Experimental Methods: Lecture 4

Effect Heterogeneity and Power

Raymond Duch

May 19, 2021

University of Oxford

Road Map

- Effect heterogeneity: theory
- Power

Effect heterogeneity: theory

Motivation

- Recall the fundamental assumption about treatment effects for the RI confidence interval estimator
- What does "constant treatment effects" really mean?
- More importantly, is the average treatment effect the same for every single observation in the sample?
- Furthermore, we are often interested in the "generalizability" of experimental findings and their policy relevance
- Treatment effect heterogeneity is one way to address these issues

Ferraz and Finan 2008

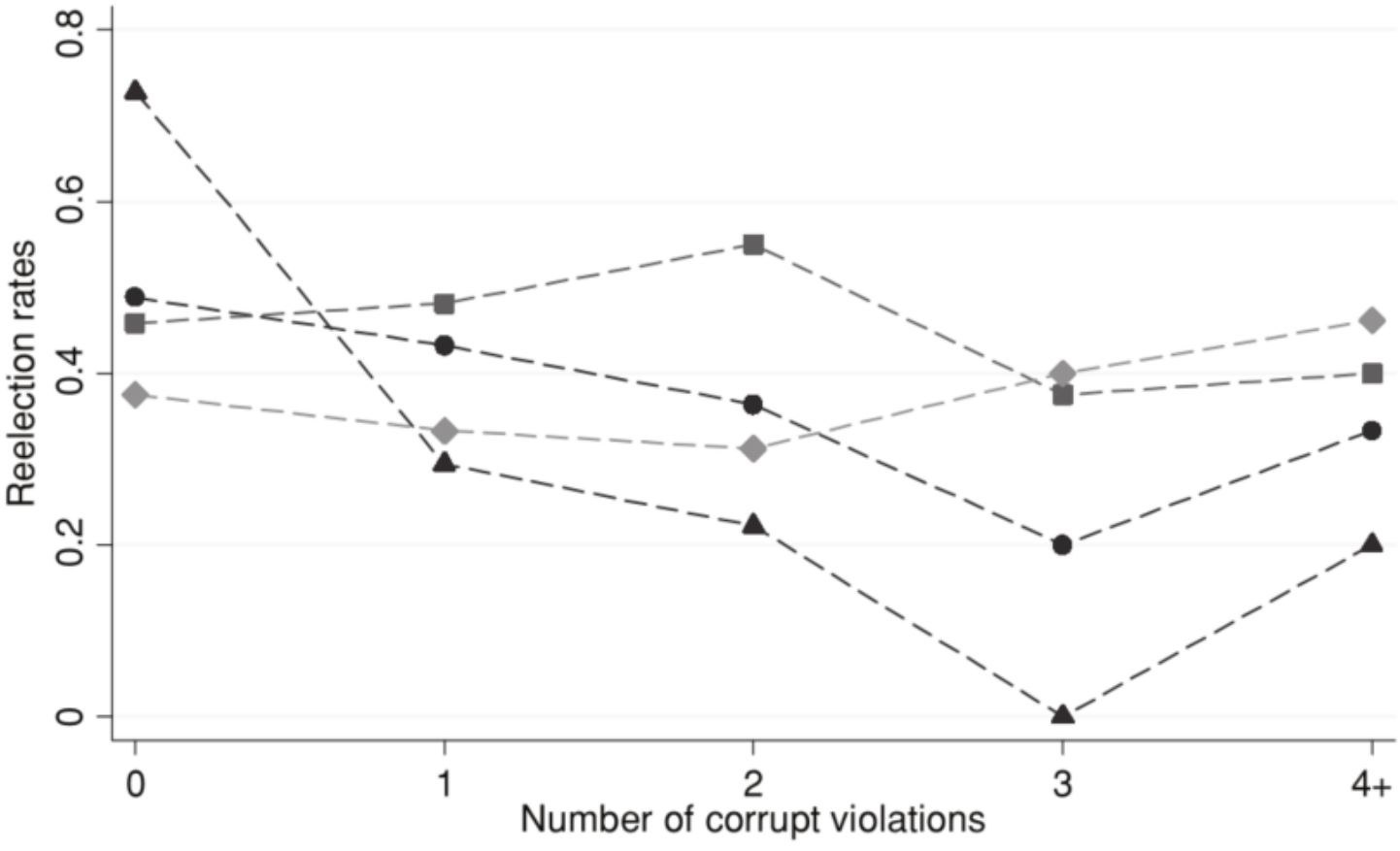


FIGURE IV
Relationship between Reelection Rates and Corruption Levels

Theory

We move away from constant treatment effects and therefore define

$$\tau_i \equiv Y_i(1) - Y_i(0) \tag{1}$$

The fundamental interest under treatment effect heterogeneity is in

$$\begin{aligned} Var(\tau_i) &= Var(Y_i(1) - Y_i(0)) \\ &= Var(Y_i(1)) + Var(Y_i(0)) + 2Cov(Y_i(1), Y_i(0)) \end{aligned} \tag{2}$$

Informally, we define treatment effect heterogeneity as *variance of the treatment effect τ_i across subjects*.

What is the problem with Eq. 2?

Theory

- This is an old and now for us very familiar problem:
- Any experiment does not allow us to estimate every component of $Var(\tau_i)$
- We have information about the marginal distributions of $Y_i(1)$ and $Y_i(0)$, but not about the joint distribution of these potential outcomes
- So what should we do?

Bounding $Var(\tau_i)$

- Recall that by randomization, $E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$
- We can pair each observed $Y_i(1)$ with one of the observed $Y_i(0)$
- But which one? Many combinations possible
- We place bounds suggesting how large or small $Var(\tau_i)$ may be
- Pair values of $Y_i(0)$ and $Y_i(1)$ such that implied $Cov(Y_i(0), Y_i(1)$ is as large (upper bound) or as small (lower bound) as possible
- Sort values in ascending-ascending / ascending-descending order

Testing for heterogeneity

Suppose $H_0 : Var(\tau_i) = 0$ What if we compared $Var(Y_i(1))$ and $Var(Y_i(0))$?

Note that

$$\begin{aligned} Var(Y_i(1)) &= Var(Y_i(0) + \tau_i) \\ &= Var(Y_i(0)) + Var(\tau_i) + 2Cov(Y_i(0), \tau_i) \end{aligned} \tag{3}$$

Then, the Null of constant τ_i implies that

$$Var(\tau_i) = -2Cov(Y_i(0), \tau_i) = 0 \tag{4}$$

These two terms therefore cancel in Eq. 3 and we have shown that testing $H_0 : Var(\tau_i) = 0$ is the same as testing $Var(Y_i(1)) = Var(Y_i(0))$

Observed Outcome Local Budget

We can test this with randomization inference

	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
Mean	16	22.5
Variance	17.5	112.5

Variance in control:

$$\frac{1}{7-1}2(15-16)^2+2(20-16)^2+(10-16)^2=17.5$$

Variance in treatment: $\frac{1}{2-1}(15-22.5)^2+(30-22.5)^2=112.5$

Interaction

- These approaches test *whether* τ_i varies
- But we want to know more: conditions under which τ_i varies
- We are interested in a different estimand: Conditional Average Treatment Effect (CATE) = ATE for a defined subset of subjects $\tau_i(x) = E[Y_i(1) - Y_i(0)|X_i = x]$ (individual), and, if distribution of X_i is known, $E[\tau_i(X_i)]$ is identified (average)
- Change in treatment effect that occurs from one subgroups to the next is the difference between 2 CATEs
- These subgroups can either be defined by covariate values (*treatment-by-covariate interactions*) or by design (*treatment-by-treatment interactions*)

Treatment-by-covariate interactions

- What is the H_0 here?
- We can test the difference in CATEs with randomization inference or in a regression framework

$$Y_i = a + bI_i + cP_i + dI_iP_i + u_i \tag{5}$$

When $P_i = 0$, the CATE is b:

$$Y_i = a + bI_i + u_i \tag{6}$$

When $P_i = 1$, the CATE is $b + d$:

$$Y_i = a + bI_i + c + dI_i + u_i = (a + c) + (b + d)I_i + u_i \tag{7}$$

where d yields the change in CATEs that occurs when P_i changes

Treatment-by-covariate interactions

- An alternative is to conduct an F test via regression
- Compares sum of squared residuals from the two nested models (alternative model is Eq. 5 and null model is $Y_i = a + bI_i + cP_i + u_i$)
- If there are interaction affects, Eq. 5 should reduce SSR
- Simulate random assignments and calculate fraction of F-statistics at least as large as the observed F-statistic
- H_0 is that 2 CATEs are the same

Treatment-by-covariate interactions

- We can also use randomization inference!
- Recall estimated ATE from teacher incentives experiment is 3.5
- Does the treatment effect vary by level of parent literacy?
- $CATE_{submedian} = 11.14 - 7.83 = 3.31$
- $CATE_{abovemedian} = 12.26 - 8.57 = 3.69$
- We conduct a 2-tailed test to assess whether the difference in CATEs could have occurred by chance
- H_0: CATEs in both groups are equal to estimated ATE
- Full schedule of potential outcomes assuming constant $ATE = 3.5$ and assign subjects to treatment and control a 100,000 times
- How often does one obtain an observed difference at least as large as $|3.69 - 3.31| = 0.38$?

Caveats

- Multiple comparisons problem:
  - With 20 covariates, the probability of finding at least 1 that significantly interacts with the treatment at $\alpha = 0.05$ is $1 - (1 - 0.05)^{20} = 0.642$
  - Bonferroni correction (divide target p-value by number of hypothesis tests h)
  - Pre-register your design! (lab)
- Subgroup analysis is non-experimental: groups that are not formed by random assignment, but pre-assignment
- Teacher incentives and teacher education

Treatment-by-treatment interactions

- Manipulate treatment *and* contextual factor / personal characteristic (e.g. COVID and community infection levels)
- Define a factorial experiment as an experiment involving factors 1 and 2, with factor 1 conditions being A and B, and factor 2 conditions being C and D and E
- Then, allocate subjects at random to every possible combination of experimental conditions
- $\{AC, AD, AE, BC, BD, BE\}$

Gottlieb et al. 2018: EGAP Metaketa II: Taxation

Jessica Gottlieb, Adrienne LeBas, Nonso Obikili: "Formalization, Tax Appeals, and Social Intermediaries in Lagos, Nigeria"

T1. Control condition, not encouraged

T2. Encouraged, but not receiving a follow-up visit

T3. Encouraged, and receiving one of the following four follow-up visit combinations:

- T3a. Public goods message from state representative
- T3b. Enforcement message from state representative
- T3c. Public goods message from marketplace representative
- T3d. Enforcement message from marketplace representative

Figure 2: Research Design and Assignment Probabilities

				Message Type	
				Public Goods	Enforcement
Control	Formalization Intervention only	Delivery Type	State Rep.	T3a: 5/36	T3b: 5/36
T1: 1/6	T2: 5/18		Market Association	T3c: 5/36	T3d: 5/36

Multiple treatment arms

From Rosen 2010

	Colin		Jose	
	Good grammar	Bad grammar	Good grammar	bad grammar
% Received reply	52	29	37	34
(N)	(100)	(100)	(100)	(100)

This design requires us to be especially careful with defining the causal estimand – what quantity are we interested in in this application?

Multiple treatment arms

Quiz: Why would these two models estimate the same quantities from the Rosen 2010 experiment?

$\{NG, HG, NB, HB\}$ are indicator variables for each of the 4 treatment groups

$J_i = 1$ if Jose Ramirez; $G_i = 1$ if good grammar

$$Y_i = b_1 CG + b_2 JG + b_3 CB + b_4 JB + u_i$$
$$Y_i = a + bJ_i + cG_i + d(J_i G_i) + u_i$$

What quantity in the table do each of the coefficients represent?

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

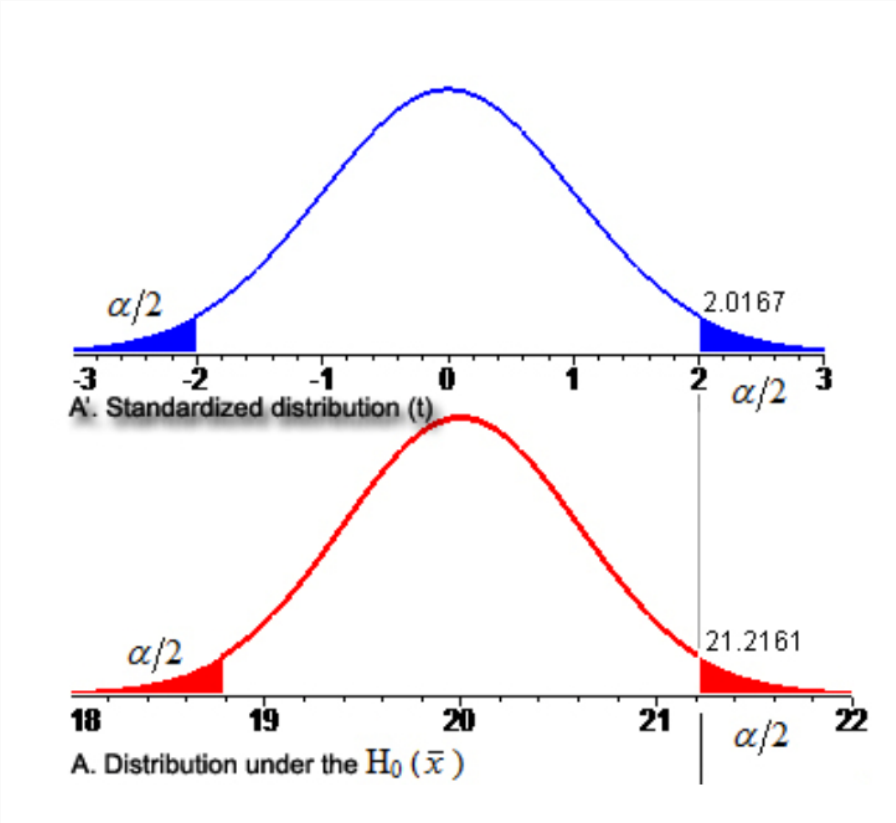
Statistical Power

- 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

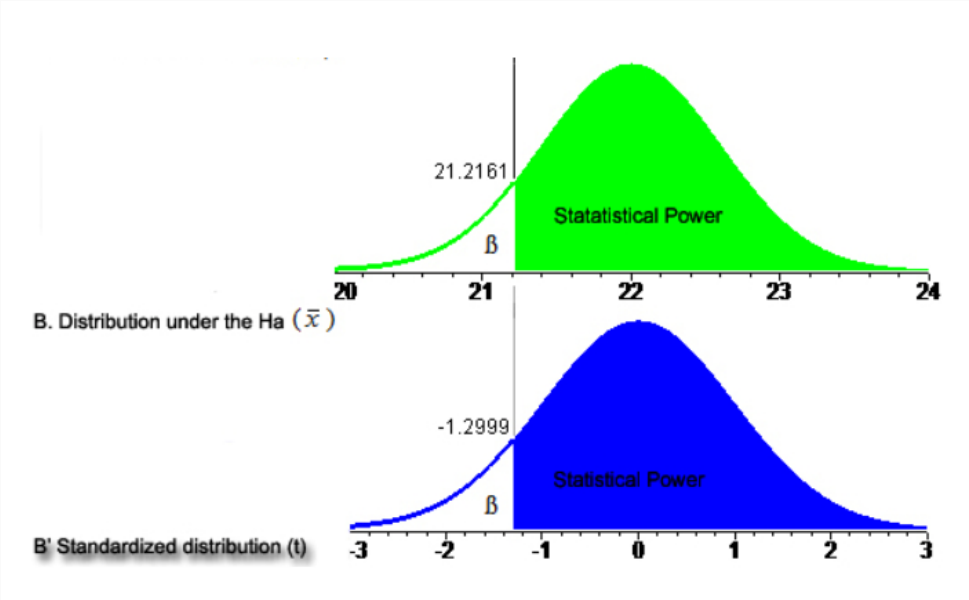
- What is the effect of losing Medicaid on infant mortality?
- $H_0 = 20$ deaths per 1,000 live births (assumed known without uncertainty here)
- True effect is an increase of 2 deaths per 1,000 live births
- Standard deviation in population is 4, we have N=44 observations; sampling distribution yields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



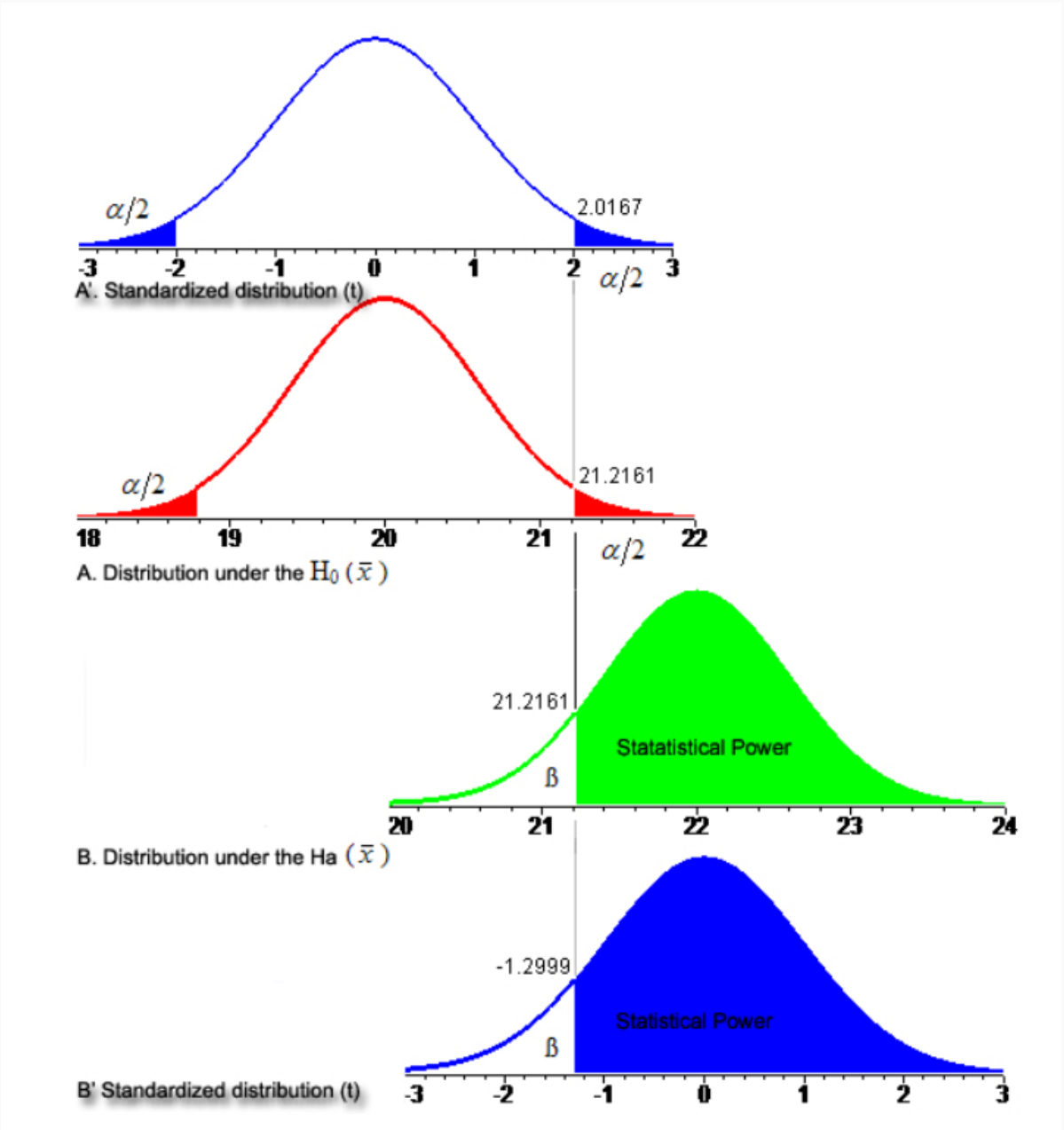
- Say for our test $\alpha = 0.05$
- Can rescale via Z-transformation
- What does this graphic mean?
- For $\hat{x} = 22$,
- t-stat=3.32, $p < 0.01$

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)

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

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
```

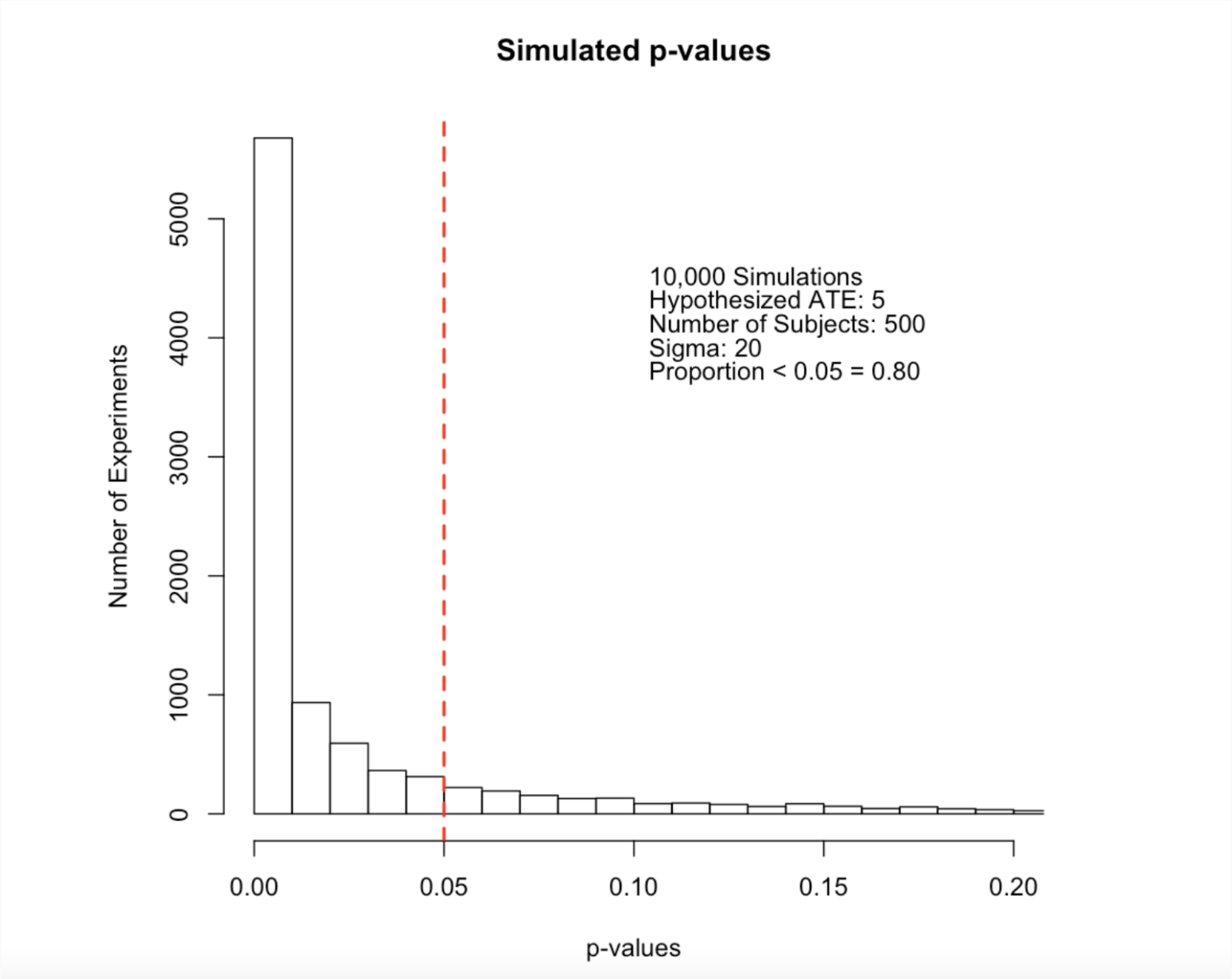
Simulation to Estimate Power

```
#### Inner loop to conduct experiments "sims"
    times over for each N ####

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



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

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 + τ (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

```
# 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)

}
```

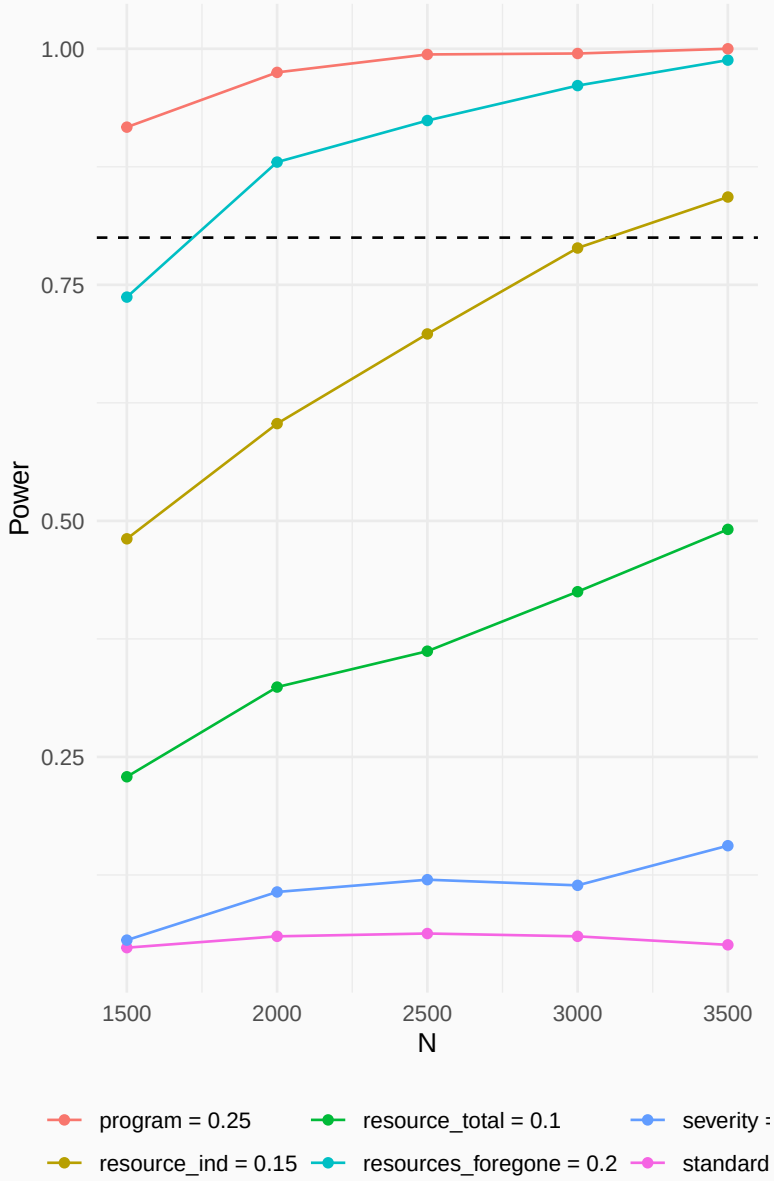
Duch & Torres Simulation to Estimate Power

```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



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

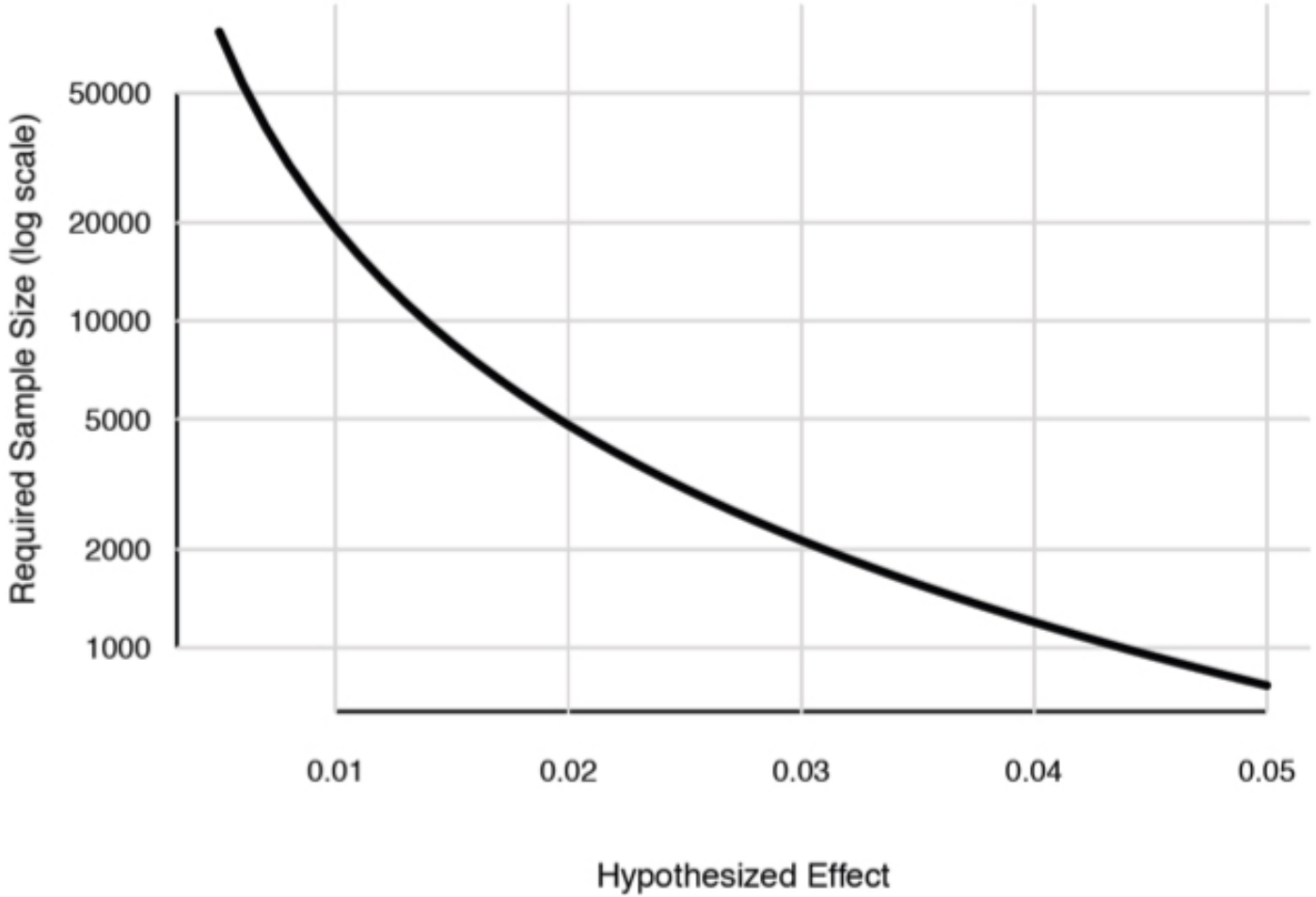
alternative = one.sided

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}
```

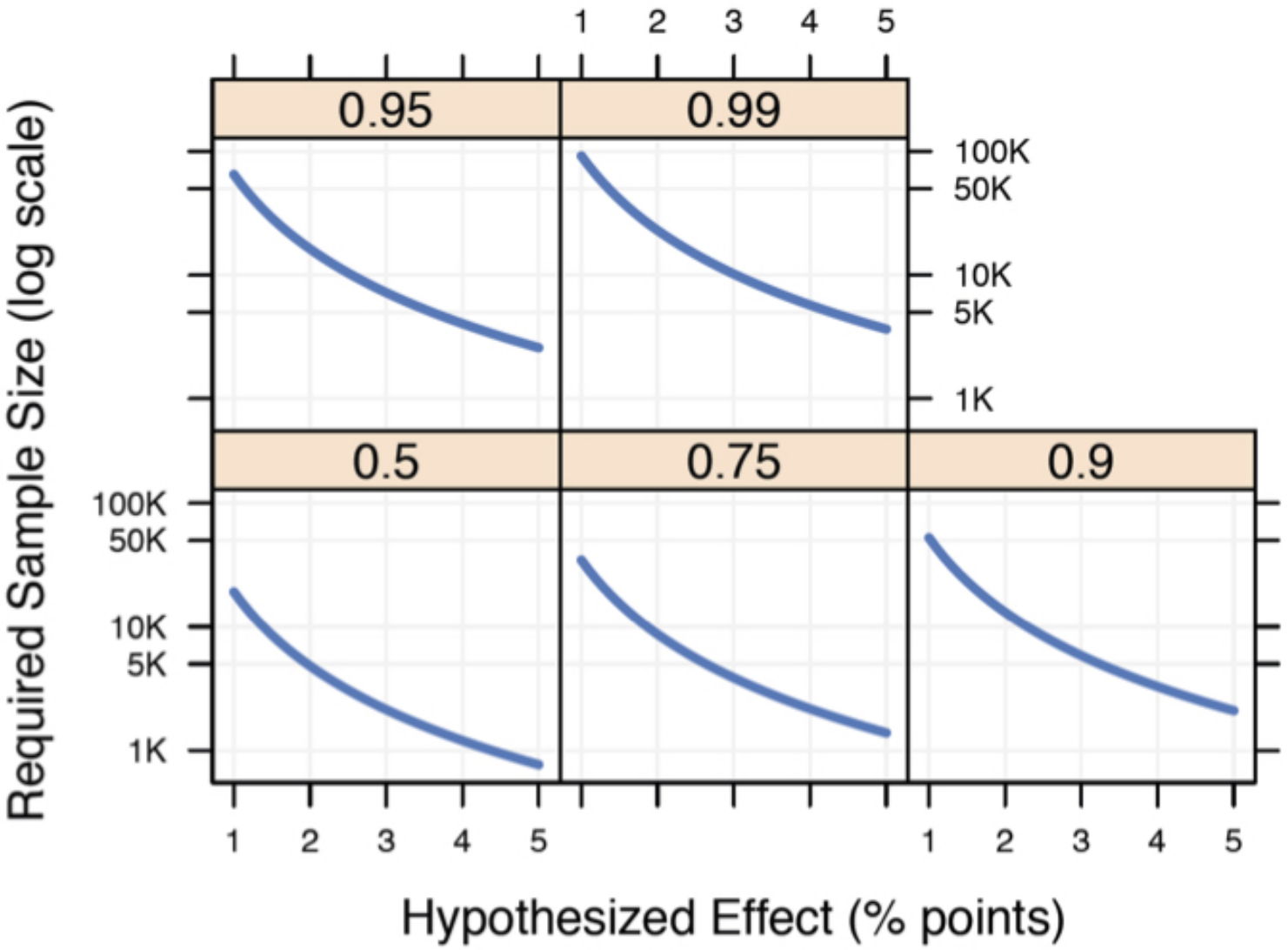
Power Curves



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?