Chapter 5: Optimal Experimental Design

A Course in Experimental Economics

John List

Key Ideas

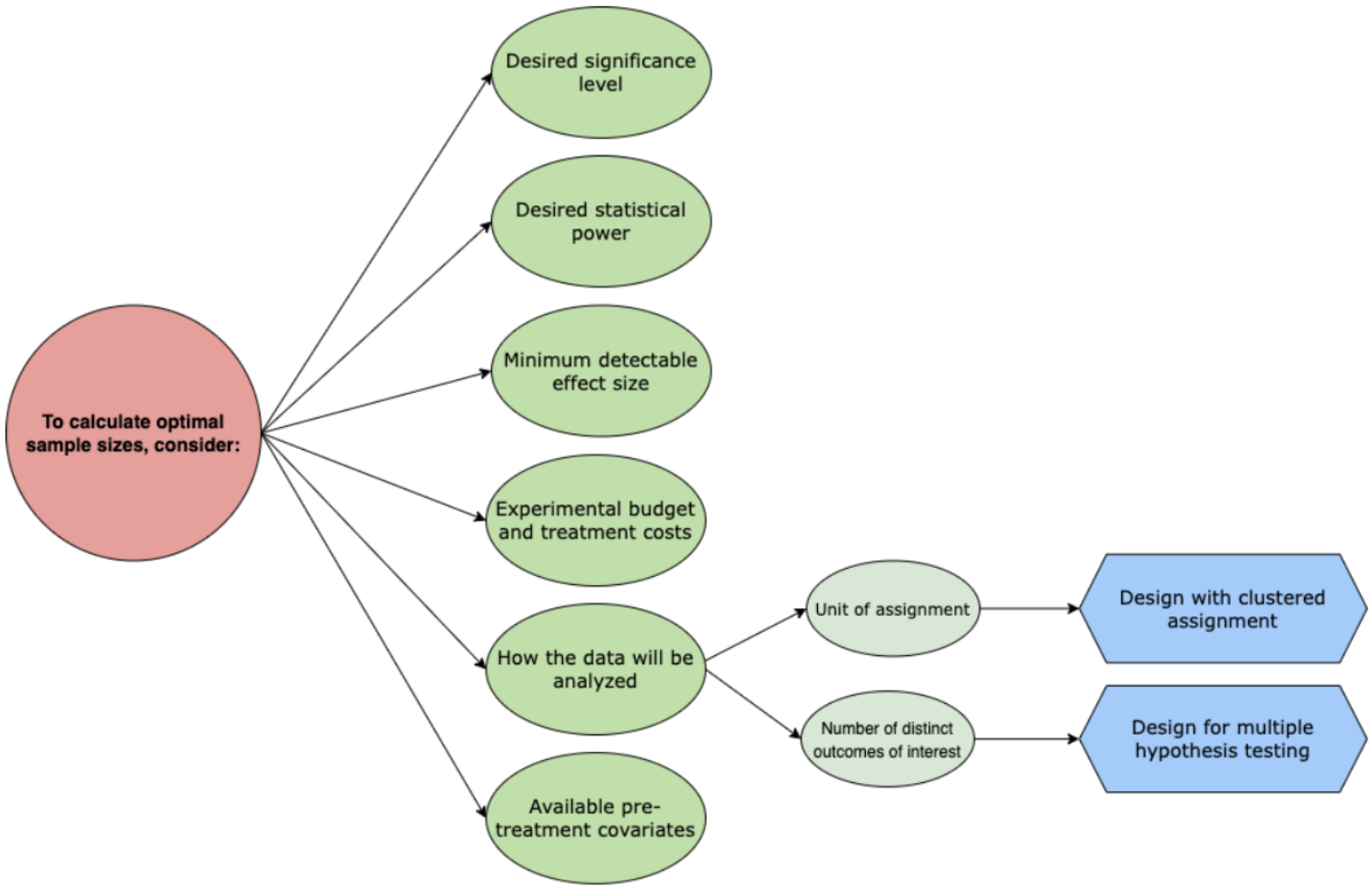
1. Statistical power provides an indication of the likelihood of detecting a true effect when there is one

2. There are several approaches to improving statistical power of an experiment, from initial design to final data analysis

3. Sample size, variation in treatment, cost per observation, treatment response variability, covariates collected pre-treatment, and the modeling of data all play a key role in optimal design

4. When the unit of randomization is different from the unit of observation, clustering is an important consideration

Mind's Eye: Optimal Experimental Design



This chapter delves into optimal experimental design, emphasizing the calculation of optimal sample sizes and the creation of efficient experiments. Key considerations for determining optimal sample sizes include the desired significance level and statistical power, the minimum detectable effect size, budget and costs, the data analysis plan, and available pretreatment covariates. When the unit of randomization differs from the unit of observation, clustering becomes necessary. Additionally, multiple-testing considerations should be integrated into the design

Introduction

- The assignment mechanism plays a fundamental role in determining which data points are missing and which are observed
- Efficient experimental design begins with an understanding of statistical power, as we discussed in chapter 4
- Alongside power, significance level, minimum detectable effect size, available pre-treatment covariates, the experimental budget, and how the data will be analyzed play a central role
- All of these (besides budget) are determined by the experimenter making assignment mechanism and deisgn choices
- We will find that minor changes to any of these features can impact statistical power in unintuitive ways not well reflected in simple rule-of-thumb approaches

Introduction

- This chapter provides a nuts-and-bolts discussion of carrying out this task in an optimal manner
- To make matters concrete, I use three running examples
- This chapter showcases how both benefits (e.g., power to make stronger inference) and costs (e.g., resource outlays to generate data) affect optimal design choices
- While nonoptimal designs will not typically affect the internal validity of causal estimates, they do prevent the analyst from making the sharpest inference, leading to lost opportunities for optimal knowledge creation

Three Running Examples

- The first study is a natural field experiment (NFE) of Holladay, LaRiviere, Novgorodsky, and Price (2019). The authors partnered with an energy utility in a midisze southern US state to improve the energy efficiency of households
  - Households in the various treatment groups were sent letters informing them of the in-home energy audits (IHEA) program. Beyond this information, the letters randomized social norm (nudge) content and whether a subsidy for the IHEA was offered
  - In contrast, the control group received no letter, information, or subsidy. The key outcome variable the authors explored is a binary one: whether or not the household would take up the IHEA program
  - From their NFE, the authors found that nudges could be effective (relative to financial subsidies) at inducing IHEAs, but broadly speaking, neither approach appreciably changed audit rates
  - This study plays an important role in this chapter by highlighting an example of the experimental design problem where the outcome variable (i.e., audits) is binary

Three Running Examples

- Bessone, Rao, Schilbach, Schofield, and Toma (2021) took to the task of learning the causal impact of increased quality sleep hours on a range of outcome variables
  - In their FFE, the authors randomized a set of treatments that have a similar flavor to those utilized in Holladay et al. (2019)
  - Along the main nighttime sleep dimension, the authors randomized some people to receive information (strategies to improve sleep and potential benefits of sleep) as well as access to additional sleep aid device
  - Furthermore, among these treated households, some were randomly offered incentives to increase their minutes asleep relative to a baseline measurement
  - The authors found that while nighttime treatments increased time spent in bed (suggesting individuals can change their schedules), hours asleep did not change with information, tools, or incentives
  - In our discussion of designs with continuous outcomes, we will refer to this suite of Bessone et al. (2021) results

Three Running Examples

- Our third example is the Oregon Health Insurance Experiment (OHIE), a framed field experiment in which Oregon expanded Medicaid for low-income uninsured adults in 2008 through a random-lottery selection. This design allowed researchers to study the causal effects of Medicaid under strong internal validity
  - Finkelstein, Taubman, Allen, Wright, and Baicker's (2016) analysis highlights the surprising result that Medicaid coverage increased emergency department (ED) visits by 40% in the 15 months after subjects won the lottery
  - The authors supplemented their experimental data with administrative data on ED visits from 2007 to 2010 to understand the effects of Medicaid on ED in the two-year period after the lottery
  - Specifically, the administrative data allowed the authors to understand whether the increase in ED use was due to a "pent-up demand" that would dissipate over time
  - They found the effect persisted at least through the first two years of coverage

Table 5.1: Essential Elements of Our Three Running Examples

	Type	Units	Control Condition	Treatment Conditions	Outcome
Holladay, LaRiviere, Novgorodsky, and Price (2019)	NFE	Households in a midsize southern US state	No letter, nudge, or subsidy around in-home energy audits (IHEAs)	Nudges and/or subsidies for IHEAs	Binary: completion of IHEA
Bessone, Rao, Schilbach, Schofield, and Toma (2021)	FFE	Temporary data entry workers in Chennai, India, aged 25 to 55, who had numerical literacy, had relatively low earnings, and had willingness and ability to have sleep monitored	Sleep monitoring	Sleep monitoring together with access to naps, sleep encouragement, and sleep subsidies	Many continuous, including sleep amount, sleep quality, worker productivity, earnings, and memory
Oregon Health Insurance Experiment (OHIE; Finkelstein, Taubman, Allen, Wright, and Baicker [2016])	FFE	Uninsured low-income adults	No Medicaid coverage	Medicaid coverage	Many, with the focus on emergency room visits

Basic Principles of Statistical Power

- Our framework thus far reveals that a key piece of identification and measurement of relevant counterfactuals is appropriate randomization or establishing statistical independence

- However, optimal design involves much more than simply satisfying assumption 3.4 from chapter 3

- The chore of optimal data generation brings several new dimensions that are not present in explorations of naturally occurring data, including choosing how many units to randomize, which treatments to randomize them into, the nature and form of the treatments themselves, and what outcomes to measure

- In this chapter, we illustrate several trade-offs experimenters face, using the general economic framework of marginal benefits and marginal costs to guide the assignment problem

Basic Principles of Statistical Power

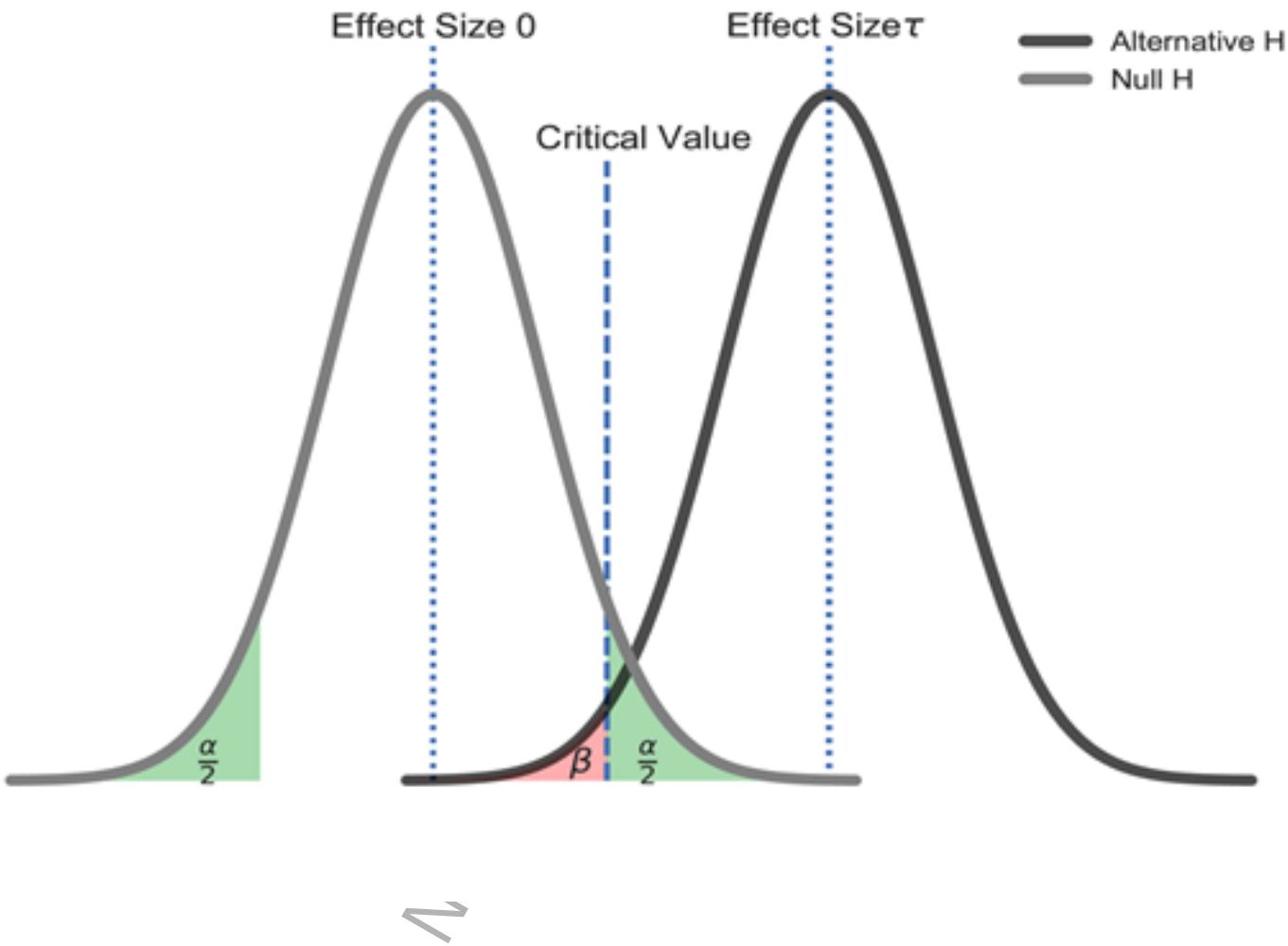
- To calculate optimal sample sizes, an experimenter first considers three key elements that we discussed in chapter 4:

1. **Significance level**: also known as the probability of Type 1 error, is the probability of falsely rejecting the null hypothesis (i.e., a "false positive")

2. **Minimum detectable effect size**: the smallest magnitude of the treatment effect that the analyst desires to detect

3. **Statistical power**: the probability of detecting an effect if there is an effect to be detected; or stated differently, the probability that we can reject the null hypothesis of no effect, given a specific significance and effect size

Figure 5.1: An Ocular View of Type 1 and Type 2 Errors and Power



The figure visually captures the two errors – Type 1 and Type 2 -— that a researcher is concerned with in the context of hypothesis testing. The darker gray areas correspond to the probability of a Type 1 error (α), given the lighter gray distribution of ATEs under a true null hypothesis of no effect ($\tau = 0$). Conversely, the lighter gray area represents the probability of a Type 2 error (β), with area ($1 - \beta$) being the power of the test, when we have a strictly positive effect ($\tau = \delta > 0$)

The Case of a Binary Treatment with Continuous Outcomes

- One common experimental design is to use a binary treatment and explore how its application affects a continuous outcome
- These types of experiments consider the outcome Y_i of the subject i under treatment $D = 1$ and control $D = 0$ that can be modeled as a function of:
  - Covariates, $\mathbf{X}_i$
  - An unobserved person-specific effect, α_i
  - An average treatment effect, $\overline{\tau}$
  - A person-specific treatment effect τ_i, where $E(\tau_i) = 0$
  - An error term, ε_i, which is assumed to be i.i.d.

$$Y_{iD} = \alpha_i + X_i\beta + \overline{\tau}D + \tau_i D + \varepsilon_i \tag{5.1}$$

- As randomization ensures that the assignment to treatment is independent of other sources of variation, the ATE estimate is unbiased; thus

$$\begin{aligned}\hat{\tau} &\equiv \mathbb{E}(Y_{i1}|D = 1) - \mathbb{E}(Y_{i0}|D = 0) \\ &= \mathbb{E}(Y_{i1}) - \mathbb{E}(Y_{i0}) = \mathbb{E}(Y_{i1} - Y_{i0}) = \overline{\tau}\end{aligned}$$

The Case of a Binary Treatment with Continuous Outcomes

- Note that in this framework the variance of the ATE is given by

$$var(\hat{\tau}) = \frac{\sigma^2}{N} = \frac{var(\varepsilon)}{N * var(D)} \tag{5.2}$$

- The variance of the treatment effect is increasing in the variance of the unobserved component, $var(\varepsilon)$, and is inversely related to the sample size, N, and the variance of the treatment propensity, $var(D)$

- If there is only one treatment, then $var(D) = p(1-p)$, with p as the proportion under treatment

- A single treatment results in (conditional) outcomes Y_{i0} if $D = 0$ where $Y_{i0}|X_i \sim N(\mu_0, \sigma_0^2)$ and Y_{i1} if $D = 1$ where $Y_{i1}|X_i \sim N(\mu_1, \sigma_1^2)$

The Case of a Binary Treatment with Continuous Outcomes

- Thus, $\sigma_1^2 - \sigma_0^2 = var(\tau|X)$ only if the variance of the individual-specific treatment effect equals zero, that is, only if the treatment effect is homogeneous across units, $var(\tau|X) = 0$. In this case, the variances of the treatment and control group are equivalent

- Since the experiment has not yet been conducted, the experimenter must form beliefs about the variance of outcomes, which may come from theory, prior evidence, or a pilot

- The experimenter must also make a choice about the minimum difference between the mean of the control and the treatment outcomes that the experiment is able to detect

- The minimum detectable effect (MDE) is the minimum ATE $\overline{\tau}$ that the analyst can detect at a given significance level and power

The Case of a Binary Treatment with Continuous Outcomes

- Beyond the *MDE*, to calculate the optimal sample size, the experimenter must also specify a null hypothesis and an alternative hypothesis
- In the typical case, the null hypothesis is that there is no treatment effect

$$H_0 : \mu_1 = \mu_0$$

- The alternative hypothesis is that the effect size differs from zero, or takes one specific value

$$H_1 : \mu_1 \neq \mu_0$$

- The optimal sample size considers the trade-off between Type 1 and Type 2 error, as shown in figure 5.1

Putting It All Together to Create an Optimal Design

- If our data generation satisfies assumption 3.4, then sample means $\overline{Y_1} - \overline{Y_0}$ (not yet observed) must satisfy two conditions:

  1. A probability α of committing a Type 1 error in a two-sided test is given by

$$\frac{\overline{Y_1} - \overline{Y_0}}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} = t_{\frac{\alpha}{2}} \Rightarrow \overline{Y_1} - \overline{Y_0} = t_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \tag{5.3}$$

     where n_D and σ_D^2 are the sample size and the conditional variance of the outcome for $D = \{0, 1\}$, respectively

  2. A probability β of committing a Type 2 error in a one-sided test is given by

$$\frac{\overline{Y_1} - \overline{Y_0} - MDE}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} = -t_\beta \Rightarrow \overline{Y_1} - \overline{Y_0} = MDE - t_\beta \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \tag{5.4}$$

Putting It All Together to Create an Optimal Design

- From equations 5.3 and 5.4, it follows that the *MDE* is given by

$$MDE = (t_{\frac{\alpha}{2}} + t_{\beta})\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \tag{5.5}$$

- These parameters determine the smallest value of $|\overline{\tau}| > 0$ for which the experiment will correctly reject the null hypothesis with probability $1 - \beta$ at significance level α
- If $\sigma_1^2 = \sigma_0^2 = \sigma^2$ and we define $N = n_0 + n_1$ as the total number of units in treatment and control, then
  - $\varphi = \frac{n_1}{N}$ is the proportion of units allocated to the treatment group
  - $(1 - \varphi) = \frac{n_0}{N}$ is the proportion of units allocated to the control group
- The *MDE* can therefore be written as

$$MDE = (t_{\frac{\alpha}{2}} + t_{\beta})\sqrt{\frac{1}{\varphi(1-\varphi)}\frac{\sigma^2}{N}} \tag{5.6}$$

where $\frac{1}{\varphi(1-\varphi)}\frac{\sigma^2}{N}$ is the exact sample variance of the treatment effect estimator, $var(\hat{\tau})$

Putting It All Together to Create an Optimal Design

- The smallest sample sizes that solve the equality in equation 5.6 satisfy $n_0 = n_1 = n$ are given by

$$n_0^* = n_1^* = n^* = 2(t_{\frac{\alpha}{2}} + t_\beta)^2 \left(\frac{\sigma}{MDE} \right)^2 \tag{5.7}$$

- If the variances of the outcomes are not equal, this solution becomes:

$$N^* = \left(\frac{t_{\frac{\alpha}{2}} + t_\beta}{MDE} \right)^2 \left(\frac{\sigma_0^2}{\pi_0^*} + \frac{\sigma_1^2}{\pi_1^*} \right) \tag{5.8}$$

with:

- $\pi_0^* = \frac{\sigma_0}{\sigma_0 + \sigma_1}$
- $\pi_1^* = \frac{\sigma_1}{\sigma_0 + \sigma_1}$
- $N = n_0 + n_1$
- $\pi_0 + \pi_1 = 1$
- $\pi_0 = \frac{n_0}{n_0 + n_1}$

Putting It All Together to Create an Optimal Design

- If the sample sizes are large enough, then the normal distribution is a good approximation for the t-distribution, and the above equations are closed-form solutions for the optimal sample sizes
- The optimal sample size increases proportionally with the variance of outcomes, increases nonlinearly with the significance level and the power of the test, and decreases proportionally with the square of the *MDE*
- Beyond yielding design insights to show the intuition behind the marginal benefits of varying unit allocation, equation 5.8 can be used to compute the sample sizes necessary to detect different treatment effect sizes
  - For example, assuming equal outcome variances across treatment and control, and setting the significance level $\alpha = 0.05$ and the power $1 - \beta = 0.8$ then from the standard normal tables we have $t_{\frac{\alpha}{2}} = 1.96$ and $t_{\beta} = 0.84$, one can generate insights on unit allocation across treatment cells
  - Table 5.2 provides several examples

Table 5.2: Simple Rules of Thumb for Sample Size by Minimum Detectable Effect

MDE (in Standard Deviation Units)	n^* (Per Cell)
1/50	39,244.4 $\approx$ 39,245
1/20	6,279.1 $\approx$ 6,280
1/10	1,569.8 $\approx$ 1,570
1/5	392.4 $\approx$ 393
1/3	141.3 $\approx$ 142
1/2	62.8 $\approx$ 63
1	15.7 $\approx$ 16

The table reports several possible *MDEs* desired by the experimenter and their corresponding number of necessary observations for both the treatment and the control group, n^*. As the *MDE* decreases, the required sample size increases. For example, to detect a $1/50^{th}$ standard deviation change, 39,245 observations are required in treatment and in control, whereas to detect 1 standard deviation, only 16 observations are required. The required sample size is rounded up in each example to account for the non-divisibility of experimental units

Putting It All Together to Create an Optimal Design

- How practical is having this continuous outcome – binary treatment – in mind for mental back-of-the-envelope calculations?
- Let us consider the study on sleep deprivation in Bessone et al. (2021) and focus on one of their key outcomes – hours of night sleep
- As this is a continuous outcome, our approach is useful
- As the incentives increase, the marginal benefit of night sleep relative to other uses of time increases, thus we expect to see night sleep increase

Putting It All Together to Create an Optimal Design

- From table 2 in Bessone et al. (2021), we observe that in the control group, mean time asleep is 5.61 hours with a standard deviation of 1.2
- These are two crucial inputs to the back-of-the-envelope design
- In fact, if we use the research standard of level $\alpha = 0.05$ and power $1 - \beta = 0.8$, then all that remains is an assessment of a relevant *MDE*
- Alternatively, we can do things in reverse and ask: For a sample size of 150 units in treatment and 150 in control, what is the *MDE*?
- If the (known) control standard deviation is common to both groups, then these sample sizes allow for an increase in time asleep of approximately 20 minutes to be detected reliably, given the researchers' desired significance and statistical power

The Case of a Binary Treatment with Binary Outcomes

- The formulas derived in section 5.3.1 for the continuous outcome case can be adapted easily for other common experimental designs, including continuous treatment, binary outcomes, and cluster designs
- As in the continuous outcomes case, we assume the normal approximation can be used to the given approximation
- In the case of binary outcomes, such as Holladay et al. (2019), the variance depends on the mean
- In equation 5.3, as the null hypothesis is true, the treatment and control groups have equal means and therefore equal variances and equal optimal sample sizes
- In equation 5.4, as the alternative hypothesis is true, the treatment and control groups have different means and therefore different variances

The Case of a Binary Treatment with Binary Outcomes

- Using the normal approximation to the binomial distribution for binary data, the variance is equal to $p(1-p)$ where p is the mean of the outcome variable

- When we set the null hypothesis of equal proportions under treatment and control, $H_0 : p_0 = p_1$, and when we use equation 5.3 for the null hypothesis of equal means (and thus variances) for sample sizes and significant test and use equation 5.4 for the alternative hypothesis $H_1 : p_0 \neq p_1$ of different means (and thus variances) for power test, then the optimal sample sizes become

$$n_0^* = n_1^* = n^* = \left(t_{\frac{\alpha}{2}} \sqrt{2\overline{p}(1-\overline{p})} + t_{\beta} \sqrt{p_0(1-p_0) + p_1(1-p_1)} \right)^2 MDE^{-2} \tag{5.9}$$

where $\overline{p} = \frac{p_0 + p_1}{2}$

The Case of a Binary Treatment with Binary Outcomes

- Since the variance $p(1-p)$ is maximized for $p = 0.5$, optimal sample sizes increase as $\overline{p}$ approaches 0.5 (i.e., the sample sizes decrease in $|\overline{p} - 0.5|$

- Similarly, if the null hypothesis is of the form $H_0 : p_1 = kp_0$ with $k > 0$, then the sample size arrangement is dictated by k in the identical manner as in the continuous case using equation 5.8:
  - The closer p_1 is to 0.5 relative to p_0, the larger the proportion of the total sample size that should be allocated to p_1

- Unlike in the continuous outcome case, the relationship between the mean and variance in binary outcomes means we do not have a "common variance" case

Varying Treatment Levels with Continuous Outcomes

- Many firms need information on the nature of the demand curve they face to lend insights into a key problem they commonly face: optimal pricing
- To uncover this, the experimenter might experimentally vary price for a specific route over a range of prices
- Let us return to the empirical specification in section 5.1, but assume that $\tau_i = 0$ for all i, thus treatment and control outcomes have the same variance

$$Y_i = X_i\beta + \overline{\tau} D_i + \varepsilon_i$$

- Let Y_i be measurable in continuous units and the treatment variable is over the range $[0, \overline{D}]$
- Recall that $var(\hat{\tau}) = \frac{var(\varepsilon)}{N * var(D)}$
- Therefore, in this case, to increase precision we can:
  1. Decrease the variance of the unobserved component $var(\varepsilon)$
  2. Increase the sample size N
  3. Increase the variance of the treatment $var(D)$

Varying Treatment Levels with Continuous Outcomes

- The most dominant approach in the literature is to increase the sample size, but this is also potentially the costliest approach
- Yet, as our equations make clear, increasing the variation in D can have a considerable effect on precision
- If the effect of the treatment is linear, then the variance of D is maximized by placing half of the sample in treatment cell $D = 0$ and half of the sample in treatment cell $D = \overline{D}$
  - This design allocation maximizes the variance of the treatment variable and therefore minimizes the standard error of the treatment effect estimate
  - Hence, if a linear treatment effect is to be identified, the optimal sample design is to place half of the sample at each of the extremes of the range of potential treatment intensities
- The overall sample size can then be calculated using equation 5.7 obtaining $\left(\frac{\sigma}{MDE}\right)^2$

Varying Treatment Levels with Continuous Outcomes

- If theory or hypothesis testing calls for nonlinearities, however, or if the experimenter believes that the intensity of treatment D has a nonlinear effect on the outcome variable, allocating the entire sample to the extremes is inappropriate since curvature cannot be detected using such an approach

- In general, identification in such instances requires that the number of treatment cells used be equal to the highest polynomial order plus one

- For example, if there are priors of a quadratic relationship, then three treatment cells should be chosen in a feasible range; and two of those three cells should be placed at the extremes and one at the midpoint of the treatment variable range: $\mathrm{D} = \{0, \frac{\overline{D}}{2}, \overline{D}\}$

Varying Treatment Levels with Continuous Outcomes

- Intuitively, the test for a quadratic effect then compares the mean of the outcomes at the extremes to the mean of the outcome at the midpoint

- And, as before, the variance is maximized when equal proportions are allocated to the these two means: the midpoint and the extremes (and the observations at the extremes are also equally divided)

- For this reason, the optimal proportion in each of these treatment cells is $\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\}$

- If both a linear and a quadratic effect are included, then the problem becomes considerably more complicated, with the solution being a weighted average of the linear and quadratic optimal allocations

Table 5.3: Simple Rules of Thumb for Sample Size for Polynomial Outcome Models

Highest and Only Polynomial Order	Number of Treatment Cells	Sample Allocation
1	2	$\left\{\frac{1}{2}, \frac{1}{2}\right\}$
2	3	$\left\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right\}$
3	4	$\left\{\frac{1}{6}, \frac{1}{3}, \frac{1}{3}, \frac{1}{6}\right\}$
$\vdots$	$\vdots$	$\vdots$
10	11	$\left\{\frac{1}{20}, \frac{1}{10}, \cdots, \frac{1}{10}, \frac{1}{20}\right\}$

The table reports the optimal sample allocations for various powers of polynomial responses. Generally, the number of treatment cells is equal to the highest polynomial order plus one. The sample allocation for a quadratic relationship, for example, has three treatment cells: two cells at the extremes and one cell at the midpoint. If multiple effects are included, then the experimenter must take the weighted average of the optimal allocations for each effect

Expanding the Tool Kit

- Thus far, we have learned the nuts and bolts of optimal experimental design and a useful framework that nearly every advanced design will build from

- In this section, we build on that model using three key departures that mirror the design considerations necessary in modern experimentation in economics

    1. We begin by considering the case of unequal cost to generate data across experimental cells

    2. Our second departure relates to cases when clustering is necessary; for example, when the unit of randomization is at a level different from that of the unit of observation

    3. Our final departure studies the role that corrections for MHT have on optimal experimental design

Heterogeneity in Participant Costs

- In our previous sections, we make the critical assumption that the cost of data generation is homogeneous across treatment and control groups
- Without this assumption, determining the optimal sample sizes turns out a bit more complicated, but out intuition from above remains
- Assume that the cost of applying control c_0 and treatment c_1 must be considered
- The total experiment cost is given by $M = c_0 n_0 + c_1 n_1$

Heterogeneity in Participant Costs

- The maximum power is achieved by finding n_0 and n_1 that maximize the MDE size as given by equation 5.5 (reproduced below for convenience):

$$MDE = (t_{\frac{\alpha}{2}} + t_{\beta})\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}$$

subject to the total cost constraint

- Solving for n_0 and n_1 when taking into account the cost per observation, the optimal sample sizes are now given by

$$\frac{n_1^*}{n_0^*} = \frac{\pi_1^*}{\pi_0^*} = \sqrt{\frac{c_0}{c_1}}\frac{\sigma_1}{\sigma_0}$$

Heterogeneity in Participant Costs

- As before when we assumed isomorphic costs per unit across treatment and control, the optimal sample sizes are:
  - again proportional to the standard deviations of the respective outcomes
  - inversely proportional to the square root of the relative sampling cost
- In general, this implies that if sampling costs for the control group are smaller than for the treatment group, then the control group should contain more units than the treatment group, ceteris paribus
- Since the optimal sample sizes are proportional to the square root of the cost of sampling, the importance of this result grows in import with cost heterogeneity across treatment and control

Clustered Experimental Designs

- Built into our description of sampling frameworks thus far is one key assumption: the unit of observation coincides with the unit of randomization

- In many practical settings of interest, the unit of observation will be finer than the unit of randomization. Social science researchers refer to such experimental designs as **cluster-randomized experiments** where the cluster g (which is the unit at which researchers administer treatment) is an aggregation of N_g observed units

- Conceptually, such cases require a unique sampling framework for at least two reasons:

    1. There is no longer a single ATE as, in theory, we can now define potential outcomes specific to each unit i in cluster g

    2. In principle, one can consider multistage sampling schemes that draw i.i.d. samples of clusters followed by random or nonrandom samples of units within each cluster

Clustered Experimental Designs

- To highlight these two elements, we start with a slight modification to our previous equations

$$Y_{ig} = Y_{ig}(1)D_g + Y_{ig}(0)(1 - D_g) \tag{5.10}$$

where Y_{ig} is the outcome for unit i in cluster g and D_g is a binary treatment indicator, but now applying to all units in cluster g (hence we drop the i subscript)

- We consider the following regression specification where each unit is also a member of a cluster g, and outcomes for $D = \{0, 1\}$ are given by

$$Y_{igD} = \alpha + \bar{\tau} D_g + \nu_g + \varepsilon_{ig} \tag{5.11}$$

with ε_{ig} the individual specific i.i.d. error term and ν_g a cluster specific i.i.d. error term. For simplicity X_i, α_i and τ_i are ignored

- In this scenario, suppose that sampling is by cluster, where each cluster is of size m for both control and treatment groups

Clustered Experimental Designs

- Assuming equal costs and equal variances (and therefore equal sample sizes) accross treatment and control groups, the optimal sample sizes in cluster designs can be calculated as follows

$$n_0^* = n_1^* = n^* = 2(t_{\frac{\alpha}{2}} + t_\beta)^2 \left(\frac{\sigma}{MDE}\right)^2 (1 + (m-1)\rho) \qquad (5.12)$$

with $2(k-1)$ degrees of freedom (assuming no other covariates), where $k = \frac{n}{m}$ is the number of clusters, σ^2 is the common variance, and $\rho = \frac{var(\nu_j)}{var(\nu_j)+var(\varepsilon_{ij})}$ is the coefficient of intracluster correlation

- Equation 5.12 shows that the total sample size needed in a cluster design increases near proportionally with both the size of each cluster and the intracluster correlation
- Also, as the degrees of freedom in a cluster design are far smaller, the necessary sample size further increases
- Hence, in the presence of intracluster correlation, $\rho > 0$, it is in many cases optimal to randomize over many, small clusters to maximize the efficiency of the experiment

Clustered Experimental Designs

- The total cost of collecting the data in a cluster design is given by $2(c_m m + c_k)k = M$ where c_m is the cost of each unit and c_k is the fixed cost per cluster.

- Maximizing the *MDE* from equation 5.12 subject to this budget constraint leads to an expression for the optimal size of each cluster

$$m^* = \sqrt{\frac{(1-\rho)}{\rho}}\sqrt{\frac{c_k}{c_m}} \tag{5.13}$$

- The expression for the optimal size indicates that on a limited budget, the experimenter should first work out how many units to sample from each cluster and then sample as many clusters as possible

- The optimal number of clusters k^* can be found by substituting m^* into equation 5.12, recalling that $n = mk$

Table 5.5: Simple Rules of Thumb for Sample Size for Clustered Units

This table summarizes how the optimal choice of overall sample size, clusters, and units per cluster changes depending on the intracluster correlation coefficient, ρ

ρ	m	n^*	k^*
0	10	62.79 $\approx$ 63	6.28 $\approx$ 7
0	30	62.79 $\approx$ 63	2.09 $\approx$ 3
0.25	10	204.07 $\approx$ 205	20.4 $\approx$ 21
0.25	30	518.02 $\approx$ 519	17.26 $\approx$ 18
0.5	10	345.35 $\approx$ 346	34.53 $\approx$ 35
0.5	30	973.26 $\approx$ 974	32.44 $\approx$ 33
0.75	10	486.63 $\approx$ 487	48.66 $\approx$ 49
0.75	30	1428.50 $\approx$ 1429	47.61 $\approx$ 48
1	10	627.91 $\approx$ 628	62.79 $\approx$ 63
1	30	1883.73 $\approx$ 1884	62.79 $\approx$ 63

*T*he table reports several intracluster correlation coefficient values (ρ) and the corresponding optimal number of units in each cluster (m), the optimal overall sample size (n^*), as well as the optimal number of clusters (k^*). The table demonstrates how the optimal number of clusters increases as the intracluster correlation increases.

Optimal Design with Multiple Hypothesis Adjustment

- In chapter 4, we discussed appropriate methods to analyze experimental data and control error rates
- One key consideration was adjusting for multiple hypothesis testing (MHT)
- The intuition is parallel to our discussion concerning clusters: within optimal design, we need to adjust the relevant sample sizes to take into account any multiple testing
- Adjusting for multiple testing creates an important tension for experimentalists to control false discovery, as maintaining proper Type 1 error rates increases as the number of hypotheses under consideration grows
- Thus, it creates analogous issues for statistical power and design

Optimal Design with Multiple Hypothesis Adjustment

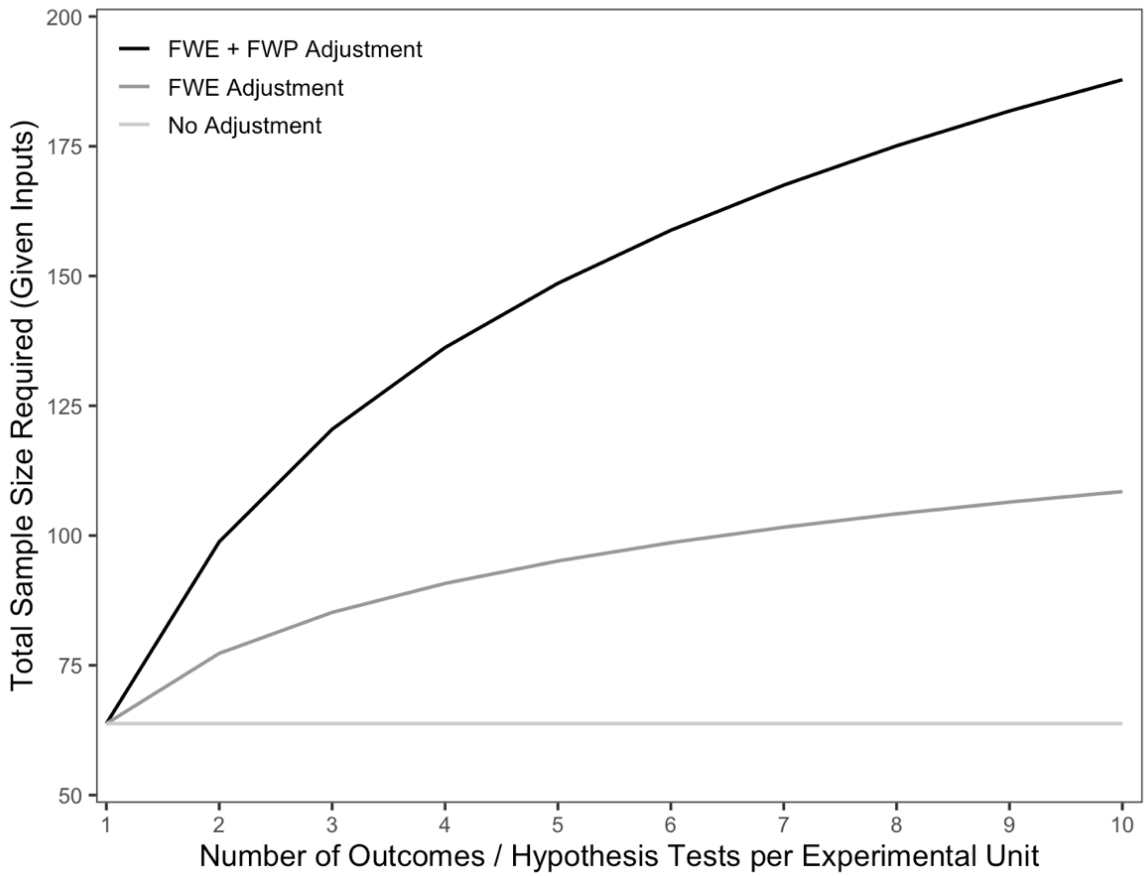
- To provide intuition into how MHT corrections affect optimal design, it is useful to recall that in the context of MHT, the notion of a Type 1 error rate is generalized to the incorrect rejection of *any* of a family of hypotheses (i.e., the family-wise error rate, or FWER)
- Similarly, we can have an analogous extention of our notion of statistical power, called the family-wise power (FWP)
- To illustrate this concept, we revisit equation 5.6
- We consider a setting where a researcher observes multiple outcomes for everyone, such as Bessone et al. (2021), to illustrate the MHT problem
- The simplest extension of the basic design approach in equation 5.6 to the multiple-testing setting that allows us to retain the closed-form solution for intuition is to treat the tests across outcomes as independent

Optimal Design with Multiple Hypothesis Adjustment

- Analogously, the correct rejection of false nulls factors into the question of statistical power
- This independence assumption yields simple Bonferroni-type corrections to both Type 1 and Type 2 error rates
- Meaning that we can simply divide the significance level by the number of tests, and similarly, exponentiate the statistical power by the reciprocal of the number of tests
- Otherwise, the expression follows in equation 5.6, meaning that we simply must adjust the critical values used in the calculation

Figure 5.2: MHT and Statistical Power

We illustrate the consequences that multiple tests have for sample size arrangement in the context of a specialized version of our earlier example, using an *MDE* of 0.5 standard deviations



The graph reports the required sample size on the y-axis versus the number of distinct hypothesis tests (number of outcomes) on the x-axis using an MDE of 0.5 standard deviations. The light gray line represents no adjustment, the darker gray line introduces the Bonferroni correction for the family-wise error rate (FWER), and the black line corrects for both the family-wise error rate (FWER) and the family-wise power (FWP). The darker gray line shows that the FWER adjustment increases the required sample size for any given number of outcomes greater than one, and the black line indicates that the FWER + FWP adjustment increases sample size requirements even further.

Less Considered Design Choices to Enhance Statistical Power

- Thus far, we have discussed how power is critically related to N, outcome variances, a richer treatment space, MHT, cluster analysis, and the relative cost of each observation
- Yet, optimal design dimensions can potentially run even deeper than these traditional considerations
- In this section we discuss four further design choices that can have a disproportionate effect on power but are often not utilized

Including Covariates in the Estimation Model

- The use of covariates in the analysis of data from randomized experiments is a contentious issue dating back at least to their inclusion in regression models for treatment effects in Fisher (1935)
- The rationale for their inclusion in a homogeneous effect model is quite compelling:
  - In a randomized experiment with homogeneous ATE, controlling for the pre-treatment values of covariates that are predictive of the outcome does not affect the expected value of the estimator of τ but can reduce its variance, yielding gains in statistical power
- In cross-sectional settings, these covariates tend to be factors predetermined at the time of random assignment
- For outcomes with strong dependence over time, such pretreatment outcome measures may be particularly attractive covariates to include to reduce residual variation in the outcome not due to treatment
- More generally, a new body of work is developing tools to allow researchers to learn what the important, predictive covariates might be using pilot experiments together with administrative data

Including Covariates in the Estimation Model

- As one simple illustrative example, we follow the machine-learning approach of List, Muir, and Sun (2022), who solve for the efficient regression adjustment in a large class of adjustments
- Using this approach, we revisit the data from OHIE to explore the gains of using covariates
- When doing the analysis, we find that across various specifications, the machine-learning approach making use of four simple covariates improves standard errors by about 1.5%
- While these gains are modest fixing sample size, they imply that for a similar level of statistical power, researchers could reduce sample sizes by about 3%, with a commensurate reduction in the experimental variable costs
- The exercise also reveals that relevant gains would be much larger with a richer covariate set, such as using pre-treatment *Ys*, which we discuss in chapter 9

Including Covariates in the Estimation Model

- While this discussion makes covariates sound like the mythical free lunch, researchers must take care in cases where their covariates are *not* likely to be predictive of the outcome
- In such scenarios, the covariates are exhausting degrees of freedom (as their contribution to the outcome equation must be estimated) without accounting for any residual variation in the outcome not explained by treatment
- As a result, covariates can limit statistical power in a finite sample sense; with an infinite sample, the distinction between estimated and observed is no longer relevant
- In practice, this issue is substantive only when sample sizes are sufficiently small that a modest change in the degrees of freedom substantially impacts the estimated uncertainty around the estimate of the treatment effect
- As a rule of thumb, the Green Lab standard operating procedures advise that the number of covariates included in the regression model not exceed 1/20 of the overall sample size to avoid the potential finite-sample bias

Including Covariates in the Estimation Model

- In the context of CATEs, the message is broadly similar, although functional form takes on greater import as the treatment effect may vary as a function of the covariates

- In such cases, the analyst should include treatment-by-covariate interactions to ensure both that the regression model recovers the ATE, and that the inclusion of predictive covariates (weakly) improves the precision of the estimate of the CATEs as the sample size increases

Designs to Maximize Compliance

- Given a particular question and a budget constraint, the researcher must choose the elements that maximize power, or the precision of their treatment effect estimates

- This implies determining key aspects such as:
  - The level of randomization (individual or cluster)
  - The optimal sample size
  - The number of treatments
  - Whether the outcome variable(s) is (are) binary or continuous
  - Whether heterogeneity will be explored
  - Which covariates to collect

Designs to Maximize Compliance

- Beyond these dimensions, there may be choices that the researcher makes in the precise nature of the administration of the treatment that may positively or negatively impact statistical power of their eventual test

- A key choice is how to maximize compliance with treatment assignment

- Recall from chapter 3 that full compliance is an identification assumption, and when this assumption is violated, researchers must move to study the assignment mechanism's causal effect rather than the treatment's causal effect (i.e., the effect of Z on Y rather than D on Y)

- While in that case we can make valid causal inference about Z, such causal effects are typically not of primary interest

Designs to Maximize Compliance

- For example, consider the prototypical case of a public program, such as a job training program or an educational program for parents of young children
- To receive treatment, the individual must attend meetings to develop the skills being taught
- Examples of considerations in such experiments that may help to maximize compliance include financial and nonfinancial compensation for attending the meetings, surveying subjects assigned to treatment to identify times of day and days of the week that are the most flexible, providing on-site babysitting (for any young children), reminding subjects of the schedule frequently, backloading incentives, and so on
- Each such design choice plays a crucial role in reducing the costs, or increasing the benefits, of attending the meetings
- We have found that such inducements help to encourage attendance, and, hence, compliance with treatment assignment (to ensure that $Z = D$)

Designs to Maximize Compliance

- A related set of issues arises in certain field experiments where treatment takes the form of information
- Typically, such treatments are administered through modes of communication such as physical mail, electronic mail, and text message
- Any letters, emails, or text messages that are never opened or read are sources of treatment noncompliance
- Here, again, the fundamental idea is the same: Take steps to reduce the costs or increase the benefit of the treatment

Designs to Maximize Compliance

- Finally, note that by their very nature, laboratory experiments and some artefactual and framed field experiments have the advantage of very tight control over the link between treatment assignment and treatment receipt

- However, depending on the precise definition of treatment receipt, such tightly controlled environments may still face these same design choices, say, in an effort to encourage experimental units not only to hear or read the informational treatment, but also to comprehend and internalize it

Designs to Maximize Compliance

- What are the potential costs of not designing with such potential noncompliance in mind? One simple way to answer this question in the context of homogenous treatment effects is to calculate the consequences of noncompliance for the statistical power of the test
- For example, in its most recent set of internal analyses of email marketing campaigns, MailChimp found that only slightly more than 20% of the generic emails sent through its email marketing platform are opened
- The consequences for our simple rule of thumb are straightforward: Our effect sample size becomes only 1/5 the size it would have been had we continued to make inference on D
- Thus, for a 0.5 standard deviation effect, assuming that email open rates are similar between to-be-treated and control groups, whereas before we needed 63 units total, now we require five times as many to arrive at an effective sample size of 63 units

The Nature of the Sample

- The nature of the subject pool itself can affect the power of the experiment
- For example, heterogeneity matters greatly: If units in the $P = 1$ group respond to treatment identically, such as units like silicon chips, or Thing 1 and Thing 2 (the human-like twins from Dr. Seuss's book *The Cat in the Hat*), then the design might be constructed much differently compared to the case with vast heterogeneities
- As homogeneity increases in the $P - 1$ group, variance of the residuals becomes smaller
- One might imagine that homogeneity is lower among a subject pool of university students, or a group of donors to a particular charity, than among a random sample of US citizens or even buyers at Amazon or users of Google, Facebook, or Uber

The Nature of the Sample

- Recall from equation 5.2 that the variance of the ATE is increasing in the variance of the unobserved component $var(\varepsilon)$
- Thus, we learn directly that increased subject homogeneity reduces the variance of the estimator, which increases the power of the statistical test
- This potentially raises a key trade-off between experimental power and external validity
- For example, in a traditional lab experiment the recruited subjects are college students, of similar ages and backgrounds, and probably engaged in similar activities

The Nature of the Sample

- Because they are similar, the residuals are likely to be smaller, and the experiment can obtain greater power with the same sample size compared to a quite heterogeneous sample
- Yet, if the college students have characteristics very different from those of the target population of interest, then even though the experiment can detect smaller effect sizes with a given sample size, it might not provide relevant information about the true parameter of interest should the experimenter have primary interest in Experimental Problem 2 (EP2)
- In this sense, selecting a homogeneous sample can be viewed as a quite useful efficacy test because the researcher is testing whether the theory works in one specific context even though that might not be reflective of the true parameter of interest
- Put another way, sample selection choices may flow directly out of a desire to "give theory its best chance" with a given sample size

Measurement Choices

- In Chapter 2 we discussed that when human subjects are the population of interest, the researcher must be aware of potential observer effects
- A related effect is measurement choice
- Within economics experiments, measurement choices may curb or enhance statistical power to the extent that they exacerbate or mitigate measurement error in the outcome variable
- This measurement error can take the form of random noise, such as the kind that we typically associate with physical measurement implements like thermometers

Measurement Choices

- Here, the guidance is straightforward: Minimize sources of such noise as part of data collection
- Otherwise, we might be observing a noise-augmented version of our outcome of interest, meaning that the precision of our estimated treatment effect decreases
- A general result is that we require a larger sample size for any given desired statistical power and MDE
- Minimizing such random measurement error is a more difficult order in some types of designs than in others

Measurement Choices

- Perhaps more concerning than random measurement error is the systematic measurement error that arises in various experimental settings
- Addressing systematic measurement error is typically a difficult task
- For example, a researcher analyzing the results of an early education program may ask the question: Did the labor supply of parents increase in response to having their child offered free preschool?
- With the ability to link the identities of experimental units to administrative data such as Social Security Administration earnings records or state unemployment insurance data, one can answer this question utilizing data that are relatively free of noise through the combined forces of high incentives for accuracy and internal audits

Measurement Choices

- In the absence of such administrative data, the researcher must rely on self-reported labor supply information
- Such self-reported data may be noise ridden for various reasons that we discuss in chapter S5
- But, perhaps more pernicious is measurement error that is systematic and related to our outcome of interest
- In this case, various notions of social desirability bias may differentially push parents of children receiving free preschool to say that they have returned to work, biasing any estimate of the effect of child preschool on parental labor supply

Factorial Designs

- A key design consideration is the exact number of experimental cells to populate
- Consider Bessone et al. (2021), who as part of their FFE design explore the impact of two factors—information and sleep aid devices—on hours of sleep. Assume that each of the two factors can take on binary values
- A **full factorial design** is a type of experimental design wherein researchers measure responses for all possible combinations of the factor levels
- In this case, with two factors, each with two levels, a full factorial design has four treatment combinations in total. This is often referred to as a 2×2 factorial design

Factorial Designs

- In the case of Bessone et al. (2021), as in many situations in the social sciences, the number of factors can grow rapidly, making decision choice considerably more complicated
- Consider that 8 factors would need 256 treatment cells, and 9 factors demands 512. Populating that many cells becomes infeasible with sufficient power to detect realistic effect sizes
- What is necessary is a **fractional factorial design**. This type of experimental design involves populating only a carefully selected subset (or "fraction") of the potential experimental cells
- The choice of how to whittle down a full factorial design is part art, part science
- In the end, the choice of which factorial design to choose is based on a trade-off between information gained and resource constraints

Factorial Designs

- Fractional factorial designs offer the researcher the chance to implement a more efficient design if there is reason to believe certain, or all, factors are independent (i.e., the baseline levels of other variables are not important)

- Key considerations in making such choices include theory and research needs as well as available budget

- If a fractional factorial design is chosen, then a key consideration is the resolution of how well the design can distinguish between main effects and interactions

Factorial Designs

- Higher-resolution designs are preferred because they minimize confounding between effects
- We strongly prefer no confounding of key main effects, but researchers with other needs might wish to consider alternative resolutions
- For example, it is common in the hard sciences to admit confounding effects with main effects to strike a balance between estimating main effects and capturing interaction estimates
- For economics experiments, we advise using higher resolution than is typically used in the hard sciences

Conclusions

- Optimal design is to an experimentalist what an exquisite recipe is to a fine chef
- Both offer guidelines for creation of something immediate that will ultimately be woven into a creation that delights
- And, when each is complete, it can be replicated exactly as it was carried out the first time
- Sure, this replication might reveal slightly different causal relationships quantitatively, or the dish will never taste exactly the same, but at least the tracks are laid for replication and further trials provide deeper insights into the real beauty of the first creation

Conclusions

- Beyond discussing how to arrange the various pieces correctly in your design, this chapter shows that one key goal of optimal experimental design is to reap the "biggest bang for your buck"

- Economic thinking using marginal analysis goes a long way here, and this chapter highlights several key elements that the data generator must consider

- Optimal experimental design demands experimenters have control over basic statistics, power, key design elements, knowledge of outcome variances, how much each data point costs to generate, available covariates, and how the data will be analyzed after the experiment

Conclusions

- Viewed in this light, the burden of generating data is unlike any other empirical exercise in the sciences
- Optimally overlaying all of these constraints onto the design, while achieving EP1 and EP2, is the ultimate designer's chore
- Generating data is simple. Generating data optimally is not
- The goal of this chapter is to make optimal data generation a bit less abstract and more approachable
- Upon construction of an optimal design, the data generation from the experiment begins, and that starts with the assignment mechanism and choice of randomization approach, which we turn to next