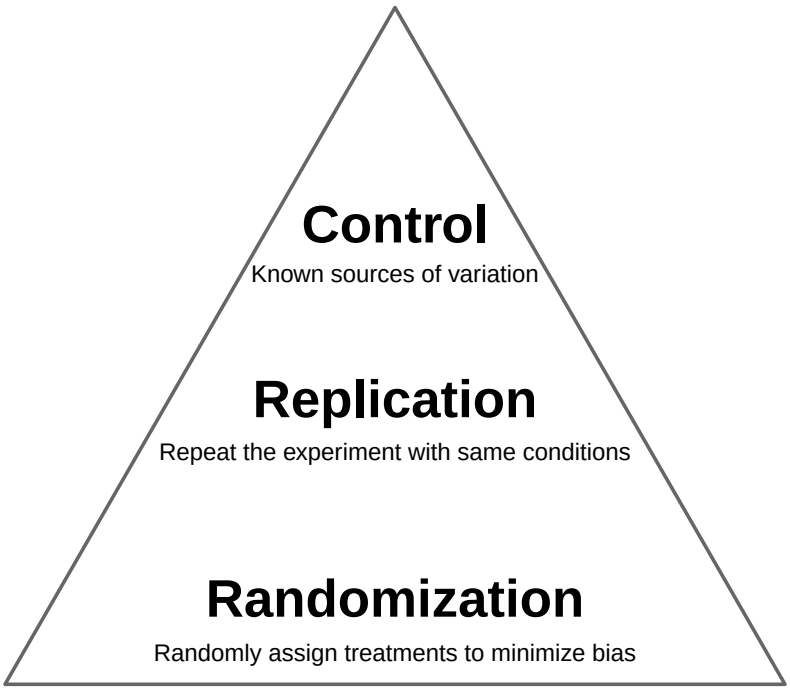
Chapter 2.3 - Designed Experiments

> "Hiring a statistician after the experiment is done is like hiring a physician when your patient is in the morgue. He may be able to tell you what went wrong, but he is unlikely to be able to fix it." - Sir Ronald A. Fisher

Designed experiments are the second pillar of data sourcing and by far the most universally recognized component of any scientific inquiry. Applied statisticians tend to fall into two categories, those who work with designed experiments and those who do not. I say to warn you that design of experiments is a tricky field that heavily mixes the idea of art with science; you shouldn't be disheartened if the topic isn't clicking for you.

The entirety of statistics for designed experiments owes its existence to Sir Ronald A. Fisher who developed the major principals for the field during his time in agricultural research (Working with wheat, which should make him an honorary Kansan). In his original text on the topic *The Design of Experiments* (Fisher, 1935) he highlighted three components for experiments:



We'll develop familiarity with the vocabulary for designed experiments, the types of designed experiments, and how Fisher's components fit into everything.

Design Vocabulary

You're a researcher for an agricultural research company. The company is testing a new pesticide that's only harmful to Japanese beetles (*Popillia japonica*), the primary pest their consumer market deals with. Your team is asked to validate that the pesticide:

a. Works to prevent beetles from damaging 5 different types of common garden flower.

b. Is effective at each of the 2 recommended treatment levels (mild and severe infestation) for consumers as well as the 2 recommended for commercial farming.

Naturally, you develop an experiment. Each of the 5 flower types are planted into 5 different plots, where each plot is isolated from one another, 4 plots receive differing levels of pesticide, one received no pesticide, and all are exposed to the same level of beetle infestation. After 2 weeks of using the recommended treatment plans on the packages all the flowers are cut and inspected for damage. Your team counts up the number of damaged leaves from each plot and records them as a sum total.

What was just outlined is an involved, but fairly traditional experimental design. It contained all the elements that statistics instructors love to see when testing vocabulary (albeit lacking in some characteristics needed to make it a proper design question).

First off, this was all easily distinguishable from an observational study because the researchers were able to **directly control** which individuals should have a response. This is the core definition for a **designed experiment**.

> **Designed Experiment**: A study where the researcher controls the conditions under which observations are taken and imposes treatments on individuals in order to observe the responses.

When we're developing an experiment we need to define the subjects with which we're exerting our influence over. In the case of the pesticide study those subjects were the *plots of flowers*, since we controlled how much pesticide was used per plot. We refer to those plots as **experimental units**.

> **Experimental Unit** (EU): The smallest unit to which a treatment is independently assigned/applied.

The experimental unit wasn't what we made our inference from, however. When we *observed* our response we were looking at individual leaves of our plants and counting up how many were damaged. Inline with standard statistical practices, the leaves are referred to as **observational units**.

> **Observational Unit** (OU): The smallest unit on which observations are made.

The application of the pesticide to each plot was the way that we controlled the general outcome of our experiment, and it's referred to as the **treatment** in the experiment.

> **Treatment** (Trt): Experimental condition applied to experimental unit.

Sometimes we have characteristics beyond the treatment that have an affect on the response. If we had decided that some plots received treatment once per week and some once per day we would refer to that as the *level* of treatment. Similarly, we can imagine exposing some plots to an extreme infestation of beetles and some to only a few. We would consider that exposure to be just like a treatment, in that it has an effect on the response and is measured by its levels, but since it **isn't** the treatment we refer to it as the **factor**.

Factor: A controlled independent variable; a variable whose levels are set by the experimenter.

All field experiments conclude with a measurement step: Counting colonies on an agar plate, weighing masses of dead mosquitos, measuring heights of saplings, and so on. When these measurements are the major outcome of interest, they're referred to as the **response**.

Response: The thing we measure to determine the effect of the treatments

---

Advantages of Designed Experiments

During the COVID-19 pandemic the entire world found themselves in a psuedo-experiment with face mask policies. At the time of writing this the public health field is still unsure just how effective mask policies are. One of the problems of that is that we can't run a **proper** experiment. Even if we could somehow have one state enforce a militant mask policy and have another outlaw them entirely without violating every ethical known ethical principal, it still wouldn't be enough.

Researchers would have two design a grand-scale experiment the size of two cities. Each city would have to have roughly the same number of people or at least demographically proportional populations. They would have one city operate on a mask policy with all participants willingly wearing masks anytime they might encounter another person. The other would have all participants go about their days as if masks weren't an option. Both cities would then have a pathogen introduced to their participants and the disease outbreak would be tracked day by day.

If it weren't for how terrible this is from a human rights and medical ethics standpoint, it would be incredibly informative. The researchers would gain a lot of **information dense** data. They'd be able to control all environmental conditions which would allow them to shuffle out un-explainable noise in the data, there would be information of combined effects (i.e., The interaction between *mask usage* and *age* relative to *disease prevalence*), and analyses could generate real sources of causation.

These are the major advantages of experiments:

- **Experiments avoid and tease out confounding effects**.
- **Experiments control the environment and remove factors we're not interested in**.
- **Experiments can study interaction effects (the combined effects of multiple factors)**.

The reasons behind why this hypothetical experiment is impossible are also the disadvantages of experiments:

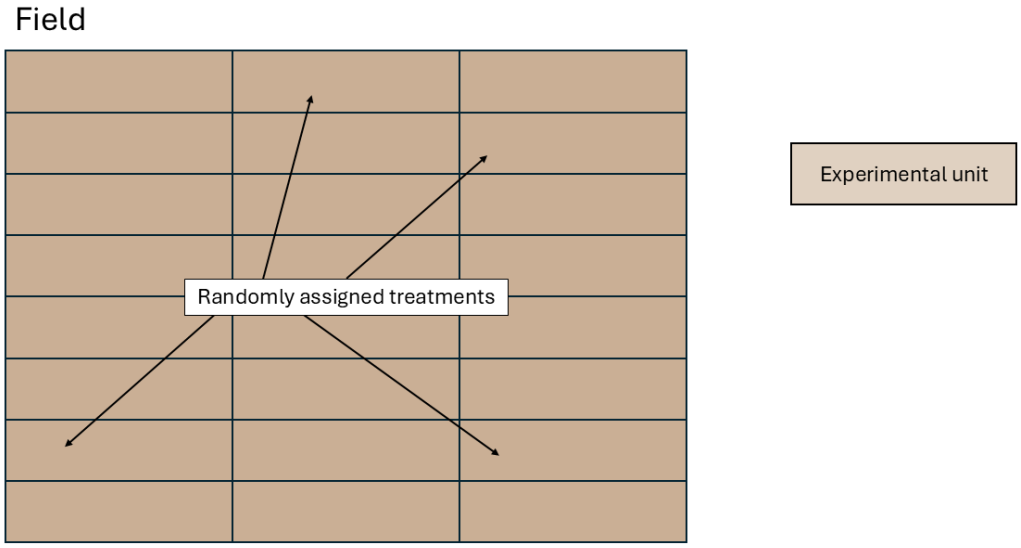
- **Experiments easily stumble into ethical problems**.
- **Experiments can be extremely resource intensive**.
- **The information from an experiment is only as good as its design**.

Common Experimental Designs

This next section will cover a series of named experimental designs. Methods become named because they're common— and they tend to be common because they're *very useful*. So while these aren't the full extent of possibilities in experimental design you should be *very aware* of these methods prior to developing your own.

> You want to determine if a new form of diet could help with weight loss. You gather 30 lab rats and assign each of them a random number (without replacement) between 1 and 30. Then you order them from 1 to 30, and randomly generate 15 numbers between that interval. Those numbers are given the new diet and the remaining are giving a standard diet. You homogenize their exercise and activity throughout the trial. The rats are weighed before and after the trial.

What you've performed is called a **Completely randomized design (CRD)**

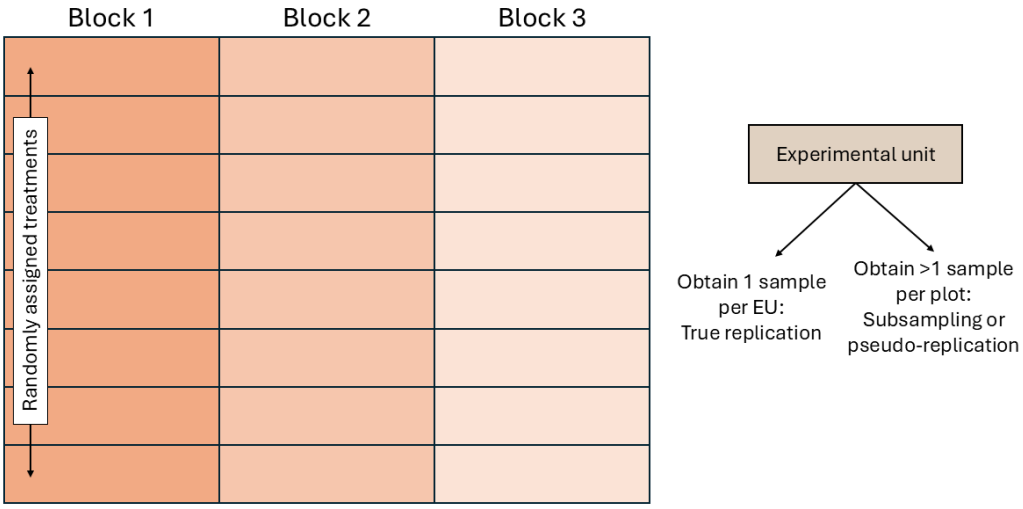


Individuals are randomly assigned to groups, then the groups are assigned to treatments completely at random. It doesn't need to be an equal number of individuals in each treatment group, but balanced designs are mathematically convenient.

CRDs are the fundamental framework for all experimental design, meaning that at the core of every design is a CRD. This is also the simplest form of experiment so while it lacks many features that more nuanced designs bring to the table, CRDs can still provide good enough evidence for different treatments **causing** different responses.

You want to determine the affect of a specific headache medication for both biological males and females. You acquire 50 volunteers and separate them by sex (30 males and 20 females). You perform a CRD on each group, assigning half of each group the trt of the headache medication and the other half a control. You compare the affect for each group.

You've performed a **Randomized Complete Block Design (RCBD)**:



Realistically this is a special case of stratification. Since experimental design happens before any observations happen we use slightly different language, but for simplicity sake blocks are identical to strata.

- **Blocking**: Taking a group of individuals known prior to the experiment, who share some attribute that is expected to affect the response.

While the blocks are not randomly assigned, treatment administered to or even within blocks **is** randomly assigned.

If you wanted to observe the outcome of two different trts, one of the ways you could observe that would be by assigning a different trt to two EUs of the exact same characteristic. Another could be by assigning a different trt to the same EU.

In this case we're met with two different designs: **Matched Pair** and **Repeated Measure**.

Matched Pairs: Two individuals are determined to be similar in the ways that are important to a study, one is given treatment A and the other is given treatment B.

Identical twin studies are the "classic" example of matched pairs:

> Pairs of identical twins are selected for a study of blood pressure medicine. One of the twins is given medicine A, and the other is given medicine B, neither are aware of which one they received. Treatment assignment is determined completely at random between the twins. Each twin has their blood pressure measured after one week.

The problem with matched pair designs is rooted in the loss of independent observations, which is discussed later in this chapter and fully detailed in Chapter 3.

Repeated Measures: Each individual is given both treatments with randomized order and assignment of order to avoid bias.

The (in)famous Pepsi vs. Coke experiments were a huge publicity stunt meant to close the debate on which cola was "better". The problem is that the designs always included some form of (intentional) bias. The ideal framework for this kind of study really wouldn't have been hard to achieve either:

> Thirty subjects are selected to compare Pepsi to Coke via double-blind tasting of each cola. The cola tasted first is determined by a coin toss prior to the colas being placed in the room for the subject to taste. Since the experiment is blind, the subjects aren't informed which drink is which before or during the tasting process. The double-blind component means that the researcher tossing the coin then rearranging the cups is not the same person as the surveyor recording the subjects scores, and the surveyor isn't informed the order of the colas. Subjects give scores to each on scale of 1 to 10 to indicate how much they like the taste, and the researcher who arranged the cups would "decode" the scores for each cup based off of their notes on how they ordered them.

This is a great way to generate a more objective conclusion on such a hot debate, but repeated measures have their own flaws.

They're not applicable to a wide variety of scientific inquiry (How we could possible perform repeated measures in vaccine trials?) and they're vulnerable to bias at every avenue. Ideally, all outcomes from all experimental units have nothing to do with the outcomes from one another or factors outside of the experiment. Repeated measures have to be cautiously developed in order to avoid this idea of *independence* being ripped apart.

Independence

Mathematical statistics is the sort of course that collegiate nightmares are made from. Dense material leading students down a pitch-black rabbit hole of convolution and asymptotic results. But as one of the most influential philosophers of the modern era, Uncle Iroh, once said:

> "... You can't always see the light at the end of the tunnel, but if you just keep moving... you will come to a better place."

Independence is that better place.

> Imagine a random sample of 10 ponds. You want to understand the affect of global climate change on the number of fish in each pond. If you assume that each pond is "iid" or "Independent and Identically Distributed" then you can treat each pond as if they're the *exact same*.

Thanks to the power and convenience that the assumption of independence provides, the majority of the statistical sciences revolve around it. Prior to the emergence of computational statistics (usage of physical computers rather than human ones) analyses almost **demanded** independence in order to proceed. As such we should consider some rules of thumb with independence:

- If we're **certain** we lack independence, we **must** address that in our analysis.
- If we aren't certain we can *usually* **assume** we have independence.

Choosing a design also means deciding how important the assumption of independence is to your study.

CRD and Matched Pair designs differ fundamentally since the matched pair groups are **not** independent. Each pair was intentionally decided and the analysis hinges on them being linked. In a CRD all of the individuals are independent of one another since everything was assigned via **fair** random chance.

Blocks in RCBDs *should* be independent, but the magnitude of the response is often **dependent** on the block itself.

Repeated measures can encounter the problem of each of the responses being affected by the previous ones—this breaks the assumption of independence entirely. Note: this is not an *intended* feature of repeated measures but it's common enough it should be considered in the design and execution of the experiment.

A clinical trial is being conducted to compare four diets(A,B,C,D), by assessing their ability to reduce LDL cholesterol in humans as measured in (mg/dl) from a 10-ml blood sample. Forty subjects total are available for this trial and will be assigned at random so that each diet has equal numbers of subjects.

- Factor(s)?
- Trt(s)?
- EU?
- Response?
- How many EUs are there per trt?

Suppose we have a total of 20 mares (female horses). Ten of them will be implanted with a capsule containing fish oil. The other ten mares will be implanted with a capsule containing sterile water. At the end of a week a blood sample will be taken from each mare and the serum fatty acid concentration (in mg/l) will be measured. Mares will be kept in separate stalls for the week of the experiment.

- Factor(s)?
- Trt(s)?
- EU?
- Response?
- How many EUs are there per trt?