Unit 3: Collecting Data

Merrick Fanning

November 17, 2025

Unit 3 Outline: Collecting Data

1. Introducing Statistics: Do the Data We Collected Tell the Truth?
2. Introduction to Planning a Study
3. Random Sampling and Data Collection
4. Potential Problems with Sampling
5. Introduction to Experimental Design
6. Selecting an Experimental Design
7. Inference and Experiments

Goal: Learn how to design studies that produce reliable, unbiased data, and understand when it's appropriate to make conclusions about populations or cause-and-effect.

Observational Studies vs. Experiments

Observational Study:

- Researchers observe individuals and measure variables **without imposing treatments**.
- Useful for discovering associations, but **cannot establish causation**.
- Example: Measuring average sleep and GPA among students.

Experiment:

- Researchers **actively impose treatments** to measure the effect on a response variable.
- Well-designed experiments can support **causal conclusions**.
- Example: Randomly assigning students to receive extra study sessions and measuring test scores.

Observational Study	**Experiment**
No treatment imposed	Treatment imposed
Can show association	Can show cause-and-effect
No control over confounding variables	Can control confounding via design

Prospective vs. Retrospective Studies

Both are types of observational studies - no treatment is imposed, but data are collected in different ways.

Prospective Study:

- Identifies a group of subjects and **follows them forward in time**.
- Data are collected as events unfold.
- Example: Track a group of smokers and non-smokers over 10 years to compare lung cancer rates.

Retrospective Study:

- Looks **backward in time**, using existing records or past data.
- Subjects are grouped based on current outcomes.
- Example: Compare the past diets of patients who currently have heart disease with those who don't.

Prospective	**Retrospective**
Follows subjects forward	Looks back at past data
More control over data quality	Often uses historical records
Takes longer, often more expensive	Faster and cheaper

Sampling Bias

Definition: Sampling bias occurs when the method of selecting a sample **systematically favours certain outcomes**. It threatens the **representativeness** of the sample.

Common Types of Sampling Bias:

- **Undercoverage:** Some groups in the population are left out of the sampling frame. *Example: A survey that only contacts people with landlines.*
- **Voluntary Response Bias:** People choose to respond - often those with strong opinions. *Example: Online polls asking for public opinion.*
- **Convenience Sampling:** Individuals are chosen based on ease of access. *Example: Surveying students in the cafeteria during one lunch period.*
- **Question Wording Bias:** Questions that are confusing or leading.

Consequences:

- Biased samples lead to estimates that are **systematically wrong**.
- Results cannot be generalized to the population.

AP Tip: Sampling bias is about how the sample is selected - not about whether the data were collected correctly.

Sampling Bias and Unbiased Estimators

Sampling Bias: A sampling method exhibits bias if it **systematically overestimates or underestimates** the true value of the parameter we're trying to measure.

Key Idea: If the sampling method is biased, then the **sampling distribution** of the statistic will be centered **away from** the true population parameter.

Formal Definition: We say that a statistic $\hat{\theta}$ is an **unbiased estimator** of a parameter θ if:

$$\mathbb{E}[\hat{\theta}] = \theta$$

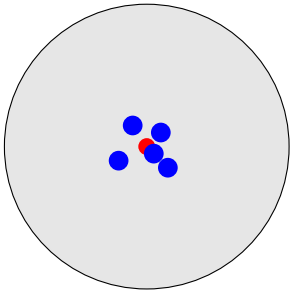
That is, the **mean of the sampling distribution** of the statistic equals the true parameter.

Examples of Unbiased Estimators:

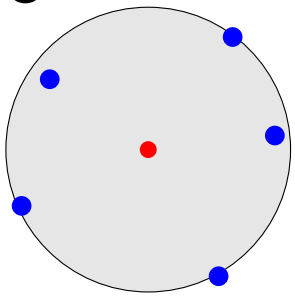
- Sample mean $\bar{x}$ is an unbiased estimator of population mean μ
- Sample proportion $\hat{p}$ is an unbiased estimator of population proportion p

Bias and Variability Illustrated

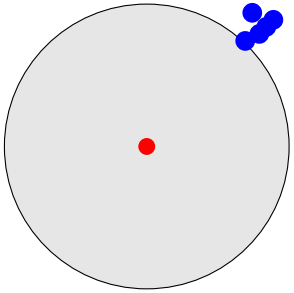
Low Bias
Low Variability



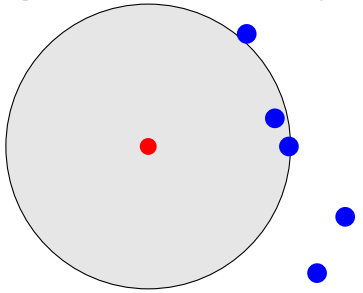
Low Bias
High Variability



High Bias
Low Variability



High Bias
High Variability



Red dot = true parameter

Blue dots = sample statistics (e.g., sample means)

Simple Random Sampling (SRS)

- **Definition**: Simple Random Sampling (SRS) is a method of sampling where every individual in the population has an equal chance of being selected.
- **Steps to Carry Out SRS**:
  1. **Label each individual**: Assign each individual in the population a unique number (e.g., from 1 to N).
  2. **Randomly select numbers**: Use a random number generator, random digit table, or drawing to randomly select numbers corresponding to the individuals.
  3. **Sample the individuals**: The individuals with the selected numbers make up the sample.
- **Important Notes for the AP Statistics Exam**:
  - Make sure the method is random (not biased).
  - Clearly explain your process of random selection.
  - Specify the population and ensure every individual has an equal chance.

Stratified Random Sampling

- **Definition**: Stratified Random Sampling is a method where the population is divided into subgroups, or strata, that share similar characteristics. A random sample is then taken from each stratum.
- **Steps to Carry Out Stratified Sampling**:
  1. **Identify the strata**: Divide the population into mutually exclusive subgroups (e.g., based on age, gender, income, etc.).
  2. **Randomly sample from each stratum**: Perform Simple Random Sampling within each stratum to select participants.
  3. **Combine samples**: The final sample consists of the individuals selected from each stratum.
- **Why Use Stratified Sampling?**
  - Ensures that each subgroup of the population is well-represented.
  - Helps improve precision and reduce variability within the sample.
- **Important Notes for the AP Statistics Exam**:
  - Clearly define the strata and how they relate to the population.
  - Explain why stratified sampling was chosen (e.g., to ensure representation of all groups).
  - Specify how you will randomly sample from each stratum.

Cluster Random Sampling

- **Definition**: Cluster Random Sampling is a method where the population is divided into smaller groups, or clusters. Entire clusters are then randomly selected for the sample, rather than selecting individuals.
- **Steps to Carry Out Cluster Sampling**:
  1. **Divide the population into clusters**: Create natural or convenient clusters (e.g., by geographic area, school districts, or neighborhoods).
  2. **Randomly select clusters**: Use a random method to select whole clusters to be included in the sample.
  3. **Sample all individuals in selected clusters**: Include every individual in the selected clusters for the sample.
- **Why Use Cluster Sampling?**
  - Useful when the population is large and spread out geographically.
  - Reduces travel and logistical costs by focusing on a smaller number of clusters.
- **Important Notes for the AP Statistics Exam**:
  - Clearly define the clusters and explain how they are formed.
  - Specify how clusters will be randomly selected.
  - Ensure the method selects whole clusters, not individuals within the clusters.

Examples for Sampling Techniques

Example 1: Political Scientist and Voting Behavior

- A political scientist is interested in estimating the proportion of Albertans who will vote for the conservative in the upcoming election.
- The goal is to select a representative sample of voters to determine the proportion of support for the conservative party.
- Different sampling techniques will be used to obtain this estimate.

Example 2: Sloth Researcher and Lifespan Estimation

- A researcher aims to estimate the lifespan of the three-toed sloth.
- While every sloth may be tagged and tracked, sampling techniques can still be used to confirm data consistency and speed up analysis.
- This example demonstrates the challenges and considerations when using sampling with a well-documented population.

Simple Random Sampling (SRS)

Example 1: Political Scientist and Voting Behavior

- **Steps to Carry Out SRS**:
  1. Label all eligible voters in Alberta from $1 - N$.
  2. Use a random number generator or random digit table to generate a sample of n numbers. Select individuals from the list that correspond with the numbers.
  3. Survey the selected individuals to estimate the proportion of voters who will vote conservative.

Example 2: Sloth Researcher and Lifespan Estimation

- **Steps to Carry Out SRS**:
  1. Label all the three-toed sloths in the database (e.g., tagged and tracked sloths). Is this realistic?
  2. Use a random number generator or table to select a subset of sloths.
  3. Estimate the average lifespan based on the sample.

Stratified Random Sampling

Example 1: Political Scientist and Voting Behavior

- **Steps to Carry Out Stratified Sampling**:
  1. Divide the population into strata based on relevant factors (e.g., age, income, urban/rural).
  2. Randomly sample from each stratum (e.g., randomly select voters from each age group).
  3. Combine the samples to get an estimate of the proportion of voters supporting conservatives.

Example 2: Sloth Researcher and Lifespan Estimation

- **Steps to Carry Out Stratified Sampling**:
  1. Divide the sloth population into strata based on factors such as age, sex, or location.
  2. Randomly sample from each stratum (e.g., select sloths from different age groups).
  3. Combine the samples to estimate the lifespan of the sloths.

Cluster Sampling

Example 1: Political Scientist and Voting Behavior

- **Steps to Carry Out Cluster Sampling**:
  1. Divide Alberta into smaller geographic clusters (e.g., cities, towns or districts).
  2. Randomly select several clusters.
  3. Survey all voters in the selected clusters to estimate the proportion of conservative supporters.

Example 2: Sloth Researcher and Lifespan Estimation

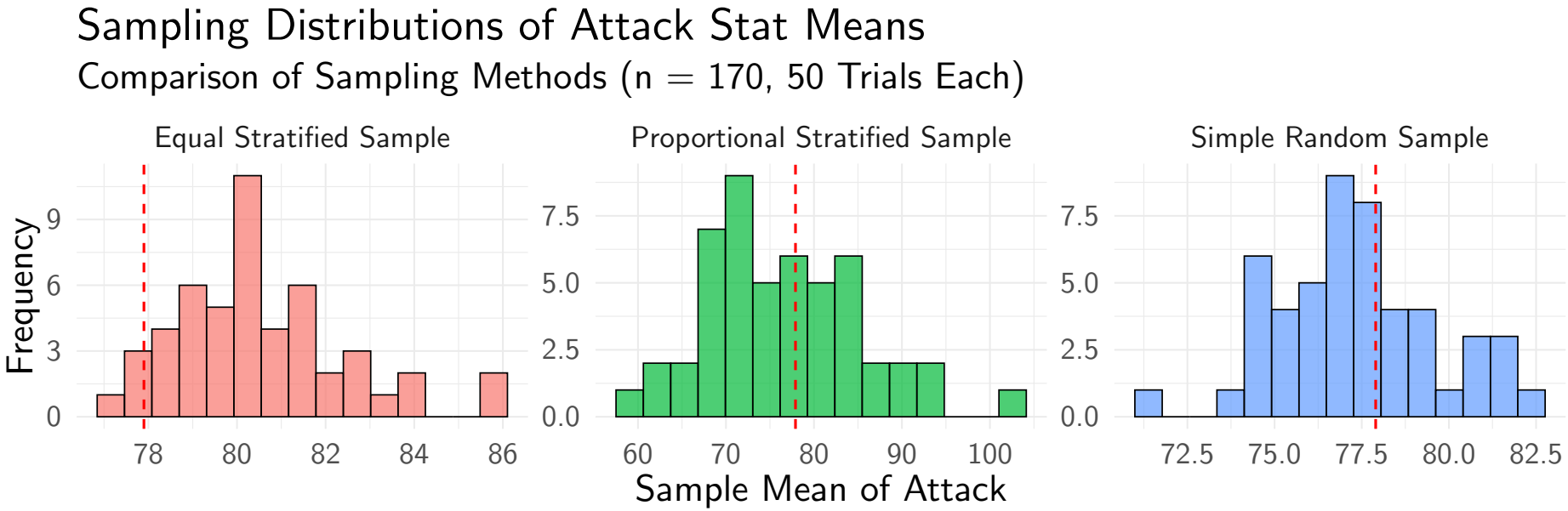
- **Steps to Carry Out Cluster Sampling**:
  1. Divide the population of sloths into clusters (e.g., by wildlife reserves or regions).
  2. Randomly select a few clusters.
  3. Measure the lifespan of all sloths within the selected clusters.

Full Marks on Sampling Questions

To earn full marks on a sampling question, ensure you do the following:

- **Assigning Numbers**:
  - Assign a unique number to each individual in the population. For example, label all the students in the class with numbers 1, 2, 3, ..., N.
- **Using a Random Number Generator**:
  - Use a random number generator or a random digit table to select distinct numbers from a specified range (e.g., 1 to n).
  - Ensure that the numbers are randomly chosen and no duplicates occur.
- **Linking Selected Numbers with Corresponding Individuals**:
  - Once random numbers are selected, link them to the corresponding individuals from the population list.
  - These individuals form your sample for further analysis.

Pokémon Example (Kaggle Dataset)



1. Which sampling technique led to a biased sampling distribution?
2. Which sampling distribution has the least variation?

Source: Pokémon dataset from Kaggle

Introduction to Experiments

Example:
Last year, a school offered an after-school Physics 30 diploma prep class that students could volunteer to attend. 44 students took the prep class and later wrote the Physics 30 diploma exam. The average diploma exam score for these students was 82%. The average diploma score for students who did not take the prep class was 68%.

Discussion Questions:

- Is the situation described an observational study or an experiment?
- What are the explanatory and response variables for this example?
- Can you conclude that taking the prep course will cause a student's test score to increase? Why or why not?
- What other variables might explain the difference in average scores?
- How could we design a better study to investigate the effect of the prep class?

Confounding Variables

A **Confounding Variable** is a variable that is related to both the independent variable (explanatory variable) and the dependent variable (response variable). It can distort the apparent relationship between these variables.

- **Key Features of Confounding Variables**:
  - A confounding variable is associated with both the independent and dependent variables.
  - It makes it difficult to determine the true effect of the independent variable on the dependent variable.
  - Confounding variables can either strengthen or weaken the apparent relationship between the variables.
- **Example of Confounding**:
  - In a study testing the effect of exercise on weight loss, a confounding variable could be diet. If people who exercise also follow a specific diet, it's hard to isolate whether weight loss is due to exercise alone or a combination of exercise and diet.
- **How to Handle Confounding Variables**:
  - Randomization: Ensures that potential confounders are equally distributed across treatment groups.
  - Control: Use statistical methods to control for confounding variables, such as blocking or regression.

Lurking Variables

A **Lurking Variable** is a variable that has an effect on both the independent and dependent variables but is not included in the study. Lurking variables can create a false impression of a relationship between the variables being studied.

- **Key Features of Lurking Variables**:
  - Lurking variables are not included or measured in the study but affect both the independent and dependent variables.
  - They can create a false or misleading relationship between the studied variables.
- **Example of Lurking**:
  - In a study showing a correlation between ice cream sales and drowning rates, the lurking variable might be temperature. Warmer weather leads to more people buying ice cream and also increases the likelihood of swimming and drowning.
- **How to Handle Lurking Variables**:
  - Identifying and including lurking variables in the analysis is important to avoid erroneous conclusions.
  - Use randomization or control for known lurking variables to improve study accuracy.

Overview of Experiments

An experiment is a study in which the researcher actively manipulates one or more variables to observe the effect on another variable.

- **Key Features of Experiments**:
  - The researcher manipulates the independent variable (the factor).
  - Participants are randomly assigned to different treatment groups (e.g., control vs. treatment group).
  - The effect of the manipulation is measured on the dependent variable.
- **Identifying Experiments vs. Observational Studies**:
  - **Experiment**: The researcher manipulates the variable of interest and randomly assigns subjects to **treatments**. The goal is to establish causality.
  - **Observational Study**: The researcher simply observes and records data without manipulating any variables. The goal is to identify associations or correlations, not causality.

Factors, Levels, and Treatment Groups in Experiments

- **Factors**:
  - A factor is a variable that is manipulated by the researcher in an experiment.
  - Examples of factors include treatment types, dosage amounts, or methods used in an experiment.
- **Levels**:
  - The levels are the different values or categories that a factor can take.
  - For example, if the factor is "dosage," the levels could be 0mg, 5mg, and 10mg.
- **Determining the Number of Treatment Groups**:
  - The number of treatment groups is found by multiplying the number of levels for each factor.
  - Example: If there are 3 levels of factor A (e.g., low, medium, high) and 2 levels of factor B (e.g., control, treatment), then the number of treatment groups is $3 \times 2 = 6$ treatment groups

Completely Randomized Design

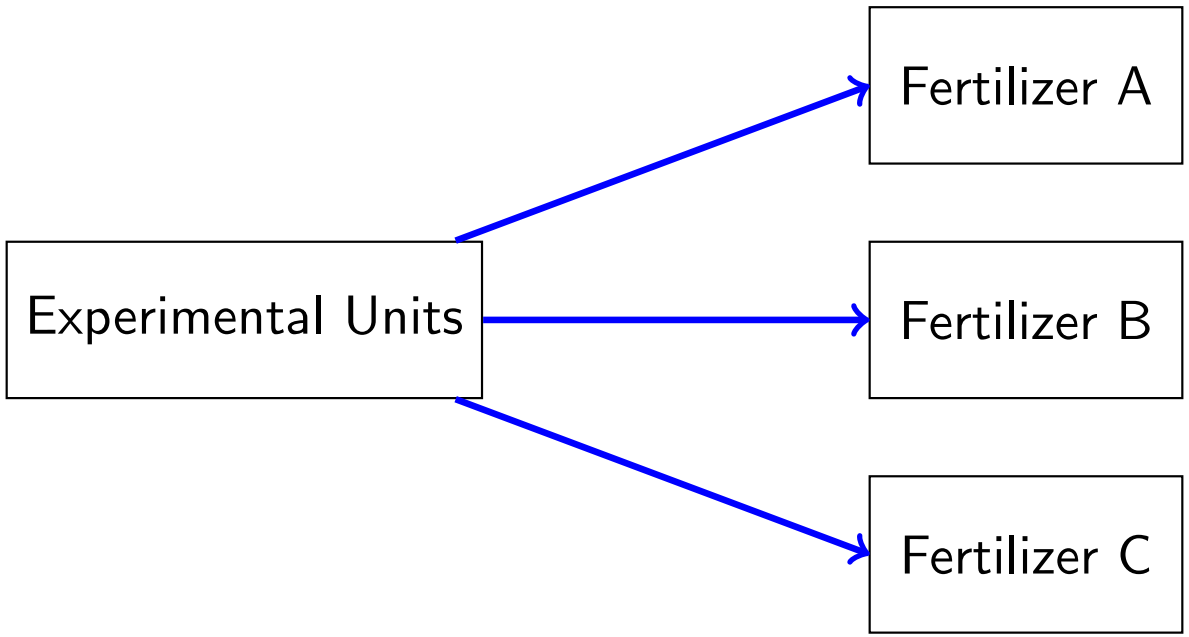
A **Completely Randomized Design** (CRD) is an experimental design where all experimental units are randomly assigned to treatment groups.

- **Key Features**:
  - Random assignment of experimental units to treatments.
  - Each unit has an equal chance of being assigned to any treatment group.
  - Often used in experiments with a single factor.
- **Steps**:
  - Randomly assign experimental units to treatment groups.
  - Measure the outcome for each treatment group.
  - Analyze the data to compare the effects of each treatment.
- **Example Features**:
  - The treatment groups are created through random assignment.
  - Often used when the goal is to test the effect of different treatments on a single outcome.

Example and Experimental Design Chart

Example: A researcher is testing the effect of three types of fertilizer (A, B, and C) on plant growth. 30 plants are randomly assigned to one of the three fertilizers. The plants' growth will be measured after a set period.

- The experiment involves three treatments (fertilizer types).
- Plants are randomly assigned to each fertilizer treatment.
- The outcome being measured is the plant growth after a fixed period.



Randomized Block Design

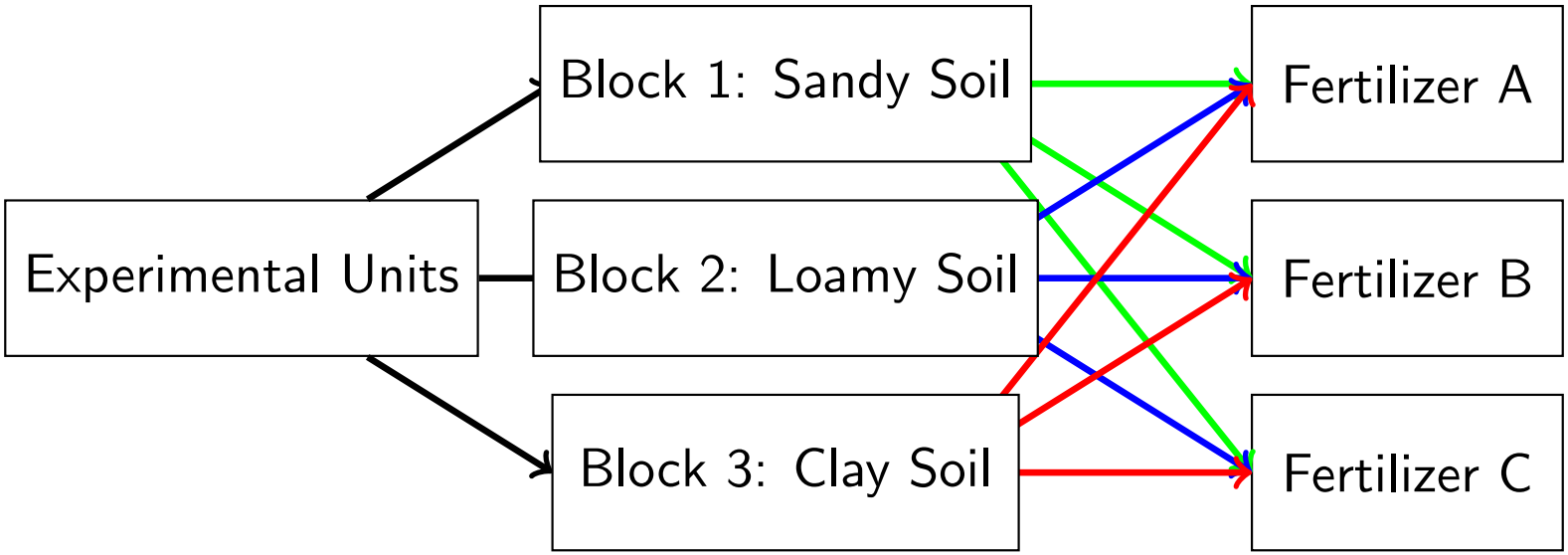
A **Randomized Block Design** (RBD) is an experimental design where experimental units are first grouped into blocks based on a known characteristic, and then treatments are randomly assigned within each block.

- **Key Features**:
  - Experimental units are grouped into blocks based on a specific characteristic (e.g., age, gender, soil type).
  - Random assignment of treatments is done within each block.
  - Used to reduce variability within the treatment groups and increase precision.
- **Steps**:
  - Divide experimental units into blocks based on a characteristic (e.g., plant species, initial size).
  - Randomly assign treatments to experimental units within each block.
  - Measure the outcome for each treatment in each block.
  - Analyze the data to compare treatments within blocks and overall.
- **Example Features**:
  - The blocks control for variability due to a known factor.
  - Treatment comparison is made within each block to isolate the effect of the treatment.

Example and Experimental Design Chart

Example: A researcher is testing the effect of three types of fertilizer (A, B, and C) on plant growth. Plants are grouped into blocks based on soil type (e.g., sandy, loamy, clay). Each block contains plants that will receive one of the three fertilizers. The growth of the plants is measured after a set period.

- The experiment involves three treatments (fertilizer types).
- The plants are grouped into blocks based on soil type.
- The outcome being measured is plant growth after a fixed period.



Matched-Pairs Experimental Design

A **Matched-Pairs Design** is a type of experimental design where experimental units are paired based on some similarity or characteristic. Each pair is then randomly assigned to different treatments, or two treatments are applied to the same unit.

- **Key Features of Matched-Pairs Design**:
  - Paired experimental units are similar in some way (e.g., same age, gender, or pre-treatment condition).
  - Two treatments are applied to each pair or one treatment is applied to both members of a pair.
  - The goal is to reduce variability by comparing the difference in responses within each pair.
- **Steps in Matched-Pairs Design**:
  - Pair the experimental units based on a relevant characteristic (e.g., similar height, weight, or pre-treatment score).
  - Randomly assign one treatment to one member of each pair and the other treatment to the other member.
  - Measure and compare the outcomes for the two treatments within each pair.

Example of Matched-Pairs Design

Example: A researcher is testing the effect of a new diet on weight loss. 10 participants are selected, and each participant is paired with another based on gender and initial weight. One member of each pair is given the new diet, and the other is given a placebo diet. The weight loss after 4 weeks is compared between the two treatments.

- The same pair of participants are used to compare the two diets.
- Randomly assign which diet each participant will receive.
- The outcome being measured is the difference in weight loss between the two diets.

Other Miscellaneous Stuff

1. Single and Double Blind Experiments
2. Confirmation Bias
3. Placebos – Check out video here.