TYPES OF DATA

Data is divided into several categories according to their characteristics

1. **DISCRETE/ CATEGORICAL /NONPARAMETRIC/ QUALITATIVE/ SYMBOLIC**
   A. NOMINAL
   B. ORDINAL

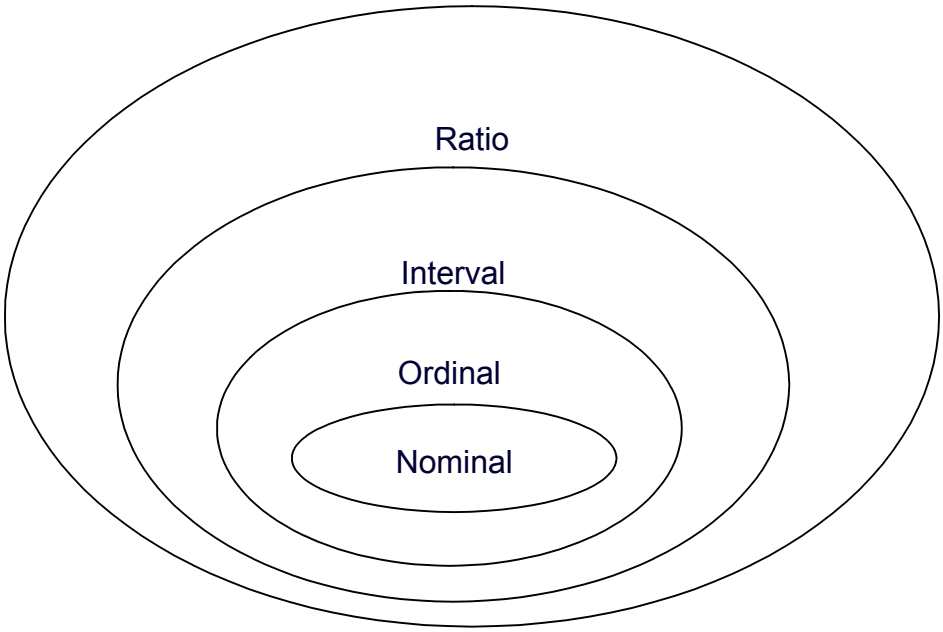
A nominal scale is an order-less scale, which uses different symbols, characters, and numbers to represent the different states (values) of the variable being measured. An example of a nominal variable, a utility, customer-type identifier with possible values is residential, commercial, and industrial. These values can be coded alphabetically as A, B, and C, or numerically as 1, 2. or 3, but they do not have metric characteristics as the other numeric data have. The numbers used to designate different attribute values have no particular order and no necessary relation to one another.

2. **CONTINUOUS/ SCALE /PARAMETRIC/QUANTITATIVE/ NUMERIC/ METRIC**
   C. INTERVAL / INTEGER
   D. RATIO

Numeric values include real-value variables or integer variables such as age, speed, or length. A feature with numeric values has two important properties: its values have an order relation ($2 < 5$ and $5 < 7$) and a distance relation ($d(2.3, 4.2) = 1.9$).

The difference between these two scales lies in how the zero point is defined in the scale. The zero point in *the interval scale* is placed arbitrarily and thus it does not indicate the complete absence of whatever is being measured. The best example of the interval scale is the temperature scale, where zero degrees Fahrenheit does not mean a total absence of temperature. Because of the arbitrary placement of the zero point, the ratio relation does not hold true for variables measured using interval scales. For example, 80 degrees Fahrenheit does not imply twice as much heat as 40 degrees Fahrenheit. In contrast, *a ratio scale* has an absolute zero point and, consequently, the ratio relation holds true for variables measured using this scale. Quantities such as height, length, and salary use this type of scale. Continuous variables are represented in large data sets with values that are numbers-real or integers.

There are four types of data that may be gathered in social research, each one adding more to the next. Thus ordinal data is also nominal, and so on.



NOMINAL

The name 'Nominal' comes from the Latin nomen, meaning 'name' and nominal data are items which are differentiated by a simple naming system.

The only thing a nominal scale does is to say that items being measured have something in common, although this may not be described.

Nominal items may have numbers assigned to them. This may appear ordinal but is not -- these are used to simplify capture and referencing.

Nominal items are usually categorical, in that they belong to a definable category, such as 'employees'.

Example

The number pinned on a sports person.

A set of countries.

BINARY

A Special Case of Nominal Data is the Binary which is a Nominal attribute with only 2 states (0 and 1).

It can be categorized into 2 types

SYMMETRIC BINARY: both outcomes equally important
e.g., gender
ASYMMETRIC BINARY: outcomes not equally important.
e.g.: Convention: assign 1 to most medical test (positive vs. negative) important outcome (e.g., HIV positive)

ORDINAL

An ordinal variable is a categorical variable for which an order relation is defined but not a distance relation

Items on an ordinal scale are set into some kind of order by their position on the scale. This may indicate such as temporal position, superiority, etc. The order of items is often defined by assigning numbers to them to show their relative position. Letters or other sequential symbols may also be used as appropriate.

Ordinal items are usually categorical, in that they belong to a definable category, such as '1956 marathon runners'.

You cannot do arithmetic with ordinal numbers -- they show sequence only.

Example

The first, third and fifth person in a race.

Pay bands in an organization, as denoted by A, B, C and D.

A special class of discrete variables is periodic variables. A periodic variable is a feature for which the distance relation exists but there is no order relation. Examples are days of the week, days of the month, or year. Monday and Tuesday, as the values of a feature, are closer than Monday and Thursday, but Monday can come before or after Friday.

INTERVAL

Interval data (also sometimes called *integer*) is measured:
on a scale of **equal-sized units** with no true zero-point. or along a scale in which each position is **equidistant** from one another.
This allows for the distance between two pairs to be equivalent in some way.
This is often used in psychological experiments that measure attributes along an arbitrary scale between two extremes.
Interval data **cannot be multiplied or divided**.

Example

My level of happiness, rated from 1 to 10.

Temperature in Fahrenheit.

RATIO

In a ratio scale, numbers can be compared as multiples of one another. Thus one person can be twice as tall as another person. Important also, the number **zero** has meaning.
Thus the difference between a person of 35 and a person 38 is the same as the difference between people who are 12 and 15. A person can also have an age of zero.
Ratio data can be multiplied and divided because not only is the difference between 1 and 2 the same as between 3 and 4, but also that 4 is twice as much as 2.
Interval and ratio data measure quantities and hence are quantitative. Because they can be measured on a scale, they are also called *scale data*.

Example

A person's weight

The number of pizzas I can eat before fainting

PARAMETRIC VS. NON-PARAMETRIC

Interval and ratio data are *parametric*, and are used with parametric tools in which distributions are predictable (and often Normal).

Nominal and ordinal data are *non-parametric*, and do not assume any particular distribution. They are used with non-parametric tools such as the Histogram.

CONTINUOUS AND DISCRETE

CONTINUOUS measures are measured along a continuous scale which can be divided into fractions, such as temperature. Continuous variables allow for infinitely fine sub-division, which means if you can measure sufficiently accurately, you can compare two items and determine the difference.

DISCRETE variables Has only a finite or countably infinite set of values
Discrete variables are measured across a set of fixed values, such as age in years (not microseconds). These are commonly used on arbitrary scales, such as scoring your level of happiness, although such scales can also be continuous.

Finally, one additional dimension of classification of data is based on its **behavior with respect to time**.

STATIC DATA: the data values do not change with time

DYNAMIC OR TEMPORAL DATA: the attribute values change with time.

The majority of the data-mining methods are more suitable for static data, and special consideration and some preprocessing are often required to mine dynamic data.

ATTRIBUTE VALUES

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
    - Same attribute can be mapped to different attribute values
      Example: height can be measured in feet or meters
- Different attributes can be mapped to the same set of values
      Example: Attribute values for ID and age are integers
      But properties of attribute values can be different
      ID has no limit but age has a maximum and minimum value

The type of an attribute depends on which of the following properties it possesses:

Distinctness: = ≠

Order: < >

Addition: + -

Multiplication: * /

Nominal attribute: distinctness

Ordinal attribute: distinctness & order

Interval attribute: distinctness, order & addition

Ratio attribute: all 4 properties

ATTRIBUTE TYPE	DESCRIPTION	EXAMPLES	OPERATIONS
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	employee ID numbers, eye color, sex: {male, female}	mode, entropy, contingency correlation, $\chi 2$ test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {good, better, best}, grades, street numbers	median, percentiles, rank correlation
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., new_value = f(old_value) where f is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	new_value =a * old_value + b where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	new_value = a * old_value	Length can be measured in meters or feet.

Monotonic: if for all x and y such that $x \leq y$ one has $f(x) \leq f(y)$

BIBLIOGRAPHY

(LAROSE, 2005)

Bibliography

David, O. L., & Delen, D. (2008). *Advanced Data Mining Techniques.* Springer-Verlag Berlin Heidelberg.

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms.* John Wiley & Sons.

LAROSE, D. T. (2005). *DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining.* New Jersey.: A JOHN WILEY & SONS.