ÉCOLE SUPÉRIEURE EN SCIENCES ET TECHNOLOGIES DE L'INFORMATIQUE ET DU NUMÉRIQUE

# FUNDAMENTALS OF DATA SCIENCE AND DATA MINING

## CHAPTER 5:

## EXPERIMENTAL DESIGN

Dr. Chemseddine Berbague

2024-2025

# OUTLINE

- Introduction to experimental design
  - Types of experimental design
- Key concepts.
  - Independent variable, dependent variable.
  - Randomization, replication.
- Controlled group Vs Experimental group
- A/B Test
  - Step of A/B test
  - Application of A/B test
- Causality and correlation
- Statistical tests

# INTRODUCTION

- **The Importance of Experimental Design in Data Science**
  - **Identify Causal Relationships**
    - Helps distinguish between correlation and causation, leading to more reliable predictions.
  - **Minimize Bias**
    - Ensures unbiased, accurate results by controlling confounding variables.
  - **Optimize Resources**
    - Efficiently tests multiple factors, maximizing time and resource effectiveness.
  - **Improve Model Accuracy**
    - Provides structured methods to validate and fine-tune models for better performance.
  - **Ensure Ethical Practices**
    - Promotes transparency and reproducibility, ensuring trustworthy and ethical outcomes.

# MOTIVATION

- **1. Identify Causal Relationships**
  - **Case**: A study found a correlation between ice cream sales and drowning incidents, but failed to account for warm weather being the true cause of both.
- **2. Minimize Bias**
  - **Case**: A drug trial only included young, healthy participants, making the results unrepresentative of the general population.
- **3. Optimize Resources**
  - **Case**: Small sample sizes in A/B tests led to inconclusive results, wasting time and resources without meaningful insights.
- **4. Improve Model Accuracy**
  - **Case**: A model trained and tested on the same data showed high accuracy but failed when applied to new data, due to overfitting.
- **5. Ensure Ethical Practices**
  - **Case**: Facebook manipulated user feeds without consent to study emotional impact, violating ethical standards and user trust.

# DEFINITION

- **Experimental design (ED)** is a systematic approach to planning, conducting, and analyzing experiments to ensure reliable and valid results.
- **ED** involves making thoughtful decisions about how to manipulate and control variables to investigate cause-and-effect relationships or compare different treatments or interventions.
- **ED** is commonly used in scientific research, data science, and various fields to draw meaningful conclusions from experiments.

# COMMON EXPERIMENTAL DESIGN

- **Quasi-Experimental Design**:
  - In situations where randomization is not feasible or ethical, quasi-experimental designs are used.
  - They still involve a control group and an intervention but lack random assignment.
- **Within-Subjects Design**:
  - In this design, the same group of participants is exposed to all experimental conditions, allowing researchers to compare participants' responses under different treatments.
- **Between-Subjects Design**:
  - Different groups of participants are exposed to different experimental conditions.
  - This design compares the average responses of different groups.

# COMMON EXPERIMENTAL DESIGN

- **Randomized Controlled Trial (RCT)**: This is a gold standard experimental design used to evaluate the effectiveness of interventions or treatments.
  - Participants are randomly assigned to either a control group (receives no treatment) or an experimental group (receives the treatment). The outcomes of the two groups are compared to assess the treatment's impact.
- **Pretest-Posttest Control Group Design**: This design includes both a pretest (measurement before treatment) and a posttest (measurement after treatment) for both the experimental and control groups.
  - It helps assess whether the treatment caused changes in the dependent variable.
- **Factorial Design**: This design involves studying the effects of multiple independent variables (factors) simultaneously. It allows researchers to examine the main effects of each factor and their interactions.

# KEY CONCEPTS IN EXPERIMENTAL DESIGN

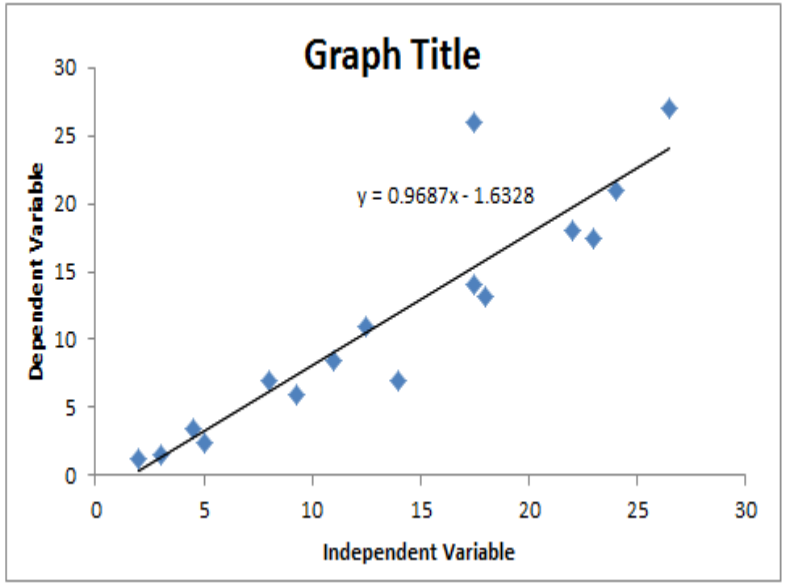
- **Independent Variable (IV):** The variable that the experimenter manipulates or controls to observe its effect on the dependent variable. It is also called the "treatment" or "factor."
- **Dependent Variable (DV):** The variable that is measured or observed to assess the effect of the independent variable. It is the outcome of interest and responds to changes in the independent variable.
- **Control Group:** In experiments with multiple groups, a control group serves as a baseline to which the experimental groups are compared. The control group does not receive the treatment or intervention being tested, helping researchers understand the impact of the independent variable.
- **Randomization:** Random assignment of participants or samples to different experimental groups helps reduce biases and ensures that the groups are comparable before the treatment is applied.
- **Replication**: Repeating the experiment multiple times with different samples or participants helps increase the reliability of the results and assess the consistency of the effects.

# COMPARISON TABLE

| Aspect | Independent Variable (IV) | Dependent Variable (DV) |
|---|---|---|
| **Purpose** | Represents the cause or factor that is presumed to influence the dependent variable. | Represents the effect or outcome that is influenced by the independent variable. |
| **Control** | Controlled or deliberately varied by the researcher. | Not controlled by the researcher; instead, it is observed or measured. |
| **Variation** | Changes in this variable are introduced by the researcher to observe their effect. | Changes in this variable occur as a result of changes in the independent variable. |
| **Role in Experiment** | Acts as the input or predictor variable. | Acts as the output, response, or predicted variable. |
| **Type of Data** | Can be categorical (e.g., group assignments) or continuous (e.g., dosage levels, time). | Can be continuous (e.g., test scores, revenue) or categorical (e.g., pass/fail outcomes, yes/no responses). |
| **Example (Study Time vs. Scores)** | Study time: The researcher manipulates the amount of time students spend studying to observe its effect on exam performance. | Exam scores: The researcher measures the scores to determine how they are affected by changes in study time. |
| **Relationship** | Changes in the independent variable are presumed to cause changes in the dependent variable. | Changes in the dependent variable are the result of changes in the independent variable. |
| **Control Over Values** | The researcher has full control over the values or levels assigned to this variable. | The values depend on how the independent variable interacts with the experiment or study environment. |

# EXAMPLE OF IV AND DV

- A real estate company wants to predict the price of houses based on various features.
  - **Independent Variable (IV)**: The **square footage** of the house (the size of the house in square feet).
    - The company believes that larger houses will likely have higher prices, so square footage is the variable they manipulate or use to predict the outcome.
  - **Dependent Variable (DV)**: The **price of the house** (how much the house is sold for).
    - The price depends on various factors, including the square footage, and it is what the company is trying to predict based on the house's size.



# CONTROL GROUP

- **Definition**:
  - A control group is a group of participants or samples in an experiment that does not receive the intervention or treatment being tested. It serves as a baseline for comparison to evaluate the effects of the intervention.
- **Purpose**:
  - The primary purpose of a control group is to provide a reference point against which the experimental groups are compared. By isolating the effect of the independent variable, researchers can assess whether the intervention causes a change in the dependent variable.
- **Experiment Design:**
  - Control groups are commonly used in experimental and quasi-experimental designs. In an experiment, participants or samples are randomly assigned to either the control group (no treatment) or the experimental group(s) (receiving the intervention).
- **Example**:
  - In a drug trial, the control group would receive a placebo (a substance with no therapeutic effect), while the experimental group receives the actual drug. By comparing the outcomes between the two groups, researchers can determine the drug's effectiveness.

# EXPERIMENTAL DESIGN

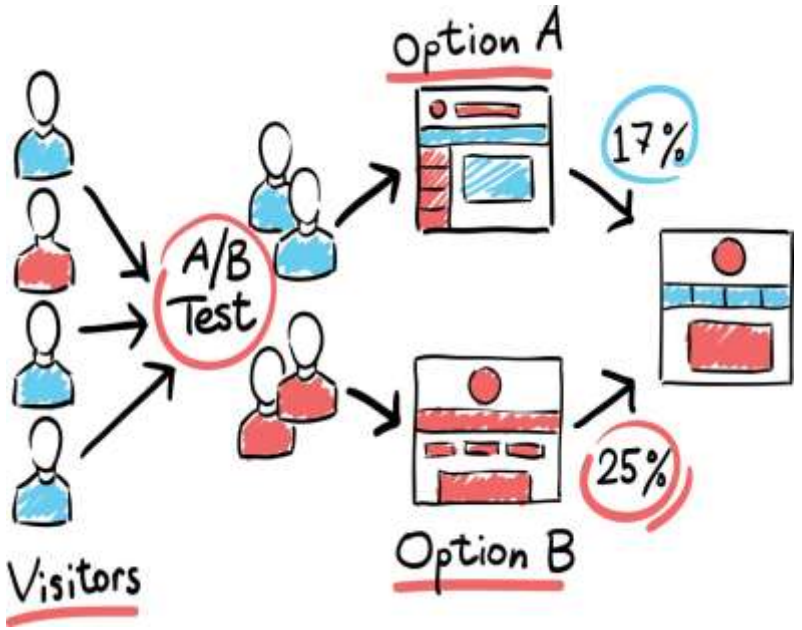
- **Definition:**
  - An **experimental group** is the group in an experiment that receives the treatment or intervention being tested, exposing it to changes in the independ variable.
- **Purpose:**
  - To assess the effect of the treatment by comparing its outcomes to a control group.
- **Experiment Design:**
  - Participants are assigned to the experimental group to receive the interventio while the control group serves as a baseline.
- **Example:**
  - In a drug trial, the experimental group receives the drug, and the control grou receives a placebo to evaluate the drug's effectiveness.

# EXAMPLES OF CONTROL GROUP AND EXPERIMENTAL GROUP

- **Scenario 1: Email Marketing Campaign**
  - A data science team wants to test whether a **personalized email** increases customer purchases:
    - **Control Group**: Customers receive a **generic email** without personalization.
    - **Experimental Group**: Customers receive a **personalized email** with their name and tailored recommendations.
  - The team measures the **conversion rate** (percentage of customers who make a purchase) for both groups.
  - **Task for Students**: Identify the control group, experimental group, and the independent and dependent variables.
- **Scenario 2 Predicting Employee Productivity**
  - A company tests whether **flexible work hours** affect employee productivity:
    - **Control Group**: Employees work **standard 9-to-5 hours**.
    - **Experimental Group**: Employees are allowed **flexible work hours**.
  - The team measures the **average number of tasks completed** per week by both groups.
  - **Task for Students**: Define the groups, the variable being tested, and the outcome
- **Scenario 3: Machine Learning Algorithm Performance**
  - A data scientist wants to test whether adding a **new feature** to a machine learning model improves prediction accuracy:
    - **Control Group**: The model runs with the **original feature set**.
    - **Experimental Group**: The model runs with the **new feature added**.
  - The scientist compares the **accuracy scores** of the two models on a validation dataset.
  - **Task for Students**: Determine which group is control, which is experimental, and what metric is used to evaluate performance.

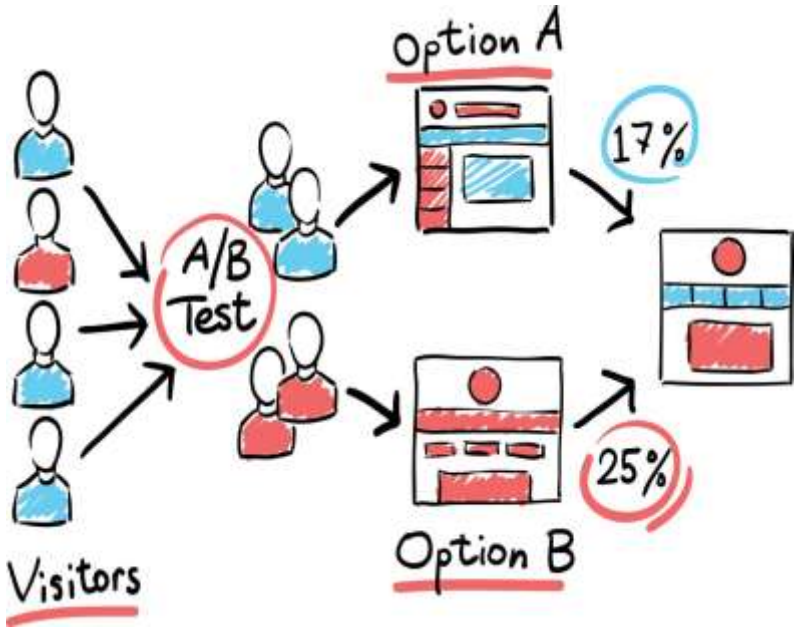
# A/B TEST

- **Definition**:
  - A/B testing, also known as split testing, is a specific experimental method used to compare two or more variations of an element (e.g., webpage, email, advertisement) to identify the most effective version.
- **Purpose**:
  - The primary purpose of A/B testing is to optimize a specific variable by evaluating which version leads to better outcomes. It is often used in marketing, user experience design, and website optimization to make data-driven decisions.
- **Experiment Design:**
  - A/B testing typically involves dividing participants or users into two (or more) groups and exposing each group to a different variation (A and B) of the element being tested. The performance of each variation is then compared based on a predefined metric.
- **Example**:
  - In website design, A/B testing may involve showing half of the website visitors the original webpage layout (Version A) and the other half an updated version with some changes (Version B). The performance metric, such as click-through rate or conversion rate, is then analyzed to determine which version performs better.



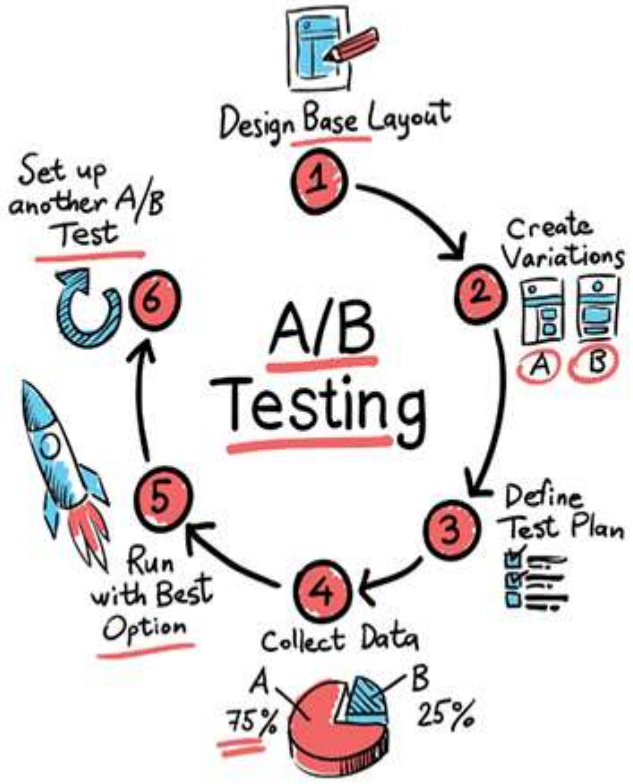
# WHEN TO USE A/B TEST ?

- A/B testing can be used in various scenarios, including:
  - Optimize websites, apps, and marketing campaigns.
  - Test new features and design elements.
  - Evaluate machine learning models and algorithms.
  - Improve user experience and conversion rates.



# PROCEDURE OF A/B TESTING

- Steps in **A/B Testing**:
  - **Define the Objective:** Clearly outline the specific metric you want to improve and set a well-defined goal for your A/B test.
  - **Create Variations:** Design two or more versions (A and B) of the element you want to test. Ensure that the variations are distinct and have a single, isolated difference between them.
  - **Randomly Assign Participants:** Randomly divide your audience or users into groups and ensure they are mutually exclusive. Each group will see only one variation.
  - **Run the Test:** Implement the variations and collect data on the chosen metric for a predetermined period. The test should be conducted simultaneously for both groups to avoid time-based biases.
  - **Analyze the Results**: Compare the performance of variation A and B based on the collected data. Use statistical analysis to determine if the observed differences are significant or due to random chance.
  - **Draw Conclusions**: Based on the results, decide which variation performed better and achieved the defined objective.
  - **Implement the Winner**: Implement the better-performing variation and monitor its performance to ensure it delivers the expected results.



# IMPETRATING THE RESULTS

- **Statistical Significance:**
  - Results must be statistically significant to be reliable.
  - This means that the differences observed between the variations are not due to chance but are likely to be representative of the larger population.
- **Sample Size:**
  - A larger sample size generally provides more reliable results.
  - Smaller sample sizes can lead to inconclusive or misleading outcomes.
- **Practical Significance:**
  - Even if a result is statistically significant, consider its practical impact on your business objectives.
  - Sometimes, small changes may not be worth implementing if they don't make a meaningful difference.

# A/B FOR DATA SCIENCE

- Tips for A/B Testing in Data Science:
  - Ensure that the sample sizes for each variation are large enough to produce reliable results.
  - For machine learning A/B tests, consider using cross-validation to obtain more robust performance estimates.
  - Monitor potential confounding factors and external variables that could influence the results.
  - Be mindful of ethical considerations, especially when conducting A/B tests involving human participants.
  - Document the A/B testing process thoroughly to facilitate replication and transparency.

# EXAMPLE

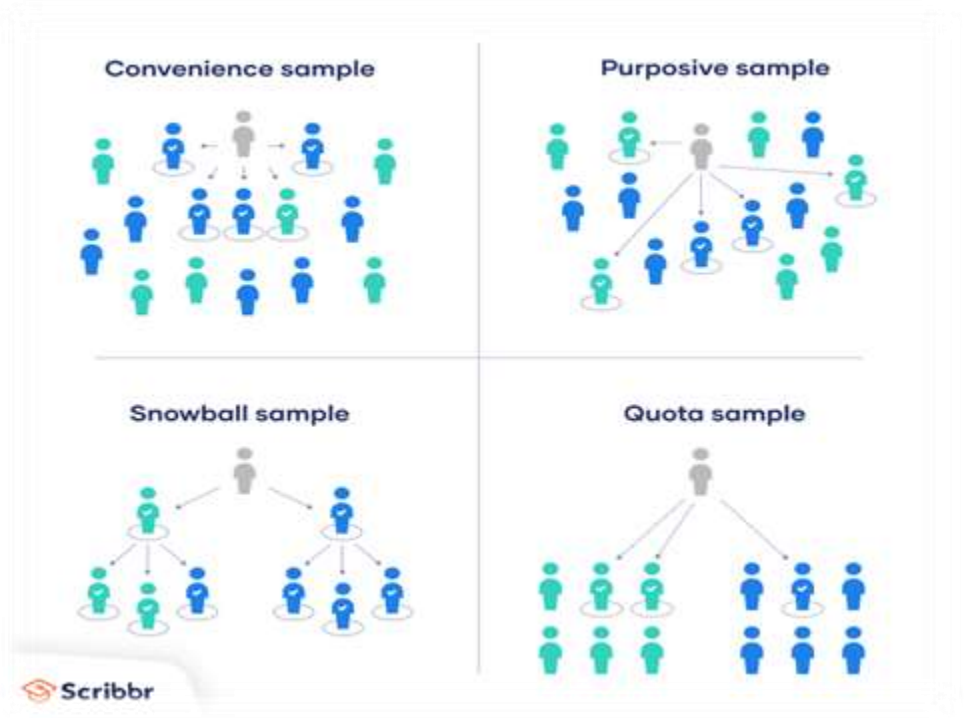
- Objective: Increase the click-through rate (CTR) of the marketing email.
  - **Step 1:** Define the Objective The objective is to improve the click-through rate (CTR) of the marketing email sent to a target audience.
  - **Step 2:** Identify the Variations
    - Variation A: The original email with the existing design and content.
    - Variation B: An updated version of the email with a more compelling subject line and a different call-to-action (CTA) button color.
  - **Step 3:** Randomly Assign Participants Suppose we have a total of 20,000 subscribers in our email list. We randomly split them into two groups:
    - Group A (Variation A): 10,000 subscribers
    - Group B (Variation B): 10,000 subscribers
  - **Step 4:** Run the Test Send the emails to the respective groups simultaneously. Group A receives Variation A email, and Group B receives Variation B email.

# EXAMPLE

- **Step 5:** Collect Data and Analyze Results After running the campaign, we collect the data on the number of clicks for each variation:
  - Group A (Variation A): 300 clicks out of 10,000 emails (CTR = 3%)
  - Group B (Variation B): 500 clicks out of 10,000 emails (CTR = 5%)
- **Step 6:** Statistical Analysis To determine if the difference in CTR is statistically significant, we perform a chi-squared test.
  - Let's assume the chi-squared test results in a p-value of 0.002 (indicating a statistically significant difference between the variations).
- **Step 7**: Draw Conclusions With a statistically significant difference in CTR between Variation A and Variation B, we can conclude that Variation B's changes had a meaningful impact on the CTR. Therefore, Variation B is the preferred choice for improving the email's performance.
- **Step 8**: Implement the Best Variation Implement the changes from Variation B (compelling subject line and different CTA button color) in future marketing emails to improve the overall click-through rate.
- **Step 9**: Continuously Improve Monitor the performance of the updated email design and continue to run A/B tests on other elements (e.g., email content, visuals, etc.) to further optimize the marketing campaign's effectiveness.

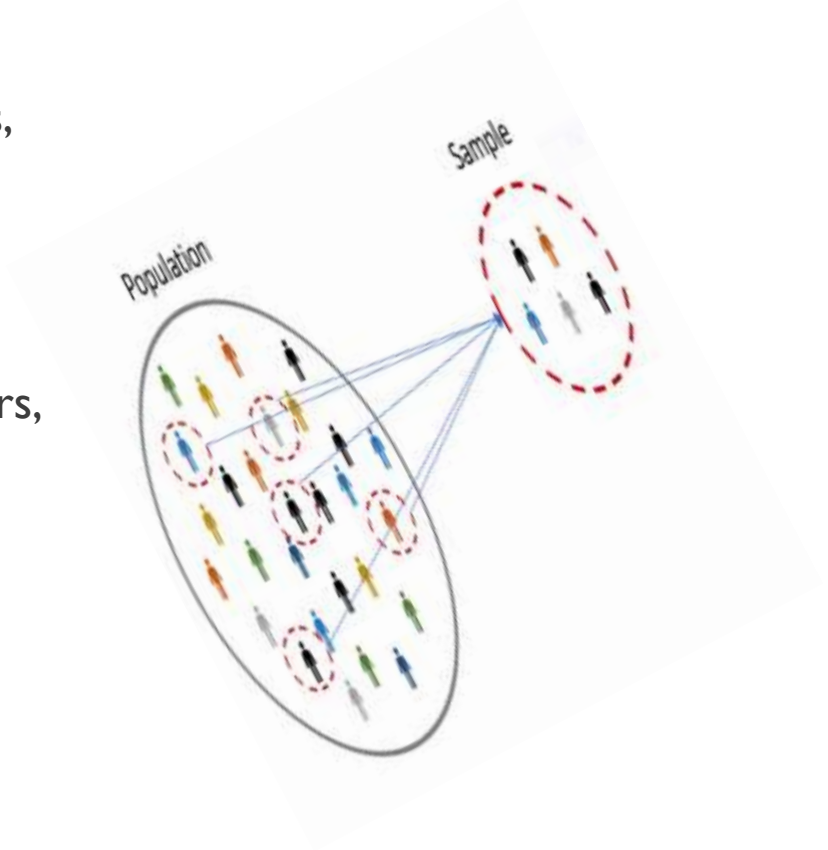
# RANDOMIZATION

- **Randomization** is a fundamental concept in experimental design, and its importance lies in its ability to reduce biases and enhance the validity of experimental results.
  - It ensures that the experimental and control groups are **comparable before the treatment or intervention** is applied.



# IMPORTANCE OF RANDOMIZATION

- **Minimizing Bias**: Random assignment ensures equal chances for participants, reducing selection and experimenter biases.
- **Establishing Causal Relationships**: It isolates the independent variable as the only difference between groups, helping establish cause-and-effect.
- **Controlling Confounders**: Randomization evenly distributes external factors, strengthening the experiment's validity.
- **Enhancing Generalizability**: Results from randomized groups are more representative of the larger population.
- **Statistical Validity**: It supports valid statistical analysis, enabling reliable conclusions about causality.



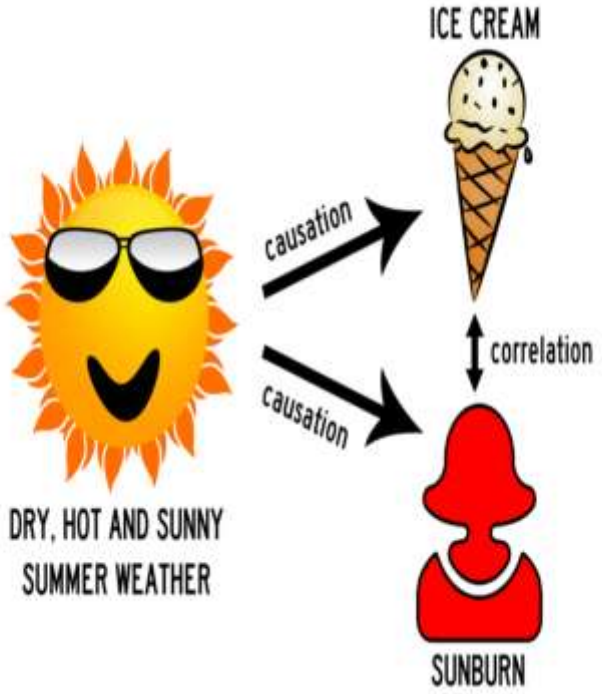
# EXAMPLE OF RANDOMIZATION

## Testing a New Recommendation Algorithm for E-commerce

- **Participants**:
  - The experiment involves 1,000 users of the e-commerce platform who have agreed to participate in the study.
- **Randomization**:
  - The users are randomly assigned to two groups using a randomization technique.
  - Half of the users (500) are assigned to the experimental group, and the other half (500) are assigned to the control group.
- **Experimental Group (Treatment Group):**
  - This group uses the platform with the new recommendation algorithm, which suggests products based on updated machine learning techniques.
  - Control Group:This group uses the platform with the current recommendation algorithm as a baseline for comparison.
- **Blinding**:
  - To avoid bias, users are not informed about whether they are experiencing the new or the current algorithm.
- **Data Collection:**
  - Data such as click-through rates, items added to the cart, and purchases are collected for both groups over two weeks.
- **Analysis**:
  - After the experiment, sales data from both groups are analyzed to determine if the new algorithm resulted in a statistically significant increase in sales compared to the current algorithm.

# CAUSALITY VS CORRELATION

- **Causality** and correlation are two distinct concepts used to describe the relationships between variables in research and data analysis.
  - Understanding the difference between causality and correlation is crucial for drawing accurate conclusions from observational studies and experimental research.
- **Correlation** indicates a statistical relationship between two variables, while causality establishes a cause-and-effect relationship between them.
  - While correlation is important for identifying associations and patterns in data causality requires rigorous experimentation and the elimination of alternative explanations to establish a clear cause-effect relationship between variables.



# CORRELATION

- **Correlation** refers to a statistical relationship between two variables where changes in one variable are associated with changes in the other variable. When two variables are correlated, they tend to move together or show a consistent pattern, but this does not necessarily imply a cause-and-effect relationship.
- It is quantified using correlation coefficients such as **Pearson's correlation coefficient or Spearman's rank** correlation coefficient.
- The correlation coefficient ranges from -1 to +1, where:
  - a positive value indicates a positive correlation (both variables increase together),
  - a negative value indicates a negative correlation (one variable increases while the other decreases),
  - a value close to 0 indicates little to no correlation.

# EXAMPLE OF CORRELATION

- **Scenario**: A researcher examines the relationship between daily hours of sunlight and ice cream sales in a city over several months.
- **Data Collection:** The researcher collects data on daily hours of sunlight and daily ice cream sales from various ice cream parlors in the city.
- **Results**: The data analysis shows a positive correlation between daily hours of sunlight and ice cream sales. On days with more sunlight, ice cream sales tend to be higher, and on days with less sunlight, ice cream sales tend to be lower.
- **Interpretation**: In this scenario, a positive correlation between daily hours of sunlight and ice cream sales is observed. However, it is important to note that correlation does not imply causation. While the data shows that the two variables are related, it does not necessarily mean that increased sunlight directly causes higher ice cream sales or vice versa.

# CAUSALITY

- **Causality** refers to a cause-and-effect relationship between two variables, where changes in one variable directly influence or cause changes in the other variable. Causality implies that one variable is responsible for producing changes in the other variable.
- Establishing causality **requires** evidence from rigorous experimental designs, such as randomized controlled trials (RCTs), where one variable is manipulated (independent variable) to observe its effect on the other variable (dependent variable) while controlling for confounding variables.
- **The three criteria for establishing causality** are:
  - temporal precedence (the cause precedes the effect in time),
  - association (a correlation exists between the variables),
  - and ruling out alternative explanations (confounding variables are controlled or eliminated).

# EXAMPLE OF CAUSALITY

- **Scenario**: A researcher wants to investigate whether a new exercise program causes improvements in cardiovascular fitness.
- **Experimental Design:** The researcher recruits two groups of participants with similar characteristics and fitness levels.
- **Group A (Experimental Group):** This group participates in the new exercise program for 8 weeks, engaging in regular aerobic and strength-training exercises.
- **Group B (Control Group):** This group does not participate in the new exercise program and maintains their regular daily activities during the 8 weeks.
- **Data Collection:** At the end of the 8-week period, the researcher measures the cardiovascular fitness of both groups using standardized fitness tests.
- **Results**: The data analysis reveals that Group A, the one exposed to the new exercise program, shows a statistically significant improvement in cardiovascular fitness compared to Group B.

# KEY DIFFERENCES

- **Nature of Relationship:**
  - Correlation: A statistical relationship between variables, showing how they are associated and vary together.
  - Causality: A cause-and-effect relationship, indicating that changes in one variable directly cause changes in another variable.
- **Directionality:**
  - Correlation: Indicates the direction of the relationship (positive or negative), but does not imply a cause-effect direction.
  - Causality: Establishes a specific cause-effect direction, with one variable being the cause and the other being the effect.
- **Experimental Evidence:**
  - Correlation: Can be determined from observational studies or data analysis but does not provide evidence of causation.
  - Causality: Requires evidence from well-designed experimental studies to establish causation.
- **Temporal Relationship:**
  - Correlation: Does not necessarily imply a temporal sequence; variables could change together simultaneously.
  - Causality: Requires a temporal sequence, where the cause precedes the effect.

# STATISTICAL TESTS

- **Statistical tests** are methods used in data analysis to draw conclusions from data, determine the significance of relationships or differences, and make inferences about populations based on sample data. These tests help researchers and analysts make data-driven decisions and draw reliable conclusions from their findings.
- **Statistical tests** play a vital role in various fields, including scientific research, social sciences, healthcare, business, and more. They enable researchers to make evidence-based decisions, test hypotheses, identify significant relationships, and generalize findings to broader populations. However, it's crucial to choose the appropriate statistical test based on the research question, data type, and assumptions to ensure accurate and valid conclusions.

# HYPOTHESIS TESTING

- **Hypothesis** testing is a fundamental statistical technique used to test hypotheses about population parameters based on sample data.
- It involves formulating a null hypothesis (H0) that represents the status quo or no effect, and an alternative hypothesis (Ha) that suggests a specific effect or relationship.
- Common hypothesis tests include t-tests, chi-square tests, ANOVA (Analysis of Variance), and z-tests.

# TYPE OF TESTS

- **Parametric Tests**:
  - Parametric tests assume that the data follow a specific probability distribution, typically the normal distribution.
  - Examples of parametric tests include t-tests, ANOVA, and Pearson's correlation coefficient.
- **Non-Parametric Tests**:
  - Non-parametric tests do not assume a specific probability distribution and are used when the data do not meet the assumptions of parametric tests.
  - Examples of non-parametric tests include Mann-Whitney U test, Wilcoxon signed-rank test, and Kruskal-Wallis test.

# CONCLUSION

- **Causal Inference:** Enables identification of cause-and-effect relationships.
- **Bias Reduction**: Randomization minimizes selection bias for reliable results.
- **Baseline Comparisons:** Control groups ensure observed changes are due to interventions.
- **Real-World Validation:** A/B testing evaluates algorithms in practical scenarios.
- **Generalizability**: Improves applicability of findings to larger populations.
- **Optimization**: Drives continuous system and algorithm improvements.
- **Evidence-Based Decisions**: Supports data-driven and actionable insights.