Chapter 7

Experimental Design

This chapter focuses on how to design experiments. We bring the assumptions we've learned into play and discuss key ideas and principles of good experimental design. Good design concepts are best illustrated by examples, so we provide many in this chapter, mostly involving clinical trials or improving education in developing countries.

■ 7.1 Core elements of good design

While experimental design is a broad topic that is often difficult to get right, there are a few guiding principles that all good designs are built on top of:

1. **Replication**: Any good experiment should be reproducible, and in particular, replication should yield similar results. Shockingly, many published scientific papers fail at this tenet![1] Meanwhile, anecdotal evidence is not scientific proof, and we've seen so far that most of the methods we've discussed improve as the number of samples increases. This is often difficult to achieve because of cost or time constraints in a study: while gathering an infinite amount of data might be theoretically ideal, it's practically impossible.

2. **Comparison/control/baseline**: In any experiment where you're measuring the effect of a treatment, it's impossible to assess that effect without having a reference value.

   For example, a remedial summer program in a poor school district might result in little to no improvement in student performance between the beginning and end of the summer. But what would have happened if we hadn't intervened? In disadvantaged schools, academic performance can sometimes *decline* over the summer – in this case, our intervention would be classified as an improvement!

   As another example, when an experimental medical treatment undergoes clinical trials, it is standard to compare the treatment against a **placebo**, which refers to a sham treatment that is advertised as effective. In these trials, the goal is to show that the experimental treatment significantly outperforms the placebo. The exception to using a placebo in these cases is when the disease is extremely debilitating where it makes

[1] See The Economist, "Unreliable research: trouble at the lab", October 19, 2013.

sense to compare against a standard treatment currently in use (but that perhaps isn't very effective) rather than a placebo. Despite a placebo being a sham treatment, it can often actually make a subject feel better compared to not giving any treatment at all! This is explained in the next example panel.

Example: The Placebo and Hawthorne Effects

Beware the **placebo effect**! When you apply a treatment to a subject, that treatment may be ineffective, but can still produce a significant effect simply due to the existence of the treatment. For example, a fake placebo surgery can actually do as well as a common knee surgery![a] And giving people nonalcoholic drinks but telling them that the drinks are alcoholic can result in a decline in memory powers![b]

Because of the placebo effect, development of medical treatments demands the stronger standard of outperforming a placebo rather than outperforming not giving any treatment at all since a placebo alone could already result in a startling improvement in a test subject's condition, often largely due to psychological factors.

To examine the benefit of a placebo, an experiment could have a control group that receives no treatment, a placebo group that receives a sham treatment, and a treatment group that recieves the actual treatment under study. By doing this, the experimenter can measure both the effect of the placebo over the baseline and the effect of the treatment over the placebo.

Closely related to the placebo effect is the **Hawthorne effect**: in a behavioral study, the behavior of subjects might be due to their reaction to being studied. In a famous experiment in the early 20th century, a factory called the Hawthorne Works wanted to measure the effect of lighting on productivity. An experimental group had their light bulbs changed, and the experimenters wanted to measure the effect on productivity. A control group saw workers change their bulbs, but the new bulbs were identical to the old ones. However, both groups improved after the "new" bulbs were put in: the control group improved purely due to their perception of an effect.

These effects can turn up where you least expect them to! For example, suppose you're designing an experiment to measure the effect of fertilizers on a farm. While the plants probably aren't vulnerable to the placebo effect, the farmers could be. A farmer whose field is fertilized might work harder and be more motivated simply by being part of the study. As a result, in such a study, it might be a good idea to have a placebo farm that receives plain dirt. There are also other confounding factors: seemingly uninteresting quantities such as the overall moisture level of the fertilizer may have an impact on the result, so it's important to make the placebo group as equalized as possible!

[a] See Baylor College Of Medicine, "Study Finds Common Knee Surgery No Better Than Placebo," ScienceDaily, July 12, 2002.

[b] See BBC News, "Being drunk 'a trick of the mind"', January 7, 2003.

3. **Blocking/controlling for confounds**: In an experiment with possible external sources of variability, it's always best to control for these factors (recall Simpson's paradox from Chapter 5!). Controlling for confounds is best achieved through *blocked design*, where we divide subjects into groups corresponding to levels of a confounding factor, and repeat the experiment for each group. By accounting for the effect of confounding variables, we can avoid being misled by our data.

   For example, an educational intervention program may have different effects on students of different gender. If we aren't interested in the confounding effect of gender, then we can analyze the blocks separately or even include this confound in our analysis (e.g., using ANCOVA).

4. **Randomization**: Most theoretical analysis assumes that data points are independent. Randomization is often the key ingredient to satisfying this assumption! For example, it's better to have randomly selected data points, and to randomly assign those data points to different groups/treatments, and so on. Skipping randomization can often lead to *bias* in data!

A general rule of thumb, attributed to the famous statistician George Box, is to "block what you can, randomize what you cannot."

We'll see concrete examples of how these principles come into play.

■ 7.2 Gathering Samples

A critical part of any experiment is gathering samples or data points. In all of these examples, we assume some underlying "population". For example, if you're conducting a poll for all of the U.S., then your population could be all U.S. residents. If you're studying the effect of a new pilot bilingual immersion program at a high school, then your population would be students from that high school.

Here are a few ways to gather data:

1. **Simple Random Sample** (SRS): In a simple random sample, we draw members of the population uniformly at random *without replacement*. This is like putting the names of everyone in the population into a hat and then drawing a few names out of the hat, assuming of course that the drawing is fair. The "without replacement" part just means that once a name has been drawn from the hat, we don't put it back into the hat. Effectively we can't draw the same name twice. We'll see in Section 7.3 that an SRS consists of points that are *not* independent! However, if the number of data points in the SRS is much smaller than the total population size, then we can safely treat the samples as approximately independent.

   Unfortunately, collecting an SRS is often difficult to carry out in practice: if we want to randomly sample the population of people in New York City with landline phones, then we can take a phonebook, choose names at random, and call them. But, if we're sampling the population of students in a developing country, it's almost impossible to find a list, let alone obtain access to people that are uniformly sampled.

   Another issue with simple random samples is that it's often difficult to make conclusions about smaller subpopulations. For example, if a particular subgroup is relatively small, a uniform sample may not capture any members of that group. For example, if we're trying to estimate the proportion of the population that has an extremely rare disease (say, 1 in a million), then chances are that from sampling, say, 1000 people, none of them are going to have the disease.

2. **Stratified Random Sample**: (First off, by convention, the abbreviation "SRS" refers to a simple random sample, and *not* a stratified random sample!) As for stratified random samples: suppose we know that the population consists of several different non-overlapping groups, and that there isn't much variation within each group. Then we can divide the population into these groups and within each group collect an SRS. These groups are called **strata**, with each group called a **stratum**.

   In a scheme known as *proportional allocation*, the number of subjects per stratum is usually chosen to match that stratum's true relative size in the population. For example, if 60% of the population of interest is female, and we have two strata, one per gender, then if we want our study to have a total of 1,000 people, proportional allocation would ask that we collect two separate SRSs, one with 600 women and one with 400 men.

   In *Neyman's optimal allocation*, the number of subjects per stratum is determined by both the stratum's true relative size in the population as well as the variance within the stratum. If either the relative size of the stratum in the population is larger, or the variance within the stratum is higher, then we'll collect more samples from this stratum. Formally, if W_ℓ is the true proportion of the population that is in stratum ℓ, and σ_ℓ is the true standard deviation within stratum ℓ, then Neyman's optimal allocation says that the size of the SRS for stratum ℓ should be proportional to $W_\ell \sigma_\ell$.

   This technique allows us to accurately measure the effects of small groups that may have otherwise been missed in an SRS over the whole population (i.e., without stratification). For example, we may want to sample the performance of students in different types of schools. If some school categories are larger (i.e., have more students) than others, then an SRS over the whole population may miss the small categories. A stratified sample would list the categories and sample randomly within each type of school.

3. **Cluster Sampling**: The two methods above require samples from either the entire population or every single stratum. This may not always be cost-effective or even feasible. Cluster sampling is based on the idea of dividing the population into natural, heterogeneous groups that are relatively similar to each other. Each group should be well-representative of the population. Instead of sampling from all of the groups, we'll randomly sample a few, and then do random sampling within each one. Since they're all similar to each other, then a random sample from one should be representative of a random sample of the population.

   For example, if we're polling a city, we might divide it up into city blocks. Then we randomly choose some number of blocks to sample. Finally, within each block, we collect an SRS. As long as there are no large differences between each block, and each block represents the overall city population well, then this technique is often more cost-effective than an SRS over the whole population.

However, all of these frameworks have issues:

- Getting an unbiased list of subjects to sample from, even within a stratum or a cluster, can often be difficult.

- We may have *non-response bias*: in a study of people such as a survey, there will almost always be people who choose not to respond. Unfortunately, different groups often have different non-response rates. For example, in an approval survey, more enthusiastic people are more likely to respond to questions, which can bias the results toward the extremes. As another example, a poll asking about the workload of students may run into non-response bias where overworked students are too busy to respond and, as a result, the collected responses may suggest that people work fewer hours than they actually do on average.
- For surveys, how questions are worded can make a huge difference in how people respond! We see this in the following example.

Example: Wording matters!

In October 2004, Stanley Presser ran a poll for *The New Yorker*, where half of respondents were asked "Do you think the United States should allow public speeches against democracy?" and the other half were asked the same question except with "allow" replaced by "forbid". Whereas 56% answered no to "forbidding", 39% answered yes to "allowing" despite the two answers corresponding to the same response.

Generally speaking, it is a good idea to word questions as neutrally as possible, and if the questions don't have some order dependence, to randomize their ordering.

The above issues often make it hard to extend conclusions beyond a study: any analysis we can do is only valid for the population that we sampled from. This is highlighted by the following example, which led to the downfall of the magazine *The Literary Digest*.

Example: The Literary Digest poll for the 1936 US Presidency

Republican candidate Alfred Landon was running against Democrat Franklin Delano Roosevelt. *The Literary Digest* projected that Landon would win by a huge margin: a 57% to 43% victory. The magazine had polled 10 million people and received a whopping 2.4 million responses! Yet Franklin Roosevelt won by a landslide, carrying 46 states while Landon only carried 2 states. The win wasn't just in the electoral college either: Roosevelt won 61% of the popular vote.

What had happened? Of course, a mix of things happened including non-response bias and likely wording issues, but the main issue was selection bias: the questionnaires were sent out to readers of *The Literary Digest*, those that were in a phone listing, and those on a listing of car owners. But all these lists contain more rich people than poor people, which led to a heavily skewed poll result.

In contrast, a Gallup poll that same year predicted that Roosevelt would receive 56% of the popular vote using only a sample size of 50,000, which turned out to be far more accurate than *The Literary Digest*'s poll results.

■ 7.3 Simple Random Samples: are samples really independent? (optional)

Suppose we have a population of N people and we're measuring their heights. Let the true population mean height be μ with variance σ^2. We draw a simple random sample (SRS) of size n from the population. Let $x_1, x_2, \ldots, x_n$ be the heights we measure of n people in the SRS. Let the sample average height be

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

First off, note that the x_i's are *not* independent! The intuition is as follows. Let's say the first person we sample from the population is Alice, so x_1 is Alice's height. Since we are sampling without replacement, the second height x_2 cannot come from Alice! So how we selected x_1 affects how we select x_2, which means they're not completely independent. This reasoning extends to all the x_i's.

The sample average (which, as a reminder, is a random variable) turns out to have mean and variance given by

$$\begin{aligned} \mathbb{E}[\bar{x}] &= \mu, \\ \text{var}[\bar{x}] &= \frac{\sigma^2}{n} \underbrace{\left(\frac{N-n}{N-1}\right)}_{\text{correction factor}}, \end{aligned}$$

where we note that there is a correction factor due to the samples not being independent.

If $n \ll N$, then the correction factor is approximately 1, and furthermore, the samples can be approximated as independent and as if they were drawn *with replacement*. The intuition is that if we're drawing n names from a hat of N names, and $n \ll N$, then even if each time we draw a name, we put the name back into the hat, the chance of us drawing the same name twice is negligible — whether we put the name back in the hat or not doesn't really affect the result!

Confidence intervals

For sufficiently large n, the sample average $\bar{x}$ is approximately normal, but *not* because of the central limit theorem which we saw earlier (recall that the theorem required that the random variables we're summing to be independent, which isn't the case here). Instead, one needs a fancier central limit theorem that tolerates some dependence between random variables. In any case, the approximate normality of $\bar{x}$ allows us to construct an approximate confidence interval for μ.

As with our earlier excursions into computing confidence intervals, what we need is an estimate of the standard error of our estimator $\bar{x}$. It turns out that an unbiased estimator for $\text{var}[\bar{x}]$ is:

$$\frac{s^2}{n}\left(\frac{N-n}{N}\right).$$

Thus, an estimate of the standard error is

$$s_\mu = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}},$$

from which we derive a 95% confidence interval for the mean height μ:

$$\bar{x} \pm 2s_\mu = \bar{x} \pm 2\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}.$$

As a reminder, the coefficient 2 comes from the fact that within 2 standard deviations of a standard normal random variable lies 95% of the probability mass centered around the mean.

If instead the x_i's had been binary random variables taken on values 0 or 1 (e.g., we ask each of n people a yes/no question, where "yes" is encoded as a 1), then one could show that an approximate 95% confidence interval is

$$\bar{x} \pm 2\sqrt{\frac{\bar{x}(1-\bar{x})}{n-1} \cdot \frac{N-n}{N}},$$

where in this context, note that $\bar{x}$ estimates the fraction of the population that has value 1 (e.g., the proportion of people who answer "yes" to a poll).

■ 7.4 Some sample designs

This section covers more experimental designs that are useful for more complex experiments.

■ 7.4.1 Paired tests and repeated measures

Whenever possible, if applying a treatment, it's best to have **paired** data, where we obtain measurements for each subject before and after treatment. As we saw with t-tests, paired tests often give us the most power.

A generalization of paired tests is **repeated measures** design. In such a design, we may have multiple (i.e., 2 or more) treatments, and each subject will receive all the treatments. This way, each subject can be thought of as its own control.

For example, suppose we measure the effect of caffeine (in the form of tea and coffee) on student performance. In a repeated measures design, each student would spend a month drinking coffee, a month drinking tea, and a month with no caffeine intake (for control). We may also want to add a month with a decaffeinated drink as a placebo. In such designs, it's important to randomize the order, and to be wary of temporal effects. In this example, stopping caffeine treatment might lead to worse performance due to withdrawal. As a result, it might be worthwhile to wait in between each "measure". We can sometimes model these temporal effects with *autocorrelation* models, where the errors are no longer assumed to be independent, but rather to depend on each other in sequence.

■ 7.4.2 Randomized complete block design

What do we do when we have multiple factors to block on? If the factors don't depend on each other, then we'll probably have the same number of sub-blocks with in each block. For example, in an experiment where we block on gender and handedness (left or right), we'll have left-handed and right-handed groups for men, and left-handed and right-handed groups for women. Such a design is called *complete*, because each sub-block is being tested. We'll focus here on cases where we have two blocking factors, although the ideas we'll discuss can be generalized. In a randomized complete block design, we may not have enough data points to replicate within sub-blocks, so we must assign different sub-blocks to different treatment conditions.

Example

For example, suppose we want to measure the effect of giving tablets to students in developing countries. Our experimental condition might be providing students with tablets and giving them an extra hour every day to use them. We would need a control group that receives the normal curriculum, and a placebo group that receives an extra hour of unstructured time (but no tablets) every day. This gives us three levels for the treatment factor: tablet (T), unstructured hour (U), and control (C). Suppose this is a one-year study where we have three terms (fall, spring, and summer), and three (mostly-similar) schools in which to run the experiment. Such a setup is known as a **row-column design**, and the experimental setup can be illustrated by the following table:

		Time of year		
		Fall	Spring	Summer
Location	1			
	2			
	3			

We'll fill in each entry with the treatment we use for that setup. A first attempt at this design (where T, U, and C stand for tablet, unstructured hour, and control respectively) might look like this:

		Time of year		
		Fall	Spring	Summer
Location	1	T	U	C
	2	T	U	C
	3	T	U	C

Unfortunately, this design does not properly take into account the time of year: if we were to run the experiment and see a significant improvement from the tablets, it might have been entirely due to the confounding effect of having the tablet conditions all in the fall! As a result, our ideal design would have each condition appear exactly once per row and once per column (like a Sudoku). Grids that satisfy constraints like this are called *latin squares*, and we can produce one by taking the table above and shifting each row:

		Time of year		
		Fall	Spring	Summer
	1	T	U	C
Location	2	C	T	U
	3	U	C	T

This way, we'll try each experimental condition in every location and during every time of year.

Example: How hard is experimental design, really?

Let's take a seemingly simple example, and see how complicated things can get. Suppose we want to bake the best possible loaf of bread. After some preliminary experimentation, we come up with 2 brands of flour, 2 brands of yeast, and 3 oven temperatures, and want to find the optimal combination (out of the 12 possibilities). We find 5 volunteer chefs willing to bake the bread, and 20 volunteer tasters willing to help us evaluate how good it tastes.

Exercise: How would you design an experiment to find the best combination of conditions?

Exercise: Suppose you have each chef bake each of the 12 loaves 4 times (to do this, you'd probably have to upgrade them from volunteeres to paid experimenters!). What are sources of variability within

(a) one loaf of bread?

(b) two loaves with the same recipe and ingredients from the same chef?

(c) two loaves with the same recipe and ingredients from different chefs?

How might you account for these sources of variability?