Experimental Design

Throughout this course, we've discussed how data professionals use hypothesis testing to determine whether the results of an experiment are statistically significant. In previous scenarios, we analyzed the results of experiments such as clinical trials and A/B tests. For instance, we imagined a clinical trial that tests the effectiveness of a new medicine and an A/B test that examines how changing the design of a web page affects the average time customers spend on the page.

Data professionals often work with experimental data previously collected by other researchers. However, the right data for a specific project might not always be available or accessible. In this case, data professionals can design their own experiments and collect their own data.

In this reading, we'll discuss how data professionals design experiments to collect data, test hypotheses, and discover relationships between variables. You'll learn more about the basic concepts and procedures of experimental design.

Context: Experimental design

Experimental design refers to planning an experiment in order to collect data to answer your research question.

Researchers conduct experiments in many fields: medicine, physics, psychology, manufacturing, marketing, and more. The typical purpose of an experiment is to discover a cause-and-effect relationship between variables. For example, a data professional might design an experiment to discover whether:

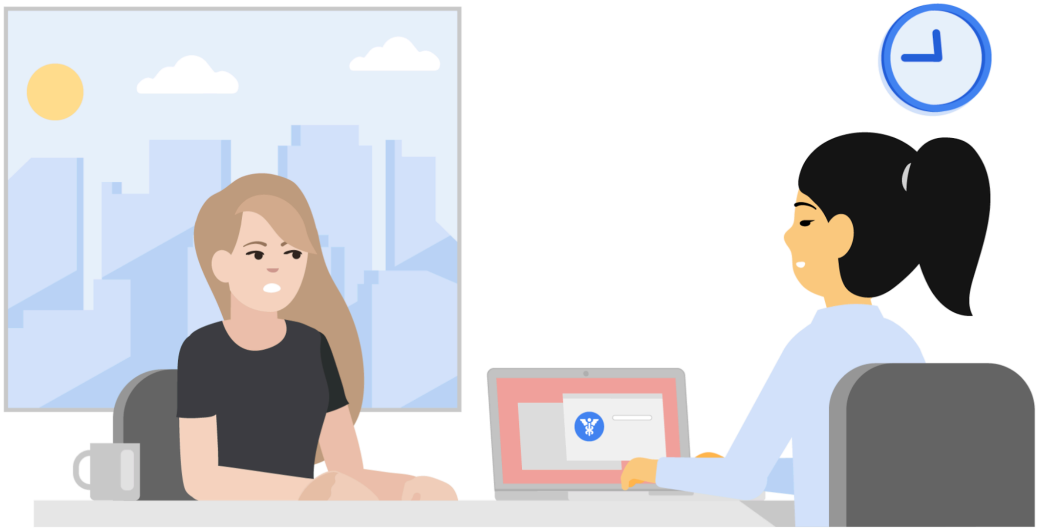
- A new medicine leads to faster recovery time
- A new website design increases product sales
- A new fertilizer increases crop growth
- A new training program improves athletic performance

It's important to understand experimental design because it affects the quality of your data, and the validity of any conclusions you draw based on your results. A poor design might lead to invalid results, which can be costly for companies and consumers. Based on the results of a flawed experiment, a company can spend years developing a medicine that is ineffective or invest heavily in a manufacturing process that is inefficient. A well-designed experiment will give you reliable data that helps answer your research question.

You can explore an example to get a better understanding of experimental design.

Example: Clinical trial

Imagine you're a data professional who works for a pharmaceutical company. The company invents a new medicine to treat the common cold. Your team leader asks you to design an experiment to test the effectiveness of the medicine. You want to find out whether taking the medicine leads to faster recovery time.



There are at least three key steps in designing an experiment:

1. Define your variables
2. Formulate your hypothesis
3. Assign test subjects to treatment and control groups

Note: These are basic steps that apply to controlled experiments (more below). Experimental design is a complex topic, and a more detailed discussion is beyond the scope of this course.
Next, examine each step in more detail using our example.

Step 1: Define your variables

Data professionals often begin by defining the independent and dependent variables in their experiment. This helps clarify the relationship between the variables.

- The **independent variable** refers to the cause you're interested in investigating. A researcher changes or controls the independent variable to determine how it affects the dependent variable. "Independent" means it's not influenced by other variables in the experiment.
- The **dependent variable** refers to the effect you're interested in measuring. "Dependent" means its value is influenced by the independent variable.

In your clinical trial, you want to find out how the medicine affects recovery time. Therefore:

- Your independent variable is the medicine—the cause you want to investigate.
- Your dependent variable is recovery time—the effect you want to measure.

In a more complex experiment, you might test the effect of different medicines on recovery time, or different doses of the same medicine. In each case, you manipulate your independent variable (medicine) to measure its effect on your dependent variable (recovery time).
Note: Later in this certificate program, when we discuss regression analysis, you'll have the chance to learn about independent and dependent variables in more detail.

Step 2: Formulate your hypothesis

The next step is to formulate a hypothesis. Your hypothesis states the relationship between your independent and dependent variables and predicts the outcome of your experiment. Earlier, you learned that data professionals formulate both null and alternative hypotheses when they conduct research that involves statistical testing. Recall that the null hypothesis typically assumes that there is no effect on the population, and the alternative hypothesis assumes the opposite. For your clinical trial:

- Your null hypothesis (H0) is that the medicine has no effect.
- Your alternative hypothesis (Ha) is that the medicine is effective.

Step 3: Assign test subjects to treatment and control groups

Treatment and control groups

Experiments such as clinical trials and A/B tests are controlled experiments. In a **controlled experiment**, test subjects are assigned to a treatment group and a control group. The **treatment** is the new change being tested in the experiment. The **treatment group** is exposed to the treatment. The **control group** is not exposed to the treatment. The difference in metric values between the two groups measures the treatment's effect on the test subjects.

In your clinical trial, the treatment is the medicine that the subjects in the treatment group are given. The subjects in the control group are not given the medicine. Imagine your results show that mean recovery time is lower in the treatment group (6.2 days) than in the control group (7.5 days). The difference between the two groups, 7.5 - 6.2 = 1.3 days, measures the treatment's impact. In other words, the medicine decreases mean recovery time by 1.3 days.

Note: After a data professional designs and runs their experiment, they use statistical testing to analyze the results. As a next step, you might conduct a two-sample t-test to determine whether the observed difference in recovery time is statistically significant or due to chance.

Ideally, exposure to the treatment is the only significant difference between the two groups. This design allows researchers to control for other factors that might influence the test results and draw causal conclusions about the effect of the treatment.

For example, imagine the subjects in your treatment group have a much healthier diet than the subjects in your control group. Any observed decrease in recovery time for the treatment group might be due to their healthier diet—and not to the medicine. In this case, you cannot say with confidence that the medicine alone is the *cause* of the faster recovery time.

Randomization

Typically, data professionals randomly assign test subjects to treatment and control groups. Randomization helps control the effect of other factors on the outcome of an experiment. Two common methods for assigning subjects to treatment and control groups are completely randomized design and randomized block design.

In a **completely randomized design**, test subjects are assigned to treatment and control groups using a random process. For example, in a clinical trial, you might use a computer program to label each subject with a number and then randomly select numbers for each group.

Sometimes, however, a completely randomized design might not be the most effective approach. When designing an experiment, data professionals must account for **nuisance factors**. These are factors that can affect the result of an experiment, but are not of primary interest to the researcher.

Researchers can use a **randomized block design** to minimize the impact of known nuisance factors. **Blocking** is the arranging of test subjects in groups, or blocks, that are similar to one another. In a block design, you first divide subjects into blocks, and then you randomly assign the subjects within each block to treatment and control groups.

For example, suppose you know that age is a significant factor in recovery time from the common cold. In particular, you know that people under the age of 35 tend to recover faster than older people. In this scenario, age is a nuisance factor because it might affect the results of your experiment. For example, in a clinical trial with a completely randomized design and a smaller sample size, you might randomly get a large proportion of young people in the treatment group. This will make it more difficult to determine whether any observed decrease in recovery time is due to the treatment (medicine) or to the nuisance factor (age).

In this case, blocking for the age factor is a more effective way to design your experiment. First, you divide the test subjects into blocks based on age, such as 21-35, 36-50, and 51-65. Next, you randomly assign the subjects within each block to treatment and control groups. This way, if there is a significant difference in recovery time within a specific block, you can be more confident that this result is due to the treatment (medicine) and not to the nuisance factor (age).