

---

## 12. Experiments

---

Experiments are the single most important tool available to the researcher. While not every research project can be fully framed as an experiment, the basic principles of experimental design can often still be applied to increase reliability.

The essence of an experiment involves exposing subjects to a treatment and observing the results.

The **treatment** is often fundamentally connected to the researcher's **hypothesis**. The hypothesis is falsifiable, so there are explicit predictions that can be deduced from the hypothesis. The treatment is then selected to see if those predictions are correct.

There are three fundamental principles of experimental design.

- Control
- Randomization
- Replication

Every well-designed experiment will attend to each of these principles. Each principle contributes in a powerful way to the reliability of the conclusions. In order to provide a context for the application of the principles, we will consider the following hypothesis.

### 12.1. Example.

*Hypothesis: Low carbohydrate diets result in greater weight loss for males over age 30 than do low fat diets.*

## 12.2. Control.

Notice that our hypothesis actually involves comparing two different diets: low carbohydrate diets and low fat diets. This is not an accident. Almost every well-formed hypothesis will involve some kind of comparative conclusion, whether between treatments (as in this case) or between treatment groups. For example, another possible hypothesis might be

*Hypothesis B: Low carbohydrate diets are more effective for males than for females.*

In this latter case our hypothesis there is only one treatment, but the conjecture is that it has different effects on different groups.

Even a seemingly direct hypothesis such as

*Hypothesis C: Low carbohydrate diets result in weight loss.*

is really a comparative statement. If we simply expose subjects to a low carbohydrate diet and observe weight loss, we don't know that the weight loss would not have occurred anyway, for example it might be an artifact of the subjects being observed (the **Hawthorne effect**). In order to properly test this hypothesis, we should have a second group exposed to no particular diet, and compare the results of both groups. The only difference between the first group (the **the experimental group**) and the second group (the **control group**) is the special diet to which the first group is exposed. If we see a difference between these two groups, which are the same in every way except the diet, then and only then do we have reliable evidence that supports Hypothesis C.

Thus using **experimental** and **control** groups is an important feature of experimental design and one of the ways in which the researcher exercises **control**.

Sometimes the experimental and control groups use the same subjects. In our original hypothesis

*Hypothesis: Low carbohydrate diets result in greater weight loss for males over age 30*

we might divide the subject pool into two groups, A and B. We might then expose subjects in Group A to a low carbohydrate diet and subjects in Group B to a low fat diet for three months. At the end of the first three months, we could then reverse the diets, with Group A exposed to a low fat diet and Group B exposed to a low carbohydrate diet. In this way we guarantee that any difference between the outcomes in "low fat" and "low carb" diets is due to the diet and not due to differences in the subjects.

Other ways in which the researcher exerts **control** over the research process involve the **research protocol**. The protocol is method by which subjects are exposed to the treatments, how the researcher interacts with the subjects, and how the measurements are taken.


Part of any protocol with human subjects involves assurances of ethical treatment. The most basic elements of ethical treatment include

- **Informed Consent.**
- **Information about potential risks and benefits.**
- **Ability to withdraw from the study at any time without penalty.**
- **Basic information about the purpose of the experiment and what the subjects will experience.**

Of course the researcher should take steps to avoid exposing subjects to undue risk, and should consider whether the potential benefits of the proposed research justify any potential risks to the subjects. Universities are required by law to have independent review boards that approve all research involving human or animal subjects.

In our diet example, all of the publicity about the low-carb Atkins diet might influence the outcomes. The subjects' expectations about weight loss might influence their compliance with the diet or might influence weight loss all by itself (the so-called **placebo effect**). Thus in our example we might not inform the subjects about the sequencing of the diets.

Presumably in our example the dependent variable would be **weight loss**. This means that a member of the research team will weigh the subjects at least at the beginning and end of each phase of the study. It is possible that the expectations of the researcher could also influence the measurements. This is particularly true in the case of drug trials where the control group is often receiving a placebo and the experimental group is receiving the experimental drug. Thus the member of the research team taking the measurements also is usually not informed as to which group is being observed.



This kind of design is said to **double-blind** since neither the subjects nor the observers know which treatment is being measured. This is another fundamental way in which the researcher exerts control over the experiment.

In order to assure that the experiment measures the difference between the treatment groups, the researcher will often take steps to standardize all interactions with subjects. Intake interviews, exit interviews, processing questions and all other interactions with the subjects are often carefully scripted and members of the research team are not permitted to deviate from the script. Even the physical setting – subject sitting or standing – can influence results and is therefore standardized. This is another aspect of **control**.




### 12.3. Randomization.

Experiments will almost always involve samples rather than census data. When dealing with samples, error is unavoidable since the researcher necessarily has incomplete information. Good experimental design avoids **bias** or systematic error. Systematic error favors one outcome over another in the experiment and thus can lead to false conclusions.

**Random error** however does not favor one outcome over another, but is neutral with respect outcomes. Thus in our diet example we would randomly select the test subjects.


In a random sample, every member of the population has an equally likely chance of being selected for the sample.



There are many challenges with constructing a truly random sample. Properly speaking the researcher should have a complete list of all members of the population and then randomly select the sample from that list: similar to a giant lottery.

There are many ways in which sampling bias can occur. For example, running an ad in a newspaper might result in persons more motivated to lose weight or to persons who are otherwise not representative of the population. Similarly, randomly selecting potential subjects from a phone directory limits the subject pool to those who have listed telephone numbers, missing those who only own cell phones, who do not have a phone or whose numbers are unlisted. These sampling methods do not involve **randomization** and are hence subject to bias.

Other sampling techniques involve **stratified random samples**, **cluster samples**, and **multi-phase sampling**. All of these are designed to increase the likelihood that the sample is similar to the population being studied and discussed more fully on the course website.



In our diet example, one proposed strategy was to divide the subjects into two groups, alternating the diet plans between the two groups. The division into the groups could be done randomly – for example by flipping a coin. This element of randomization is much easier to manage since we are now dealing with a smaller group, the sample. Note that this approach has the effect of randomizing the sequence in which any individual subject is exposed to the two diet plans.

By randomizing the sequence in which the subjects are exposed we are also exerting control over the sequence. It is possible that one diet is more effective if followed by the other, so having half our subjects randomly selected to be exposed to the diets in inverted order controls for this.

**Blocking** is a concept similar to stratified random samples. In stratified random samples, the population is divided into strata and then the sample is constructed by sampling from each strata. The strata are defined in ways that are relevant to the variables: for example, subject weight might be useful strata in our example.

In blocking, the sample is already constructed but there might still be differences in the subjects that could influence the results. In our example, early weight loss tends to be higher for persons with higher weight. Thus if one group started with more persons of higher weight, this could bias the outcomes. Thus the researcher might **block** the sample by initial weight, then randomly sequence the diets in each block. Ultimately the researcher still has two groups which are exposed to the diets in inverse order, but the groups are constructed in a way that makes them more similar.

Once again, the goal is to assure that our measurements are sensitive to differences in the treatments rather than unplanned differences in the subjects.

## 12.4. Replication.

This is the simplest principle: make the sample as large as possible.

This will make your measurements more sensitive to differences in the treatments and less sensitive to differences in the subjects.

As we have already seen, larger samples have smaller sampling variance. This is another way of stating the above observation. While larger samples are certainly more reliable, we shall see that relatively small samples can provide highly accurate and reliable conclusions. Properly constructed samples and surveys have repeatedly proven to provide reliable and accurate predictions regarding many phenomena, including elections.

## 12.5. Clinical Trials.

Clinical trials are a particular kind of experiment. These trials, which are required for new pharmaceuticals, are designed to occur in three phases, each testing a different hypothesis:


- **Phase One Trials** only look for harmful side-effects.
- **Phase Two Trials** test for efficacy.
- **Phase Three Trials** are longer term and test for both harmful side-effects and efficacy.



Phase One trials tend to be *cross-sectional studies*, relatively short-term and involve relatively small samples. Phase one trials look only for harmful side-effects. Aspirin would most likely not be approved for sale if it were introduced today due to harmful side-effects, namely aspirin allergy in a significant part of the population.

Phase Two trials tend to be also be *cross-sectional studies*, somewhat longer-term and can have samples that are quite large (at least 10,000). Phase Three trials are *longitudinal studies* and often have extremely large sample sizes.

Typically Phase Two and Phase Three trials are expected to identify harmful side-effects that affect as few as 0.5% of the population with at least 99% reliability. This level of accuracy and reliability requires samples of at least 10,000.



In 1950 Jonas Salk spent 18 months in human trials before the Salk Polio vaccination was approved for use in the general population. Today it typically takes as much as 18 **years** for all three phases of a clinical trial to complete and a new drug to be approved for sale. Fewer than one drug in one thousand that enters Phase One trials successfully completes Phase Three trials.

Clinical Trials were introduced in the 1960's. What happened between the Salk vaccine trials in the 1950's and the introduction of clinical trials 1960's that led to the introduction of more stringent protocols? Why is this being re-thought today?