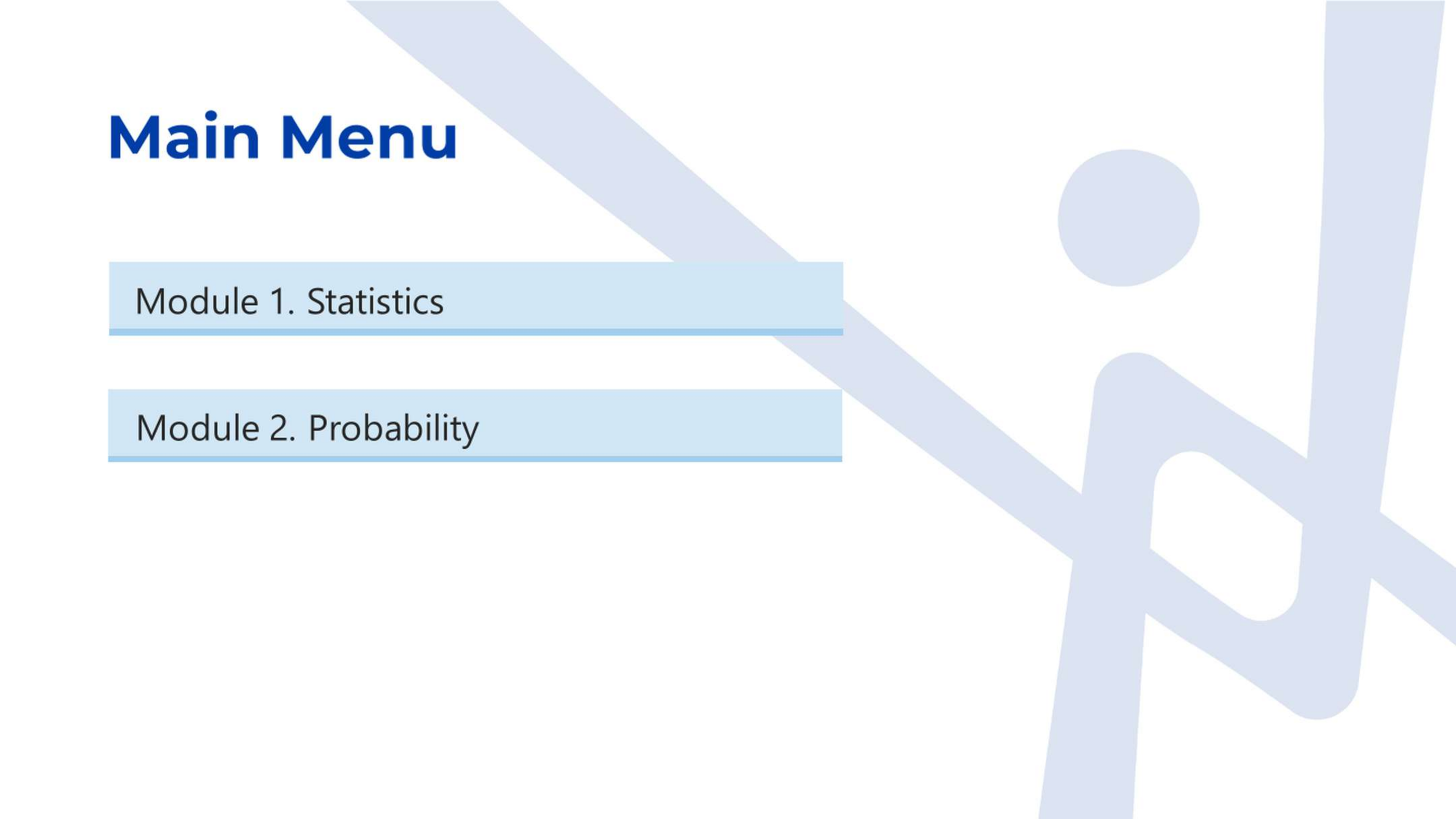
Fundamental Statistics

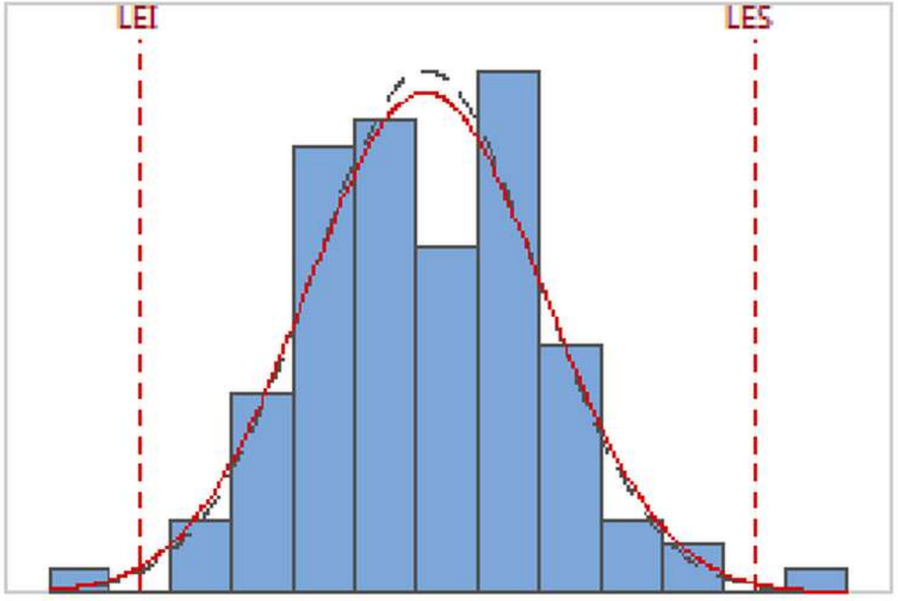
Main Menu

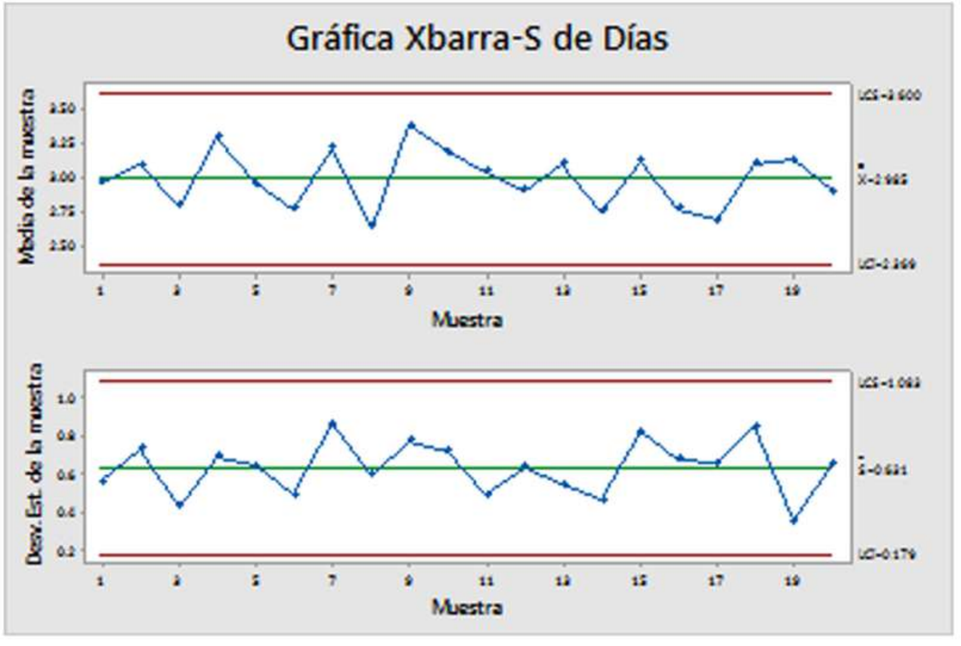


Module 1

Statistics

Statistics is science discipline dedicated to "the development and application of the theory, and best techniques for the **collection**, **classification**, **presentation**, **analysis** and **interpretation** of quantitative information obtained by observation and experimentation".





Why is Statistics important?

- Quantify risk and uncertainty
- Reliable conclusions
- Better and more informed decision making
- Identify biases
- Choosing the correct data

Some (un)familiar concepts

Population

Is any group of objects/products which are the subject of the study

Sample

Is a subset, or part of the population selected to represent the whole population

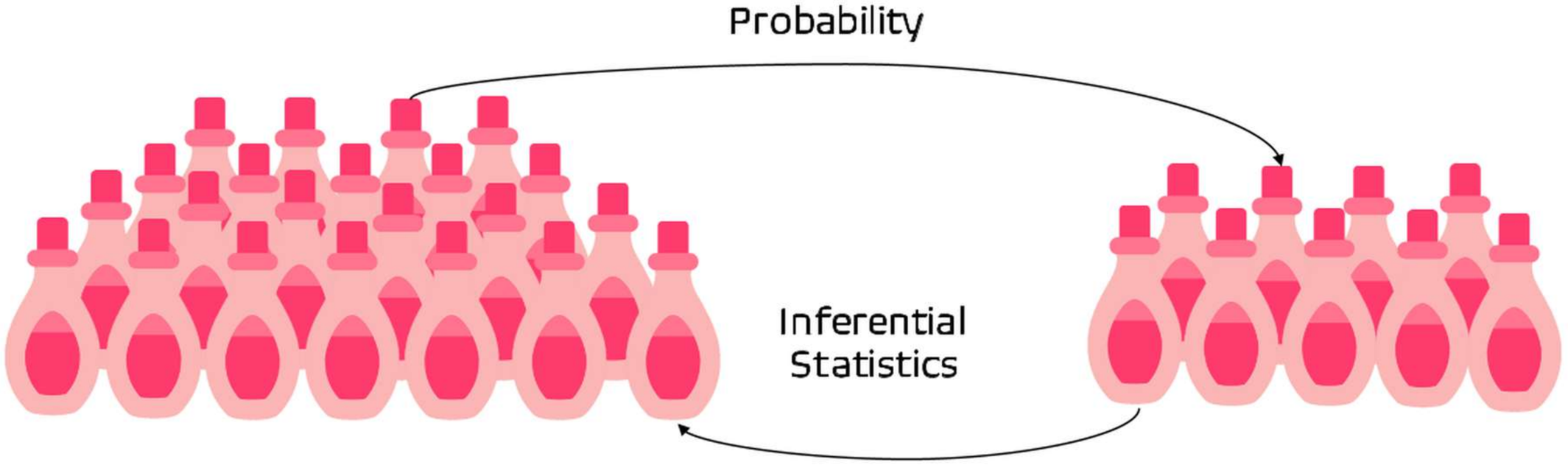
Some (un)familiar concepts

- Variable: the object of our study. (Ex: Weight)
- Observation: each person/date/product we're collecting data from. (Ex: Uncle Ben's weight)
- Data Point: each observation's value. It is represented as X_i. (Ex: 160 lb.)

Branches of Statistics

Descriptive Statistics

Inferential Statistics



Branches of Statistics

Descriptive Statistics

A measure that describes the data

Descriptive Statistics is about summarizing the data at hand through certain numbers like mean, median etc. to make the understanding and interpretation of the data easier.

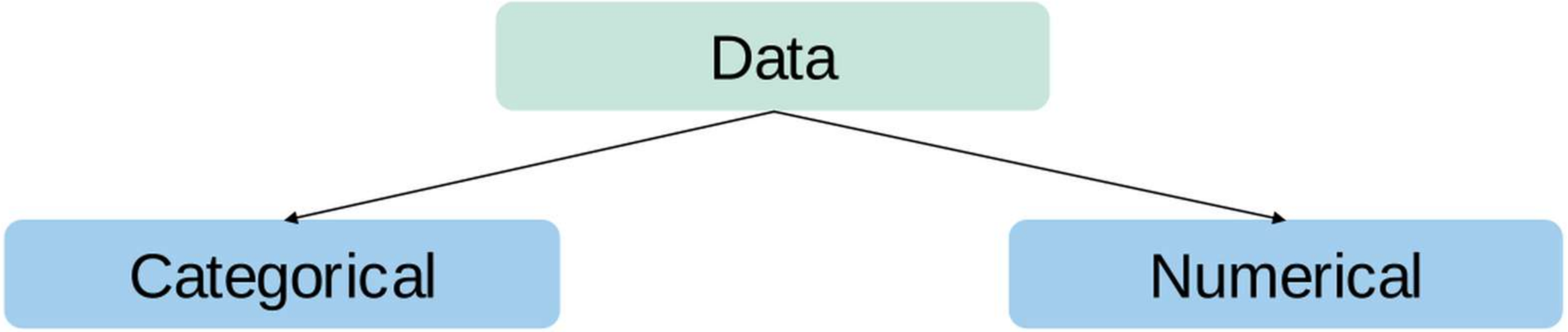
Branches of Statistics

Inferential Statistics is about using data from the sample and making inferences about the larger population from which the sample is drawn.

Inferential Statistics

Describe and make inferences about the population

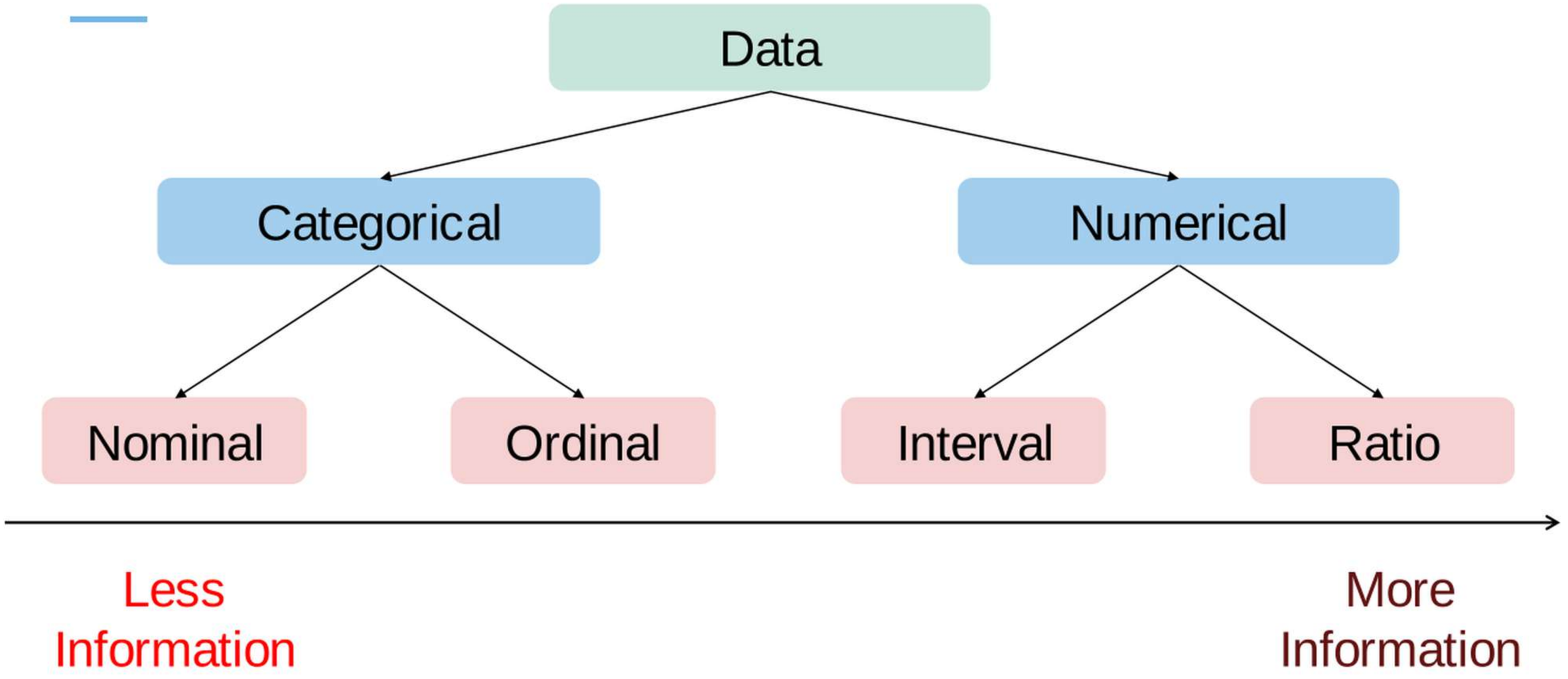
Data types



Categorical data represents characteristics. It can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don't have a particular mathematical meaning.

Numerical data represents a numerical data group. It can be Discrete data (information that can be categorized into a classification) or Continuous Data (measurements).

Data types


Data
Categorical
Numerical
Nominal
Ordinal
Interval
Ratio
Less Information
More Information

Categorical Data

Nominal Data

- **Discrete** units
- Used to label variables (**no quantitative value**)

Example

a. Which discipline did you study at University?
   - Engineering (1)
   - Business (2)
   - Pharmacy (3)
   - Law (4)
   - Other (5)

Categorical Data

Ordinal Data

- Discrete and ordered units
- Order matters

Example

b. What is your highest education level achieved?

- 1 - Elementary
- 2 – High School
- 3 – University
- 4 - Postgraduate

Numerical Data

Interval Data

Represent **ordered units that have the same difference.**

Don't have a **true zero**

Examples:

- Year
- Temperature in Fahrenheit

Ratio Data

Represent **ordered units that have the same difference.**

Do have a **true zero**

Examples:

- Age
- Height
- Weight

How do I evaluate any data set?

- Quality of data
- How was the data gathered: methodology
  - Primary data sources (first-hand)
  - Secondary data source (second-hand studies from other researchers)
- Flaws in data
- Impact of methodology on results

Determines the value of the information

Data Storytelling

Transforming data into information, through the use of **visualization tools**, in an appealing way that will help you communicate your ideas to others to allow them to make **better decisions**.

Some available tools:

- Frequency tables
- Pie Charts
- Bar graphs

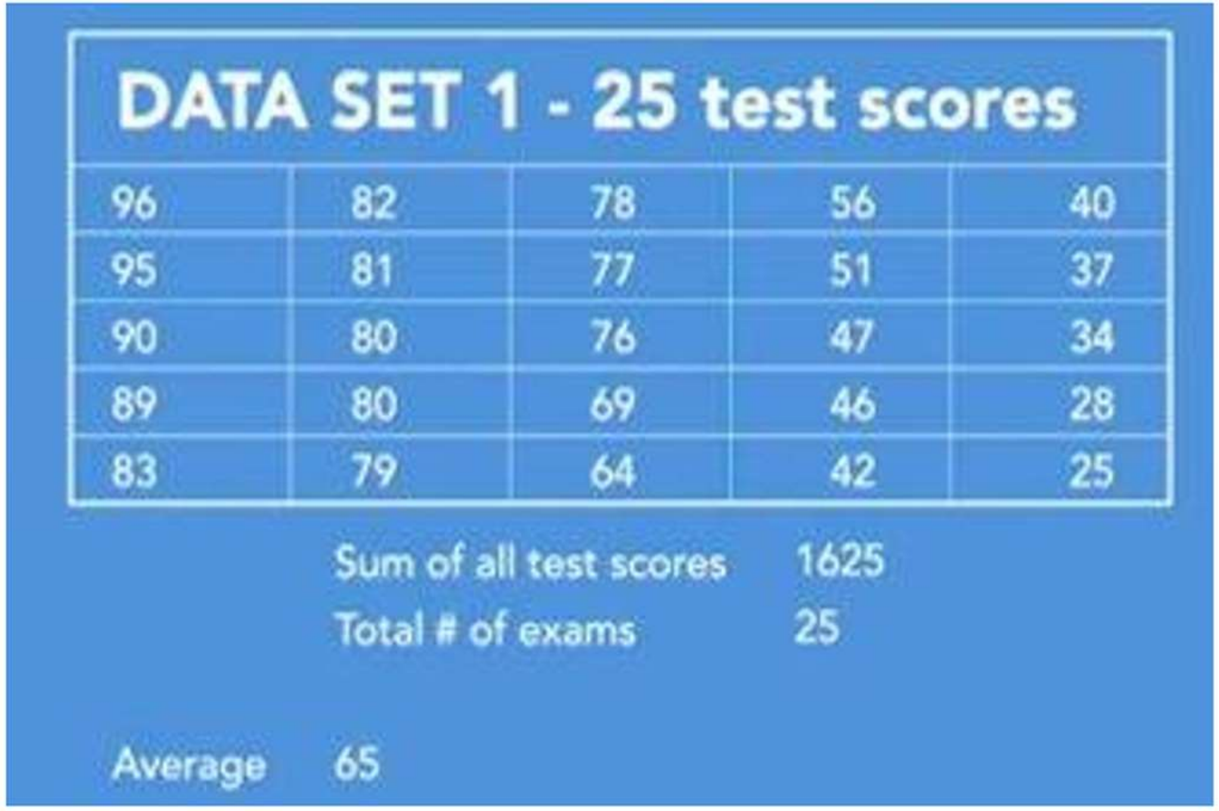
Check out our lessons on Data Visualization

Module 1

Statistics: Measures of Central Tendency

Mean (Average)

- Number around which the entire data set is spread
- Not necessarily the middle of the data
- Affected by outliers



Weighted Mean

- Signals what's valued the most (%)
- Need to decide how categories and weights are chosen
- It's simple and flexible, but arbitrary

Calculating Weighted Mean

Category	Weight	Score	Weight Score	
EXAMS 1	30%	90	27.0	.30 x 90 = 27.0
EXAMS 2	30%	80	24.0	.30 x 80 = 24.0
Quizzes	10%	75	7.5	.10 x 75 = 7.5
Homework	10%	100	10.0	.10 x 100 = 10.0
Term Paper	20%	85	17.0	.20 x 85 = 17.0
		Weighted Mean	85.5	

Median

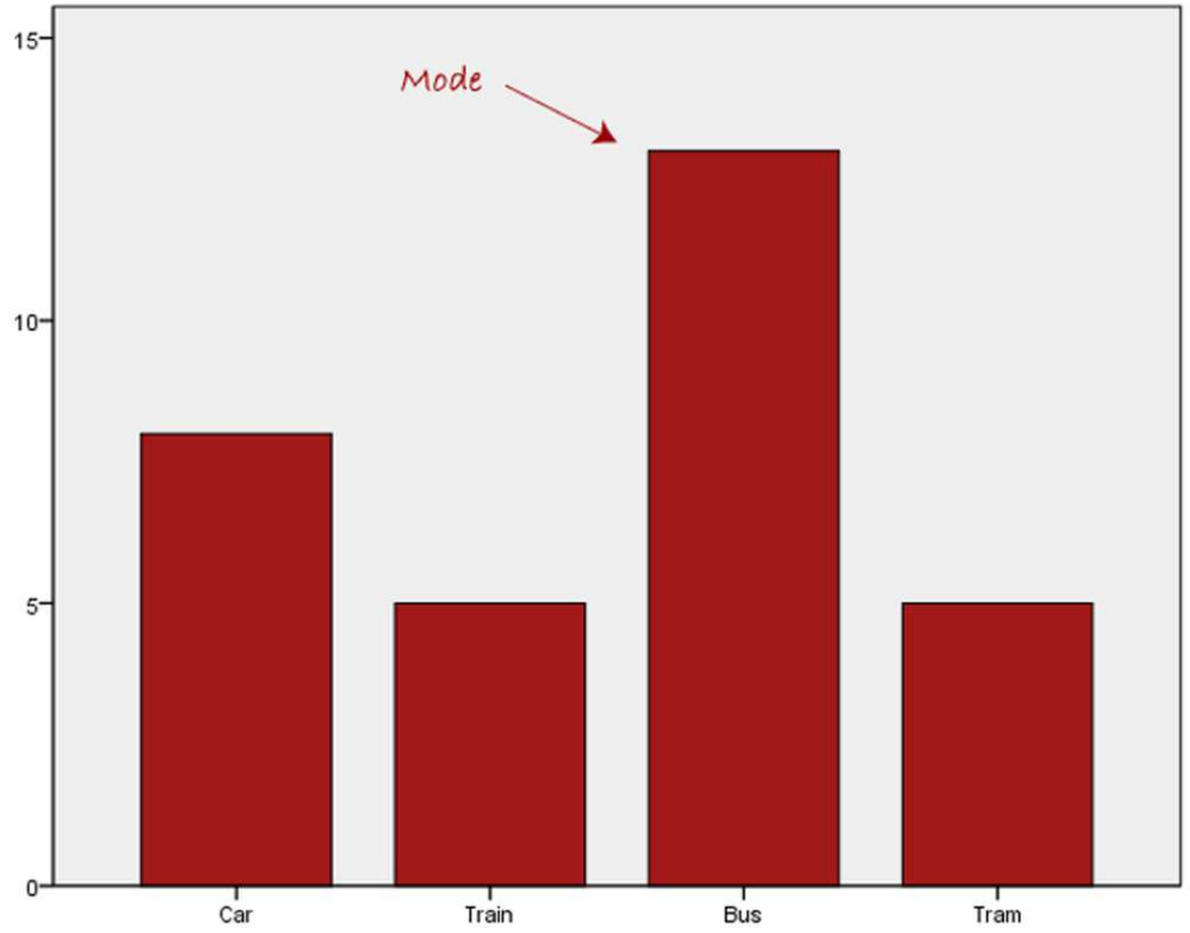
- Midpoint of the data
- Organize from top to bottom
- Depends on # of observations

FIVE DATA POINTS	
Data Point	
1	80
2	70
3	60
4	50
5	40

TEN DATA POINTS	
Data Point	
1	10
2	15
2	15
4	20
5	20
6	30
7	35
8	35
9	40
10	45

Mode

- Most frequent value



Module 1

Statistics: Measures of Dispersion (or Variability)

Range

- Difference between the maximum value and the minimum value in the data set

$$Range = Maximum - Minimum$$

107	137	76	154	32
169	185	65	23	175
154	15	76	32	27
111	176	76	83	36
20	28	18	149	35
28	67	183	70	137

Minimum	15
Maximum	185
Range	170

Variance and Standard Deviation

Variance

- How far are data points spread out from the mean
- High variance = spread widely
- Low variance: closer to the mean of the data set

Standard Deviation

- Averaged square from the mean
- Gives information about level of dispersion

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

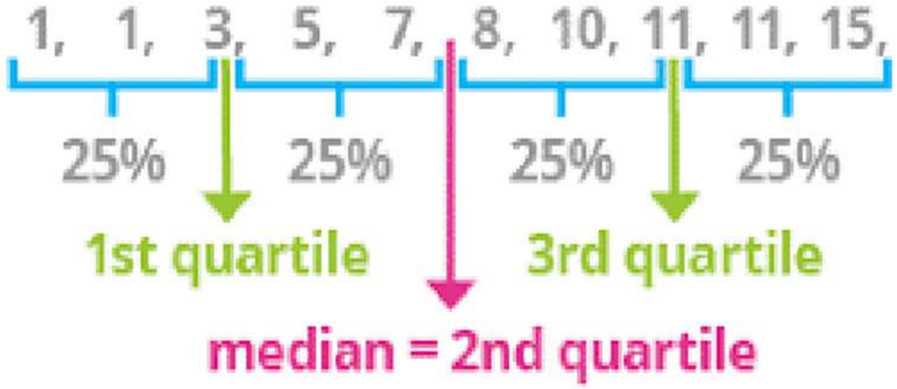
Coefficient of variation (Cov)

- The higher the Cov = the more dispersion there is
- Allows comparing distributions with different scales of measurement
- Expressed as a percentage
- Must be used only with data on a ratio scale

$$CV = \frac{\sigma}{\mu}$$

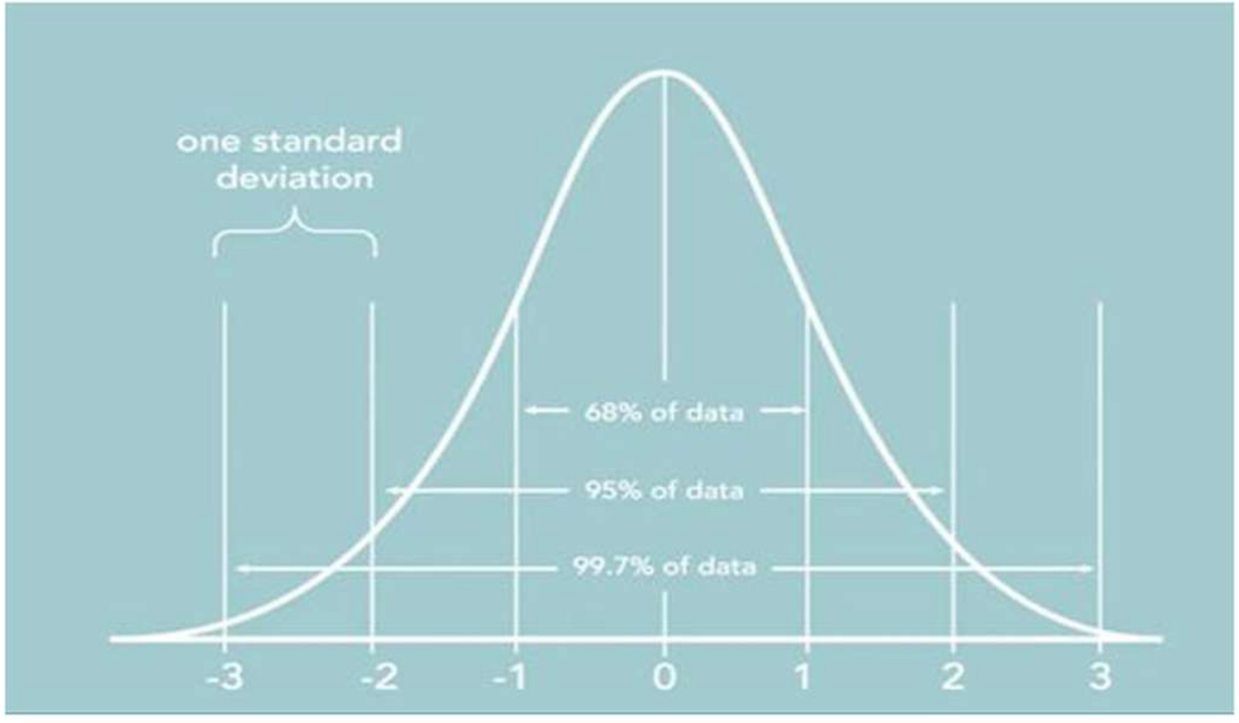
Coefficient of variation (Cov)

- Quantiles: Cut points dividing the range of a probability distribution into continuous intervals with equal probabilities.
- Types:
  - Percentiles: 100 cut points
  - Deciles: 10 cut points
  - Quartiles: 4 cut points
  - Median: 1 cut point (Midpoint of the data)
  - Quintiles: 5 cut points



Normal distribution

This is the Normal or Bell-Shaped distribution, which is roughly symmetrical. This distribution is what we call a "continuous distribution", which means that its values can vary from negative to positive numbers, with or without decimals.



Empirical rule (3σ): empirical evidence has demonstrated that data histograms with frequency can be approximated to normal curve.

Measuring distribution

- Empirical rule
- Outliers: Data points at an abnormal distance from others in a data set
  - What to do?
    - Delete them
    - Ignore them
    - Adjust them
    - Study them: seek root causes – improvement!

We can **identify outliers** using the **Empirical Rule** (3σ)

Module 2

Probability

What is Probability?

Likelihood that some event will occur

- Varies between 0 and 1, or 0% and 100%
- It's never negative
- Sum of probabilities of all possible outcomes in the same universe must be 100% or 1

Probability

- Even odds: each event is equally as likely. Examples:

Ace of Hearts
1/52

Getting a 1
1/6

Purple ball
1/3

Probability

- Uneven Odds: each event's odds as weighted, depending on other factors

For example:

  - odds of raining tomorrow
  - odds of Uncle Ben's car being stolen

- They require a more in-depth analysis, based on historic info, new info, new behaviors, etc.

Probability of two events

- Probability of A and B

$$p(A \text{ and } B) = p(A) * p(B)$$

Requires the two events to be independent.

- Probability of A or B

$$P(A \text{ or } B) = P(A) + P(B)$$

Requires the two events to be mutually exclusive.

Central Limit Theorem

- One of the most important concepts in the field of statistics.
- Average of your sufficient sample means will be approximately equal to the population mean.
- Approximates a normal distribution (where the data is symmetric about the mean and data near the mean are more frequent in occurrence than data far from the mean)

$$z\text{-}score = \frac{x_i - \bar{x}}{s}$$

x_i = data point

$\bar{x}$ = mean

s = standard deviation

Thank you

VIA
leading via learning