Chapter II: Data

Knowledge Discovery in Databases

Luciano Melodia M.A.
Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg
Summer semester 2021

Chapter II: Getting to know your data

This is our agenda for this lecture:

- **Data objects and attribute types.**
- Basic statistical descriptions of data.
- Data visualization.
- Measuring data similarity and dissimilarity.
- Summary.

Types of data sets

Records:

Relational records.
Data matrix, e.g. numerical matrix, crosstabs.
Document data: text documents, **term-frequency vectors**.
Transaction data.

	team	couch	play	ball	score	game
Document1	3	0	5	0	2	6
Document2	0	7	0	2	1	0
Document3	0	1	0	0	1	2

Graph and network:

World wide web.
Social of information networks.
Molecular structures.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Coke, Diapers, Milk

Types of data sets

Ordered data:

Video data: sequences of images.
Temporal data: time series.
Sequential data: transaction sequences.
Genetic sequence data.

Spatial, image and multimedia:

Spatial data: maps.
Image data.
Video data.

Important characteristics of structured data

Dimensionality:
Curse of dimensionality (sparse high-dimensional data spaces).

Sparsity:
Only presence counts.

Resolution:
Patterns depend on the scale.

Distribution:
Centrality and dispersion.

Data objects

Data sets are made up of data objects.
A data object represents an entity.

Examples:

- Sales database: customers, store items, sales.
- Medical database: patients, treatments.
- University database: students, professors, courses.

They are also called:
Sampels, examples, instances, data points, objects, tuples, ...

Data objects are described by attributes:

- Database rows $\rightarrow$ data objects.
- Columns $\rightarrow$ attributes.

Attributes

Attribute:
Sometimes also in other context: field, dimension, feature, variable, ...

A data field encodes the property of an entity or feature of a data object.
E.g. `customer_ID`, `name`, `address`.

Types:

- Nominal.
- Binary.
- Ordinal.
- Numerical:
  - Interval scaled.
  - Ratio scaled.

Attribute types

Nominal:

Categories, states, or "names of things".
E.g. `hair_color` $= \{$auburn, black, blond, brown, grey, red, white$\}$.
Other examples: `marital status`, `occupation`, `ID`, `ZIP code`.

Binary:

Nominal attribute with only two states (0 and 1).
Symmetric binaries: both outcomes equally important, such as `gender`.
Asymmetric binary: outcomes not equally important.
E.g. medical test (positive vs. negative).
Convention: assign 1 to most important outcome (e.g. HIV positive).

Ordinal:

Values have a meaningful order (ranking),
but magnitude between successive values is not known.
E.g. `size` $= \{$small, medium, large$\}$, grades, army rankings.

Numerical attribute types

Numerical: Quantity (integer- or real-valued).

Interval scaled:

Measured on a scale of **equally sized** units.
Values have order.
E.g. temperature in C or F, calender dates.
No true zero-point.

Ratio scaled:

Inherent **zero point**.
We can speak of values as being an order of magnitude larger than the unit of measurement.
E.g. $10K$ is twice as high as $5K$.
E.g. temperature in Kelvin, length, counts, monetary quantities.

Discrete vs. continuous attributes

Discrete attribute:

Has finite or countably infinite elements.
E.g. ZIP code, profession, or the set of words in a collection of documents.
Sometimes represented as integer variables.
Note: Binary attributes are a special case of discrete attributes.

Continuous attribute:

Has real numbers as attribute values.
E.g. temperature, height, or weight.
Practically, real values can only be measured and represented using a finite number of digits.
Continuous attributes are typically represented as floating-point variables.

Chapter II: Getting to know your data

Data objects and attribute types.

Basic statistical descriptions of data.

Data visualization.

Measuring data similarity and dissimilarity.

Summary.

Basic statistical descriptions of data

Motivation:

To better understand the data: central tendency, variation and spread.

Data dispersion characteristics:

Median, max, min, quantiles, outliers, variance etc.

Numerical dimensions correspond to sorted intervals.

Data dispersion: analyzed with multiple granularities of precision.

Boxplot or quantile analysis on sorted intervals

Dispersion analysis on computed measures.

Folding measures into numerical dimensions.

Boxplot or quantile analysis on the transformed cube.

Measuring the central tendency

Mean:

N denotes the amount of samples within the data set.
The **sample mean** is given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

While the **population mean** is defined by

$$\mu = \sum x \cdot p(x|\theta) \cdot \cdots .$$

Measuring the central tendency (2)

Median:

The median $\tilde{m}$ minimizes the sum of absolute deviations for any x of a sample X:

$$\sum_{i=1}^{n} |\tilde{x} - x_i| \leq \sum_{i=1}^{n} |x - x_i|. \qquad (1)$$

Age	Frequency
$1-5$	200
$6-15$	450
$16-20$	300
$21-50$	1500
$51-80$	700
$81-110$	44

Measuring the central tendency (3)

Median for interval grouped data:

Let n be the total amount of data points, n_i the respective number of the ith group and l_i or u_i the lower or upper interval limit. We determine the group to which the median belongs and denote it as mth group. It is determined by

$$\sum_{k=1}^{m-1} n_k < \frac{n}{2}, \text{ but } \sum_{k=1}^{m} n_k \geq \frac{n}{2}. \tag{2}$$

If there is no information about the underlying distribution, we just assume that data is equally distributed and use linear interpolation to estimate the median:

$$\tilde{x} = l_m + \frac{\frac{n}{2} - \sum_{k=1}^{m-1} n_k}{n_m} \cdot (u_m - l_m). \tag{3}$$

Age	Frequency
$1 - 5$	200
$6 - 15$	450
$16 - 20$	300
$21 - 50$	1500
$51 - 80$	700
$81 - 110$	44

Measuring the central tendency (3)

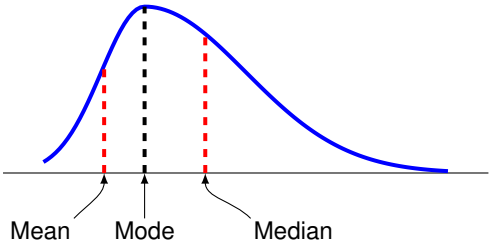
Mode:

Value that occurs most frequently within the data set.
Can be unimodal, bimodal, trimodal etc.
Empirical formula:

$$\overline{x} - \text{mode} \approx 3\left(\overline{x} - \tilde{x}\right). \tag{4}$$

Age	Frequency
$1 - 5$	200
$6 - 15$	450
$16 - 20$	300
$21 - 50$	1500
$51 - 80$	700
$81 - 110$	44

Example of mode, median and mean



$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{5}$$

Example of mode, median and mean

Quartiles, outliers and boxplots:

Quartiles: $\mathbf{Q_1}$ (25^{th} percentile), $\mathbf{Q_3}$ (75^{th} percentile).

Inter quartile range: IQR $= Q_3 - Q_1$.

Five number summary: min, Q_1, median, Q_3, max.

Boxplot: ends of the box are the quartiles;
median is marked; add whiskers and plot outliers individually.

Outlier: usually assigned to values higher/lower than $1.5 \cdot$ IQR.

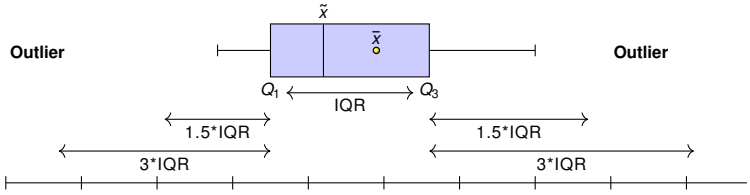
Variance σ^2 and standard deviation σ:

Empirical sample variance: $\overline{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Empirical population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$.

Standard deviation is the square root $\sigma = \sqrt{\sigma^2}$.

Boxplot analysis



Five number summary of a distribution:
Minimum, Q_1, median, Q_3, maximum.

Boxplot:

Data is represented with a box.

The ends of the box are at the first and third quartiles, i.e. the height of the box is IQR.

The median is marked by a line within the box.

Whiskers: two lines outside the box extended to minimum and maximum.

Outliers: points beyond a specified outlier threshold, plotted individually.

Properties of normal distribution curves

The normal distribution:

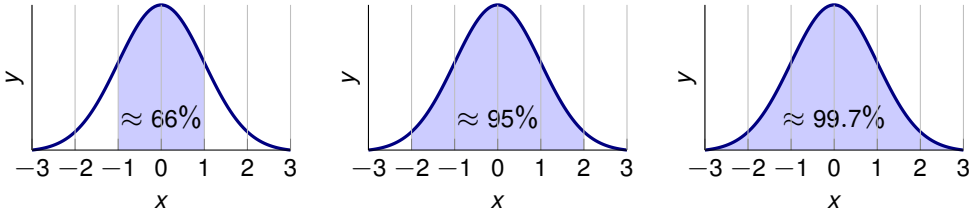
From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements.

μ: mean,

σ: standard deviation.

From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of the surface under the curve.

$\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of the surface under the curve.



Visualization of basic statistical descriptions

Boxplot: Visualization of five number summary.

Histogram: x-axis are values, y-axis represent frequencies.

Quantile plot: Each value x_i is paired with some q_i indicating that approximately $q_i \cdot 100\%$ of data are $\leq x_i$.

Quantile-quantile (q-q) plot: Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

Scatter plot: Each pair of values is a pair of coordinates and plotted as points in the plane.

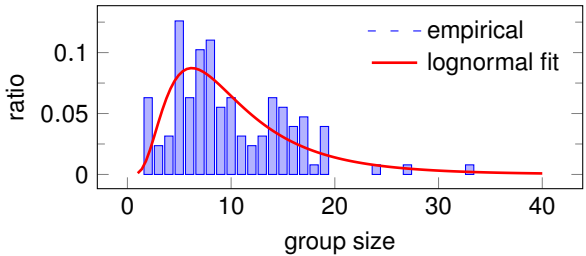
Histogram analysis

Histogram: Visualization of tabulated frequencies, shown as bars.

It shows what proportion of cases fall into each of several categories.

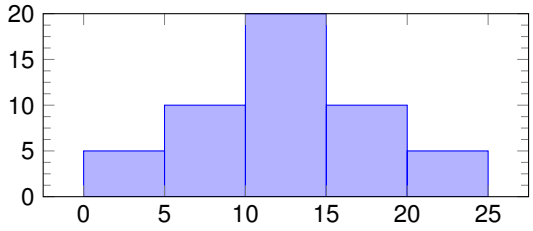
Differs from a **bar chart** in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width.

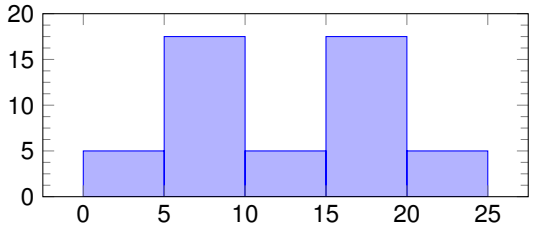
The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent.



Histograms often tell more than boxplots

The two histograms shown below may have the same boxplot representation, thus the same values for min, Q_1, median, Q_3 and for the max. But they have rather different underlying distributions.





Quantile plot

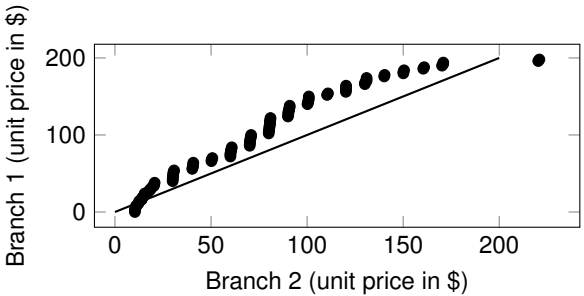
Displays all of the data.
A quantile plot allows the user to assess both the overall behaviour and unusual occurrences.

Plots quantile information.
For some data point x_i, sorted in increasing order, q_i indicates that approximately $q_i \cdot 100\%$ of the data are below or equal to the value of x_i.
TODO

Quantile-quantile (q-q) plot

Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

View: Is there is a shift in going from one distribution to another?
Example shows unit price of items sold at Branch 1 vs. branch 2 for each quantile. Unit prices of items sold at branch 1 tend to be lower than those at branch 2.



References I

References I

W. Cleveland: Visualizing Data, Hobart Press, 1993.

T. Dasu and T. Johnson: Exploratory Data Mining and Data Cleaning, John Wiley, 2003.

U. Fayyad, G. Grinstein, and A. Wierse: Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001.

L. Kaufman and P. J. Rousseeuw: Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.

H. V. Jagadish et al.: Special Issue on Data Reduction Techniques, Bulletin of the Tech. Committee on Data Eng., 20(4), 1997.

D. Keim: Information visualization and visual data mining, IEEE Trans. on Visualization and Computer Graphics, 8(1), 2002.

F. Naumann: Data Profiling Revisited, ACM SIGMOD Record, 32(4), 2013, 40-49.

D. Pyle: Data Preparation for Data Mining, Morgan Kaufmann, 1999.

References II

S. Santini and R. Jain: Similarity measures, IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999.

E. R. Tufte: The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001.

C. Yu et al.: Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009.

Thank you for your attention.
Any questions about the second chapter?

Ask them now, or again, drop me a line:
luciano.melodia@fau.de.