

# Scientific Methodology in Computer Science

MO430A

Prof. Dr. Bruno B. P. Cafeo

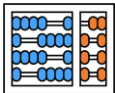
Institute of Computing  
University of Campinas



# Agenda

---

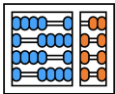
- Variables
  - Type of data
    - Quantitative
    - Categorical
  - Independent vs Dependent Variables
  - Types of Independent variables
  - Recognizing Independent variables
  - Explanatory vs Response variables
  - Mediator vs Moderator variables
- Extraneous variables
- Confounding variables
- Control variable
  - Random assignment
  - Standardized procedure
  - Statistical control
- Correlation vs Causation
- How to operationalize concepts



# Variables

---

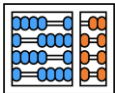
- In statistical research, a variable is defined as an attribute of an object of study.
- Choosing which variables to measure is central to good experimental design.



# Example

---

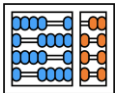
- Let's suppose you want to test the performance of different sorting algorithms. What key variables could you measure?



# Example

---

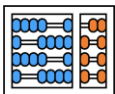
- Let's suppose you want to test the performance of different sorting algorithms. What key variables could you measure?
  - **Sorting Algorithm:** This variable represents the specific sorting method being tested, such as bubble sort, quicksort, or merge sort.
  - **Input Data:** The input data can vary in size and order. Variables related to input data might include the length of the array to be sorted and whether it is already sorted, reversed, or randomized.
  - **Execution Time:** This variable measures the time it takes for a sorting algorithm to complete the sorting process.
  - **Comparisons:** The number of comparisons made by the algorithm during the sorting process is an important variable to assess algorithm efficiency.
  - **Swaps or Exchanges:** The number of data element swaps or exchanges made by the sorting algorithm is another efficiency measure.
  - **Memory Usage:** Some sorting algorithms might require additional memory for temporary storage, so memory consumption can be a variable of interest.
  - **Stability:** This variable indicates whether the sorting algorithm maintains the relative order of equal elements in the sorted output.
  - ...



# Types of data: Quantitative vs categorical variables

---

- Data is a specific measurement of a variable – it is the value you record in your data sheet. Data is generally divided into two categories:
  - Quantitative data represents amounts
  - Categorical data represents groupings
- A variable that contains quantitative data is a quantitative variable; a variable that contains categorical data is a categorical variable.
- Each of these types of variables can be broken down into further types.

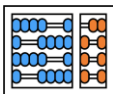


# Quantitative data

---

- The numbers you record represent real amounts that can be added, subtracted, divided, etc.
- There are two types of quantitative variables: discrete and continuous.

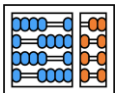
Type of variable	What does the data represent?	Examples
Discrete variables (aka integer variables)	Counts of individual items or values.	<ul style="list-style-type: none"><li>• Number of students in a class</li><li>• Number of different tree species in a forest</li></ul>
Continuous variables (aka ratio variables)	Measurements of continuous or non-finite values.	<ul style="list-style-type: none"><li>• Distance</li><li>• Volume</li><li>• Age</li></ul>



# Categorical data

- Categorical variables represent groupings of some kind.
- They are sometimes recorded as numbers, but the numbers represent categories rather than actual amounts of things.
- There are three types of categorical variables: binary, nominal, and ordinal variables.

Type of variable	What does the data represent?	Examples
Binary variables (aka dichotomous variables)	Yes or no outcomes.	<ul style="list-style-type: none"><li>• Heads/tails in a coin flip</li><li>• Win/lose in a football game</li></ul>
Nominal variables	Groups with no rank or order between them.	<ul style="list-style-type: none"><li>• Species names</li><li>• Colors</li><li>• Brands</li></ul>
Ordinal variables	Groups that are ranked in a specific order.	<ul style="list-style-type: none"><li>• Finishing place in a race</li><li>• Rating scale responses in a survey, such as <a href="#">Likert scales</a>*</li></ul>

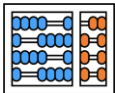




# Independent vs. Dependent Variables

---

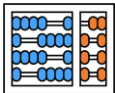
- Researchers often manipulate or measure independent and dependent variables in studies to test cause-and-effect relationships.
  - The independent variable is the cause. Its value is independent of other variables in your study.
  - The dependent variable is the effect. Its value depends on changes in the independent variable.



# Independent vs. Dependent Variables

---

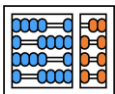
- Independent variables: Sorting algorithm, Input data
- Dependent variables: Execution time, Comparison, Swaps, Memory usage, Stability



# Independent Variables

---

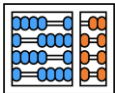
- An independent variable is the variable you manipulate or vary in an experimental study to explore its effects.
- It is called “independent” because it is not influenced by any other variables in the study.
- Independent variables are also called:
  - Explanatory variables (they explain an event or outcome)
  - Predictor variables (they can be used to predict the value of a dependent variable)
  - Right-hand-side variables (they appear on the right-hand side of a regression equation).



# Types of Independent Variables

---

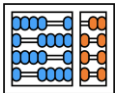
- There are two main types of independent variables:
  - **Experimental independent variables** can be directly manipulated by researchers.
  - **Subject variables** cannot be manipulated by researchers, but they can be used to group research subjects categorically.



# Experimental Independent Variables

---

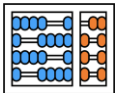
- In experiments, you manipulate independent variables directly to see how they affect your dependent variable.
- The independent variable is usually applied at different levels to see how the outcomes differ.
- You can apply just two levels in order to find out if an independent variable has an effect at all.
- You can also apply multiple levels to find out how the independent variable affects the dependent variable.



# Subject Variables

---

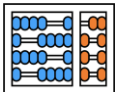
- Subject variables are characteristics that vary across participants, and they can't be manipulated by researchers.
- For example, gender identity, ethnicity, race, income, and education are all important subject variables that social researchers treat as independent variables.
- It is not possible to randomly assign these to participants, since these are characteristics of already existing groups
- Instead, you can create a research design where you compare the outcomes of groups of participants with characteristics.



# Dependent Variables

---

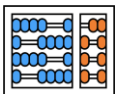
- A dependent variable is the variable that changes as a result of the independent variable manipulation.
- It is the outcome you're interested in measuring, and it “depends” on your independent variable.
- Dependent variables are also called:
  - Response variables (they respond to a change in another variable)
  - Outcome variables (they represent the outcome you want to measure)
  - Left-hand-side variables (they appear on the left-hand side of a regression equation)



# Identifying independent vs. dependent variables

---

- Distinguishing between independent and dependent variables can be tricky when designing a complex study or reading an academic research paper.
- A dependent variable from one study can be the independent variable in another study, so it's important to pay attention to research design.

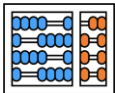




# Recognizing independent variables

---

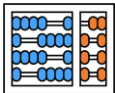
- Is the variable manipulated, controlled, or used as a subject grouping method by the researcher?
- Does this variable come before the other variable in time?
- Is the researcher trying to understand whether or how this variable affects another variable?



# Recognizing dependent variables

---

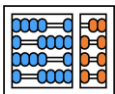
- Is this variable measured as an outcome of the study?
- Is this variable dependent on another variable in the study?
- Does this variable get measured only after other variables are altered?



# Explanatory and Response Variables

---

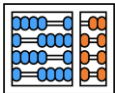
- In research, you often investigate causal relationships between variables using experiments or observations.
  - An **explanatory** variable is the expected cause, and it explains the results.
  - A **response** variable is the expected effect, and it responds to explanatory variables.
- Explanatory variables and independent variables are very similar, but there are subtle differences between them.
- Independent variables supposedly are not affected by or dependent on any other variable—they are manipulated or altered only by researchers.
- The term “explanatory variable” is preferred over “independent variable”, because in real world contexts, independent variables are often influenced by other variables. That means they’re not truly independent.



# Explanatory and Response Variables

---

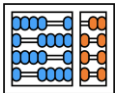
- There is a causal relationship between the variables that may be indirect or direct.
- In an indirect relationship, an explanatory variable may act on a response variable through a *mediator*.
- If you are dealing with a purely *correlational relationship*, there are no explanatory and response variables. Even if changes in one variable are associated with changes in another, both might be caused by a *confounding* variable.



# Mediator vs. Moderator Variables

---

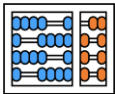
- Mediating variable (or mediator) explains the process through which two variables are related.
- Moderating variable (or moderator) affects the strength and direction of that relationship.



# Mediating Variable

---

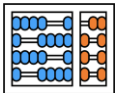
- Example: Imagine that we are investigating how the size of input data affects the execution time of sorting algorithms.
- Mediating Variable: Algorithm complexity, which includes the number of comparisons and swaps performed by the algorithm.
- Role of the Mediating Variable: Algorithm complexity mediates the relationship between input data size and execution time. This means that algorithm complexity explains part of the reason why different algorithms perform differently with larger input sizes.



# Moderating Variable

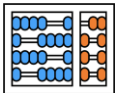
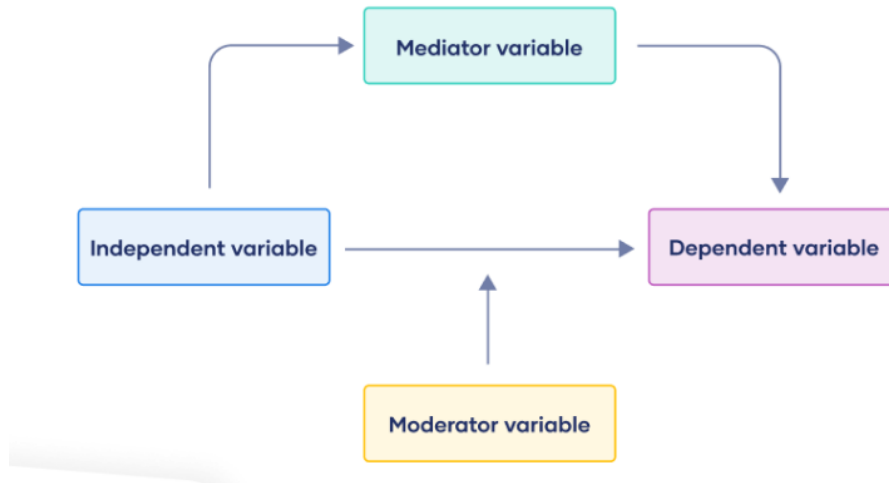
---

- Example: Now, we are exploring how the performance of sorting algorithms varies across different computer architectures.
- Moderating Variable: Machine architecture, such as high-speed CPU versus low-speed CPU.
- Role of the Moderating Variable: Machine architecture moderates the relationship between the sorting algorithm and execution time. This implies that the impact of the sorting algorithm on execution time can be more significant on one machine architecture than on another, making architecture a factor that influences the relationship.



# Mediator and Moderator Variables

---

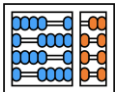




# Extraneous Variable

---

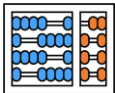
- Extraneous variable is any variable that you're not investigating that can potentially affect the outcomes of your research study.
- Extraneous variables can threaten the *internal validity* of your study by providing alternative explanations for your results.
- When not accounted for, this type of variable can also introduce many biases to your research



# Confounding Variable

---

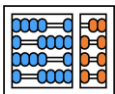
- In research that investigates a potential cause-and-effect relationship, a confounding variable is an unmeasured third variable that influences both the supposed cause and the supposed effect.
- It is important to consider potential confounding variables and account for them in your research design to ensure your results are valid.
- Left unchecked, confounding variables can introduce many research biases to your work, causing you to misinterpret your results.



# Confounding Variable

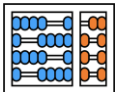
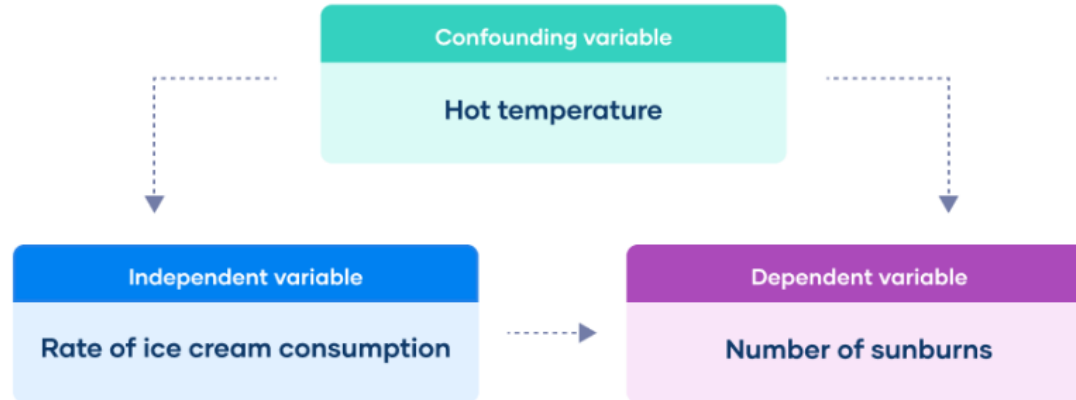
---

- Confounding variables (a.k.a. confounders or confounding factors) are a type of extraneous variable that are related to a study's independent and dependent variables.
- A variable must meet two conditions to be a confounder:
  - It must be correlated with the independent variable. This may be a causal relationship, but it does not have to be.
  - It must be causally related to the dependent variable.



# Confounding Variable

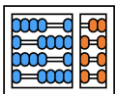
---



# Confounding Variable

---

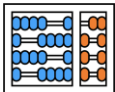
- Scenario: You are conducting research to determine whether a particular sorting algorithm's performance is influenced by the choice of programming language used to implement the algorithm.
- Variables:
  - Sorting Algorithm: The specific sorting algorithm under investigation.
  - Programming Language: The programming language used to code the sorting algorithm.
  - Execution Time: The time it takes for the algorithm to sort a given dataset.
- Confounding Variable:
  - Developer's Experience Level: A confounding variable that you didn't account for in your study. This variable represents the experience level of the developers who implemented the sorting algorithms in different programming languages.



# Control Variable

---

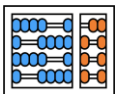
- A control variable is anything that is held constant or limited in a research study.
- It is a variable that is not of interest to the study's objectives, but is controlled because it could influence the outcomes.
- Variables may be controlled directly by holding them constant throughout a study (e.g., by controlling the room temperature in an experiment), or they may be controlled indirectly through methods like *randomization* or statistical control.



# In our sorting algorithm example

---

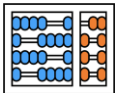
- **Machine Hardware:** To ensure consistency, control the hardware specifications, such as processor, memory, storage, and network speed, in the testing machine or server.
- **Execution Environment:** Controlling the execution environment, including CPU load, other running processes, and ambient temperature, can be important to maintain consistent conditions across algorithm executions.
- **Test Data Set:** Using a consistent and representative test data set is crucial, so control the data set characteristics, such as size, distribution, randomness, and the presence of pre-sorted data.
- **Software Versions:** If you're using different versions of sorting algorithms in different languages, control the software versions to ensure that all implementations are up-to-date and comparable.
- **Time Measurements:** Ensure that the time measurement method is consistent and accurate across all executions. This may involve using appropriate performance measurement tools.
- **Number of Repetitions:** Run tests a sufficient number of times to obtain reliable results and reduce the influence of random fluctuations. Control the number of repetitions to be the same in all comparisons.



# Control Variable

---

- There are several ways to control extraneous variables in experimental designs, and some of these can also be used in observational studies or quasi-experimental designs.
  - Random assignment
  - Standardized procedures
  - Statistical controls

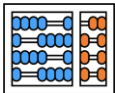




# Random assignment

---

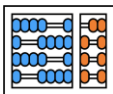
- In experimental studies with multiple groups, participants should be randomly assigned to the different conditions.
- Random assignment helps you balance the characteristics of groups so that there are no systematic differences between them.
- This method of assignment controls participant variables that might otherwise differ between groups and skew your results.



# Random assignment

---

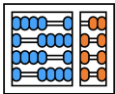
- **Experiment Objective:** To compare the performance of two different sorting algorithms, such as "QuickSort" and "MergeSort," to determine which one is more efficient in terms of execution time.
- **Random Assignment:**
  - **Random Selection of Input Data:** To ensure that the selection of input data does not introduce bias into the experiment, you can generate a random dataset that will be used to test both algorithms. This involves creating a variety of input sizes and random data configurations (sorted, reversed, random).
  - **Random Assignment of Algorithms:** The algorithms to be tested, in this case, "QuickSort" and "MergeSort," are randomly assigned to each input dataset. For a specific dataset, you randomly choose which algorithm will be applied.
  - **Multiple Executions:** Repeat the experiment with the same random assignment for multiple datasets, and execute each algorithm multiple times to calculate average execution times.



# Standardized procedures

---

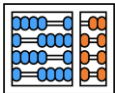
- It is important to use the same procedures across all groups in an experiment.
- The groups should only differ in the independent variable manipulation so that you can isolate its effect on the dependent variable (the results).



# Standardized procedures

---

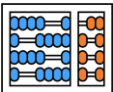
- **Input Data Generation:** Use standardized procedures to generate input data. Ensure that the data creation process is consistent and controlled, such as using a well-defined random data generation algorithm with specified data size and distribution.
- **Algorithm Implementation:** Implement both sorting algorithms according to standardized coding practices. Use the same coding standards and guidelines for each algorithm to ensure they are developed consistently.
- **Execution Environment:** Set up a controlled execution environment with standardized hardware and software configurations. This ensures that the experiment is conducted under consistent conditions, minimizing potential confounding factors.
- **Randomized Order of Execution:** While random assignment can be applied to the choice of algorithms for each dataset, the order of execution should be randomized to prevent sequence bias. Use a randomized order of execution for each algorithm on the same dataset.
- **Multiple Repetitions:** Conduct multiple repetitions of the experiment for each algorithm on different datasets. Standardize the number of repetitions to ensure a sufficient sample size for robust results.



# Statistical controls

---

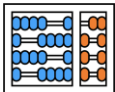
- You can measure and control for extraneous variables statistically to remove their effects on other types of variables:
  - **Execution Time Measurements:** Use a high-precision timer or profiling tool to measure the execution times of both algorithms. Ensure that the timing measurements are consistent and repeatable.
  - **Control Chart:** Create a control chart to monitor the execution times. A control chart tracks the performance of the algorithms over time, allowing you to identify any trends, shifts, or outliers that might indicate external factors affecting the results.
  - **Control Limits:** Establish control limits on the control chart based on statistical analysis. Control limits define the range within which the execution times are expected to fall if the process is in statistical control. If execution times fall outside these limits, it may indicate external factors at play.
  - **Statistical Analysis:** Apply statistical tests to analyze the data, such as t-tests or analysis of variance (ANOVA). These tests help determine whether any observed differences in execution times are statistically significant.



# Correlation vs. Causation

---

- Correlation means there is a statistical association between variables.
- Causation means that a change in one variable causes a change in another variable.
- In research, you might have come across the phrase “correlation doesn’t imply causation.”
- Correlation and causation are two related ideas, but understanding their differences will help you critically evaluate sources and interpret scientific research.

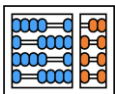


# What`s the difference?

---

- Correlation describes an association between types of variables: when one variable changes, so does the other.
- A correlation is a statistical indicator of the relationship between variables. These variables change together: they covary. But this covariation isn't necessarily due to a direct or indirect causal link.
- Causation means that changes in one variable brings about changes in the other; there is a cause-and-effect relationship between variables. The two variables are correlated with each other and there is also a causal link between them.
- A correlation doesn't imply causation, but causation always implies correlation.

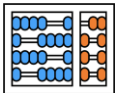
Spurious correlation: <https://www.tylervigen.com/spurious-correlations>



# How to operationalize concepts?

---

- There are 3 main steps for operationalization:
  - Identify the main concepts you are interested in studying.
  - Choose a variable to represent each of the concepts.
  - Select indicators for each of your variables.

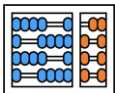


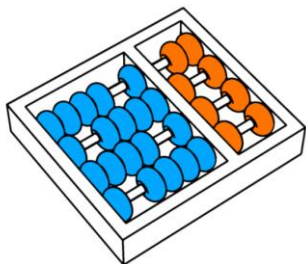


# How to operationalize concepts?

---

- **Sorting Algorithm Performance:** This is a broad concept, and it needs to be operationalized in terms of specific measures. We can operationalize it with the following variables:
  - **Execution Time:** Measure the time it takes for each sorting algorithm to sort a given dataset.
  - **Comparisons:** Count the number of comparisons made by each algorithm.
  - **Swaps or Exchanges:** Count the number of data element swaps or exchanges made by each algorithm.
  - **Stability:** Measure the extent to which the algorithm maintains the relative order of equal elements in the sorted output (e.g., using a metric like Kendall's tau or Spearman's rank correlation).
- **Input Data Characteristics:** Operationalize the concept of input data by specifying the characteristics to be measured, which could include:
  - **Data Size:** The size or length of the dataset, usually in terms of the number of elements.
  - **Data Distribution:** Categorize data as sorted, reversed, or randomized.
  - **Data Distribution Patterns:** Quantify the randomness or degree of order in the data (e.g., using metrics like inversion count or degree of randomness).





**INSTITUTO DE  
COMPUTAÇÃO**



**Prof. Dr. Bruno B. P. Cafeo**

Sala 04  
Instituto de Computação - Unicamp  
Av. Albert Einstein, 1251  
Cidade Universitária  
Campinas – SP  
13083-852

<https://ic.unicamp.br/~cafeo/>  
[cafeo@ic.unicamp.br](mailto:cafeo@ic.unicamp.br)