# SAMPLING DESIGN

This lesson is about sampling a statistical population. By the end, you will be able to:

* Describe how a sampling design fits into the overall framework of statistical inference
* List the four goals of an ideal sampling design and apply them to a given study
* Define bias and sampling independence
* Identify errors in sampling design

## Four goals of an ideal sampling design

1. **All sampling units are selectable** This goal is about the sampling unit. Every sampling unit in the statistical population must have some non-zero probability of being included in your sample.
2. **Selection is unbiased** The probability of selecting a particular sampling unit cannot depend on any attribute of that sampling unit and—on average—the sampling units must have the same attributes as the statistical population.
3. **Selection is independent** Selection of a particular sampling unit must not increase or decrease the probability that any other sampling units is selected.
4. **All samples are possible** This goal is about the sample composition. All samples that could be created from the statistical population are possible.

## Observational studies

Overarching goal is to collect data from an existing statistical population that allows us to investigate relationships among variables - where the researcher has no control over independent variable.

1. Simple random survey - A straightforward approach is to create a list of all sampling units in the statistical population and then randomly draw from that list.

2. Stratified survey - In a stratified survey, we first break the statistical population into strata and then sample within each strata. As such, each strata has equal weighting in the sample.

3. Cluster survey - used by researchers to remove heterogeneity in the statistical population that is not relevant to the research question. The cluster is the sampling unit and the only scale with true statistical replication.

4. Case-control studies are used to compare data between two groups. The first group is called the case and contains sampling units with a particular response variable. The second group is the control and contains sampling units without the response variable of the case group. This type of sampling is purposefully biased in that it aims to select sampling units for the case group based on a measured response variable and compare that to the control group.

5. *Cohort surveys* follow sampling units over time, looking for the development of a particular response variable. The goal of this type of survey is to select a random set of sampling units (simple random survey) and then observe their fate over time. The key factor that distinguishes *cohort surveys* from *case-control surveys* in that the outcome is unknown when the sampling units are selected.

## Classifying observational survey designs

*retrospective* versus *prospective*

Retrospective studies are ones where the outcome is already known, which comes with an increased risk of spurious relationships if you are selecting groups based on the outcome. Case-control studies are a good example of a retrospective study. Prospective studies are ones where the outcome is not yet known. These are typically more effort because you need to follow the sampling units for a period of time, but these studies suffer less from spurious relationships. Cohort studies are a good example of a prospective study.

*cross-sectional* versus *longitudinal* studies

*Cross-sectional* studies are ones that study a response variableat only a single snapshot in time. In contrast, *longitudinal* surveys are ones that study a *response* variableat multiple points in time.

## Experimental studies

Each manipulated variable is called a factor. For example, manipulating soil nitrogen in the above example would be the *factor* in the experiment. Each *factor* then has at least two *levels*, which are different values of the *factor*. In the soil nitrogen example, there were two *levels*. The first levelhad no added nitrogen, and the second had added nitrogen.

There are two key things that distinguish *experimental studies* from *observational studies*. The first is that the explanatory variable is manipulated by the researcher. The second is that sampling units are randomly assigned to each *level* in each *factor*. As a result, there are two steps where sampling units are selected at random in *experimental studies*. The initial step is selecting sampling units to ensure that they are an independent and unbiased subset of the statistical population

*Replication* is the number of times a *treatment* is repeated on independent, representative and randomly selected units. In statistics, this is the sampling unit. Thus, the number of *replicates* is the number of sampling units in an experimental study.

Pseudoreplicationis an error in the design of an *experimental study* where the observation units are analyzed rather than the sampling units. In the soybean study, the plot of land is the sampling unit that is randomly allocated to each *treatment*. The individual soyplants are the observation units. Since the plot of land is the sampling unit, the individual soya plants within the plot are not statistically independent of each other. One way to think of this type of non-independence is that once the plot was chosen, the plants within the plot were also selected.

**Blocking** is analogous to stratified sampling, but for *experimental studies*. It is used to control for variation among the sampling units that is not of interest to the researcher.

In a single blind design, the sampling unit does not know the treatment they are assigned to. In a double blind design, both the researcher and sampling unit do not know what *treatment* they are assigned to. The primary benefit of a blinded design is to remove accidental bias caused from the sampling unit or the researcher knowing what *treatment* is being applied.

**Placebo** is a method often used in medical trials for the *control treatment* that helps accomplish a *blinded* design. It is substance, or *treatment*, that has no effect on the response variable.

**Sham treatment** is similar to *placebo* in that it is a method used in *control treatments*. However, the purpose of a sham treatment is slightly different in that it aims to account for the effect of delivery of a *treatment* that is not of interest to the researcher.