Software Development for Data Analysis

Bibliography:

1. **Vințe C.**, **Furtună T.F.**, *Multivariate Data Analysis in Python*, Editura ASE, București 2024, https://www.editura.ase.ro/Multivariate-data-analysis-in-Python-Analiza-multivariata-a-datelor-n-Python-/
2. **Benzecri J. P.** , *L'analyse des données*, Dunod, Paris, 1979, Franţa
3. **Harman, H.H.** , *Modern Factor Analysis*, University of Chicago Press, Chicago, Ill., 1967
4. **Gelman A., Hill J.** , *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2007
5. **Cherkassky V., Mulier F.** , *Learning from Data: Concepts, Theory and Methods*, John Wiley & Sons, Inc., New York, 1998
6. **Murtagh, F., Heck, A.** , *Multivariate data analysis*, Dordrecht, 1987, Olanda
7. **Ruxanda G.** , *Analiza multidimensională a datelor*, Editura ASE, București 2009
8. **Vințe C.**, **Furtună T.F.**, *Python pentru analiza datelor*, Editura ASE, București 2020, https://www.editura.ase.ro/Python-pentru-analiza-datelor/

What is Data Analysis?

Is a process of:

- inspecting,
- cleaning,
- transforming,
- and modeling data

having the goal of:

- discovering (new) useful information,
- suggesting conclusions,
- and supporting decision-making.

What is Data Analysis?

Data Analysis:

- has multiple facets and approaches,
- encompassing diverse techniques,
- under a variety of names,
- employed in different business, science, and social science domains (*the nitty-gritty investigation of what information a given data set may contain*).

Data Mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes.

Business Intelligence covers data analysis that relies heavily on aggregation, focusing on business information.

What is Data Analysis?

Pattern Recognition

- are data analysis systems that in many cases are trained from labeled "training" data (supervised learning)
- when no labeled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning).

Machine Learning is the common term for supervised learning methods and originates from artificial intelligence, whereas knowledge discovery in databases (KDD) and data mining have a larger focus on unsupervised methods and stronger connection to business use.

Fundamental concepts

Population and sample

- Population or general collectivity, represents the set of all the measurements with relevance for researcher or experiment. A sample is a subset of a population.

Attribute or characteristic

- The properties, features of each entity within the observed population.

Fundamental concepts

Variable

- An abstract concept which allows for assigning numerical or non-numerical values to an attribute or characteristic. It must have an univocal (unambiguous) syntax and a precise semantic.

Fundamental concepts

Qualitative variables

- differ by ***type***, and refer to non-numerical properties, characteristics of the elementary units belonging to a population, modalities.

Quantitative variables

- differ by ***size***, and refer to numerical properties of the elementary units belonging to a population.

Fundamental concepts

Qualitative variables

- Nominal (colour names, music genres, architectural styles, etc.)
- Ordinal (more, less, equal, indifferent, etc.)

Quantitative variables

- Continuous
- Discrete

Fundamental concepts

Qualitative variables

- Nominal or Categorical – nominal, categorical set of features
- Ordinal - ordinal scale, order matters

Quantitative variables

- Continuous – range, interval of values
- Discrete – finite set of values

Fundamental concepts

Probability density function (PDF)

- Measures the possibility for a variable to take a certain value. It is a function defined on the set of all possible values of the variable, with values within the interval [0,1]:

$$f(x) = P(X=x),$$

where X is the variable, and x is one of the values that X may take.

Fundamental concepts

Cumulative distribution function (CDF)

- Measures the possibility for continuous random variable X to take values within a certain interval.

$$\mathrm{F}(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

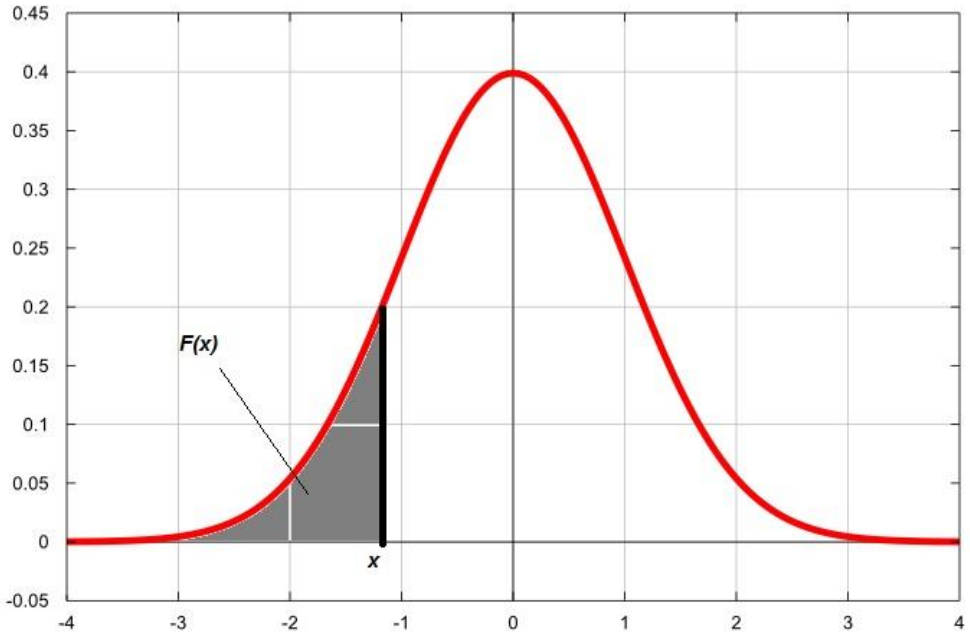
- If X is a purely discrete random variable, then it attains values x_1, x_2, ... with probability $p_i = P(x_i)$.

$$\mathrm{F}(x) = \mathrm{P}(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} P(x_i) = \sum_{x_i \leq x} p_i$$

Fundamental concepts

Cumulative distribution function

- Geometrically, F(x) represents the area below the probability density curve.



Random variable properties

Position (where?)

- Mean (average, typical, central tendency)
- Moments (the *k*-th moment)
- Median (middle)
- Percentile
- Quartiles
- Mode (the most probable value, the value that appears most often in a data set)

Random variable properties

Spread - dispersion, variability, scatter (how?)

- Amplitude
- Variance
- Mean Absolute Deviation (MAD)
- Standard Deviation (SD)
- Coefficient of Variance

Shape of the distribution (why?)

- Symmetry
- Skew

Random variable properties

Mean (average, typical, expected value, central tendency)

Discrete case:

$$E(X) = \mu = \sum_{x \in R} x \cdot f(x) \qquad E(X) = \mu = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where *f(x)* is the probability for variable X to attain value *x*, (probability density); for a normal distribution of *n* subjects.

Continuous case:

$$E(X) = \mu = \int_R x \cdot f(x) dx$$

Random variable properties

Moments (the *k*-th moment of a variable)

If the points represent probability density, then:

- the zeroth moment is the total probability (i.e. one),
- the first moment is the mean,
- the second central moment is the variance,
- the third central moment is the skewness,
- and the fourth central moment (with normalization and shift) is the kurtosis.

Random variable properties

<u>Moment</u> (the *k*-th moment of a variable)

Discrete case:

$$M_k(X) = M_k = \sum_{x \in R} (x - c)^k \cdot f(x) \qquad M_k = \frac{1}{n} \sum_{x \in R} x^k$$

The first (raw, or crude) moment is the mean (c=0, moment about zero); for a uniform distribution of *n* values.

Continuous case:

$$M_k(X) = M_k = \int_{-\infty}^{\infty} (x-c)^k f(x) dx = \int_R x^k \cdot f(x) dx$$

The *k*-th moment of a real-valued continuous function *f(x)* of a real variable about a value *c* (*c=0*, moment about zero; *c=mean*, moment about mean).

Random variable properties

Median

Splits the set of possible values in 2 subsets:

50% less than it, and 50% higher.

$P(X \leq x_{me}) = 0.5$; where x_{me} is the median value.

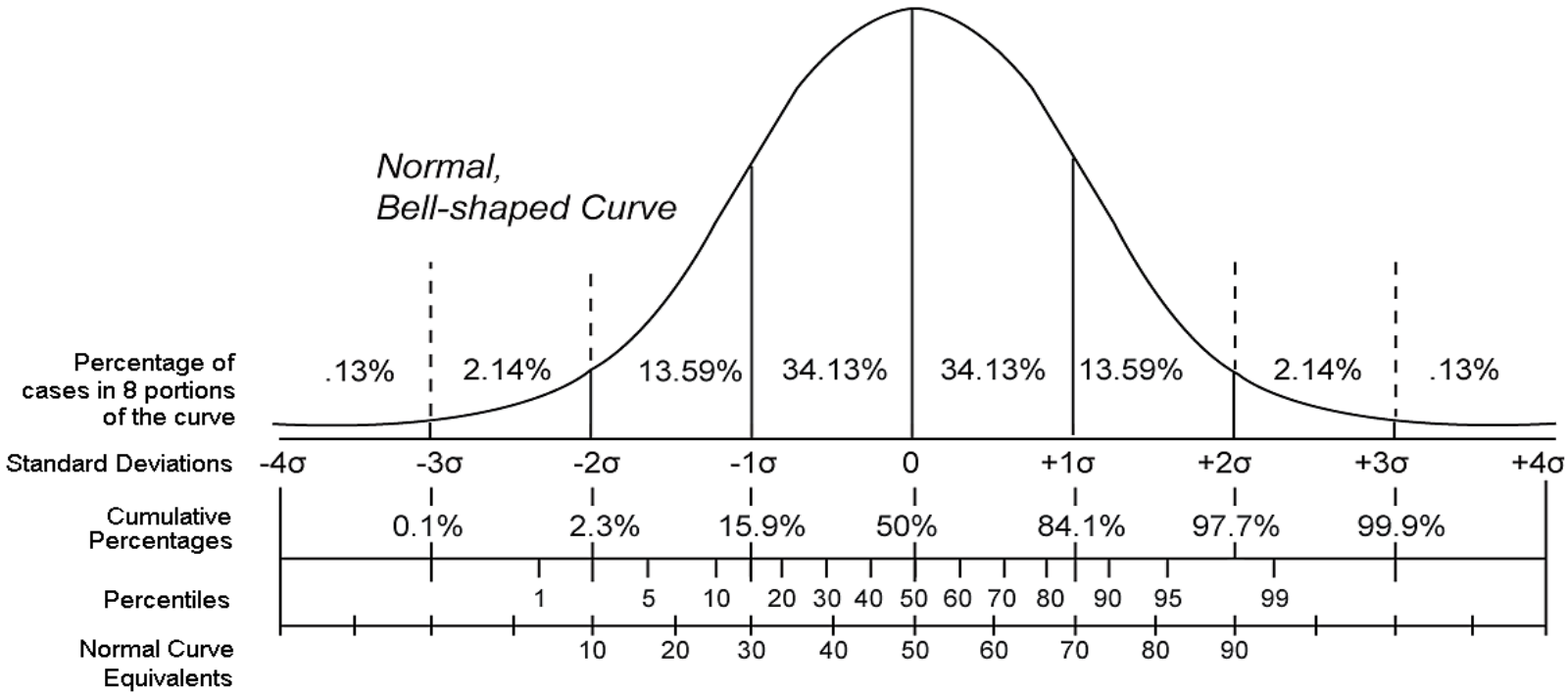
Percentile (centile)

- The value below which a given percentage of observations in a group of observations fall (*inclusive or exclusive*).
- The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3).

Random variable properties

Percentile rank

Percentile rank is the value ***p*** having the property that up to p% of the set values are less than it, and (100-p)% values a greater.



Random variable properties

Percentile rank

Having given a sample set Y_i, $i=1,n$

$Y[k]$ is the rank k element, i.e. there are $k-1$ values less than it.

If $y(p)$ is percentile rank p, then its value is determined as follows:

$$y(p) = Y[k] + d \cdot (Y[k+1]-Y[k]),$$

where:

- k is the integer part of $p \cdot (n+1)/100$, representing the number of values less than p;
- d is $[p \cdot (n+1)/100] - k$, representing the decimal part of $p \cdot (n+1)/100$, or the percentage distance of p from $Y[k]$ to $Y[k+1]$.

Random variable properties

Amplitude

The magnitude of the difference between the random variable's extreme values: maximum and minimum;

$A = X_{max} - X_{min}$

Mean Absolute Deviation (MAD)

Characterizes the spread, dispersion, of a random variable:

- Discrete case: $d = \sum_{x \in R} \left| x - \bar{x} \right| \cdot f(x)$
- Continuous case: $d = \int_{R} \left| x - \bar{x} \right| \cdot f(x) dx$

Random variable properties

<u>Mean Absolute Deviation</u> (MAD)

- For a uniform distribution, cu f(x) = 1/n, where *n* is the number of possible values in discrete case:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Random variable properties

Variance

Is the expectation of the squared deviation of a random variable from its mean. Measures how far a set of (random) numbers are spread out from their mean.

- Discrete case: $\sigma^2 = \sum_{x \in R} \left(x - \bar{x}\right)^2 \cdot f(x)$
- Continuous case: $\sigma^2 = \int_R \left(x - \bar{x}\right)^2 \cdot f(x)dx$
- For a uniform distribution, cu f(x) = 1/n, where *n* is the number of possible values:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

Random variable properties

Standard deviation (SD)

It is preferred for compatibility with the values of the random variable, from the perspective of units of measure:

$$\sigma = \sqrt{\sigma^2}$$

Coefficient of variance (CV)

Also known as relative standard deviation (RSD), is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean.

$$C_v = \frac{\sigma}{\bar{x}}$$

Random variable properties

Moments

- Discrete case: $MC_k(X) = MC_k = \sum_{x \in R} (x - \mu)^k \cdot f(x)$
- Continuous case: $MC_k(X) = MC_k = \int_R (x - \mu)^k \cdot f(x) dx$
- μ being the mean or the expected value of X
- The 2-nd central moment about the mean is the variance:

$$\sigma^2 = \sum_{x \in R} (x - \bar{x})^2 \cdot f(x)$$