Data Analytics

Lecture #2

Data Categorization

NOIR

Classification of scales of Measurement

NOIR classification

- The mostly recommended scales of measurement are

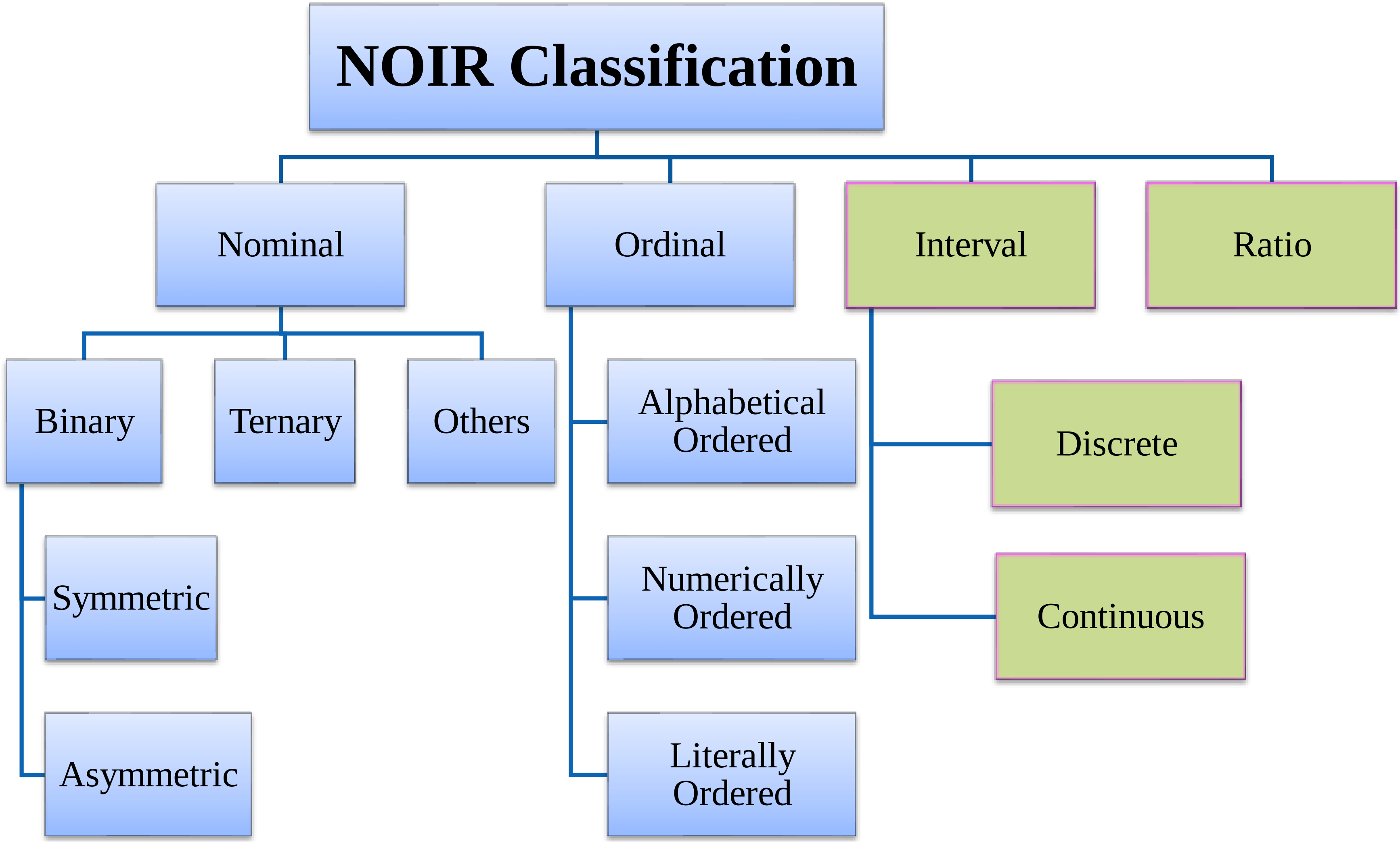
N: **<u>N</u>**ominal

O: **<u>O</u>**rdinal

I: **<u>I</u>**nterval

R: **<u>R</u>**atio

The NOIR scale is the fundamental building block on which the **extended data types** are built.



Categorical (Qualitative)

Numeric (Quantitative)

Properties of data

- Following FOUR properties (operations) of data are pertinent.

#	Property	Operation	Type
1.	Distinctiveness	$=$ and $\neq$	Categorical (Qualitative)
2.	Order	$<$, $\leq$, $>$, $\geq$	
3.	Addition	$+$ and $-$	Numerical (Quantitative)
4.	Multiplication	$*$ and $/$	

Properties of Measurement

Identity: Identity refers to each value having a unique meaning.

Magnitude: Magnitude means that the values have an ordered relationship to one another, so there is a specific order to the variables.

Equal intervals: Equal intervals mean that data points along the scale are equal, so the difference between data points one and two will be the same as the difference between data points five and six.

A minimum value of zero: A minimum value of zero means the scale has a true zero point. Degrees, for example, can fall below zero and still have meaning. But if you weigh nothing, you don’t exist.

NOIR summary

- ✓ Nominal (with distinctiveness property only)
- ✓ Ordinal (with distinctive and order property only)
- ✓ Interval (with additive property + property of Ordinal data)
- ✓ Ratio (with multiplicative property + property of Interval data)

- Further, nominal and ordinal are collectively referred to as categorical or qualitative data. Whereas, interval and ratio data are collectively referred to as quantitative or numeric data.

Nominal scale

- **Definition**

A variable that takes a value among a set of mutually exclusive codes that have no logical order is known as a nominal variable.

- **Examples**

Gender Used letters or numbers

{ M, F} **or** { 1, 0 }

Blood groups Used string

{A , B , AB , O }

Rhesus (Rh) factors Used symbols

{+ , - }

Country code ??

????

Nominal with order: Some nominal data can be sub-categorised in order, such as “cold, warm, hot and very hot.”

Nominal without order: Nominal data can also be sub-categorised as nominal without order, such as male and female.

Dichotomous: Dichotomous data is defined by having only two categories or levels, such as “yes’ and ‘no’.

Nominal scale

Note

- The nominal scale is used to label data categorization using a consistent naming convention.
- The labels can be numbers, letters, strings, enumerated constants or other keyboard symbols.
- Nominal data thus makes "category" of a set of data.
- The number of categories should be two (binary) or more (ternary, etc.), but countably finite.

Nominal scale

Note

- A nominal data may be numerical in form, but the numerical values have no mathematical interpretation.
  - For example, 10 prisoners are 100, 101, … 110, but; 100 + 110 = 210 is meaningless. They are simply labels.

- Two labels may be identical (=) or dissimilar (≠).

- These labels do not have any ordering among themselves.
  - For example, we cannot say blood group B is better or worse than group A.

- Labels (from two different attributes) can be combined to give another nominal variable.
  - For example, blood group with Rh factor (A+ , A- , AB+, etc.)

Binary scale

- **Definition**

A nominal variable with exactly two mutually exclusive categories that have no logical order is known as binary variable

- **Examples**

Switch: {ON, OFF}

Attendance: {True, False}

Entry: {Yes, No}

etc.

Note

- A Binary variable is a special case of a nominal variable that takes only two possible values.

Symmetric and Asymmetric Binary Scale

- Different binary variables may have unequal importance.

- If two choices of a binary variable have equal importance, then it is called symmetric binary variable.
  - Example: Gender = {male, female}

    // usually of equal probability.

- If the two choices of a binary variable have unequal importance, it is called asymmetric binary variable.
  - Example: Food preference = {V, NV}

Operations on Nominal variables

- Summary statistics applicable to nominal data is mode.
- Arithmetic (+,-,*and/) and logical operations (<,>,≠, etc.) are not permitted.
- The allowed operations are : accessing (read, check, etc.) and re-coding (into another non-overlapping symbol set, that is, one-to-one mapping), etc.
- Nominal data can be visualized using line charts, bar charts or pie charts etc.
- Two or more nominal variables can be combined to generate other nominal variable.
  - Example: Gender (M,F) × Marital status (S, M, D, W)

Ordinal scale

- **Definition**

Ordered nominal data are known as ordinal data and the variable that generates it is called ordinal variable.

- Example:

  Shirt size = { S, M, L, XL, XXL}

Note

The values assumed by an ordinal variable can be ordered among themselves as each pair of values can be compared literally or using relational operators ($<$, $\leq$, $>$, $\geq$).

Operation on Ordinal data

- Usually relational operators can be used on ordinal data.
- Summary measures mode and median can be used on ordinal data.
- Ordinal data can be ranked (numerically, alphabetically, etc.) Hence, we can find any of the percentiles measures of ordinal data.
- Calculations based on order are permitted (such as count, min, max, etc.).
- Spearman's R can be used as a measure of the strength of association between two sets of ordinal data.
- Numerical variable can be transformed into ordinal variable and vice-versa, but with a loss of information.
    - For example, Age [1, … 100] = [young, middle-aged, old]

Interval scale

- **Definition**

Interval-scale variables are continuous measurements of a roughly linear scale.

- Example:

  weight, height, latitude, longitude, weather, temperature, calendar dates, etc.

Note

- Interval data are with well-defined interval.
- Interval data are measured on a numeric scale (with +ve, 0 (zero), and –ve values).
- Interval data has a zero point on origin. However, the origin does not imply a true absence of the measured characteristics.
  - For example, temperature in Celsius and Fahrenheit; 0^0 does not mean absence of temperature, that is, no heat!

Operation on Interval data

- We can add to or from interval data.
  - For example: date1 + x-days = date2
- Subtraction can also be performed.
  - For example: current date – date of birth = age
- Negation (changing the sign) and multiplication by a constant are permitted.
- All operations on ordinal data defined are also valid here.
- Linear (e.g. $cx + d$) or Affine transformations are permissible.
- Other one-to-one non-linear transformation (e.g., log, exp, sin, etc.) can also be applied.

Operation on Interval data

Note

- Interval data can be transformed to nominal or ordinal scale, but with loss of information.
- Interval data can be graphed using histogram, frequency polygon, etc.

Ratio scale

- **Definition**

  Interval data with a clear definition of "zero" are called ratio data.

  - Example:

  Temperature in Kelvin scale, Intensity of earth-quake on Richter scale, Sound intensity in Decibel, cost of an article, population of a country, etc.

Note

- All ratio data are interval data but the reverse is not true.
- In ratio scale, both differences between data values and ratios (of non-zero) data pairs are meaningful.
- Ratio data may be in linear or non-linear scale.
- Both interval and ratio data can be stored in same data type (i.e., integer, float, double, etc.)

Operation on Ratio data

- All arithmetic operations on interval data are applicable to ratio data.

- In addition, multiplication, division, etc. are allowed.

- Any linear transformation of the form $(ax + b)/c$ are known.

Types of Data

Quantitative

Data that can be measured with numbers, such as duration or speed

Discrete

Whole numbers that can't be broken down, such as a number of items

Continuous

Numbers that can be broken down, such as height or weight

Interval

Numbers with known differences between variables, such as time

Ratio

Numbers that have measurable intervals where difference can be determined, such as height or weight

Qualitative

Non-numerical data that is categorical, such as yes/no responses or eye colour

Nominal

Data used for naming variables, such as hair colour

Ordinal

Data used to describe the order of values, such as 1 = happy, 2 = neutral, 3 = unhappy

Questions of the day…

1. Consider an image as an entity.
   - What are the attributes you should think to represent an image?
   - Categorize each attribute according to the NOIR data classification.
   - Suppose, two images are given. Give an idea to check if two images are identical or not.

1. How you can convert a data of interval type to ordinal type? Give an example. What are the issues of such transformation? Whether the reverse is possible or not? Justify you answer.

Questions of the day…

3. What are the different properties used to categorize the data according to NOIR data categorization?

3. Given an entity say “STUDENT” with the following attributes. Identify the NOIR category to which each of them belongs.

Scholarship amount	Name	RollNo	DoB	Aaadhar No.	Gender	Mobiloe No.	Email Id