

Collecting Data & Planning Studies Vocabulary

Data Collection Basics

1. **Good Samples** mirror the population.
2. **Poorly collected data** cannot be fixed and is useless-no conclusions can be drawn
3. **Randomization:** Equalizes the effects of outside variables by using chance to assign experimental units to treatment groups.

Population vs. Sample

Population: the entire group of subjects or individuals that is the subject of interest.

- **Census**-the collection of data from each unit in the population. (*Difficult if not impossible*)
- **Parameter**-a numerical measurement of a population (*rarely known*).
- **If parameter μ is known**-No need for a confidence interval to estimate the mean.

Sample: a portion of a population which is studied to draw conclusions & about the characteristics of the whole population.

- Must be representative of the population to be useful.
- No conclusions should be drawn from poorly collected data or badly designed experiments.
- **Statistic**-a numerical measurement of a sample.

Bias

Bias-Consistently over or under-representing a component of the population in the sample.

- **Voluntary Response Bias**-Occurs when anyone is permitted to choose to respond to a general invitation. (*radio call-in; write in; internet polls*)

Issue: over-represents those with strong opinions.

- **Under-coverage Bias**-Occurs when members of a population cannot be chosen in a sample. (*homeless, inmates, no land-line, college students*)
- **Non-response Bias**-Occurs when a member in the population **cannot be reached** when called or **refuses to participate**.

Note: Survey non-response often exceeds 50%

- **Response Bias**-Non-neutral or poorly worded questions lead people to a particular response.

Note: Question order can influence responses.

Sampling Techniques

Simple Random Sample- A number is assigned to all subjects in the population. The numbers are then selected at random. **Every possible sample of a desired size has an equal chance of being selected.**

Acceptable methods: random number generators; mixed numbers in a hat; random number tables.

Note: If a random number table is used, it is necessary to describe the **selection process** and provide the **stopping point**.

Convenience Sample-A sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher.

Systematic Sample-A random starting point is selected and **every nth subject is chosen** for the sample.

- The starting point may be chosen by dividing the population size by the sample size.
- **Issue:** The order of the subjects is often related to the variable under consideration.

Cluster Sample-The population is divided into **heterogeneous** groups that represent the population. A cluster(s) is randomly selected. Within a cluster(s) a **census** is performed.

- **Issue:** Difficult to identify clusters that represent the population.

Stratified Sample-The population is divided into **homogeneous** groups called strata. A simple random sample from each strata is collected. The SRS from each strata are combined together to form the sample.

- Ensures each strata are represented.
- **Issue:** Not all samples have equal probability of being chosen.

Note: The sample size from each strata can be based on the strata's proportional representation within the total population.

Multi-Stage Sample-Involves 2 or more sampling methods.

Collecting Data & Planning Studies Vocabulary

Experiments

Experiment- A controlled study which assigns treatments to groups and observes the responses and satisfies the principles of **control, randomization, and replication.**

- **Control-**Conditions should be as similar as possible for all treatment groups by keeping other variables which might impact the outcome the same for all treatments.

Note: Group types are determined while assignment to the groups is done by chance.

Control is not synonymous with control group.

- **Randomization-**the random allocation of subjects to treatment groups to "even out" the effects of uncontrolled or unknown variables

Note: Randomization does not refer to the selection of a subjects for the study, but does refer to the placement/allocation of selected subjects to groups within the study.

- **Replication-** Replication means using a large enough number of subjects to reduce chance variation in a study.
- **Response variable-**variable that measures the outcome of a study; **outputs (Y's)**
- **Explanatory variable(s)-**Variables that may help explain or predict changes in the response variable; **inputs; (X's).**
- **Blind-**Subjects (**usually**) or data collectors are unaware as to what treatment is being received.
- **Double Blind-**Data collector and subjects are both unaware of the treatment received.
- **Placebo effect-** People not receiving treatment, believe they are and respond accordingly.
- **Completely Randomized Design-**Experimental units are assigned treatments by chance alone.
- **Blocking-**experimental units are divided into groups due based on a similar characteristic to bring out certain differences.
- **Randomized Block Design-**Subjects assigned to groups based on a similar characteristic and are then randomly assigned treatments.

Contrasts & Comparisons

SRS vs Random Allocation-*(not synonymous)*

- **SRS** refers to how a sample was obtained
- **Random allocation** refers to how subjects were assigned to treatments.

Lurking vs. Confounding

- **Lurking-**a variable not considered in the study, affecting **the response and the explanatory variables.** *(addressed by random assignment)*
- **Confounding variable-**an outside variable which **impacts the response variable** and confuses the effect of explanatory variable of interest.
(can be controlled by blocking)

Sampling Error vs. Bias-*(not synonymous)*

- **Sampling Error-**the difference between the actual population parameter and the statistical estimate of a **well-designed study.** *(Sample error decreases as sample size increases.)*
- **Bias-**Consistently over or under-representing a portion of a population due to **poorly designed sampling techniques.**
(a large sample size will not fix-start over)

Stratified & Blocking-Both divide the population into groups based on a similar trait.

- **Stratified-**refers to surveys
- **Blocking-**refers to experiments

Control vs. Control Group-*(not synonymous)*

- **Control-**balancing the effects on the response variable of variables outside the study.
- **Control Group-**provides the baseline for the experiment-often the placebo or status quo.

Experiment vs. Observational Studies

Experiments/controlled study-Treatments assigned to subjects & responses compared.

- Not always feasible due to ethical concerns
- **May Justify a cause and effect** conclusion

Observational Study-Measures & compares variables of interest without assigning treatments.

- Most data collection are observational studies & are good sources of primary data.
- **Surveys** are a type of observational Study
- May show a relationship or correlation.
- **Issue: Does Not prove cause and effect.**

Collecting Data & Planning Studies Vocabulary

Simulations

Simulation: a way to model random events, such that the outcomes closely match the real-world.

- Explain how to model the outcomes (Assign numbers to the possible outcomes)
- Explain how to simulate the trial-what are you going to do.

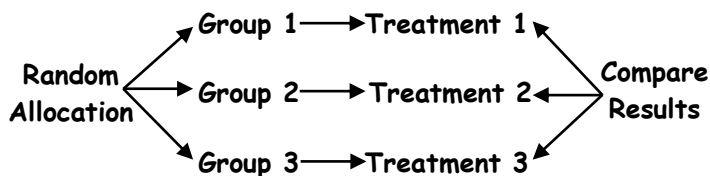
Caution: Don't forget to address repetition.

- Clearly State what the Response Variable is—

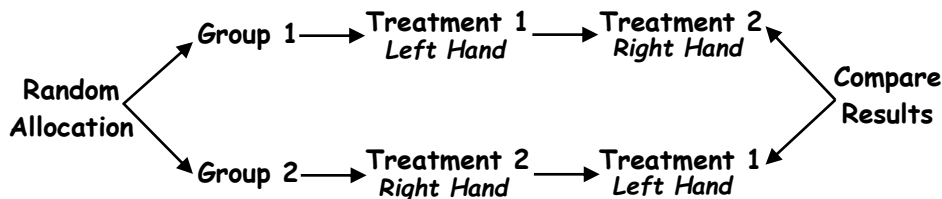
Hint: (How many random selections did it take to complete a trial?)

- Run Several Trials (chart the results)
- Analyze the Response Variable (take an average)
- State your conclusion in context.

Completely Randomized Design



Matched Pair Design



Randomized Block Design

