

Statistics

| Getting to know your data

Prof. Dr. Mostafa Elhosseini

<https://youtube.com/ElhosseiniAcademy>

| Agenda

- Variable
- Types
- Why does it matter?

| Variability

- Variability (Dispersion - spread) refers to **how spread out** or scattered the data values are in a dataset.
 - It tells us **how much the values differ** from each other and from the central value (like the mean).
-

 التباين (التشتت أو الانتشار) يشير التباين إلى مدى تباعد أو تشتت القيم في مجموعة البيانات. بمعنى آخر، يُظهر لنا مدى اختلاف القيم عن بعضها البعض وعن القيمة المركزية مثل المتوسط الحسابي.

| What is a variable?

- A **variable** is a data field, representing a characteristic or feature of a subject. The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature
- Listed in the columns of a dataset, with the rows referring to different observations on a variable
- The term **dimension** is commonly used in **data warehousing**.
- **Machine learning** literature tends to use the term **feature**, while **statisticians** prefer the term **variable**.

| Variable Types

- The type of an attribute is determined by the set of possible values the attribute can have

- **Qualitative**
 - Nominal Attributes
 - Binary Attributes
 - Ordinal Attributes

- **Quantitative:**
 - Numeric attributes

| Nominal Attributes (اسمي)

- The values of a nominal attribute are **symbols or names of things**.
- Nominal attributes are also referred to as **categorical**.
- The values do not have any meaningful order.
- In computer science, the values are also known as enumerations.
 - **Hair color**: black, brown, blond, red, auburn, gray, and white.
 - **Marital status**: single, married, divorced, and widowed.
 - **Occupation**: teacher, dentist, programmer, farmer, and so on.

| Nominal Attributes (اسمي)

- The values of a nominal attribute are symbols or “names of things,” it is possible to represent such symbols or “names” with **numbers**.
- With hair color, for instance, we can assign a code of **0 for black**, 1 for brown, and so on.
- Another example is **customer ID**, with possible values that are all numeric.
- The numbers are not intended to be used quantitatively. That is, mathematical operations on values of nominal attributes are not meaningful - **makes no sense** to subtract one customer ID number from another

| Nominal Attributes (اسمي)

Example Variable	Sample Values
Blood Type	A, B, AB, O
Gender	Male, Female, Other
Marital Status	Single, Married, Divorced
Nationality	Saudi, Egyptian, Indian, French
Favorite Color	Red, Blue, Green, Yellow

| Binary Attributes

- A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where **0** typically means that the attribute is **absent**, and **1** means that it is **present**.
- Binary attributes are referred to as **Boolean** if the two states correspond to true and false.
- Smoker describing a patient object, 1 indicates that the patient smokes, while 0 indicates that the patient does not

| Ordinal Attributes (ترتيب)

- An ordinal attribute is an attribute with possible values that have a **meaningful order or ranking** among them, but the magnitude between successive values is not known
- Drink size: small, medium, and large.
- Grade: A+, A, A−, B+
- Professional rank. Assistant, associate, and full for professors
- Army ranks: Private, private first class, specialist, corporal, and sergeant

| Ordinal Attributes (ترتيب)

- Ordinal attributes are useful for **registering subjective assessments** of qualities that cannot be measured objectively; thus ordinal attributes are often used in **surveys for ratings**.
- In one survey, participants were asked to rate how satisfied they were as customers.
- Customer satisfaction had the following ordinal categories: **0: very dissatisfied**, 1: somewhat dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied

| Ordinal Attributes (ترتيب)

Example Variable	Sample Values
Education Level	High School < Bachelor's < Master's < PhD
Customer Satisfaction	Very Unsatisfied < Neutral < Very Satisfied
Pain Level	None < Mild < Moderate < Severe
Socioeconomic Status	Low < Middle < High
Letter Grades	F < D < C < B < A

| Note: ID number

- An ID number (such as a student ID, employee ID, or national ID) is typically classified as: ► Qualitative (Categorical) – Nominal Data

✓ Why?

- It labels or identifies an individual or item.
- The numbers do not represent a quantity and have no mathematical meaning (e.g., ID 123 $\neq$ more than ID 122).
- There is no natural order or ranking among IDs.

| Note: ID number

ID Number	Name
101	Ahmed
102	Fatima
103	Mohammed

- You should not calculate an average of ID numbers — they are merely identifiers, not numerical values in the statistical sense.

| Numeric Attributes

- A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values.

| Quantitative Variable

- if observations on it take **numerical values** that represent different **magnitudes** of the variable
- Measure “how much” of something (that is, quantity or magnitude)
 - Age, number of siblings, annual income, and number of years of education completed
 - some numerical variables, such as area codes or bank account numbers, are not considered quantitative variables
 - ↳ because they do not vary in quantity (take average to check)
 - key features to describe quantitative are the center and the variability (Spread)
 - Discrete or Continuous

| Quantitative Variable

- Discrete
 - A count (“the number of ...”)
 - Any variable with a finite number of possible values is discrete

Example Variable	Sample Values
Number of Children	0, 1, 2, 3, ...
Number of Cars Owned	1, 2, 3
Number of Pages in Book	100, 200, 350
Number of Goals Scored	0, 2, 4
Number of Courses Taken	1, 3, 5

| Quantitative Variable

- Continuous (any value in a range)
 - Height, weight, age, and the amount of time it takes to complete an assignment
 - have an infinite region of values – interval - Measurable

Example Variable	Sample Values
Height of Students (cm)	165.2, 170.5, 181.3
Temperature (°C)	36.5, 37.0, 39.2
Weight (kg)	60.1, 75.8, 82.3
Time Taken (seconds)	12.4, 13.0, 15.78
Distance Traveled (km)	5.2, 10.75, 22.1

| Why do we care?

Choose the Correct Statistical Methods

- Different data types require different techniques.
 - Mean, standard deviation → valid for quantitative data, not for nominal.
 - Chi-square tests → used for categorical data.
 - T-tests, ANOVA → for numerical data.

| Why do we care?

Apply Proper Visualization Techniques

- Each data type needs a suitable graph:
 - Bar charts, pie charts → for categorical data.
 - Histograms, boxplots, scatterplots → for quantitative data..

| Why do we care?

Avoid Incorrect Calculations

- It prevents invalid operations, like:
 - Calculating the average of ID numbers or zip codes (not meaningful).
 - Sorting ordinal data without recognizing their order.

| Why do we care?

Improve Data Quality and Interpretation

- Understanding data types helps:
 - Prevent misinterpretation (e.g., treating a rating scale as interval when it's ordinal).
 - Ensure valid conclusions are drawn.

| Why do we care?

Select the Right Machine Learning or AI Model

- In AI/ML:
 - Categorical features need encoding (e.g., one-hot encoding).
 - Numerical features might need normalization or scaling.
 - Some models (like decision trees) handle both types well; others (like linear regression) expect numeric inputs..