Università degli Studi di Padova

Department of Industrial Engineering

Machine Learning
Lesson #3

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 lectures

- Currently, in the process industry (and also in everyday life) we are plenty of data from different sources
  - however, we are not interested in having a lot of data. We prefer to have informative data!
- Data are structured in a way in which they sum up 2 components:
  - **systematic part**, descriptive of the system under study
  - **noise**, unavoidable, but unwanted part of the signals due to disturbances
- Dealing with data means dealing with some challenges:
  - **variability**
  - **complexity**
  - **nature**
- We started from the **simplest case**: **study of a single variable**
  - to understand how, through probability theory and univariate statistics, it is possible to describe data (**descriptive statistics**)
  - to extract the most important information about a population from a sample of a limited number of observations (**inferential statistics**)
  - to effectively describe a variable by means of few effective graphical tools and a parsimonious number of parameters describing the population features on **how data are distributed**

Today's lesson

- Today we will introduce:
  - **Gaussian (i.e., normal) distribution**
    - inferred statistics for distributions
  - visual interpretations of distributions
  - **tests of normality**
  - density functions
    - **probability density function**
    - **cumulative density function**
    - **inverse density function**
  - probability in Minitab® and Matlab®

Sampling distribution

- The **sampling distribution** is the probability distribution of a statistic
- The probability distribution of a particular statistic can be determined if the probability distribution of the population, from which the sample was collected, is known
- Several *useful sampling distributions* are available in the scientific Literature and widely utilized

- We will introduce some distributions that will be very useful in (multivariate) statistical process control and design of experiments:
  - **Gaussian distribution**
  - **t distribution**
  - **chi-squared distribution**
  - **F distribution**

Gaussian (normal) distribution

Normal (Gaussian) distribution

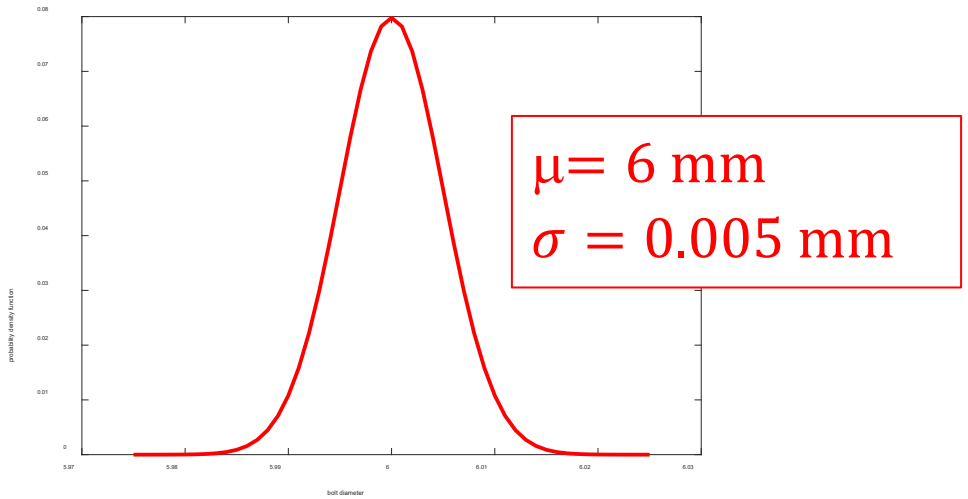
- The **normal probability distribution** of a random variable x is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for} \quad -\infty < x < \infty$$

with $\quad -\infty < \mu < \infty$

$$\sigma^2 > 0$$

- The typical shape of a normal distribution is the one of a bell or a «Mexican hat»

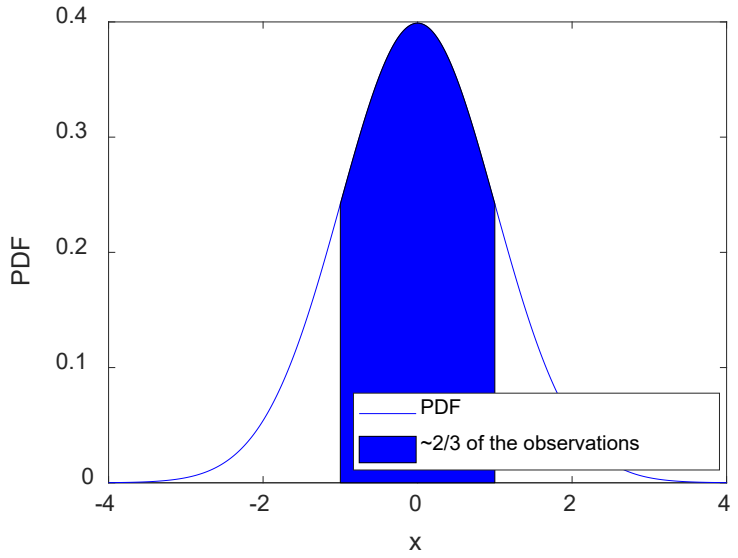


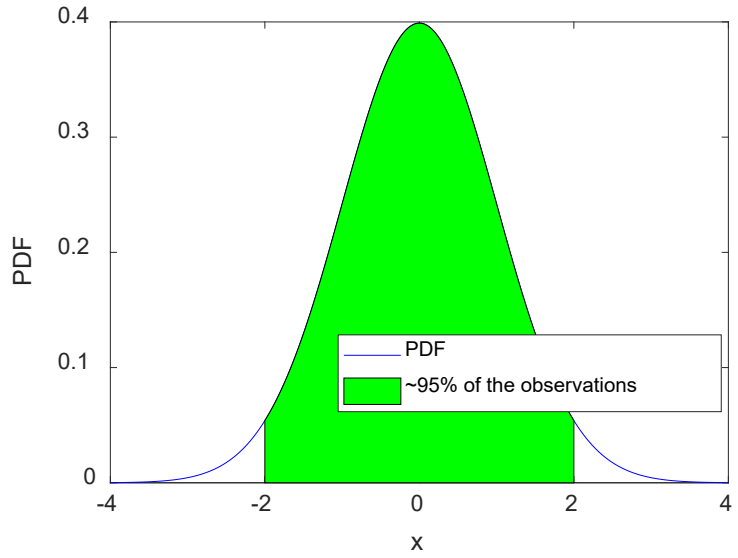
Some details on normal distributions

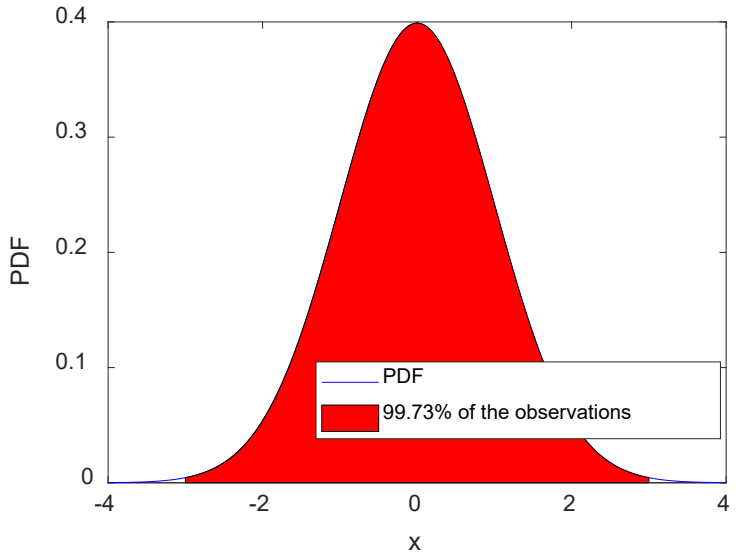
- Normal distributions are simple because they are **fully characterized by only 2 parameters**:
  - **mean μ** indicating data central tendency
  - **variance σ^2** (or standard deviation σ) indicating data dispersion around the location
- The notation to identify a normal distribution of mean μ and standard deviation σ is: $x \sim \mathrm{N}(\mu, \sigma^2)$
  - the Matlab® command: **`randn`**
    generates samples from a standard normal distribution $x \sim \mathrm{N}(0,1)$
- It is defined for $x \in \mathbb{R}$, the set of the real numbers
  - however, the tails are really thin
  - the probability of finding a value that deviates from the mean $> 4/5\sigma$ is very small

Interesting property of the normal distribution

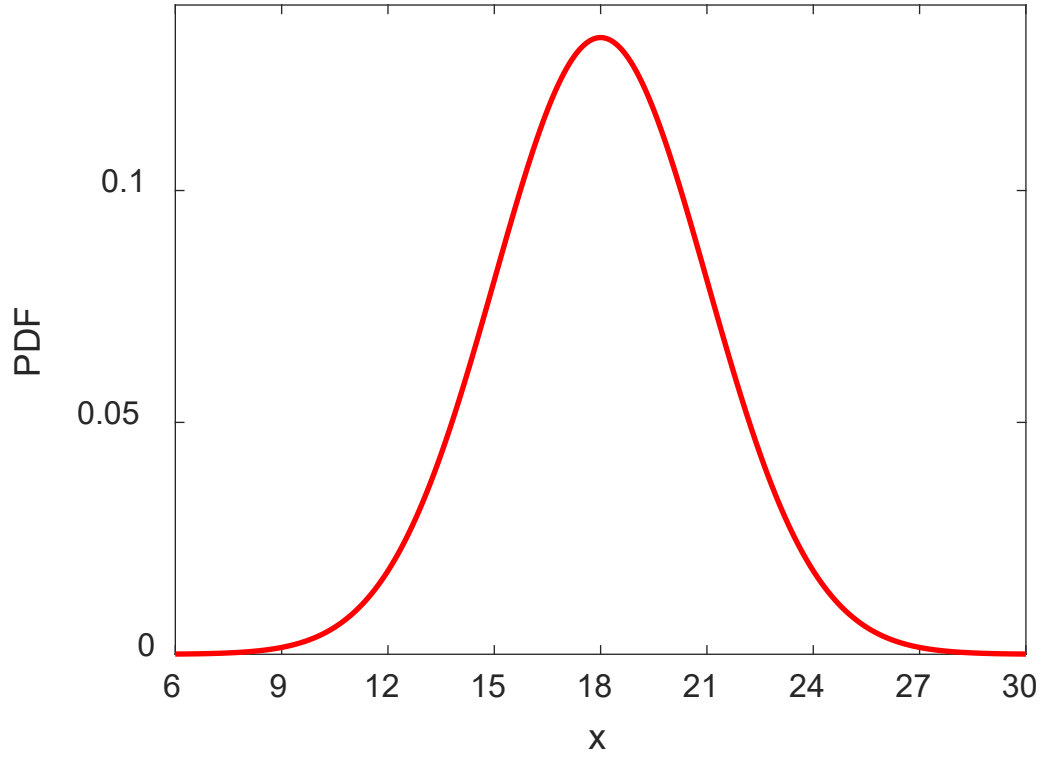
- The normal distribution has a remarkably interesting property for its application to **statistical process control** (SPC):
  - approximately **2/3 of the observations** deviate from μ less than $\mathbf{1\sigma}$ (exactly: 0.9674σ)
  - approximately **95% of the observations** deviate from μ less than $\mathbf{2\sigma}$ (exactly: 1.96σ)
  - **99.73% of the observations** deviates from the mean less than $\mathbf{3\sigma}$





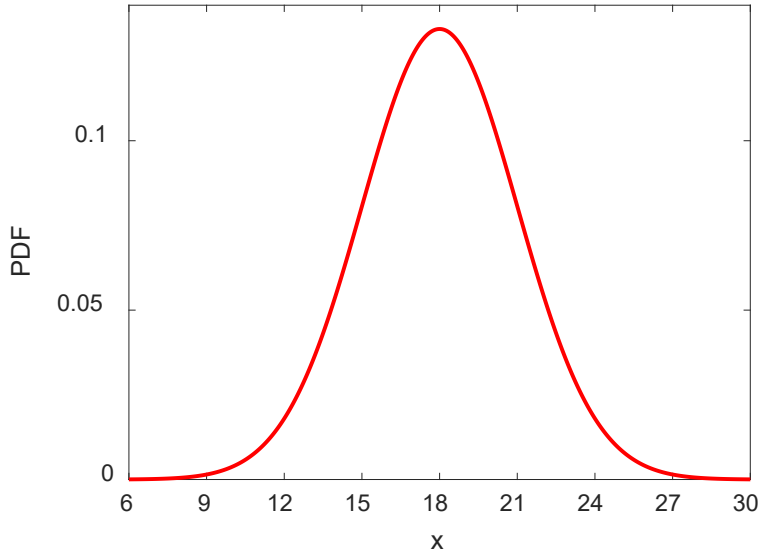


Teamwork



- What is the mean in the example of the figure?
- And the variance?
- What are the mode and the median?
- What are the expected value of the skweness and the kurtosis?
- What is the range of values which retains approximately 95% of the observations of the distribution? And the range which contains approximately 66.67% of the observations?

Teamwork

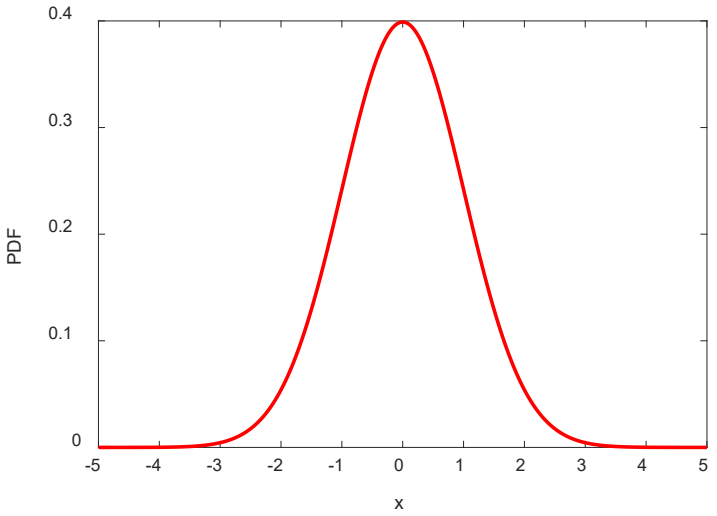


- This is a Gaussian distribution with:
  - $\mu = 18$
  - $\sigma = 3 \rightarrow \sigma^2 = 9$
- Mode and median coincide with the mean μ
- The skeweness is: $\frac{\mu_3}{\sigma^3} = 0$
- The kurtosis is: $\frac{\mu_4}{\sigma^4} = 3$
- The range of values which retains approximately 95% of the observations is **approximately**: $x \in [12,24]$
- The range of values which contains approximately 2/3 of the observations is **approximately**: $x \in [15,21]$

Standard normal distribution

- The **standard normal distribution** is an important special case of normal distribution in which:
  - the mean is zero: $\mu = 0$
  - the standard deviation is one: $\sigma = 1$

- Examples of standard normal distributions in Matlab®:
  - `randn`
  - `normpdf`
  - `pdf`



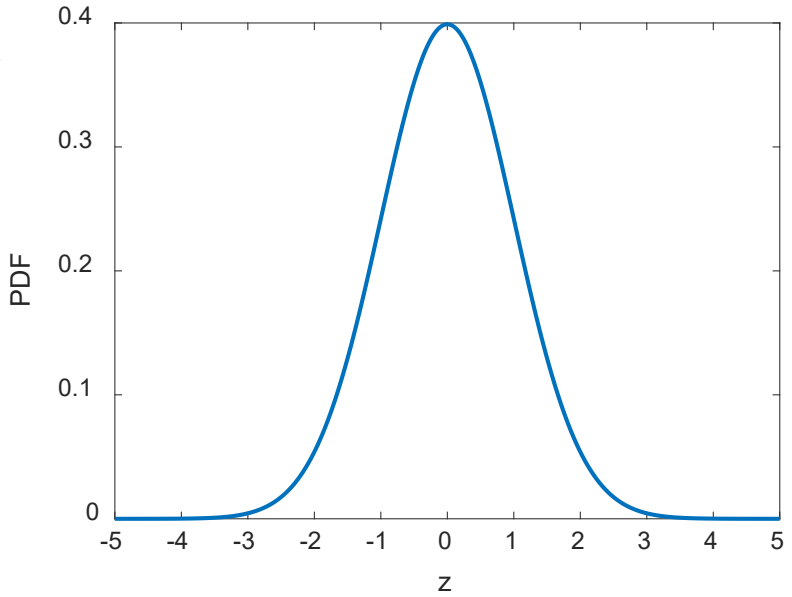
Standard normal distribution

- Define the **standardization** (called also **autoscaling**) of a variable $x \sim \mathrm{N}(\mu, \sigma^2)$ as:
  - **mean centering**: centering to the mean (i.e., subtracting the mean μ)
  - **scaling to unit variance**: dividing it by the standard deviation σ
- If x is a random variable which follows a normal distribution with mean μ and variance σ^2: $x \sim \mathrm{N}(\mu, \sigma^2)$ then:

$$z = \frac{x - \mu}{\sigma}$$

  - where z, the standardized version of x, follows a standard normal distribution:

$$z \sim \mathrm{N}(0,1)$$



Open issue

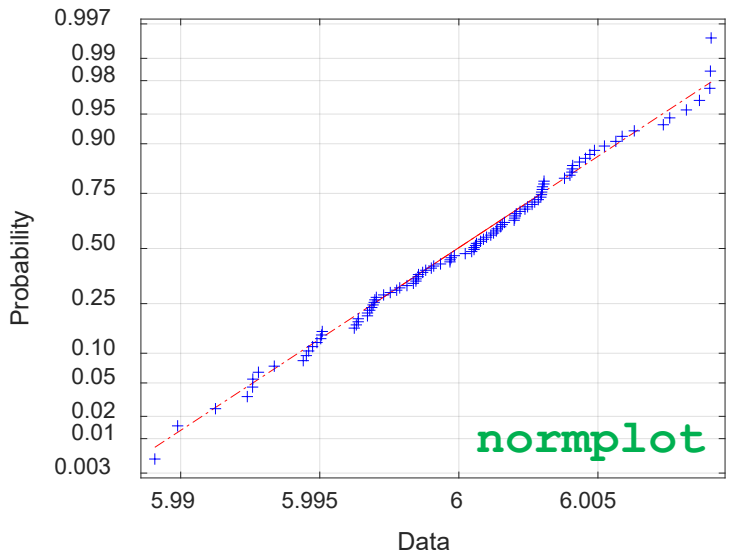
- One problem remain unsolved visualizing boxplots and histograms:
  - when may we consider a distribution to be Gaussian?

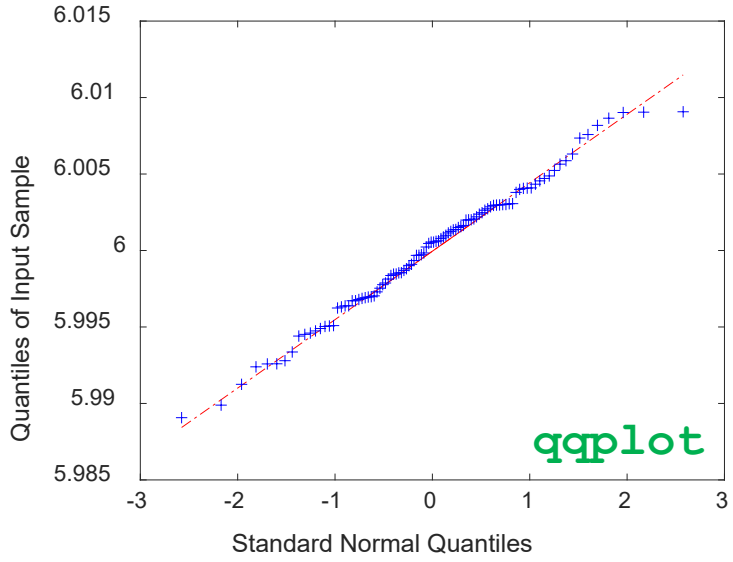
Normality test

- A **normality test** is a statistical test to verify the hypothesis of data normality (Gaussian distribution)
  - verifies if the conjecture that the analyzed set is normally distributed could be acceptable from the statistical point of view
  - it is performed comparing the frequency distribution experimentally observed with the theoretical Gaussian distribution with the same mean and standard deviation
- The p-value is the parameter which allows discriminating if the distribution is normal or not:
  - $p \geq 0.05$: the distribution can be approximated to a **Gaussian** one
  - $p < 0.05$: the distribution cannot be approximated to a Gaussian one
- Most common normality tests:
  - **Anderson–Darling test**
    - reliable for sample numerosity $N > 30$
  - **Ryan–Joiner test**
    - less reliable, utilized when $N < 30$
  - **Kolmogorov–Smirnov test**
    - reliable for $N > 1000$

Graphical normality tests

- **Normal probability plot**:
  - data are displayed against their theoretical normal distribution
  - data (blue crosses) should stay sufficiently close to the theoretical (red) line for normal distributions
- **qq-plot** (quantile-quantile plot): verifies the "normal" behavior of a distribution:
  - quantiles of the sample data x vs. theoretical quantiles values from a normal distribution
  - if the distribution of x is normal, then the plot appears approximately linear





How to compute the probability that a determined event occurs?

Recalling what are probability and PDF…

- From Lesson 2 we remember that…

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:

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

  - 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

Cumulative density function

- The **cumulative density function** $\mathrm{F}(X)$ of a continuous random variable $x \in R$ is the **integral of the probability density function** as follows:

$$\mathrm{F}(X) = \int_{-\infty}^{x_1} \mathrm{f}(x)dx$$

  - non-decreasing function of x
  - used to calculate the **probability** that the variable takes a value less than or equal to x_1:

$$\mathrm{F}(X) = \mathrm{P}[X \leq x_1]$$

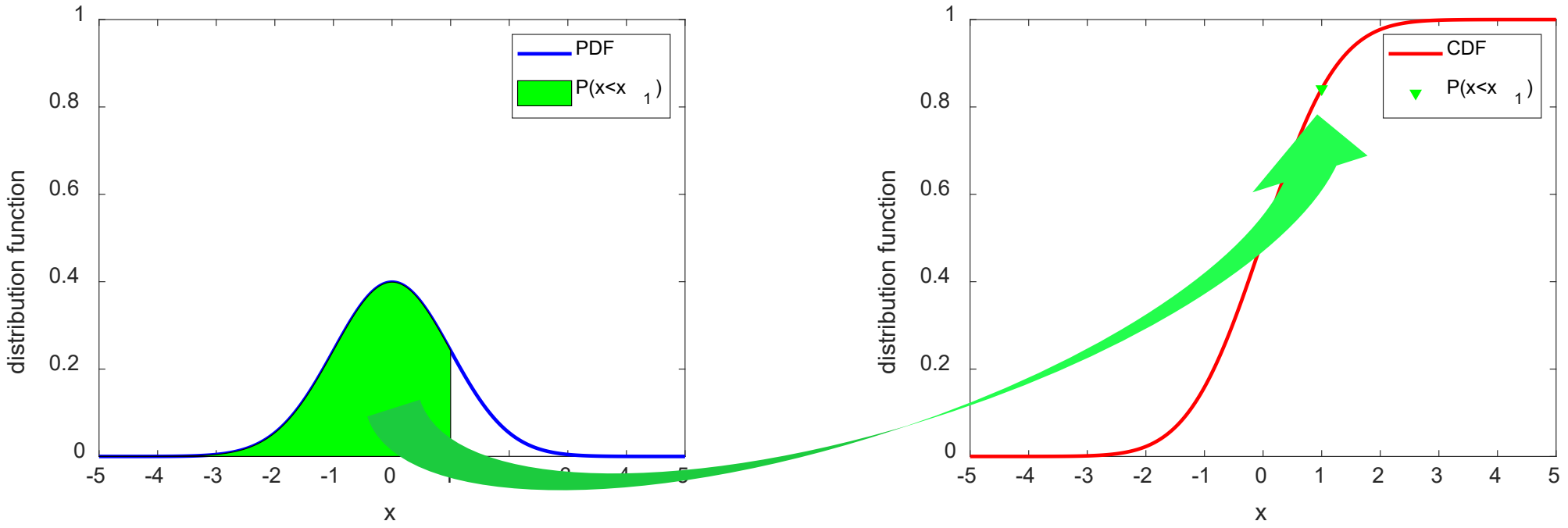
- see the Matlab® command: `normcdf`

Relation among CDF and PDF (1/2)

$$F(x_1) = \int_{-\infty}^{x_1} f(x)dx$$

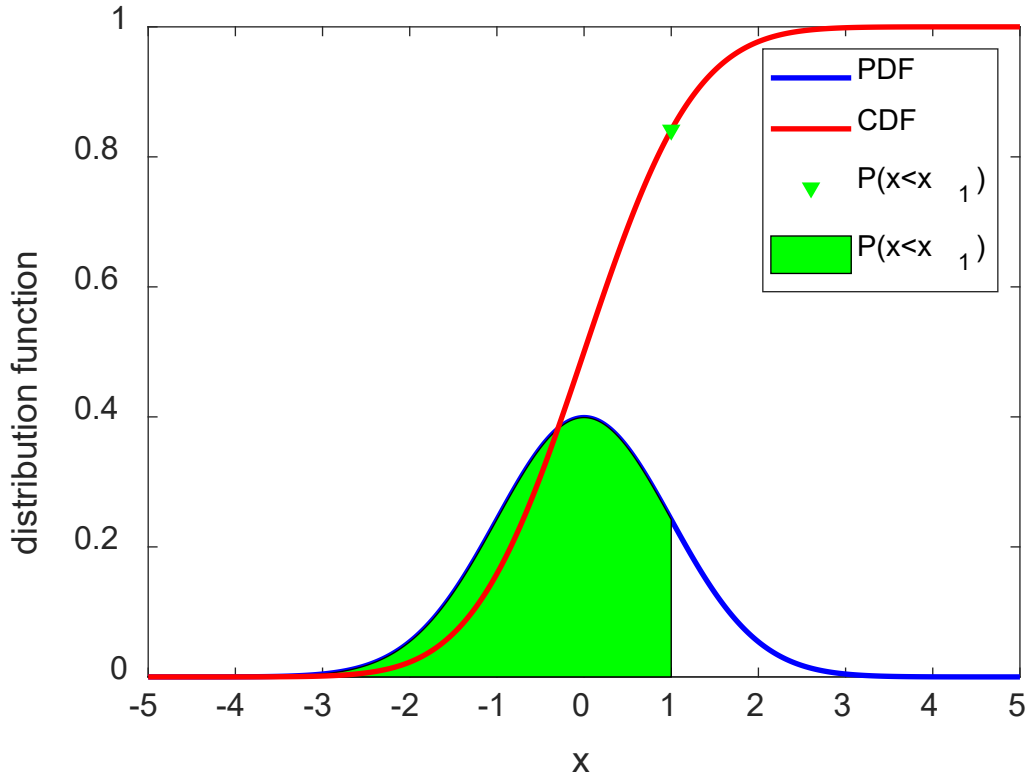
probability density function PDF

cumulative density function CDF



Relation among CDF and PDF (2/2)

- The CDF $\mathrm{F}(x_1)$ in each point x_1 corresponds to the integral $\int_{-\infty}^{x_1} \mathrm{f}(x)dx$ of the PDF $\mathrm{f}(x)$ in the interval $[-\infty\,, x_1]$



Cumulative distribution tables

- **Cumulative distribution tables** are commonly available in the textbooks of statistics and probability theory
- They relate:
  - the value of the **standard normal random variable**
  - the **cumulative probability** that the value occurs

z	0.00	0.01	0.02	0.03	0.04	z
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.0
0.1	0.53983	0.54379	0.54776	0.55172	0.55567	0.1
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.2
0.3	0.61791	0.62172	0.62551	0.62930	0.63307	0.3
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.4
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.5
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.6
0.7	0.75803	0.76115	0.76424	0.76730	0.77035	0.7
0.8	0.78814	0.79103	0.79389	0.79673	0.79954	0.8
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.9
1.0	0.84134	0.84375	0.84613	0.84849	0.85083	1.0
1.1	0.86433	0.86650	0.86864	0.87076	0.87285	1.1
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	1.2
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	1.3
1.4	0.91924	0.92073	0.92219	0.92364	0.92506	1.4
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	1.5
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	1.6
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	1.7
1.8	0.96407	0.96485	0.96562	0.96637	0.96711	1.8
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	1.9
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	2.0

z	0.05	0.06	0.07	0.08	0.09	z
0.0	0.51994	0.52392	0.52790	0.53188	0.53586	0.0
0.1	0.55962	0.56356	0.56749	0.57142	0.57534	0.1
0.2	0.59871	0.60257	0.60642	0.61026	0.61409	0.2
0.3	0.63683	0.64058	0.64431	0.64803	0.65173	0.3
0.4	0.67364	0.67724	0.68082	0.68438	0.68793	0.4
0.5	0.70884	0.71226	0.71566	0.71904	0.72240	0.5
0.6	0.74215	0.74537	0.74857	0.75175	0.75490	0.6
0.7	0.77337	0.77637	0.77935	0.78230	0.78523	0.7
0.8	0.80234	0.80510	0.80785	0.81057	0.81327	0.8
0.9	0.82894	0.83147	0.83397	0.83646	0.83891	0.9
1.0	0.85314	0.85543	0.85769	0.85993	0.86214	1.0
1.1	0.87493	0.87697	0.87900	0.88100	0.88297	1.1
1.2	0.89435	0.89616	0.89796	0.89973	0.90147	1.2
1.3	0.91149	0.91308	0.91465	0.91621	0.91773	1.3
1.4	0.92647	0.92785	0.92922	0.93056	0.93189	1.4
1.5	0.93943	0.90462	0.94179	0.94295	0.94408	1.5
1.6	0.95053	0.95154	0.95254	0.95352	0.95448	1.6
1.7	0.95994	0.96080	0.96164	0.96246	0.96327	1.7
1.8	0.96784	0.96856	0.96926	0.96995	0.97062	1.8
1.9	0.97441	0.97500	0.97558	0.97615	0.97670	1.9
2.0	0.97982	0.98030	0.98077	0.98124	0.98169	2.0

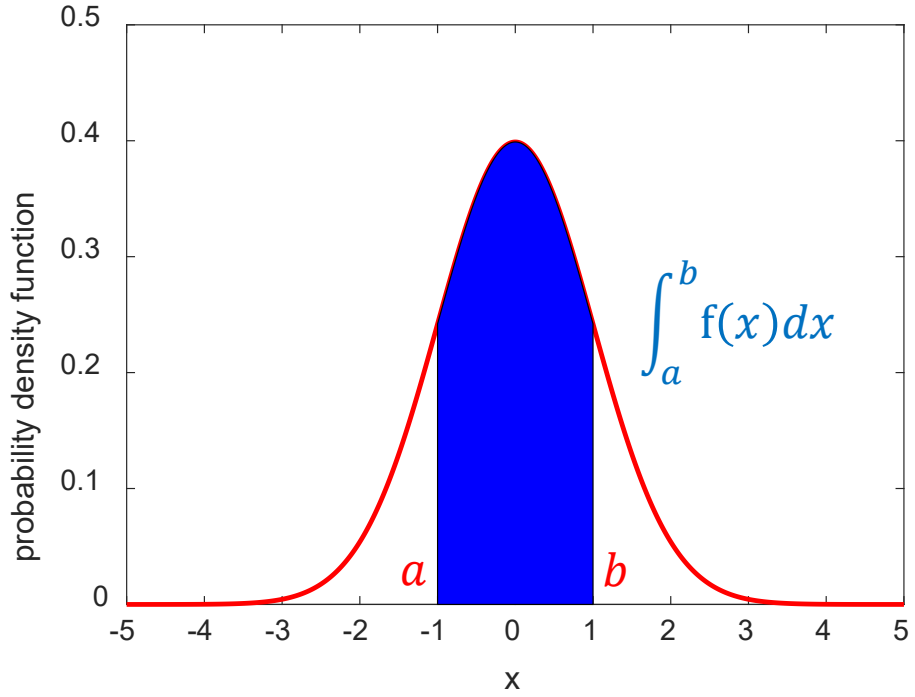
Use of the cumulative distribution tables

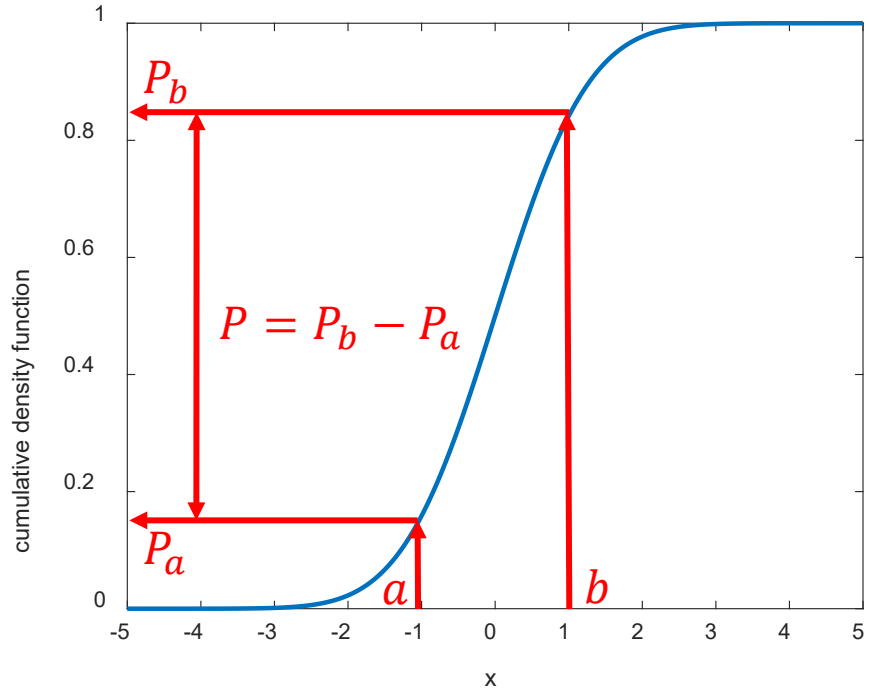
- The table gives the probability of occurrence of a determined value of a standard normal distribution of *z*
  - rows are different values of the standard normal random variable *z*
  - columns correspond to the second decimal of *z*
- Example: the probability for $z \leq 1.47$ is $P = 0.92922$
  - `normcdf(1.47,0,1)`
  - small differences can be found among values in tables and values calculated with software

z	0.05	0.06	0.07	0.08	0.09	z
0.0	0.51994	0.52392	0.52790	0.53188	0.53586	0.0
0.1	0.55962	0.56356	0.56749	0.57142	0.57534	0.1
0.2	0.59871	0.60257	0.60642	0.61026	0.61409	0.2
0.3	0.63683	0.64058	0.64431	0.64803	0.65173	0.3
0.4	0.67364	0.67724	0.68082	0.68438	0.68793	0.4
0.5	0.70884	0.71226	0.71566	0.71904	0.72240	0.5
0.6	0.74215	0.74537	0.74857	0.75175	0.75490	0.6
0.7	0.77337	0.77637	0.77935	0.78230	0.78523	0.7
0.8	0.80234	0.80510	0.80785	0.81057	0.81327	0.8
0.9	0.82894	0.83147	0.83397	0.83646	0.83891	0.9
1.0	0.85314	0.85543	0.85769	0.85993	0.86214	1.0
1.1	0.87493	0.87697	0.87900	0.88100	0.88297	1.1
1.2	0.89435	0.89616	0.89796	0.89973	0.90147	1.2
1.3	0.91149	0.91308	0.91465	0.91621	0.91773	1.3
1.4	0.92647	0.92785	0.92922	0.93056	0.93189	1.4
1.5	0.93943	0.90462	0.94179	0.94295	0.94408	1.5
1.6	0.95053	0.95154	0.95254	0.95352	0.95448	1.6
1.7	0.95994	0.96080	0.96164	0.96246	0.96327	1.7
1.8	0.96784	0.96856	0.96926	0.96995	0.97062	1.8
1.9	0.97441	0.97500	0.97558	0.97615	0.97670	1.9
2.0	0.97982	0.98030	0.98077	0.98124	0.98169	2.0

Cumulative density function and probability (1/4)

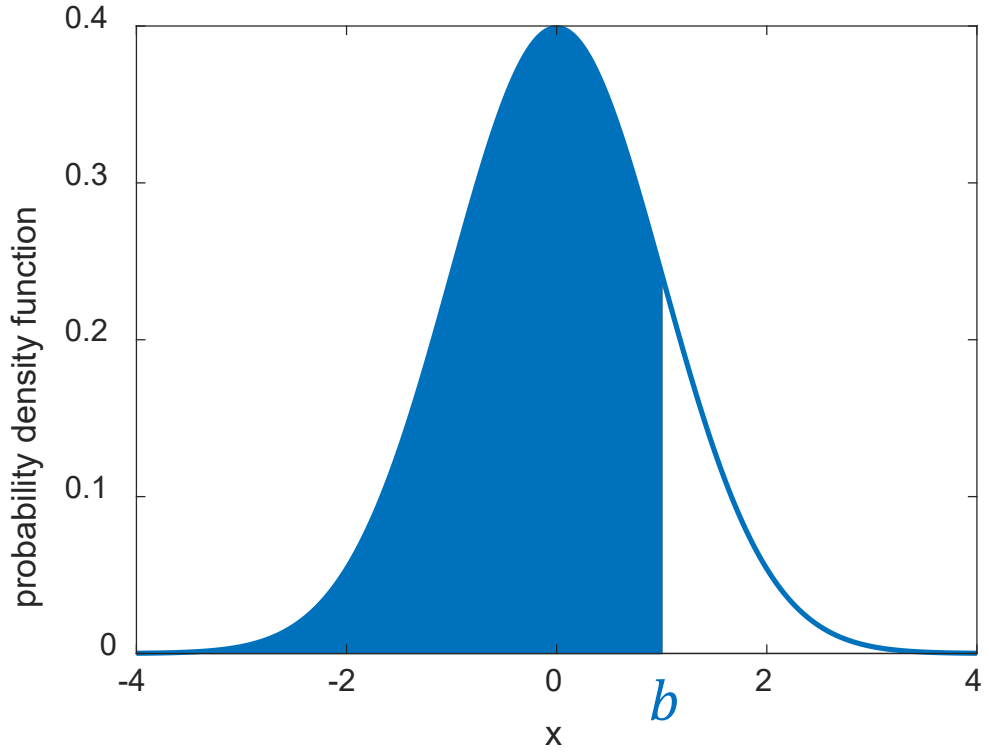
- The CDF can be used to compute what is the probability P of finding values of x within a predetermined interval $[a, b]$, also with $a = -\infty$ or $b = +\infty$

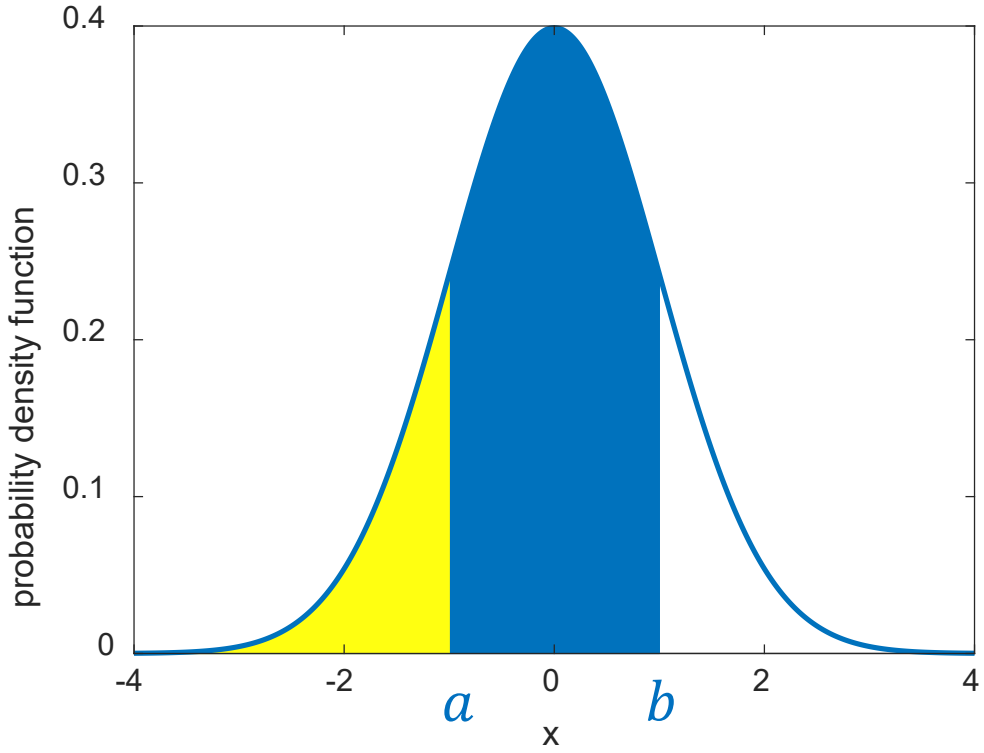




Cumulative density function and probability (2/4)

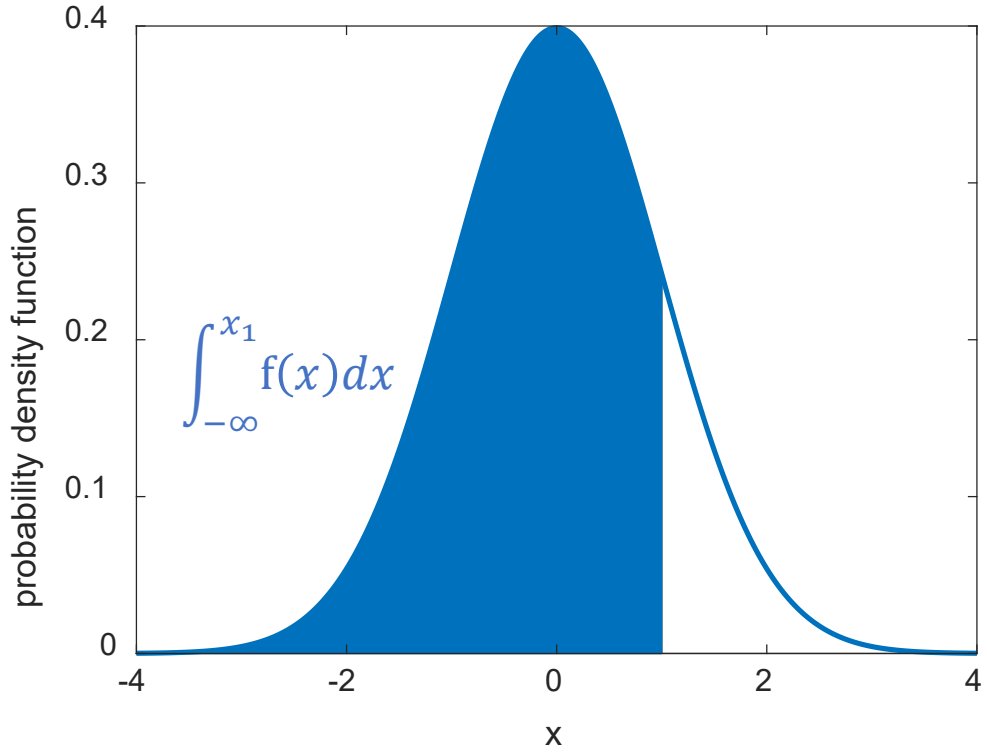
- To calculate the probability that $x \in [a, b]$:
  1. calculate the area under the PDF for $x < b$ (**blue area**): $\int_{-\infty}^{b} f(x)dx$
  2. subtract the area under the PDF for $x < a$ (**yellow area**): $\int_{-\infty}^{a} f(x)dx$

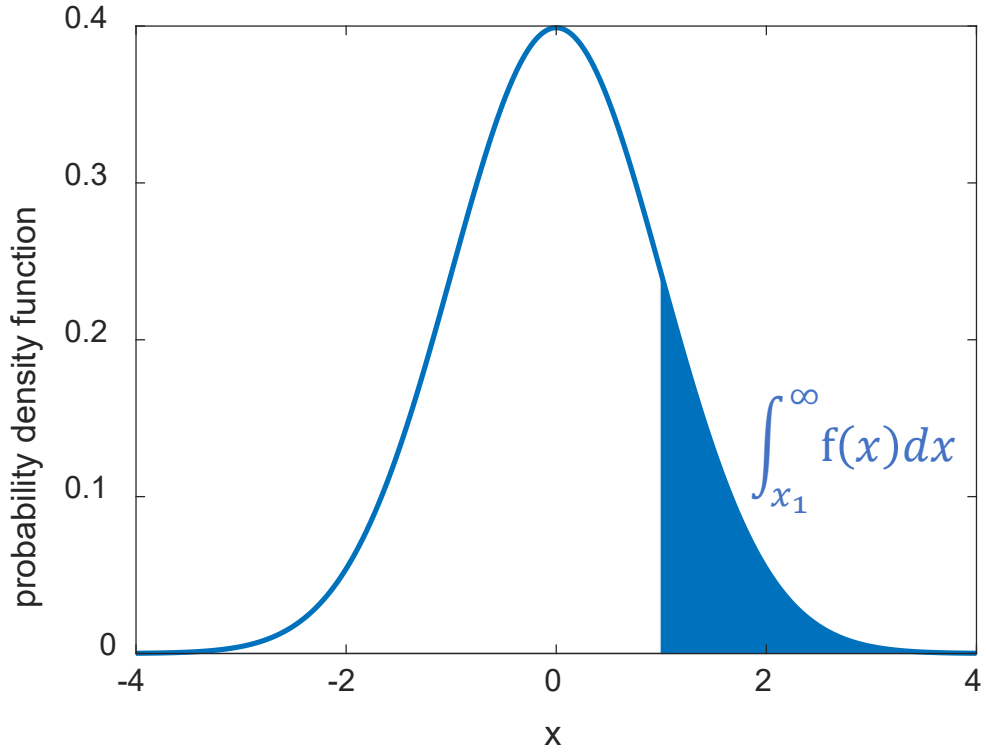




Cumulative density function and probability (3/4)

- To calculate the probability $\mathrm{P}(x > x_1) = \int_{x_1}^{+\infty} \mathrm{f}(x)dx$ that $x > x_1$:
  1. calculate the area under the PDF for $x < x_1$: $\int_{-\infty}^{x_1} \mathrm{f}(x)dx$
  2. subtract $\int_{-\infty}^{x_1} \mathrm{f}(x)dx$ to 1: $\mathrm{P}(x > x_1) = \int_{x_1}^{+\infty} \mathrm{f}(x)dx = 1 - \int_{-\infty}^{x_1} \mathrm{f}(x)dx$

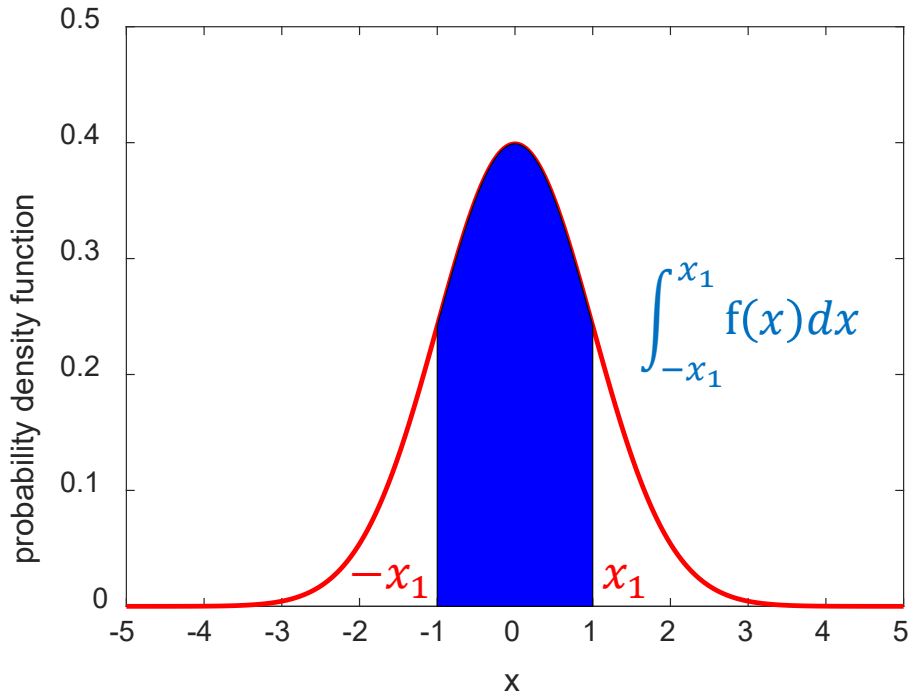


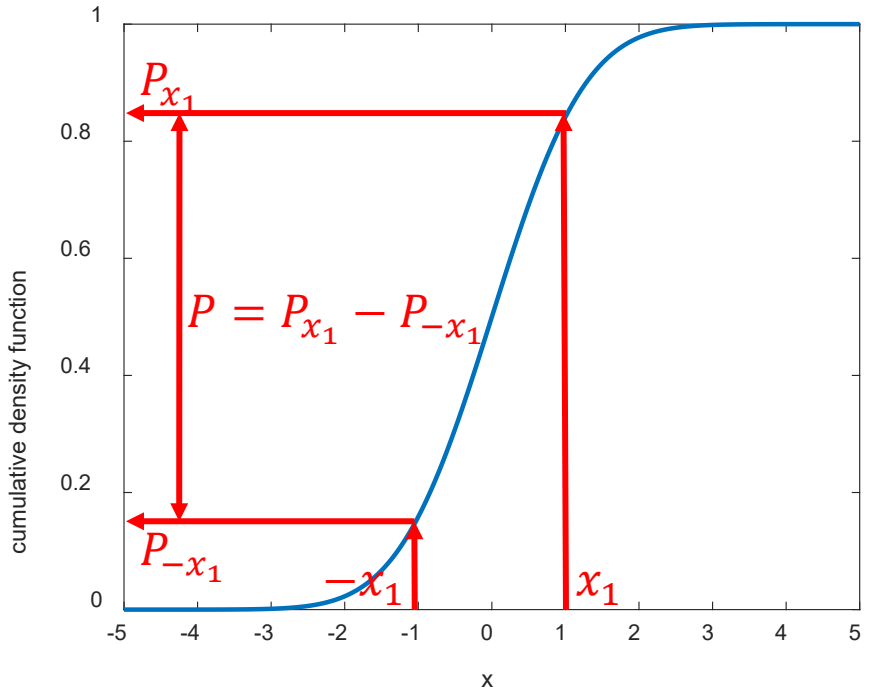


Cumulative density function and probability (4/4)

- We know that the normal distribution is symmetric, so to identify the α% **most extreme values** of the samples one should consider:
  - $\alpha/2$% in the upper tail
  - $\alpha/2$% in the lower tail

  } these two aliquots are the same





How to identify the event that has a predetermined chance of occurring?

Inverse distribution function

- If the CDF is strictly increasing and continuous then a function:

$$\mathrm{F}^{-1}(\mathrm{P}) \text{ with } P \in [0,1]$$

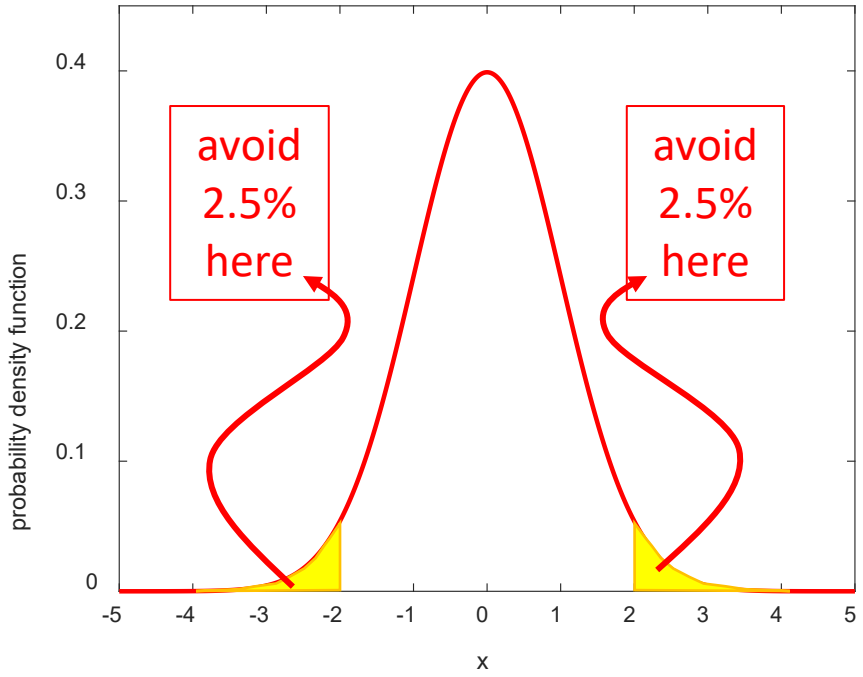
exists and is the unique real number x such that:

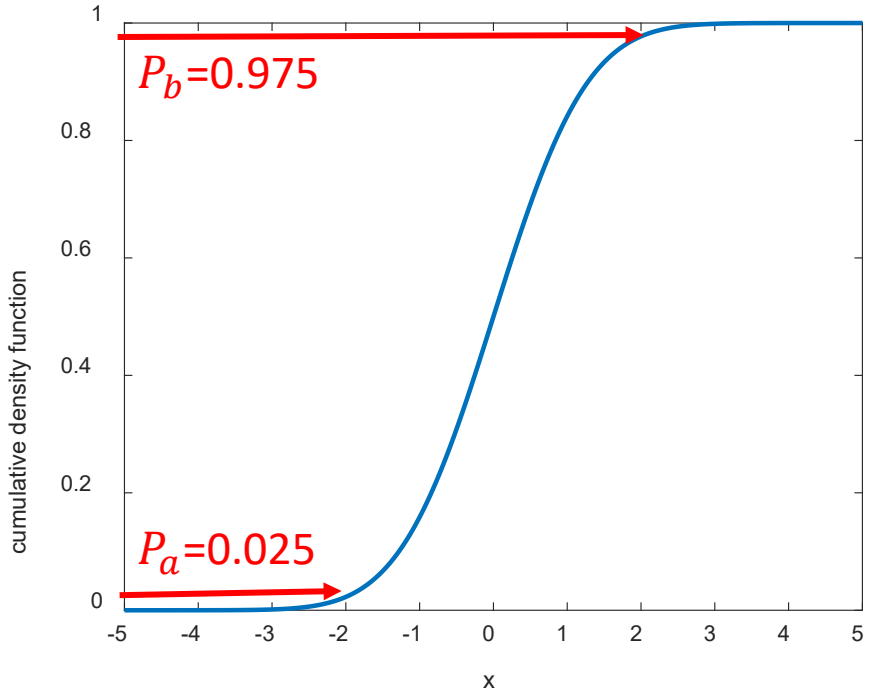
$$\mathrm{F}(x) = P$$

- This is the definition of the **inverse distribution function** IDF or **quantile function**
  - the Matlab® command for this is: `norminv`

Example: IDF (1/2)

- Where are located (i.e., what is z?) the limiting values of a standard normal random variable that guarantee that the 95% percent of the samples are within those limiting values?
  - `norminv(0.975,0,1)` ➡ 1.96
  - `norminv(0.025,0,1)` ➡ -1.96





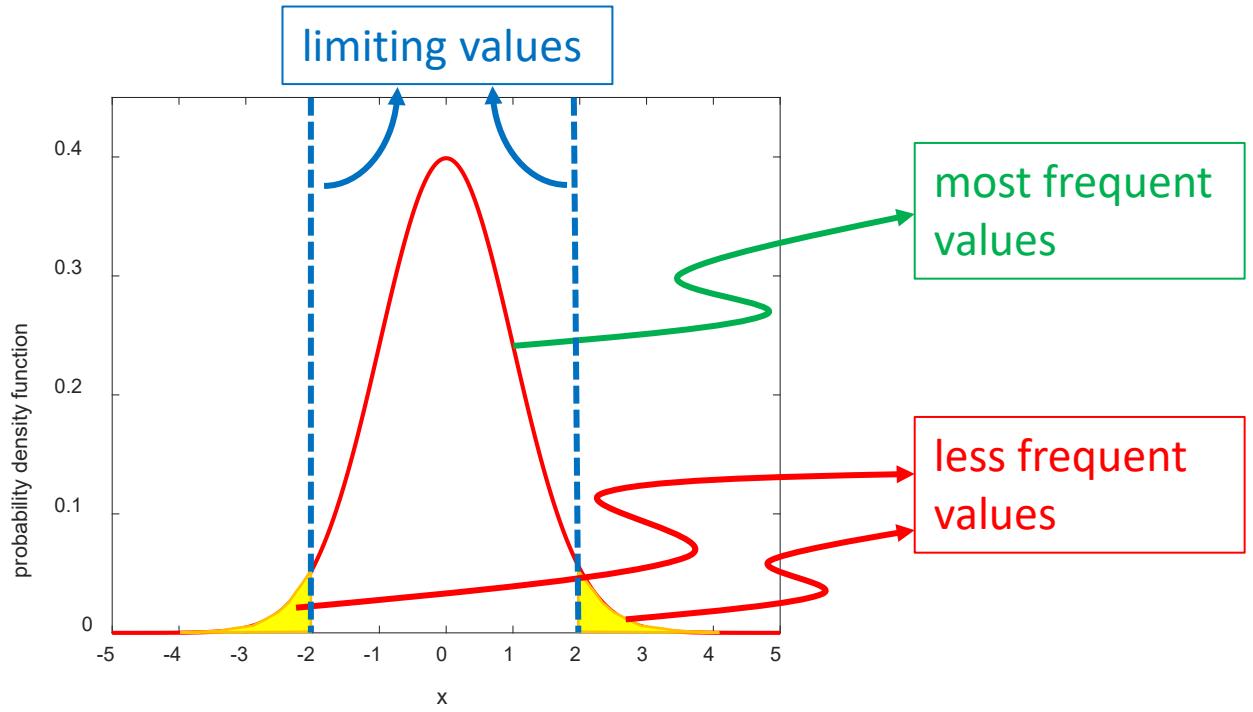
Example: IDF (2/2)

- Where are located (i.e., what is z?) the limiting values of a standard normal random variable that guarantee that the 95% percent of the samples are within that limiting values?
  - cumulative distribution tables can be used to this purpose

z	0.05	0.06	0.07	0.08	0.09	z
0.0	0.51994	0.52392	0.52790	0.53188	0.53586	0.0
0.1	0.55962	0.56356	0.56749	0.57142	0.57534	0.1
0.2	0.59871	0.60257	0.60642	0.61026	0.61409	0.2
0.3	0.63683	0.64058	0.64431	0.64803	0.65173	0.3
0.4	0.67364	0.67724	0.68082	0.68438	0.68793	0.4
0.5	0.70884	0.71226	0.71566	0.71904	0.72240	0.5
0.6	0.74215	0.74537	0.74857	0.75175	0.75490	0.6
0.7	0.77337	0.77637	0.77935	0.78230	0.78523	0.7
0.8	0.80234	0.80510	0.80785	0.81057	0.81327	0.8
0.9	0.82894	0.83147	0.83397	0.83646	0.83891	0.9
1.0	0.85314	0.85543	0.85769	0.85993	0.86214	1.0
1.1	0.87493	0.87697	0.87900	0.88100	0.88297	1.1
1.2	0.89435	0.89616	0.89796	0.89973	0.90147	1.2
1.3	0.91149	0.91308	0.91465	0.91621	0.91773	1.3
1.4	0.92647	0.92785	0.92922	0.93056	0.93189	1.4
1.5	0.93943	0.90462	0.94179	0.94295	0.94408	1.5
1.6	0.95053	0.95154	0.95254	0.95352	0.95448	1.6
1.7	0.95994	0.96080	0.96164	0.96246	0.96327	1.7
1.8	0.96784	0.96856	0.96926	0.96995	0.97062	1.8
1.9	0.97441	0.97500	0.97558	0.97615	0.97670	1.9
2.0	0.97982	0.98030	0.98077	0.98124	0.98169	2.0

Highlight on the IDF

- The IDF helps determining the “limiting values” that identify the most frequent (“normal”) occurrences, also discriminating the least frequent ones
  - this concepts are very important in the **Statistical Process Control** (SPC)



Pen and paper homework

Example: fishing ponds

- You enter a fishing contest. The contest takes place in a pond where the fish lengths have a normal distribution with mean m = 16 inches and standard deviation s = 4 inches
  - **Question 1**: What's the chance of catching a small fish — say, less than 8 inches?
  - **Question 2**: Suppose a prize is offered for any fish over 24 inches. What's the chance of catching a fish at least that size?
  - **Question 3**: What's the chance of catching a fish between 16 and 24 inches?

- see the solution in: `pond_fishing.m`

Example: fishing ponds

z	0.00	0.01	0.02	0.03	0.04	z
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.0
0.1	0.53983	0.54379	0.54776	0.55172	0.55567	0.1
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.2
0.3	0.61791	0.62172	0.62551	0.62930	0.63307	0.3
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.4
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.5
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.6
0.7	0.75803	0.76115	0.76424	0.76730	0.77035	0.7
0.8	0.78814	0.79103	0.79389	0.79673	0.79954	0.8
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.9
1.0	0.84134	0.84375	0.84613	0.84849	0.85083	1.0
1.1	0.86433	0.86650	0.86864	0.87076	0.87285	1.1
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	1.2
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	1.3
1.4	0.91924	0.92073	0.92219	0.92364	0.92506	1.4
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	1.5
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	1.6
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	1.7
1.8	0.96407	0.96485	0.96562	0.96637	0.96711	1.8
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	1.9
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	2.0

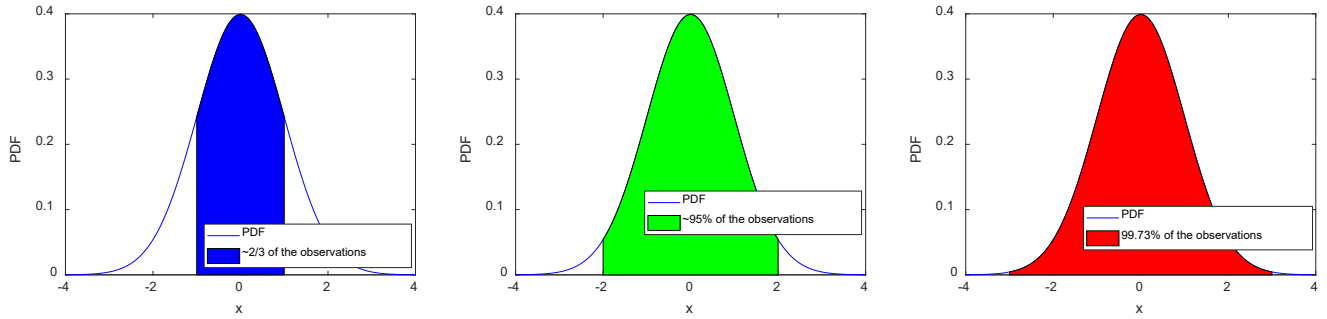
z	0.05	0.06	0.07	0.08	0.09	z
0.0	0.51994	0.52392	0.52790	0.53188	0.53586	0.0
0.1	0.55962	0.56356	0.56749	0.57142	0.57534	0.1
0.2	0.59871	0.60257	0.60642	0.61026	0.61409	0.2
0.3	0.63683	0.64058	0.64431	0.64803	0.65173	0.3
0.4	0.67364	0.67724	0.68082	0.68438	0.68793	0.4
0.5	0.70884	0.71226	0.71566	0.71904	0.72240	0.5
0.6	0.74215	0.74537	0.74857	0.75175	0.75490	0.6
0.7	0.77337	0.77637	0.77935	0.78230	0.78523	0.7
0.8	0.80234	0.80510	0.80785	0.81057	0.81327	0.8
0.9	0.82894	0.83147	0.83397	0.83646	0.83891	0.9
1.0	0.85314	0.85543	0.85769	0.85993	0.86214	1.0
1.1	0.87493	0.87697	0.87900	0.88100	0.88297	1.1
1.2	0.89435	0.89616	0.89796	0.89973	0.90147	1.2
1.3	0.91149	0.91308	0.91465	0.91621	0.91773	1.3
1.4	0.92647	0.92785	0.92922	0.93056	0.93189	1.4
1.5	0.93943	0.90462	0.94179	0.94295	0.94408	1.5
1.6	0.95053	0.95154	0.95254	0.95352	0.95448	1.6
1.7	0.95994	0.96080	0.96164	0.96246	0.96327	1.7
1.8	0.96784	0.96856	0.96926	0.96995	0.97062	1.8
1.9	0.97441	0.97500	0.97558	0.97615	0.97670	1.9
2.0	0.97982	0.98030	0.98077	0.98124	0.98169	2.0

Today's homework

- Verify the «interesting properties» of the normal distributions with Matlab® and Minitab®

Interesting property of the normal distribution

- The normal distribution has a remarkably interesting property for its application to **statistical process control** (SPC):
  - approximately **2/3 of the observations** deviate from μ less than 1σ (exactly: 0.9674σ)
  - approximately **95% of the observations** deviate from μ less than 2σ (exactly: 1.96σ)
  - **99.73% of the observations** deviates from the mean less than 3σ



28/02/2025 Prof. Pierantonio Facco - Machine learning for process engineering 9

Central limit theorem

- **Central limit theorem**:
  - if $w_1, w_2, \ldots, w_N$ is a sequence of N independent and identically distributed random variables with $E(w_i) = \mu$ and $V(w_i) = \sigma^2$ (both finite);
  - if $x = w_1 + w_2 + \cdots + w_N$:

    then

    the limiting form of the distribution of:

    $$z_N = \frac{x - N\mu}{\sqrt{N\sigma^2}} \quad ,$$

    is the **standard normal distribution** if $N \to \infty$
  - the assumption of independence among the w_n is essential, while they can have different distributions
- Stated differently:
  - for any a and b with $a < b$:

$$\lim_{N \to +\infty} P\left(a \leq \frac{w_1 + w_2 + \cdots + w_N - N\mu}{\sigma\sqrt{N}} \leq b \right) = \varphi(b) - \varphi(a)$$

  - where the standard normal distribution function is:

$$\varphi(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} e^{-\frac{1}{2}w^2} dw$$

Consequences of the central limit theorem

- This result states essentially that **the sum of N independent and identically distributed random variables is approximately normally distributed**
  - this approximation is often good for very small N, say $N < 10$, in many cases
  - in other cases, large N is required, say $N > 100$
- Many statistical techniques assume that random variables are normally distributed
  - the central limit theorem is often a justification of **approximate normality**
  - frequently, the **errors in the experiments** arise in an additive manner from several independent sources
    - consequently, the normal distribution becomes a plausible model for the **combined experimental error**

…other consequences of the central limit theorem

- This property allows **using tests based on normal distributions in several cases**
- **Even when a population does not have a normal distribution the mean of repeated samples have (approximately) a normal distribution**
  - stated formally: if $w_1, w_2, \ldots, w_N$ are independent random variables each having the same distributions with expected value μ and standard deviation σ, the sample mean

$$\overline{w} = \frac{1}{N}(w_1 + w_2 + \cdots + w_N)$$

  approximately has a normal distribution with expected value μ and standard deviation $\frac{\sigma}{\sqrt{N}}$ when N is sufficiently large
  - this holds also for samples of $N < 5$
  - verify the central limit theorem in **Matlab®**

… and now let’s practice

Probability in Minitab® and Matlab®

Example on sausage staphylococcus contamination

- **Problem**: a food industry, HappySausage Ltd., produces and commercializes sausages.
  They want to know the results of a wide experimental campaign to evaluate the sausage contamination due to staphylococcus.

- **Available data**:
  - dataset: `sausage.xlsx`

- **Questions**:
  - what is the natural variability of the staphylococcus contamination in their sausages?
  - what is the probability of having an optimum product, where optimum means <200 ufc/g?
  - given the specification of 10 000 ufc/g from the regulatory agencies, what is the probability of having a product which is not acceptable because it is out of specification (with a contamination level higher than the specification limit)?

Descriptive data analysis

- Utilizing graphical and statistical indices for:
  - understanding how data are structures
    - position
    - variability
    - distribution
    - shape
  - visualize samples in a straightforward manner
    - individual value plot
    - histogram
    - boxplot

Preliminari visual analysis of the data

- To have a preliminary and qualitative idea on the data, always start with a **visual preliminary analysis** through:
  - histograms
  - dot-plot
  - etc.
- The visualization step misses:
  - numerical descriptive indices
  - the information about the relation between data variability and specifications

Descriptive data analysis

Matlab® applications

Descriptive data analysis

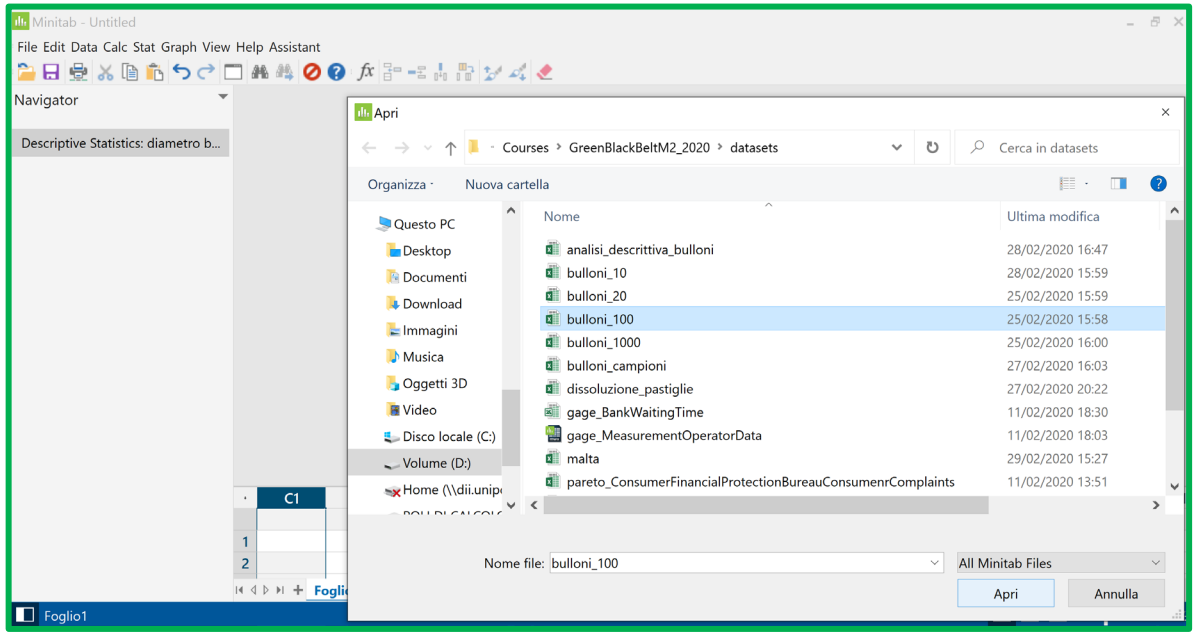
```
1.  no=10000;
2.  x=randn(no,1);

3.  % descriptive statistics
4.  m=mean(x);
5.  s=std(x);
6.  s2=var(x);
7.  ss=realsqrt(var(x));
8.  minx=min(x);
9.  maxx=max(x);
10. sk=skewness(x);
11. k=kurtosis(x);
12. q1=prctile(x,25);
13. q2=prctile(x,50);
14. q3=prctile(x,75);
15. intqr1=q3-q1;
16. intqr2=iqr(x);

17. % plots
18. figure;plot(x);xlabel('observation');ylabel('x')
19. figure;hist(x,40);xlabel('x');ylabel('count')
20. figure;
21. plot(-5:0.001:5,normpdf(-
    5:0.001:5,0,1));xlabel('x');ylabel('PDF')

22. boxplot(x);xlabel('sample');ylabel('x');

23. figure;normplot(x);xlabel('x');ylabel('probability');

24. [ns,ct]=hist(x,40);
25. d=diff(ct);
26. figure;
27. bar(ct,ns./sum(d(1,1).*ns));hold on;box on;
28. plot(-5:0.001:5,normpdf(-
    5:0.001:5,0,1));xlabel('x');ylabel('PDF');hold off

29. figure;
    p1=plot(-5:0.001:5,normpdf(-5:0.001:5,0,1));hold on;
    p2=plot(-5:0.001:5,normcdf(5:0.001:5,0,1));
30. xlabel('x');ylabel('DF');legend('PDF','CDF');hold off

31. figure;
    plot(0:0.001:1,norminv(0:0.001:1,0,1));xlabel('P');ylabel('x');
```

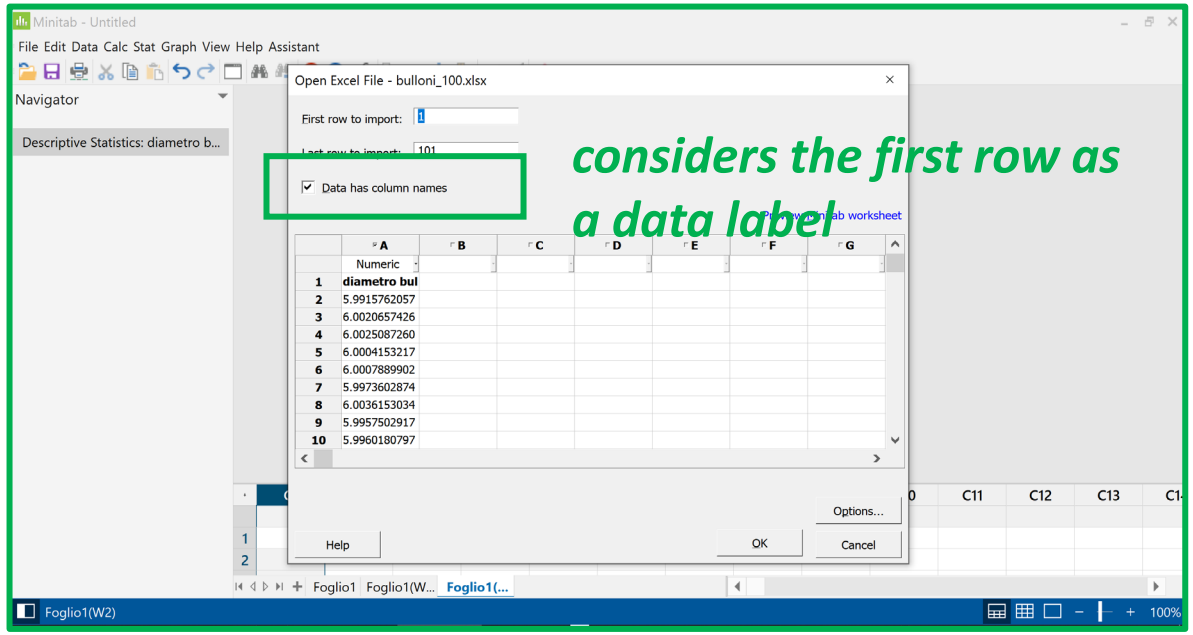
Descriptive data analysis

Minitab® applications

Import file in Minitab®

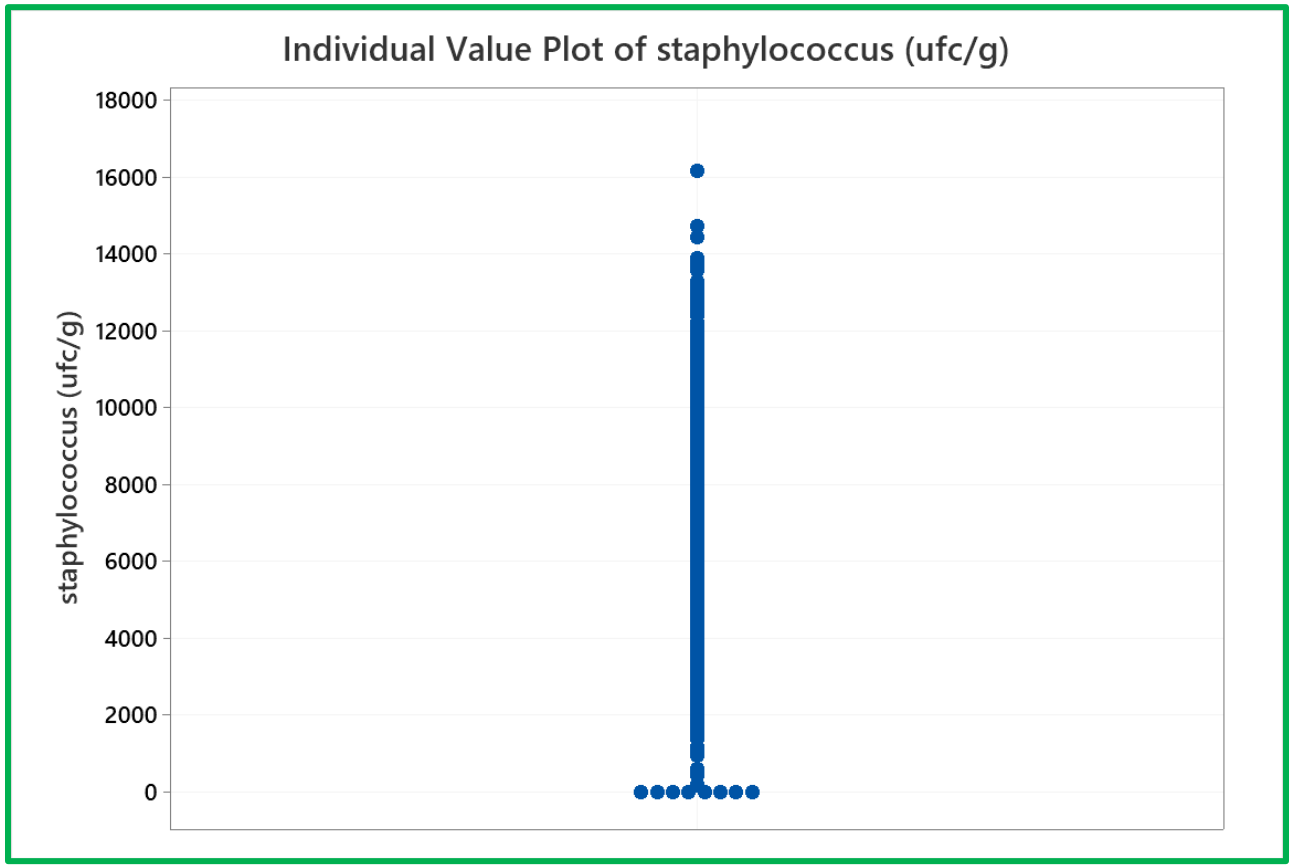
- Select:
  - **File**
  - **Open**
  - select file: **sausage.xlsx**
  - click OK
  - select **Data has column names**
  - click OK





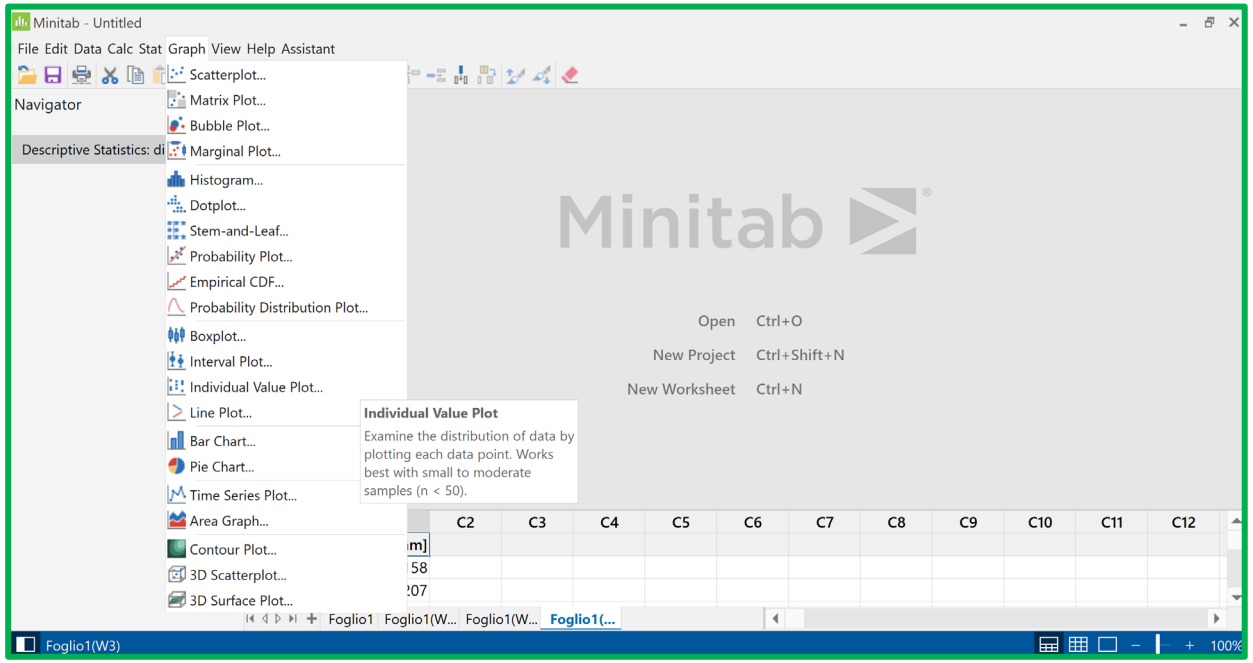
Individual value plot

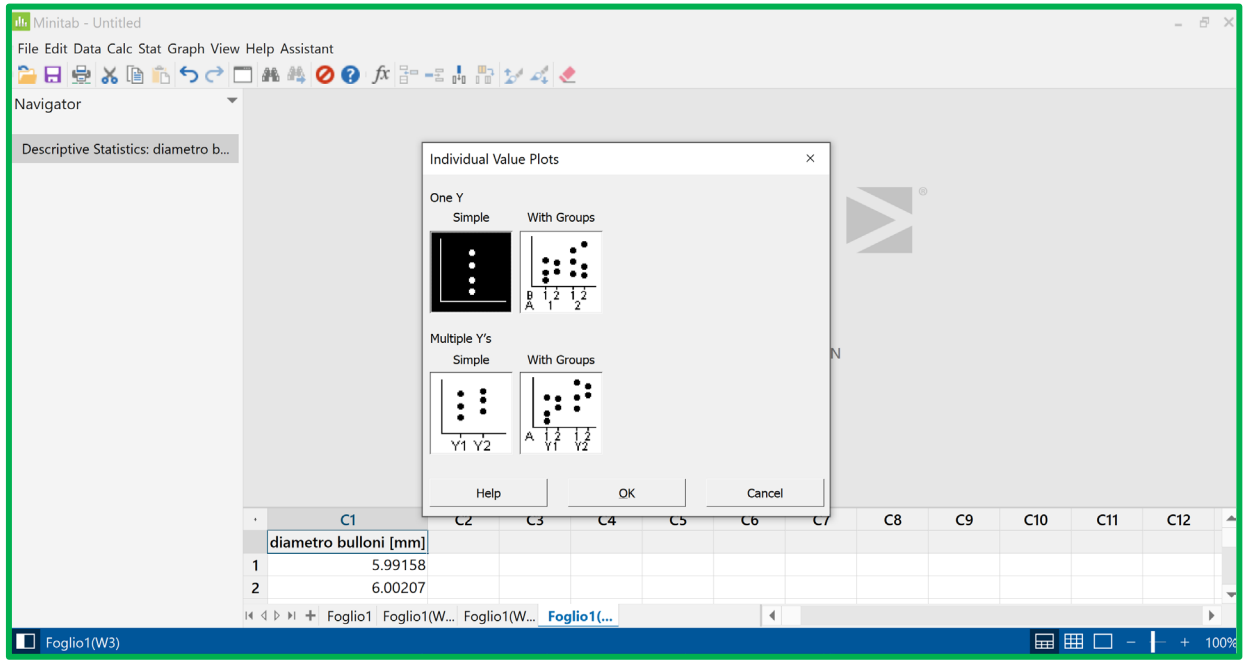
- A preliminary and simple way to display data is using **individual value plot**
  - shows all single values
  - determines in a qualitative way:
    - position
    - variability
    - shape
  - cons:
    - not very effective with large samples



Individual value plot in Minitab® (1/2)

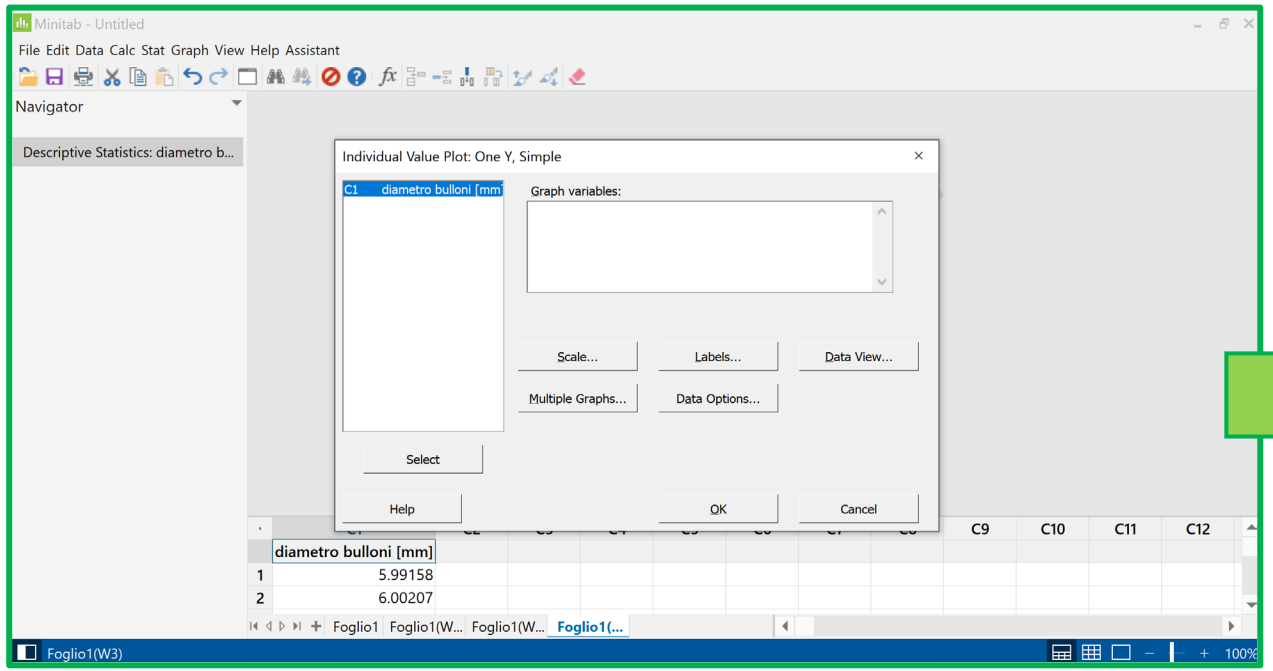
- **Graph**
- **Individual value plot**
- Click OK
- Select: **One Y**
- **Simple**
- Click OK

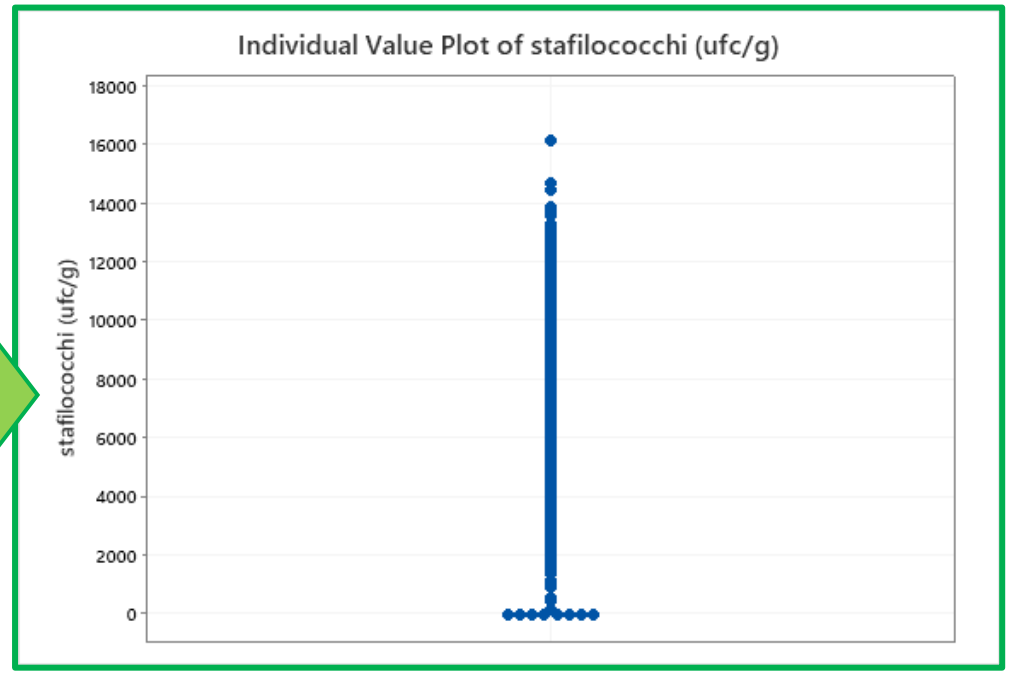




Individual value plot in Minitab® (2/2)

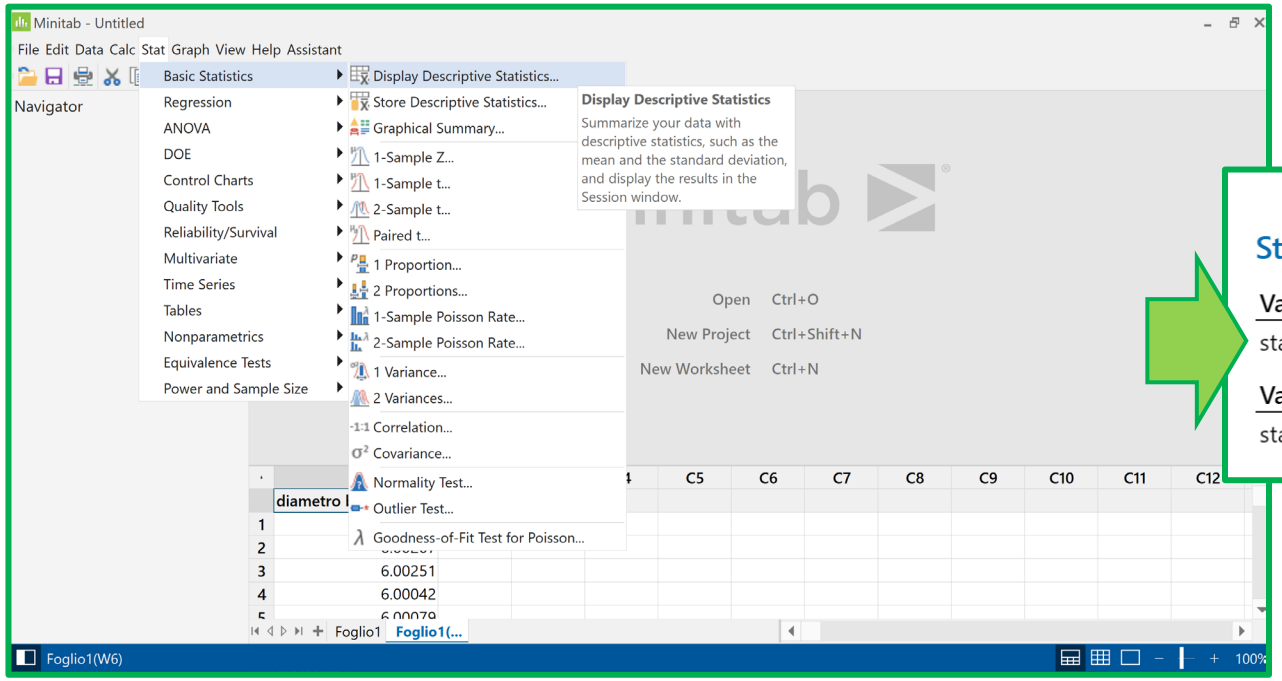
- Select the variable to plot
- Click: **Select**
- Click OK





Descriptive analysis in Minitab®

- The most important statistical indiced are automatically calculated in the following manner:
  - **Stat**
  - **Basic Statistics**
    Display Descriptive Statistics



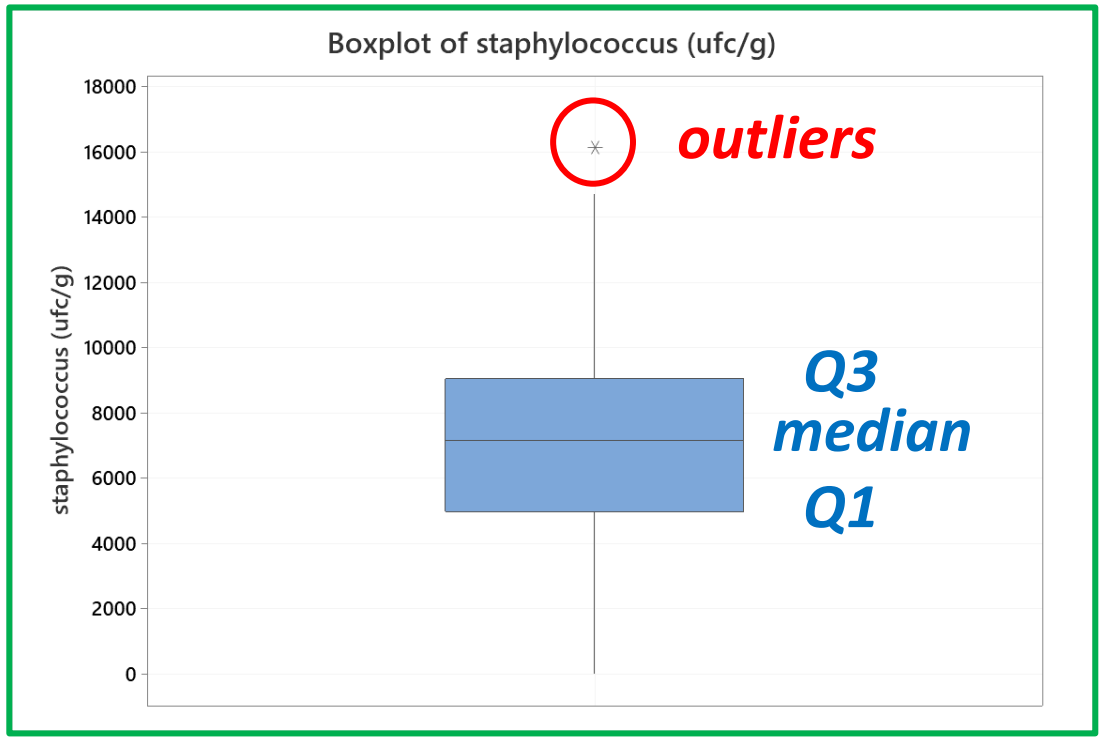
Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
staphylococcus (ufc/g)	1000	0	6992,5	90,6	2865,0	0,0	4970,4	7152,2	9049,7

Variable	Maximum
staphylococcus (ufc/g)	16146,9

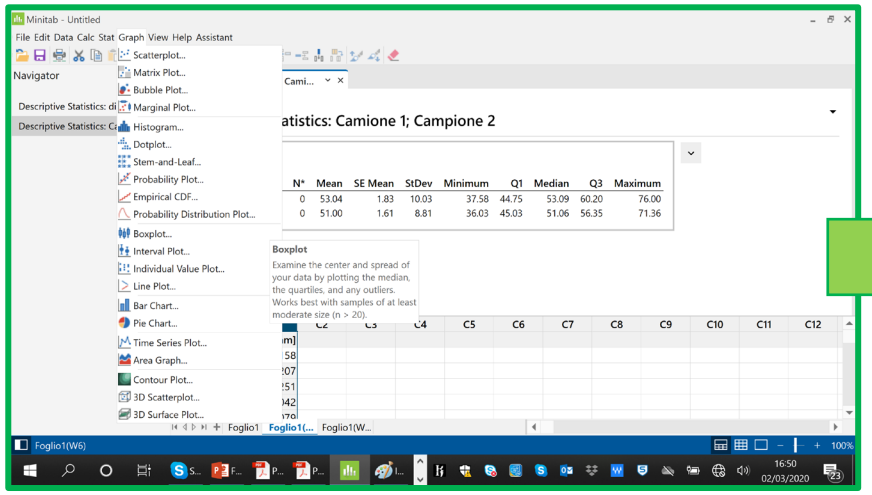
Boxplot

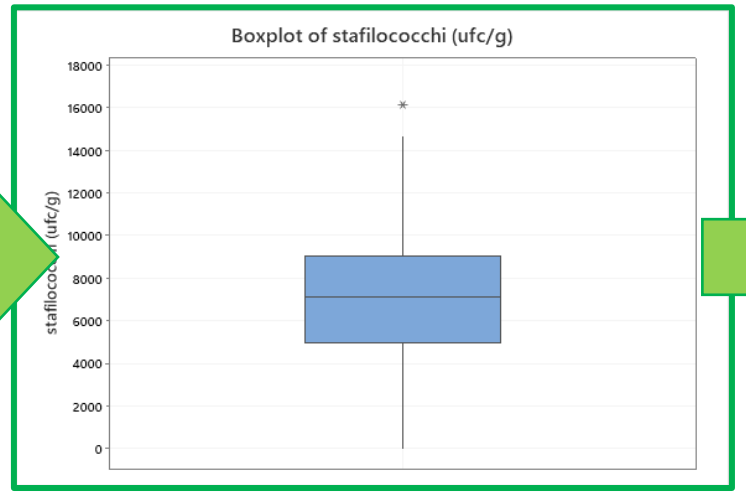
- Boxplot allows comparing distributions in a visual manner:
  - position
  - variability
  - shape
- Cons:
  - not effective for:
    - low numerosity of samples
    - uniformly distributed observations
- Outlier are observations which do not conform to the rest of the population
  - could be of a different population
  - could be a normal event with a very low probability of occurring

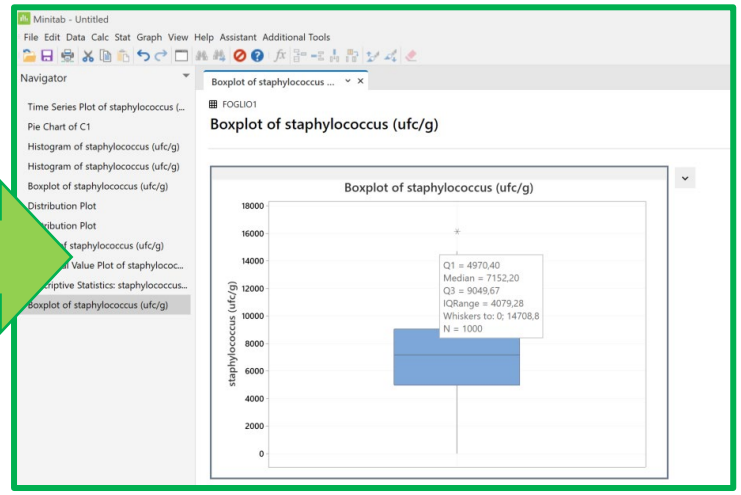


Boxplot in Minitab®

- **Graph**
- **Boxplot**
- Click OK
- Select the variable
- Click OK
- Important statistical indices can be visualized





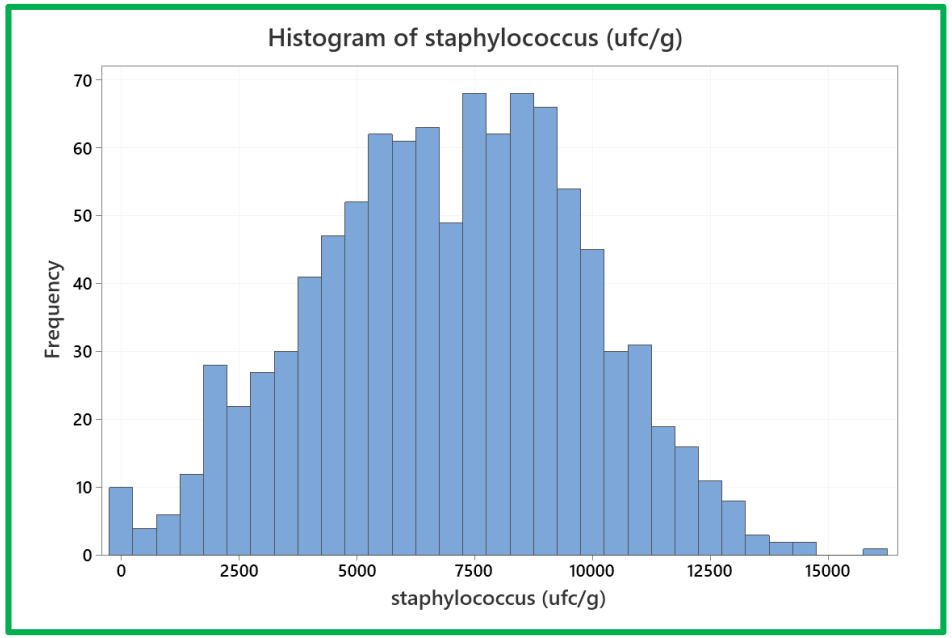


Histogram

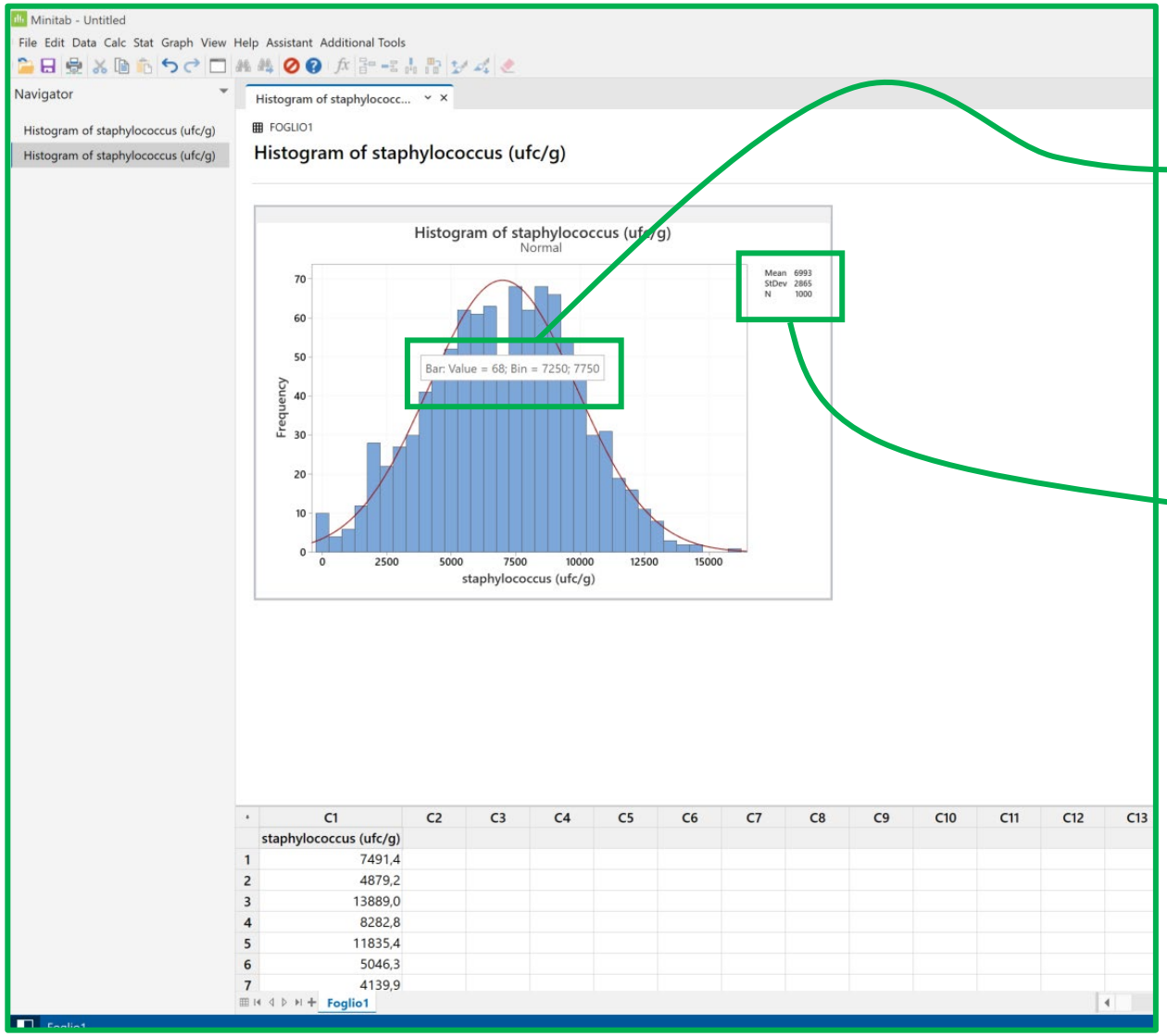
- The histogram is a simple bar plot of the distribution shape in terms of:
  - mode
  - variability (through the range)
  - shape of the probability density function
- It is built in the following manner:
  - define an appropriate number of classes (bins)
    - two simple rules to determine the number of classes C is:

$$C = \sqrt{N}$$
$$C = 1 + 3.332\log(N)$$

  - calculate the class width A:
    - $A = \frac{R}{C}$ dividing the range R by the number of classes C
  - identify the class limits
  - count the number of observations within each interval
- In Minitab®:
  - **`Graph`**
  - **`Histogram`**
  - **`Simple`** or **`With fit`**
  - etc.



Details on histograms

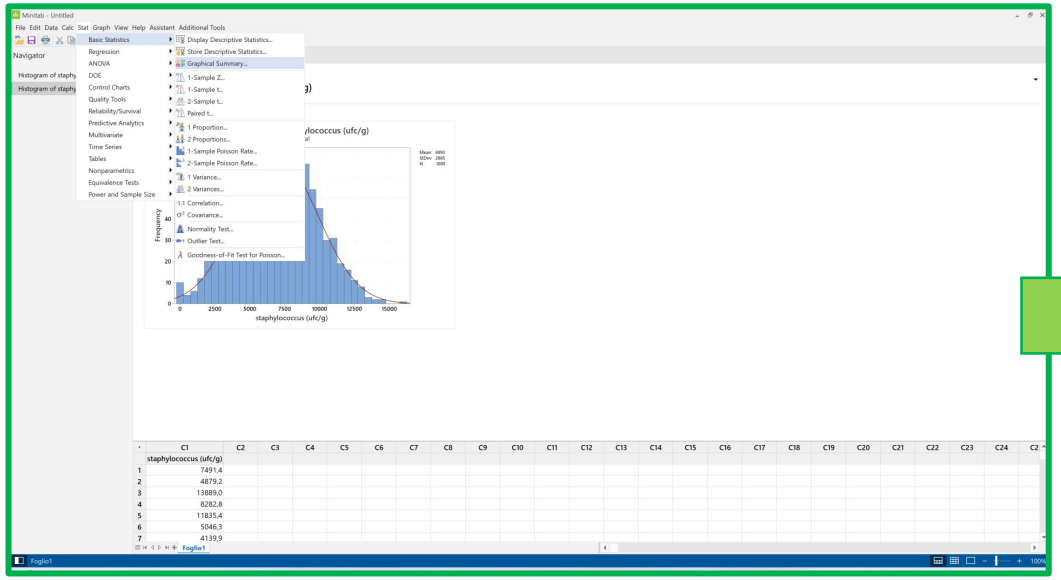


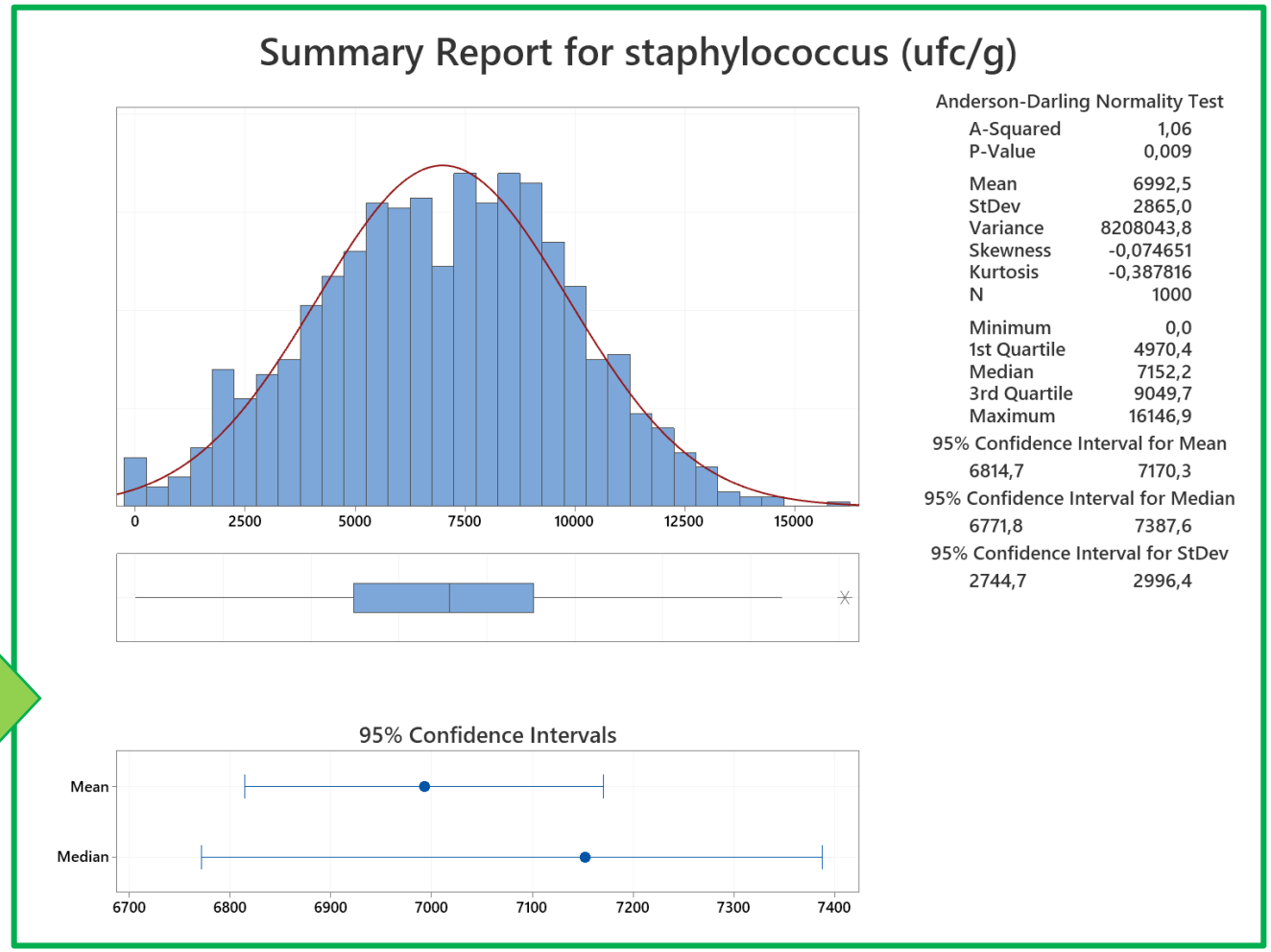
- Information on the **bins** limits and count can be seen directly on the plot

- In histogram with fit **sample numerosity, mean and standard deviation** are calculated, and a Gaussian distribution superimposed to the bar plot

Graphical summary

- To have an effective summary of the descriptive statistics a graphical summary can be assessed in Minitab®:
  - **Stats**
  - **Basic Statistics**
    Graphical Summary
  - select dataset





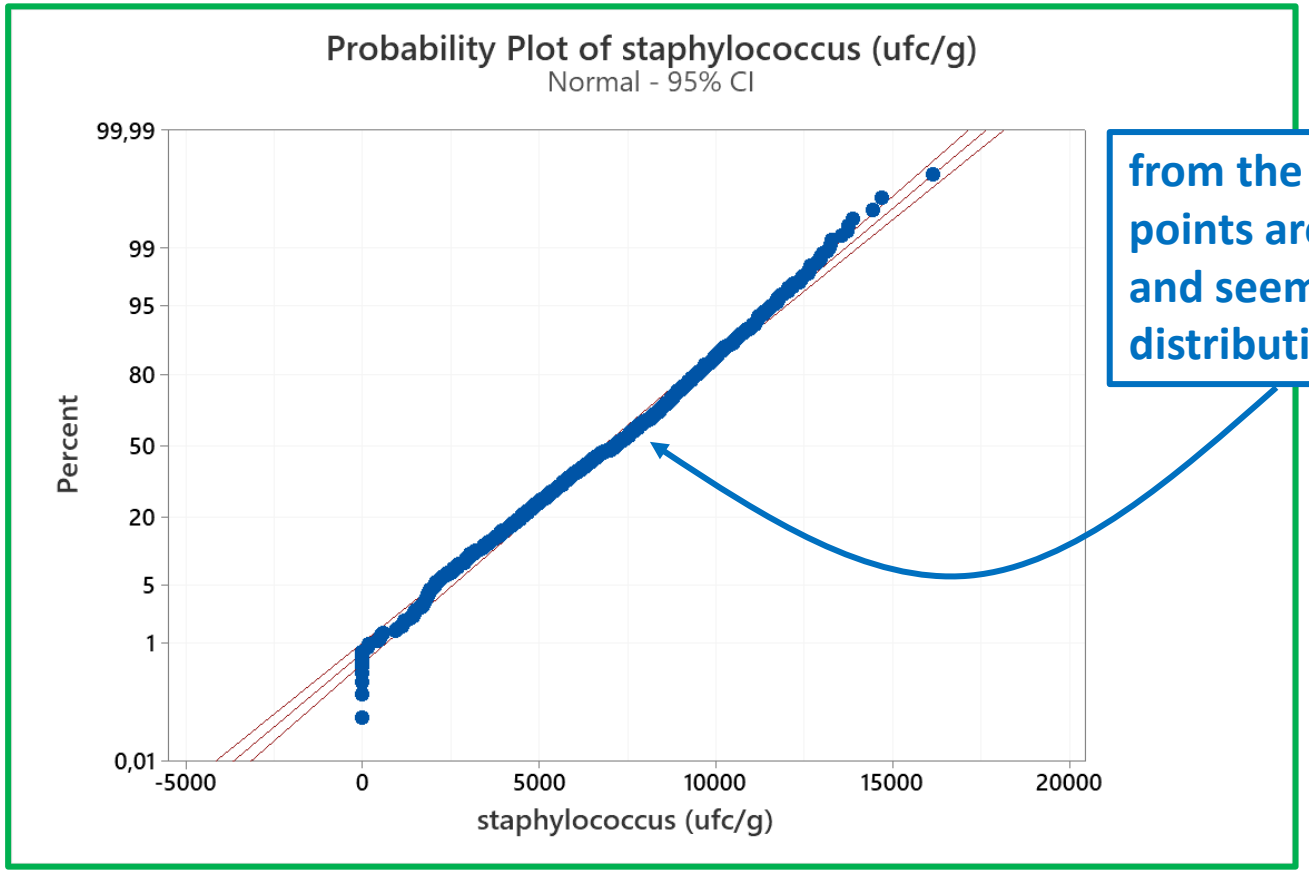
Open issue

- One problem remain unsolved visualizing boxplots and histograms: may we consider a distribution to be Gaussian?

Normality test for single variables

Visual normality test through normality plot in Minitab®

- **Graph**
- **Probability plot**
- **Single**
- Select variable
- Click OK

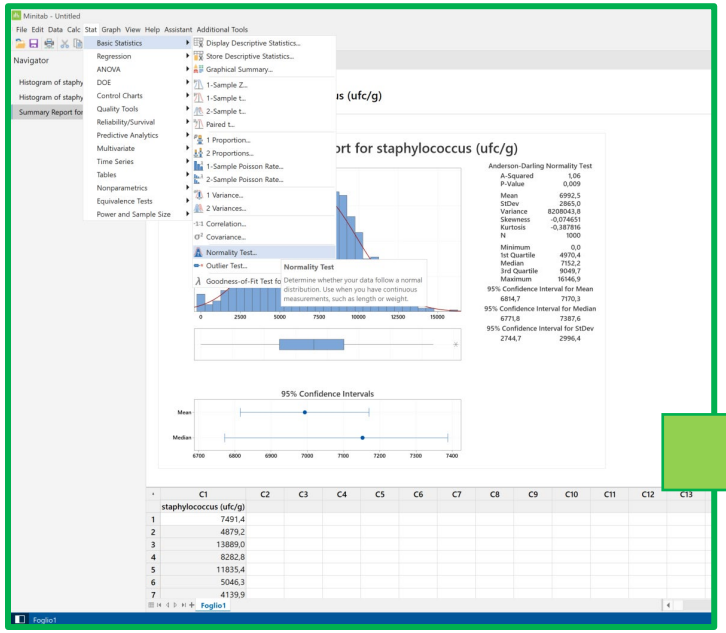


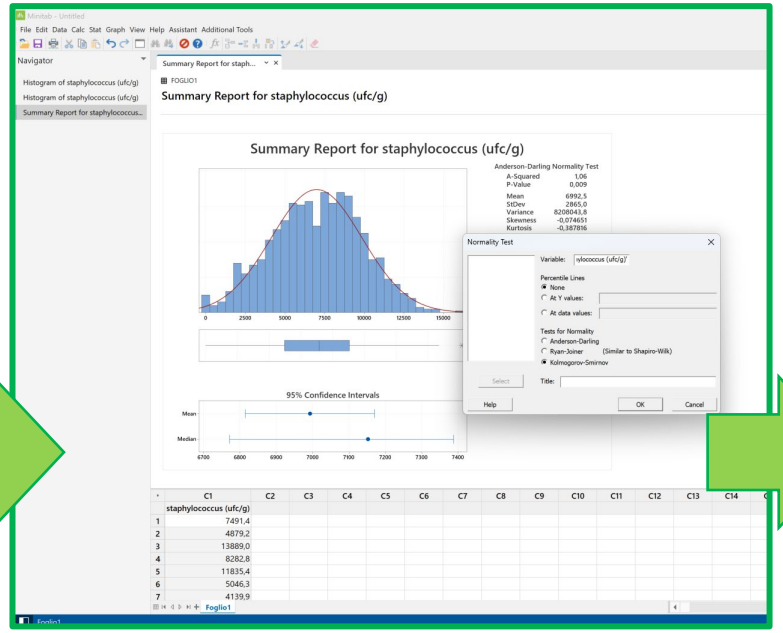
from the visual point of view the points are close enough to the line and seem to identify a Gaussian distribution

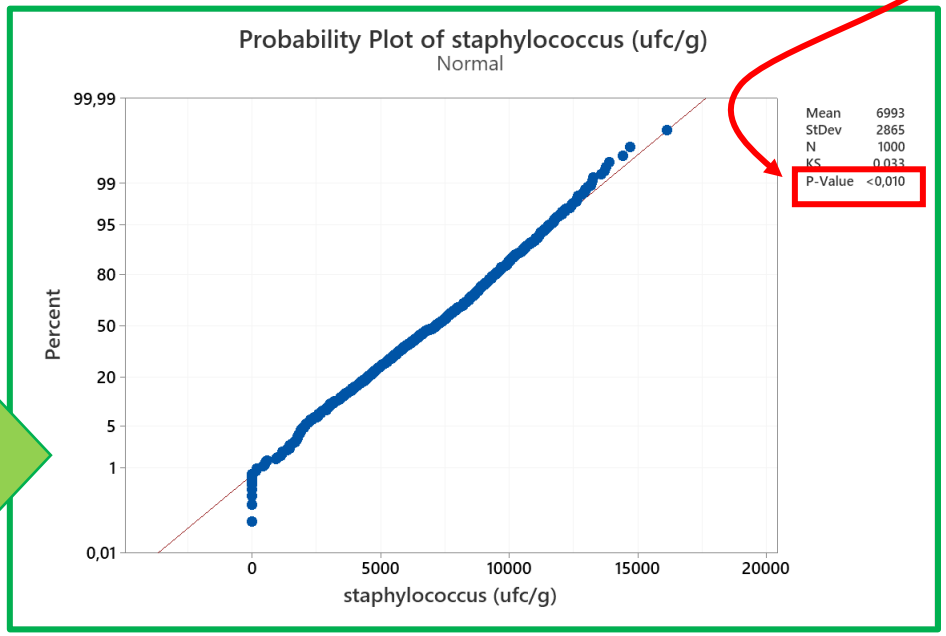
Formal normality test in Minitab®

- **Stat**
- **Basic Statistics**
- **Normality Test**
- Select variable
- **Kolmogorov-Smirnov**
- Click OK

p<0.05 means that the distribution of the observations of the Staphylococcus case study cannot be approximated with a Gaussian distribution

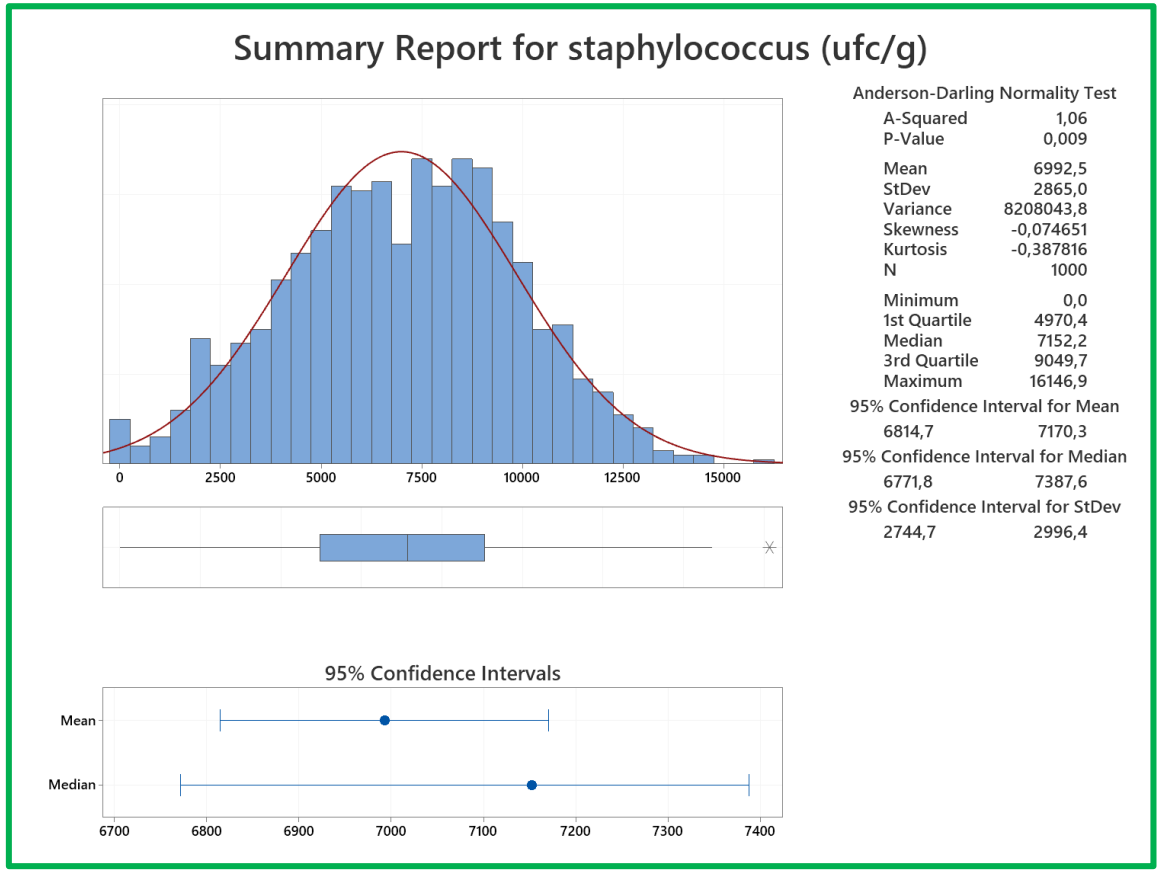






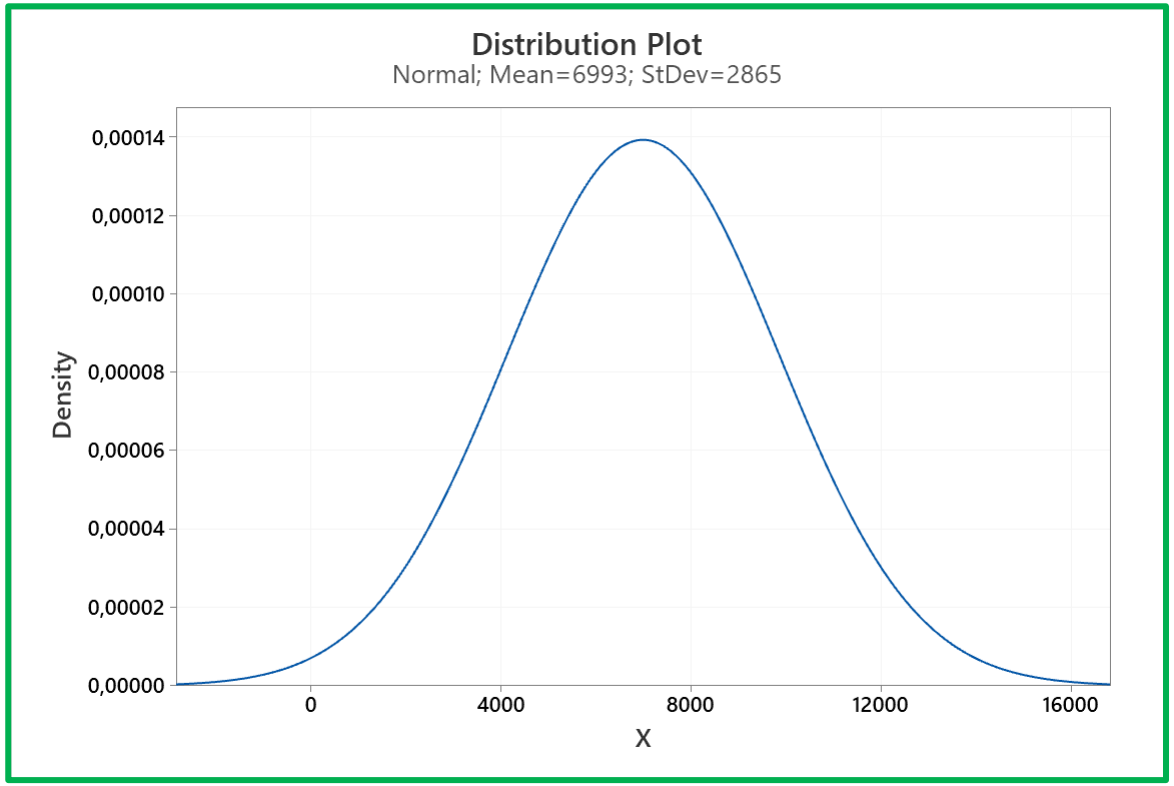
Exploratory analisis

- Descriptive statistics aid to summarize data features
  - we already know how to calculate descriptive indices in the graphical summary



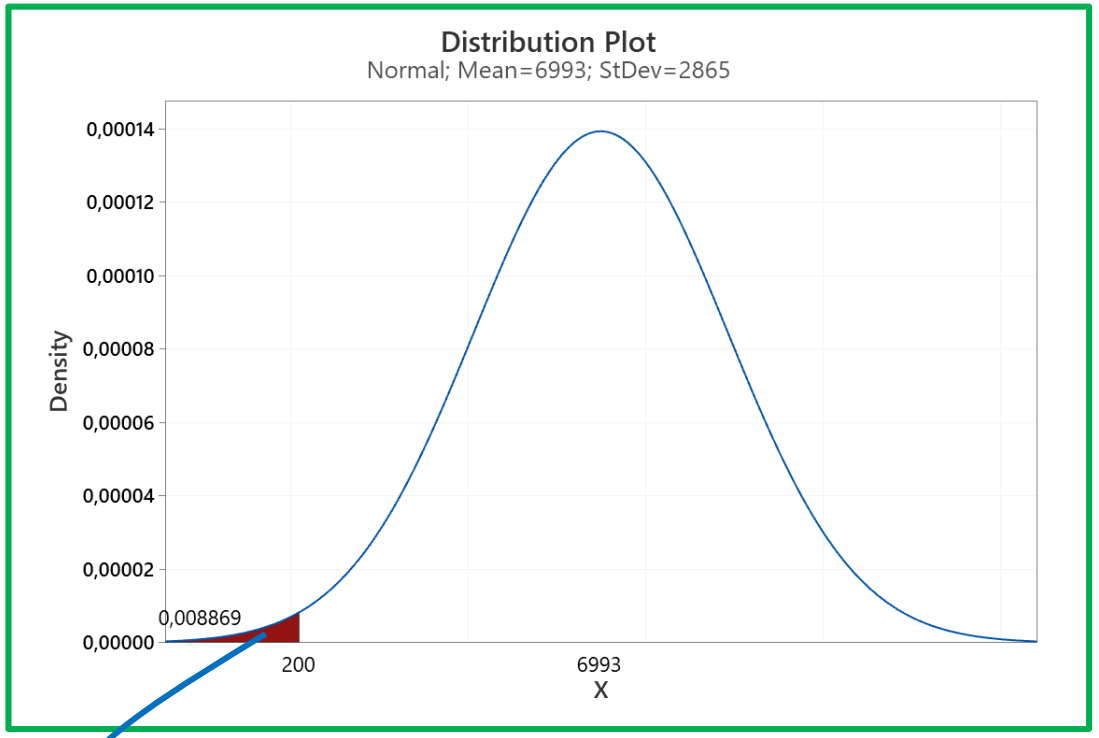
Gaussian normal PDF for the staphylococcus case

- A Gaussian normal distribution is built on the sample mean and standard deviation
  - the model identifies the natural variability of the contamination



Probability of optimal product

- The probability of obtaining an optima product (<200 ufc/g) is lower than **0.9%**
  - **however, few sausages are optimal**
- In Minitab®:
  - `Graph`
  - `Probability distribution plot`
  - select `View probability`
  - click OK
  - in the `Distribution` tab:
    - select Normal
    - input the mean ands standard deviation
  - in the `Shaded Area` tab:
    - select X value
    - select Right tail
    - input the X value: `200`
  - click OK
- In Matlab®:
  - `mx=mean(sausages);`
  - `sx=std(sausages);`
  - `normcdf(200,mx,sx)`

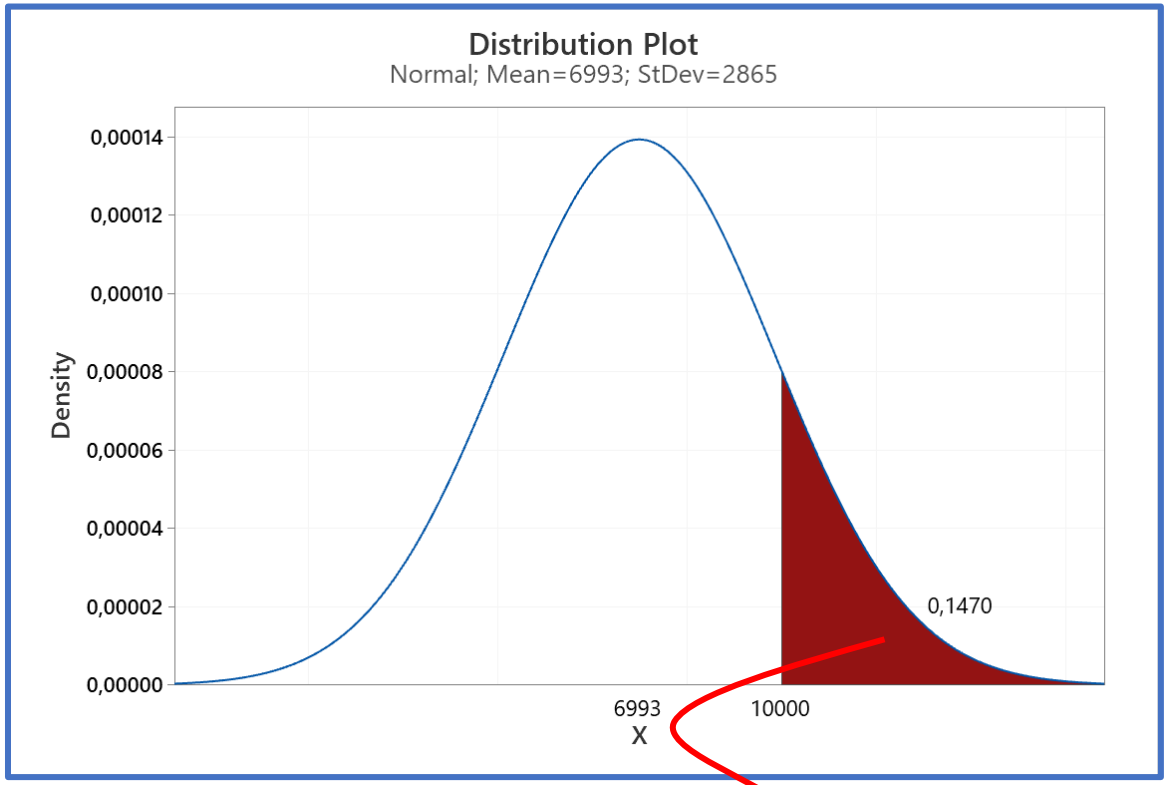


in the dataset it can be verified that 10 sausages out of 1000 are optimal!

Probability of out-of-spec

- The probability of out-of-spec sausages (>10 000 ufc/g) is almost **15%**
  - **this is a criticality because a large part of the produced sausages are contaminated to dangerous levels**

- In Matlab®:
  - `mx=mean(sausages);`
  - `sx=std(sausages);`
  - `1-normcdf(10000,mx,sx)`



in the dataset it can be verified that 145 sausages out of 1000 are out of specification!

Example in fish industry

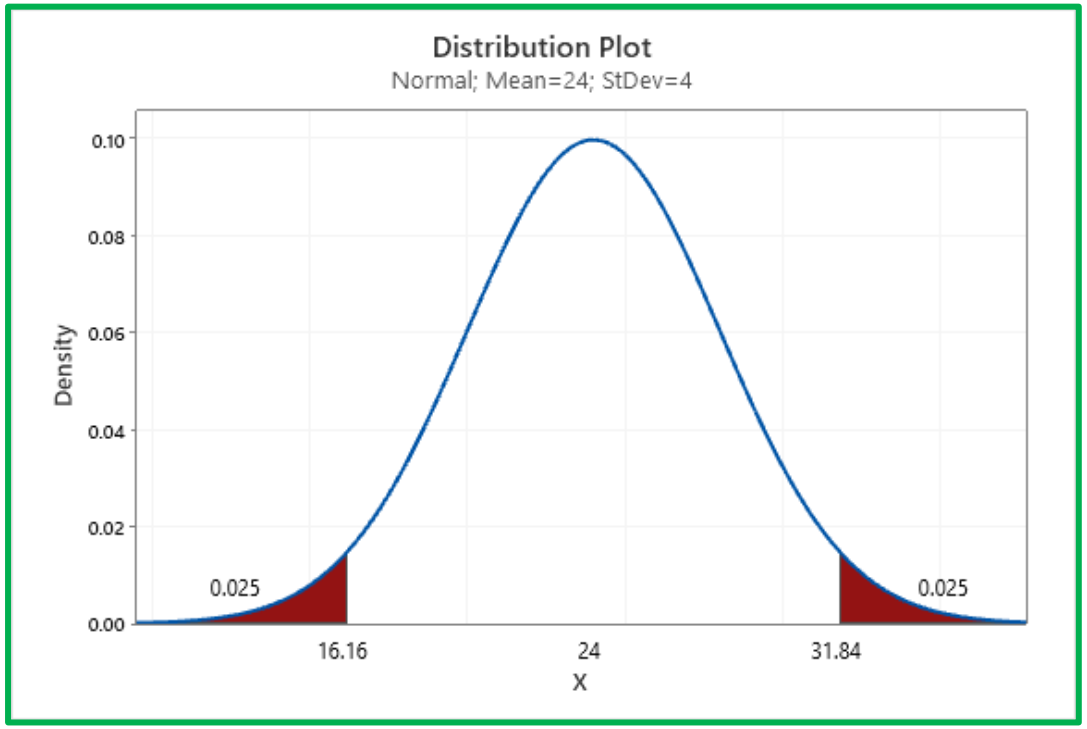
- **Problem**: the company GoodFish buys a fish farm to produce seabasses. They want to have a better understanding of the productive capability of the farm.

- **Available information**:
  - the average length of the fish is 24 cm and the standard deviation is 4 cm

- **Questions**:
  - to understand the typical length of the seabass, what are the 95% confidence limits?
  - if the shortest seabasses that can be sold are 14 cm-long, what is the probability of having a seabass whose length is lower than 14 cm?
  - what is the probability of having a fish that is larger than 42 cm (they have a really high cost on the market)?

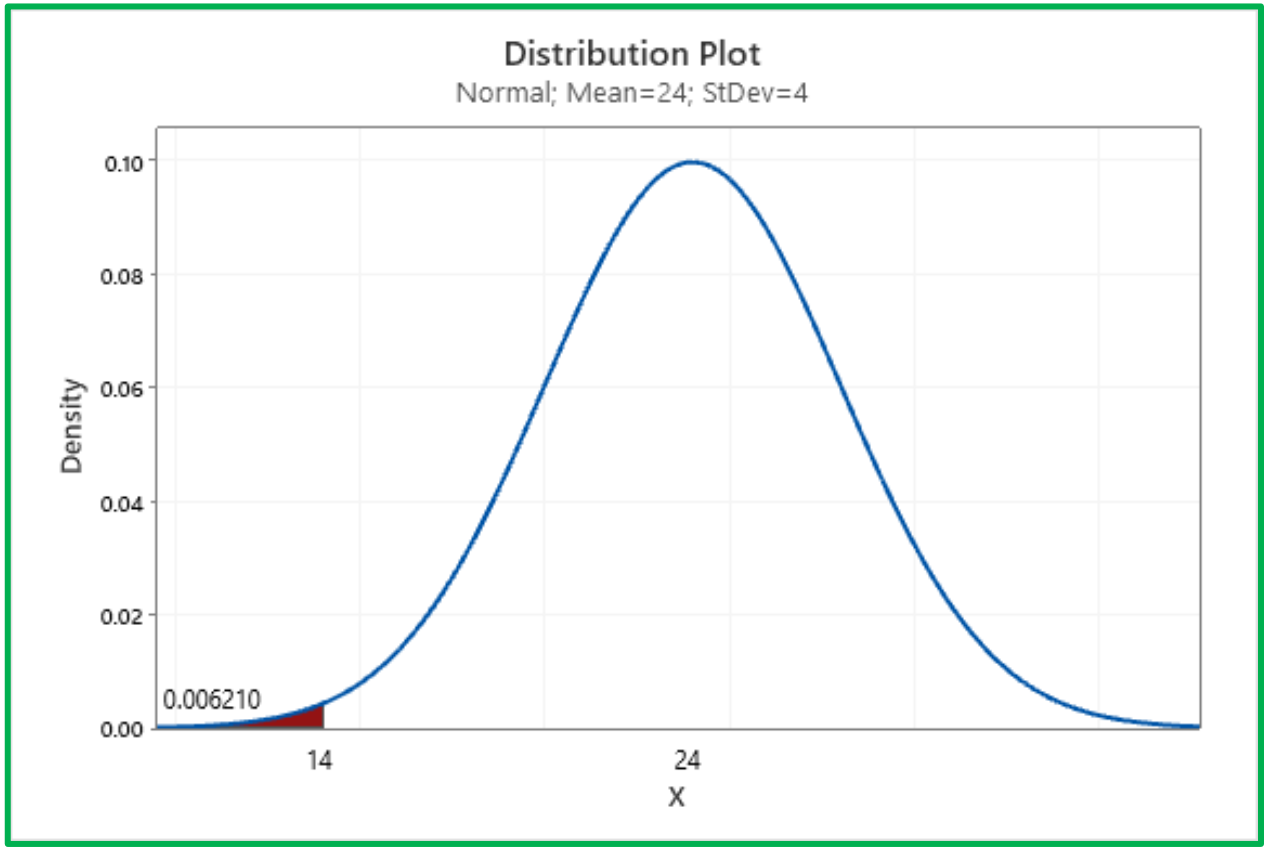
Identification of the normal variability

- The normal variability of seabass length is identified by the 95% confidence limits
  - 2.5% is removed from the upper tail
  - 2.5% is removed from the lower tail
  - the “normal” length is from 16.16 cm to 31.84 cm
- In Minitab®:
  - **Graph**
  - **Probability distribution plot**
  - **View probability**
  - click OK
  - **Distribution: normal**
  - **Mean: 24**
  - **Std. dev.: 4**
  - click **Shaded Area**
  - select **Probability**
  - **Both tails**
  - **Probability: 0.05**
  - click OK



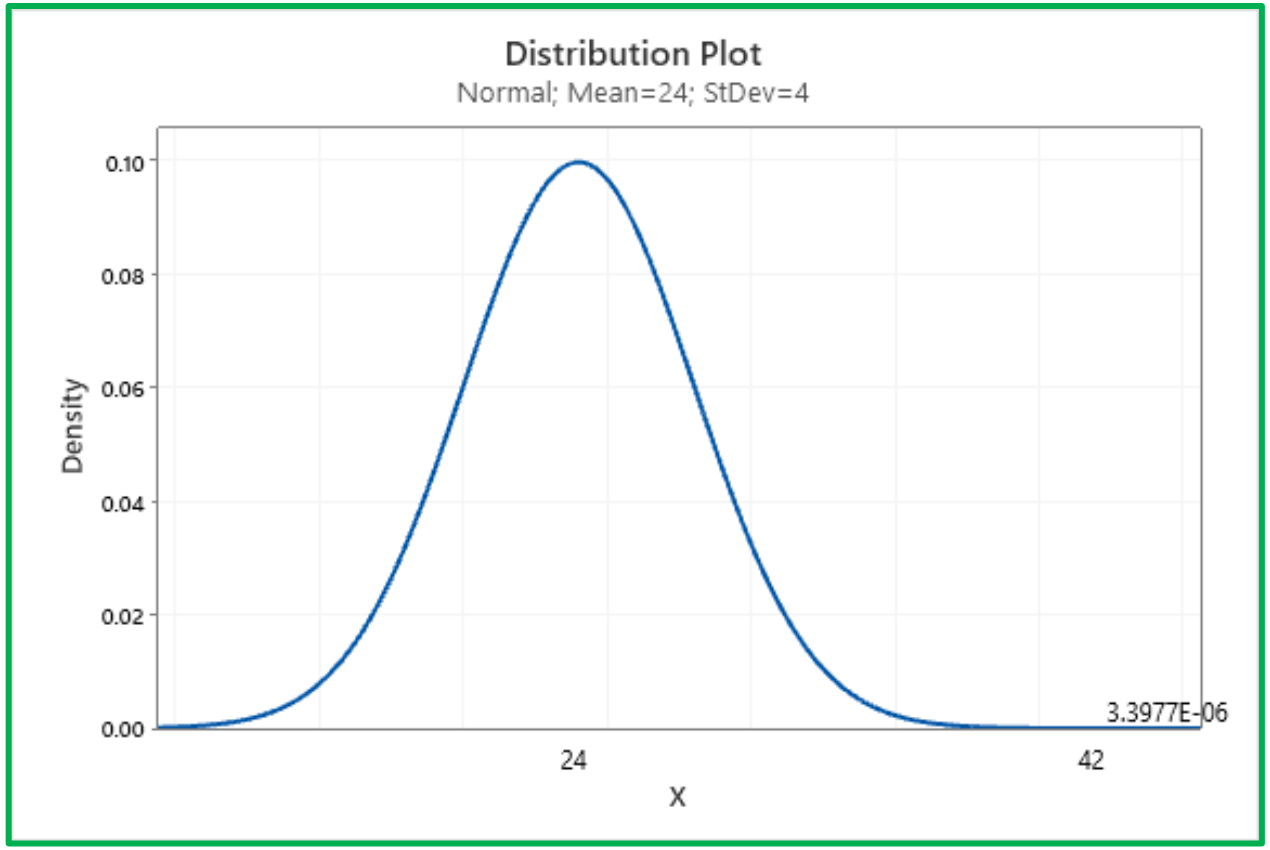
Probability of out-of-spec

- The probability of finding fishes that are too short are 0.62%



Probability of large fishes

- The probability of finding very long fishes is very very low!



Take-home message

- Today we leaned how **descriptive statistics** and **probability theory** can be used to describe the chance of occurrence of the greatest part of the physical, chemical and biological phenomena due to common cause variability
  - **Gaussian distributions**
  - tests of normality
  - **density functions**
    - probability density function
    - cumulative density function
    - inverse density function
  - use software (Minitab® and Matlab®) to:
    - describe univariate data
    - build statistical models for random variables

… per sempre a fianco a me!