Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lesson #2

Academic year 2025-2026

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab/

Recap of the previous lessons

- We live in a world in which **data** are of paramount importance
  - if we want to generate informative data, we need to plan and carry out experiments through **DoE**
  - if we already have data, we need flexible and straightforward tools to summarize, visualize and interpret them in a meaningful manner, namely we need **data analytics and machine learning**
- **Data analytics** (DA) is the process of examining data to extract useful information with the aid of specialized systems and software
- **Machine learning** (ML) exploits mathematical algorithms, statistical models and computational systems to train machines to make decisions without being explicitly programmed to perform the specific task
  - **unsupervised learning** is used to explore data and provide data summary and overview
  - **supervised learning** is used to make predictions based on the information extracted from the relation between inputs and outputs
    - classification
    - regression
- We are conscious of the fact that having (a lot of) data does not necessarily means having also good data
  - **informative data** on the system under study are needed to succeed in DA and ML

Today's lesson

- Today we will speak about:
  - data structure and challenges
  - probability theory and random variables
    - descriptive and inferential statistics
    - probability density functions (PDFs)
    - statistical indices to define a PDF

Data structure

- "Good data" are collected to capture the essential features of a system or a process
- Data contain:

$$x = \eta + \varepsilon$$

  - **systematic information** η about the system/process under study
    - deterministic component of the measurements
  - **noise** ε made of unwanted variations
    - measurement error
    - experimental error
    - sampling error
    - model error
    - etc.

Data challenges

- The main challenges on data are related to:
  1. **variability**:
     - systematic part of the signals should be distinguished from the noise
     - systematic variability can be introduced changing some factors of the system/process, for example using DoE
     - the presence of noise should be considered to avoid drawing misleading conclusion
  2. **complexity**:
     - a system is incomprehensible if the number of measured variables is $V > 3$
       - simple statistics and graphical representations are not effective with (large) multivariate datasets
  3. **nature**: data types can be categorized in several manners:
     - factors and responses
     - quantitative and qualitative
       - quantitative may assume any reasonable real value in a continuous scale
       - qualitative are categorical variables that assumed predetermined levels
     - controlled and uncontrolled
       - controlled variables can be manipulated, set to a determined value and kept there
       - uncontrolled variables are impossible to regulate, but may impact on the system/process

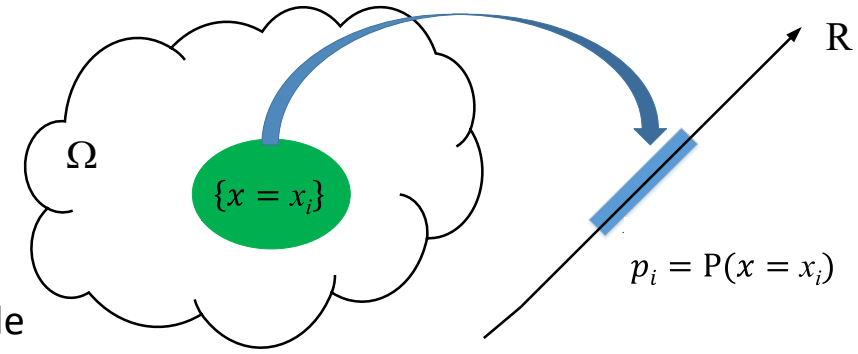
Probability theory

Just a little bit of mathematics…

Probability and random variables

- **Probability** is an assessment of the likelihood of the various possible outcomes in an experiment/ process/ situation with a *random* outcome
  - if we take a sample from a population, what is the probability of that a determined event happens?

- The **probability theory** studies **random variables** (variables which take on their values *by chance*)
  - **discrete random variables** correspond to categorical/discrete quantitative variables
    - their support (namely, their domain) is a finite or infinite list of numeric values which have non-zero probability (e.g. toss of a coin, toss of a die, etc.)
  - **continuous random variables** correspond to continuous quantitative variables
    - defined in the continuous range of real numbers (rarely disconnected ranges)
    - each outcome in the support has probability zero, but none is actually impossible



- The **probability theory** makes **predictions about the chance of occurrence of events based on a set of assumptions** about the underlying probability process:
  - studies in a deductive way what is likely to happen based on mathematical principles and on some assumptions
  - a probability measure is a function P that assigns a numerical probability to each subset of the space of samples
    - each outcome of an experiment gets a probability in the form of a real number **between 0 and 1**
      - we are often more interested in the function of the outcome than in the actual outcome

Frequency

- Suppose that a certain **chance experiment** can be carried out:
  - a very **large number of times**
  - **under exactly the same conditions**
  - in a way so that **the repetitions are independent** of each other
- Let Z be an event in the experiment
- The **relative frequency** of the event Z in N repetitions of the experiment is:
  - $n(\mathrm{Z})$ = number of times the event Z occurred (also called absolute frequency)
  - N = number of the repetitions of the experiment

$$f_n(Z) = \frac{n(\mathrm{Z})}{N}$$

  - the relative frequency with which event Z occurs will fluctuate less as the number of repetitions increases (**empirical law of large numbers**)
- **Defining the probability of occurrence of the event Z with a single repetition of the experiment would be highly desirable**
  - this can be done through **statistical modelling**

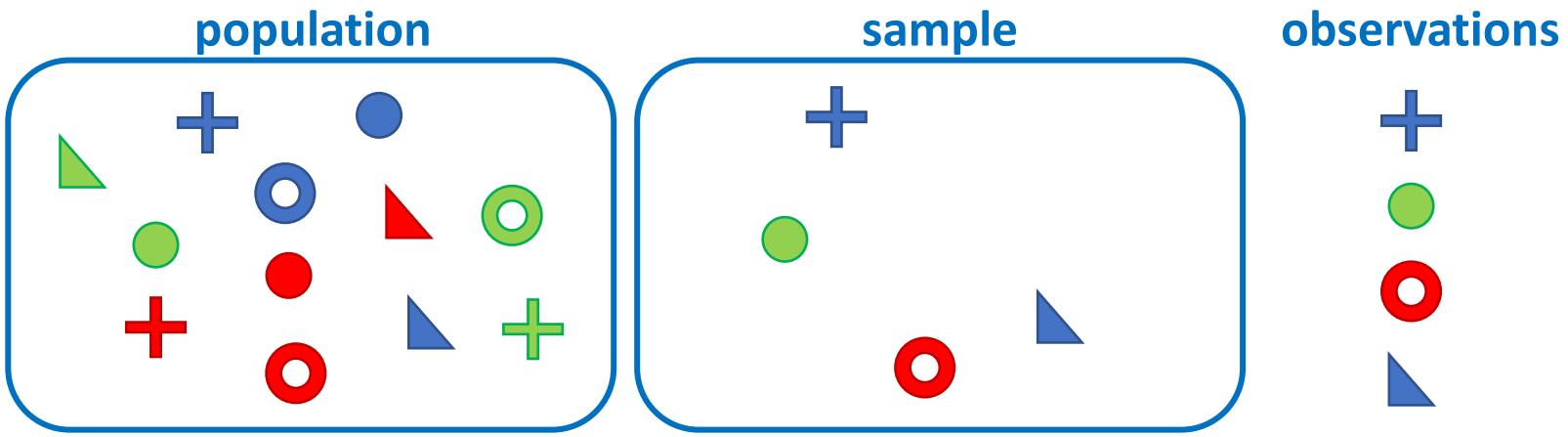
Probability model

- When the experiment can be repeated infinite times under stable conditions we have:
  - the **empirical relative frequencies** of the outcomes
  - the **assignment of the probabilities to the possible outcomes**

- A **probability model** can be built to translate the physical context of the experiment to a mathematical framework:
  - building a probability model is possible when a sample space and the probabilities assigned to the elements of the sample space are available
  - the model should represent the reality with high accuracy

Descriptive and inferential statistics

- When we want to know the characteristics of a population it is desirable:
  - avoid analyzing the whole **population**
  - the analysis of a (reduced) **sample** collected from the population is likely to be representative of the population characteristics with a good degree of confidence
- **Descriptive statistics**
  - analyzes and summarizes the information available from a sample
- **Inferential statistics**
  - estimates the characteristics of the population from the knowledge acquired from descriptive statistics



Statistical analysis

descriptive statistics

- extracts statistical indices to summarize the population features:
  - collects and organizes
  - summarizes
  - simplifies
  - presents

inferential statistics

- transfer the characteristics of a sample to the population
  - generalizes
  - forecasts
  - finds relationschips
  - verifies hypothesis

Challenges in data analytics

- A lot of data are typically available for process engineers
- Consolidated **mathematical and statistical tools** are used in data analytics and machine learning to extract information on data
- However…

- … dealing with data means facing several **challenges** in terms of:
  - **data variability**
    - we need to understand what is the systematic part of the data
    - noise have to be discarded
  - **data complexity**
    - what can we do if the number of variables is very large?
    - how can we summarize the information stored into the data
  - **data nature**
    - factors-responses; qualitative-quantitative; controllable-uncontrollable, etc…

Uni-variate data analytics

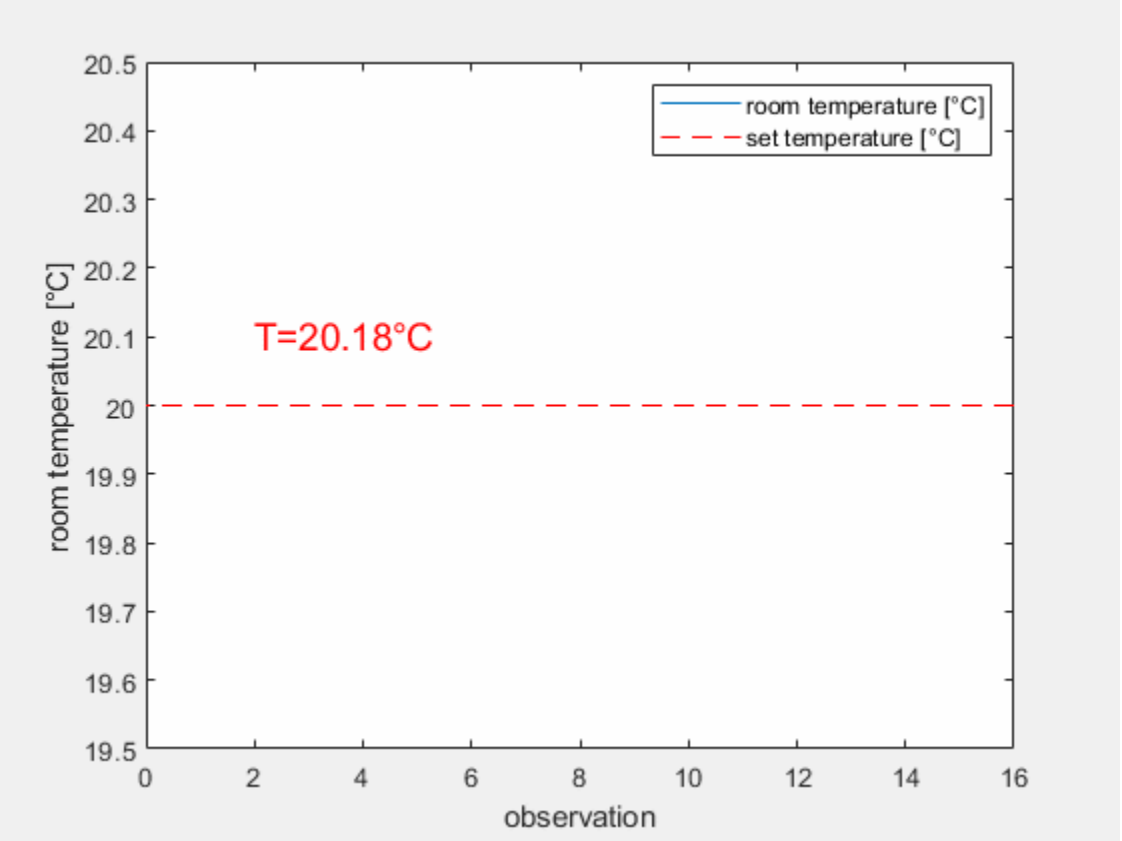
Brief introduction on practical aspects of the probability theory

Let’s start from the simplest case

- We start from the study of a simple case:
  - we want to fully describe the **variability of a random variable**:
    - this type of variables varies by chance under the action of **normal (or common cause) variability** and **no special causes** determine its variability
  - at the **lowest complexity**:
    - one single variable describes the system under study
  - in a quite general case where the **nature** of the variable is not necessarily specified

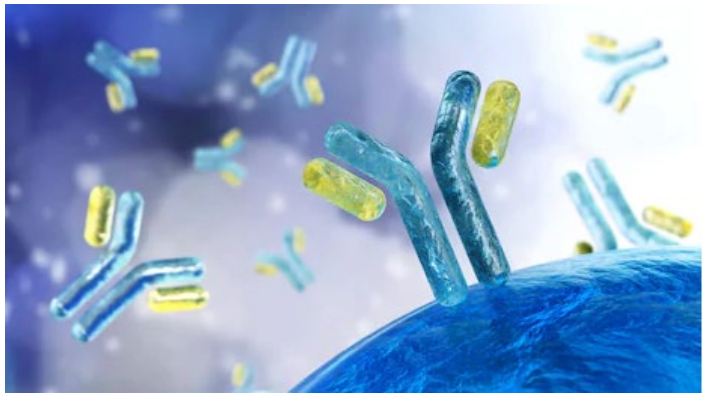
Simple starting example

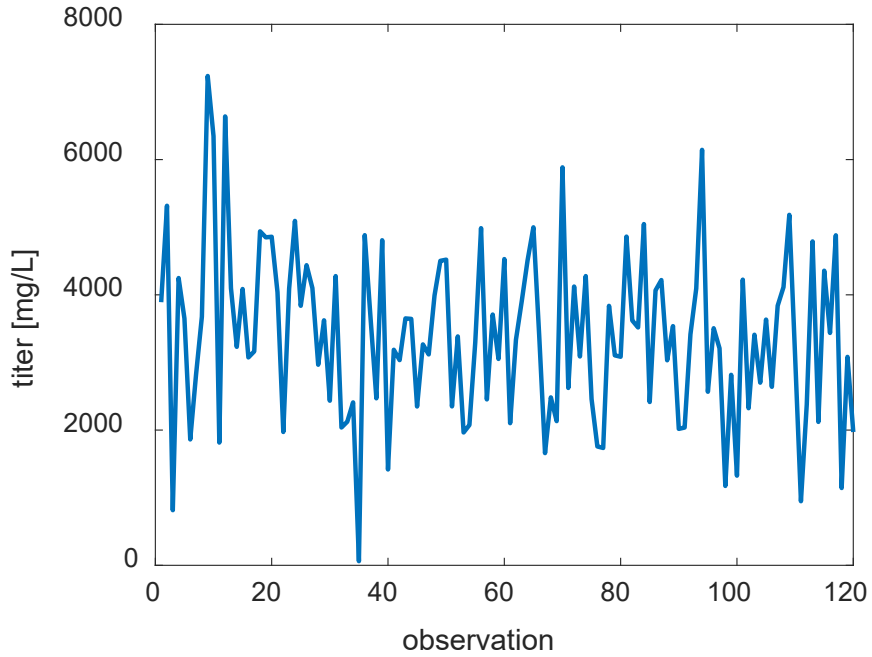
- Let’s consider the temperature of this room:
  - measurement repeated several times
- The measurement is affected by a **natural variability** which is due to several causes:
  - cumulative outcome of unavoidable phenomena due to:
    - measurement system
    - environment
    - etc.
  - **background noise** with chance causes of variation



Random variable in the biopharmaceutical industry

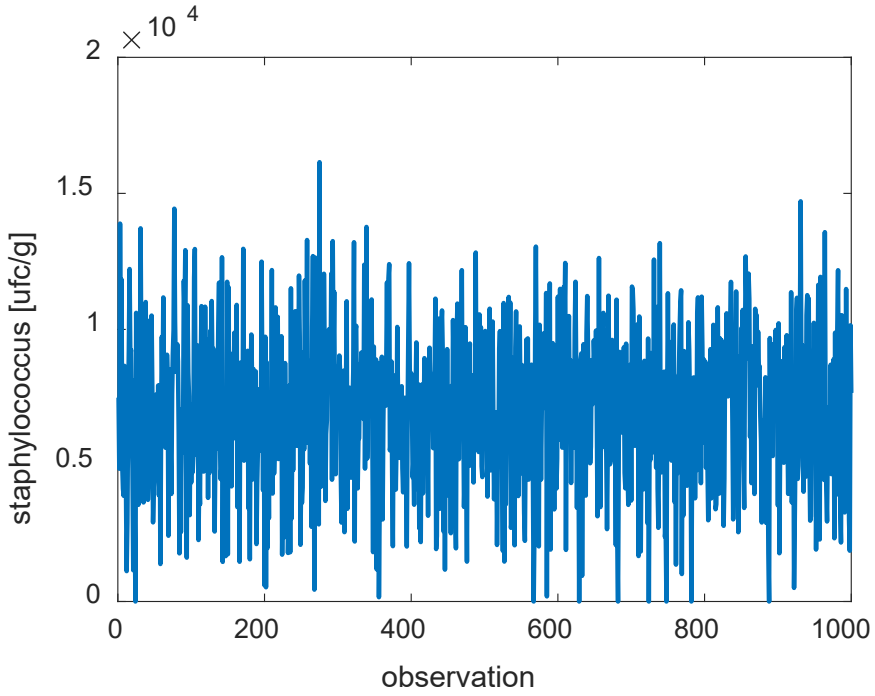
- Process of production of monoclonal antibodies
  - the productivity of clones of the same cell line is evaluated
- The final titer (namely, a measurement of productivity) at the AMBR15® scale is, as an average, 3300 mg/L
  - there is a large variability
  - 120 measurements are taken





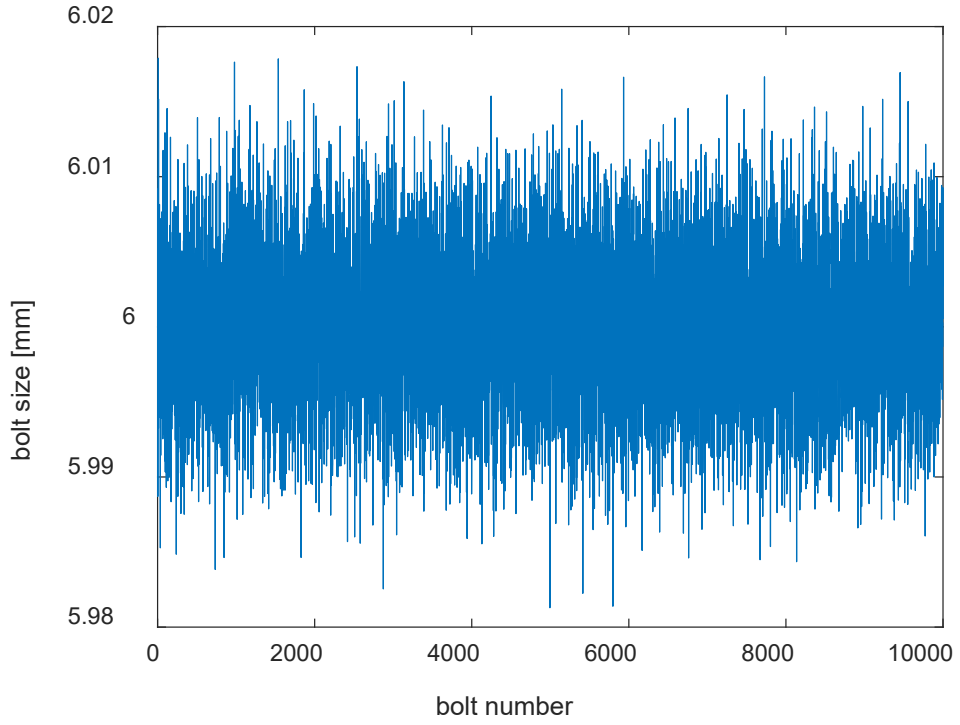
Random variable in food industry

- Staphylococcus contaminations are evaluated in a wide sample of sausages
- The contamination level stays typically close to a «typical» value
  - a **natural variability** is found (**common-cause variability**)
  - common cause variability is determined by a series of random phenomena



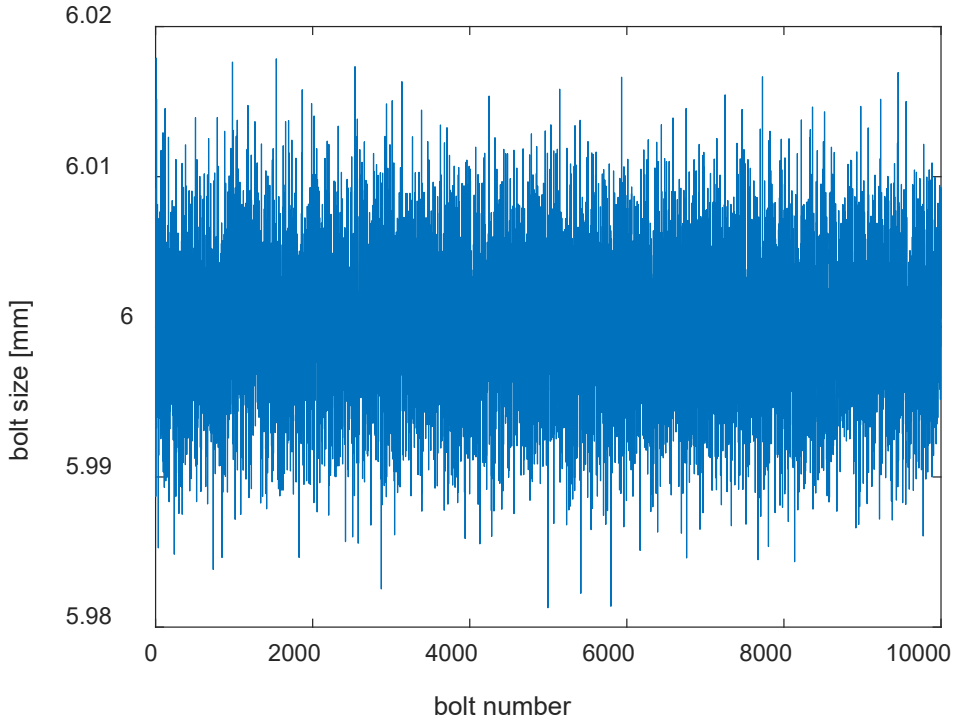
Example of the production of bolts

- Only **natural variability** affects the process of production of bolts
- The bolts desired diameter is 6 mm, however there is a tolerance in their dimension
  - 100 000 bolts are measured



Some considerations on the bolt measurements

- The measured values of the bolt diameters are not stably at the target value of 6 mm
- Natural variability determines a certain variation of the bolt diameter around the target



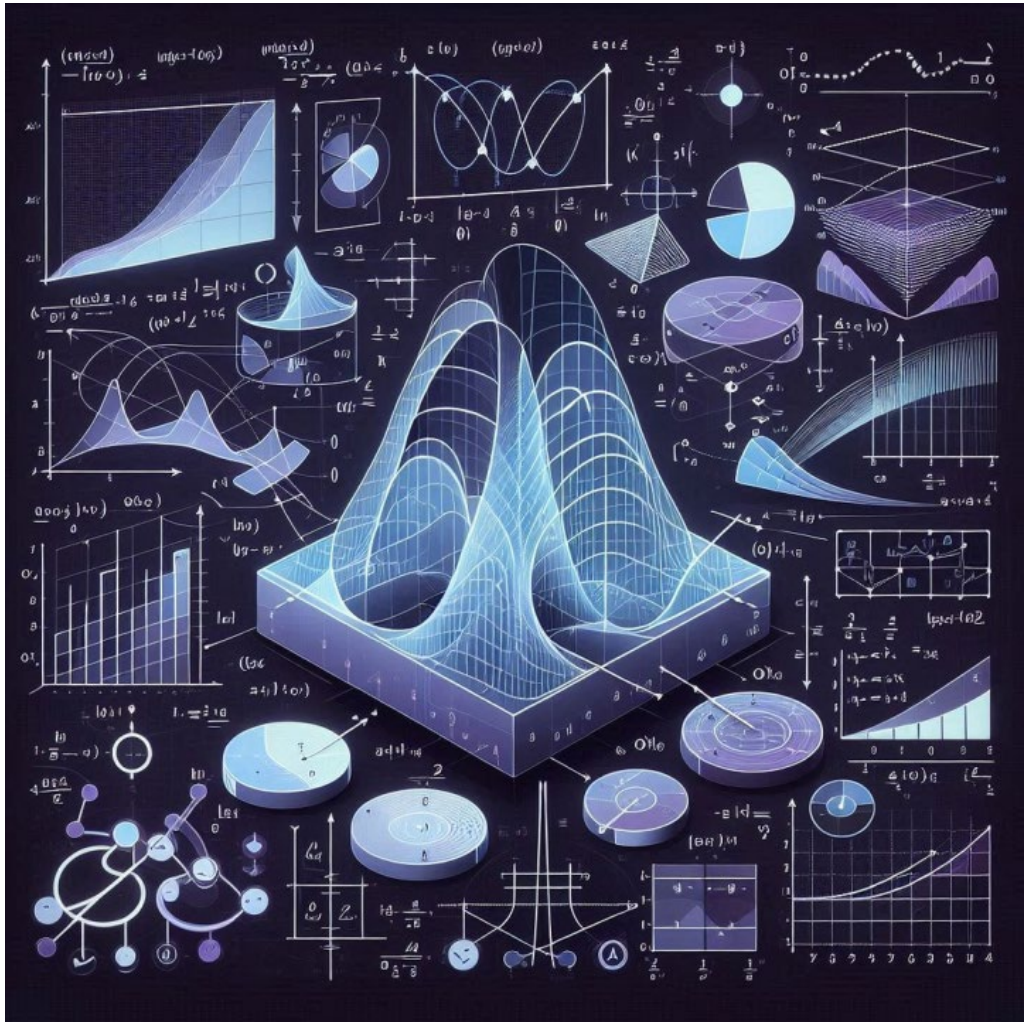
How is a sample of a random variable?

- If a sample is sufficiently wide to represent the entire population, a random variable, on which only common-cause variability acts, has the following characteristics:
  - **values are denser in proximity of a nominal value** (target, average)
    - the values which are closer to the nominal value have a larger probability of occurrence
  - **values are less likely when they are far from the nominal one**
    - the further the value are from the nominal one, the least likely they are to happen
  - typically:
    - the most likely value is the nominal one
    - two values that are at the same distance from the nominal one have often the same probability of occurring

Mathematical representation to random variables

- A **mathematical model** of a random variable should represent:
  - the frequency of occurrence of the events
  - the density of the frequency

- The probability theory describes the **probability that an event occurs for a random variable**, where the random variable assumes different values due to random mechanisms (common-cause variability)



Distribution of the single units

- List of all the observed values of contamination for all the inspected units of the sausage

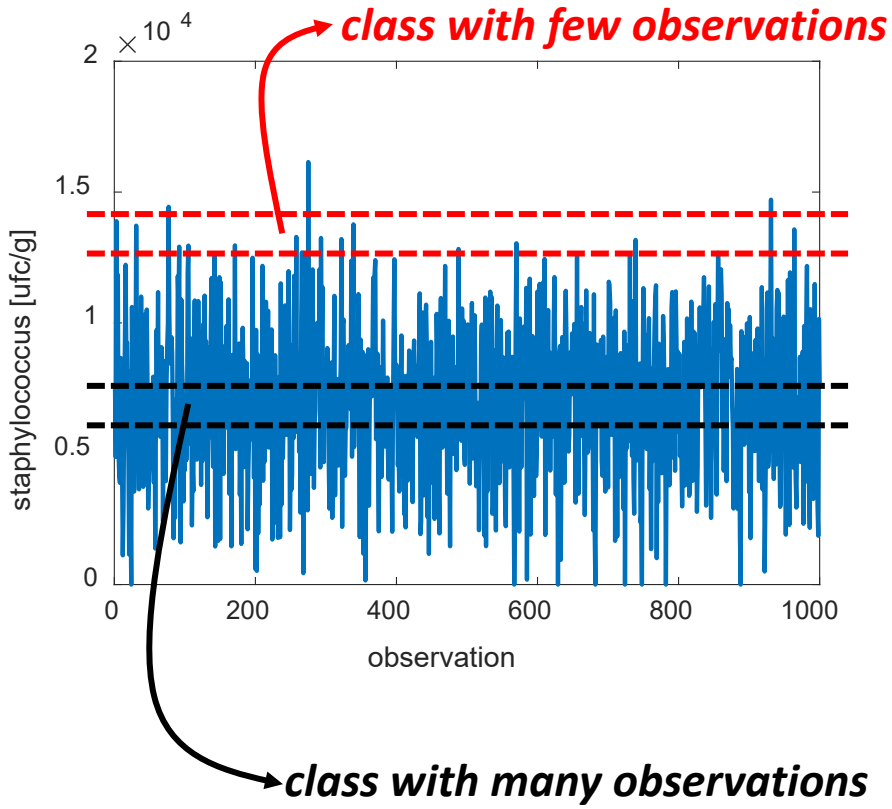
observation	staphylococcus (ufc/g)
1	7491
2	4879
3	13888
4	8282
5	11835
6	5046
7	4139
8	3884
9	8660
…	…
1000	9452

Frequency distribution

- The analysis of the phenomenon under study (namely, contamination) begins summarizing the units through a **frequency distribution**
- The frequency distribution describes how a variable is distributed in its population:
  - frequency allows summarizing important information about the variable
    - what are the most frequent values to happen
    - uniformity of distribution
    - different weight of some values
  - one can associate to each unit value:
    - **absolute frequency** = number of times the value is present in the population
    - **relative frequency** = absolute frequency divided by the number of units in the sample
    - **percent relative frequency** = 100 · relative frequency %

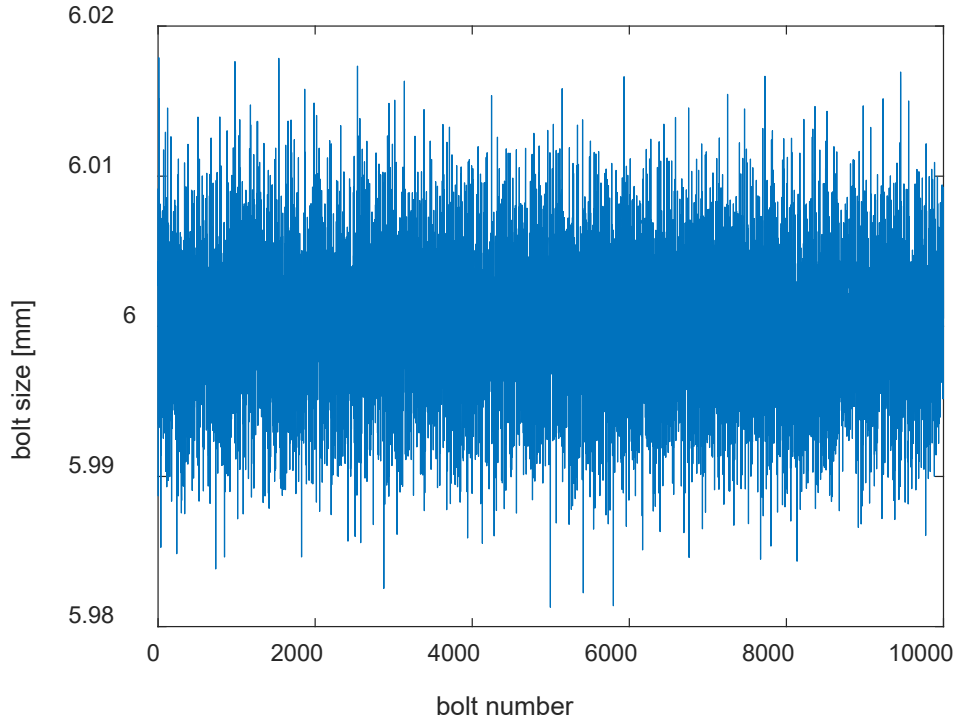
Frequency distribution calculation

- An appropriate number of classes is selected to cover the entire range of possible outcomes of the random variable under study:
  - depending on:
    - observed values dispersion
    - numerosity of the observations
    - objective of the study
  - one must:
    - include all the observed values
    - avoid overlapping between classes
  - one should consider:
    - intervals of even length if you want to pay attention to the variable distribution
    - intervals of uneven length if you want to pay attention to different types of classes
  - variable nature should be considered:
    - qualitative
      - all the variables' classes must be considered
      - the respective frequencies computed
    - quantitative and continuous
      - different classes are represented by intervals of real values
      - the respective frequencies computed counting the number of observations in that interval



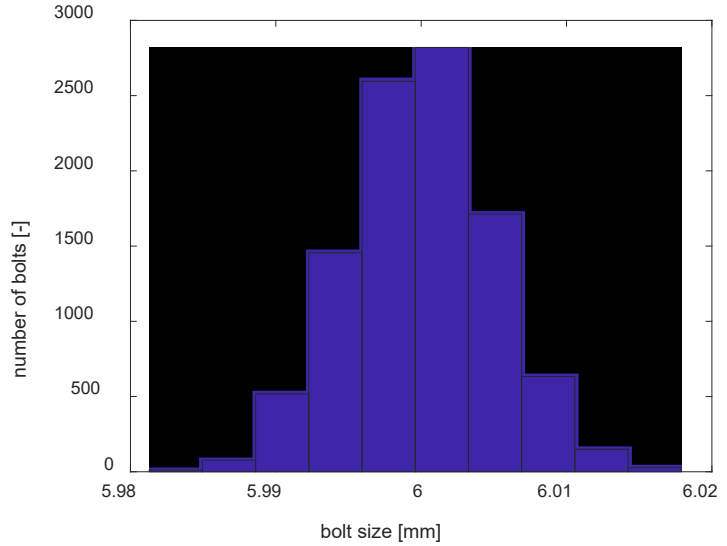
Histograms of the bolts diameter

- For example, the measurements of the bolts diameters can be summarized through histograms

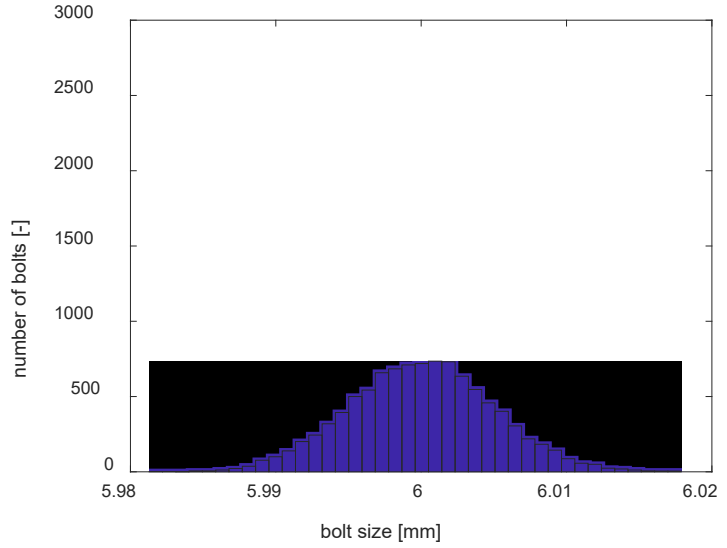


`hist` or `histogram`

low number of bins

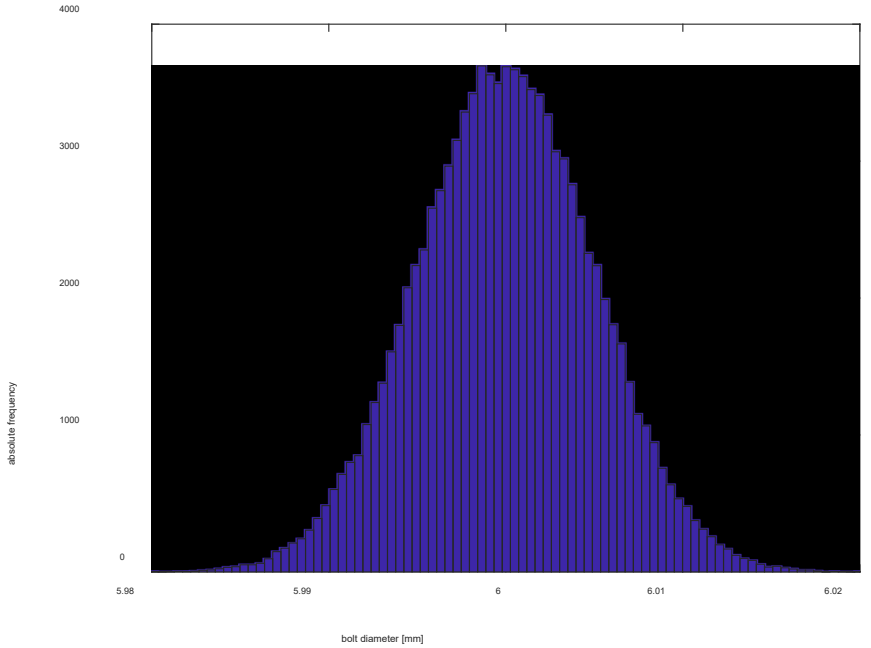


large number of bins



Histogram

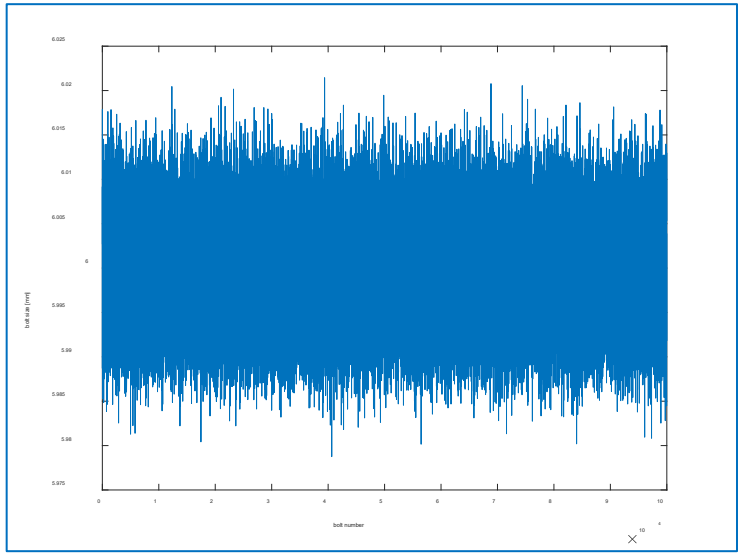
- The **histogram** is a bar diagram representing the variable under study in different bins
  - x-axis: bins of predetermined width
    - the appropriate number of bins $\sqrt{N}$ can be approximately chosen as the square root of the number N of the collected observations
  - y-axis: absolute or relative frequency of the observations in the specific bin
- The histogram gives a precise idea of the data:
  - position
  - variability
  - shape



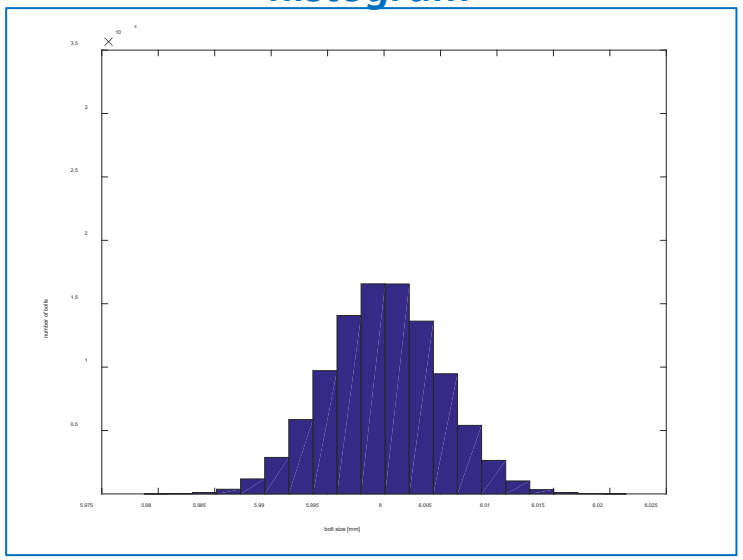
Probability density function

- The normalized histogram with bins of infinitely small dimension (namely, an infinitely large number of bins) generates a **probability density function**

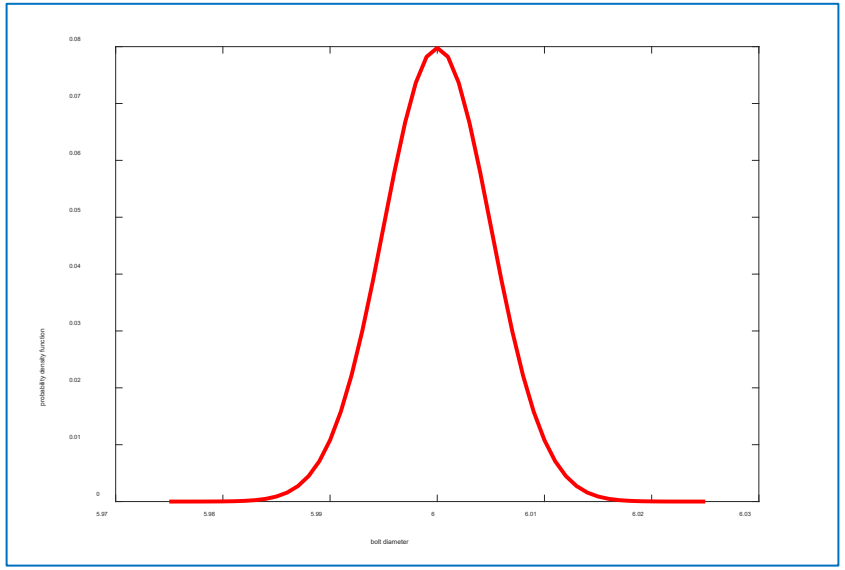
raw data



histogram



probability density function

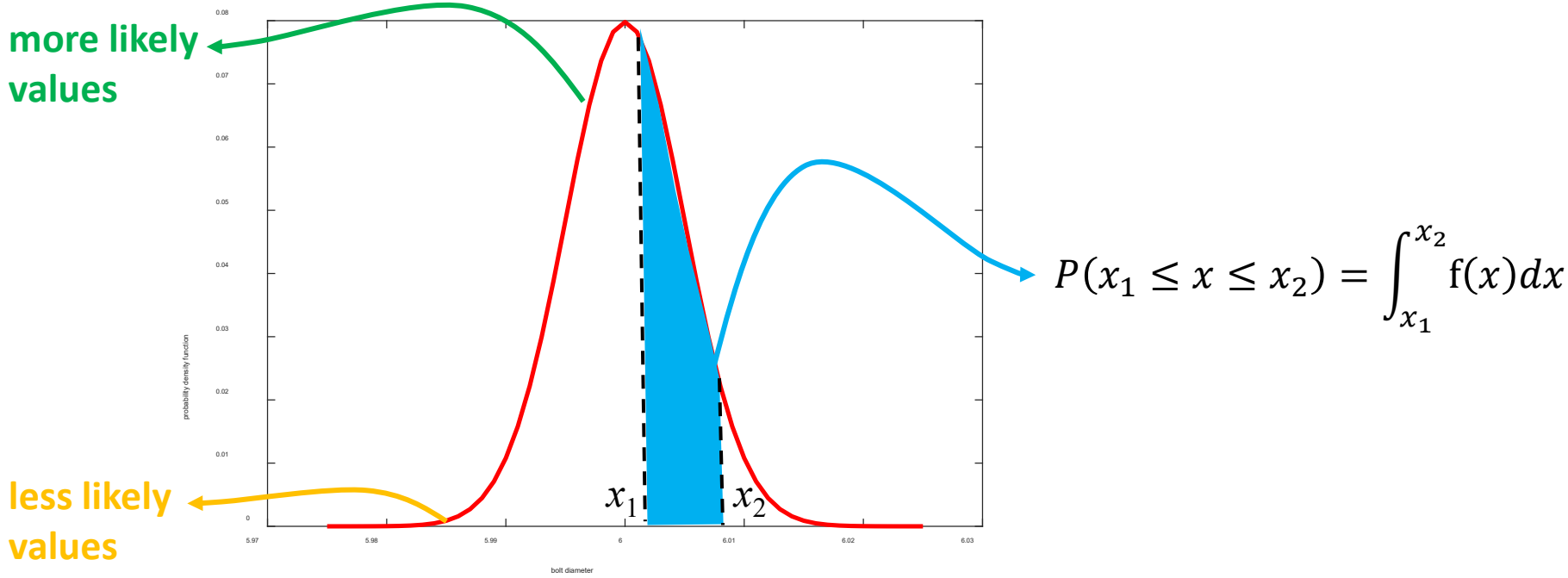


Probability measure and axioms

- The **probability function** $\mathrm{P}(A)$ is the sum of the probabilities of the individual outcomes in the set A of samples
  - a numerical value $\mathrm{P}(A)$ is assigned to each event of A in the sample space
  - $\mathrm{P}(A)$ tells how likely A is to occur
- The **probability measure** is the function that assigns the numerical probability $\mathrm{P}(A)$ to each subset A of the sample space
- **Axioms**:
  - $P(A) \geq 0$ ∀ event A
  - $P(A) = 1$ when A corresponds to the sample space
    - the total area under the curve of the PDF is 1
      - we will see later what this means
  - $P(A \cup B) = P(A) + P(B)$ for disjoint (i.e.: independent) events A and B

Reading a probability density function

- Plot of the probability density function (**PDF**):
  - $\mathrm{f}(x)$ vs. possible outcomes values
  - more likely outcomes display higher frequencies
- Many PDFs are available in families which are indexed by one or more parameters



Probability density functions PDF

- **Probability density function** $\mathrm{f}(x)$:
  - does not represent probability directly
  - the probability $\mathrm{P}(x)$ that an outcome falls within a certain range $[x_1, x_2]$ can be calculated using the following integral:

$$\mathrm{P}(x_1 \leq \mathrm{x} \leq x_2) = \int_{x_1}^{x_2} \mathrm{f}(x)dx$$

  - is applied to continuous random variables
    - in the case of discrete variables, a probability mass function can be found
- The **PDF parameters** give detailed information on the data

Teamwork: discussion with your mates

- What parameters entirely and unambiguously determine a PDF?

Probability density functions PDF

- **Probability density function** $\mathrm{f}(x)$:
  - does not represent probability directly
  - the probability $\mathrm{P}(x)$ that an outcome falls within a certain range $[x_1, x_2]$ can be calculated using the integration:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \mathrm{f}(x)dx$$

  - is applied to continuous random variables
    - in the case or discrete variables, a probability mass function can be found
- The **PDF parameters** give detailed information on:
  - **position**
  - **dispersion**
  - **shape**, etc…

Central tendency of a population

- The **population mean** of a probability distribution is a measurement of the central tendency (i.e., location) of a population:
  - for continuous variables:

$$\mu = \int_{-\infty}^{+\infty} x\mathrm{f}(x)dx$$

  - for the case of a discrete random variable with *exactly N equally likely values*, meaning that $\mathrm{p}(x) = 1/N$, it is reduced to:

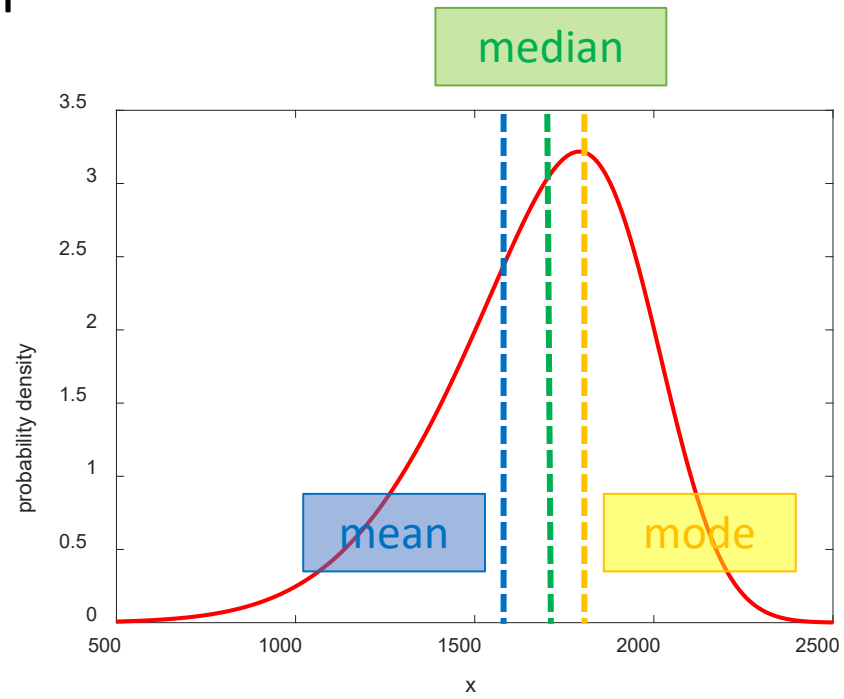
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- The mean may be expressed also in terms of the **expected value** or the long-run average value

$$\mu = E(x)$$

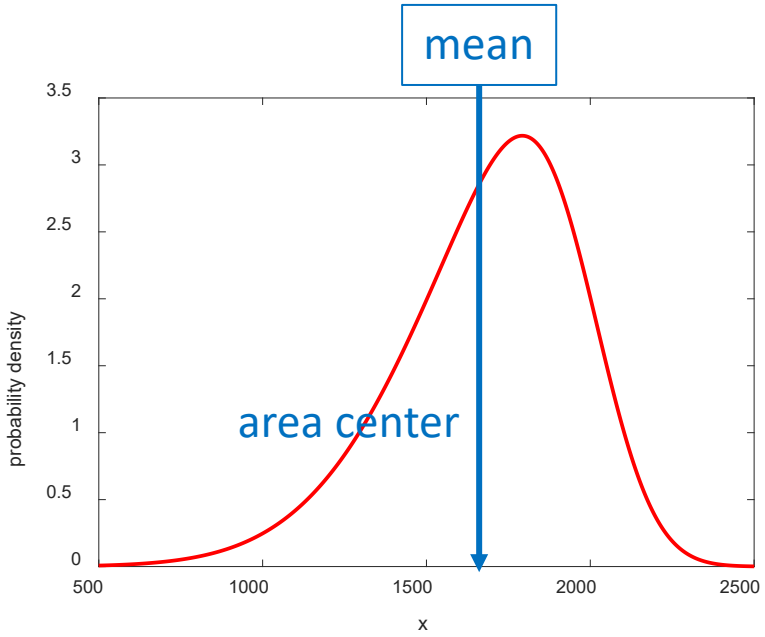
Mean and other location parameters

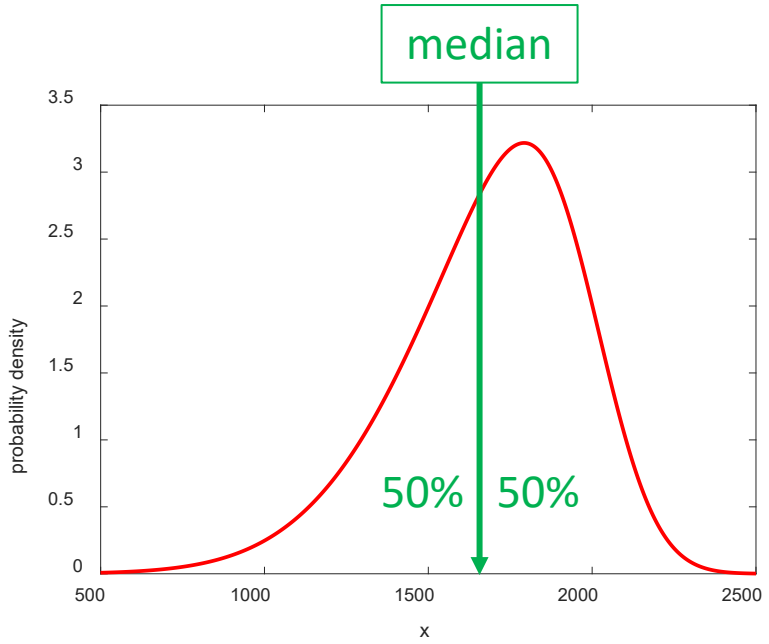
- The mean is the **center of mass** of the probability distribution:
  - it is the point at which the distribution exactly "balances"
  - it simply determines the **location** of the distribution
  - it suffers the presence of outliers
- The mean is not necessarily:
  - the **median**:
    - the 50th percentile of the distribution*
    - separates the higher half of the population from the lower half
  - the **mode**:
    - the most likely value of the variable
    - the value that appears most often

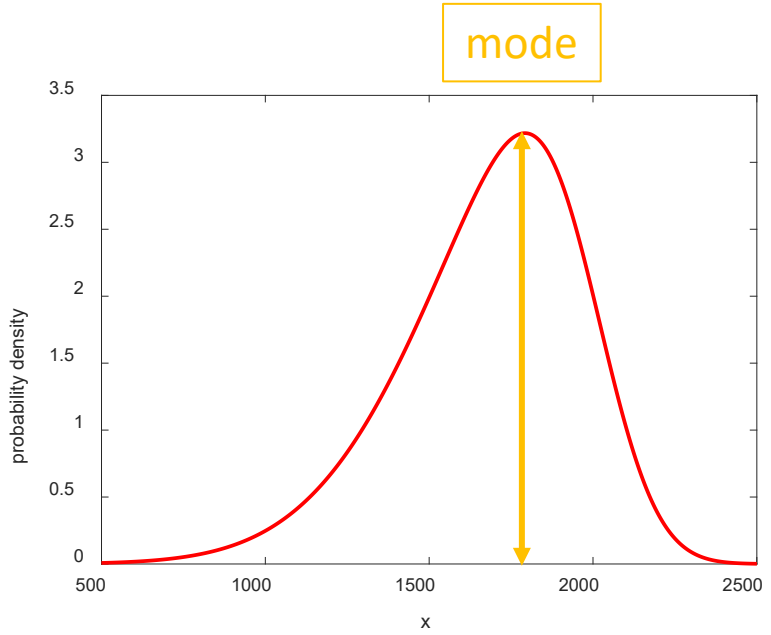


* the *n*-th **percentile** indicates the value below which *n*% of population observations falls
Example, the 20-th percentile is the value below which 20% of the observations are found

Mean vs. median vs. mode







Variability of a population

- The **population variance** indicates the variability of a probability distribution, namely the spread (or the dispersion, the scatter, etc…). It is usually the mean squared distance from the mean
  - for continuous variables:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \mathrm{f}(x) dx$$

  - for discrete variables:

$$\sigma^2 = \sum_{i=1}^{\infty} (x_i-\mu)^2 \, \mathrm{p}(x_i)$$

- **Standard deviation** is the square root of the variance
- The variance can be expressed in terms of **expectation**, as well:

$$\sigma^2 = \mathrm{V}(x_i) = E[(x_i-\mu)^2]$$

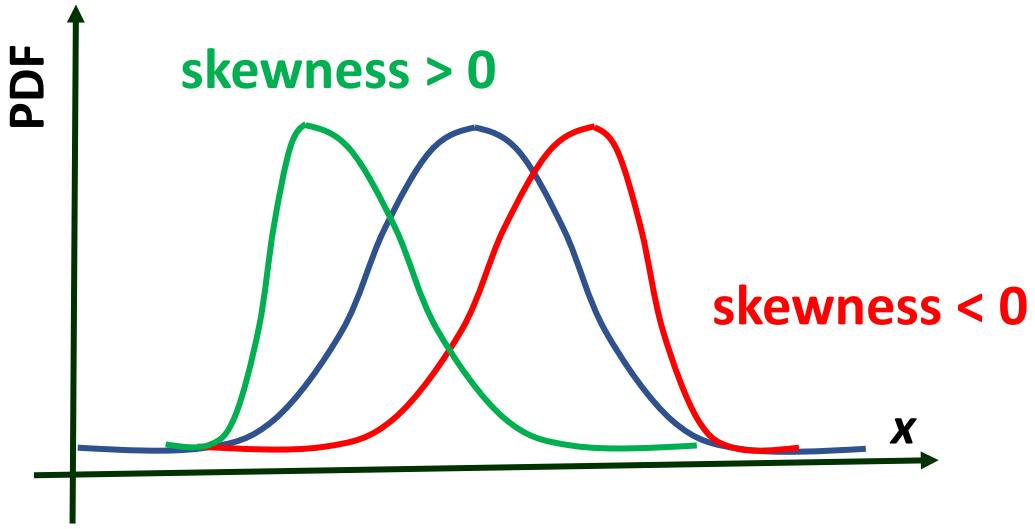
Parameters of a distribution

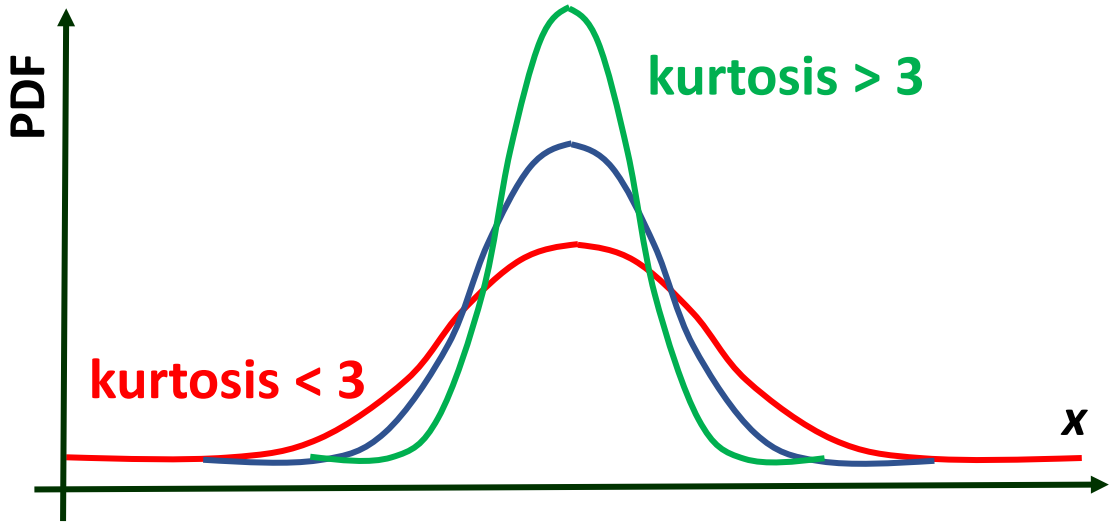
- The n^{th} **moment** of a real-valued continuous function of a real variable about the value c is:

$$\mu_n = \int_{-\infty}^{+\infty} (x-c)^n f(x) dx$$

- The moments identify the **shape** of a probability distribution:
  - the 1st raw moment is the **mean**
  - the 2nd central moment (i.e., a moment of a probability distribution of a random variable about the random variable's mean) is the **variance**
  - the normalized 3rd central moment of a random variable is the **skewness**: $\frac{\mu_3}{\sigma^3} = \frac{E[(x-\mu)^3]}{\sigma^3}$
    - the skewness is a measurement of the *asymmetry* of a distribution
    - the skewness is:
      - = 0 for a Gaussian distribution
      - > 0 (< 0) for a distribution pulled out towards lower (higher) values
  - the normalized 4th central moment of a random variable is the **kurtosis**: $\frac{\mu_4}{\sigma^4} = \frac{E[(x-\mu)^4]}{\sigma^4}$
    - the kurtosis is a measurement of the *peakedness or flatness* of a distribution
    - the kurtosis is:
      - = 3 for Gaussian distribution
      - > 3 (< 3) for sharply shaped and fat tails (rounder shoulders and thin tail) distribution

PDF shape: skewness and kurtosis





Statistical inference

- The aim of the **statistical inference** is to draw conclusions about a population using a sample from that population
  - random samples are usually assumed to be used for statistical inference
    - it is a sample selected from the population in such a way as that every possible sample has an equal probability of being selected
- A **statistic** is a function of the observations in the sample that does not contain unknown parameters
- Let $x_1, x_2, \ldots, x_N$ be the observations of a sample:
  - the **sample mean** is:

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

  - the **sample variance** is:

$$S^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}$$

  - the **sample standard deviation** is:

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

Properties of sample mean and sample variance

- An **estimator** of an unknown parameter is a statistic that corresponds to that parameter:
  - a point estimator is a random variable
  - a particular numerical value of an estimator, computed from sample data, is called an **estimate**
- The *sample mean* $\bar{x}$ is a **point estimator** of the *population mean* μ
- The *sample variance* S^2 is a point estimator of the *population variance* σ^2
- Several properties are required of good point estimators and among them:
  - the point estimator should be **unbiased**
    - the long-run average or expected value of the point estimator should be equal to the parameter that is being estimated
      - unbiasedness is desirable, but this property does not always make an estimator a good one
  - an unbiased estimator should have **minimum variance**
    - this property states that the minimum variance point estimator has a variance that is smaller than the variance of any other estimator of that parameter

Some additional inferred statistics

- Position:
  - **percentiles**: the n-th percentile is the value below which $n\%$ of observations falls in a group of observations
    - example: the 23-rd percentile is the value below which 23% of the observations may be found
    - rounding to two decimal places in a Gaussian distribution:
      - -3σ is the 0.13-th percentile
      - -2σ is the 2.28-th percentile
      - -1σ is the 15.87-th percentile
      - 0σ is the 50-th percentile
    - **quartiles**:
      - 25-th percentile = first quartile Q1
      - 50-th percentile = second quartile Q2
        - it corresponds to the median
      - 75-th percentile = third quartile Q3
- Variability:
  - **range**: difference between maximum and minimum observed value
    - it is a measure of statistical dispersion
  - **inter-quantile range IQR**: is equal to the difference between 75-th and 25-th percentiles (namely, between Q3 and Q1 quartiles)

$$IQR = Q3 - Q1$$

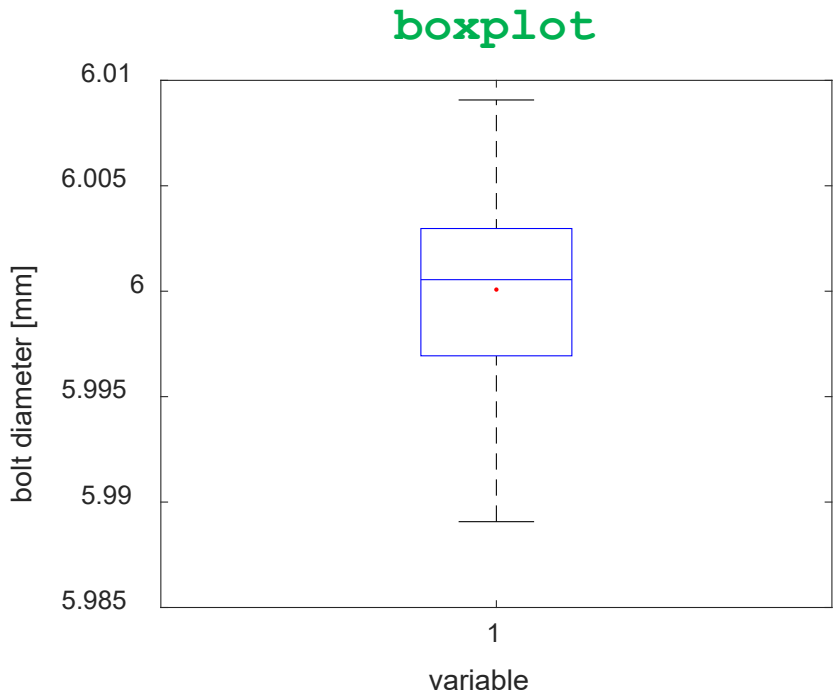
How to infer statistics from observations?

- Commercial and free software can be a valid support to infer statistics from a sample (see Matlab® examples)

statistics	Matlab®
sample mean	`mean`
sample median	`median`
sample mode	`mode`
sample variance	`var`
sample standard deviation	`std`
skewness	`skewness`
kurtosis	`kurtosis`
n^{th} percentile	`prctile`
moment	`moment`
inter-quartile range	`iqr`
minimum	`min`
maximum	`max`
range	`range`

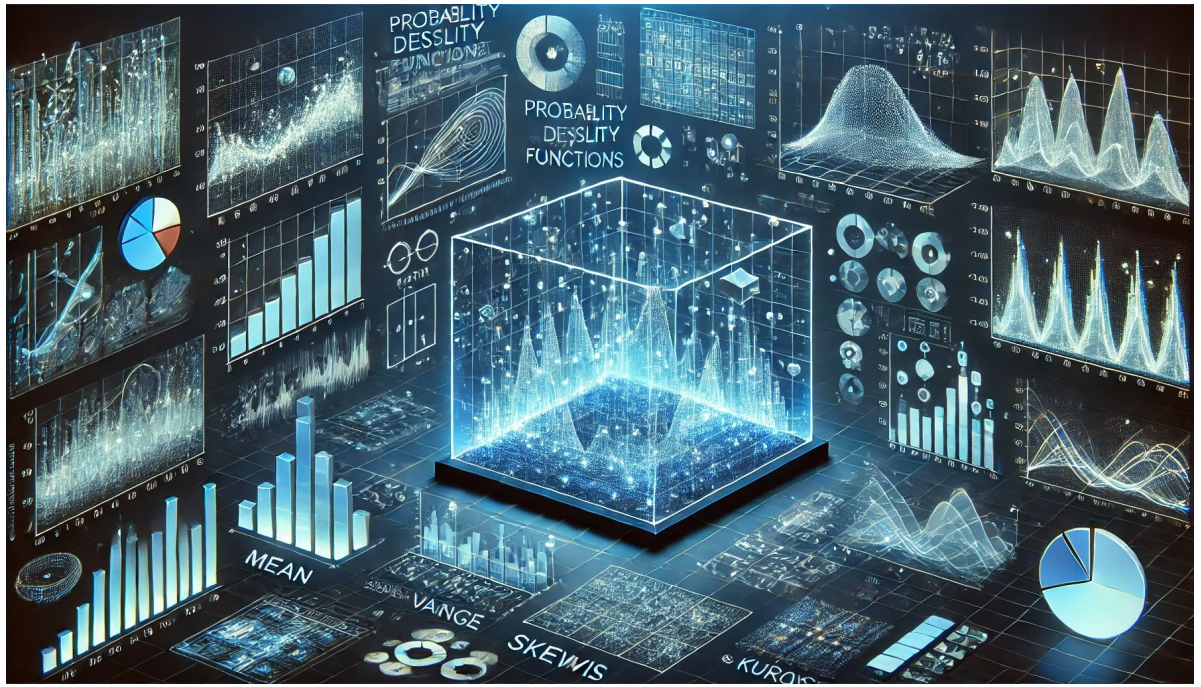
Graphical representations of the distributions

- **Box and whiskers plot**: a non-parametric method (no assumptions are made on the underlying distribution) for graphically depicting data through their quartiles:
  - box indicates the quartiles Q1 and Q3
  - whiskers indicate variability outside the upper and lower quartiles
  - indicate the degree of dispersion and skewness
  - visually estimate various parameters (e.g.: IQR is box dimension)
  - show outliers



Take-home message

- Today we leaned what are:
  - descriptive statistics
  - inferential statistics
- Furthermore, we introduced:
  - probability theory
  - probability
  - probability density function
  - descriptive statistical indices:
    - position
      - mean (median, mode, percentiles, quartiles)
    - dispersion
      - variance and standard deviation (range, IQR, etc.)
    - shape
      - moments (skewness and kurtosis)
  - simple visualization tools of data:
    - histogram
    - box plot



… and now let’s practice

A funny example:…

- Are there 6 brave students for an experiment?
  - OK! Let's start a Russian roulette with 6 cylinders and one bullet in a revolver

  - see as an example:
    `russian_roulette.m`

Random variables in practice

- Different examples of random variable generators are available in Matlab®:
  - **`randn(size_dimension_1, …, size_dimension_N)`**
    - returns a [size_along_dimension_1, …, size_along_dimension_N] array of random scalars drawn from the standard normal distribution
  - **`rand(size_dimension_1, …, size_dimension_N)`**
    - returns a [size_along_dimension_1, …, size_along_dimension_N] array of uniformly distributed random numbers
  - **`randi(size_dimension_1, …, size_dimension_N)`**
    - uniformly distributed pseudorandom integers

Examples: tossing fair coins, dice, etc…

- Flipping a coin:
  - use **`datasample`** in Matlab®
  - see as an example: **`toss_a_coin.m`**

- Tossing a die:
  - use **`randsample`**
  - see as an example: **`rolling_dice.m`**

Everyday homework

- Just play with commands:
  - `randn`
  - `datasample`
  - `randsample`
  - `randperm`

… per sempre a fianco a me!