Lecture: 03

Experimental Design

Today’s Objectives

Today, we’ll **cover the basics of experimental design**, including how to plan and conduct experiments.

The goal is to help you design and analyze experiments more effectively.

- Learn to identify variables, hypotheses, and confounding factors.

❑ What is Experimental Design?

❑ Variables & Hypothesis

❑ Confounders & Bias

❑ Briefly: What is Hypothesis

❑ Data Collection Methods

❑ Case Study: Online Retail CTR

Experimental design

The science and subfield of statistics about how to collect data effectively…

R. A. Fisher (1890-1962)
Founding father of Modern Statistics
Geneticist

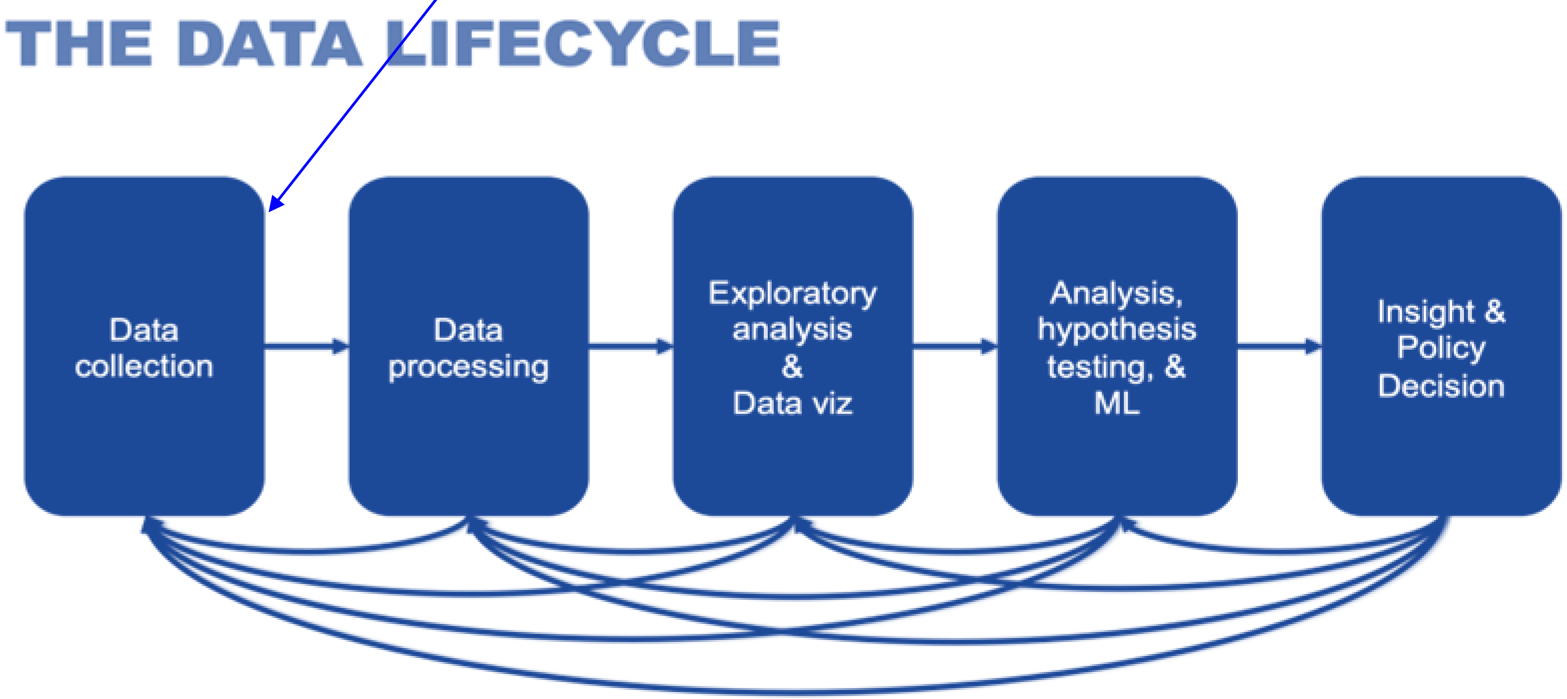
"There's a flaw in your experimental design. All the mice are scorpios."

Data science fundamentally involves making decisions **based on data**.

Experiments in Data Science require proper planning and design to best answer its questions.

Experimental Design in Data Science?

The process of **planning, conducting, and analyzing** experiments to test hypotheses and gather **meaningful data** for **data-driven decisions**.



THE DATA LIFECYCLE

Data collection
Data processing
Exploratory analysis & Data viz
Analysis, hypothesis testing, & ML
Insight & Policy Decision

- **Involves planning** how to collect, manipulate, and analyze data to answer specific research questions or test hypotheses (e.g., maximize CTR).
- **Goal:** Gather reliable, unbiased data for valid conclusions.

Maximizing the amount of data that can be gathered from an experiment while minimizing the time, costs, and mistakes that are involved is the goal of Data Science and experimental design.

Define the Problem or Research Question aims to address

What will the weather be like for the next 10 days?

Which courses are likely to have the highest demand next semester?

How many visitors will the website named "X" receive over the next week?

Which students are at risk of dropping out of an online course?

Which loan applicants are likely to default in the next year?

Upcoming movie "Avatar 4" will be a box office hit or miss?

Will flights be delayed over the next 10 days?

Which subscribers are likely to cancel their Netflix subscription soon?

Asking the right questions before solving a DS problem is a great start! And be specific!

$\underbrace{\hphantom{xxxxxxxxxx}}$

Need proper planning and good **experimental design**.

Remember that

For most data science applications, ultimately, comes down to predicting the future.

We answer the question, “Given some set of options, which option **maximizes** my **optimization criteria**.”

Objective or goal function → we want to *achieve*

Identify the option/ choice/ decision that maximizes or optimizes the desired outcome based on these predefined criteria.

- An example of Experimental Design (ED)

Identifying Variables & Population/Sample of the Study

Independent variable

Dependent variables

Hypothesis

A potential problem in ED: Confounder Variable

How to Deal with Confounder

Control

Randomization

Replication

Methods for Collecting Data

Observational studies

Cross sectional studies

Retrospective (case control) studies

Prospective (longitudinal or cohort) studies

Surveys.

Experiments

Placebo Effect

A common method to minimize bias in Experimental Design

Simulations

Example: Online Retail

Let's say you're a data scientist working for an online retailer, and you want to test whether changing the color of the "Buy Now" button on your website affects the click-through rate (CTR)

- Click-through rate (CTR) → the percentage of users who click on the ad after seeing it

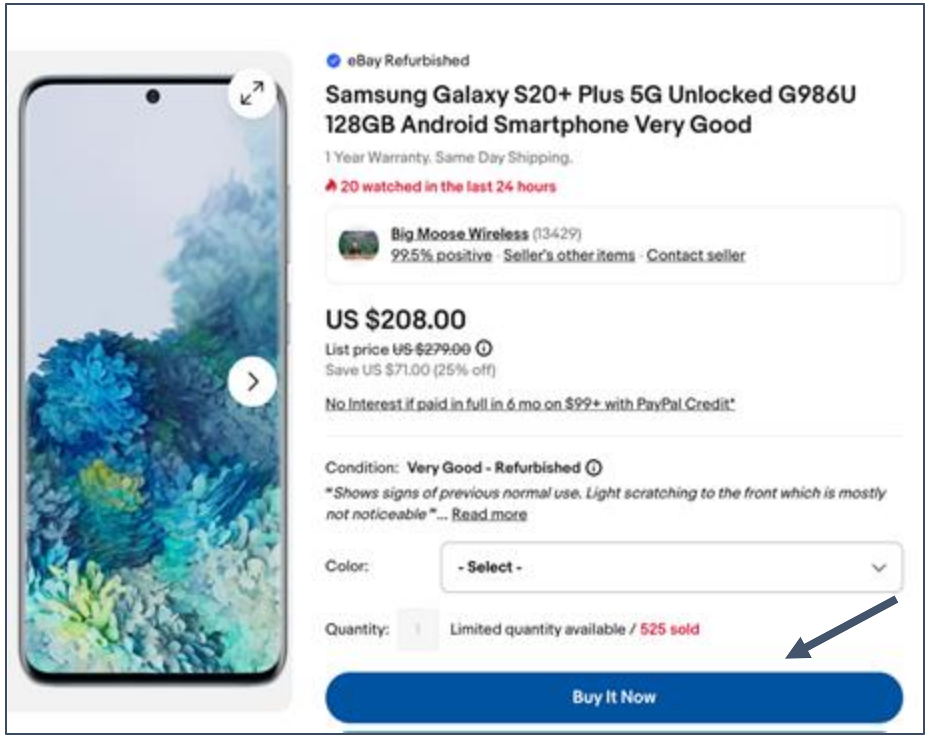
What is your problem definition?

Find which version of the button “Buy Now” (Option A (default) or Option B (red)) is more likely to maximize the CTR

What is your Optimization Criteria? What we want to maximize?

CTR → we want to select the ad options with button “buy now” that leads to **Higher CTR**

Ques: How can we set up an experiment to collect data in this case?



eBay Refurbished

Samsung Galaxy S20+ Plus 5G Unlocked G986U 128GB Android Smartphone Very Good

1 Year Warranty. Same Day Shipping.

20 watched in the last 24 hours

Big Moose Wireless (13429)
99.5% positive · Seller's other items · Contact seller

US $208.00

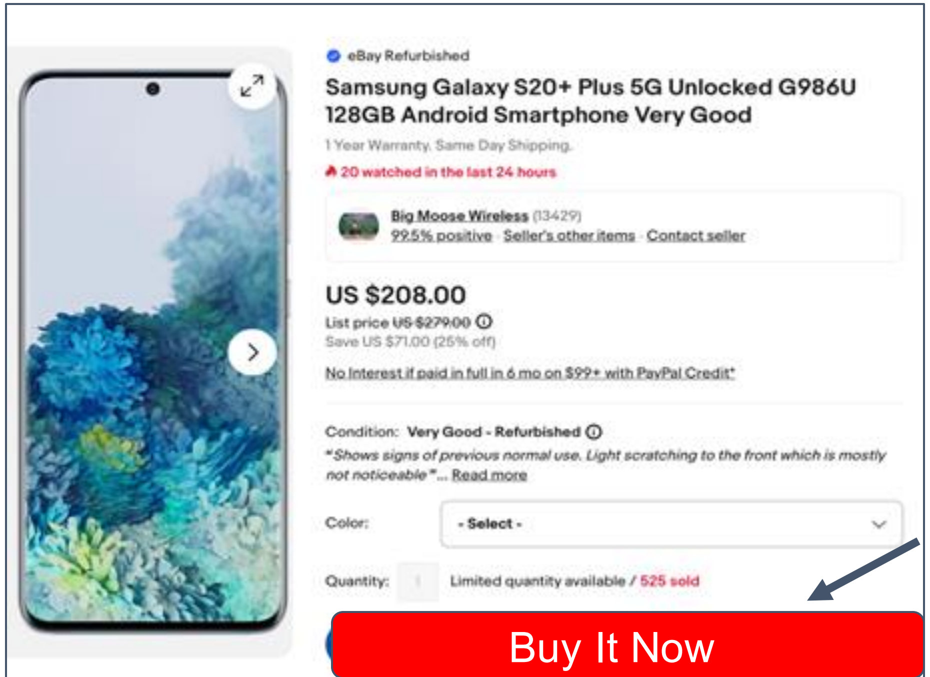
List price ~~US $279.00~~
Save US $71.00 (25% off)

No Interest if paid in full in 6 mo on $99+ with PayPal Credit*

Condition: **Very Good - Refurbished**
"Shows signs of previous normal use. Light scratching to the front which is mostly not noticeable"... Read more

Color: - Select -

Quantity: 1 Limited quantity available / 525 sold



eBay Refurbished

Samsung Galaxy S20+ Plus 5G Unlocked G986U 128GB Android Smartphone Very Good

1 Year Warranty. Same Day Shipping.

20 watched in the last 24 hours

Big Moose Wireless (13429)
99.5% positive · Seller's other items · Contact seller

US $208.00

List price ~~US $279.00~~
Save US $71.00 (25% off)

No Interest if paid in full in 6 mo on $99+ with PayPal Credit*

Condition: **Very Good - Refurbished**
"Shows signs of previous normal use. Light scratching to the front which is mostly not noticeable"... Read more

Color: - Select -

Quantity: 1 Limited quantity available / 525 sold

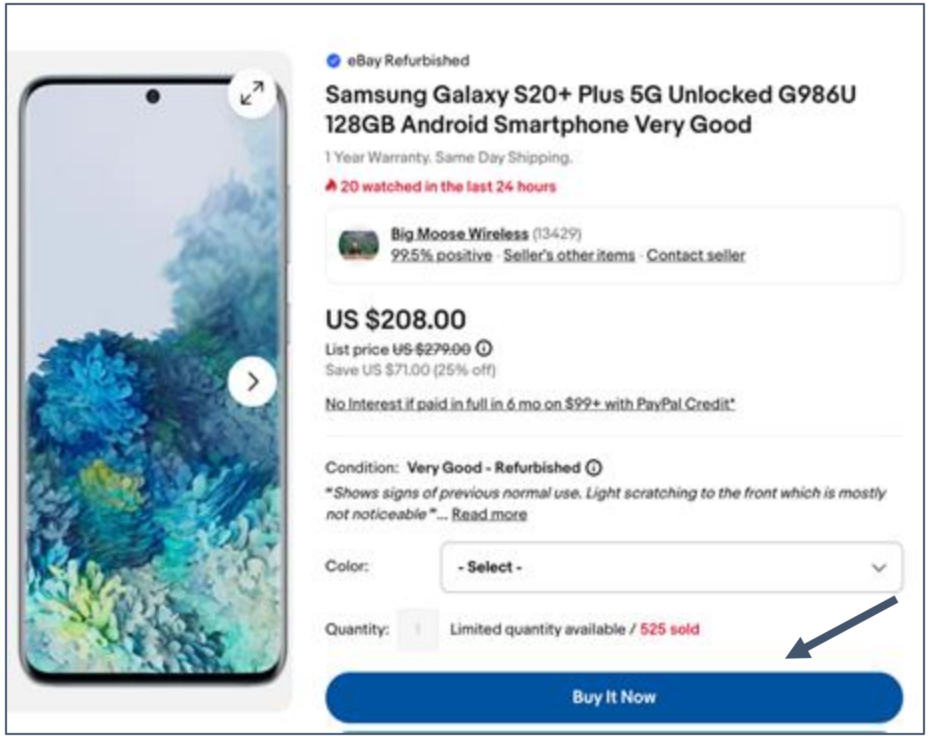
Ques: How can we set up an experiment to collect data in this case?

Data Size / Sample ? → No. of website visitors

- → Views the original website with **existing button color**.
- → Experiences no changes; baseline for comparison.

- → Sees the same website but with a **different color for the "Buy Now"** button.
- → Experiences the change you want to test.

- **Collect data** on the click-through rates for both groups over a specific period.
- **Compare** the click-through rates (CTR) between the groups **after** the experiment
- **Check** if there's a significant difference, we can infer that the **change in button color influenced the click-through rate**.



eBay Refurbished

Samsung Galaxy S20+ Plus 5G Unlocked G986U 128GB Android Smartphone Very Good

1 Year Warranty. Same Day Shipping.

20 watched in the last 24 hours

Big Moose Wireless (13429)
99.5% positive · Seller's other items · Contact seller

US $208.00

List price ~~US $279.00~~

Save US $71.00 (25% off)

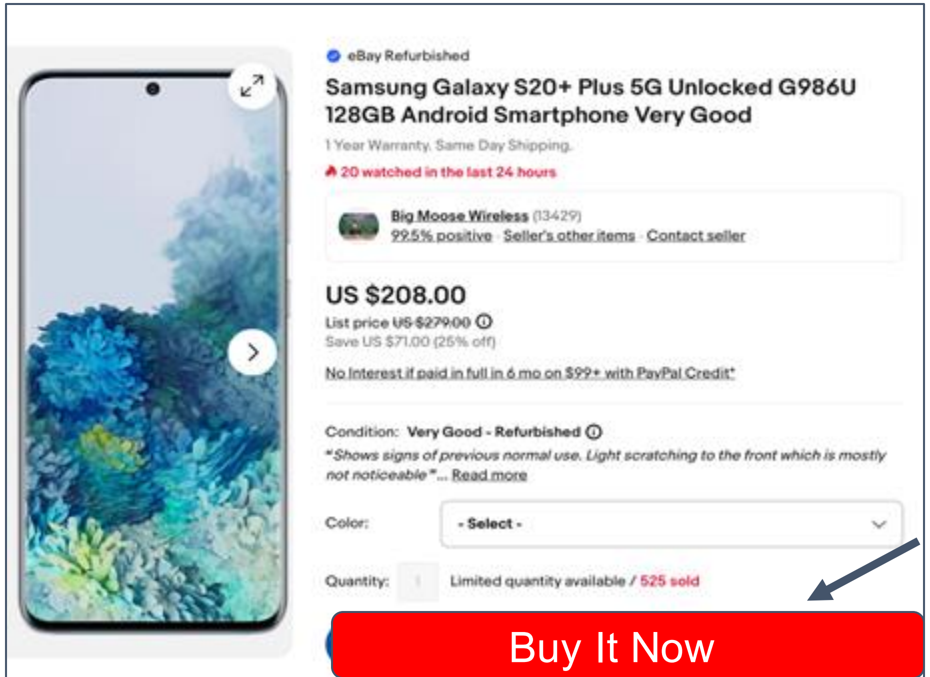
No Interest if paid in full in 6 mo on $99+ with PayPal Credit*

Condition: Very Good - Refurbished

"Shows signs of previous normal use. Light scratching to the front which is mostly not noticeable"... Read more

Color: - Select -

Quantity: 1 Limited quantity available / 525 sold



eBay Refurbished

Samsung Galaxy S20+ Plus 5G Unlocked G986U 128GB Android Smartphone Very Good

1 Year Warranty. Same Day Shipping.

20 watched in the last 24 hours

Big Moose Wireless (13429)
99.5% positive · Seller's other items · Contact seller

US $208.00

List price ~~US $279.00~~

Save US $71.00 (25% off)

No Interest if paid in full in 6 mo on $99+ with PayPal Credit*

Condition: Very Good - Refurbished

"Shows signs of previous normal use. Light scratching to the front which is mostly not noticeable"... Read more

Color: - Select -

Quantity: 1 Limited quantity available / 525 sold

Ques: How can we set up an experiment to collect data in this case?

Data Size / Sample ? → No. of website visitors

Views the original website with **existing button color**. This group experiences no changes.

Sees the same website but with a **different color for the "Buy Now".** This is the group that experiences the change you want to test.

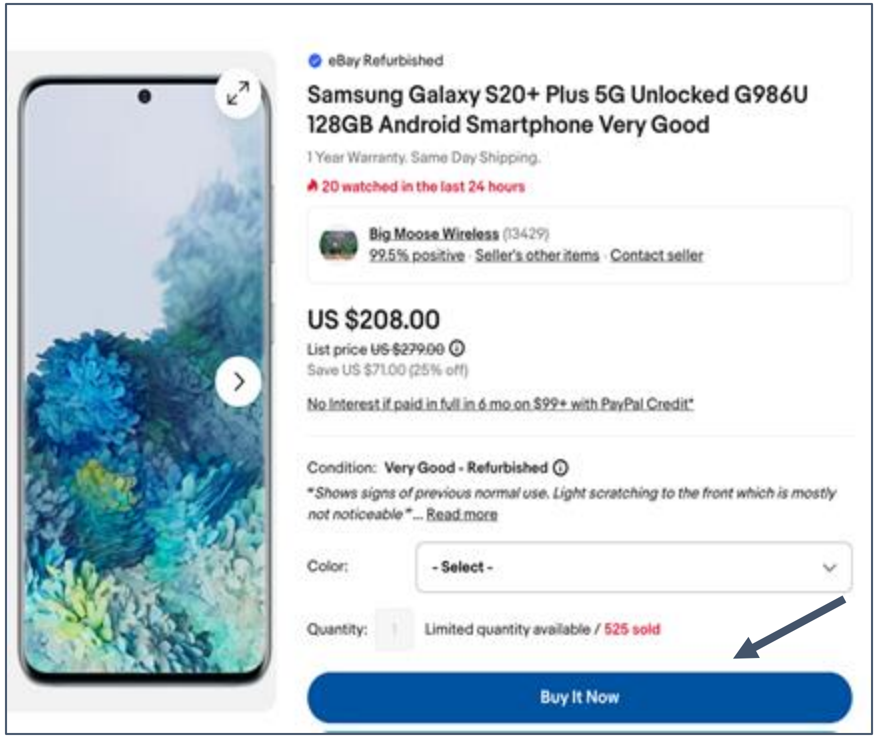
CONTROL GROUP

TREATMENT GROUP

Question: What are the variables here?

- Click-through rate(what we measure: outcome) ← Dependent Variable
- Color of the "Buy Now" (what we manipulate) ← Independent Variable

Draw more reliable conclusions about the impact of the independent (manipulated) variable.



eBay Refurbished

Samsung Galaxy S20+ Plus 5G Unlocked G986U 128GB Android Smartphone Very Good

1 Year Warranty. Same Day Shipping.

20 watched in the last 24 hours

Big Moose Wireless (13429)
99.5% positive · Seller's other items · Contact seller

US $208.00
List price ~~US $279.00~~
Save US $71.00 (25% off)

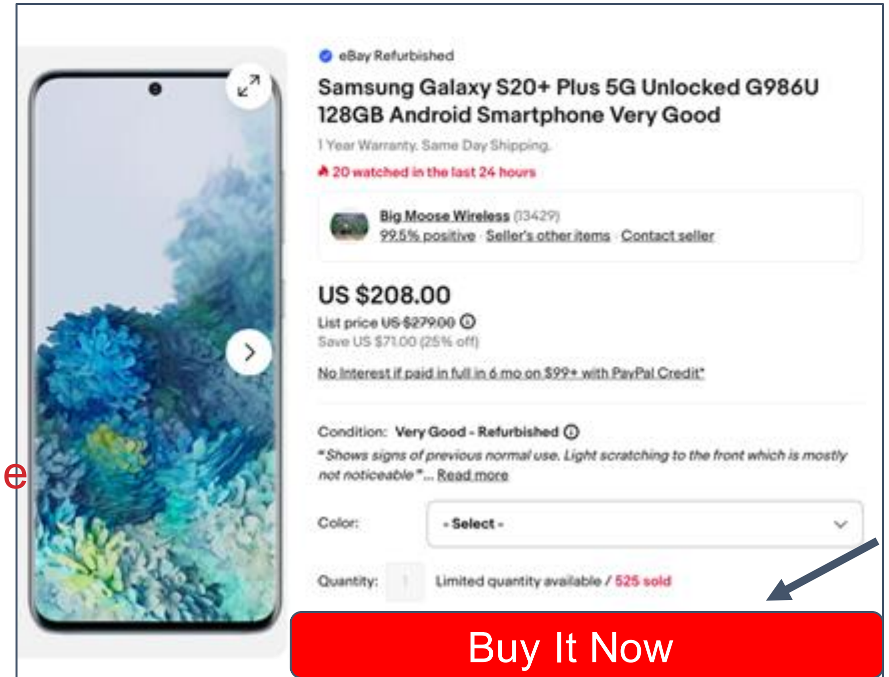
No Interest if paid in full in 6 mo on $99+ with PayPal Credit*

Condition: Very Good - Refurbished
"Shows signs of previous normal use. Light scratching to the front which is mostly not noticeable"... Read more

Color: - Select -

Quantity: 1 Limited quantity available / 525 sold

Buy It Now



eBay Refurbished

Samsung Galaxy S20+ Plus 5G Unlocked G986U 128GB Android Smartphone Very Good

1 Year Warranty. Same Day Shipping.

20 watched in the last 24 hours

Big Moose Wireless (13429)
99.5% positive · Seller's other items · Contact seller

US $208.00
List price ~~US $279.00~~
Save US $71.00 (25% off)

No Interest if paid in full in 6 mo on $99+ with PayPal Credit*

Condition: Very Good - Refurbished
"Shows signs of previous normal use. Light scratching to the front which is mostly not noticeable"... Read more

Color: - Select -

Quantity: 1 Limited quantity available / 525 sold

Buy It Now

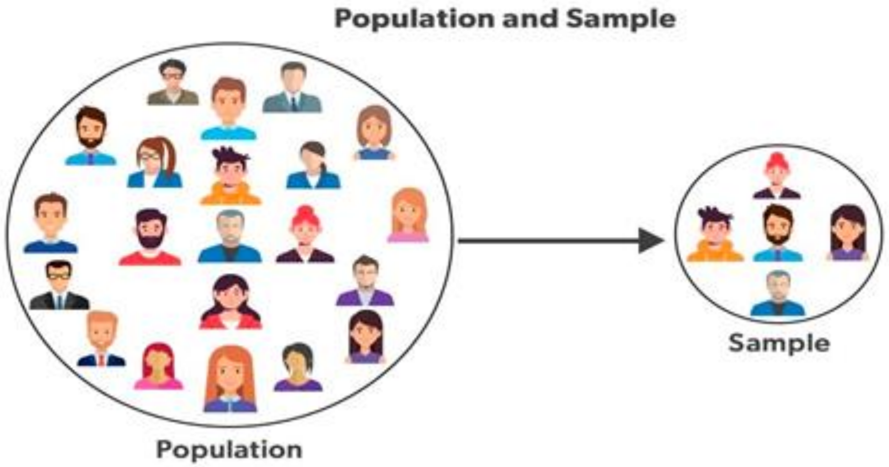
- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Identify the variable (or variables) of interest and the population of the study.

Once the problem is defined, identify the variable(s) of interest that are relevant to your research question.

- ➔ Independent variable (IV)
- ➔ Dependent variable (DV)

Also, specify the **population or sampl**e that your study will focus on.



Identify the variable (or variables) of interest: IV and DV

2 types:

- **Independent variable (IV):** The variable that is manipulated or changed by the researcher.
  - Example: A new algorithm, a marketing strategy, or a drug dosage
  - Why manipulate? To observe its effect on the **dependent variable (DV)**.
- **Dependent variable (DV):** outcome of interest (what we measure)
  - Expected to change as a result of changes in the independent variable
  - Example: User engagement, sales, or patient recovery rate.

- An example of Experimental Design (ED)
  Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables

  Hypothesis

  A potential problem in ED: Confounder Variable

  How to Deal with Confounder
  - Control
  - Randomization
  - Replication

  Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Come up with a Hypothesis

The hypothesis of your experiment is the statement you want to test.

“if hypothesis X is right, then Y should be true…..”

- **A testable explanation** that provides an **educated guess** about the relationship between ***variables and outcomes*** of the experiment**.**

How to Test a Hypothesis is correct? Conduct experiments or make observations.

- If results match prediction, the hypothesis is supported;
- If not, revise the hypothesis or form a new one.

Example (1)

Hypothesis:

"If the amount of **average study time** (___ variable?) is increased, then **average exam scores** (___ variable?) will also increase."

Recap:

- **Independent variable (IV):** are the factors/cause that may influence the outcome.
- **Dependent variable (DV):** outcome of interest (what we measure)

Examples (2)

Hypothesis:

As **books read** increases, **average literacy** also increases

If the **exercise duration** is extended, then the **average calories burned** will also increase

If the **temperature** rises, then **average ice cream sales** will also increase.

Ques: What are the optimization goal/ criteria here? How independent and dependent variables are related?

Brainstorming Time!

- While independent variables affect dependent variables, multiple independent variables ________________ (can / can not) influence each other.
- Do you think is it a problem or not? (Yes/No)
  - Why?

Good experimental design aims to **minimize correlated variables**.

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Next:

Select a Data Collection Method and Collect Data to Test your Hypothesis

Before collecting data to test your hypothesis, you first have to consider the problems that can cause errors in your result, one of them being a **confounder**.

Analyze if any other variables (beyond the ones you intended to manipulate or measure) could have impacted / influenced your results (Note that, if is true, we may need to redesign the experiment for ensuring more accurate and reliable conclusions).

What is a confounder?

Extraneous variables/ External factors (e.g., seasonality, user demographics) that may influence the outcome (DV).

- Can affect IV-DV relationships and distort results if not controlled.
- Not the study focus but must be managed for valid conclusions.

A **confounder** can **lead to incorrect conclusions** about the true effect of the independent variable on the dependent variable.

Examples: Find Potential Confounder

- If the **exercise duration** is extended, then the average **calories burned** will also increase → Metabolic Rate
- As **books read** increases, average **literacy** also increases →

Age, Socioeconomic status etc.

More Example: Experimental Design Flow : Polling

1. **Problem Formulation:** Imagine you're a data scientist tasked with predicting the outcome of a political election using a dataset of voter preferences. Your goal is to **design an experiment** that **accurately represents the entire population's voting behavior.**

   Questions: How do we know which candidate is ahead!

   - Can we create the IDEAL POLLING?

Eliminate confounding variables as much as possible → Sample Bias, geographic representation, Population Proportion Bias, Demographic Mismatch and many more to make the data as accurate as feasible.

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Some Ways to Deal with Confounder

- Control
- Sampling via Randomization
- Replication

1. Control

Manage or eliminate the impact of confounding variables (to isolate the effect of the independent variable(s) on the dependent variable).

Some ways:

- **Holding Variables Constant:** Keep potential confounding variables constant across all experimental conditions.
  - Example:
    - Conduct an experiment in the same season to control for seasonal effects.
    - Ensure all participants have similar health conditions when testing a new drug to avoid health status as a confounder.

1. Control

Manage or eliminate the impact of confounding variables (to isolate the effect of the independent variable(s) on the dependent variable).

Some ways:

- **Involves the use of control groups and treatment groups:** Include a control group that does not receive the experimental treatment.
    - This helps differentiate the effects of the treatment from other influences.

2. Randomization

Helps make groups fair by deciding things randomly.

Randomly assign participants to

- The control group (**won’t receive the drugs**) or
- The treatment group (**receive the drugs**)

Why? To minimize systematic confounding, reduce the risk of bias in either group being enriched for confounders, and help distribute confounding variables equally

❑ This minimizes the likelihood of systematic confounding;
  - Reduces the risk of bias in either group being enriched for confounders.
  - Each participant is assigned to one of the two groups, not both simultaneously.

3. Replication

Doing your experiment more than once to be more certain.

- Repeat the experiment multiple times to assess the consistency and reliability of results, helping to identify and account for potential confounding factors.

> If replication is done(with a new set of data) and it **produces the same conclusions**, this shows that the experiment is strong and has a good design and suggests that **confounders are less likely to be influencing** the outcomes.

Example: Control Confounder Variable in an Experiment

Example: Control Confounder Variable in an

If the amount of **study time** (independent variable) is increased, then **exam scores** (dependent variable) will also increase.

Potential Confounder? Prior Knowledge

Experiment Design Idea 1: Random Sampling: Randomly assign participants to two groups:

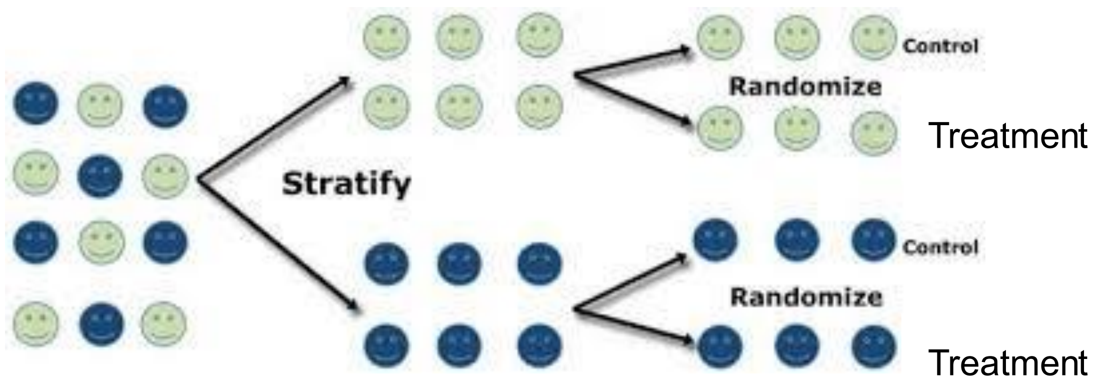
- **Control group:** Normal study time.
- **Treatment group:** Increased study time.

Randomization (Randomized Controlled Trial (RCT)) helps ensure that prior knowledge is evenly distributed between the two groups, minimizing its influence as a confounder

Example: Control Confounder Variable in an

If the amount of **study time** (independent variable) is increased, then **exam scores** (dependent variable) will also increase.

Experiment Design Idea 2: Stratified Randomization:



Stratify by prior knowledge (high, medium, low) and Randomly assign to

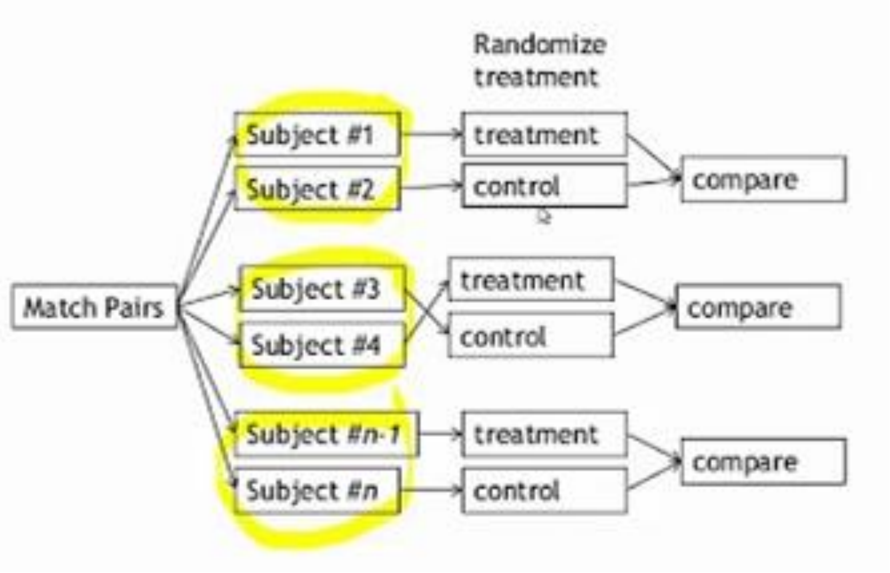
- **Control group:** Normal study time.
- **Treatment group:** Increased study time.

This ensures that prior knowledge is balanced across both groups.

Example: Control Confounder Variable in an

If the amount of **study time** (independent variable) is increased, then exam scores (dependent variable) will also increase.

Experiment Design Idea 3: Block Design (Matched Pair)



- ❑ Pair participants by prior knowledge.
- ❑ Randomly assign one to control, one to treatment.

- **Control group:** Normal study time.
- **Treatment group:** Increased study time.

This ensures that prior knowledge is controlled within each pair.

Try by yourself: control the effect of “age”

"As **books read** increases, **avg. literacy** also increases.”

We can measure the age of each individual; to see the effects of age on literacy.

Experimental Design: How to design

- Random Sampling
- Stratified Randomization
- Block Design (Match Pair)

- An example of Experimental Design (ED)
  Identifying Variables & Population/Sample of the Study
    - Independent variable
    - Dependent variables

  Hypothesis
  A potential problem in ED: Confounder Variable
  How to Deal with Confounder
    - Control
    - Randomization
    - Replication

  Methods for Collecting Data
    - Observational studies
      - Cross sectional studies
      - Retrospective (case control) studies
      - Prospective (longitudinal or cohort) studies
    - Surveys.
    - Experiments
      - Placebo Effect
      - A common method to minimize bias in Experimental Design
    - Simulations

Methods for Collecting Data

Below are some common methods/way to collect data if pre-existing datasets are not available.

- Observational studies
- Surveys.
- Experiments → ex. AB Testing
- Simulations

Methods for Collecting Data

A. Observational studies → Observe and record data (variables) without intervening or manipulating variables (Observe; don't change anything intentionally).

E.g. Observing animal behavior in a natural habitat without any external influence.

B. Surveys → Collect information through structured questionnaires or interviews.

E.g. Conducting a survey to gather opinions on a political issue.

C. Experiments → We actively change something to see what happens.

D. Simulations → Create artificial scenarios to model real-world situations for data collection.

E.g. Using a computer simulation to study traffic patterns in a city.

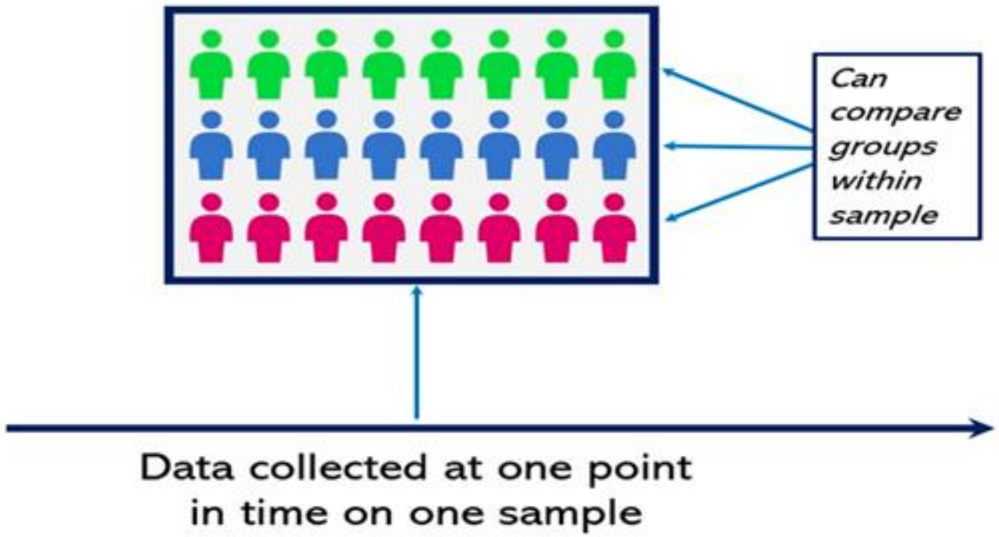
A. Observational Studies

- Cross sectional studies
- Retrospective (case control) studies
- Prospective (longitudinal or cohort) studies

A. Observational Studies

Cross sectional studies: Collects data from many different individuals at one specific single point of time

- Taking a picture of a group **right now** to see what they're like.
- Surveying individuals of various age groups **on a single day**

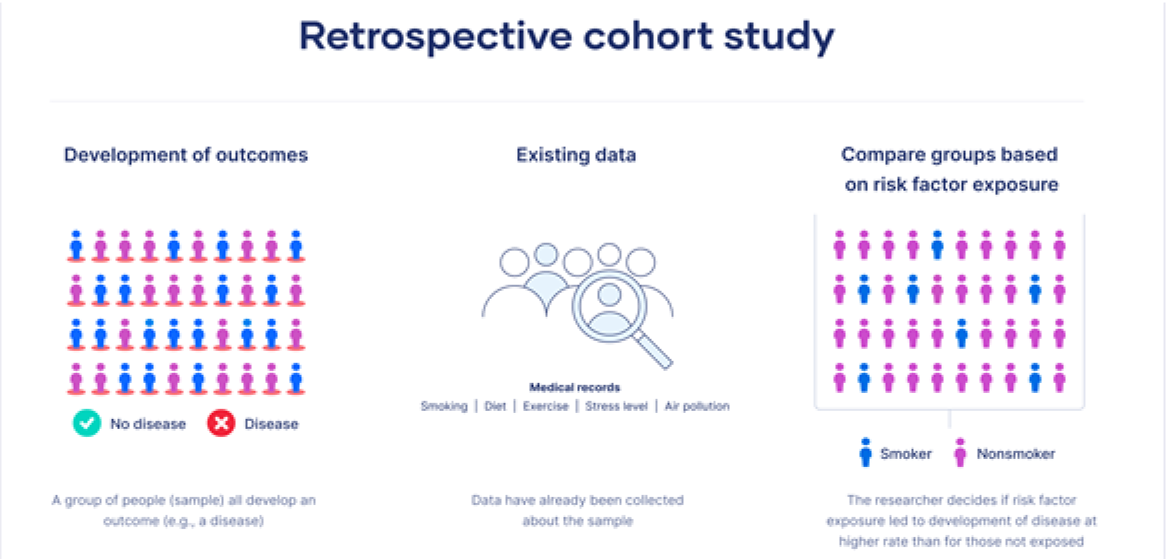

Can compare groups within sample
Data collected at one point in time on one sample

A. Observational Studies

Retrospective (case control) studies: Look back at past events to examine the relationship between exposure and outcome. **E.g., Investigating past events to identify causes.**

Interviewing severe dengue patients to study disease spread.

Studying past smoking history in lung cancer patients: Compare smoking history in cancer vs. non-cancer patients.



Retrospective cohort study

Development of outcomes
Existing data
Compare groups based on risk factor exposure
No disease
Disease
Medical records
Smoking | Diet | Exercise | Stress level | Air pollution
Smoker
Nonsmoker

A group of people (sample) all develop an outcome (e.g., a disease)

Data have already been collected about the sample

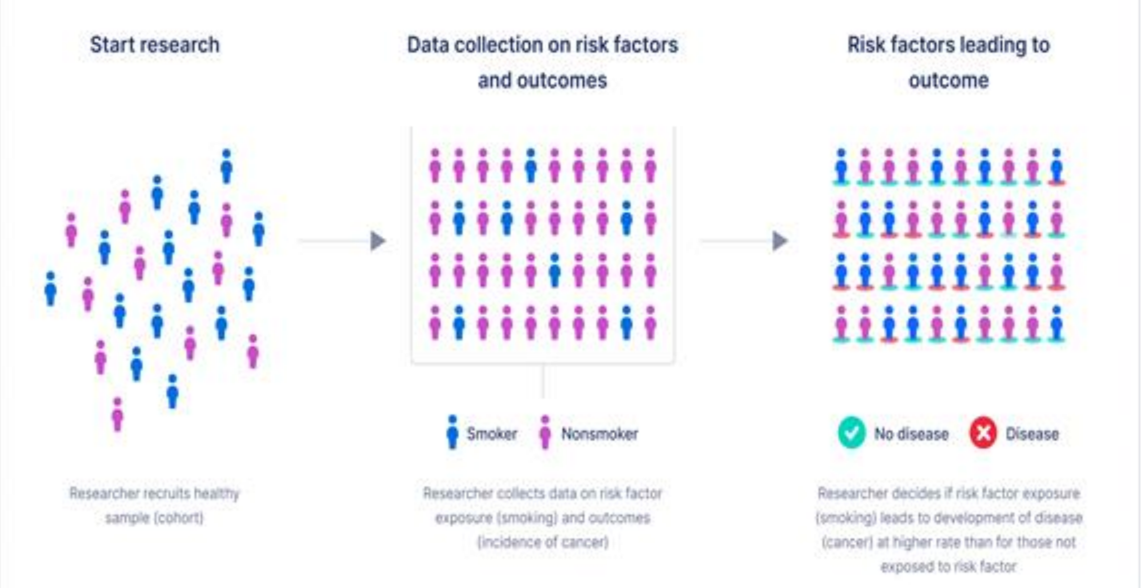
The researcher decides if risk factor exposure led to development of disease at higher rate than for those not exposed

A. Observational Studies

- **Prospective (longitudinal or cohort) studies:** Researchers follow (called a cohort) a group closely ***over a period of time*** to track exposure and identify potential causes.

Track a cohort's smoking habits and health outcomes over time to study the link between smoking and respiratory diseases.

Challenge: High dropout rates can complicate results.


Start research
Data collection on risk factors and outcomes
Risk factors leading to outcome
Smoker
Nonsmoker
✓ No disease
✗ Disease

Researcher recruits healthy sample (cohort)

Researcher collects data on risk factor exposure (smoking) and outcomes (incidence of cancer)

Researcher decides if risk factor exposure (smoking) leads to development of disease (cancer) at higher rate than for those not exposed to risk factor

B. Surveys (A specific type of observational study)

A survey is used to investigate characteristics of a population. It is frequently used when the subjects are people, and questions are asked of them.

- When designing a survey, you must be very careful of wording (and sometimes ordering) the questions so that the results are **not biased**.

C. Experiment

In an experiment, a researcher assigns **a treatment** and observes the response.

- Observe effects on subjects after the application of some treatment
  - Might want to compare a treatment versus a control or multiple treatments -

receives intervention/actual treatment

receives no intervention or receive placebo (fake treatment)

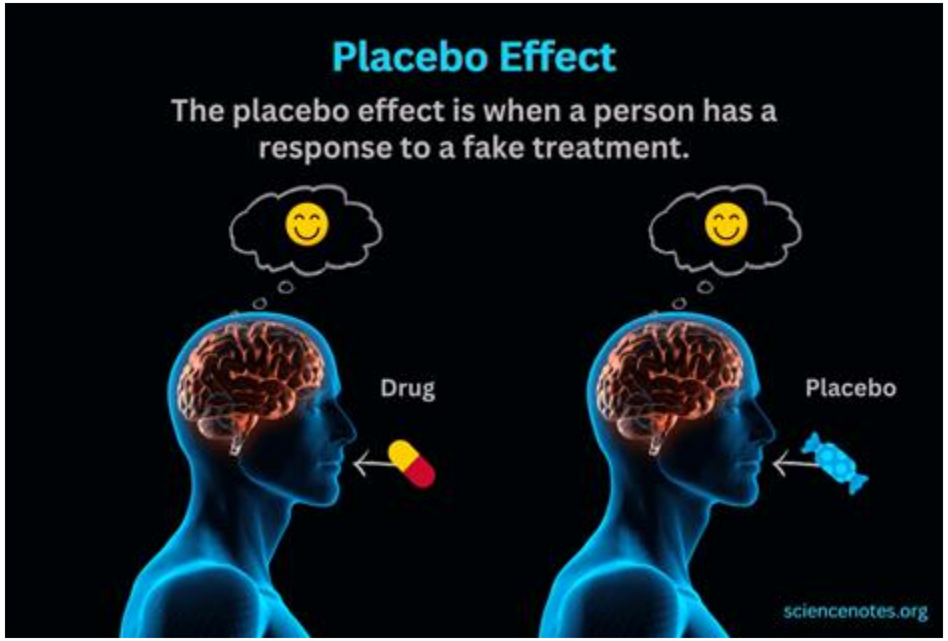
- The more variables you can control for, the better your experiment

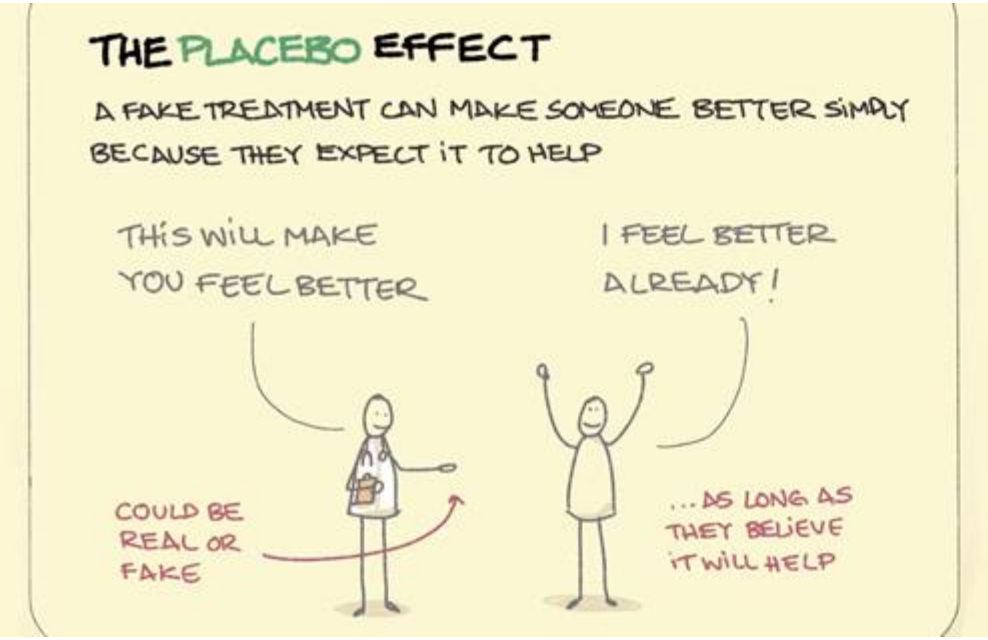
Placebo effect

Placebo Effect: Belief in treatment leads to feeling better even without actual treatment.

- A person believes psychologically that a certain treatment is positively affecting them, even though no treatment was given at all.

E.g: The placebo effect is when a person's physical or mental health appears to improve after taking a placebo or 'dummy' treatment.


Placebo Effect
The placebo effect is when a person has a response to a fake treatment.
Drug
Placebo
sciencenotes.org


THE PLACEBO EFFECT
A FAKE TREATMENT CAN MAKE SOMEONE BETTER SIMPLY BECAUSE THEY EXPECT IT TO HELP
THIS WILL MAKE YOU FEEL BETTER
I FEEL BETTER ALREADY!
COULD BE REAL OR FAKE
...AS LONG AS THEY BELIEVE IT WILL HELP

A common method to minimize bias in Experimental Design

Blinding: (Subjects would be blinded) when the people involved in an experiment **don't know who's getting the real treatment and who's not.** Prevent Biases. It is a technique used to make the subjects “blind” to which treatment (or placebo) they are being given

- **Single-blinding** is when either the participants or the researchers don't know
- **Double-blinding** is when both don't know.

The use of placebos contributes to the blinding of experiments

D. Simulation

A simulation uses a mathematical, physical, or computer model to replicate the conditions of a process or situation.

- This is frequently used when the actual situation is too expensive, dangerous, or impractical to replicate in real life

The Fundamental Rule of Data Collection

Your data must representative of the population you want to study.

Keep in mind that

It is almost impossible to be certain that your experiment has completely removed all forms of bias. It is necessary to consider possible sources of bias and highlight them in your analysis. Ideally, future experiments would improve upon your method by iteratively eliminating those sources of bias.

Quick Class Task

Identify which method for collecting data (observational study, an experiment, a simulation, or a survey) is **best** in each of the following situations and **explain your answer.**

1. The effect of a severe earthquake would have on the Northern Areas.
2. Whether or not a certain coupon attached to the outside of a package makes recipients more likely to order products from a courier company.
3. Whether or not smoking has an effect on coronary heart disease.
4. Determining the average household income of homes in Lahore.