Experimental Methods: Lecture 3

Non-Compliance and Power

Raymond Duch

May 12, 2021

University of Oxford

Road Map

- Noncompliance
- Statistical power

Noncompliance

Intuition

- If $ATE = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$, how can we ever know that subjects were actually treated?
- More importantly, *what does it mean to be "treated"?*
- Let us distinguish between
  - Assignment of treatment (Z)
  - Receipt of treatment (D)
- Yes, by the exclusion restriction, $Y_i(z, d) = Y_i(d)$
- However, in many applications, $z_i \neq d_i$
- Noncompliance with treatment assignment $=$ subjects do not receive the treatment to which they were assigned

Kalla and Broockman 2020

- Kalla and Broockman (2020): Reducing exclusionary attitudes through interpersonal conversation (APSR)
- 230 canvassers *are assigned to* have face-to-face conversations with 6,869 voters deploying non-judgmental exchange of narratives on a range of topics
- Outcome: Exclusionary immigration policy and prejudicial attitudes
- What can go wrong?
- In this example, when does a subject comply with the treatment assignment?

TABLE 1. Summary of Differences Between Conditions and Results in Previous Study and Experiments 1–3

Study	Broockman and Kalla (2016)	Experiment 1		Experiment 2		Experiment 3
Topic	Transphobia	Unauthorized immigrants		Transphobia		Transphobia
Condition name	Full Intervention	Full Intervention	Abbreviated Intervention	Participants' and Video Narratives	Video Narratives Only	Participants' Narratives by Phone
Intervention contents						
Non-judgmental exchange of narratives…						
○ From participants (voter and canvasser)	YES	YES	NO	YES	NO	YES
○ In video	YES	NO	NO	YES	YES	NO
Address concerns and deliver talking points	YES	YES	YES	YES	YES	YES
Results						
ITT[a]	Positive effects ($d = 0.16$, $p < 0.001$)	Positive effects ($d = 0.08$, $p < 0.001$)	Null effects ($d = 0.02$, $p = 0.27$), statistically distinguishable from Full Intervention ($d = 0.06$, $p < 0.01$)	Positive effects ($d = 0.08$, $p < 0.001$)	Positive effects ($d = 0.08$, $p < 0.001$)	Positive effects ($d = 0.04$, $p < 0.001$)
CACE[b]	$d = 0.22$	$d = 0.12$	$d = 0.03$ (Abbreviated vs. Placebo)	$d = 0.10$	$d = 0.10$	$d = 0.08$

Notes: Each Experiment also contained a Placebo condition not shown in the table. These Placebo conditions contained no persuasive content on the topics but are used as a baseline for comparison when estimating the effect sizes shown in the table.

[a]To summarize the results of each study, we first average the pre-specified Overall Index in each study across survey waves to compute a pooled Overall Index. We then report intent-to-treat (ITT) effects on this pooled Overall Index, which represents the mean difference between individuals assigned to each condition among all individuals who identified themselves at their doors, regardless of whether the conversation continued after that point. The ITT estimates represent the average causal effect of attempting to treat people who open their doors, even if they refuse to converse soon after. This means the ITT estimates are "diluted" by the presence of individuals who open the door but do not enter into the conversation.

[b]To estimate the implied Complier Average Causal effect (CACE), or the effect among those who received the intervention, we estimate compliance under a conservative definition of compliance, whether participants got to the "first rating" part of the conversation where they initially told canvassers how they felt about the policy. The CACE estimates represent the average causal effect of treating the people who do

Definition and formalization

- Where is the *ATE* row? What are *ITT* and *CACE*?
- Let $d_i(z)$ denote whether subject i is actually treated when treatment assignment is z
- There are different types of compliance and noncompliance with the treatment
- **Compliers**: $d_i(1) = 1$ and $d_i(0) = 0$ or $d_i(1) > d_i(0)$
- **Never-Takers**: $d_i(1) = 0$ and $d_i(0) = 0$
- **Always-Takers**: $d_i(1) = 1$ and $d_i(0) = 1$
- **Defiers**: $d_i(1) = 0$ and $d_i(0) = 1$ or $d_i(1) < d_i(0)$

Definition and formalization

- These groups are formed *after* random assignment, not formed *by* random assignment $\rightarrow$ they might differ systematically in ways that bias *ATE* estimator
- 2 types of noncompliance
  - One-sided: $d_i(1) = 0$ for some i but $d_i(0) = 0 \ \forall i$ (only compliers and never-takers)
  - Two-sided: additionally, $d_i(0) = 1$ for some i (these can be defiers or always-takers)
- In any experiment facing noncompliance, which subjects *could* make up the treatment group, and which the control group? How might that look like in Kalla and Broockman (2020)?
- What is the problem of naively comparing treated and untreated subjects, i.e. estimate *ATE*?

Estimation of treatment effects under noncompliance

- What groups *could* we compare to unbiasedly estimate a treatment effect?
- 2 estimands
    - Intent-to-treat effect (*ITT*)
    - Complier average causal effect (*CACE*)
- Choice of estimands depends, of course, on your research question and goal of causal inference

Intent-to-Treat Effect

$$\text{ITT} \equiv E[Y_i(z = 1)] - E[Y_i(z = 0)]$$
$$= E[Y_i(z = 1, d(1))] - E[Y_i(z = 0, d(0))]$$

- ITT captures the average effect of being assigned to the treatment group regardless of the proportion of the treatment group actually treated

- Which causal inference method does this setup remind you of?

Complier Average Causal Effect

$$\text{CACE} \equiv \underbrace{E[(Y_i(d=1) - Y_i(d=0))}_{\text{average treatment effect}} \, | \, \underbrace{d_i(1) - d_i(0) = 1]}_{\text{among Compliers}}$$

Let

$$\pi_C = E[d_i(z=1) - d_i(z=0)]$$

be the proportion of compliers in the sample.

Then, the sample analog of the *CACE* estimand is

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

9

- Assumptions: Non-interference, excludability, and, under 2-sided noncompliance, monotonicity (no defiers, i.e. $d_i(1) \geq d_i(0)$)
- CACE also referred to as Local Average Treatment Effect (LATE) and, under one-sided noncompliance, Treatment on Treated (TOT)
- ATE among Compliers

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 ITT, 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 (always-takers will 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
- $\pi_C = 395/1445 = 0.273$
- CACE = ITT$/\pi_C$ = 3.84/0.273 = 14.1

14

Kalla and Broockman (2020)

FIGURE 1. Experiment 1 Results: Intent-to-Treat Effects
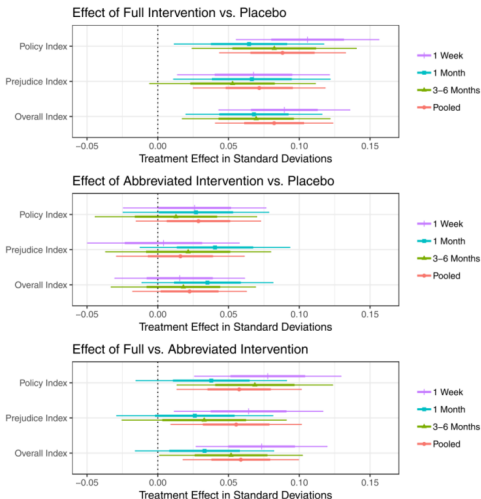
Notes: Each panel shows the estimated intent-to-treat effects when comparing the two experimental conditions described in the panel title (e.g., the top panel compares the Full Intervention condition to the Placebo condition). Within each panel, we show treatment effects on the pre-specified primary outcome indices. Results are average treatment effects with 1 standard error (thick) and 95% confidence intervals (thin). To form each pooled index, we average each respondent's values for the corresponding index across all post-treatment survey waves. See Online Appendix Tables OA.9–11 for numerical point estimates and standard errors.

Broader takeaways

1. Carefully define the treatment itself
2. Carefully define treatment assignment and treatment receipt
3. Carefully define and try to identify compliant and non-compliant subgroups of subjects

Design implications

Bear in mind that

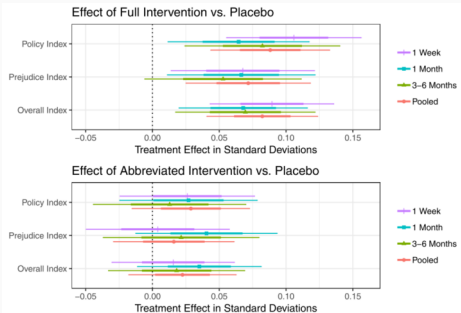$$\mathrm{SE}(\widehat{CACE}) \approx \frac{\mathrm{SE}(\widehat{ITT})}{\pi_C}$$

- Increase π_C; rule out defiers
- 1-sided noncompliance: Placebo design

Placebo Design

- Researchers attempt to contact individuals assigned to receive the treatment
- Those reached are then randomly allocated to two different groups
    - Treatment group
    - Placebo group receiving a "non-treatment"
- Kalla and Broockman (2020) canvassing experiment
    - Narratives (treatment)
    - Housing in Orange County (placebo)
- CACE estimated by comparing the outcomes for those in the treatment group to those in the placebo group
    - Random sample of Compliers whose untreated potential outcomes can be measured

Kalla and Broockman (2020)

Placebo Design

- Logic is that placebo design screens out Never-Takers (since they are, in addition to compliers, part of the control group under 1-sided noncompliance)

- Compliers in the treatment group are compared directly to Compliers in the untreated group

- Reduces noise from Never-Takers in both treatment and control groups

- Moves us to a world of "full compliance"

Placebo Design

- Downside is that not all Compliers receive the treatment
- Resources are wasted on those receiving the placebo
- Opportunity to collaborate with someone studying an unrelated topic

Placebo Design

- The placebo and conventional design both allow estimation of the CACE
- Choice depends on the budget and compliance rate
- Under a fixed budget, the conventional design is preferable if compliance rate $> 50\%$ ($\pi_C > 1/2$)
- Canvassing studies often have a lower rate
- A pilot study may give a better idea of the expected compliance rate

Voter Turnout Example

- Gerber, Green, and Larimer (2008) interested in the effect of communication on turnout
- U.S. has voters files, anyone know what they are?
- 180,000 Michigan households in experiment
- 100,000 in control group (no postcards), other groups 20,000 each
- **Civic duty**: "It's your civic duty to vote"
- **Hawthorne**: "It's your civic duty to vote, we're doing a study and will check public records"
- **Self**: "You should vote, here's your recent voting record"
- **Neighbors**: "You should vote, here's your neighbors' voting records and your own"

	Control	Civic	Hawthorne	Self	Neighbors
Pct Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N	191,243	38,218	38,204	38,218	38,201

Anyone here know how Gerber followed up on this study?

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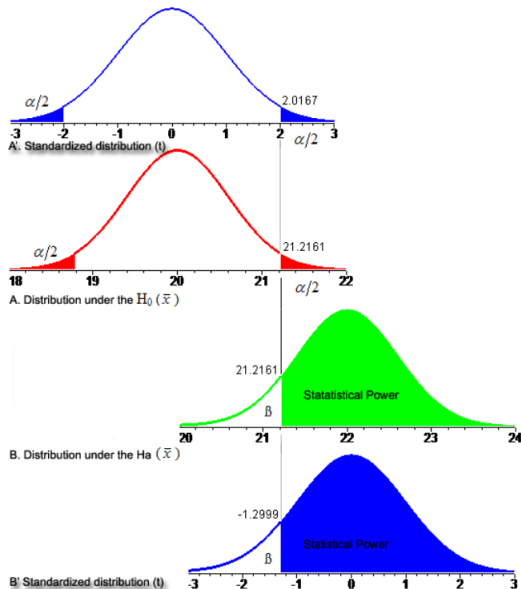
Sampling Distribution Under Null
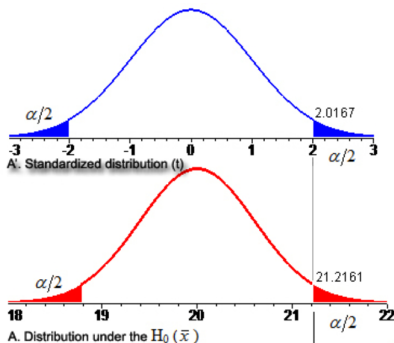
- Say for our test $\alpha = 0.05$
- Can rescale via Z-transformation
- What does this graphic mean?
- For $\hat{x} = 22$,
- t-stat=3.32, $p < 0.01$

Sampling Distribution of $\hat{x}$
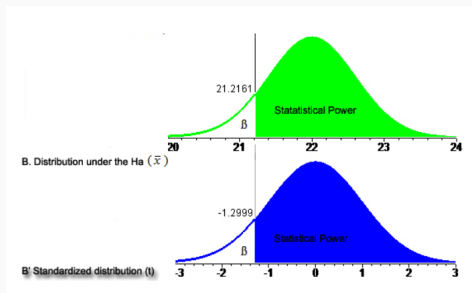
- Interpret this graphic
- $1 - \beta$ is fraction of estimates that reject null hypothesis
- Power of the test
- What x_{true} yields $1 - \beta = 0.5$?
- What parameters are needed?

The Relationship Between α and β

30

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 $T = \hat{\theta} \big/ \sqrt{\hat{V}(\hat{\theta})}$
- Example: Mean of normal distribution θ, data $y = (y_1, ..., y_n)$, iid

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

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

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

- 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

R Code for formula

```
power_calculator <- function(mu_t, mu_c,
  sigma, alpha=0.05, N){
lowertail <- (abs(mu_t - mu_c)*sqrt(N))/
  (2*sigma)
uppertail <- -1*lowertail
beta <- pnorm(lowertail - qnorm(1-alpha/2)
  , lower.tail=TRUE) + 1- pnorm(
  uppertail - qnorm(1-alpha/2), lower.
  tail=FALSE)
return(beta)
}
```

Simulation to Estimate Power

```
possible.ns <- seq(from=100, to=2000, by=40) # The
    sample sizes we'll be considering
stopifnot(all( (possible.ns %% 2)==0 )) ## require
    even number of experimental pool
powers <- rep(NA, length(possible.ns)) # Empty
    object to collect simulation estimates
alpha <- 0.05 # Standard significance level
sims <- 500 # Number simulations conduct for each N
#### Outer loop to vary the number of subjects ####
for (j in 1:length(possible.ns)){ N <- possible.ns[
    j] # Pick the jth value for N
  Y0 <- rnorm(n=N, mean=60, sd=20) # control
      potential outcome
  tau <- 5 # Hypothesize treatment effect
  Y1 <- Y0 + tau # treatment potential outcome
  significant.experiments <- rep(NA, sims) # Empty
      object to count significant experiments
```
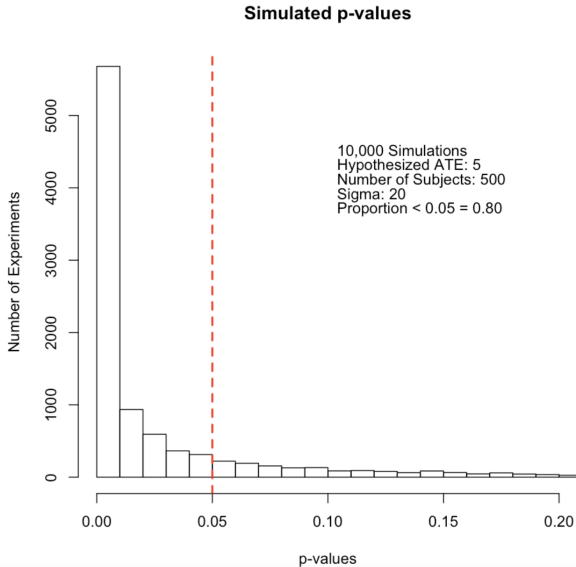
Simulation to Estimate Power

```
#### Inner loop to conduct experiments "sims"
    times over for each N ####
        Y0 <- rnorm(n=N, mean=60, sd=20) # control
            potential outcome
        tau <- 5 # Hypothesize treatment effect
        Y1 <- Y0 + tau # treatment potential
            outcome
for (i in 1:sims){
        ## Z.sim <- rbinom(n=N, size=1, prob=.5) #
            Do a random assignment by coin flip
        Z.sim <- sample(rep(c(0,1),N/2)) ## Do a
            random assignment ensuring equal sized
            groups
        Y.sim <- Y1*Z.sim + Y0*(1-Z.sim) # Reveal
            outcomes according to assignment
```

Simulation to Estimate Power

```
        fit.sim <- lm(Y.sim ~ Z.sim) # Do analysis
            (Simple regression)
        p.value <- summary(fit.sim)$coefficients
            [2,4] # Extract p-values
        significant.experiments[i] <- (p.value <=
            alpha) # Determine significance
            according to p <= 0.05
        }
  powers[j] <- mean(significant.experiments) #
      store average success rate (power) for each N
  }
powers
```

Simulated p-values

10,000 Simulations
Hypothesized ATE: 5
Number of Subjects: 500
Sigma: 20
Proportion < 0.05 = 0.80

Power Analysis: Duch & Torres

Motivation

- Malfeasance messaging experiments often result in null findings – subjects may not be updating their priors

- Choice Architecture provides some suggestions as to why

- We messaging treatment experiment to identify optimal framing of malfeasance messages

Metric Treatments

- *Standard*: presents the total number of irregularities reported for the subject's municipality
- *Severity*: subjects informed about the number of serve irregularities
- *Resources - Total*: total cost of irregularities and expresses this as a percent of the total municipality budget.
- *Resources - Individual*: expresses malfeasance costs in terms of the tax burden of individuals – expressed as the share of every $1,000 Chilean pesos that the municipal budget spends is lost due to irregularities.
- *Resources - Foregone Loses* total cost of malfeasance in terms of lost funding for influenza vaccines in the municipality

40

Benchmarking Treatments

- *spatial*" subjects learn how the reported irregularities for their municipality compare to those of other municipalities in their region.

- *temporal*: the irregularities reported for their municipality are compared to those reported in the municipality's previous Contraloria audit report.

Standard Evaluation Questions

- The content of the video is reliable (Strongly agree, Agree, Neutral, Disagree, Strongly Disagree)

- The content of the video is trustworthy (Strongly agree, Agree, Neutral, Disagree, Strongly Disagree)

- The content of the video is convincing (Strongly agree, Agree, Neutral, Disagree, Strongly Disagree)

- The content of the videos credible (Strongly agree, Agree, Neutral, Disagree, Strongly Disagree)

Frame	Metric	Outcome audit	Sample	Treatment
Spatial	Standard	Positive	160	T_1
		Negative	160	T_2
	Severity	Positive	160	T_3
		Negative	160	T_4
	Resource total	Positive	160	T_5
		Negative	160	T_6
	Resource individual	Positive	160	T_7
		Negative	160	T_8
	Foregone loss	Positive	160	T_9
		Negative	160	T_{10}
	Program	Positive	160	T_{11}
		Negative	160	T_{12}
Temporal	Standard	Positive	500	T_{13}
		Negative	160	T_{14}
	Severity	Positive	160	T_{15}
		Negative	160	T_{16}
	Resource total	Positive	160	T_{17}
		Negative	160	T_{18}
	Resource individual	Positive	160	T_{19}
		Negative	160	T_{20}
	Foregone loss	Positive	160	T_{21}
		Negative	160	T_{22}
	Program	Positive	160	T_{23}
		Negative	160	T_{24}

Table 3: Factorial design

Duch & Torres Power

- 1000 times generate a treatment schedule according to the number of individuals in the sample (1,500 to 3,500)
- Assume treatment effect τ is 0.0 0.05, 0.1, 0.15, 0.20 and 0.25 each of six treatment arms
- Each subject gets 6 randomly assigned videos
- effect size (ite) function treatment assignment (τ)
- Outcome; $Y =$ Individual Fixed Effect $+ \tau$ (treatment) $+$ draw from random normal (mean 0 and sd 0.4)
- for samples 1000, 2000, 2500, 3000, 3500) 1000 draws from normal and estimate distribution of outcomes
- regress outcomes on treatment assignment and retain the p value of the estimated coefficient
- proportions of p values $< 0.05 =$ Power!

Duch & Torres Simulation to Estimate Power

```
library(tidyverse)

set.seed(89)
taus_metric <- c(0,0.05,0.1,0.15,0.2,0.25)
n_vids <- 6

# Schedule of treatment effects by treatment arm
treat_effects <- data.frame(arm = 1:24,
                            comparison = rep(c("
                                spatial","temporal"
                                ), each = 12),
                            metric = c("standard","
                                severity","resource
                                _total",
```

Duch & Torres Simulation to Estimate Power

```
                    "resource_ind","resources_foregone"
                       ,"program"),
                            outcome = rep(c("
                                positive","negative
                                "), each = n_vids),
                            tau = c(taus_metric, -
                                taus_metric))

# Function to estimate power for a given number of
    subjects
calc_power <- function(subjects) {

  B <- 1000 # No. of iterations
  power_results <- matrix(ncol = 24, nrow = B) #
    Matrix to store results

  for (b in 1:B) {
```

Duch & Torres Simulation to Estimate Power

```
# Generate distribution of treatment
    assignments and individual-level fixed
    effects
fake_data <- data.frame(id = rep(1:subjects,
    each = n_vids),
                        id_fe = rep(rnorm(
                            subjects), each = n
                            _vids),
                        assign = as.vector(
                            replicate(subjects,
                             sample(1:24, n_
                            vids))))
# The last line of code here takes separate
    samples (without replacement) of the
    treatment
```

Duch & Torres Simulation to Estimate Power

```r
# Get corresponding treatment effect
fake_data$ite <- treat_effects$tau[fake_data$
    assign]

# Generate outcome
fake_data$Y <- fake_data$id_fe + fake_data$ite
    + rnorm(subjects*n_vids, 0, 0.4)

# Convert assignment and ids to factors for
    easy handling
fake_data$assign <- as.factor(fake_data$assign)
fake_data$id <- as.factor(fake_data$id)

# Generate vector of indicators if p < 0.05
power_results[b,] <- summary(lm(Y ~ assign,
    fake_data))$coefficients[,4] < 0.05 #
    Extract the p-values
```

Duch & Torres Simulation to Estimate Power

```
# Collapse simulation results into proportion of
    times p < 0.05 per treatment arm
power_vec <- apply(power_results, 2, function(x)
    sum(x)/B)

return(power_vec)

}
```

Duch & Torres Simulation to Estimate Power

```r
# Potential Ns
Ns <- c(1500,2000,2500,3000,3500)

# Simulate power for each N
results <- sapply(Ns, calc_power)
colnames(results) <- Ns

# Coerce results to make it easy to plot graph
plot_df <- as.data.frame(results) %>%
  mutate(arm = 1:24) %>%
  pivot_longer(-arm) %>%
  rename(N = name, Power = value) %>%
  mutate(N = as.numeric(N)) %>%
  left_join(treat_effects, by = "arm") %>%
  filter(arm < 7) %>%
  mutate(print_label = paste0(metric,"_=_",tau))
```
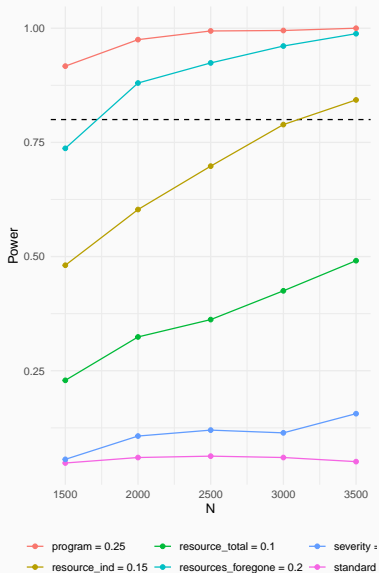
Duch & Torres Simulation to Estimate Power

```
# Plot
ggplot(plot_df, aes(x = N, y = Power, color = print
    _label)) +
  geom_hline(yintercept = 0.8, linetype = "dashed")
      +
  geom_point() +
  geom_line() +
  theme_minimal() +
  labs(color = "") +
  theme(legend.position = "bottom") +
  ggsave("contraloria_power.pdf", width = 8.5,
      height = 4.5)
```

Power Curves

52

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
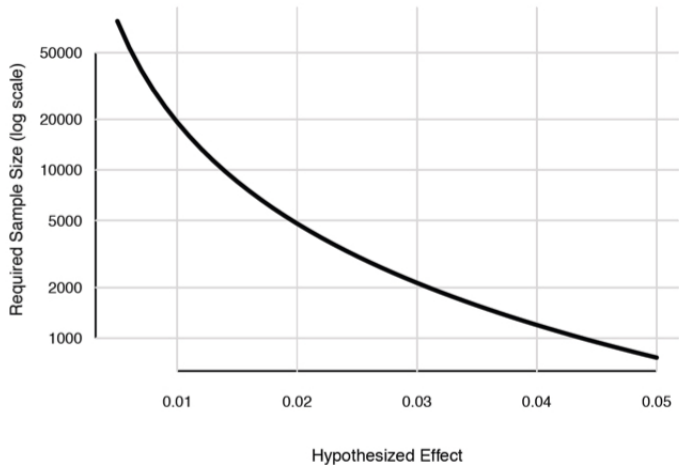power = 0.5
alternative = one.sided
NOTE: n is number in *each* group

Power Curves

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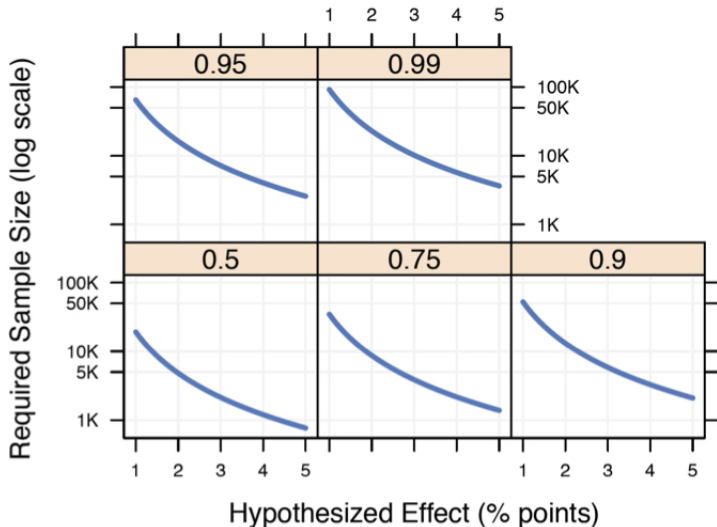
```
> power <- c(0.5, 0.75, 0.9, 0.95, 0.99)
> effects <- seq(0.01, 0.05, by = 0.001)
> base <- 0.5
> m <- c(length(power), length(effects))
> n <- matrix(NA, m[1], m[2])
> for (i in 1:(m[1])) {
+ for (j in 1:(m[2])) {
+ n[i, j] <- power.prop.test(p1 = base, p2
  = base + effects[j],
+ power = power[i])$n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

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?