Describing and visualising data

Ivan Heibi
ivan.heibi2@unibo.it – https://orcid.org/0000-0001-5366-5194 – https://ivanhb.it

Computational Management of Data – Part II (A.Y. 2025/2026)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

Summary of the previous lectures (1/2)

A datum is a declarative statement subject-predicate-object that, through the predicate, either **attributes** a **literal** (i.e. a value such as a string, a number, etc.) to a subject **entity** *or* it **relates** such a subject **entity** with another **entity**

Each entity, being used either as subject or object of a statement, is characterised by **a unique identifier**

The **same entity** can be used as subject or object in one or more data, while a literal **cannot be used** as subject in any datum

An attribute is intrinsically **part of** the entity to which it is associated – modifying the value of an attribute affect **only** the entity to which it refers to

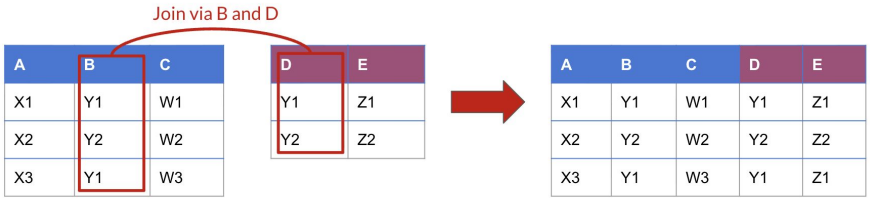
A **data model** is an abstract, simplified and formal representation of some data related to a system or a real domain, and enables us to describe what a data collection is about and to check data correctness

A **data model** permit one to specify classes of entities, their attributes and relations

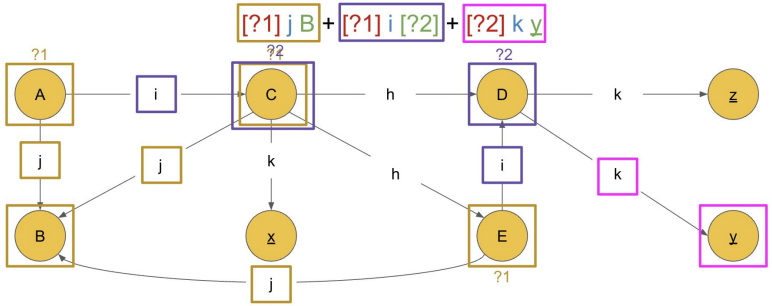
Summary of the previous lectures (2/2)

Depending on the structure in which data are stored , queries to datasets are approached differently:

- With **tabular data**, often you have to combine tables between them to obtain bigger tables which contain the query requirements and the related answer



- With **graph data**, you explore the graph starting from fixed points (i.e. known entities, values, predicates) to find a pattern that is compliant with the query



Descriptive statistics

Descriptive statistics are a series of statistics which aim at describing quantitatively a collection of data

Such statistics do not infer new information from a given population, since it does not use probability at all, but it provides measure to summarise data as they are

Often, such statistics are accompanied by visual graphs that enable a reader to understand simply some of the aspects of a collection of data

Different kinds of measures:

- measures of central tendency: mean, median, and mode
- measures of variability: minimum, maximum, and standard deviation

Mean

In mathematics and statistics, the arithmetic mean or, simply, the **mean** is the sum of a collection of numbers divided by the count of numbers in the collection

For instance, consider the following years of publication of 10 articles

1962, 2005, 2007, 2011, 2011, 2013, 2014, 2016, 2019, 2022

the mean is

(1962 + 2005 + 2007 + 2011 + 2011 + 2013 + 2014 + 2016 + 2019 + 2022) / 10 =

2008

Median

The median is the value separating the higher half from the lower half of a data sample – it may be thought of as "the middle" value

Basic feature: it is not skewed by a small proportion of extremely large or small values, and therefore provides a better representation of a typical value

How to calculate it:

- If the count of numbers *n* in a collection is odd, the median value is at index *(n-1)/2* (starting indexing items from 0, as in Python list)
- If the count of numbers *n* in a collection is even, the median value is the mean of the value at index *(n/2)-1* and *n/2* (starting indexing items from 0, as in Python list)

1962, 2005, 2007, 2011, 2011, 2013, 2014, 2016, 2019, 2022

0 1 2 3 4 5 6 7 8 9

(n/2)-1 n/2 → (2011 + 2013) / 2 = 2012

Mode

The mode is the value that appears most often in a set of data values

For instance, consider the following years of publication of 10 articles

1962, 2005, 2007, 2011, 2011, 2013, 2014, 2016, 2019, 2022

the mode is 2011

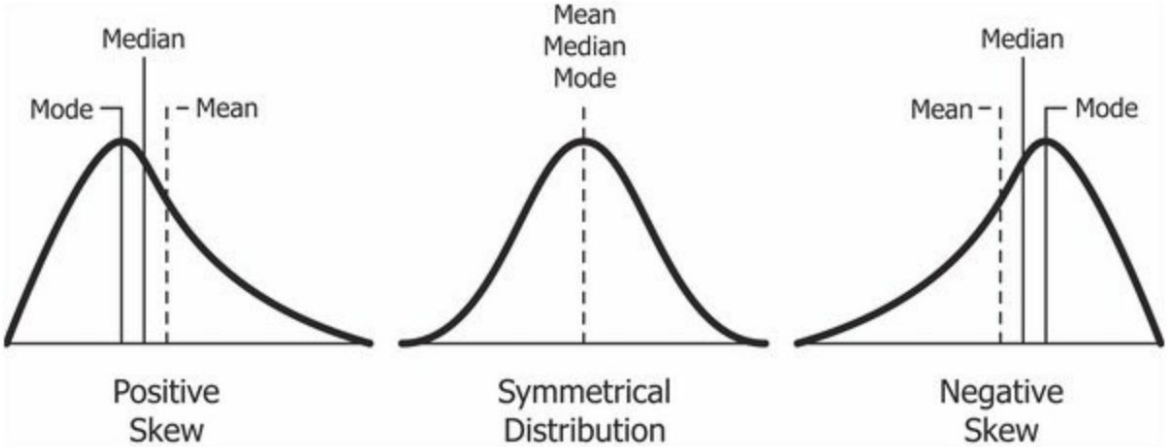
For a sample where each value occur precisely once, the usual practice is to discretize the data by assigning frequency values to intervals of equal distance, as for making a histogram, effectively replacing the values by the midpoints of the intervals they are assigned to

Why mean, median and mode can be different

The mean is largely affected by outliers, i.e. either small or large values that differ significantly from other observations

The median can be used as a measure of location when one thinks extreme values are of minimal or no importance, e.g. because a distribution is skewed

The mode is the same as that of the mean and median in a normal distribution, but it may be very different in highly skewed



Minimum and Maximum

The maximum and minimum are the values of the greatest and least elements of a collection

For instance, consider the following years of publication of 10 articles

1962, 2005, 2007, 2011, 2011, 2013, 2014, 2016, 2019, 2022

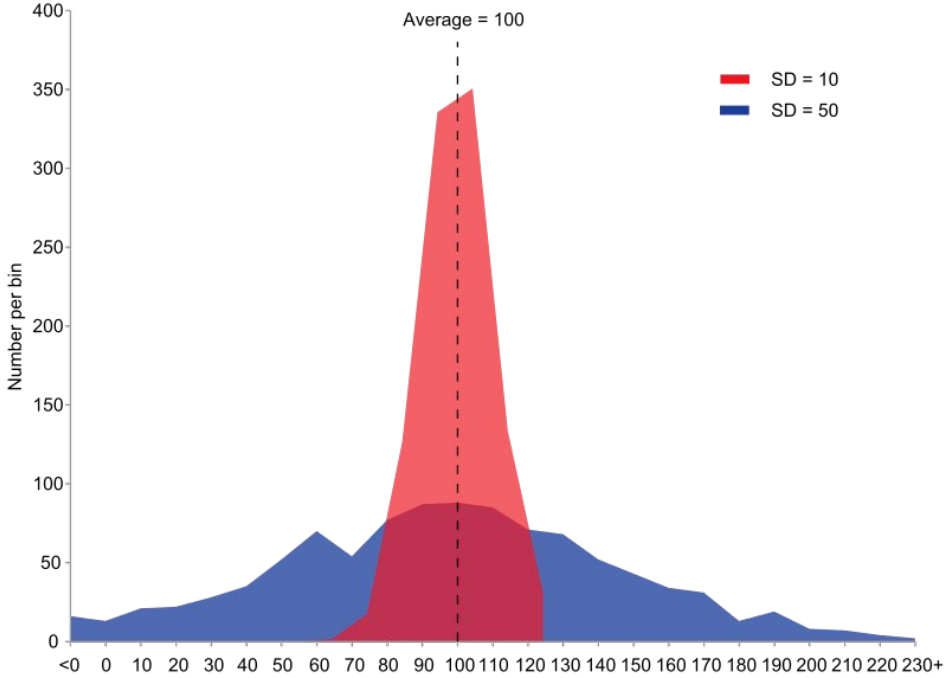
the maximum is 2022 and the minimum is 1962

If the sample has outliers, they necessarily include the sample maximum or sample minimum, or both

Standard deviation

The standard deviation measures the amount of dispersion of a set of values

- Low standard deviation: the values tend to be close to the mean
- High standard deviation: the values are spread out over a wider range



Visualisation

Visualisation techniques are of crucial important to effectively communicate a message to humans

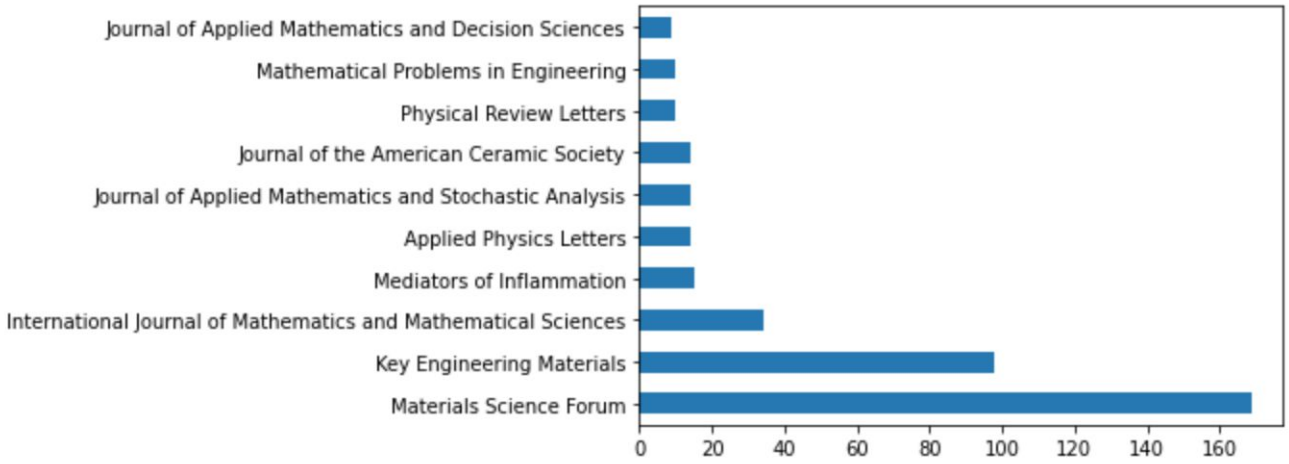
Data visualisation concerns the techniques used to communicate (often statistical) information about data, that can be categorised according to specific labels or shown as an time-oriented evolution of observations

Data visualisation techniques may be combined, when needed, with **information visualisation** techniques, that are used to represent visually numerical and non-numerical data (e.g. text or geolocated information) to support human comprehension of a phenomenon

Bar Charts

Bar charts are used to present **categorical data** – i.e. a variable that can take on one of a limited, and usually fixed, number of possible values – with rectangular bars with heights or lengths proportional to the values that they represent

The bars can be plotted vertically or horizontally

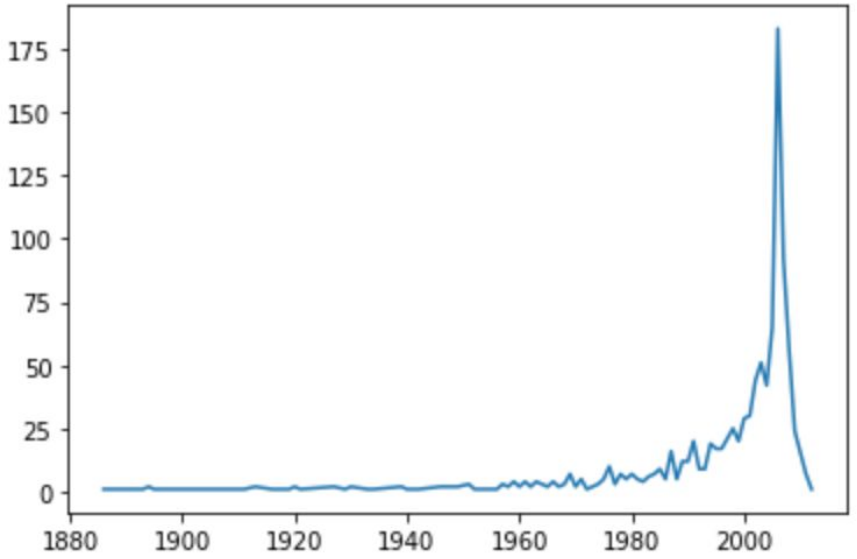


Time series

A time series is a series of data points **indexed and ordered according to time**, usually depicted as x-axis of a two dimensional graph

Commonly, a time series is a sequence taken at successive equally spaced points in time (e.g. years of publication)

A time series is very frequently plotted via a run chart (which is a temporal line chart).



END

Ivan Heibi
ivan.heibi2@unibo.it – https://orcid.org/0000-0001-5366-5194 – https://ivanhb.it

Computational Management of Data – Part II (A.Y. 2025/2026)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna