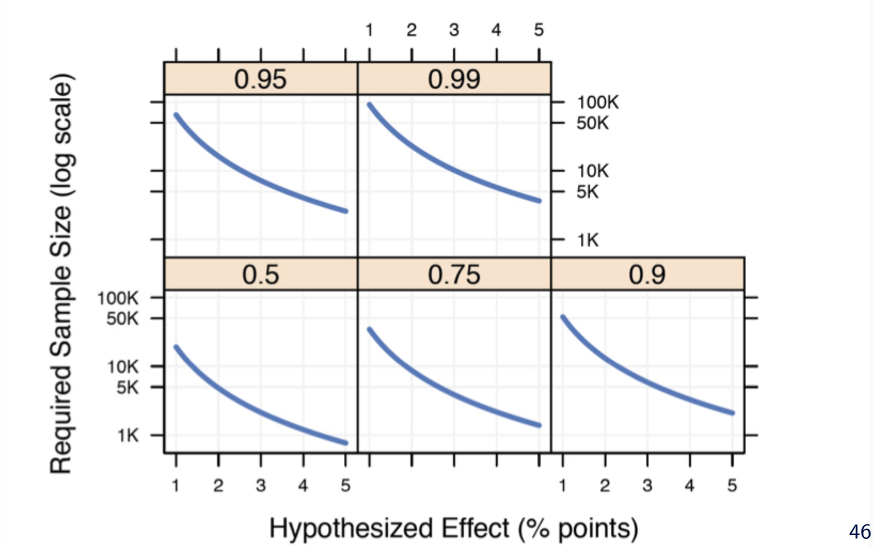
Power Curves



Looping over Power Curves

```
> power <- c(0.5, 0.75, 0.9, 0.95, 0.99)
> effects <- seq(0.01, 0.05, by = 0.001)
> base <- 0.5
> m <- c(length(power), length(effects))
> n <- matrix(NA, m[1], m[2])
> for (i in 1:(m[1])) {
+ for (j in 1:(m[2])) {
+ n[i, j] <- power.prop.test(p1 = base, p2
   = base + effects[j],
+ power = power[i])$n
+ }
+ }
```

Power Curves: different power levels



Practical Advice on Power

- What is "typical" size for effects, and how might we guess?
    - Some thoughts on later example
- Generally, experiments require $1 - \beta > 0.8$ to get funding
- Zaller's maxim: "Do your power analysis, figure out your sample size, then double it"

Practical Advice on Power

- Cost considerations: Gerber and Green turnout experiment
    - One component involved canvassing
    - $40 per hour for a pair of students, 6,000 treated
    - If 6 houses an hour, need 1000 hours, so $40k right there alone
    - Implications based on power curve slide
- In particular costs high for general population experiments
- Anyone have guesses how much surveys cost?
- How much value?