Chapter 4

Randomized Studies

4.1 Basics

A randomized study is regarded as one of the most valued research methodologies for examining the efficacy or effectiveness of interventions. A double-blind randomized controlled study/trial is the "gold standard" in treatment and intervention evaluation.

Strengths of a Randomized Study

- Ability to evaluate causal relationships
- High internal validity (the extent to which differences between intervention and control groups can be attributed to the intervention), due to minimized bias within the study
- Investigator control over patient exposure
- Prospective data collection, which allows for standardization of exposure and outcome collection
- Attempted balance, through randomization, between known and unknown confounding factors between groups

Limitations of a Randomized Study

- Higher cost than observational studies
- Limited external validity and generalizability, due to strict inclusion and exclusion criteria and application of interventions by protocol
- Ethical considerations related to assigning patients to particular care approaches
- Generally shorter-duration follow-up than observational studies
- Inefficiency of detection of rare or delayed outcomes, due to smaller sample size and shorter-duration follow-up than observational studies

What do we mean by randomization?

Randomization is the process of assigning participants to treatment and control groups, assuming that each participant has an equal chance of being assigned to any group.

Since Fisher first introduced the idea of randomization in a 1926 agricultural study, the academic community has deemed randomization an essential tool for unbiased comparisons of treatment groups. Five years after Fisher's introductory paper, the first randomized clinical trial involving tuberculosis was conducted. A total of 24 participants were paired (i.e., 12 comparable pairs), and by a flip of a coin, each participant within the pair was assigned to either the control or treatment group.

Simple randomization is randomization based on a single sequence of random assignments. This technique maintains complete randomness of the assignment of a person to a particular group. The most common and basic method of simple randomization is flipping a coin.

- This randomization approach is simple and easy to implement in a clinical trial.
- In large trials (n $>$ 200), simple randomization can be trusted to generate similar numbers of participants among groups.
- However, randomization results could be problematic in relatively small sample size clinical trials (n $<$ 100), resulting in an unequal number of participants among groups.
- For example, using a coin toss with a small sample size ($n = 10$) may result in an imbalance such that 7 participants are assigned to the control group and 3 to the treatment group (Figure 4.1).

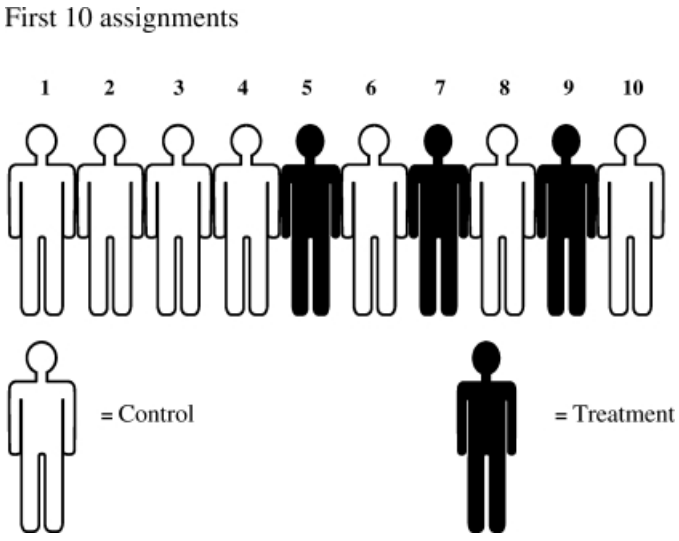

Figure 4.1: Illustration of a simple randomization

Block randomization is designed to randomize participants into groups that result in equal sample sizes. This method is used to ensure a balance in sample size across groups over time. Blocks are small and balanced with predetermined group assignments, which keeps the numbers of participants in each group similar at all times (Figure 4.2).

Stratified randomization addresses the need to control and balance the influence of covariates.

- This method can be used to achieve balance among groups in terms of participants' baseline characteristics (covariates).
- Specific covariates must be identified by the researcher who understands the potential influence each covariate has on the dependent variable.
- Stratified randomization is achieved by generating a separate block for each combination of covariates, and participants are assigned to the appropriate block of covariates.
- After all participants have been identified and assigned into blocks, simple randomization occurs within each block to assign participants to one of the groups.

In **covariate adaptive randomization** a new participant is sequentially assigned to a particular treatment group by taking into account the specific covariates and previous assignments of participants.

Cluster (or group) randomization is when the unit of randomization is a group rather than an individual. Such groups might be schools, clinics, worksites, communities, or other units. Group randomization to treatment can be an efficient strategy when an intervention is difficult to implement on an individual level without the risk of contamination, such as interventions that affect environments.

There are many other types of complex randomization methods.

What is noncompliance?

Unfortunately, all studies have potential issues of noncompliance. Noncompliance means that the individual does not comply with their assigned group protocol. In the table below, individuals in a and d are compliers; individuals in c and b are noncompliers.

Patients' noncompliance with their assigned treatment will undermine randomization and potentially bias the estimate of the treatment effect.

When there is noncompliance, there are different approaches that can be used to analyze the data.

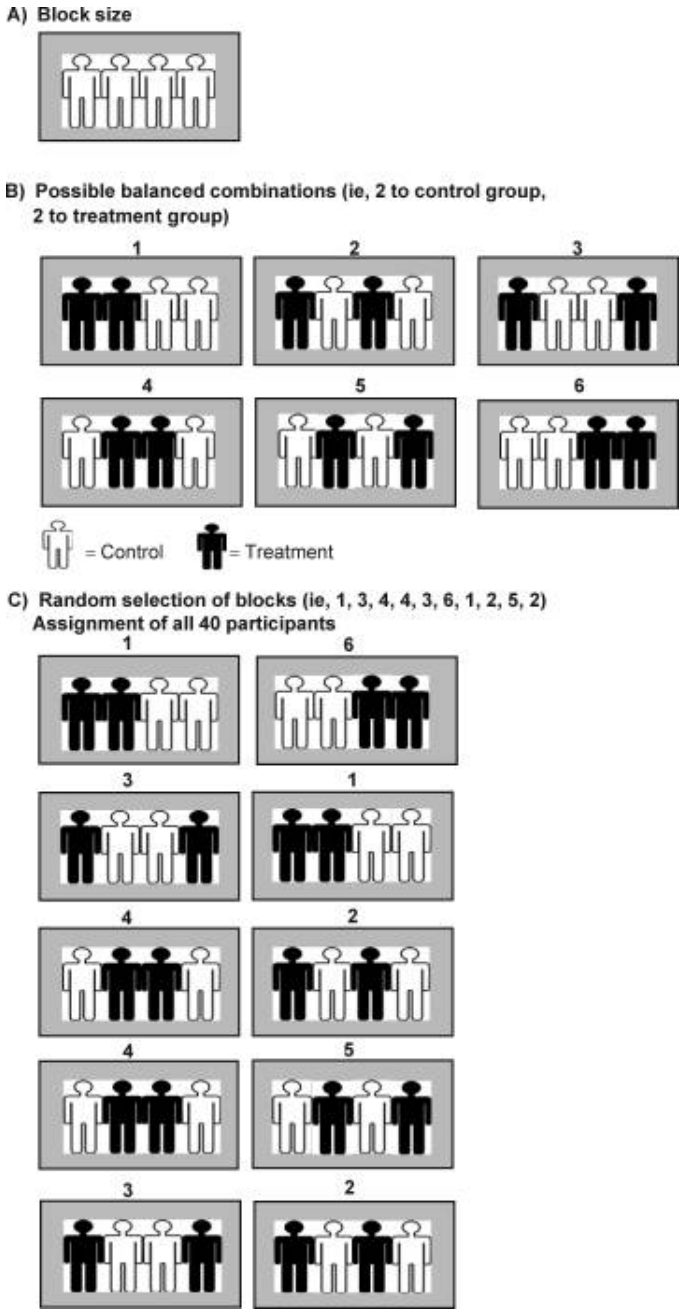

Figure 4.2: Illustration of block randomization

	Took Treatment	Did not take Treatment	
Assigned to Treatment Group	Compliers (a)	Noncompliers (b)	a+b
Assigned to Control Group	Noncompliers (c)	Compliers (d)	c+d
	a+c	b+d	

- **Intent-to-treat (ITT)** analysis means that the data are analyzed as assigned, ignoring any noncompliance issues. That is, you analyze individuals assigned to the treatment group even if you know they did not take the treatment; you analyze individuals assigned to the control group even if you know they took the treatment. This analysis compares (a+b) vs. (c+d).
  - ITT analysis is considered the gold standard. Any other approach **breaks** randomization. However, it substantially reduces your power and can result in a null result when there is truly an effect.
- An **As-treated (AT)** analysis means that the data are analyzed according to the treatment the individual actually received. If someone was assigned to the control group, but they took the treatment, then they are analyzed as if they are in the treated group. This is extremely problematic. It completely breaks the initial randomization. This analysis compares (a+c) vs. (b+d).
- A **Per-protocol (PP)** analysis excludes patients who did not fully comply with the treatment protocol. That is, you analyze individuals assigned to the treatment group who took the treatment; you analyze individuals assigned to the control group who did not take the treatment. This analysis compares a vs. d.
  - You **ignore** all noncompliers. This is also extremely problematic as it also breaks randomization. The noncompliers may differ from the compliers in ways that are not measurable. There is no assurance that the resulting groups are balanced on unmeasured confounders.
- Another class of methods to deal with noncompliance includes instrumental variable (IV) and complier average causal effect (CACE) approaches.
  - The IV approach uses the randomization indicator as an IV to adjust for the proportion of noncompliant patients.
  - The CACE method estimates the treatment effect among compliers, but using principal stratification; it is not the same as a PP analysis. There are two general approaches for estimation: maximum likelihood approach by expectation-maximization (EM) algorithm and Bayesian estimation.

4.2 Power and Sample Size

The **power** of an analysis/study is the probability that the analysis will **detect an effect when there truly is an effect.**

Recall that the **Type 1 error** is the probability that the analysis will detect an effect **when there truly is no effect.** This is α and is usually set to 0.05.

Take a look at Figure 4.3 which plots the distribution of values observed from a treated group and a control group. Both plots reflect a study with $n = 100$ in each group where the true mean in the treated group is 2 and true mean in the control group is 1. In the plot on the left, the data come from a distribution with standard deviation = 0.2; in the plot on the right, the standard deviation is 2. For which picture do you think it would be "easier" to detect a difference between the two groups?

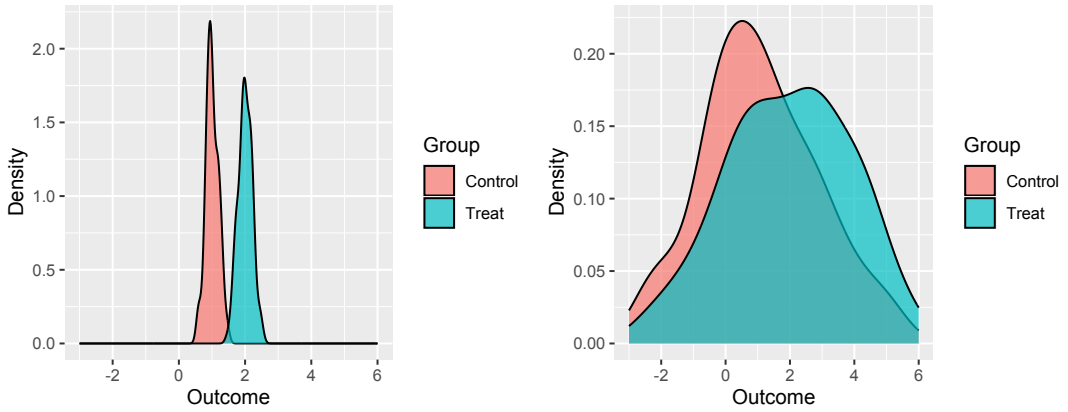

Figure 4.3: Two studies, both with $n = 100$ and same means, but different standard deviations

Studies can be **underpowered**, **overpowered**, or just right.

What is Effect Size?

Effect size is a quantitative measure of the magnitude of an effect. The larger the effect size the stronger the effect or the stronger the relationship between two variables.

The calculation of an effect size differs depending on the type of outcome. For a continuous outcome, effect size is usually quantified using **Cohen's d**. Cohen's d is defined as the difference between two means divided by the standard deviation of the data,

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

where $\bar{x}_1$ is the mean of the outcome in group 1, $\bar{x}_2$ is the mean of the outcome in group 2, and s is the standard deviation of the data. For s, it is most common to use the pooled standard deviation

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where n_1 and n_2 are the sample sizes in group 1 and group 2, respectively, and s_1 and s_2 are the standard deviations in each group.

The table below describes the magnitude of specific values of Cohen's d.

Effect Size	**Cohen's d**
Small	0.20
Medium	0.50
Large	0.80

When you are designing your study, **you want to have sufficient power to detect a meaningful effect size.** Whether small, medium, or large is meaningful depends on the setting. Usually, the goal is to detect small effects to justify carrying out the study. However, if the treatment is extremely invasive or toxic, it may be that only a large treatment effect is of clinical interest.

For example, you may want to say that your study has 80% power to detect a small effect size (Cohen's d = 0.2). This means that assuming there is a true effect/difference that is 0.2 or greater in effect size magnitude, the probability that your statistical test for the difference will detect the effect (e.g., conclude that the difference is significant) is 0.80.

How to Calculate Power?

The power is specific to the test you plan to use in your analysis. Suppose in the example above, we plan to use a two-sample t-test to test for a difference between the two groups.

Specifically, we are testing the following:

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

which is equivalent to

$$H_0 : \delta = 0$$

$$H_A : \delta \neq 0$$

where μ_1 and μ_2 are the true means in group 1 and group 2, respectively, and $\delta = \mu_1 - \mu_2$.

We would calculate the t-statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}}$$

and we would reject H_0 if $|t| > T_{1-\alpha/2,v}$, where $T_{1-\alpha/2,v}$ is the critical value with v degrees of freedom (Figure 4.4).

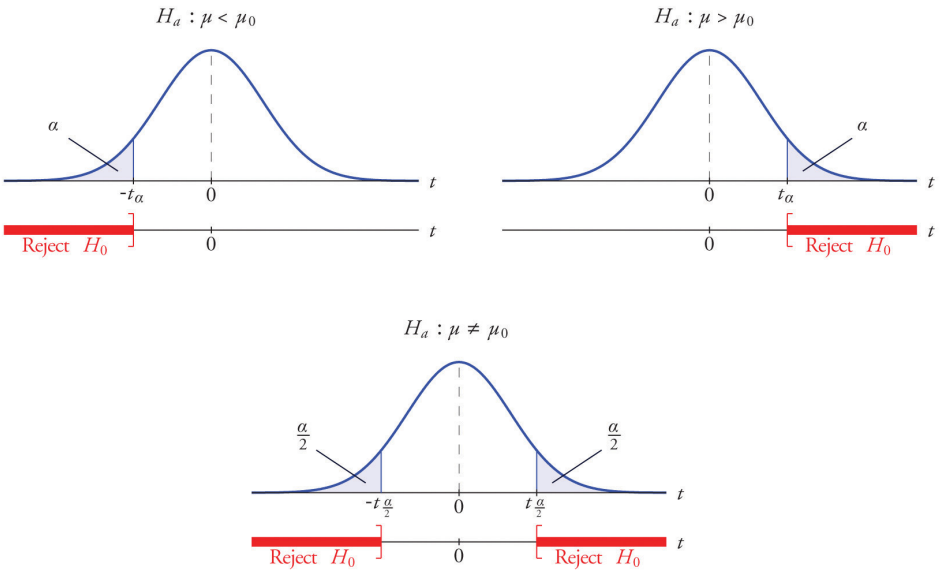

Figure 4.4: t-test Rejection Region

The power for this test, assuming the true effect is δ^*, can be shown to be:

$$\begin{aligned} \text{Power} &= 1 - \Phi\left(T_{1-\alpha/2,v} - \frac{\delta^*}{s\sqrt{1/n_1 + 1/n_2}}\right) \\ &= 1 - \Phi\left(T_{1-\alpha/2,v} - \frac{d}{\sqrt{1/n_1 + 1/n_2}}\right) \end{aligned}$$

where Φ is the Normal cumulative distribution function (Figure 4.5).

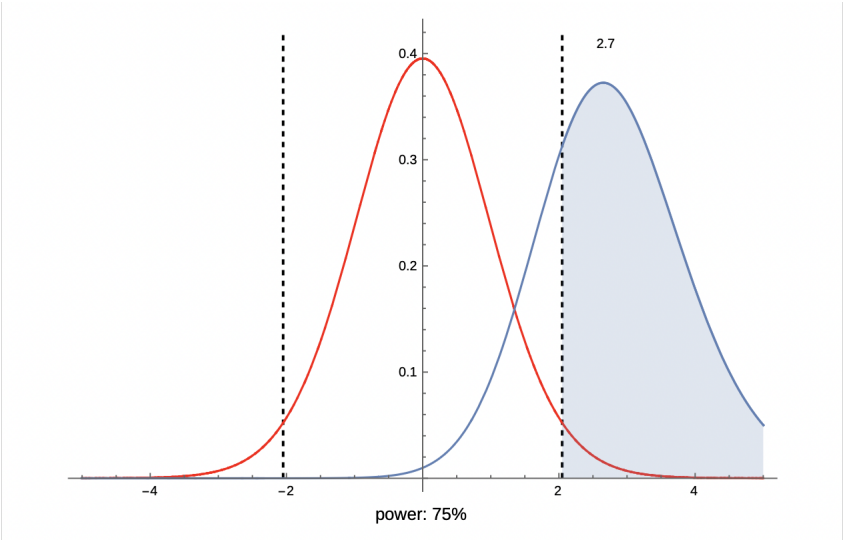

Figure 4.5: t-test Power, $\delta = 2.7$

Power depends on the sample size and the effect size. The needed sample size depends on the desired power and the expected effect size.

4.3 Power Calculations for a Complex Design/Analysis

It is often the case that your design or analysis is not a simple two sample t-test. For more complex study designs or analysis plans, formulas may not exist or may be very difficult to adapt to a particular setting.

In these situations, one approach is to conduct **simulation-based power calculations.**

- The basic idea is that you repeatedly simulate your entire experiment and calculate the proportion of experiments in which the null hypothesis is rejected; this is your estimated power. (Note that we did this in R with the t-test.)
- Simulating your entire experiment will typically involve generating a dataset, and then running an analysis that involves a hypothesis test. Randomness is usually introduced into the process through the dataset generation, although sometimes you will fix a population dataset and induce randomization by taking samples from that population.
- Often, the most difficult part is to simulate a dataset that accurately reflects the nuances (e.g. the correlation structure) of your real dataset.
- If you want to calculate sample size at a fixed power level (e.g. 90%), you can use a "guess and check" approach. With this approach, you select a sample size n and run the simulation to estimate your power. If power is estimated to be lower than 90%, n^* that is larger than n and run the simulation again. You repeat this procedure until the estimated power is roughly 90%.

Another option is to **simplify** your setting as much as possible to get an approximate power calculation. For example, can your primary hypothesis be considered as a test between two independent groups? If so, you may be able to frame it as a t-test.

A useful tool here is the concept of the **effective sample size (ESS)**. The effective sample size is the sample size of your data after you account for complexities that exist either in the data, or are induced by your analysis.

The effective sample size is defined as

$$ESS = \frac{n}{DEF}$$

where n is your true sample size and DEF stands for the **design effect** (see Figure 4.6).

The design effect is usually (but not always) greater than 1 and thus will **penalize** your sample size.

- Essentially, the design effect quantifies how much more the variance has increased (or decreased, in some cases) because our sample was drawn and adjusted to a specific sampling design (e.g., using weights, or other measures) as it would be if instead the sample was a simple random sample.

- There are many ways of calculating DEF depending on the parameter of interest, the estimator used, and the sampling design.

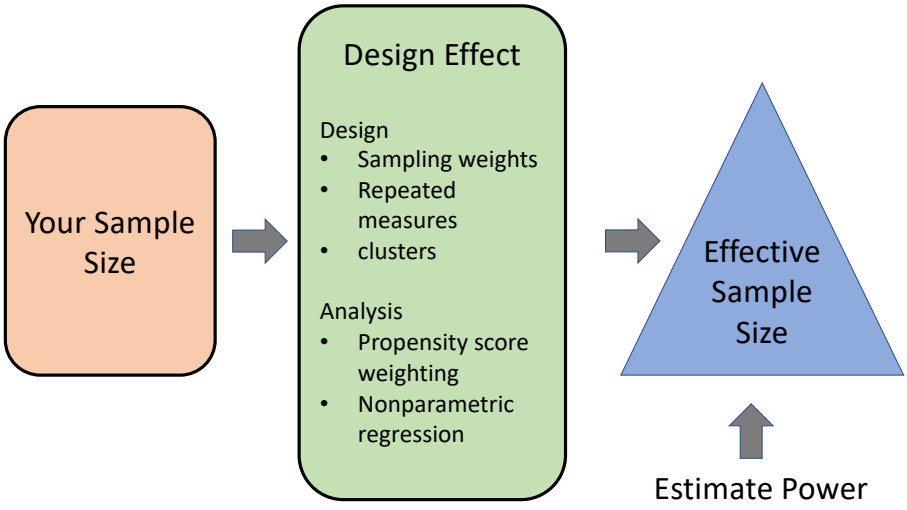

Figure 4.6: Sample Size vs. Effective Sample Size

Many settings and analysis can induce a design effect. We will discuss two that we have covered in class: **propensity score weighting and longitudinal data.**

- Using **propensity score weights** induces a design effect. In fact, weights of any kind induce a design effect (e.g. nonresponse weights, stratification weights, survey sampling weights). When you use propensity score weighting, you have to pay a price in terms of precision.

- Suppose that in your analysis you are using a weight, w_i, for each person i in your sample. The DEF based on Kish's design effect is

$$DEF_w = \frac{n\sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2}$$

  and thus the effective sample size is

$$ESS = \frac{n}{DEF_w} = \frac{n}{\frac{n\sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2}} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$$

- Repeated measures on the same individual or unit, as we have in **longitudinal data**, induces a design effect which is based on the intraclass correlation (ICC). Recall that the ICC was a quantity that was between 0 and 1 describes how strongly measurements in the same group/person resemble each other.

- The DEF for such "clustered" data is

$$DEF_c = 1 + (n^* - 1)ICC$$

  where n^* is the size of the clusters. The effective sample size is

$$ESS = \frac{n}{DEF_c} = \frac{n}{1 + (n^* - 1)ICC}$$

Once you have determined your effective sample size, it is often possible to calculate power assuming a simplified test (e.g. a t-test) but the calculation is done with the effective sample size.

References:

Houle, S. (2015). An introduction to the fundamentals of randomized controlled trials in pharmacy research. The Canadian journal of hospital pharmacy, 68(1), 28.

Fisher, R. A. (1992). The arrangement of field experiments. Breakthroughs in statistics: Methodology and distribution, 82-91.

Kang, M., Ragan, B. G., & Park, J. H. (2008). Issues in outcomes research: an overview of randomization techniques for clinical trials. Journal of athletic training, 43(2), 215-221.

Ye, C., Beyene, J., Browne, G., & Thabane, L. (2014). Estimating treatment effects in randomized controlled trials with non-compliance: a simulation study. BMJ open, 4(6), e005362.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences.

Kish, Leslie (1965). Survey Sampling. New York: John Wiley & Sons, Inc. ISBN 0-471-10949-5.