Econ 272 / MGTECON 607: Intermediate Econometrics III – Section 1

April 3, 2025 *

Teaching Assistant: Amar Venugopal
Contact Information: amar.venugopal@stanford.edu
Sections: Thursdays from 4:30 pm - 6:20 pm in STLC 114
Office Hours: Wednesdays from 2 - 4 pm in Econ 149

1 Outline Today

1. General information
2. Problem sets
3. Review of concepts
4. Practice problems

2 Announcements

Don't forget to email me and Guido whether you wish to take the final exam! Econ PhD students don't need to email us – they have to take the exam. If you are a pre-doc and you wish to waive the class next year, you will have to take the final exam as well. If you need any special accommodations for the exam, please email us as well.

3 Resources (non-metrics related)

Below I am highlighting two of the many resources for you to make use of if you are experiencing a conflict or issue that is interfering with your personal, academic or work life. Grad school can be pretty tough on its own already, not to mention if any other issues additionally arise. Make sure to make use of the resources Stanford has to offer!!

Resources

- Stanford's Ombuds office (https://ombuds.stanford.edu/): this is a **confidential** resource!
- Graduate Life Office (https://glo.stanford.edu/): this is a **non-confidential** resource

Lastly, if you ever need anyone to chat to regarding any of the issues mentioned above, please feel free to reach out to me as well! Always happy to grab a coffee :)

*Most of these section notes are drawn from those prepared by Lea Bottmer, as well as Guido's lecture slides and notes.

4 Section Norms

Creating an inclusive classroom where everybody feels welcome is very important to me. Thus, I would like to reiterate our **community norms**

- One microphone: we will allow everyone to finish their thought without interruption.
- Everyone got accepted to Stanford and is here for a reason. There is no need prove anything to anyone.
- Asking questions is a great way to deepen your understanding and resolve any confusion you have left. It is not a way to signal.
- There is no such thing as a stupid or dumb question.
- Questions asked during the section will be answered by the TA.

I have set up two Google forms to help with section. Section Question Form and Section Feedback Form. Use the former to submit any questions/requests to review specific material that you would like me to address in section (please submit by 11:59pm Tuesday in order for me to have time to prepare to discuss in section on Thursday). Use the latter to submit any feedback you might have on section (e.g. things that you think are working/not working, potential improvements, etc.). Both forms are anonymous so please use them if you have any questions or feedback!

5 Homework

- **Due Dates:** No late homework submissions will be accepted. Problem set 1 is due on Sunday, April 6, 11pm sharp.
- **Submission Requirements:** Homework assignments submissions should be typed. Now is a good time to learn LaTeX or LyX; Microsoft Word is acceptable as well though not ideal. Please submit the final product as a .pdf file printable to letter size (8.5 inch $\times$ 11 inch) paper. Note that this is a slightly different size than A4 paper which is the default for some LaTeX document types. All homework is submitted through the Canvas page for the assignment. When submitting your homework, you should submit no more than two files: (1) a .pdf file with the homework writeup and (2) a .zip file (or other compressed file format) with all of your code and user-created programs.
- Make sure your write-up is self-contained. That means the grader will not have to look into your code and the output console to find the relevant numbers, but rather they are all reported and interpreted in your write-up.
- Output should be presented in an easy-to-read format. When there is a single quantity to report, it should be reported in a sentence. When there is more than one quantity to report, it should be presented in a table. Output should also be explained and interpreted clearly in the text when appropriate. For empirical questions, try to make economic sense of your results (the "econ" part of econometrics) and point out where they do not make sense and why you think this may be the case.
- For "theoretical" questions, make sure to explain your steps clearly (including citing theorems, references, etc.) to obtain full credit.

- Code should be well-commented, include all files that are called by the main file, and submitted with your submitted homework compressed into a .zip file. The grader should be able to run the code on his or her computer and get the same results that you are getting.
- For problem set related questions, please come to my office hours or talk to me during/after class. If you can't make office hours, I am happy to set up another time to chat. For any problem set related emails, please make sure to cc Guido on the emails too, so he can help answer questions. I will also not respond to problem set related emails that are sent the day off the problem set deadline.

6 Clarifications and Hints for Problem Set 1

The lecture notes (Part I Chapter I) might be helpful if you get stuck at estimating the homoskedastic variance.

7 Review of Concepts from Class

7.1 Road Map for the Entire Course

As a general road map for the rest of the quarter, we'll start off with **randomized experiments**. They are a great starting point to talk about causal inference because they are the simplest (and cleanest) possible setting to obtain credible causal estimates. Their important feature is that the assignment mechanism, that is, how each individual came to receive the treatment level received, is known and independent of any observable characteristics of the units in the experiment. After a brief aside on **identification and different approaches to causal inference**, we will then move on to **observational methods in cross sectional settings**. As opposed to experiments, in observational data settings, the assignment mechanism is generally unknown and is usually assumed to depend on observable or unobservable characteristics of the units and sometimes even the potential outcomes themselves. We will talk about several empirical approaches used in observational data settings to obtain causal estimates, including doubly robust estimation methods, instrumental variables and regression discontinuity designs. For the last section of the course, we will move on to **panel data settings**, i.e. we observe several units over multiple time periods. Specifically, we will talk about fixed effect models, difference-in-differences (and variations thereof) and synthetic control methods.

7.2 Questions

I plan to start off every section with letting you discuss questions about the topics discussed in the lectures this week with the people sitting next to you. You can use these questions as a check to figure out things you understand and things you are still uncertain about. If you know how to answer all of these questions, you know the main take-aways for all the topics discussed (and you will be in a great place for the exam if you are taking it).

7.2.1 Lecture 1: Randomized Experiments

1. What roles do *sampling*, *potential outcomes*, and the *assignment mechanism* play in the analysis of randomized experiments?
2. Why is the Stable Unit Treatment Value Assumption (SUTVA) important?

3. What is a *sharp* null hypothesis? Why does this allow us to calculate exact Fisher p-values? What are some examples of sharp null hypotheses in the context of potential outcomes?

4. How can we calculate / approximate the exact p-value? What test statistics can we think of?

5. How is the unbiased estimator of the average treatment effect calculated?

6. What is the formula for the variance of the unbiased estimator of the average treatment effect? What is the primary challenge in estimating this? What do we need to do to make the variance calculable and does this make standard errors more or less conservative?

7.2.2 Lecture 2: Stratified Randomized Experiments and Power Analyses

1. How does one use a stratified randomized experiment to calculate average treatment effects? (formulas)

2. What is meant by Adjusting by Design versus Adjusting by Analysis? What are the (dis-) advantages of a stratified, paired, fully randomized design?

3. When / How should a researcher re-randomize units to treatment and control?

4. What do we mean by power calculations? How do we perform power calculations based on the difference in means estimator?

7.3 Big Picture This Week

First, I would briefly like to get everyone on the same page about the big picture for this week. In general, for this first section on experiments, I have found this article [1] written by Susan and Guido to be incredibly helpful as a resource for gaining a deeper understanding of the topics discussed. In the first lecture, we focused on completely randomized experiments (CRE). Those experiments are the simplest experimental setting possible and probably a lot of times the experiments we think about when someone mentions experiments. One example of completely randomized experiments are clinical trials where a certain number of individuals are randomly assigned to treatment, receiving the new medication, or control, receiving a placebo. In the next lecture, we considered slightly more complicated designs: stratified and paired experiments. The overarching theme for that lecture was thinking about how to incorporate covariate information. Particularly, thinking about (1) What if we run a completely randomized experiment, but we observe that the covariates are not balanced? and (2) Is there a way we can incorporate the covariate information a priori to make our estimates more precise? In general, we can think about the inclusion of covariate information in two separate ways: (1) design and (2) analysis.

7.4 Introduction to Causality

Many of you have probably heard the saying **"correlation does not imply causation"**. But what does this actually mean? When two variables, A and B, are correlated, it means that they have a tendency to move together, positively or negatively. However, this does not imply that a change in A results in a change in B. In econometrics, we are oftentimes interested in exactly this stronger relationship: we want to know if changing a variable will lead to a change in another variable. Generally speaking, we are interested in questions like "What is the effect of changing

[1] `https://www.theigc.org/sites/default/files/2016/06/athey_imbens_june19.pdf`

someone's treatment status on some outcome?". These findings will be much more insightful for policy making than correlational evidence.

Randomized experiments are oftentimes considered the "gold standard" in causal inference. A well-conducted experiment requires few assumptions to establish causality. Randomizing units to treatment and control takes away a lot of uncertainty about the underlying mechanism how some people in real life ended up with treatment or control, allowing us to make truly causal claims about the treatment. Thus, it truly holds everything, but the treatment fixed.

Broadly, there are three areas that we are interested in when analyzing randomized experiments:

1. Point Estimation
2. Inference
3. Hypothesis Testing (related to inference)

Specifically, for the purpose of our experiments section, we are mainly looking at these four **questions**:

1. Estimate for average effect of treatment (point estimation)
2. Variance for the estimator for average treatment effect (inference)
3. Confidence intervals for the average treatment effect (inference)
4. Is there any effect of the treatment? (hypothesis testing)

In an ideal world, for our analysis, we would like to compare what happens if individual i received treatment, e.g. a drug, vs if individual i remained in control, e.g. no drug. This unfortunately is not possible as we only observe one of these happening. However, with some statistical assumptions (more on that later), we can get around that and still identify several of the quantities of interest mentioned above.

7.5 Notation

In this section, we revisit some of the notation introduced in the lecture. Some of you might not be familiar with the potential outcomes notation, so here are some quick reminders:

- Y_i : Observed outcome
- W_i : Treatment indicator
  - We focus on a binary treatment indicator, i.e. $W_i \in \{0, 1\}$
- R_i : Sample indicator
- $Y_i(w)$: Potential outcome (PO) for individual i with treatment status $W_i = w$
  - $Y_i(0)$: PO in absence of treatment
  - $Y_i(1)$: PO in presence of treatment
  - We generally treat these potential outcomes as fixed.

- We only ever observe one of the two outcomes $- >$ fundamental problem of causal inference
- We can rewrite the observed outcome as

$$Y_i = W_i \; Y_i(1) + (1 - W_i) \; Y_i(0)$$

- X_i: covariates

Table 1 shows an example of how this notation is used for an experiment with outcome Y and a sample from the super-population [2]. You can see that we only observe a treatment indicator, covariates and outcomes for individuals i included in the sample, i.e. individuals for which $R_i = 1$. For the individuals in the sample, some of them will have been assigned to treatment ($W_i = 1$) and some will not. For the ones that have been assigned to treatment, the observed outcome Y_i will be equal to their potential outcome in presence of treatment $Y_i(1)$. We don't observe their outcome in absence of treatment, $Y_i(0)$, hence we observe a question mark for these outcomes. Vice versa for the individuals who have been assigned to the control group.

Aside. By adopting Neyman's repeated sampling approach to estimation and inference in the randomized settings, we are moving a little bit away from what a lot of you might have encountered in previous econometrics classes where we usually take a population perspective and, for example, take sampling uncertainty into account in the standard errors. Specifically, we assume the potential outcomes to be fixed as opposed to random (so we are not assuming they follow a specific distribution). A place where this will be important is when you think about calculating the variance of the observed outcome, i.e. $Var(Y^{obs})$. Let's think this through

$$\begin{aligned} Var(Y^{obs}) &= Var(W_i \; Y_i + (1 - W_i) \; Y_i) \\ &= Var(W_i \; Y_i) + Var((1 - W_i) \; Y_i) + 2 \; Cov(W_i \; Y_i, (1 - W_i) \; Y_i) \end{aligned}$$

Focusing on the first term only, I want to show you how fixing the potential outcomes helps us make more progress in this decomposition.

$$\begin{aligned} Var(W_i \; Y_i) &= Var(W_i \; Y_i(1)) \\ &= Var(W_i) \; Y_i(1)^2 \end{aligned}$$

The first equality follows from the fact that we know that for $W_i \; Y_i$ the outcome Y_i will always be equal to $Y_i(1)$. In the second equality, we can take $Y_i(1)$ then out of the expectation because the potential outcome, unlike the observed outcome, is fixed.

[2] We oftentimes refer to our experimental sample as the population since we are thinking mainly about design-based inference here. Thus, we refer to the population broader than the experimental sample as the super-population. Sometimes the population will also be referred to as the sample (e.g. when we think about sample average treatment effecs vs population average treatment effects), so it can be very confusing and I apologize for that. I am trying my best to be clear while being consistent with the lecture slides/notes.

Population	**Sample**	**Treatment**	**Covariate Attributes**	**Outcome**	**Potential Outcomes**	
i	R_i	W_i	X_i	Y_i	$Y_i(0)$	$Y_i(1)$
1	1	1	$37.5, 5$	13	?	13
2	1	0	$40, -0.6$	7	7	?
3	0	-	-	-	-	-
4	0	-	-	-	-	-
5	1	0	$24.2, 7$	6	6	?
6	0	-	-	-	-	-
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Table 1: Data and Notation

7.6 Assumptions

Given we only ever observe one of the two potential outcomes for each individual, i, we need to make some assumptions if we ever hope to recover a causal estimate. For our current setting, that is randomized experiments, these are the following assumptions we are making:

- **Known treatment assignment**, independent of the potential outcomes:

$$(Y_i(0), Y_i(1)) \perp W_i \ \forall i$$

- **Stable Unit Treatment Value Assumption (SUTVA):** Treatment applied to one unit does not affect the outcome of another:

$$Y_i(\mathbf{W}) = Y_i(W_i) \ \forall i,$$

  where $\mathbf{W}$ is the treatment assignment vector for all individuals in the experiment.

7.7 Completely Randomized Experiments (CRE)

In this section, we discuss completely randomized experiment, the simplest setting for experimentation. We cover point estimation as well as inference for these experiments. One by one we will answer the four questions raised earlier for analyzing a randomized experiment. You can find a definition for completely randomized experiments below.

Definition. Completely randomized experiments consider a finite population of N units. M of the units will receive treatment and $N - M$ remain in control.

7.7.1 Point Estimation

Neyman's repeated sampling approach focuses on methods for estimating the average treatment effect. His approach contains finding an estimator for the average treatment effect and deriving its distribution under repeated sampling. In this context, repeated sampling refers to redrawing from the randomization distribution of $\mathbf{W}$, the assignment vector.

Estimand. In any analysis you conduct as a researcher, you will first have to define your estimand of interest. An estimand simply describes the object of interest for estimation. In Neyman's repeated sampling approach, this is the population level average treatment effect,

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big) = \bar{Y}(1) - \bar{Y}(0).$$

Note here that population refers to all subjects observed in the experiment. The average treatment effect on the treated or the average treatment effect on the control are two other estimands that could be of interest. [3]

Estimator. After having specified what target we want to estimate, we need to define an estimator. One intuitive estimator for the average treatment effect is the difference in the average outcomes for those assigned to treatment and those assigned to control,

$$\hat{\tau}^{DiM} = \frac{1}{M}\sum_{i:W_i=1} Y_i^{obs} - \frac{1}{N-M}\sum_{i:W_i=0} Y_i^{obs} = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}.$$

This estimator has some desirable properties, namely it is unbiased for τ, i.e. $\mathbb{E}[\hat{\tau}^{DiM}] = \tau$. It turns out that the same estimator is also an unbiased estimator for the average treatment effect on the treated in this randomized experiment setting when we think about the super-population context.

7.7.2 Variance

In order to conduct inference, we will derive the variance of this proposed estimator for the average treatment effect, $\hat{\tau}^{DiM}$. [4] Oftentimes, we use this in combination with asymptotic results (e.g. CLTs) to obtain confidence intervals. One important note in calculating the variance is that our treatment assignments are not independent, $W_i \not\perp W_j$ for i, j because of our completely randomized setting. In this setting, knowing that individual i received treatment will decrease the chance of unit j having received treatment. To make this even clearer, let's consider the situation where we have seen all but one unit, e.g. 19 out of 20 units in our experiment. Moreover, we observe each of the observed units' treatment status. Because we give treatment to exactly M of the 20 units in our experiment, we know that the one unobserved unit must be treated if we observed $M-1$ treated units or that it must be control if we observed M treated unit. Hence, the treatment status for the last unit is not independent of the treatment status of the other units.

True variance. To derive the true variance for the difference-in-means estimator for the average treatment effect, we plug in our definition for $\hat{\tau}^{DiM}$ and re-arrange, carefully thinking about which quantities in the expression are random and which are known. After some algebra, we find that the true variance of the difference-in-means estimator for the average treatment effect is

[3] The different estimands might sometimes be confusing. To draw a clear distinction between the average treatment effect and the average treatment effect on the treated, let's think about the treatment effect of being at Stanford. The average treatment effect on the treated would tell you how much the typical *Stanford student* gained as a consequence of being at Stanford while the average treatment effect would tell you how much the typical *applicant to Stanford* gained or lost.

[4] Note that the result will be different when we consider different estimands, e.g. if we want to find the variance of the difference-in-means estimator as an estimator for the average treatment effect on the treated. You will consider this case in problem set 1.

$$Var(\hat{\tau}^{DiM}) = \mathbb{E}[(\tau^{Di\hat{M}} - \tau)^2] = \frac{S_0^2}{N-M} + \frac{S_1^2}{M} - \frac{S_{01}^2}{N},$$

where $S_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(w) - \bar{Y}(w)\right)^2$ for $w = 0, 1$ and $S_{01}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - \tau\right)^2$.

Note that, again, the variance for the difference-in-means estimator here is in terms of the potential outcomes because we assumed them to be fixed throughout.

Estimator. In order to estimate this variance, we focus on each component individually. The first two are easy to get. For S_0^2, we find that

$$s_w^2 = \frac{1}{N-M-1}\sum_{i:W_i=0}\left(Y_i^{obs} - \bar{Y}_0^{obs}\right)^2$$

is unbiased. Similar derivations give us an unbiased estimate for S_1^2. The third term is difficult to estimate. Note that, since we don't ever observe both $Y_i(1)$ and $Y_i(0)$ for individual i, we don't observe S_{01}^2 in practice. However, we know that is is a positive quantity because it is squared. Thus, in practice, we just ignore it in the estimation and obtain a **conservative variance estimator**:

$$\widehat{Var(\hat{\tau}^{DiM})} = \frac{s_0^2}{N-M} + \frac{s_1^2}{M}.$$

Under special circumstances, like a constant treatment effect (i.e. $Y_i(1) - Y_i(0) = c\ \forall i$) or if we are interested in super-population inference, this cross term is equal to zero and it vanishes. In this case, we find that our variance estimator is actually an unbiased estimator for the variance.

Note. Unlike probably a lot of econometrics or statistics classes you have previously taken, we are not worried about sampling uncertainty here. Instead of answering questions like "What would the sample look like if we observed different sample indicators?", we ask "What would the sample look like if we observed a different treatment assignment?". The standard errors in this approach will capture the uncertainty due to treatment assignment as opposed to the uncertainty due to sampling variation. This thought experiment is particularly important when we observe an entire population, e.g. all states in the US. In the literature, these two approaches are generally referred to as *design-based* and *sampling-based* uncertainty.

Confidence interval. We invoke a central limit theorem to create confidence intervals. We rely on the large sample approximations (e.g. normality of the estimator) for this. More details on the derivation of this can be found in the lecture slides. We find the 95% bounds as follows

$$CI_{lower} = \hat{\tau} - 1.96 * \sqrt{\widehat{Var(\hat{\tau})}}$$
$$CI_{upper} = \hat{\tau} + 1.96 * \sqrt{\widehat{Var(\hat{\tau})}}$$

7.7.3 Hypothesis testing: Fisher Exact p-Values

Sometimes you might only be interested in figuring out whether there *is* an effect of the treatment at all instead of whether the average effect is zero or not (or any other value). Note how these two cases are quite different as in the first case we conjecture that there is no treatment effect for any

unit while in the second case we can still have a positive and negative treatment effect for some individuals, they just have to average out to zero. If we are interested in thinking about the first scenario, that is whether there is any effect at all, we can use Fisher exact p-values. We get the p-values from the Fisher exact test approach which requires two choices (i) the choice of the **sharp** null and (ii) the choice of a test statistic.

We first focus on (i) the choice of the **sharp** null. A sharp null hypothesis means that all values of the potential outcomes for the units in the experiment are either observed or can be inferred. For example, let the sharp null be that the program has absolutely no effect on earnings

$$H_0: \ Y_i(0) = Y_i(1) \ \forall i = 1, ..., n.$$

This is in contrast to a null hypothesis that assumes that the *average* treatment effect is zero. Why? This is because the treatment can still have a positive or negative effect for some individuals in the sample even if the average effect is zero.

Next, we move on to (ii) the choice of a test statistic. A test statistic is a function of the stochastic assignment vector, $\mathbf{W}$, the observed outcomes $\mathbf{Y}$, and pre-treatment variables, $\mathbf{X}$. The stochastic nature of the treatment assignment produces a **randomization distribution** of the test statistic that we will use to test for statistical significance. The most standard choice of a statistic in randomized experiments is the difference in average outcomes by treatment status:

$$T = \frac{\sum_i W_i Y_i}{\sum_i W_i} - \frac{\sum_i (1 - W_i) Y_i}{\sum_i (1 - W_i)}.$$

Alternatives include:

- **Difference in transformed average outcomes:** $T = \frac{\sum_i W_i ln(Y_i)}{\sum_i W_i} - \frac{\sum_i (1-W_i) ln(Y_i)}{\sum_i (1-W_i)}$
- **Normalized Rank Statistic:** $T = \frac{\sum_i W_i Rank_i}{\sum_i W_i} - \frac{\sum_i (1-W_i) Rank_i}{\sum_i (1-W_i)}$, where $Rank_i = \sum_{j=1}^{N} \mathbb{1}_{Y_j < Y_i} + \frac{1}{2}\left(1 + \sum_{j=1}^{N} \mathbb{1}_{Y_j^{obs} = Y_i^{obs}}\right) - \frac{N+1}{2}$

Aside. If the sample is reasonably large and the distribution is not highly skewed, a lot of these test statistic yield very similar estimates. Sometimes, if we are concerned about outliers driving our conclusions, we might prefer choosing a more robust statistic like the rank statistic over others. Generally, you do not need to choose a test statistic that has an interpretation, but simply one that helps you test your hypothesis. This is different from estimating the average effect!!
Having determined the null hypothesis as well as the test statistic, we can move on to test for statistical significance. In plain English, we are trying to find out how likely it is to observe a value of the test statistic that is as large in absolute value as the one we actually observed. If the p-value is close to zero, we reject the null hypothesis.

Summary of the procedure

1. Define the **sharp** null hypothesis you would like to test.
2. Define an appropriate test statistic, T.
3. Calculate its distribution (exact or numerical approximation).

4. Calculate the p-value accordingly: $p = \frac{\sum_a \mathbf{1}_{|T^a| \geq |T^{obs}|}}{A}$, [5] where A is the total number of treatment assignment vectors (two-sided test since we are testing the null of no effect whatsoever).

Figure 1 shows what this distribution could look like. We can use this visualization to judge significance (somewhat inaccurately) too. Assume your observed test statistic is the one in red. Given where it lies in the distribution, it seems like it is really likely to observe a value as big as the observed one from any random draw of the treatment assignment vectors where we assume the treatment has no effect whatsoever. Thus, we would not reject the null hypothesis. However, if the observed test statistic is the one in blue, then our conclusion would be different. It lies in the tail end of the distribution of the test statistics, indicating that it is not super likely to have observed such a test statistic from just randomly drawing a treatment assignment vector and assuming no treatment effect whatsoever. Hence, we would conclude that we can reject the null hypothesis.

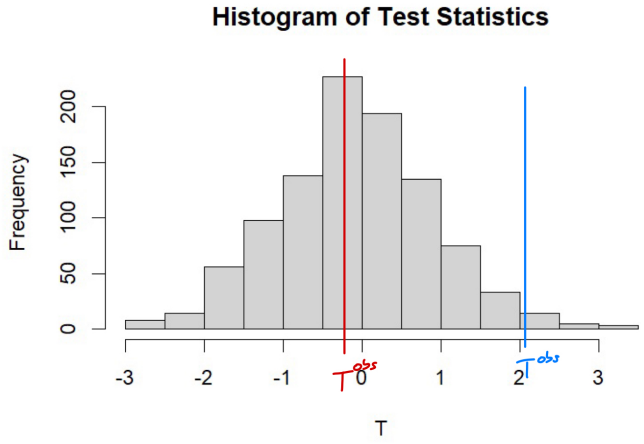

Figure 1: Potential histogram of test statistics. If T^{obs} is the red one, the p-value will be pretty high. If T^{obs} is the blue line, T^{obs} will be significant and have a small p-value.

[5] In general, this is $p = \sum_{w \in W} \mathbf{1}_{|T^w| \geq |T^{obs}|} \mathbb{P}(W = w)$, where W are the possible treatment assignment vectors and $\mathbb{P}(W = w)$ indicates each realized treatment vector's probability.

7.8 Dealing with Covariate Imbalances

Assume we observe covariates in the data set. Randomization should in theory take care of balancing the covariates, but sometimes this does not work out in practice. How can we resolve this? We can generally think of two ways to incoporate covariates in experiments: 1) ex-ante, which we will discuss in subsection 7.8.1 and 7.8.2, and 2) ex-post, which we will discuss in subsection 7.8.3. Subsection 7.8.4 includes a discussion about incorporating covariates into the experimental design and analysis.

7.8.1 Stratified and Paired Experiments

The first remedy is to ensure you avoid having the imbalance in the first place. This can be done using *stratified experiments*, which means we stratify the sample based on covariates first (e.g., imagine placing the sample into two different buckets, women and men, and then randomizing treatment within those two buckets). In other words, we group individual observations together and randomize within these groups. That way, we achieve covariate balance by design. The limiting case of stratification is to pair the units up, known as a *paired* experiment.

Naturally, when choosing the right design for your experiment, you might wonder what the right strategy is. Should I be running a completely randomized experiment, a stratified experiment or a paired experiment? We use Table 2 to highlight advantages and disadvantages for each of these designs:

	CRE	**Stratified Experiments**	**Paired Experiments**
Advantages	• Simple and easy to implement	• Lower expected squared error compared to completely randomized design (strictly if covariates predict potential outcomes) • Can establish whether there is heterogeneity in treatment effects	• Lower expected squared error compared to stratified design in finite samples • Higher power in sufficiently large samples
Disadvantages	• Makes no use of covariates		• Cannot estimate variance of average treatment effect

Table 2: Advantages and Disadvantages of Experimental Designs

Estimand. In the stratified setting, we are interested in the super-population of the finite sample,

$$\tau_{pop} = \mathbb{E}[Y_i(T) - Y_i(C)] = \mathbb{E}[\tau(X_i)] = \mathbb{E}[\mathbb{E}[Y_i(T) - Y_i(C)|X_i = x]].$$

Estimator. The strata estimator first estimates the average treatment effects within each stratum and then weights these by the relative stratum sizes, i.e. for the male and female example:

$$\hat{\tau}^{strata} = \frac{N_{0f} + N_{1f}}{N}\big(\bar{Y}_{1f}^{obs} - \bar{Y}_{0f}^{obs}\big) + \frac{N_{0m} + N_{1m}}{N}\big(\bar{Y}_{1m}^{obs} - \bar{Y}_{0m}^{obs}\big).$$

In more general terminology, let g denote a stratum, then $\hat{\tau}^{strata} = \sum_{g=1}^{G} \frac{N_g}{N} \hat{\tau}_g$. Alternatively, we can use the regression function

$$Y_i^{obs} = \alpha + \tau W_i + \beta \mathbb{1}_{X_i = f} + \varepsilon_i,$$

where we estimate τ by least squares and get $\hat{\tau}^{reg}$.

Variance. We get the variance of this estimator by calculating the variance for each stratum separately and averaging over all strata,

$$Var(\hat{\tau}^{strata}) = \sum_{g=1}^{G} \left(\frac{N_g}{N}\right)^2 Var(\hat{\tau}_g),$$

where $\hat{\tau}_g$ is the difference-in-means estimate for stratum g and $Var(\hat{\tau}_g) = \frac{S_{0,g}^2}{N_{0,g}} + \frac{S_{1,g}^2}{N_{1,g}} - \frac{S_{01,g}^2}{N_g}$ is the true variance for stratum g. We estimate this variance by estimating each $Var(\hat{\tau}_g)$ term by its conservative Neyman variance, i.e.

$$\widehat{Var(\hat{\tau}^{strata})} = \sum_{g=1}^{G} \left(\frac{N_g}{N}\right)^2 \widehat{Var(\hat{\tau}_g)}.$$

Recommendation in practice. The recommendation in the literature about which experiment to run is as follows:

- In *large* samples, and if the covariates are strongly associated with the outcomes, stratify or pair.
- In *small* samples, with weak association between covariates and outcomes, there is mixed advice.

The main take-away from this is that you should be careful in practice. Make sure to think through and be explicit about your goals: are we aiming for more precise estimators or more powerful tests? Moreover, always make sure to clearly state your estimand in your analysis: what is the population of interest?

Recommendation in practice: It is better to use small strata rather than pairs!

7.8.2 Re-Randomization

Another way to deal with covariate imbalances after randomization is to re-randomize. This procedure can improve precision of the estimates and power of tests, but make sure to do it carefully to maintain valid inference!

Two Ways to Conduct Re-Randomization

1. Decide *a priori* to randomize M times. Implement assignment vector that minimizes some criterion (e.g., maximum of the t-statistics for the K covariates).
2. Re-randomize until the criterion meets some threshold (e.g., t-statistics below 1, imbalance of covariates below a threshold a (see lecture slides)).

Careful!: (1) Choose your strategy *a priori*, so randomization inference is valid. (2) Do **not** search over all assignments for an optimal value of the criterion.

Recommendations for practice: Instead of re-randomization lay out an acceptable set of random assignments.

7.8.3 Ex-post stratification

When we talk about ex-post stratification, we usually refer to incorporating covariates into our analysis after randomization has been taken place, but not into our experimental design. For example, think about model-based adjustments like controlling for covariates in regressions. We call this *ex-post* because the randomization assignment was done without taking the covariates into account, but we want to adjust for any imbalances we might observe after randomization (ex-post). Suppose we have a single binary covariate and we ran a completely randomized experiment. Ex-post, we can use two estimators to make adjustments: either the strata estimator, $\hat{\tau}^{strata}$ discussed earlier or we can use a regression function

$$Y_i^{obs} = \alpha + \tau \; W_i + \beta \; 1\{X_i = 1\} + \varepsilon_i,$$

where we estimate τ by OLS, leading to $\hat{\tau}^{reg}$.

7.8.4 Discussion on Ex-Ante vs Ex-Post Stratification

There are three comparisons we want to make: 1) stratified design, $\mathcal{S}$, vs completely randomized design, $\mathcal{C}$, 2) the variance of the difference-in-means estimator under both designs and 3) the ex-post estimators in comparison to the difference-in-means under a completely randomized design.

We start off by contrasting the two designs. Here, the take-away is that the stratified design has lower expected squared error than the completely randomized design. This comparison is strict if the covariate is predictive of the potential outcomes in the population. This is an exact finite sample result and, thus, true irrespective of the sample size.

Next, we want to illustrate this result a bit more by comparing the variances of the difference-in-means estimator under both designs. In general, we cannot compare the conditional variance because there is no ranking, but we need to rather compare the marginal variance. While it is possible that stratification can lead to a larger variance because of negative correlations within strata in a finite sample, that is not possible on average (over repeated samples). Thus, if our objective is to get the most precise estimate of the average treatment effect, stratification dominates complete randomization even in small samples.

Lastly, we want to note that under a stratified design, if we assume that the treatment assignment probabilities are the same in the two strata, $\hat{\tau}^{reg}$, $\hat{\tau}^{strata}$ and $\hat{\tau}^{DiM}$ are identical and thus have the same variances. Under a completely randomized design, thy are generally different (see Lecture Notes, Part I, Chapters 2-3, Section 2.4.2 and 2.4.3 for more details on derivation of the variances). Now there is a difference between the finite sample variance and the asymptotic variance. If the sample is sufficiently large, the variance of $\hat{\tau}^{strata}$ under the completely randomized design (ex-post stratification) will be approximately the same as the variance under a stratified design, and thus be able to exploit the advantages of stratification. Hence, we expect the variance to be lower than the variance for the difference-in-means estimator under a completely randomized

design, especially when there is some correlation between the outcomes and the covariates that underly stratification. Intuitively, that is because we explain some of the variation of the outcome variable away by predictive covariates. However, when we are in a small sample setting, if the correlation is weak, the variance of the ex-post stratified estimator may be larger than the one of $\hat{\tau}^{DiM}$.

Main take-away: *"Design trumps analysis"* - Rubin. Make sure that you think through your analyses in advance. *Ex ante* there is no cost to doing stratification, but *ex post* adjustments, for example model-based adjustments like controlling for covariates in regressions, can be costly in finite-samples. Useful strategies to adjust for differences in particular covariates is to randomize in a way that $\hat{\tau}^{DiM} = \hat{\tau}^{reg}$ (e.g. by stratification or re-randomization).

7.9 Power Calculations

A common trade-off in practice for choosing between a completely randomized design or a stratified design or a clustered design (more on that next week) is that we are concerned about the power of our statistical tests. But what is the power of a test? You are probably very familiar with the size of the test, α. The size of the test is the probability of a type I error, which is falsely rejecting the null hypothesis. The power of the test, generally referred to as β, is the probability of correctly rejecting the null hypothesis. In other words, the power of a test is the probability of detecting an effect, if there is a true effect present to detect. One way to use these power calculations is to estimate the sample size required to detect an effect in an experiment, e.g we can answer the question whether a null effect is maybe more likely due to an underpowered experiment rather than a true null effect. A typical question would be: given the true mean and variance of the outcome as well as a specific effect size, what is the minimum sample size N I need to achieve a specific size and power of a test?

A general procedure is outlined below:

1. Specify the equation for which we reject the null hypothesis, e.g.,

$$|T| \geq \Phi^{-1}(1 - \alpha/2).$$

   (Rejecting a two-sided test of size α, where the test statistic T is standard normally distributed under the null.) This could be the mean of a distribution or the difference-in-means estimator.

2. Specify the probability of the above equation and set it equal to β, the power of the test.

3. Re-arrange the equation to solve for N. If N is not an integer, make sure you pick the closest integer *greater* than the exact value to obtain the minimum sample size, e.g. if you find $N = 13.8$, we say that the minimum sample size is $N = 14$.

We introduced the calculations for two examples in the lectures. We find the following formulas:

1. Testing a mean against zero:

$$N = \Big(\frac{\Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2)}{\mu_0/\sigma}\Big)^2$$

   (Assuming a random sample from a normal distribution).

2. Testing a DiM with unequal sample sizes and equal variances:

$$N = \frac{\left(\Phi^{-1}(\beta) + \Phi^{-1}(1-\alpha/2)\right)^2}{(\tau_0^2/\sigma^2) \cdot \gamma \cdot (1-\gamma)},$$

where γ is the share of treated units.

8 Practice Problems

8.1 Randomization Test: Fixed M vs. Random M

As in the lecture, assume that the econometrician observes $\big(W_i, Y_i(W_i)\big)$ for all i in the population of four individuals where $Y_i(W_i) = W_i Y_i(1) + (1-W_i)Y_i(0)$ results from a randomized experiment. We will use the test statistic

$$T = \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i} - \frac{\sum_{i=1}^{n} (1-W_i)\, Y_i}{\sum_{i=1}^{n} (1-W_i)}$$

The data are as follows:

i	W_i	Y_i^{obs}	$Y_i(0)$	$Y_i(1)$
1	0	0		
2	0	6		
3	1	18		
4	1	42		

1. Perform the randomization test that was discussed in lecture (fixing the number of treated units at two). What is the Fisher exact p-value implied by the data when testing the sharp null hypothesis of no treatment effect for any individual?

2. Now perform the randomization test that would result if you let the number of treated units vary stochastically according to a binomial distribution with success probability of $\frac{1}{2}$. What exact p-value (for the sharp null of no treatment effect for any individual) is implied?

Randomization Inference (Final Exam 2017, #1)

Consider a completely randomized experiment with N_0 control units and N_1 treated units.
(a) Describe how you would test the null hypothesis that $Y_i(1) = 2 \times Y_i(0) + 3$ for all i.
(b) Suppose you want to estimate the difference in medians by treatment status. Describe how you would estimate this. [*added*] Explain why we might want to use this statistic over the difference in means.

8.2 Randomized Experiment Variance Calculation (Final Exam 2018, #1)

Consider a completely randomized experiment with N_0 control units and N_1 treated units. Let $Y_i(0)$, $Y_i(1)$ denote the potential outcomes, Y_i denote the realized outcome, and $W_i \in \{0,1\}$ the treatment. The exact variance for the simple difference in means estimator

$$\bar{Y}_1 - \bar{Y}_0 = \frac{1}{N_1}\sum_{i=1}^{N} W_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1-W_i)Y_i$$

is

$$\frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{01}^2}{N}$$

where

$$S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(0) - \frac{1}{N}\sum_{j=1}^{N} Y_j(0)\right)^2$$

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - \frac{1}{N}\sum_{j=1}^{N} Y_j(1)\right)^2$$

$$S_{01}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - \left(\frac{1}{N}\sum_{j=1}^{N} Y_j(1) - \frac{1}{N}\sum_{j=1}^{N} Y_j(0)\right)\right)^2$$

(a) Calculate the exact, finite sample, variance of $\bar{Y}_1$.
(b) Calculate the exact, finite sample, covariance of $\bar{Y}_1$ and $\bar{Y}_0$.

8.3 Randomized Experiment (Final Exam 2020, #1)

Suppose we conduct a randomized experiment on a random sample of the US population. We assign the treatment randomly at the individual level. The dataset observed by the econometrician includes individual-level outcomes, location (state) for each individual, and a binary treatment indicator.
(a) Suppose you estimated the average treatment effect as the difference in means by treatment status, give an expression for the variance of the estimator, and how you could estimate the unknown components of that variance? You can ignore the fact that the population is finite.
(b) How would you use the location information to obtain a more efficient estimator? What would its variance be?

8.4 Randomized Experiments

Suppose we have an infinitely large population. We randomly sample N units from this population. We consider an experiment where we randomly assign each unit to a binary treatment (independently for each unit), with the probability of being in the treatment group equal to p. Let $W_i \in \{0,1\}$be the binary treatment. Let $Y_i(0)$ and $Y_i(1)$ be the potential outcomes (assuming SUTVA, or no-interference), and let $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ be the population average effect of the treatment, and let $\tau_s = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$ be the sample average treatment effect. Consider two estimators

$$\hat{\tau}^1 = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{W_i Y_i}{p} - \frac{(1-W_i)Y_i}{1-p}\right\}$$

and

$$\hat{\tau}^2 = \frac{1}{\sum_{i=1}^N W_i} \sum_{i=1}^N W_i Y_i - \frac{1}{\sum_{i=1}^N (1-W_i)} \sum_{i=1}^N (1-W_i) Y_i$$

(if one of the denominators is equal to zero, the estimator is defined to be zero).

1. What is the exact finite variance of the two estimators, viewed as estimators of the sample average treatment effect? (For the second estimator you can condition on the value of $\sum_{i=1}^N W_i$, and you can focus on the case where this value is between 1 and $N-1$).

2. What is the asymptotic variance of the two estimators, viewed as estimators of the population average treatment effect?

3. Suppose I also observe some covariates (but the assignment is still completely random). Describe how you can improve over the second estimator by exploiting the covariates.

8.5 Randomized Experiment (Final Exam 2023, #1)

Suppose we have an infinitely large population. We randomly sample N units from this population. We consider an experiment where we randomly assign each unit to a binary treatment (independently for each unit), with the probability of being in the treatment group equal to p. Let $W_i \in \{0,1\}$ be the binary treatment. Let $Y_i(0)$ and $Y_i(1)$ be the potential outcomes (assuming SUTVA, or no interference) and let $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ be the population average effect of the treatment and let $\tau_s = \sum_{i=1}^N (Y_i(1) - Y_i(0))/N$ be the sample average treatment effect.
(a) Show that

$$Y_i \frac{W_i - p}{p(1-p)}$$

conditional on the potential outcomes is unbiased for the unit-level treatment effect $\tau_i = Y_i(1) - Y_i(0)$.
(b) Is the average

$$\frac{1}{N} \sum_{i=1}^N Y_i \frac{W_i - p}{p(1-p)}$$

a good estimator for τ?

8.6 Power Calculations

Suppose we plan to test whether the mean of a population is different from 0 using a t-statistic, with a $\alpha = 0.05$ significance level. Assume that the variance of the observations is known to be $\sigma^2 = 100$. How many (independent) observations N do we need to have power $\beta = 0.9$ when the true mean is $\mu = 1$? (Let $\Phi^{-1}(0.8) = 0.8416, \Phi^{-1}(0.9) = 1.2816, \Phi^{-1}(0.95) = 1.6449$ and $\Phi^{-1}(0.975) = 1.9600$.)

8.7 Power Calculations

Suppose we want to run a completely randomized experiment on $N = 3000$ individuals. We plan to test whether the average treatment effect is different from 0 using the difference in means estimator, at a $\alpha = 0.05$ significance level. I want to achieve a power of $\beta = 0.8$ when the true treatment effect is $\tau = 1$.

(a) What should the variance of my outcome be for me to achieve my goal? (Let $\Phi^{-1}(0.8) = 0.8416, \Phi^{-1}(0.9) = 1.2816, \Phi^{-1}(0.95) = 1.6449$ and $\Phi^{-1}(0.975) = 1.9600$.)
(b) Suppose I observe some pre-treatment data and find that $\sigma^2 = 64$. Based on your answer to (a), will I be able to achieve the desired power in my experiment? Explain why or why not.

9 Solution Sketches for Practice Problems

Randomization Test: Fixed M vs. Random M

As in the lecture, assume that the econometrician observes $\big(W_i, Y_i(W_i)\big)$ for all i in the population of four individuals where $Y_i(W_i) = W_i Y_i(1) + (1 - W_i)Y_i(0)$ results from a randomized experiment. We will use the test statistic

$$T = \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i} - \frac{\sum_{i=1}^{n} (1 - W_i)\, Y_i}{\sum_{i=1}^{n} (1 - W_i)}$$

The data are as follows:

i	W_i	Y_i^{obs}	$Y_i(0)$	$Y_i(1)$
1	0	0		
2	0	6		
3	1	18		
4	1	42		

1. Perform the randomization test that was discussed in lecture (fixing the number of treated units at two). What is the Fisher exact p-value implied by the data when testing the sharp null hypothesis of no treatment effect for any individual?

2. Now perform the randomization test that would result if you let the number of treated units vary stochastically according to a binomial distribution with success probability of $\frac{1}{2}$. What exact p-value (for the sharp null of no treatment effect for any individual) is implied?

Solution Sketch

1. Note that we can fill in some blanks as a first step:

i	W_i	Y_i^{obs}	$Y_i(0)$	$Y_i(1)$
1	0	0	0	?
2	0	6	6	?
3	1	18	?	18
4	1	42	?	42

We specify our (sharp) null hypothesis as

$$H_0 : Y_i(1) = Y_i(0)\ \forall i.$$

We want to test the null hypothesis of no treatment effect, so specify our alternative hypothesis as

$$H_1 : Y_i(1) \neq Y_i(0) \text{for at least one} i.$$

Note that this means we will conduct a tw-sided hypothesis test.

From our null hypothesis, we can fill in the rest of the table:

i	W_i	Y_i^{obs}	$Y_i(0)$	$Y_i(1)$
1	0	0	0	0
2	0	6	6	6
3	1	18	18	18
4	1	42	42	42

The prompt defines a test statistic for us to use. That means we have everything we need to calculate our Fisher exact p-value.

To start, we note that there are $\binom{4}{2}$ possible assignment vectors. For each assignment vector, we calculate T.

$$\begin{aligned}
W = (0\ 0\ 1\ 1) &\rightarrow T = \frac{1}{2}(18+42) - \frac{1}{2}(0+6) = 27 \\
W = (0\ 1\ 1\ 0) &\rightarrow T = \frac{1}{2}(6+18) - \frac{1}{2}(0+42) = -9 \\
W = (0\ 1\ 0\ 1) &\rightarrow T = 15 \\
W = (1\ 0\ 0\ 1) &\rightarrow T = 9 \\
W = (1\ 0\ 1\ 0) &\rightarrow T = -15 \\
W = (1\ 1\ 0\ 0) &\rightarrow T = -27
\end{aligned}$$

To calculate our p-value, we compare each $|T|$ to $|T^{obs}|$:

$$\begin{aligned}
W = (0\ 0\ 1\ 1) &\rightarrow |T| = 27 = |T^{obs}| \\
W = (0\ 1\ 1\ 0) &\rightarrow |T| = 9 < |T^{obs}| \\
W = (0\ 1\ 0\ 1) &\rightarrow |T| = 15 < |T^{obs}| \\
W = (1\ 0\ 0\ 1) &\rightarrow |T| = 9 < |T^{obs}| \\
W = (1\ 0\ 1\ 0) &\rightarrow |T| = 15 < |T^{obs}| \\
W = (1\ 1\ 0\ 0) &\rightarrow |T| = 27 = |T^{obs}|
\end{aligned}$$

Thus, we find that the implied p-value is

$$p = \frac{1}{6}(1+0+0+0+0+1) = \frac{1}{3}.$$

2. What changes in this part of the question is that we also consider assignment vectors that have $M = 1$ or $M = 3$. Note that we have a total of $2^4 = 16$ possible treatment assignments. However, we exclude the vectors with all treatment and all control as the test statistic is not well defined for those. Thus, we are left with $2^4 - 2 = 14$ assignment vectors. We already calculated all the statistics for the case where $M = 2$. The cases where $M = 1$ and $M = 3$ are the same up to a sign flip which we are going to ignore when we calculate the absolute values.

For $M = 1$, we find

$$\begin{aligned}
W = (1\ 0\ 0\ 0) &\rightarrow T = -21 \ \ \& \ \ |T| = 21 < |T^{obs}| \\
W = (0\ 1\ 0\ 0) &\rightarrow T = -14 \ \ \& \ \ |T| = 14 < |T^{obs}| \\
W = (0\ 0\ 1\ 0) &\rightarrow T = 2 \ \ \& \ \ |T| = 2 < |T^{obs}| \\
W = (0\ 0\ 0\ 1) &\rightarrow T = 34 \ \ \& \ \ |T| = 34 > |T^{obs}|
\end{aligned}$$

Thus, for our p-value we find

$$p = \frac{1}{14} * (4) = \frac{4}{14} = \frac{2}{7}.$$

Note that we used $p = \frac{1}{2}$ in this calculation. if $p \neq \frac{1}{2}$, we would not be able to simply multiply our sum of indicators by 1 over the number of treatment vectors in this randomized experiment. For example, while $W = (0\ 0\ 0\ 1)$ and $W = (1\ 1\ 1\ 0)$ have the same statistic, up to a sign, their probability, which we are multiplying it by to obtain the p-value would be different if $p \neq \frac{1}{2}$. Assume $p = \frac{1}{3}$, the first W, W_1, would have $P(W = W_1) = (\frac{2}{3})^3 * \frac{1}{3}$ while the second one, W_2, would have $P(W = W_2) = (\frac{1}{3})^3 * \frac{2}{3}$ Recall the general p-value formula:

$$p = \sum_a \ P(W = a)\ \mathbf{1}_{|T^a| \geq |T^{obs}|}.$$

Thus, we would not be able to pull out $P(W = a)$ of the sum when $p \neq \frac{1}{2}$.

Randomization Inference (Final Exam 2017, #1)

Consider a completely randomized experiment with N_0 control units and N_1 treated units.
(a) Describe how you would test the null hypothesis that $Y_i(1) = 2 \times Y_i(0) + 3$ for all i.
(b) Suppose you want to estimate the difference in medians by treatment status. Describe how you would estimate this. Explain why we might want to use this statistic over the difference in means.

Solution Sketch

(a) We can use Fisher exact p-values. This is a sharp null hypothesis, so given Y_i^{obs} and W_i, we can infer the unobserved potential outcome given $H_0 : Y_i(1) = 2 \times Y_i(0) + 3\ \forall i$. We choose a test statistic, e.g. difference-in-means, and obtain its distribution for every possible treatment assignment (or approximate it depending on how large N_0 and N_1 are. Lastly, we calculate the p-value.
Note: there was some confusion in section over how a difference in means estimator would work here. The natural estimator would be to use the difference in means between the two sides of the sharp null: $T = \bar{Y}_1^{obs} - 2\bar{Y}_0^{obs} - 3$. Under the null, this statistic would be 0.

(b) We estimate the difference in medians by splitting the sample into one sample with only treated units and one sample with only control units. For each sample separately, we evaluate its median. Our estimator will simply be the difference between those values

$$\hat{\tau} = med(Y_i \ s.t. \ W_i = 1) - med(Y_i \ s.t. \ W_i = 0).$$

We might prefer this if your sample contains a lot of outliers since the median is more robust than the mean.

Randomized Experiment Variance Calculation (Final Exam 2018, #1)

Consider a completely randomized experiment with N_0 control units and N_1 treated units. Let $Y_i(0)$, $Y_i(1)$ denote the potential outcomes, Y_i denote the realized outcome, and $W_i \in \{0, 1\}$ the treatment. The exact variance for the simple difference in means estimator

$$\bar{Y}_1 - \bar{Y}_0 = \frac{1}{N_1}\sum_{i=1}^{N} W_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - W_i)Y_i$$

is

$$\frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{01}^2}{N}$$

where

$$S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(0) - \frac{1}{N}\sum_{j=1}^{N} Y_j(0)\right)^2$$

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - \frac{1}{N}\sum_{j=1}^{N} Y_j(1)\right)^2$$

$$S_{01}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - \left(\frac{1}{N}\sum_{j=1}^{N} Y_j(1) - \frac{1}{N}\sum_{j=1}^{N} Y_j(0)\right)\right)^2$$

(a) Calculate the exact, finite sample, variance of $\bar{Y}_1$.
(b) Calculate the exact, finite sample, covariance of $\bar{Y}_1$ and $\bar{Y}_0$.

Solution Sketch

(a) We have

$$\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^{N} W_i Y_i = \frac{1}{N_1}\sum_{i=1}^{N} W_i Y_i(1)$$

So

$$\begin{aligned}\text{var}(\bar{Y}_1) &= \text{var}(\frac{1}{N_1}\sum_{i=1}^{N} W_i Y_i(1)) \\ &= \frac{1}{N_1^2}\text{var}(\sum_{i=1}^{N} W_i Y_i(1))\end{aligned}$$

Note that $Y_i(1)$ is fixed, but W_i is random. If we flipped a coin for each unit, then W_i and W_j would be independent for $i \neq j$, so that the sum of the variance would be the variance of the sum. However, when we have a fixed number of treated, N_1, then if $W_i = 1$, W_j is slightly less likely to

be 1, and if $W_i = 0$, W_j is slightly more likely to be 1. Hence, in this setting, terms in the sum have non-zero covariance.
We rewrite the variance as the covariance of the sum with itself because covariances are a little easier to work with.

$$\begin{aligned}
\text{var}(\bar{Y}_1) &= \frac{1}{N_1^2}\text{var}(\sum_{i=1}^{N} W_i Y_i(1)) \\
&= \frac{1}{N_1^2}\text{cov}(\sum_{i=1}^{N} W_i Y_i(1), \sum_{i=1}^{N} W_i Y_i(1)) \\
&= \frac{1}{N_1^2}\sum_{i=1}^{N}\text{cov}(W_i Y_i(1), \sum_{j=1}^{N} W_j Y_j(1)) \\
&= \frac{1}{N_1^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\text{cov}(W_i Y_i(1), W_j Y_j(1)) \\
&= \frac{1}{N_1^2}\sum_{i=1}^{N}\sum_{j=1}^{N} Y_i(1)Y_j(1)\text{cov}(W_i, W_j)
\end{aligned}$$

So we need to figure out $\text{cov}(W_i, W_j)$. We will do just that in 2 steps:
Case 1: $i = j$
Then $\text{cov}(W_i, W_j) = \text{cov}(W_i, W_i) = \text{var}(W_i) = p(1-p)$, where $p = \frac{N_1}{N}$ is the (Bernoulli) treatment probability.
Case 2: $i \neq j$
Then

$$\begin{aligned}
\text{cov}(W_i, W_j) &= E((W_i - p)(W_j - p)) \\
&= E(W_i W_j - W_i p - W_j p + p^2) \\
&= E(W_i W_j) - E(W_i p) - E(W_j p) + p^2 \\
&= E(W_i W_j) - p^2 \\
&= E(W_i W_j | W_i = 1)\Pr(W_i = 1) + E(W_i W_j | W_i = 0)\Pr(W_i = 0) - p^2 \\
&= E(1 W_j | W_i = 1)p + E(0 W_j | W_i = 0)(1-p) - p^2
\end{aligned}$$

where the second to last line uses the law of total probability. Now, $E(W_j | W_i = 1) = \frac{N_1 - 1}{N - 1}$ because if $W_i = 1$, then only $N_1 - 1$ out of the remaining $N - 1$ units will also be assigned to the treatment. So

$$\begin{aligned}
\text{cov}(W_i, W_j) = \frac{N_1 - 1}{N - 1}\frac{N_1}{N} - \frac{N_1^2}{N^2} \\
&= \frac{N(N_1 - 1)N_1 - (N-1)N_1^2}{(N-1)N^2} \\
&= -\frac{N_1 N_0}{(N-1)N^2}
\end{aligned}$$

We could then plug everything in above.

(b)

We want the covariance between $\bar{Y}_1$ and $\bar{Y}_0$. The derivation can go along similar lines to the previous question:

$$\begin{aligned}
\text{cov}(\bar{Y}_1, \bar{Y}_0) &= \text{cov}(\frac{1}{N_1}\sum_{i=1}^{N} W_i Y_i, \frac{1}{N_0}\sum_{i=1}^{N}(1-W_i)Y_i) \\
&= \frac{1}{N_1}\frac{1}{N_0}\text{cov}(\sum_{i=1}^{N} W_i Y_i(1), \sum_{i=1}^{N}(1-W_i)Y_i(0)) \\
&= \frac{1}{N_1}\frac{1}{N_0}\sum_{i=1}^{N}\sum_{j=1}^{N} Y_i(1)Y_j(0)\text{cov}(W_i, (1-W_j)) \\
&= -\frac{1}{N_1}\frac{1}{N_0}\sum_{i=1}^{N}\sum_{j=1}^{N} Y_i(1)Y_j(0)\text{cov}(W_i, W_j)
\end{aligned}$$

where we can use the previous result for the covariance.

Randomized Experiment (Final Exam 2020, #1)

Suppose we conduct a randomized experiment on a random sample of the US population. We assign the treatment randomly at the individual level. The dataset observed by the econometrician includes individual-level outcomes, location (state) for each individual, and a binary treatment indicator.
(a) Suppose you estimated the average treatment effect as the difference in means by treatment status, give an expression for the variance of the estimator, and how you could estimate the unknown components of that variance? You can ignore the fact that the population is finite.

Solution Sketch

(a) We are going to view the difference in means estimator as an estimator for the population level treatment effect. This takes into account the fact that we are only observing a random sample of the US population. Thus, from the lecture slides, we find that the variance is

$$Var(\hat{\tau}^{DiM}) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1},$$

where σ_w^2 is the population variance of $Y_i(w)$. An estimator for this variance is

$$\widehat{Var(\hat{\tau}^{DiM})} = \frac{s^2}{N_0} + \frac{s_1^2}{N_1},$$

where s_w^2 denotes the sample variance (for the subsamples split by treatment status).

Randomized Experiments (Take-Home Final Exam 2022, #1)

Suppose we have an infinitely large population. We randomly sample N units from this population. We consider an experiment where we randomly assign each unit to a binary treatment (independently for each unit), with the probability of being in the treatment group equal to p. Let $W_i \in \{0, 1\}$be the binary treatment. Let $Y_i(0)$ and $Y_i(1)$ be the potential outcomes (assuming SUTVA, or no-interference), and let $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ be the population average effect of the

treatment, and let $\tau_s = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$ be the sample average treatment effect. Consider two estimators

$$\hat{\tau}^1 = \frac{1}{N}\sum_{i=1}^{N}\{\frac{W_i Y_i}{p} - \frac{(1-W_i)Y_i}{1-p}\}$$

and

$$\hat{\tau}^2 = \frac{1}{\sum_{i=1}^{N} W_i}\sum_{i=1}^{N} W_i Y_i - \frac{1}{\sum_{i=1}^{N}(1-W_i)}\sum_{i=1}^{N}(1-W_i)Y_i$$

(if one of the denominators is equal to zero, the estimator is defined to be zero).

1. What is the exact finite variance of the two estimators, viewed as estimators of the sample average treatment effect? (For the second estimator you can condition on the value of $\sum_{i=1}^{N} W_i$, and you can focus on the case where this value is between 1 and N - 1).

2. What is the asymptotic variance of the two estimators, viewed as estimators of the population average treatment effect?

3. Suppose I also observe some covariates (but the assignment is still completely random). Describe how you can improve over the second estimator by exploiting the covariates.

Solution Sketch

1.
First Estimator:

$$\begin{aligned}
\text{Var}\left(\hat{\tau}_1\right) &= \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i W_i}{p} - \frac{\left(1-W_i\right)Y_i}{1-p}\right) \\
&= \frac{1}{N^2}\frac{1}{[(p(1-p)]^2}\sum_{i=1}^{N}\text{Var}\left(Y_i(1)W_i(1-p) - \left(1-W_i\right)Y_i(0)p\right) \\
&= \frac{1}{N^2}\frac{1}{[(p(1-p)]^2}\sum_{i=1}^{N}\text{Var}\left(Y_i(1)W_i(1-p) - Y_i(0)p + W_i Y_i(0)p\right) \\
&= \frac{1}{N^2}\frac{1}{[(p(1-p)]^2}\sum_{i=1}^{N}\text{Var}\left(W_i\left(Y_i(1)(1-p) + Y_i(0)p\right)\right) \\
&= \frac{1}{N^2}\frac{1}{[(p(1-p)]^2}\sum_{i=1}^{N}\text{Var}\left(W_i\right)\left(Y_i(1)(1-p) + Y_i(0)p\right)^2 \\
&= \frac{1}{N^2}\frac{1}{[(p(1-p)]^2}\sum_{i=1}^{N}p(1-p)\left(Y_i(1)(1-p) + Y_i(0)p\right)^2 \\
&= \frac{1}{N^2}\frac{1}{p(1-p)}\sum_{i=1}^{N}\left(Y_i(1)(1-p) + Y_i(0)p\right)^2
\end{aligned}$$

Second Estimator:

Once we condition on $\sum_{i=1}^{N} W_i$, we can treat this estimator as the simple difference-in-means estimator that we analyzed for completely randomized experiments. Thus, from the lecture slides, we know that

$$Var(\hat{\tau}_2) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{01}^2}{N},$$

where $S_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(w) - \bar{Y}(w)\right)^2$ for $w = 0, 1$ and $S_{01}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - \tau\right)^2$.

Alternatively, we can also go throuhg the derivations again:

$$\begin{aligned}
\mathrm{Var}\left(\hat{\tau}_2\right) &= \mathrm{Var}\left(\hat{\bar{Y}}_1\right) + \mathrm{Var}\left(\hat{\bar{Y}}_0\right) - 2\,\mathrm{Cov}\left(\hat{\bar{Y}}_1, \hat{\bar{Y}}_0\right) \\
&= \frac{1}{n_1^2}\sum_{i,j=1}^{N} Y_i(1)Y_j(1)\,\mathrm{Cov}\left(W_i, W_j\right) + \frac{1}{n_0^2}\sum_{i,j=1}^{N} Y_i(0)Y_j(0)\,\mathrm{Cov}\left(W_i, W_j\right) - \frac{2}{n_1 n_0}\sum_{i,j=1}^{N} Y_i(1)Y_i(0)\,\mathrm{Cov}\left(W_i, W_j\right) \\
&= \frac{n_0}{N n_1} S_1^2 + \frac{n_1}{N n_0} S_0^2 - \frac{S_{01}^2 - S_1^2 - S_0^2}{N} \\
&= \frac{n_0 + n_1}{N n_1} S_1^2 + \frac{n_1 + n_0}{N n_0} S_0^2 - \frac{S_{01}^2}{N} \\
&= \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_{01}^2}{N}
\end{aligned}$$

where $\mathrm{Var}\left(\hat{\bar{Y}}_1\right) = \frac{n_0}{N n_1} S_1^2$ with $S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - \bar{Y}_1\right)^2$ (and analogously for the untreated) and

$$\begin{aligned}
\mathrm{Cov}\left(\hat{\bar{Y}}_1, \hat{\bar{Y}}_0\right) &= -\frac{1}{n_1 n_0}\sum_{i=1}^{N}\sum_{j=1}^{N} Y_i(1)Y_j(0)\,\mathrm{Cov}\left(W_i, W_j\right) \\
&= -\frac{1}{n_1 n_0}\sum_{i=1}^{N}\left[Y_i(1)Y_i(0)\frac{n_1 n_0}{N^2} - \sum_{j\neq i} Y_i(1)Y_j(0)\frac{n_1 n_0}{N^2(N-1)}\right] \\
&= -\frac{1}{N^2}\sum_{i=1}^{N}\left[Y_i(1)Y_i(0) - \frac{1}{N-1}\sum_{j\neq i} Y_i(1)Y_j(0)\right] = \frac{S_{01}^2 - S_1^2 - S_0^2}{2N}
\end{aligned}$$

with $S_{01}^2 = \frac{N}{N-1}\frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - \tau\right)^2$.

2. First estimator:
The only difference is that $Y_i(1), Y_i(0)$ are now random as well. For the first estimator:

$$\begin{aligned}
\mathrm{Var}\left(\hat{\tau}_1\right) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i W_i}{p} - \frac{\left(1 - W_i\right)Y_i}{1-p}\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}\left(\frac{Y_i W_i}{p} - \frac{\left(1 - W_i\right)Y_i}{1-p}\right)
\end{aligned}$$

where

$$\begin{aligned}\operatorname{Var}(\tau_1) &= \operatorname{Var}\left(\frac{W_i Y_i}{p} - \frac{(1-W_i)\,Y_i}{1-p}\right) \\ &= \frac{1}{p^2}\operatorname{Var}(W_i Y_i) + \left(\frac{1}{1-p}\right)^2 \operatorname{Var}((1-W_i)\,Y_i) - \frac{2}{p(1-p)}\operatorname{Cov}(W_i Y_i, (1-W_i)\,Y_i)\end{aligned}$$

where

$$\begin{aligned}\operatorname{Var}(W_i Y_i) &= pE\left[Y_i(1)^2\right] - p^2 E\left[Y_i(1)\right]^2 \\ \operatorname{Var}((1-W_i)\,Y_i) &= (1-p)E\left[Y_i(0)^2\right] - (1-p)^2 E\left[Y_i(0)\right]^2 \\ \operatorname{Cov}(W_i Y_i, (1-W_i)\,Y_i) &= -p(1-p)E\left[Y_i(1)\right]E\left[Y_i(0)\right]\end{aligned}$$

which yields:

$$\operatorname{Var}(\hat{\tau}_1) = \frac{1}{N}\left[\frac{E[Y(1)^2]}{p} - E[Y(1)]^2 + \frac{E[Y(0)^2]}{1-p} - E[Y(0)]^2 + 2E[Y(1)]E[Y(0)]\right]$$

Second estimator:
For the second estimator, we learned in the lecture that the S_{01}^2/N term is exactly the difference between the finite sample and population variance. That is, the variance is given by:

$$\operatorname{Var}(\hat{\tau}_2) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$$

3. If we have covariates and care about the population average effect, We might be able to improve upon the second estimator by what we called *ex-post stratification* in the lecture. That is, we estimate τ via a regression $Y_i = \tau W_i + \beta X_i + \epsilon_i$ (where we assume X_i contains a constant column to ignore the intercept). By the Frish-Waugh-Lovell theorem, we can equivalently perform first-step regressions of Y_i on X_i and W_i on X_i, and regress the resulting residuals $\tilde{Y} = (I - X(X'X)^{-1}X')Y$ on those of $\tilde{W} = (I - X(X'X)^{-1}X')W$ in a second step.

Randomized Experiment (Final Exam 2023, #1)

Suppose we have an infinitely large population. We randomly sample N units from this population. We consider an experiment where we randomly assign each unit to a binary treatment (independtly for each unit), with the probability of being in the treatment group equal to p. Let $W_i \in \{0, 1\}$ be the binary treatment. Let $Y_i(0)$ and $Y_i(1)$ be the potential outcomes (assuming SUTVA, or no interference) and let $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ be the population average effect of the treatment and let $\tau_s = \sum_{i=1}^N (Y_i(1) - Y_i(0))/N$ be the sample average treatment effect.
(a) Show that

$$Y_i \frac{W_i - p}{p(1-p)}$$

conditional on the potential outcomes is unbiased for the unit-level treatment effect $\tau_i = Y_i(1) - Y_i(0)$.
(b) Is the average

$$\frac{1}{N}\sum_{i=1}^{N} Y_i \frac{W_i - p}{p(1-p)}$$

a good estimator for τ?

Solution Sketch

(a) We find that

$$\begin{aligned}
\mathbb{E}[Y_i \frac{W_i - p}{p(1-p)}] &= \mathbb{E}[Y_i \frac{W_i - p}{p(1-p)} | W_i = 1] \cdot p(W_i = 1) + \mathbb{E}[Y_i \frac{W_i - p}{p(1-p)} | W_i = 0] \cdot p(W_i = 0) \\
&= \mathbb{E}[Y_i(1) \frac{W_i - p}{p(1-p)}] \cdot p + \mathbb{E}[Y_i(0) \frac{W_i - p}{p(1-p)}] \cdot (1-p) \\
&= Y_i(1) \frac{1-p}{p(1-p)} \cdot p + Y_i(0) \frac{0-p}{p(1-p)} \cdot (1-p) \\
&= Y_i(1) - Y_i(0) = \tau_i
\end{aligned}$$

Thus, $Y_i \frac{W_i - p}{p(1-p)}$ is unbiased for the individual treatment effet τ_i.

(b)

- The estimator $\hat{\tau}$ is unbiased for the average treatment effect which is an important property we care about.

$$\mathbb{E}[\hat{\tau}] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[Y_i \frac{W_i - p}{p(1-p)}] = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0)) = \tau.$$

- However, we know that it is not semi-parametrically efficient. This estimator is simply the AIPW estimator, thus we know that this infeasible version has a higher variance than the simple DiM estimator.

Power Calculations

Suppose we plan to test whether the mean of a population is different from 0 using a t-statistic, with a $\alpha = 0.05$ significance level. Assume that the variance of the observations is known to be $\sigma^2 = 100$. How many (independent) observations N do we need to have power $\beta = 0.9$ when the true mean is $\mu = 1$?

Solution Sketch

(a) **Note:** You can skip this entire derivation and simply use the formula from the lecture slides as well. This is useful in case you want to deepen your udnerstanding of where the formula is coming from.

Denote the outcome variable by Y_i. We estimate the population mean using the sample mean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. By the central limit theorem,

$$\sqrt{n}(\hat{\mu} - \mu) \to^d N(0, \sigma^2)$$

Our test statistic is

$$T = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n} \frac{\hat{\mu} - \mu_0}{\sigma}$$

where $\mu_0 = 0$ is the mean we want to test for.

If the true mean actually satisfies $\mu = \mu_0$, then we have, by Slutsky's theorem,

$$T \to^d N(0,1)$$

So we have that

$$\Pr(c_{\alpha/2} < T < c_{1-\alpha/2}) \approx \alpha$$

where c_α is the α quantile of the standard normal distribution. The terms are only approximately equal because it relies on the asymptotic distribution of T. For finite sample sizes, the probability that T falls between the two critical values can be (slightly) different from the asymptotic level. Hence, our test will reject if $T < c_{\alpha/2}$ or $T > c_{1-\alpha/2}$.
We want that our test has a power of $\beta = 0.9$ against the alternative $\mu = 1$:

$$\Pr(reject) = \beta = 0.9$$

when $\mu = 1$. Since the events $T < c_{\alpha/2}$ and $T > c_{1-\alpha/2}$ are disjoint (cannot happen at the same time), the probability that at least one of the two happens is equal to the sum of the individual probabilities of the events. Hence we need

$$\Pr(T < c_{\alpha/2}) + \Pr(T > c_{1-\alpha/2}) = \beta$$

What are these probabilities in the expression above?
We know that T asymptotically has a standard normal distribution *if* $\mu = \mu_0 = 0$. However, now we want to know what happens when $\mu = 1 \neq \mu_0$. What we do know:

$$\tilde{T} = \sqrt{n}\frac{\hat{\mu} - \mu}{\sigma} \to^d N(0,1)$$

where the μ above is the true population mean. The solution will therefore be to rewrite T to look more like the expression we have above. Let us inspect just one of the probabilities for a moment:

$$\begin{aligned}
\Pr(T > c_{1-\alpha/2}) &= \Pr(\sqrt{n}\frac{\hat{\mu} - \mu_0}{\sigma} > c_{1-\alpha/2}) \\
&= \Pr(\sqrt{n}\frac{\hat{\mu}}{\sigma} - \sqrt{n}\frac{\mu_0}{\sigma} > c_{1-\alpha/2}) \\
&= \Pr(\sqrt{n}\frac{\hat{\mu}}{\sigma} > c_{1-\alpha/2} + \sqrt{n}\frac{\mu_0}{\sigma}) \\
&= \Pr(\sqrt{n}\frac{\hat{\mu}}{\sigma} - \sqrt{n}\frac{\mu}{\sigma} > c_{1-\alpha/2} + \sqrt{n}\frac{\mu_0}{\sigma} - \sqrt{n}\frac{\mu}{\sigma}) \\
&= \Pr(\tilde{T} > c_{1-\alpha/2} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma})
\end{aligned}$$

Note that $\tilde{T}$ has an asymptotic normal distribution, and everything on the right-hand-side is non-random. Hence $\Pr(\tilde{T} > x) = 1 - \Pr(\tilde{T} \leq x) = 1 - \Phi(x)$ where $\Phi(x)$ is the cdf of the standard normal distribution.
We can do something very similar with the other probability. We get

$$\Phi(c_{\alpha/2} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}) + 1 - \Phi(c_{1-\alpha/2} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}) = \beta$$

We could now plug in $\alpha = 0.05$, $\mu_0 = 0$, $\mu = 1$, $\beta = 0.8$ and numerically solve the equation for n.

An additional approximation can make the formula (and solving for n or other parameters) even easier. Note that we are looking at $\mu = 1 > \mu_0 = 1$. Hence the first probability satisfies

$$\Phi(c_{\alpha/2} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}) \to 0$$

because $c_{\alpha/2} + \sqrt{n}\frac{\mu_0-\mu}{\sigma} \to -\infty$. Similarly, the derivative of the first term w.r.t. n also converges to 0:

$$\phi(c_{\alpha/2} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma})\frac{\mu_0 - \mu}{\sigma}\frac{1}{\sqrt{n}} \to 0$$

because the first and the last term converge to 0. Note that in this example, the second Φ term converges to 1 (the test has asymptotic power of 1 under any fixed alternative). The point is therefore that for the n that gives us a power of β, the first term does not matter much / is dominated by the second term. This could be seen in the density plot that we drew during the section.
When the sample size n needed is large, the first term hence is close to 0 and does not change much as we change n. We can therefore get a good approximation by ignoring the first term and only focusing on the second term. We then try to find n such that

$$\begin{aligned}
1 - \Phi(c_{1-\alpha/2} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}) &= \beta \\
\Phi(c_{1-\alpha/2} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}) &= 1 - \beta \\
c_{1-\alpha/2} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma} &= \Phi^{-1}(1 - \beta) \\
n &= \left((\Phi^{-1}(1 - \beta) - c_{1-\alpha/2})/\frac{\mu_0 - \mu}{\sigma}\right)^2
\end{aligned}$$

where Φ^{-1} is the inverse cdf of the standard normal distribution. Φ^{-1} is often easy to calculate (numerically) in software packages, while the original formula you might have to write your own code to solve for the parameter of interest.
When we use the last formula (which is equivalent to the formula from the lectures if you rewrite it), we get $n = 1050.7423$, so we would need 1051 observations to have power of at least 0.9.

Power Calculations

Suppose we want to run a completely randomized experiment on $N = 3000$ individuals. We plan to test whether the average treatment effect is different from 0 using the difference in means estimator, at a $\alpha = 0.05$ significance level. I want to achieve a power of $\beta = 0.8$ when the true treatment effect is $\tau = 1$.
(a) What should the variance of my outcome be for me to achieve my goal? (Let $\Phi^{-1}(0.8) = 0.8416, \Phi^{-1}(0.9) = 1.2816, \Phi^{-1}(0.95) = 1.6449$ and $\Phi^{-1}(0.975) = 1.9600$.)
(b) Suppose I observe some pre-treatment data and find that $\sigma^2 = 64$. Based on your answer to (a), will I be able to achieve the desired power in my experiment? Explain why or why not.

Solution Sketch

(a) We assume that we treat an equal number of units, i.e. $\gamma = \frac{1}{2}$. Note that this is the optimally powered experiment, so often our fraction of choice in practice.

Given this, we have all the parameters we need. We use the formula for calculating the minimum sample size N from lecture slide 39 from lecture 2 to find σ^2. We find

$$
\begin{aligned}
N &= \frac{\left(\Phi^{-1}(\beta) + \Phi^{-1}(1-\alpha/2)\right)^2}{(\tau_0^2/\sigma^2)\cdot\gamma\cdot(1-\gamma)} \\
\implies 3000 &= \frac{\left(\left(\Phi^{-1}(0.8) + \Phi^{-1}(1-0.05/2)\right)\right)^2}{\frac{1}{\sigma^2}\frac{1}{2}(1-\frac{1}{2})} \\
\iff \frac{1}{\sigma^2} &= \frac{4 * \left(\left(\Phi^{-1}(0.8) + \Phi^{-1}(1-0.05/2)\right)\right)^2}{3000} \\
\iff \sigma^2 &= \frac{3000}{4 * \left(\left(\Phi^{-1}(0.8) + \Phi^{-1}(1-0.05/2)\right)\right)^2} \\
\iff \sigma^2 &= 95.555
\end{aligned}
$$

(b) Yes! The power will actually be higher than $\beta = 0.8$. Since our data has a smaller variance, our treatment effect is now larger compared to the effect that we are trying to detect. Thus, it will be easier to detect it as the chance of the treatment effect being τ due to chance (/variation in the outcome) is lower. You will alternatively arrive at the same solution if you look at the formula and think about what happens when the variances decreases and everything else but the power stays constant.