# Designing Studies

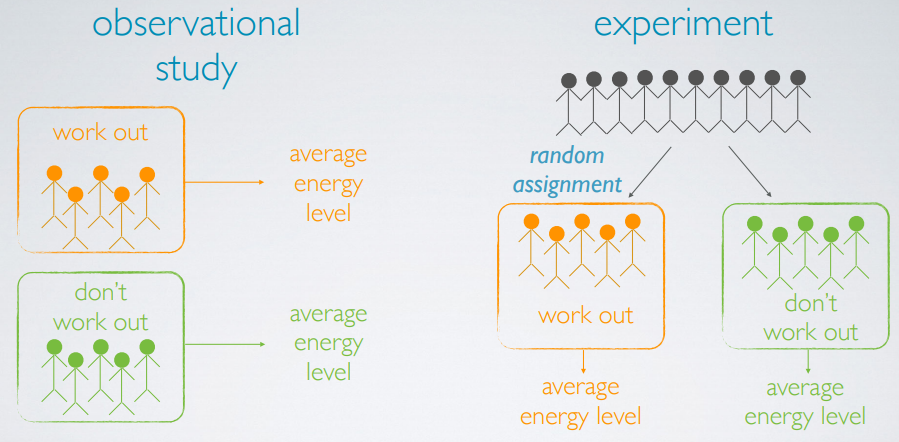
## Observational Studies & Experiments

### Observational

* collect data in a way that does not directly interfere with how the data arise/observe.
* Only establish an association
* Retrospective: use past data
* Prospective: data are collected throughout the study.

### Experiment

* Randomly assign subjects to treatments
* Establish causal connections



### Confounding variables

Extraneous variables that affect both the explanatory and the response variable, and that make it seem like there is a relationship between them.

Correlation does not imply causation.

## Sampling

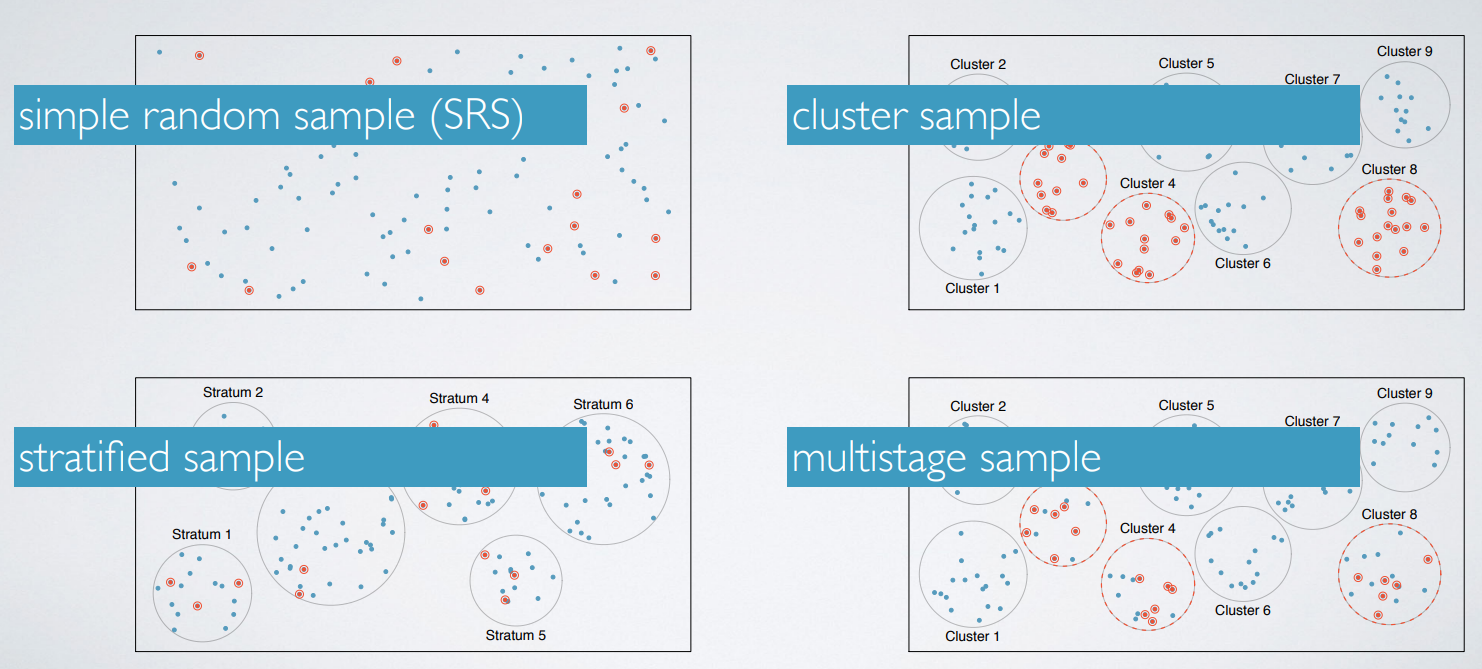
### Sampling bias

* Convenience sample: individuals who are easily accessible are more likely to be included in the sample.
* Non-response: if only a (non-random) fraction of the randomly sampled people respond to a survey such that the sample is no longer representative of the population. (there is an initial random sampling)
* Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. (no initial random sampling)

### Sampling methods

Distinguish between simple random, stratified, and cluster sampling, and recognize the benefits and drawbacks of choosing one sampling scheme over another.

* Simple random sampling: Each subject in the population is equally likely to be selected.
* Stratified sampling: First divide the population into homogenous strata (subjects within each stratum are similar, across strata are different), then randomly sample from within each strata. For example, divide the population into male/female groups first, then random sampling from each group.
* Cluster sampling: First divide the population into clusters (subjects within each cluster are non-homogenous, but clusters are similar to each other), then randomly sample a few clusters, and then randomly sample from within each cluster.
* Multi-stage sample: divide the population into clusters, randomly sample a few clusters, then randomly sample within these clusters. For example, divide a city into geographic regions that are on average similar to each other, sample randomly a few of these regions, and then sample a few people from within these regions.



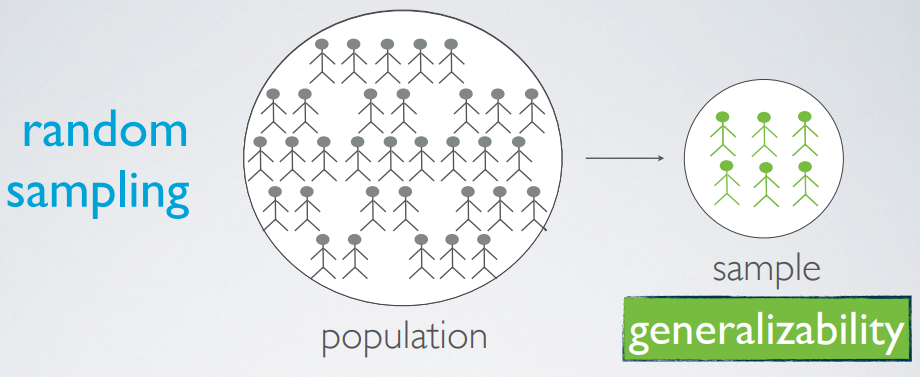
## Experimental Design

Identify the four principles of experimental design and recognize their purposes:

* control any possible confounders
* randomize into treatment and control groups
* replicate by using a sufficiently large sample or repeating the experiment
* block any variables that might influence the response.

### Random sampling

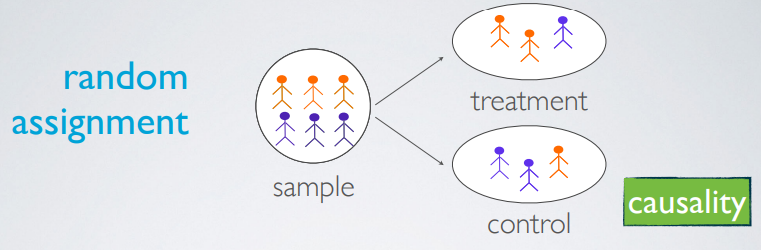
If subjects are selected randomly from the population, then each subject in the population is equally likely to be selected and the resulting sample is likely representative of the population. Therefore, the study’s results are generalizable to the population at large.



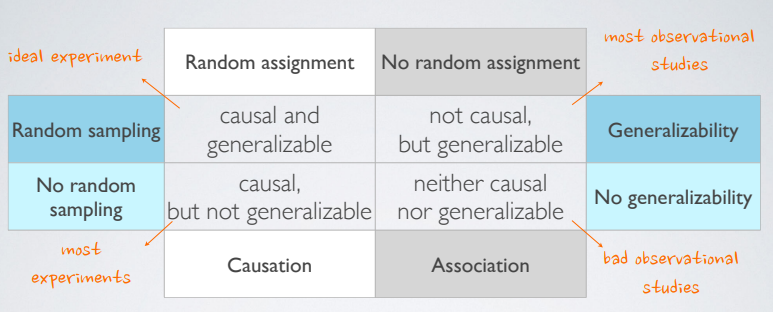
### Random assignment

Random assignment occurs only in experimental settings, where subjects are being assigned to various treatments. Taking a close look at our sample, we usually see that the subjects exhibit slightly different characteristics from one another. Through random assignment, we ensure that these different characteristics are represented equally in the treatment and control groups.

This allows us to attribute any observed difference between the treatment and control groups, to the treatment being observed on the subjects, since otherwise these groups are essentially the same. In other words, random assignment allows us to make causal conclusions based on the study.



So, random sampling happens first, then random assignment happens next.



Stratifying and blocking both allow for controlling for potential confounders, but at different stages of the study design. We stratify when we sample (divide population into strata and sample from within each stratum), and block in the process of random assignment (divide sample into blocks and randomly assign from within each block to treatment groups).