ELEMENTS OF STATISTICS FOR THE PHARMACEUTICAL QUALITY CONTROL USING MICROSOFT EXCEL®

TABLE OF CONTENTS

INTRODUCTION

INTRODUCTION

Why we need STATISTICS?

FROM A VERY GENERAL STANDPOINT:

TO DISTINGUISH SIGNAL FROM NOISE !

STATISTICS ALLOWS INFORMATION TO BE SYNTHESIZED AND CONVERTED INTO « READY-TO-USE » KNOWLEDGE

N. Silver, The Signal and the Noise: Why So Many Predictions Fail-but Some Don't, Penguin Press (2012)

INTRODUCTION

What is STATISTICS ?

In general terms, close to the use we will make of it here, **STATISTICS** can be defined as:

Set of logical and mathematical-probabilistic tools for the study of real phenomena that occur with repeated determinations characterized by *variability*

INTRODUCTION

STATISTICS can be sub-divided into two categories (DESCRIPTIVE, INFERENTIAL) which respond more to the needs of schematization: in real applications there are no such clear boundaries.

- **DESCRIPTIVE STATISTICS**: data collection and analysis by means of graphs and summary indices (position, variability and shape).
- **INFERENTIAL STATISTICS:** set of methods that allow to generalize results based on a partial observation (**sample**) : process in ***inductive inference*** **!**

DESCRIPTIVE STATISTICS WITH EXCEL®

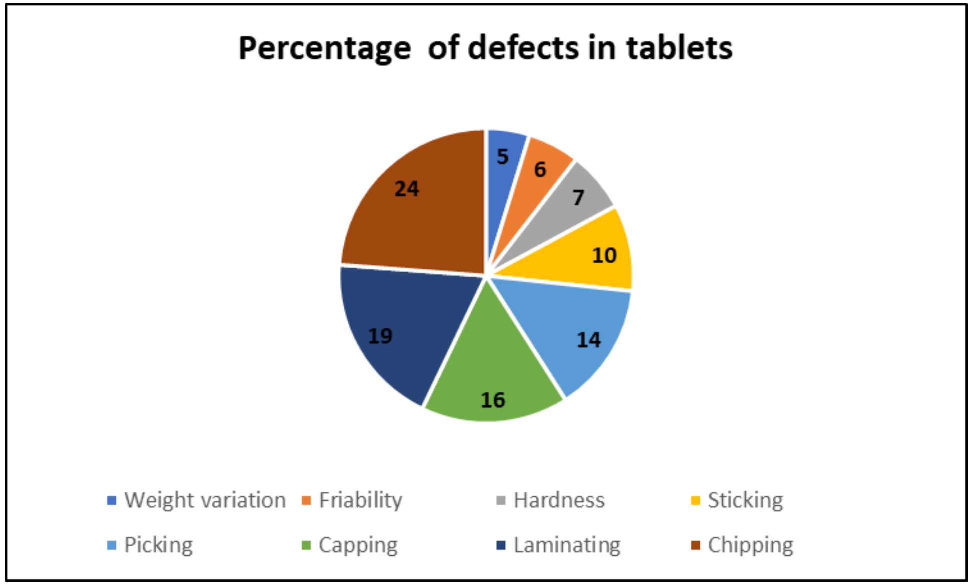
DESCRIPTIVE STATISTICS

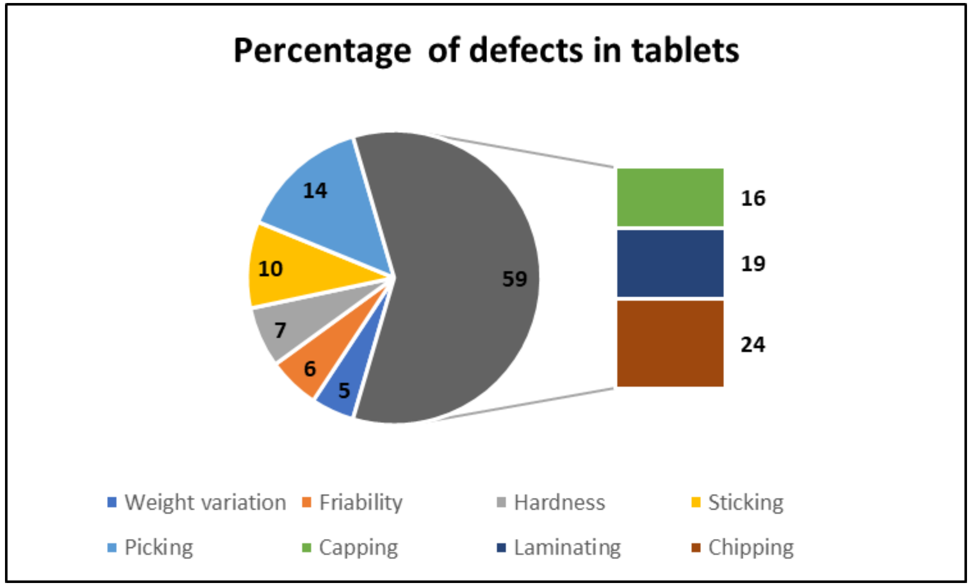
QUALITATIVE DATA are represented using **PIE CHARTS** if no order relationships can be established.

A	B	C	D
	Type of defect tablets	**Number of defects**	**Percentage of defects**
	Weight variation	5	5
	Friability	6	6
	Hardness	7	7
	Sticking	10	10
	Picking	15	14
	Capping	17	16
	Laminating	20	19
	Chipping	25	24
	Sum	105	100

=C4/C13*100

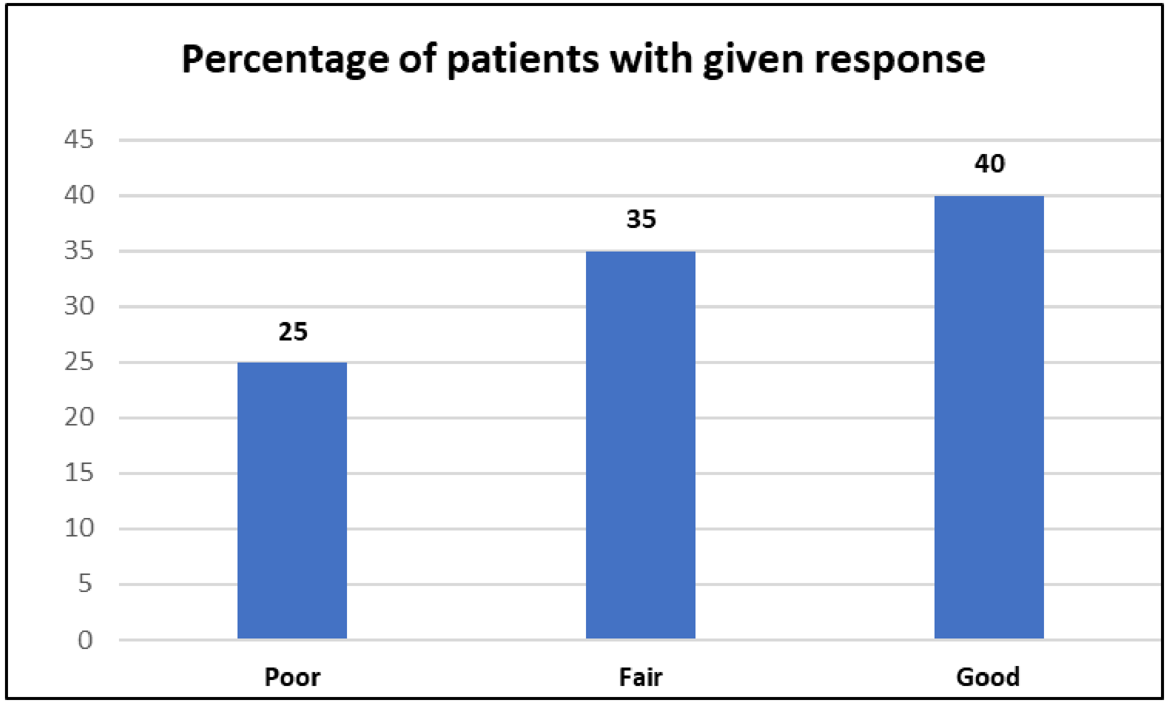
=SUM(C4:C11)

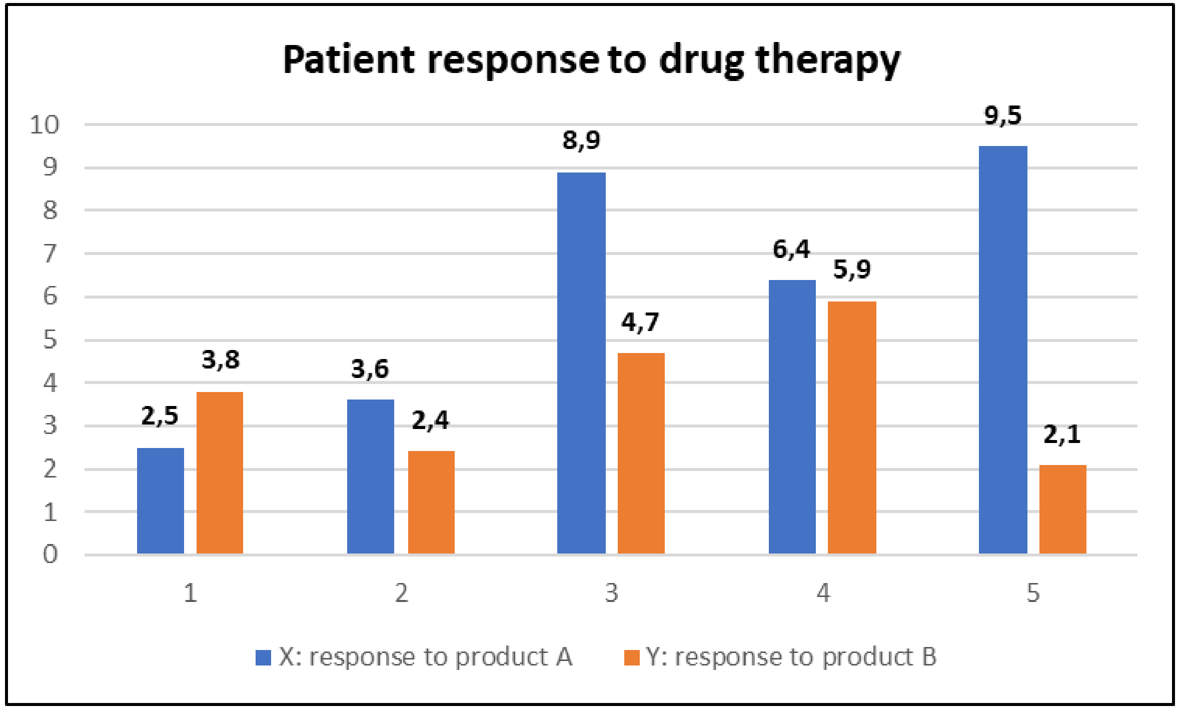




DESCRIPTIVE STATISTICS

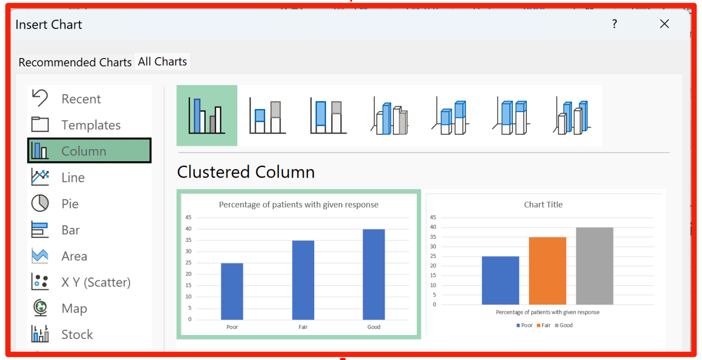
QUALITATIVE DATA are represented using **BAR CHARTS** if an order relationship can be established.

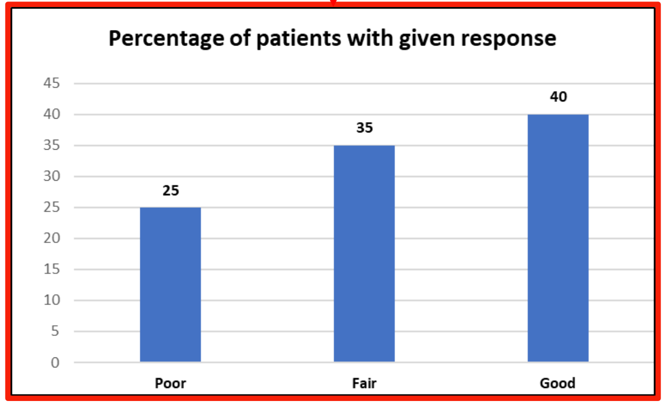




DESCRIPTIVE STATISTICS

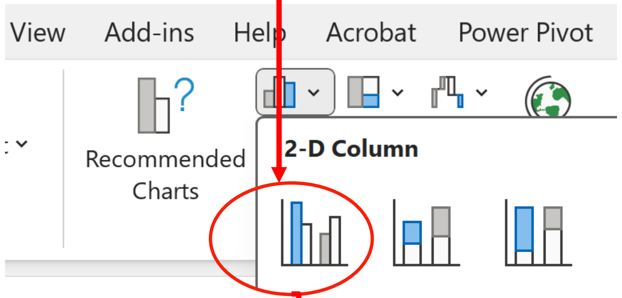
Response	Percentage of patients with given response
Poor	25
Fair	35
Good	40

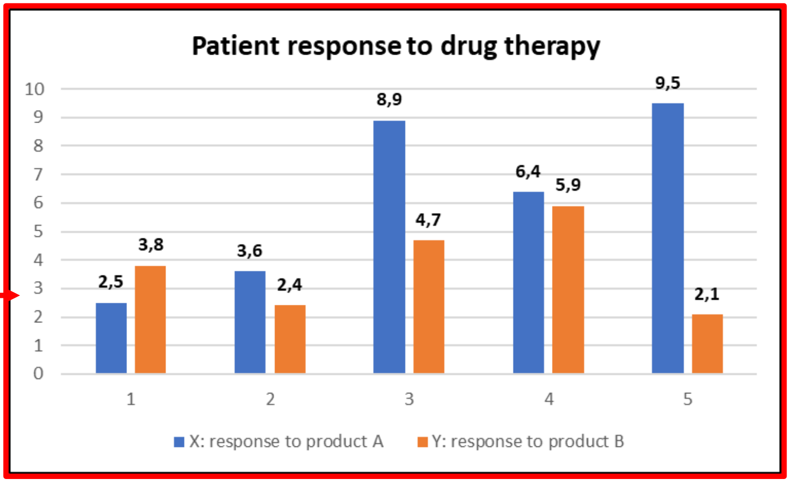




Excel convention
column on the left : *x-axis*
column(s) on the right : *y-axis*

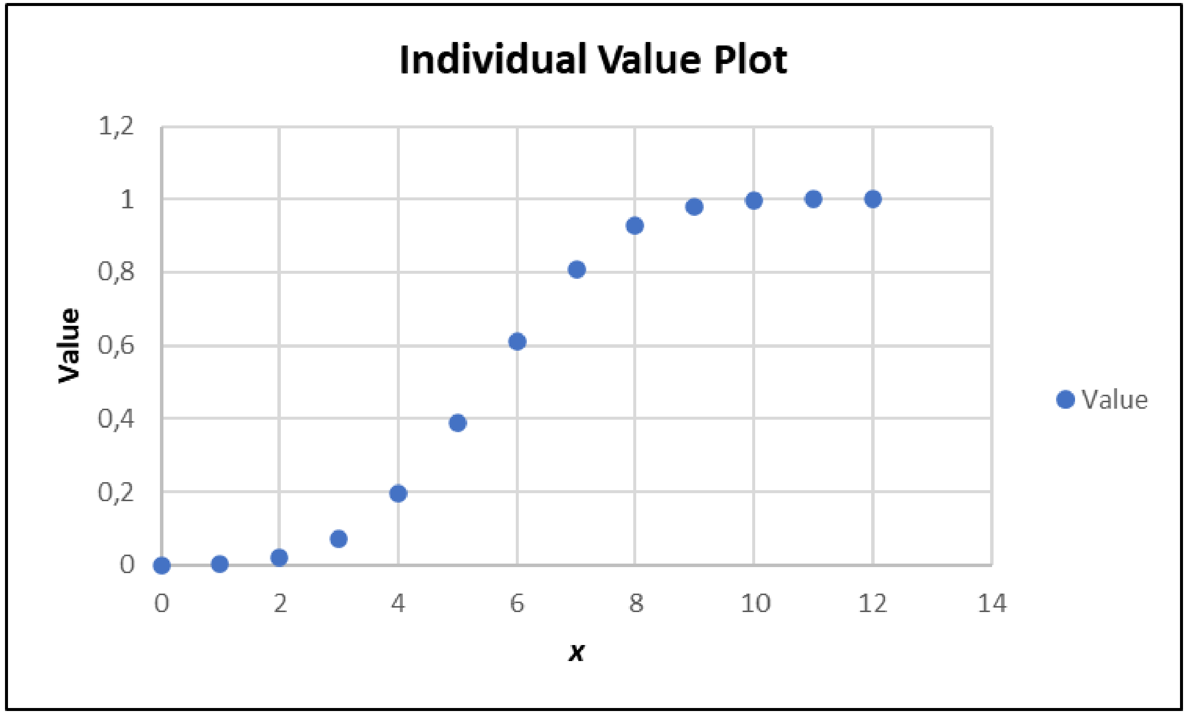
Patient	X: response to product A	Y: response to product B
1	2,5	3,8
2	3,6	2,4
3	8,9	4,7
4	6,4	5,9
5	9,5	2,1

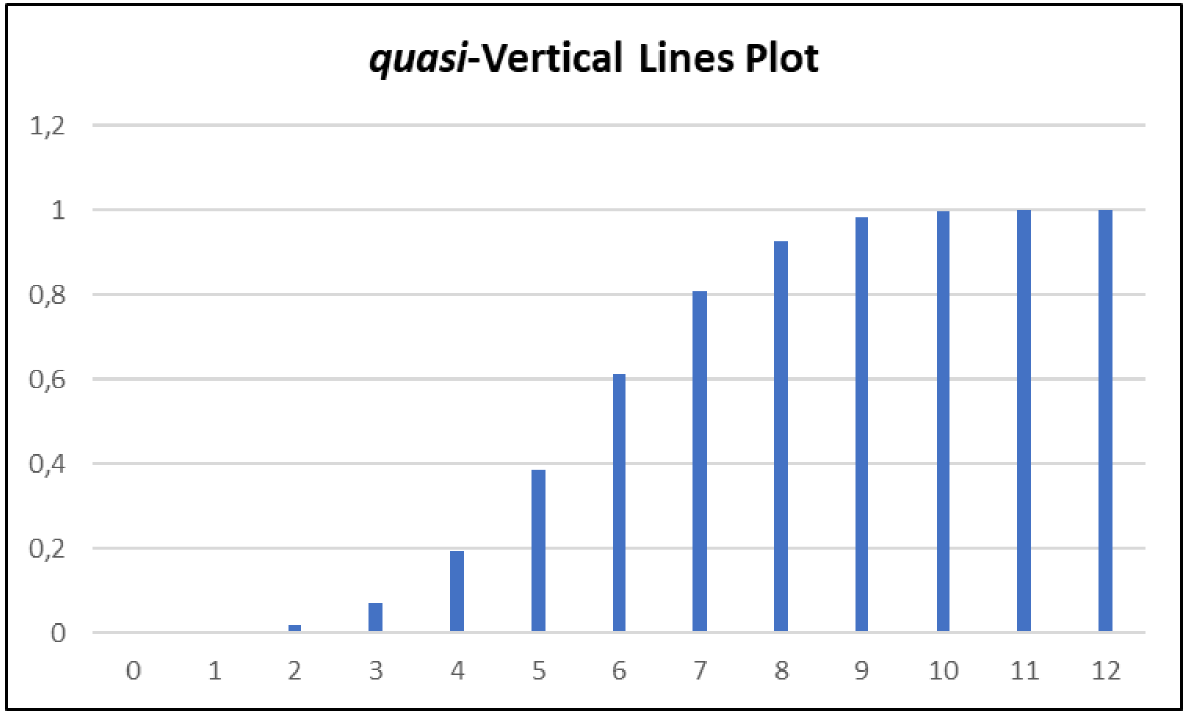




DESCRIPTIVE STATISTICS

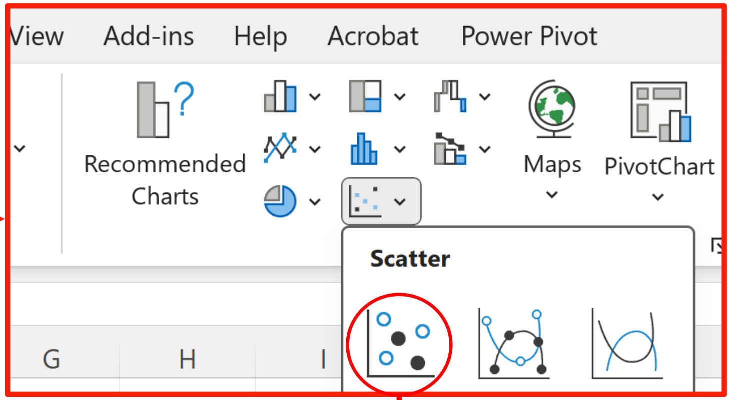
DISCRETE QUANTITATIVE DATA are represented using **INDIVIDUAL VALUE PLOTS.**

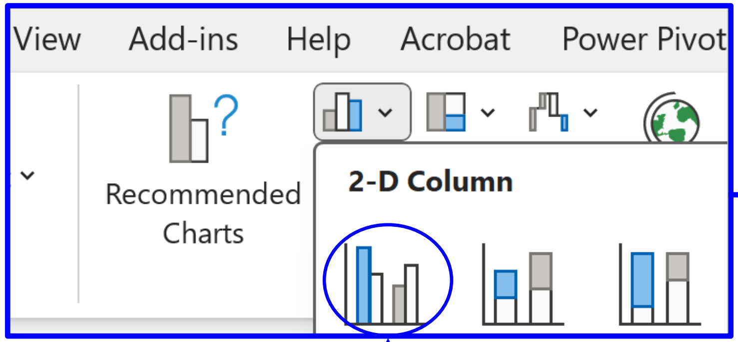


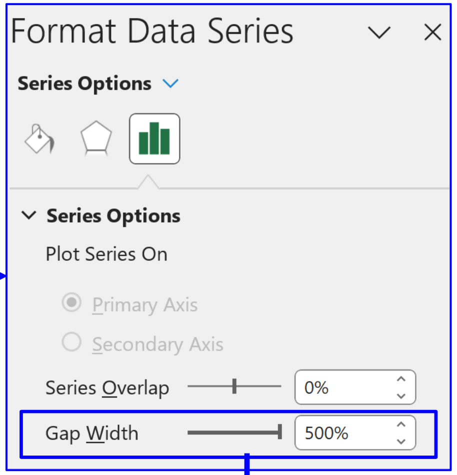


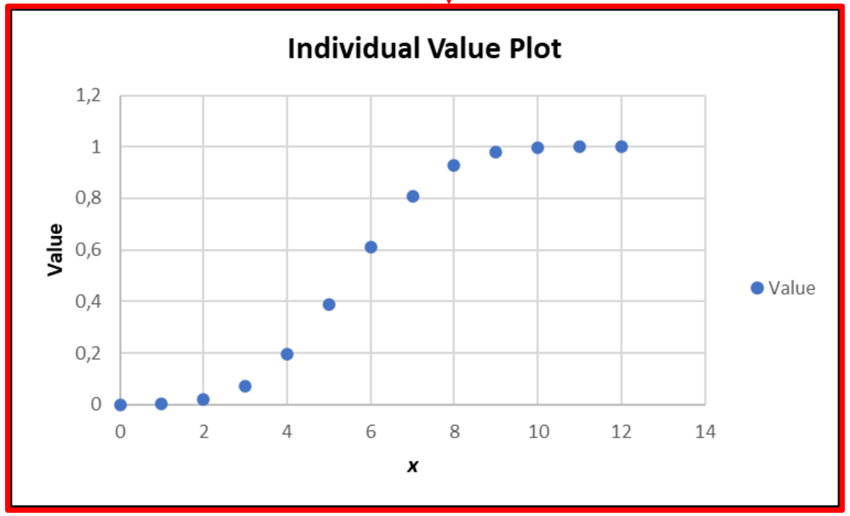
DESCRIPTIVE STATISTICS

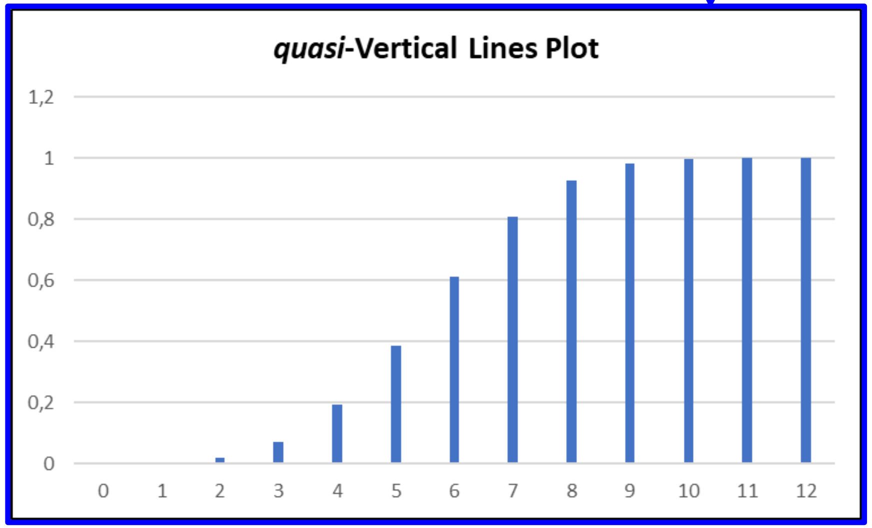
x	Value
0	0,000244
1	0,003174
2	0,019287
3	0,072998
4	0,193848
5	0,387207
6	0,612793
7	0,806152
8	0,927002
9	0,980713
10	0,996826
11	0,999756
12	1









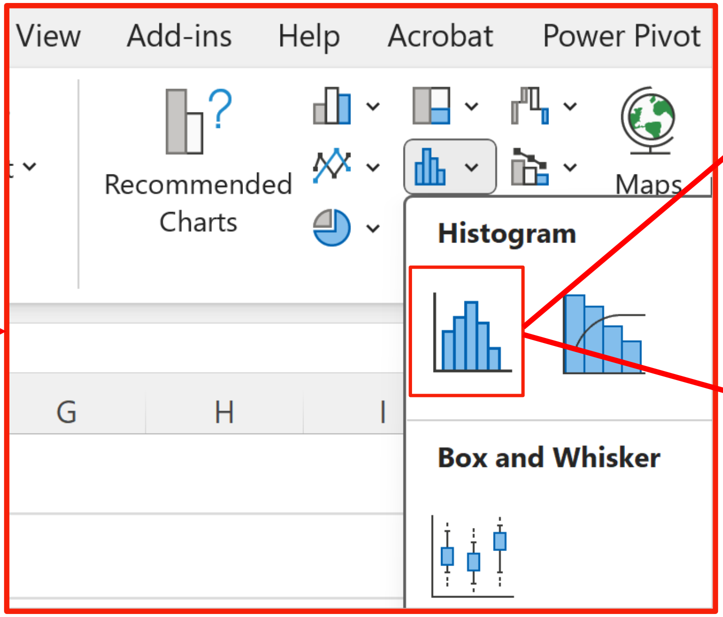


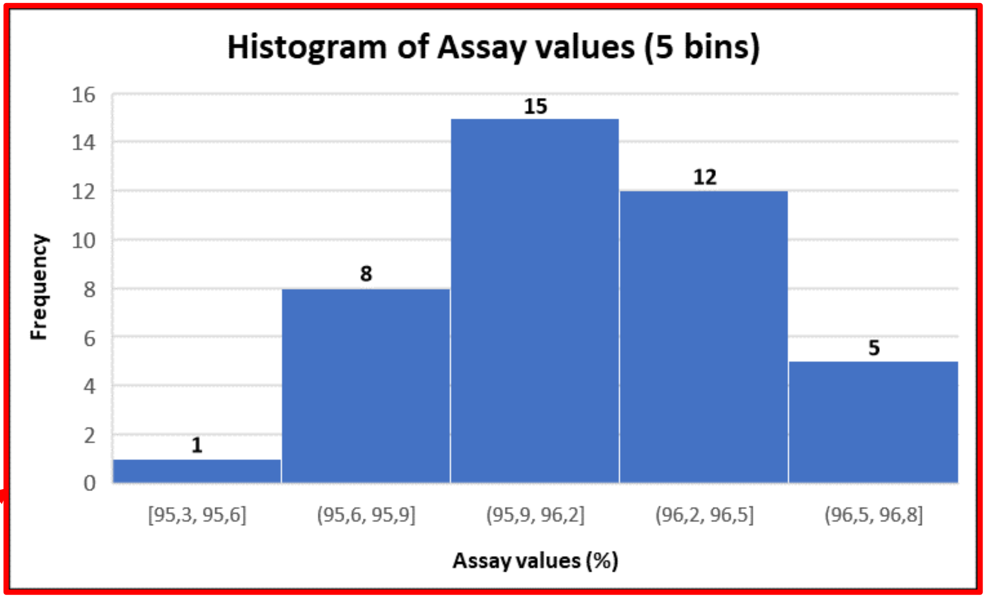
DESCRIPTIVE STATISTICS

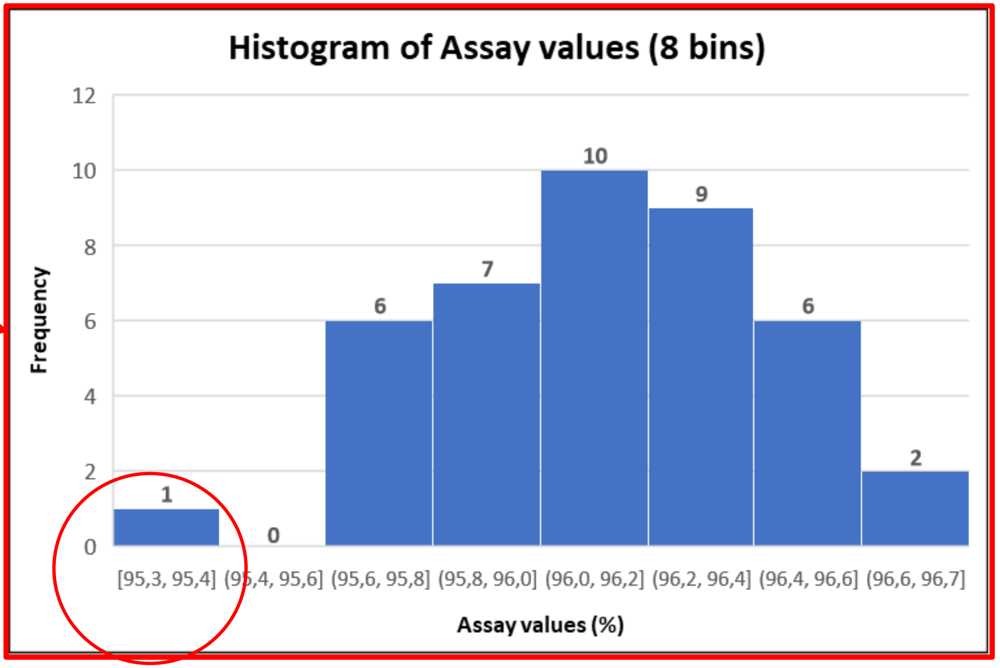
CONTINUOUS QUANTITATIVE DATA are represented using **HISTOGRAMS** which are useful not only to understand the distribution of values (*i.e.*, *central tendency*, *variability*, *shape*) and look for ***outliers***.

Assay (%)
95,9
96,5
96,3
96,3
95,8
.....
96,2
96,2
95,7
95,3
95,6
95,8
96,1
...
95,8
96,0
96,2
96,2
96,0

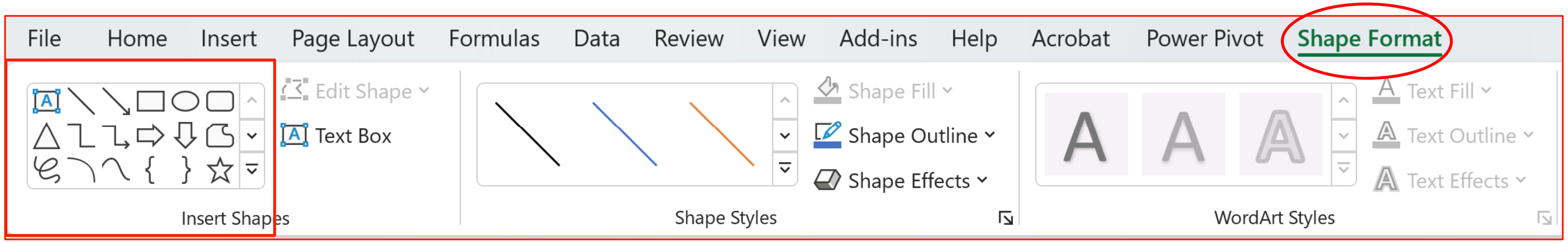
Recommendation: check different numbers of bins !



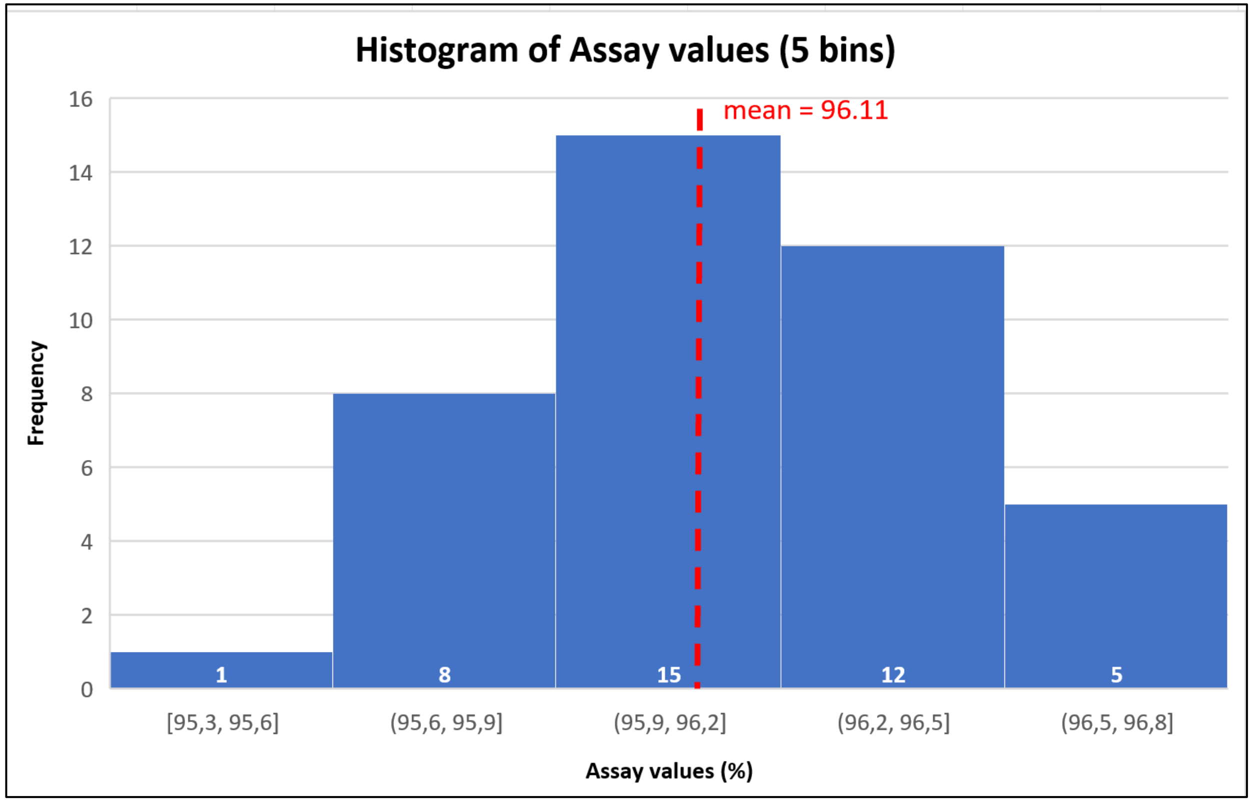




DESCRIPTIVE STATISTICS

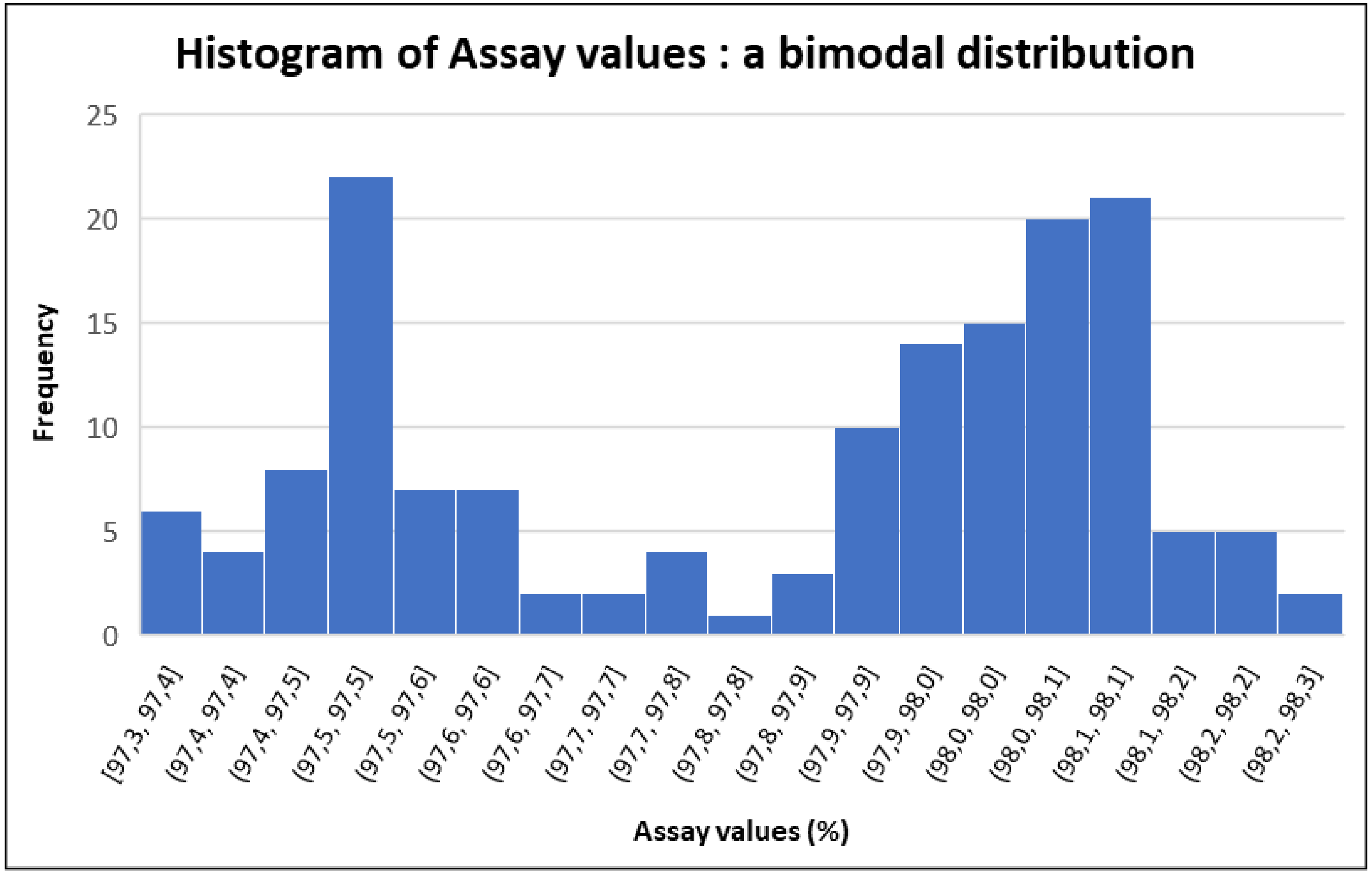


HISTOGRAM can be completed with a vertical line showing, for instance the arithmetic mean of the data set.



DESCRIPTIVE STATISTICS

HISTOGRAMS are also useful reveal *multimodal distributions*.



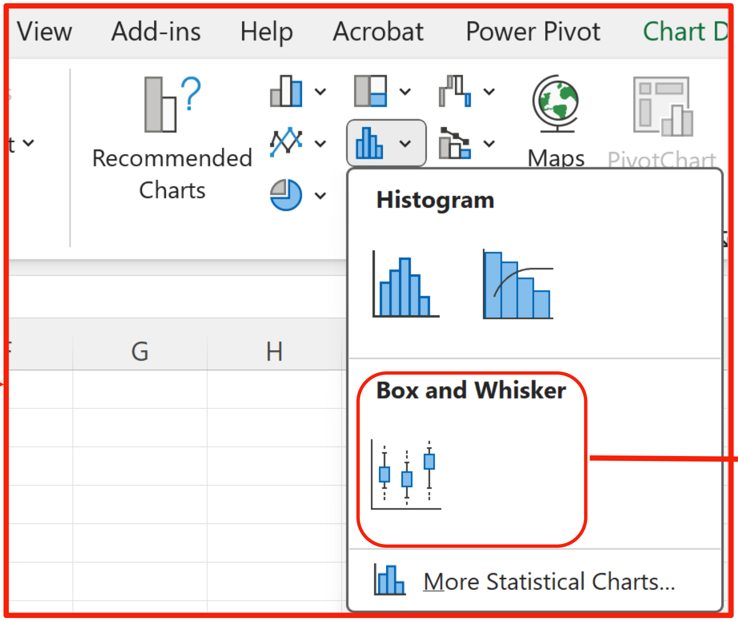
Assay values
97,86
97,91
97,89
97,90
98,10
97,97
97,94
98,03
98,01
97,82
97,96
98,17
97,93
……
97,47
97,53
97,66
97,50
97,58
97,54
97,39
97,35
97,50
…..

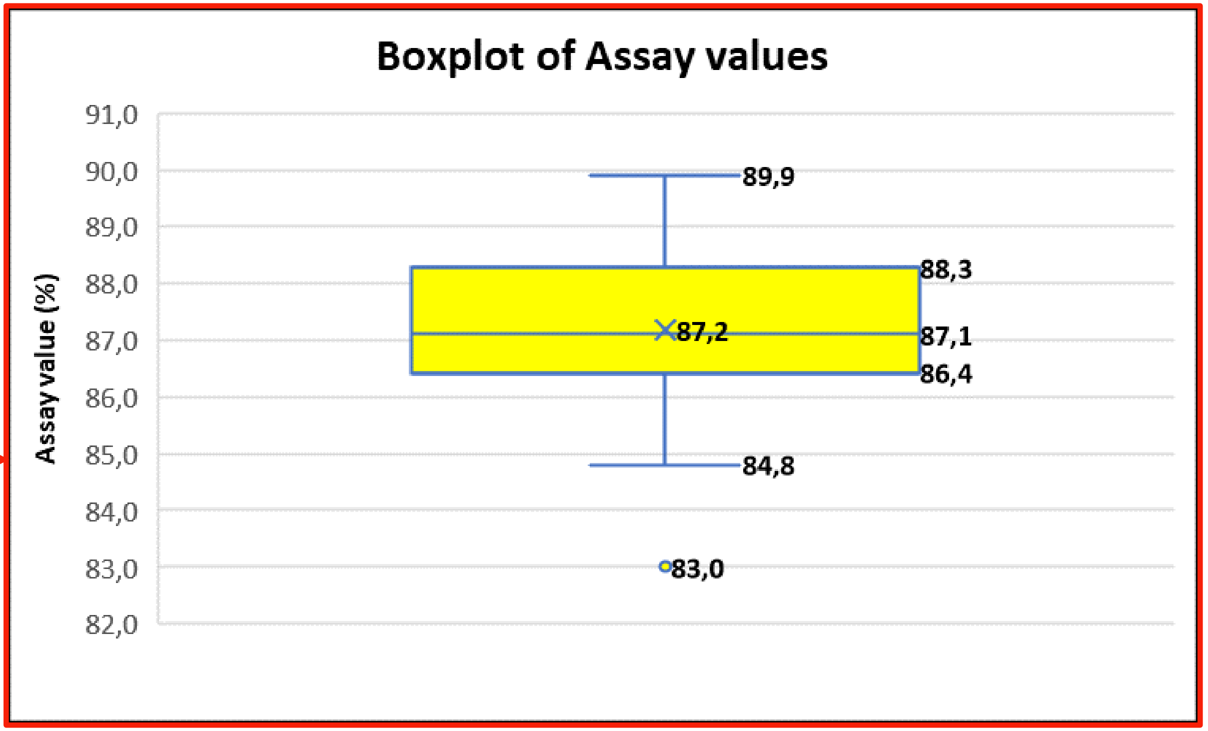
DESCRIPTIVE STATISTICS

Assay value (%)
86,6
88,2
86,4
88,3
85,4
89,9
84,8
87,0
89,6
88,8
86,1
87,9
83,0
88,5
87,2
88,0
86,5
87,5
87,0
87,0

CONTINUOUS QUANTITATIVE DATA can also be effectively represented even using

BOX PLOTS





DESCRIPTIVE STATISTICS

1st Quartile, Q1: 25% of the data ≤ this value

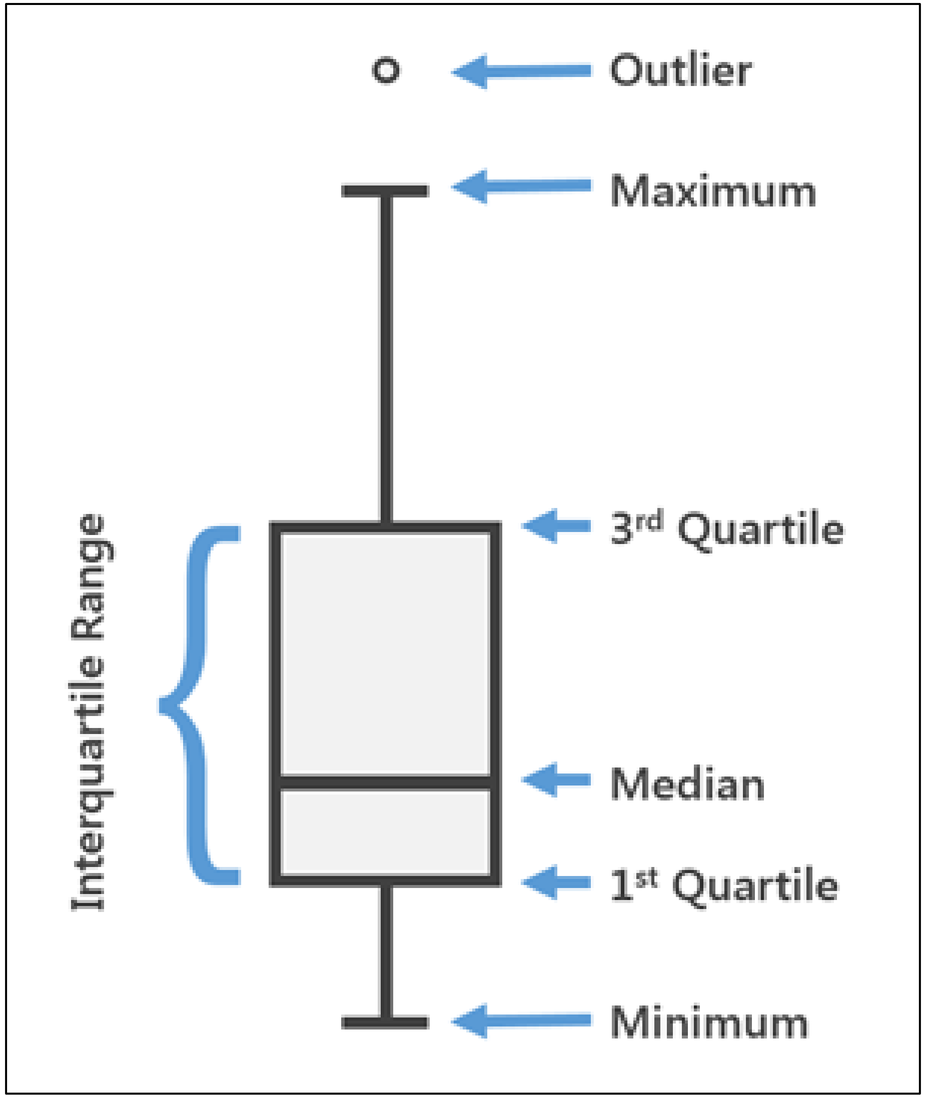
Median, Q2: 50% of the data ≤ this value

3rd Quartile, Q3: 75% of the data ≤ this value

Interquartile range: 50% of the data

Whiskers: extend to the minimum / maximum date point within 1.5 IQR from the bottom / top of the box

Outlier : observation beyond upper or lower whisker, *i.e.,* over 1.5IQR



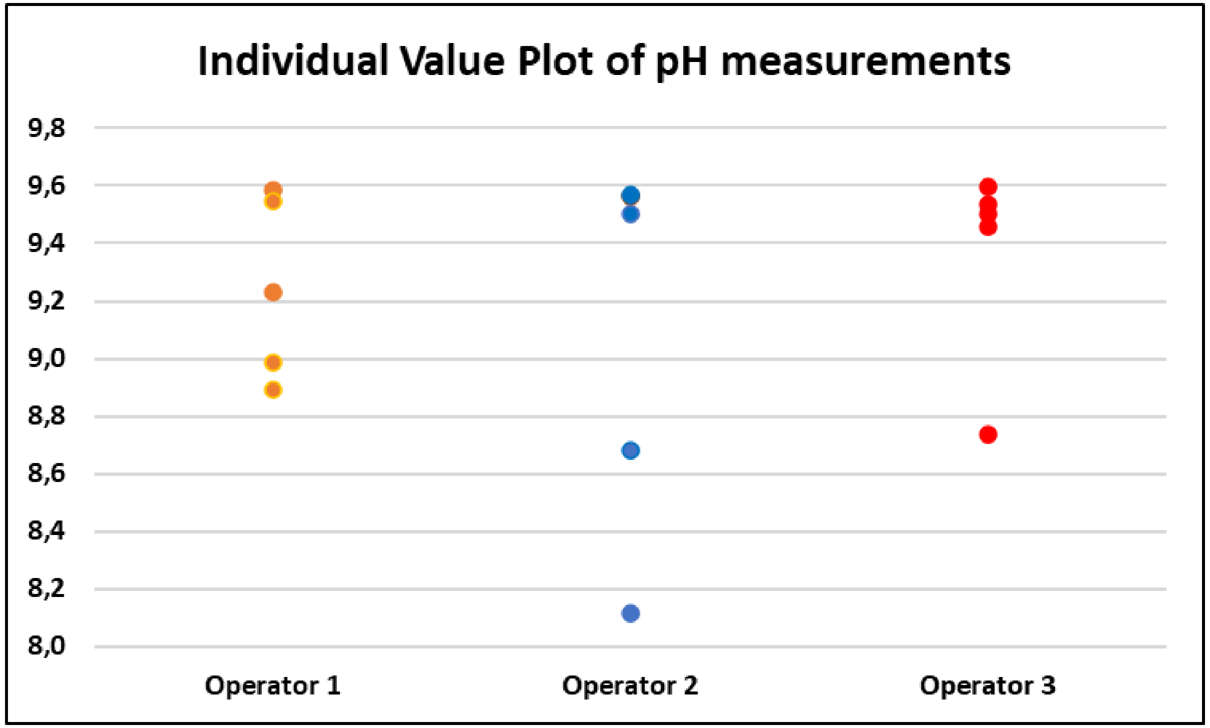
J.W. Tukey, Exploratory Data Analysis, Addison Wesley, 1977

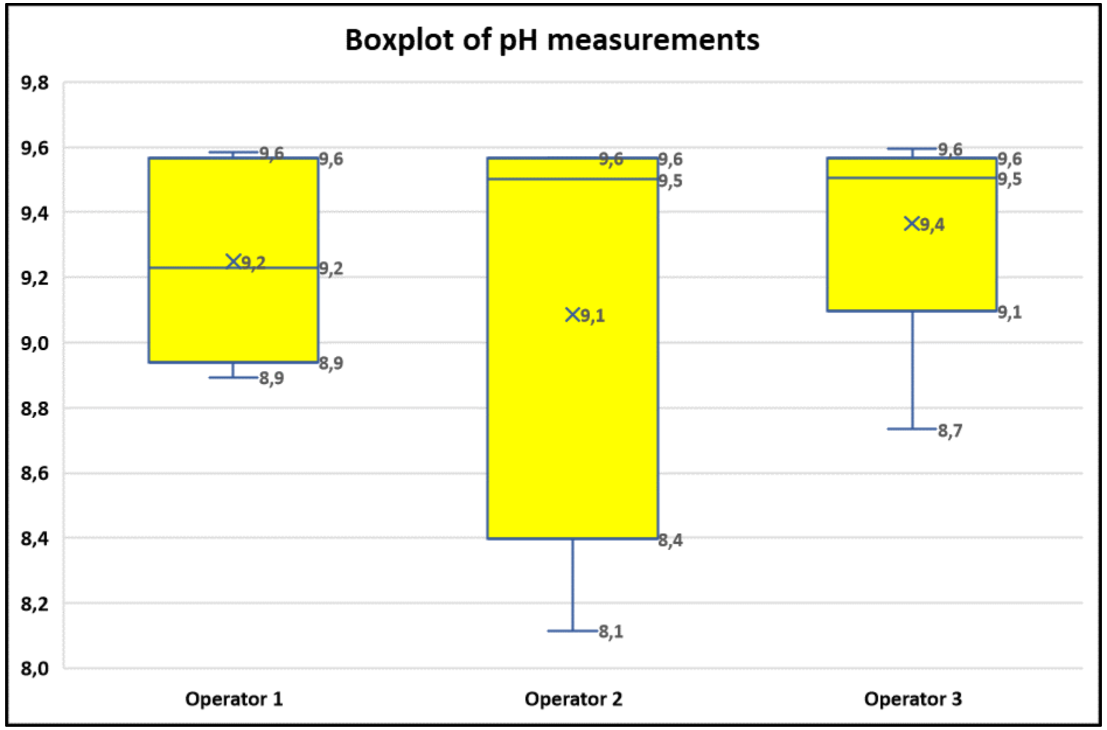
What does a Boxplot tell us at a glance?

- **If it looks «compact»** : most of the data are like each other since there are so many values in a narrow range
- **If it looks «stretched»** : most of the data are quite different from each other, as the values spread over a wide range
- **If the median is close to the bottom**: most of the data will have the lower range values
- **If the median is close to the top**: most of the data will have the higher values of the range
- **If the median is not in the center** data distribution will be « tailed »

DESCRIPTIVE STATISTICS

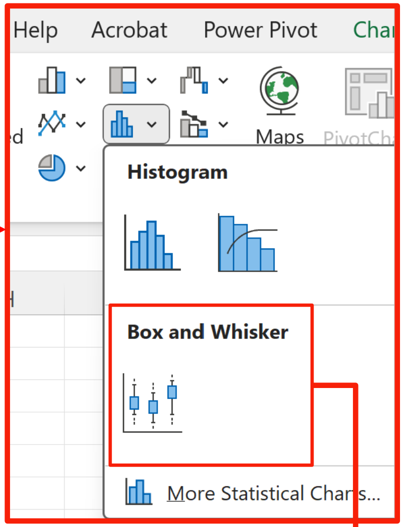
Previous types of plots are useful for multiple data sets comparisons such as, for instance:

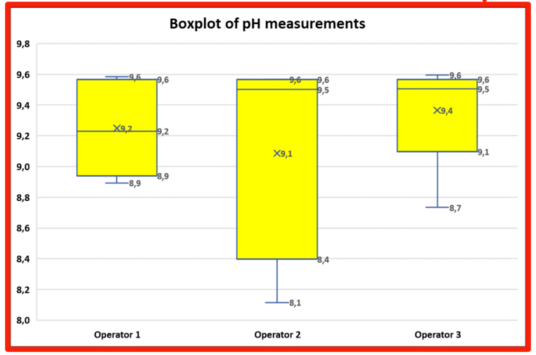




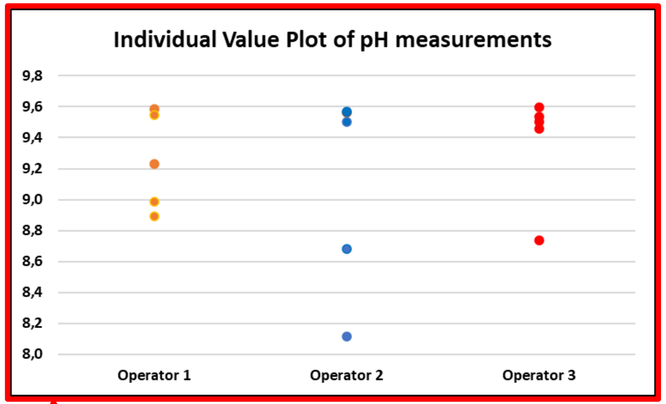
DESCRIPTIVE STATISTICS

Operator	pH value
Operator 1	9,0
Operator 1	9,2
Operator 1	9,6
Operator 1	8,9
Operator 1	9,5
Operator 2	8,7
Operator 2	8,1
Operator 2	9,6
Operator 2	9,5
Operator 2	9,6
Operator 3	9,6
Operator 3	8,7
Operator 3	9,5
Operator 3	9,5
Operator 3	9,5

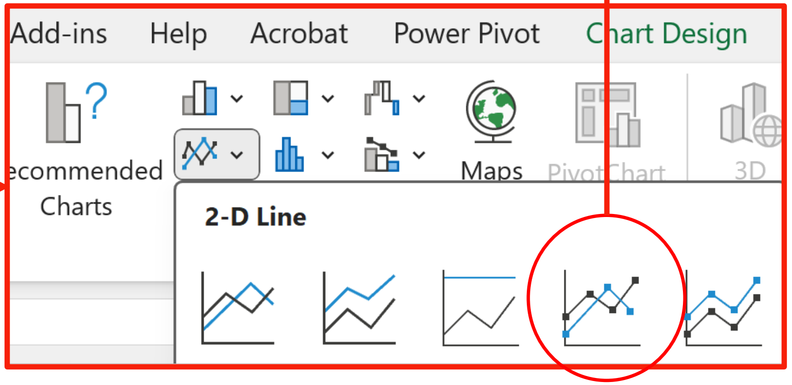




Operator 1	Operator 2	Operator 3
9,0		
9,2		
9,6		
8,9		
9,5		
	8,7	
	8,1	
	9,6	
	9,5	
	9,6	
		9,6
		8,7
		9,5
		9,5
		9,5



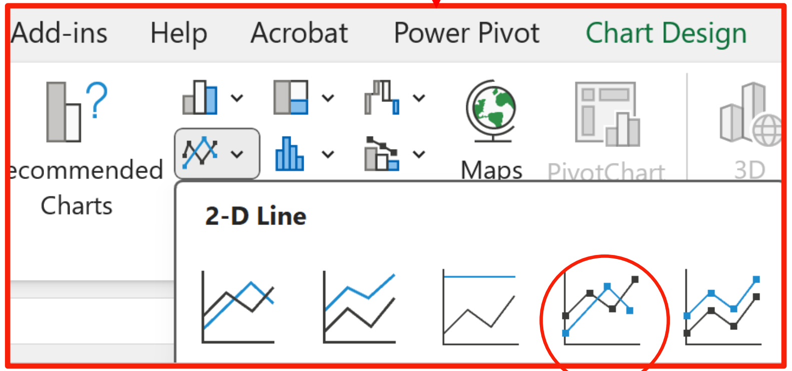
1. remove lines
2. Invert row/columns

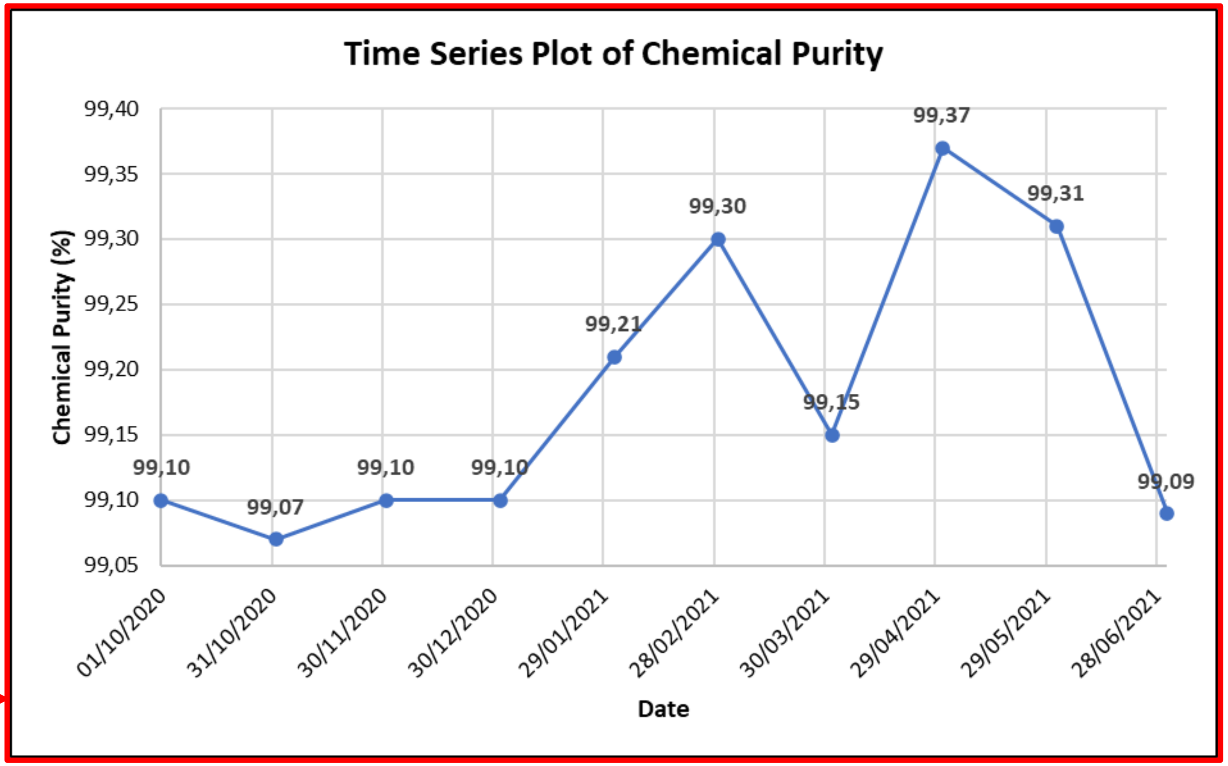


DESCRIPTIVE STATISTICS

TIME SERIES is a sequence of data points listed (or graphed) in time order. This type of graphs are also known as ***Line Graphs.***

Date	Chemical Purity
01/10/2020	99,10
01/11/2020	99,07
01/12/2020	99,10
01/01/2021	99,10
01/02/2021	99,21
01/03/2021	99,30
01/04/2021	99,15
01/05/2021	99,37
01/06/2021	99,31
01/07/2021	99,09





DESCRIPTIVE STATISTICS

Please, duly consider the following quote:

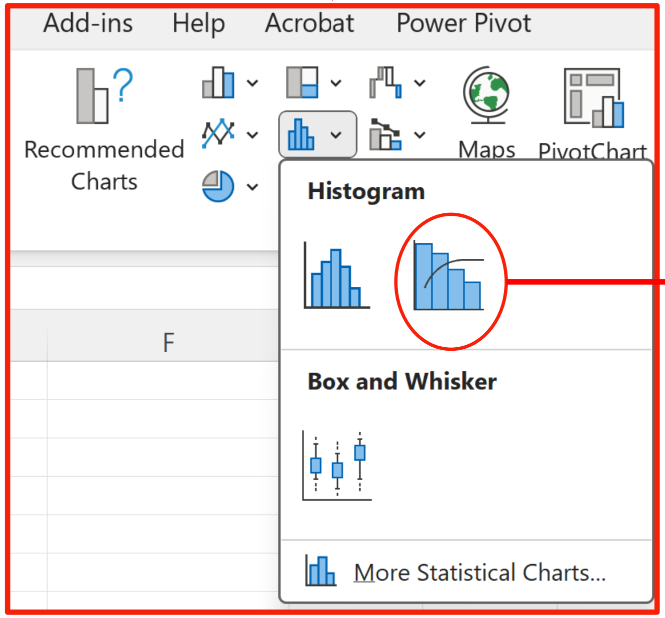
« …**TIME SERIES PLOTS and HISTOGRAMS can be thought as COMPLEMENTARY TO EACH OTHER.** While the histogram collapses all the data, showing its overall shape, the time series plot stretches out the data showing the sequential information that is obscured by the histogram. »

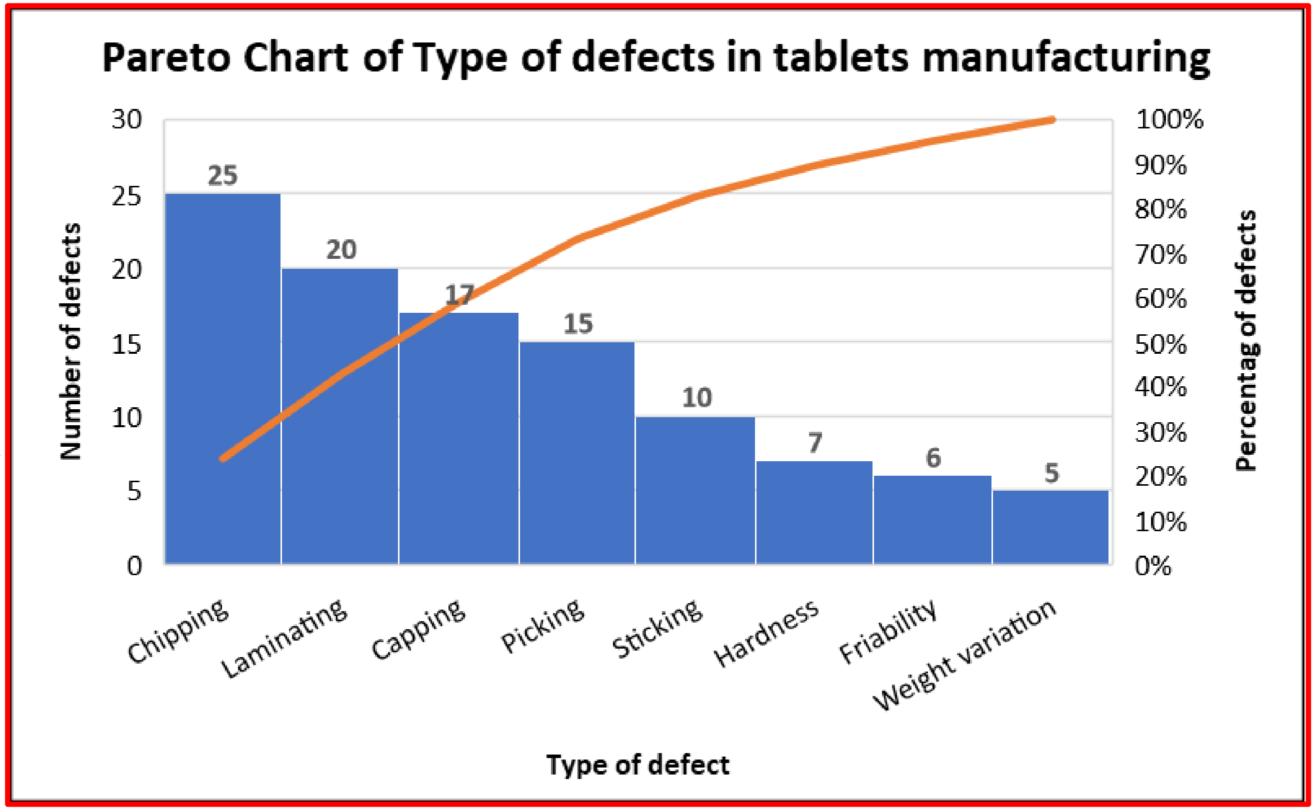
D.J. Wheeler, D.S. Chambers, Understanding Statistical Process Control, 2nd Ed., SPC Press, USA, 1992

DESCRIPTIVE STATISTICS

PARETO CHART allows you to sort the causes of defects in a process according to their relative importance.

Type of defect	Number of defects
Weight variation	5
Friability	6
Hardness	7
Sticking	10
Picking	15
Capping	17
Laminating	20
Chipping	25

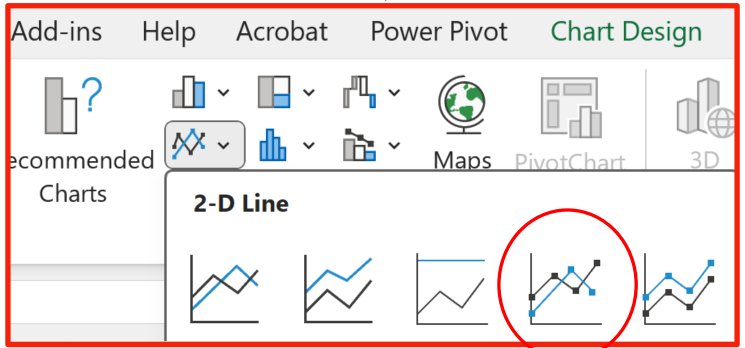


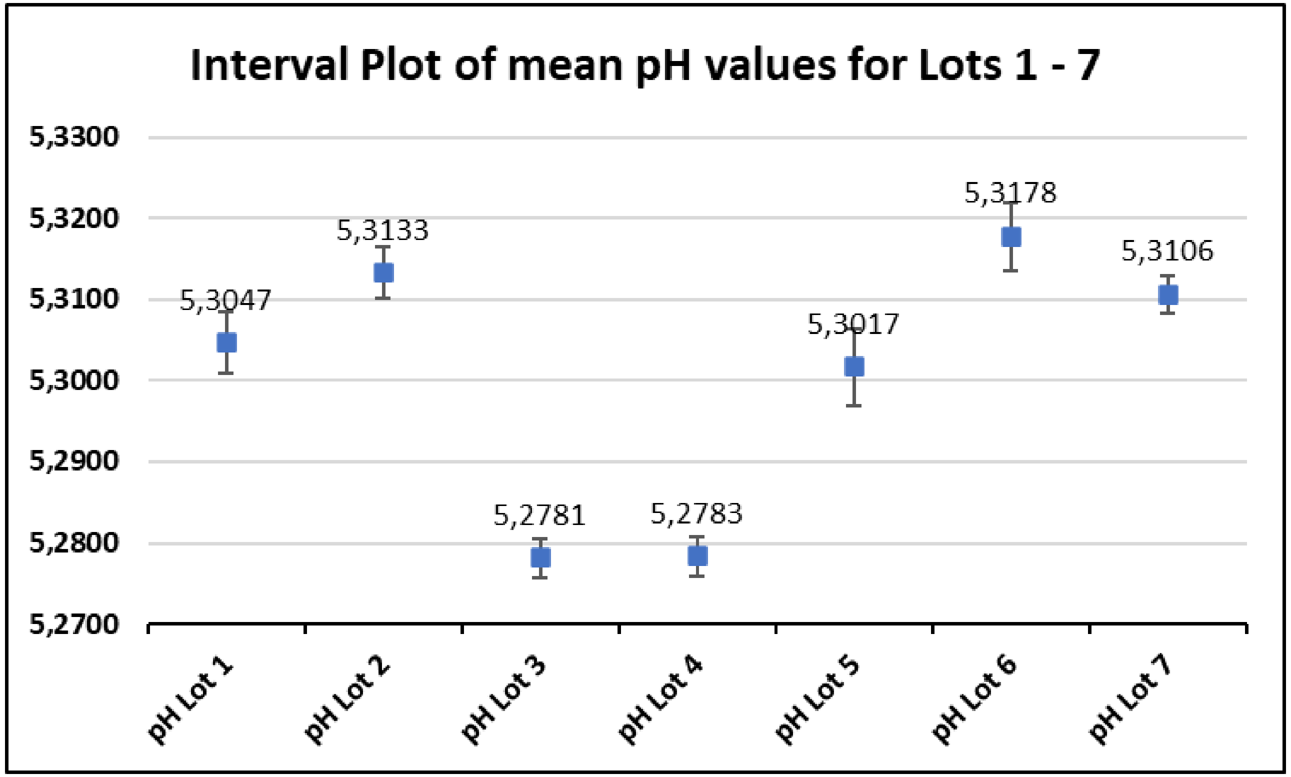


DESCRIPTIVE STATISTICS

To show mean values with margins of error: ***interval plot.***

Lot No.	Mean	Margin of Error
pH Lot 1	5,3047	0,0039
pH Lot 2	5,3133	0,0032
pH Lot 3	5,2781	0,0024
pH Lot 4	5,2783	0,0025
pH Lot 5	5,3017	0,0047
pH Lot 6	5,3178	0,0042
pH Lot 7	5,3106	0,0023



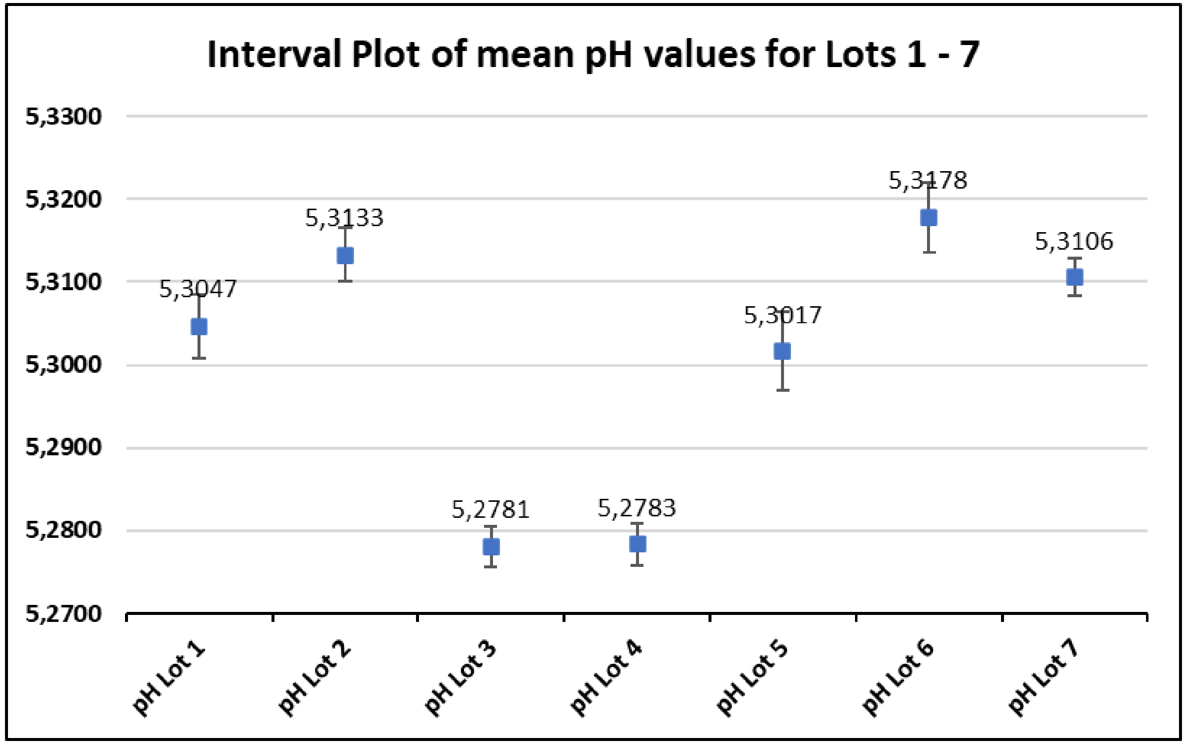


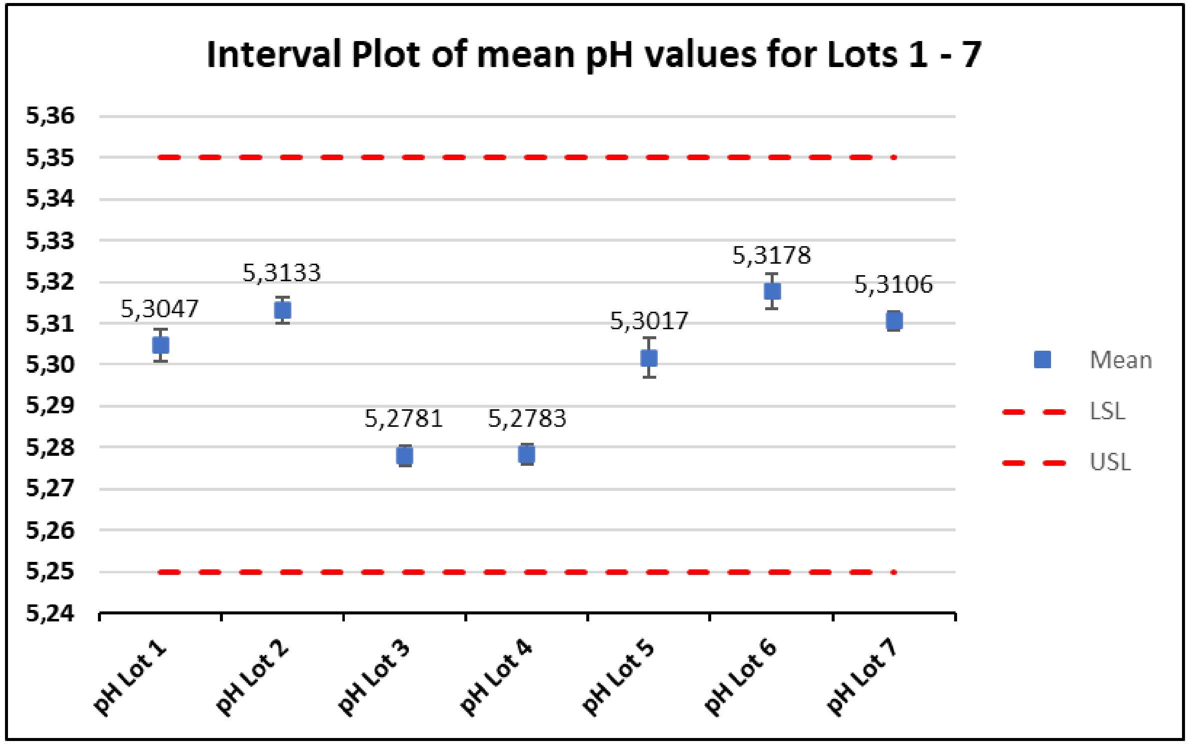
then remove the line and leave just the markers. Select *error bars* and add them.

DESCRIPTIVE STATISTICS

RECOMMENDATION

When conducting data studies, never forget to contextualize them (e.g., report specification limits)



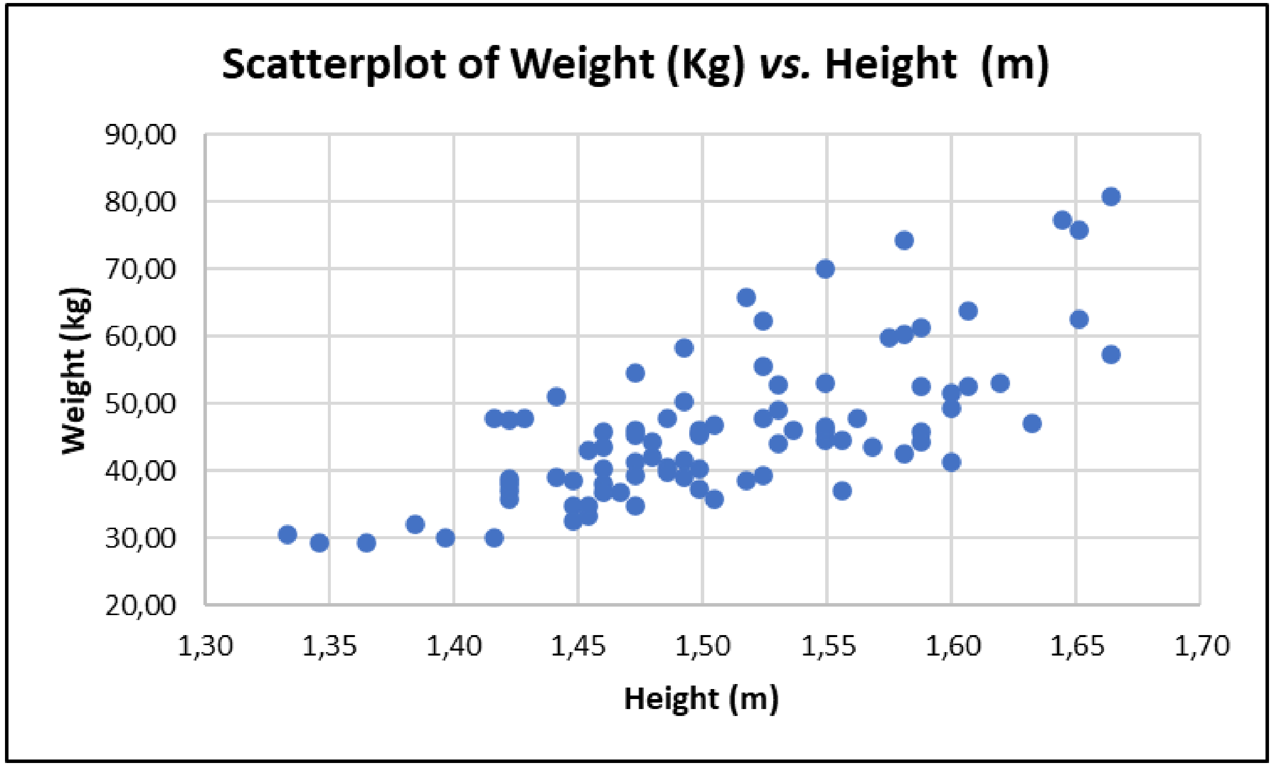


DESCRIPTIVE STATISTICS

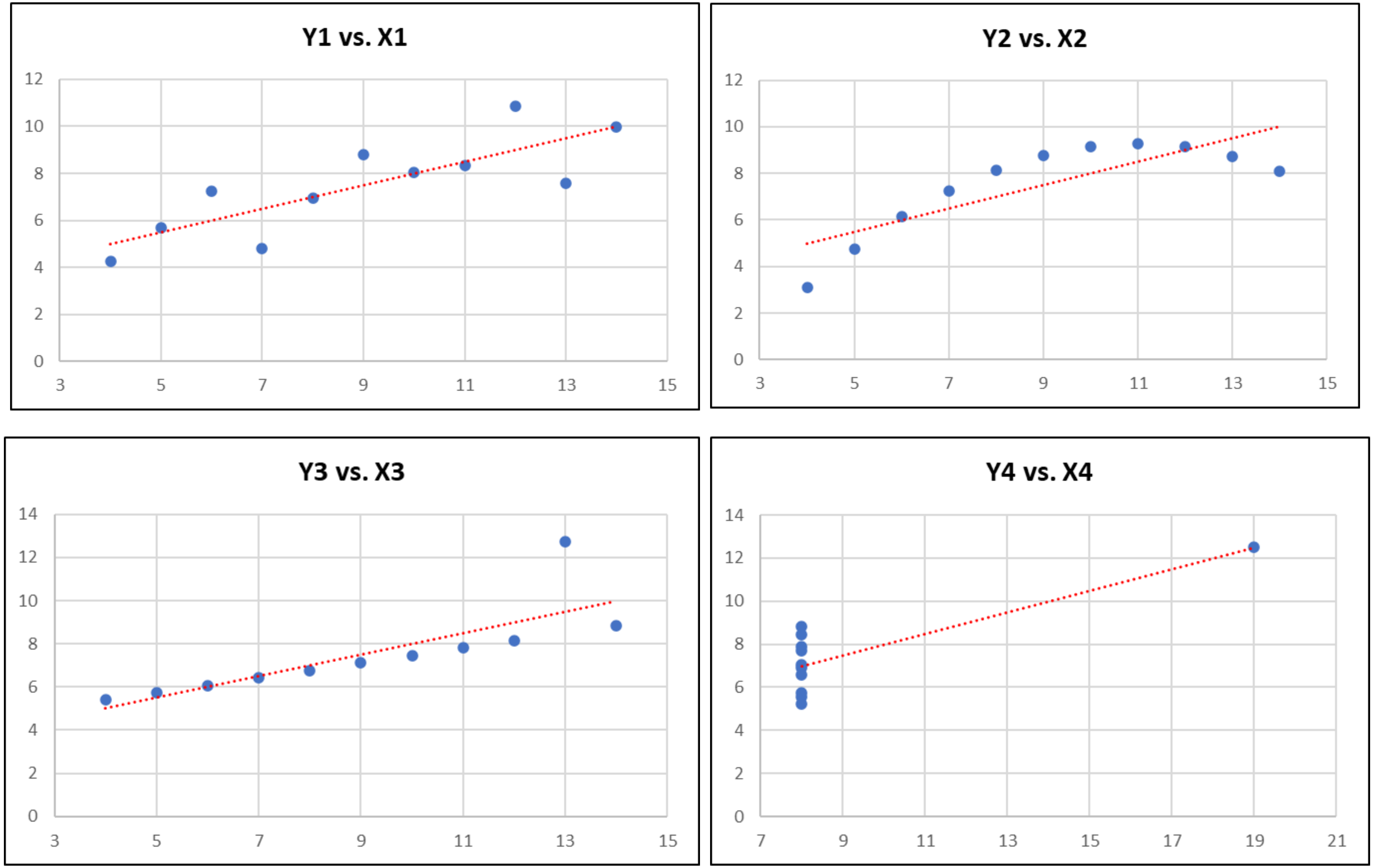
- all examples until now refer to ***one variable***
- in case of two continuous variables:

  scatterplot

- the scatterplot here on the side shows an approximately linear relationship between height and weight, but it does not give any quantitative measure of this relationship !
- Correlation <u>only</u> measures the <u>strength</u> and the <u>direction</u> of association between two variables.



DESCRIPTIVE STATISTICS



Never forget the Anscombe's Quartet ! That's a reason to plot data !

F.J. Anscombe, Graphs in Statistical Analysis, American Statistician, Vol. 27, No. 1 (1973)

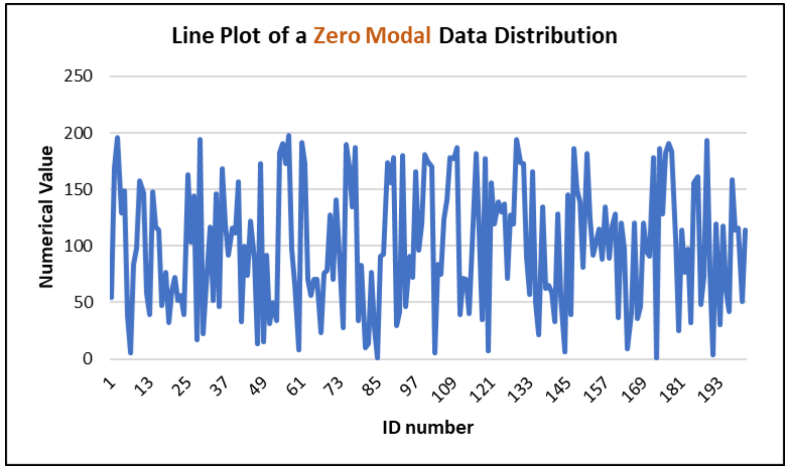
DESCRIPTIVE STATISTICS

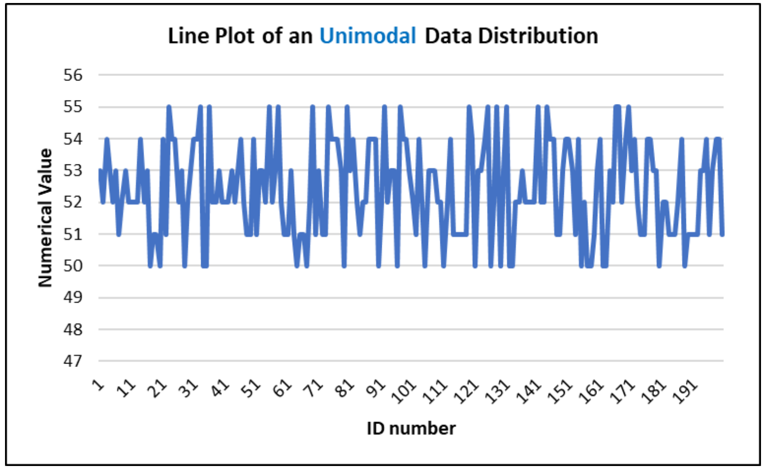
With the term ***summary indices***, or ***statistics*** we mean, in practice, *numerical indicators* that are functions of data. They are of three types:

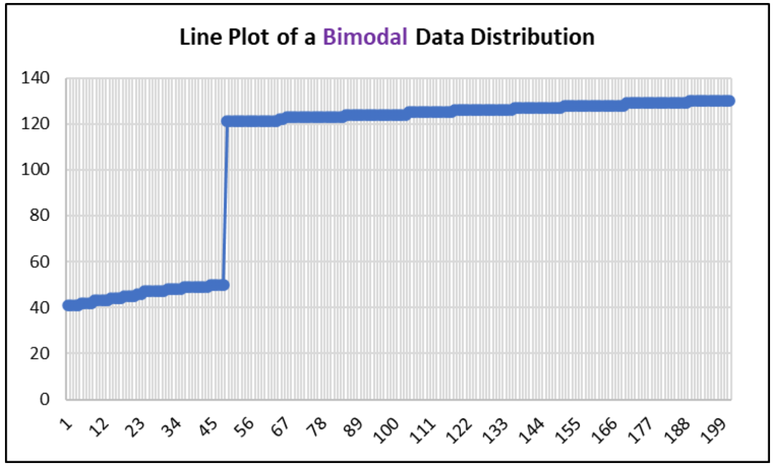
- **Position Indices**: indicators that give an idea of distribution's ***central tendency***. They are of two types:
  - ***non-analytical*** (median, mode, percentiles) and
  - ***analytical*** (analytical means)
- **Variability Indices**: indicators of the diversity / multiplicity of the values of a given variable.
- **Shape Indices**: indicators of the shape of a data distribution

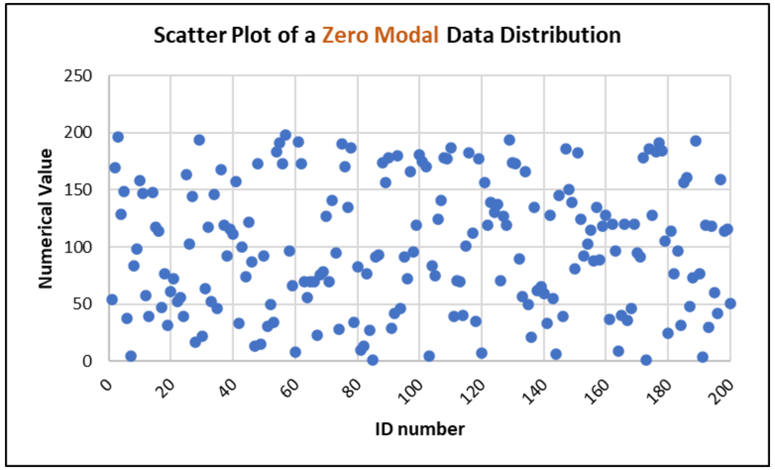
DESCRIPTIVE STATISTICS

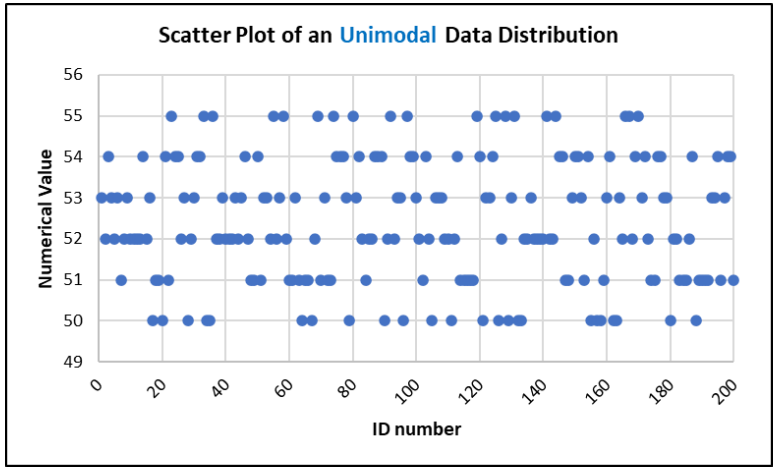
MODE : the value that appears most often in a data set, **=MODE.SNGL()** and **=MODE.MULT()**

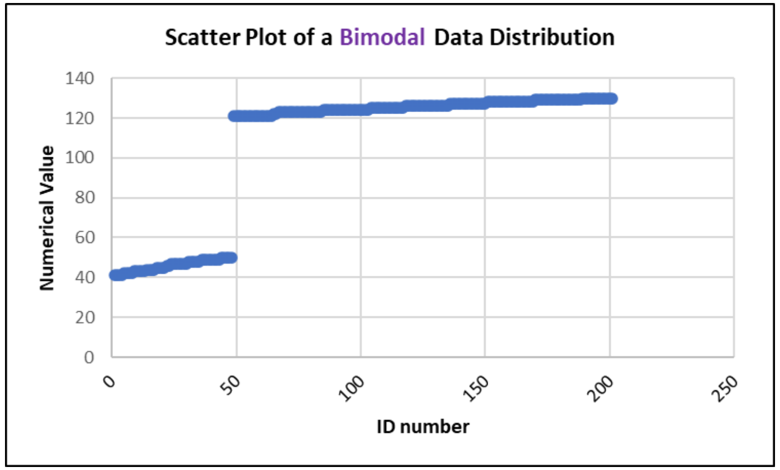






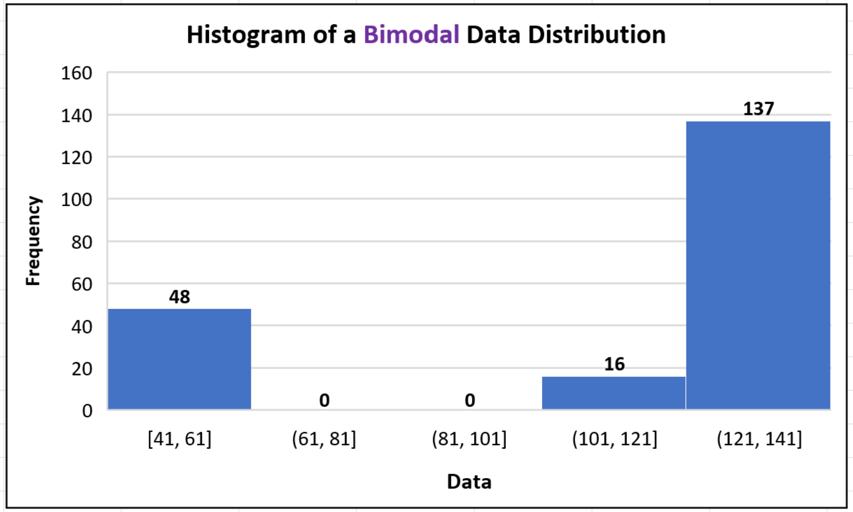


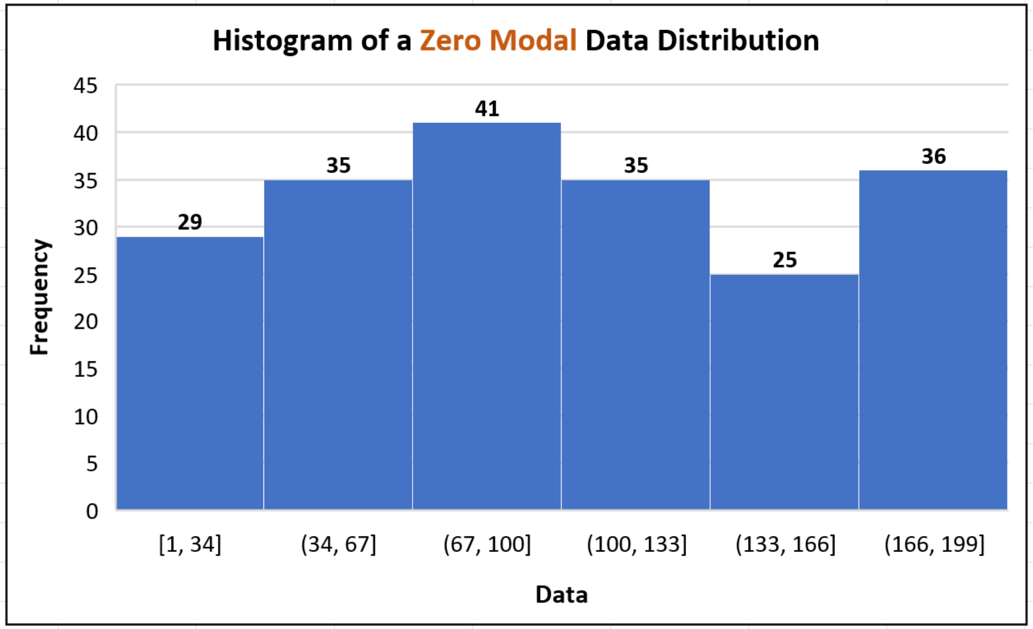


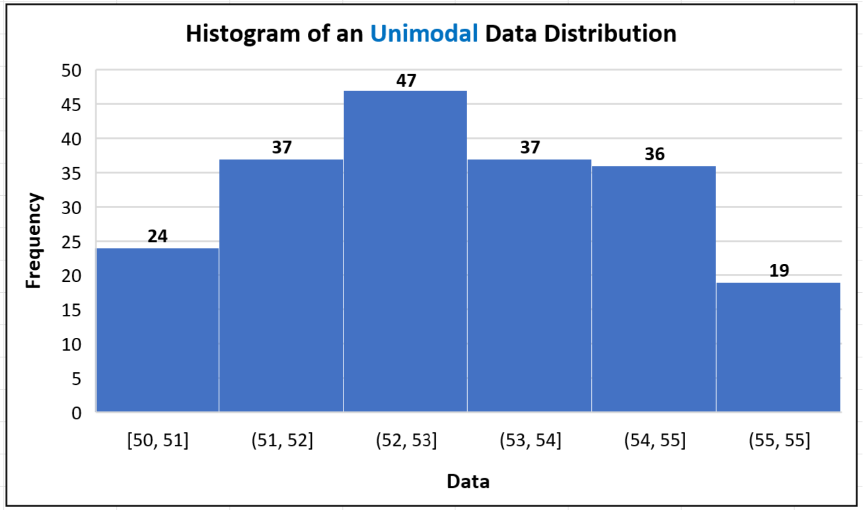


DESCRIPTIVE STATISTICS

MODE (*cont.*):



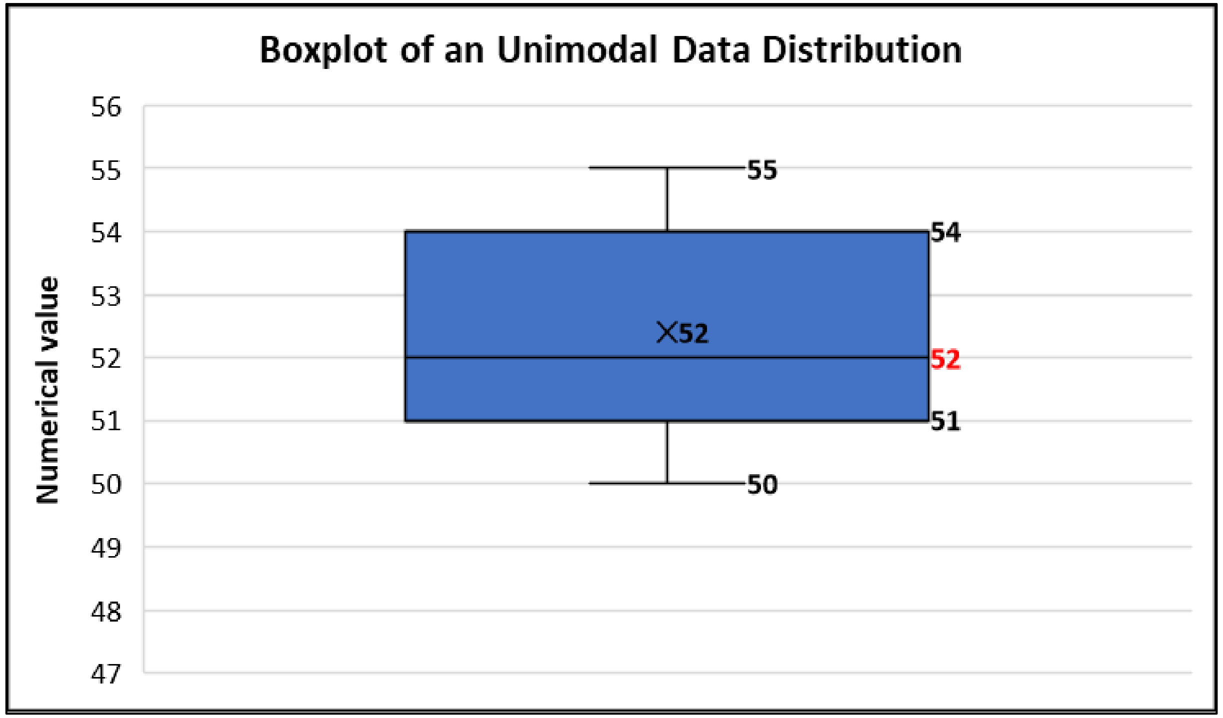


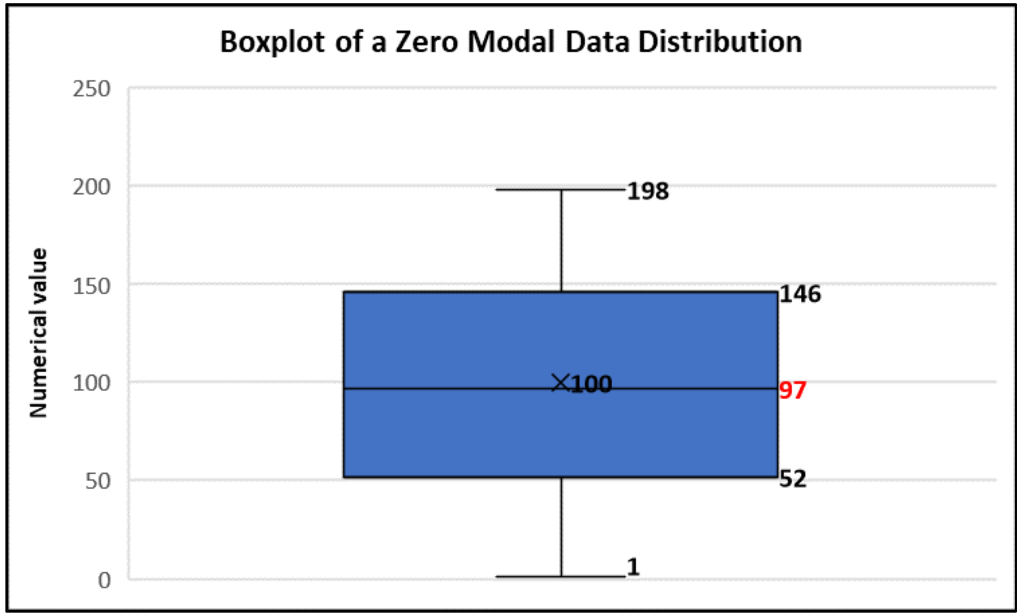


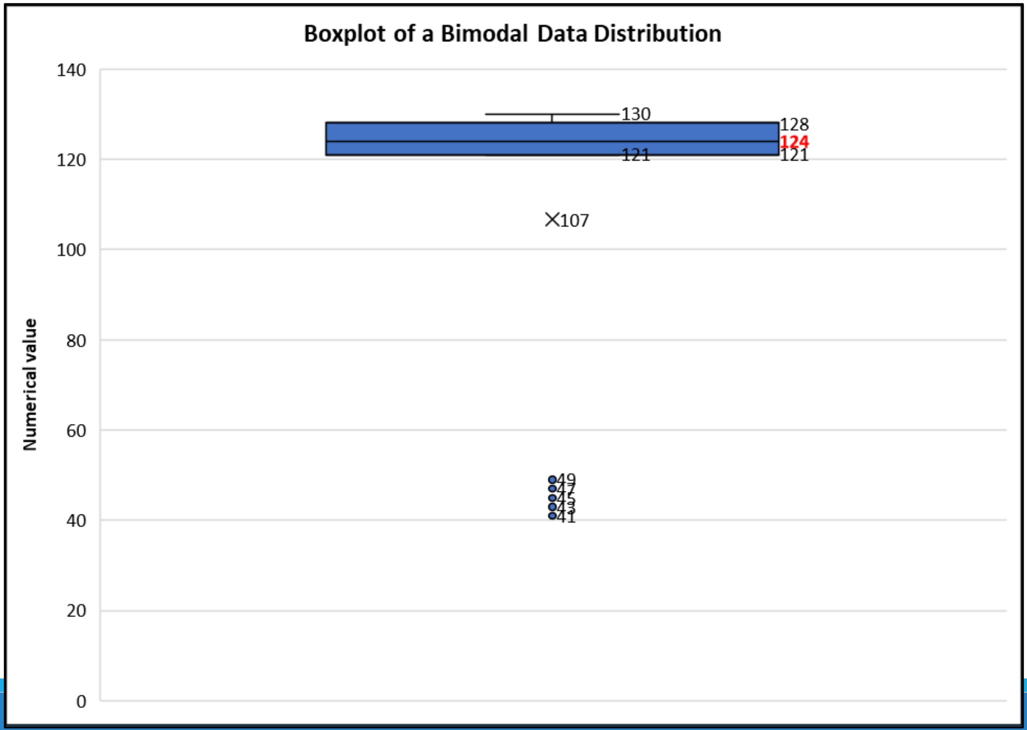
A golden rule: use multiple data visualization tools!

DESCRIPTIVE STATISTICS

MEDIAN : the middle point in a dataset, **=MEDIAN()**







DESCRIPTIVE STATISTICS

- The **ALGEBRAIC** (or **ANALYTICAL**) **MEANS** are generally defined by the formula:

$$\mu^r = \left(\frac{1}{n}\sum_{i=1}^{k} x_i^r\, n_i\right)^{1/r}$$

That for *r*=1 becomes the well-known **ARITHMETIC MEAN**:

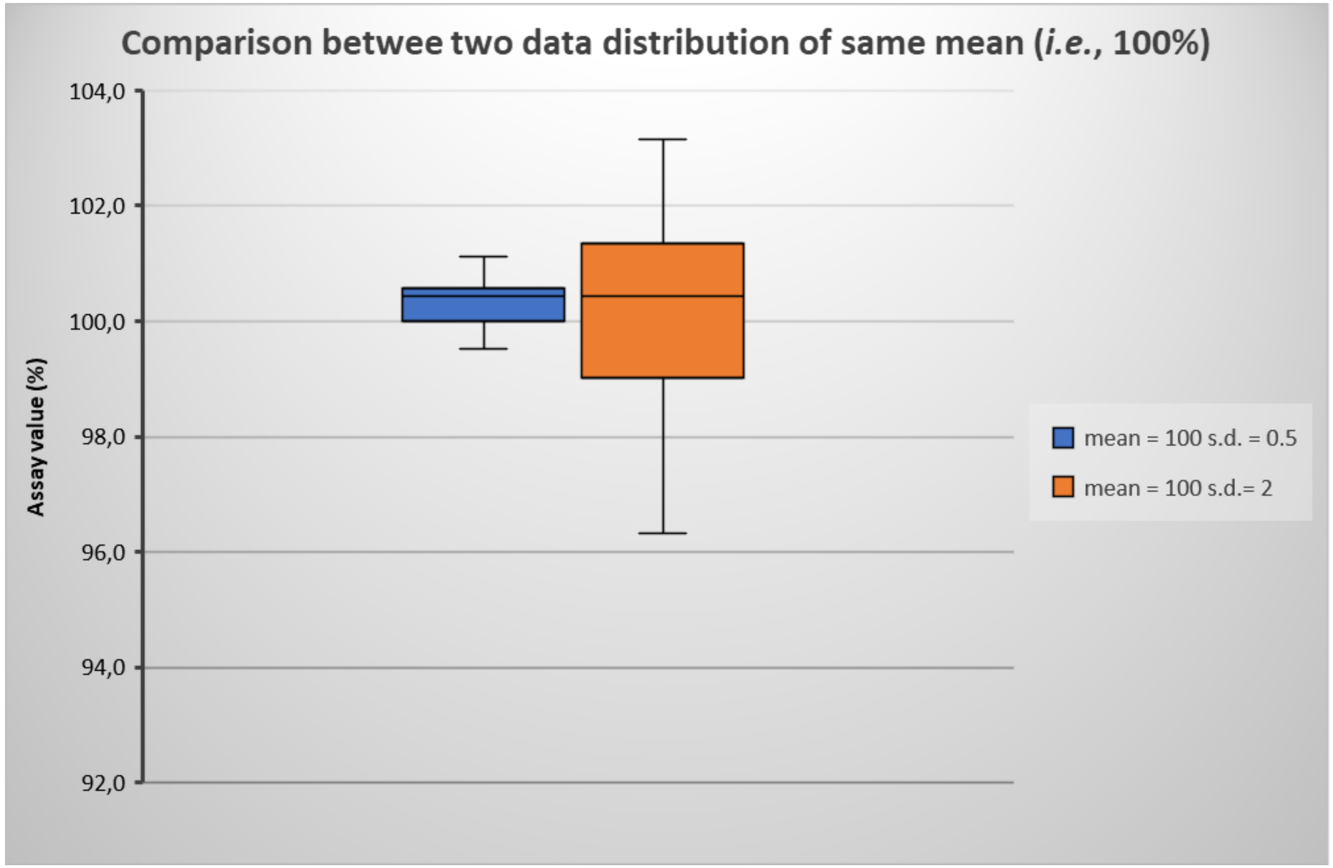
$$\mu = \frac{1}{n}\sum_{i=1}^{k} x_i\, n_i$$

e.g.: given: 3, 5, 10 the arithmetic mean is: $\mu = \frac{1}{3}\ (3 \times 1 + 5 \times 1 + 10 \times 1) = \frac{1}{3}\ (18) = 6$

DESCRIPTIVE STATISTICS

ARITHMETIC MEAN: the middle point in a dataset, **=AVERAGE()**

	mean = 100 s.d. = 0.5	mean = 100 s.d.= 2
	100,427	100,776
	99,992	99,649
	99,673	99,228
	100,267	96,328
	100,459	101,338
	100,518	101,583
	100,747	101,418
	101,116	101,269
	100,435	99,972
	100,560	100,431
	100,038	97,927
	99,527	99,010
	100,698	98,280
	99,966	101,136
	100,485	103,159
Mean	100,327	100,100



Position Indices alone are insufficient to fully describe a given distribution of data!

DESCRIPTIVE STATISTICS

The previous slides have actually introduced the need for a second type of summary indexes, namely:

VARIABILITY INDICES

whose purpose is to measure variability!

A common feature of the variability indices is that of being zero in the absence of variability and growing in value as the variability increases!

DESCRIPTIVE STATISTICS

The most widely used *« dispersion indices with respect to a center »* (*i.e.*, the arithmetic mean) are:

- ***Range***
- ***Variance***
- ***Standard Deviation***
- ***Coefficient of Variation***

DESCRIPTIVE STATISTICS

- **Range**
  - It is the simplest dispersion index.
  - It is equal to the maximum value minus the minimum value.

Range = **Maximum age — Minimum age** = 57 – 27 = **30**

DESCRIPTIVE STATISTICS

- **Standard Deviation** – measures the degree of dispersion of a dataset relative to the arithmetic mean.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

where: "n" is the number of elements forming the dataset

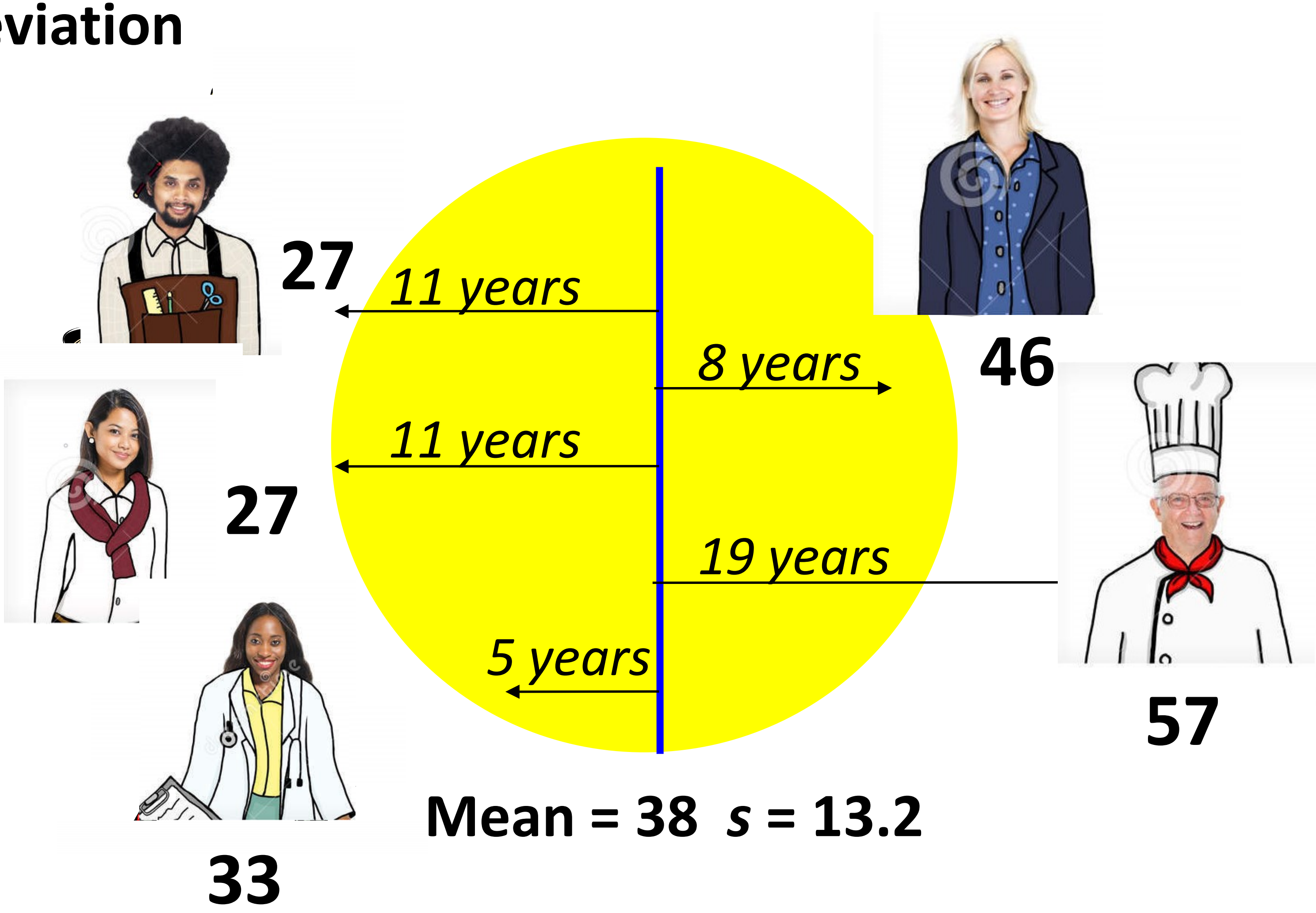
"X_i" is the value of each observation in the dataset

"$\bar{X}$" is the mean value of all observations forming the dataset

- ***The standard deviation has the same units of measurement as the variable under study !***

DESCRIPTIVE STATISTICS

- **Standard Deviation**



Mean = 38 s = 13.2

DESCRIPTIVE STATISTICS

- While ***s*** refers to the sample, ***σ*** refers to the population.

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n - 1}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n}}$$

- The reason for the difference between the two denominators is simply that if you divided by *n*, the standard deviation (or variance) of the sample would underestimate the standard deviation (or variance) of the population. That is, it would be a « *distorted statistic* ».

DESCRIPTIVE STATISTICS

- **Variance** – is the square of standard deviation.

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n - 1}$$

Where “n” is the number of the samples.

“X_i” is the value of each observation.

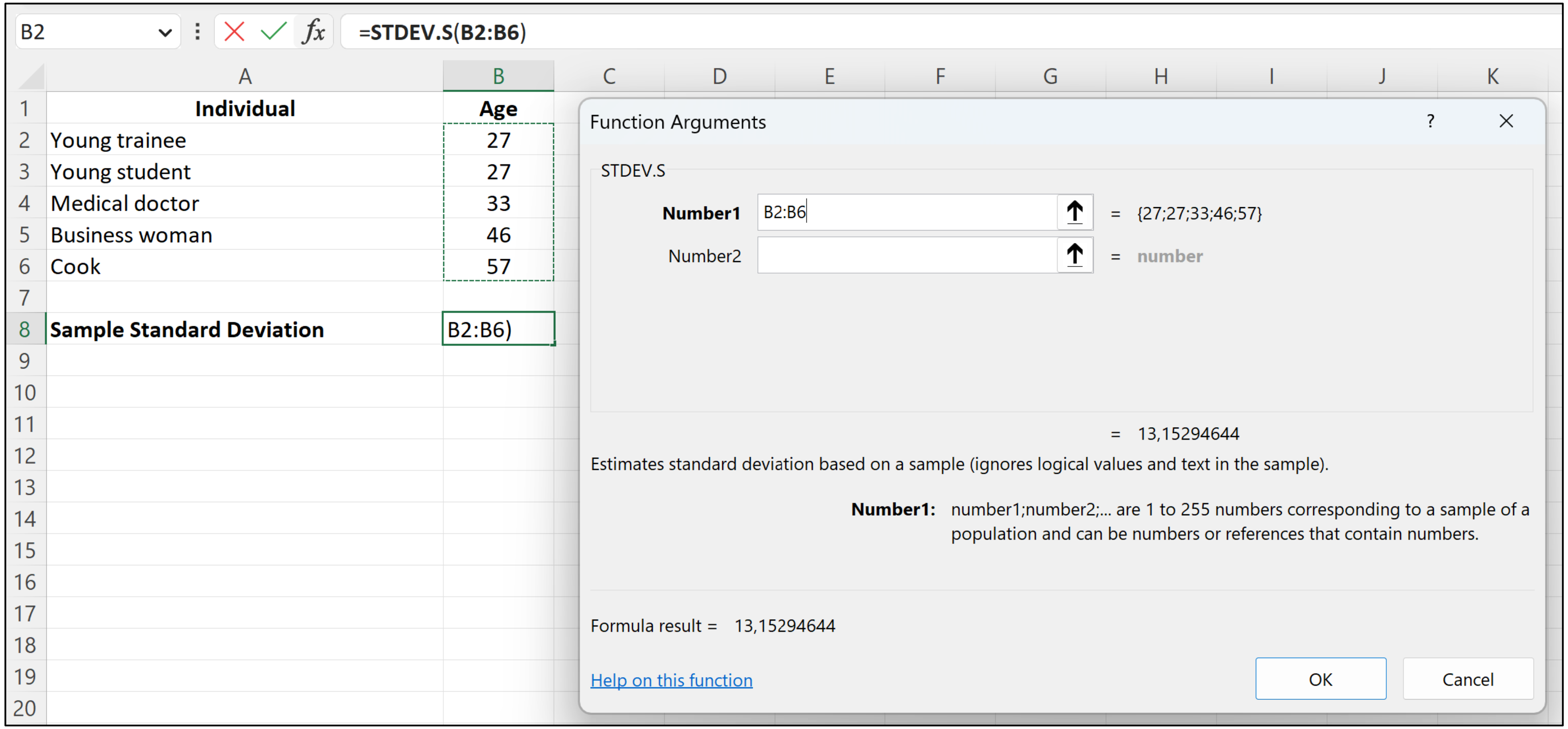
“$\bar{X}$” is the mean value of all the samples.

DESCRIPTIVE STATISTICS

Range Calculation

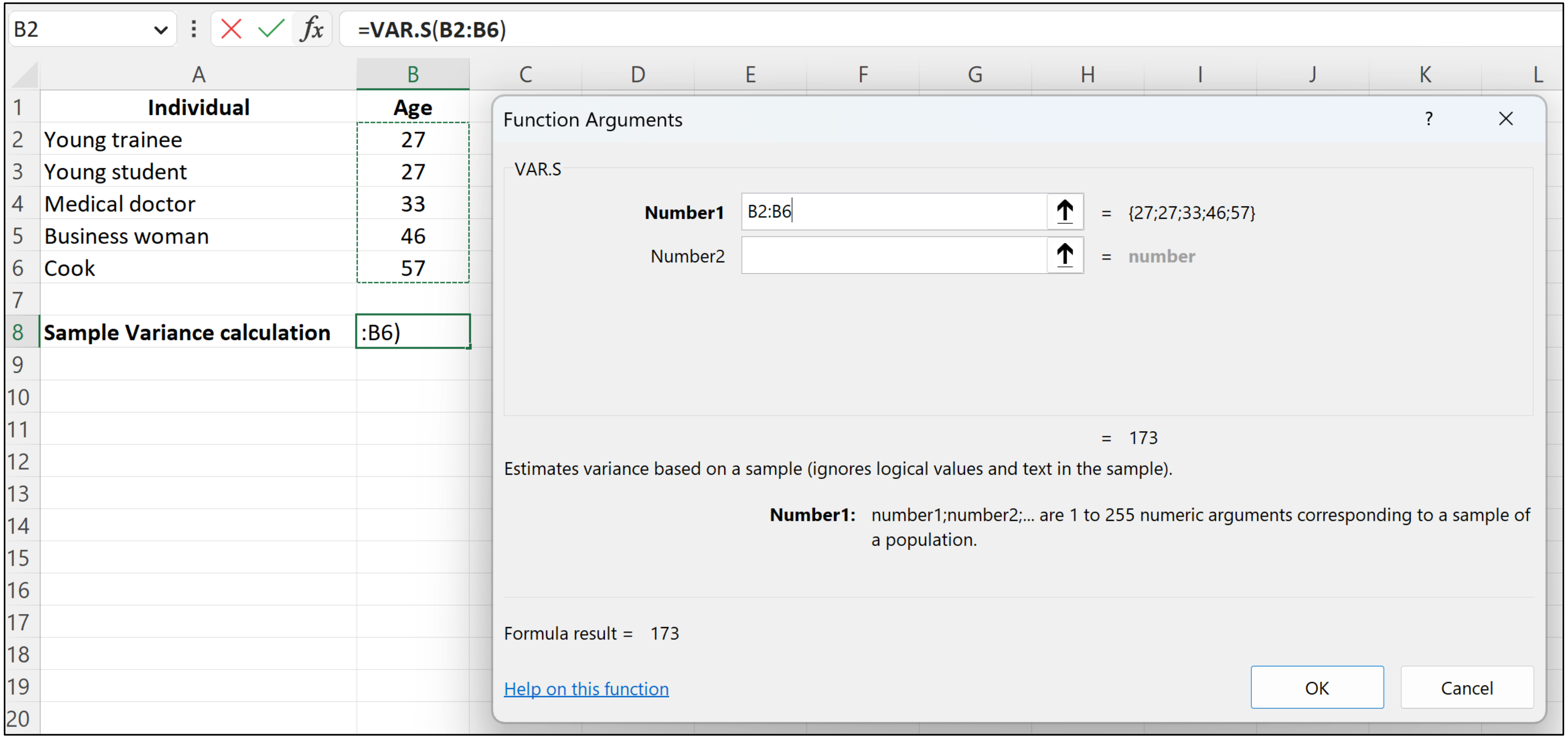
A	B
Individual	**Age**
Young trainee	27
Young student	27
Medical doctor	33
Business woman	46
Cook	57
Max	57
Min	27
Range = Max-Min	

Sample Standard Deviation Calculation



DESCRIPTIVE STATISTICS

Sample Variance Calculation



DESCRIPTIVE STATISTICS

The variance, unlike the standard deviation, has the ***property of additivity***. This means that if the elementary data form subgroups, then the total variance can be obtained as the sum of the variance "within groups" and the "variance between groups":

$$\sigma^2 = \sigma^2_{Within} + \sigma^2_{Between}$$

This « variance decomposition theorem » is the basis of the so-called

Analysis of Variance or ANOVA

DESCRIPTIVE STATISTICS

- The « between variance », $\boldsymbol{\sigma^2_{Between}}$, or « variance of group means », measures how different the group means are from each other.

- The « within variance », $\boldsymbol{\sigma^2_{Within}}$, or « mean of group variances », provides a summary of the level of variability present within each data group.

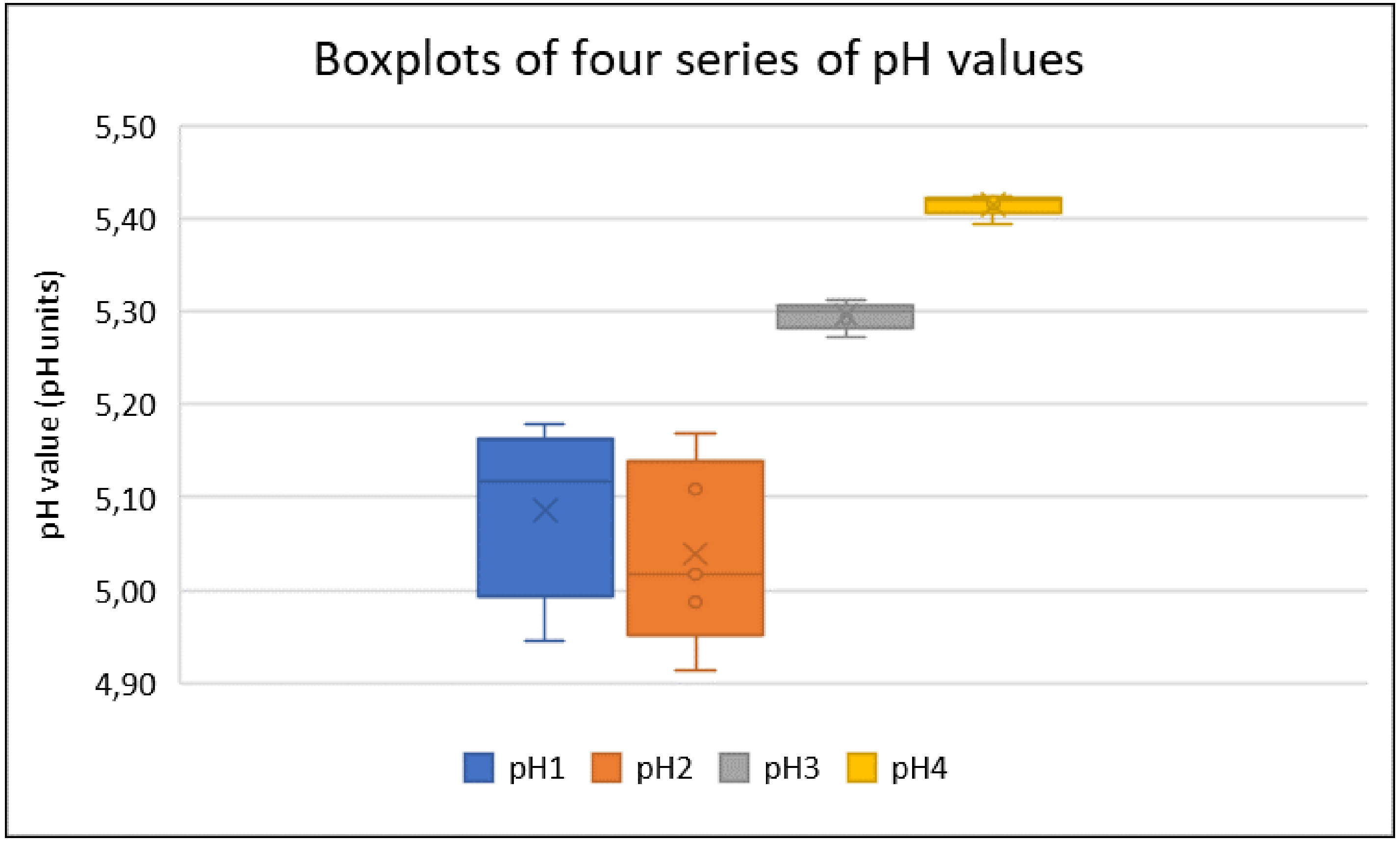
- In applying these criteria to regression analysis using the Least Squares Method, the $\boldsymbol{\sigma^2_{Between}}$ is called the ***explained variance*** while the $\boldsymbol{\sigma^2_{Within}}$ is called the ***residual variance***.

DESCRIPTIVE STATISTICS

Example: Let's consider the four series of pH values below which, at first glance, look quite similar ...

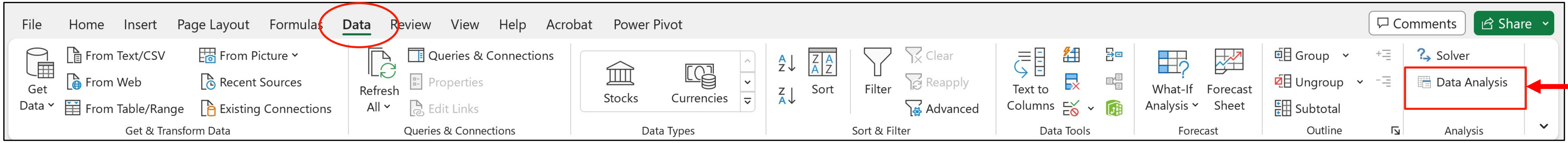
What can we say?

pH1	pH2	pH3	pH4
5,12	5,02	5,27	5,42
4,94	5,17	5,29	5,42
5,18	5,11	5,30	5,42
5,04	4,91	5,31	5,39
5,15	4,99	5,30	5,42

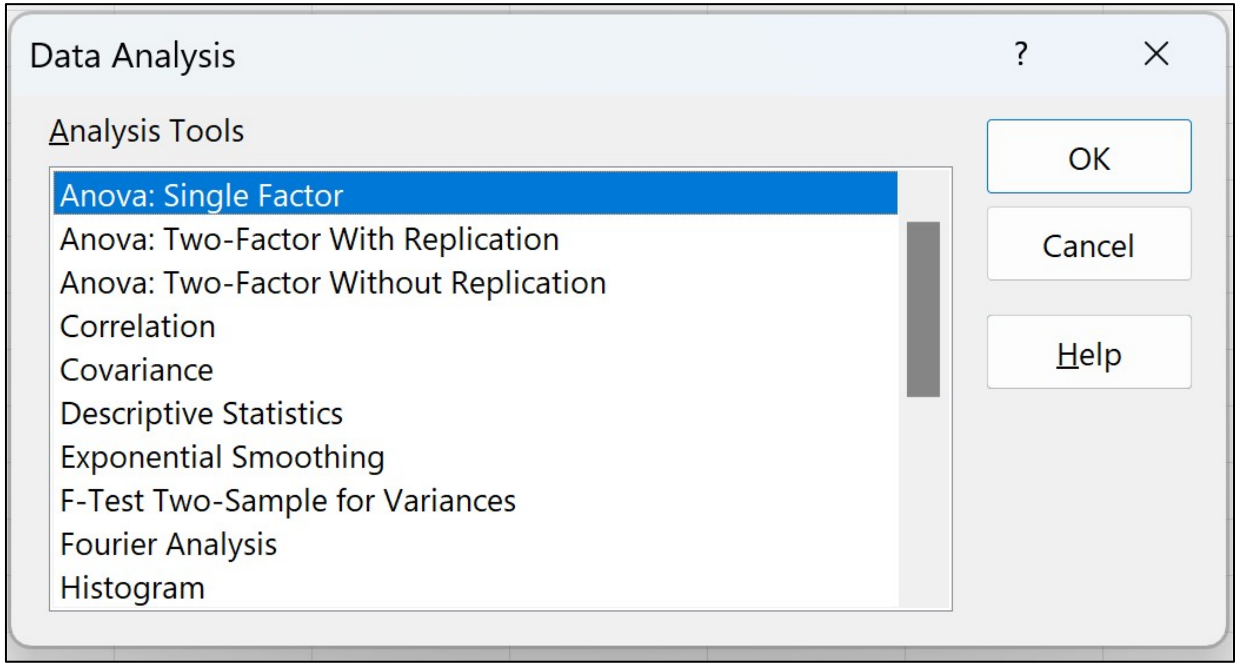


DESCRIPTIVE STATISTICS

Using the “Data Analysis” tool shown here below running ANOVA One-Way is very simple.



- Clicking on “Data Analysis” is displayed the menu shown here on the side where we have to click on “Anova: Single Factor”



DESCRIPTIVE STATISTICS

Let's see ANOVA One-Way (or One factor) results:

Groups	Count	Sum	Mean	Variance
pH1	5	25,43	5,09	0,0087
pH2	5	25,19	5,04	0,0101
pH3	5	26,48	5,30	0,0002
pH4	5	27,07	5,41	0,0001

ANOVA						
Source of variation	**Sum of Squares**	**dof**	**Mean of Squares**	**F *calculated***	**Significance value**	**F *tabulated***
Between groups	0,4690	3	**0,1563**	**32,5928**	0,0000	**3,2389**
Within groups	0,0767	16	**0,0048**			
Total	0,5458	19				

What does ANOVA One-Way tell us?

- The means of squares (or variances) are greater between data groups than within them. In other words:

variability (measured by the deviation from the mean) ***is higher between groups than within them!***

F calculated > *F tabulated* : average values of the data groups are significantly different from each other

Consider that this result is what is normally obtained by comparing series of data, such as happens for example for APQR.

ANOVA possible applications?

Comparison of multiple data series such as:

- *Yields of different lots obtained using the same process or different processes*
- *Assay values of lots listed in an Annual Product Quality Review*
- *Impact of different catalyst on chemical reaction rates*
- *Impact of fertilizer type, planting density and planting location in the field on final crop yield*
- *etc.*

DESCRIPTIVE STATISTICS

A very important and useful index of variability is the **Coefficient of Variation** which is defined as:

$$\mathrm{CV} = \mathrm{RSD} = \frac{\sigma}{\mu} \quad \text{or} \quad \mathrm{CV\%} = \mathrm{RSD\%} = \frac{\sigma}{\mu} \times 100$$

The usefulness of this index derives from the fact that it allows you to ***compare the variability*** of two different distributions of data!

This characteristic is very important if you think about how often the problem arises of comparing, for example, the variability in the yields of two processes (or of the same process but conducted in different conditions / places) or the variability of two machines, etc.

DESCRIPTIVE STATISTICS

Let’s consider, for example, the four series of pH values we have just examined using ANOVA.

CV% can be easily calculated from ANOVA’s SUMMARY adding two columns: Standard Deviation and CV% as follows:

Standard Deviation values can be obtained from corresponding Variance values just using function: SQRT(Variance)

CV% values can be calculated using the corresponding Standard Deviation and Mean values.

Anova: Single Factor

SUMMARY

Groups	***Count***	***Sum***	***Mean***	***Variance***	***Standard Deviation***	***CV%***
pH1	5	25,43	5,09	0,0087	0,0934	**1,84**
pH2	5	25,19	5,04	0,0101	0,1006	**2,00**
pH3	5	26,48	5,30	0,0002	0,0146	0,28
pH4	5	27,07	5,41	0,0001	0,0118	0,22

CV% values reflect boxplots of slide 45 !

DESCRIPTIVE STATISTICS

The third type of indices are the:

SHAPE INDICES

- In general terms it can be said that if the **Averages** give an idea of the order of magnitude of the data series, the **Variability Indices** measure the difference between the values and the **Shape Indices** describe the distancing of the data distribution from the symmetrical form (or bell).

DESCRIPTIVE STATISTICS

- **FISHER or SKEWNESS ASYMMETRY INDEX**: it is a shape index that allows to evaluate the degree of deviation of a distribution with respect to a perfectly symmetrical trend.

$$\gamma_1 = \frac{1}{\sigma^3}\left[\frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^3 \, n_i\right]$$

if $\gamma_1 > 0$: positive asymmetry or *right tail* (Mode < Median < Mean)

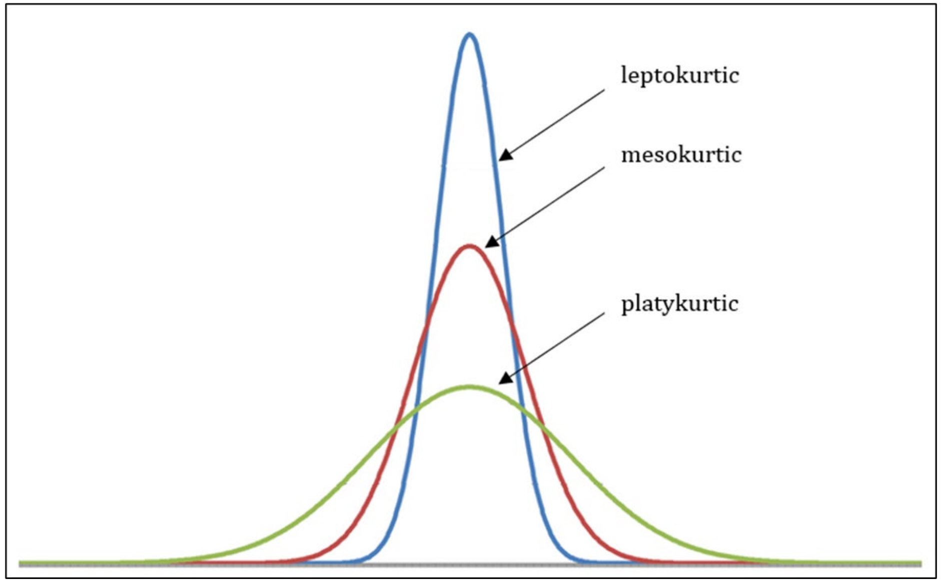
if $\gamma_1 < 0$: negative asymmetry or *left tail* (Mean < Median < Mode)

if $\gamma_1 = 0$: it's just a ***symptom*** of symmetry (Mean = Median = Mode)

DESCRIPTIVE STATISTICS

- **KURTOSIS:** is a shape index that allows you to evaluate the degree of flattening of a distribution around its central value.

$$\gamma_2 = \frac{1}{\sigma^4}\left[\frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^4 n_i\right]$$



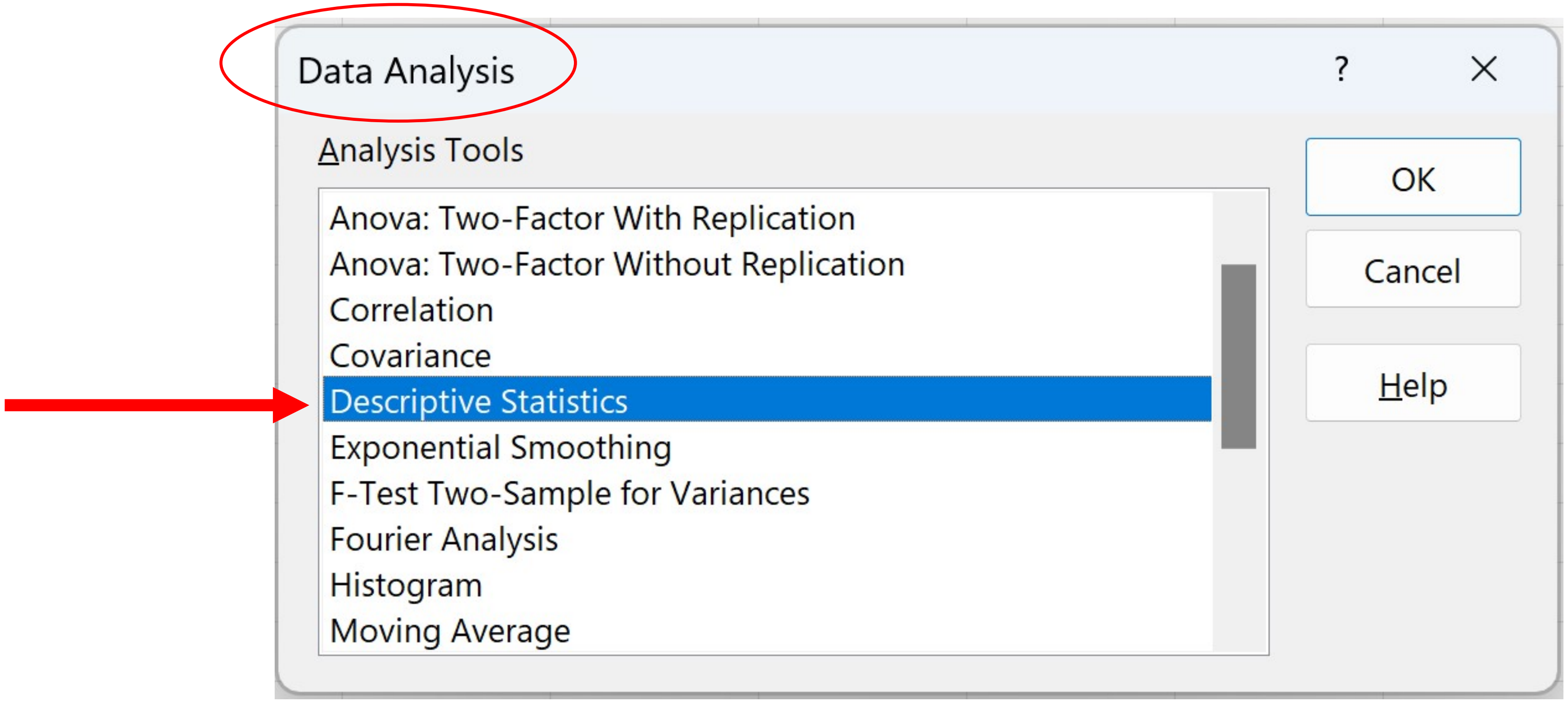
if $\gamma_2 > 3$: ***leptokurtic*** curve (pointed)

if $\gamma_2 = 3$: ***mesokurtic*** or ***normokurtic*** curve (or *Gaussian*)

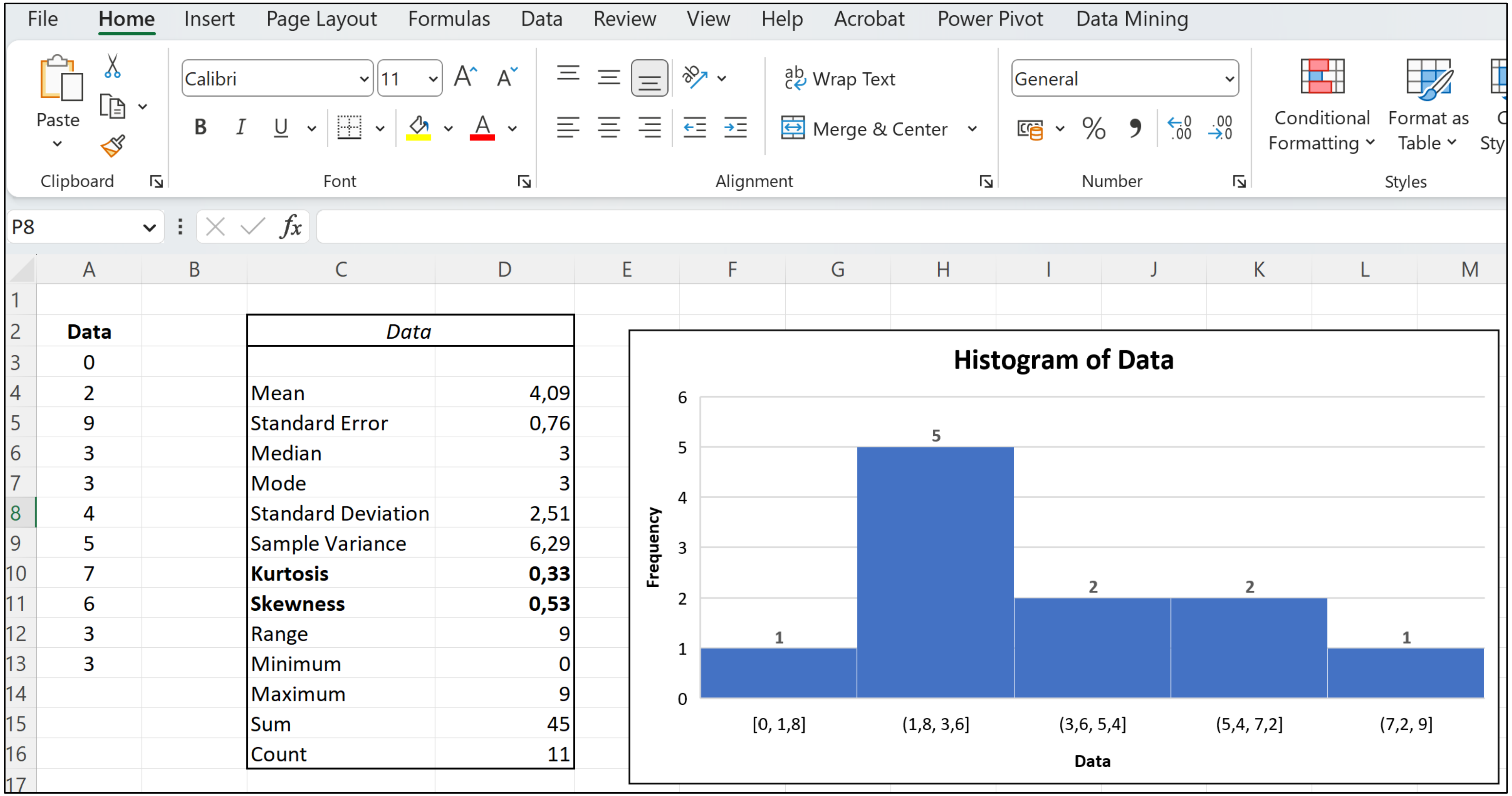
if $\gamma_2 < 3$: ***platikurtic*** curve (flattened)

DESCRIPTIVE STATISTICS

For Shape Indices Excel® provides specific functions, **SKEW()** and **KURT()**, or, alternatively, you can use the ***Data Analysis Tool:***



DESCRIPTIVE STATISTICS



Data
0
2
9
3
3
4
5
7
6
3
3

Data	
Mean	4,09
Standard Error	0,76
Median	3
Mode	3
Standard Deviation	2,51
Sample Variance	6,29
Kurtosis	**0,33**
Skewness	**0,53**
Range	9
Minimum	0
Maximum	9
Sum	45
Count	11

INFERENTIAL STATISTICS WITH EXCEL®

INFERENTIAL STATISTICS

- Is that part of the Statistics that aims to make operational decisions and choices based on limited and provisional information.

- It can be summarized as : ***FROM FEW TO ALL*** and is based on a process known as:

INFERENCE

i.e., the process of reaching a conclusion from a given set of statements (or *premises*)

- This process can be of two types: **deductive** and **inductive**

INFERENTIAL STATISTICS

- **Example 1:** **Deductive Argument** (from general to the particular)

  Premises: Socrates is a man
  All men are mortal
  Conclusion: Socrates is mortal **VALID ARGUMENT**

- **Example 2:** **Inductive Argument** (from particular to the general)

  Premises: Last September was the rainiest on record
  John's birthday is in September
  Conclusion: It rained on John's last birthday **PLAUSIBLE ARGUMENT**

The basic problem in inductive inference is
to devise ways of measuring the strength of an inductive argument!

INFERENTIAL STATISTICS

- To achieve this goal, Inferential Statistical makes use of two methodologies :
  - **Parameter Estimation** and
  - **Hypothesis Testing**

INFERENTIAL STATISTICS

To do this work, some concepts are needed the most important of which is that of

Probability and ***Probability Distribution***

WHY?

Simple ! Probability distributions, especially the "parametric" ones, are mathematical laws that represent real « reference models ».

Once demonstrated that one of them can adequately describe the behavior of the data under analysis, it becomes immediate to make extrapolations with respect to such data.

INFERENTIAL STATISTICS

According to the its ***classical definition*** (Laplace), **Probability** can be calculated dividing the number of successful times (or ways) an event occurs by the total number of possible outcomes if each outcome is equally likely.

$$P(E) = \frac{Number\ of\ ways\ E\ can\ successfully\ occur}{Total\ number\ of\ possible\ outcomes\ of\ the\ experiment} \quad (1)$$

The term ***event*** identifies any possible outcome of an experiment.

An event can be ***simple*** if it consists of just one outcome (*e.g.*, tossing a coin or a dice) or ***compound*** if it contains more than one outcome (*e.g.*, tossing a coin and a dice).

INFERENTIAL STATISTICS

- The probability value is therefore a number between 0 and 1.
- The value 0 indicates an impossible event while the value 1 indicates a certain event.
- Rolling a dice, the probability that the number "4" will come out is 1/6 since there are 6 possible events (as many as there are faces of the die) and the favorable event is only one.

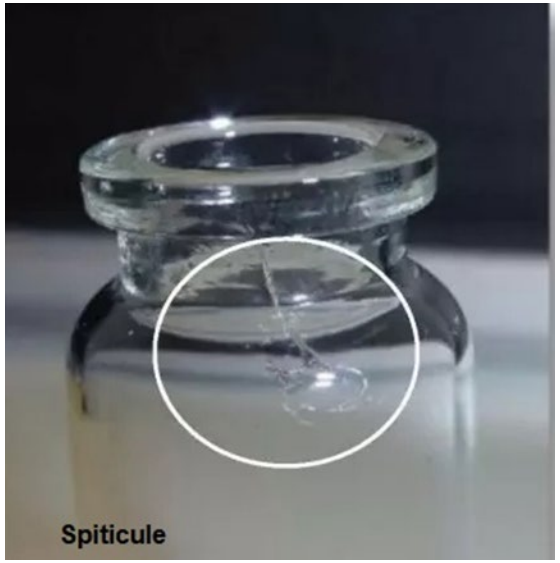
INFERENTIAL STATISTICS

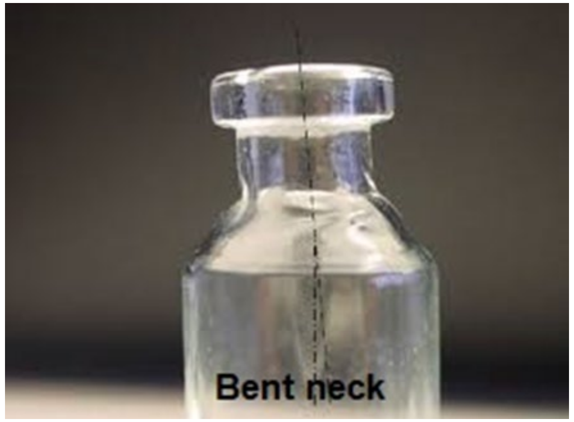
Let's consider, for example, a few types of defects that could occur in glass vials:

Class	Location	Defect type	Description
Critical	General	Crack	Fracture that penetrates completely through the glass wall.
		Spiticule	Bead or string of glass that is adhered to the interior surface.
	Finish	Broken Finish	A finish that has actual pieces of glass broken out of it
Major	Body	Ring off	A container that has separated into two pieces
	Finish	Bent neck	The finish of the container is distorted to the extent that the plane of the seal surface is not perpendicular to axis of the body
	General	Check	A discontinuity in the glass surface that does not penetrate through the glass wall
		Chipped	Container with a section or fragment broken out (other than sealing surface)
	Finish/Neck	Crizzle	A finish or neck that has several fine surface marks
….	….	…	…

INFERENTIAL STATISTICS

and assume that in a 1000000 clear glass vials batch, 30000 are flawed because of *cracks*, 10000 are flawed because of *spiticules*, 20000 are flawed because of *bent neck* and 40000 are *yellow colored*.





INFERENTIAL STATISTICS

Let assume, for simplicity, that these defects are *mutually exclusive* and that the probability of observing any one of these events for a single vial is:

Casual variable	Possible Outcomes	Probability
Glass vial defect	Crack	0.03
	Spiticule	0.01
	Bent neck	0.02
	Yellow color	0.04

INFERENTIAL STATISTICS

The probability of choosing at random an unacceptable vial (*i.e.*, *cracked*, *spiticuled*, *bent necked* or *yellow colored*) is: 0.03+0.01+0.02+0.04 = 0.10 or 10%

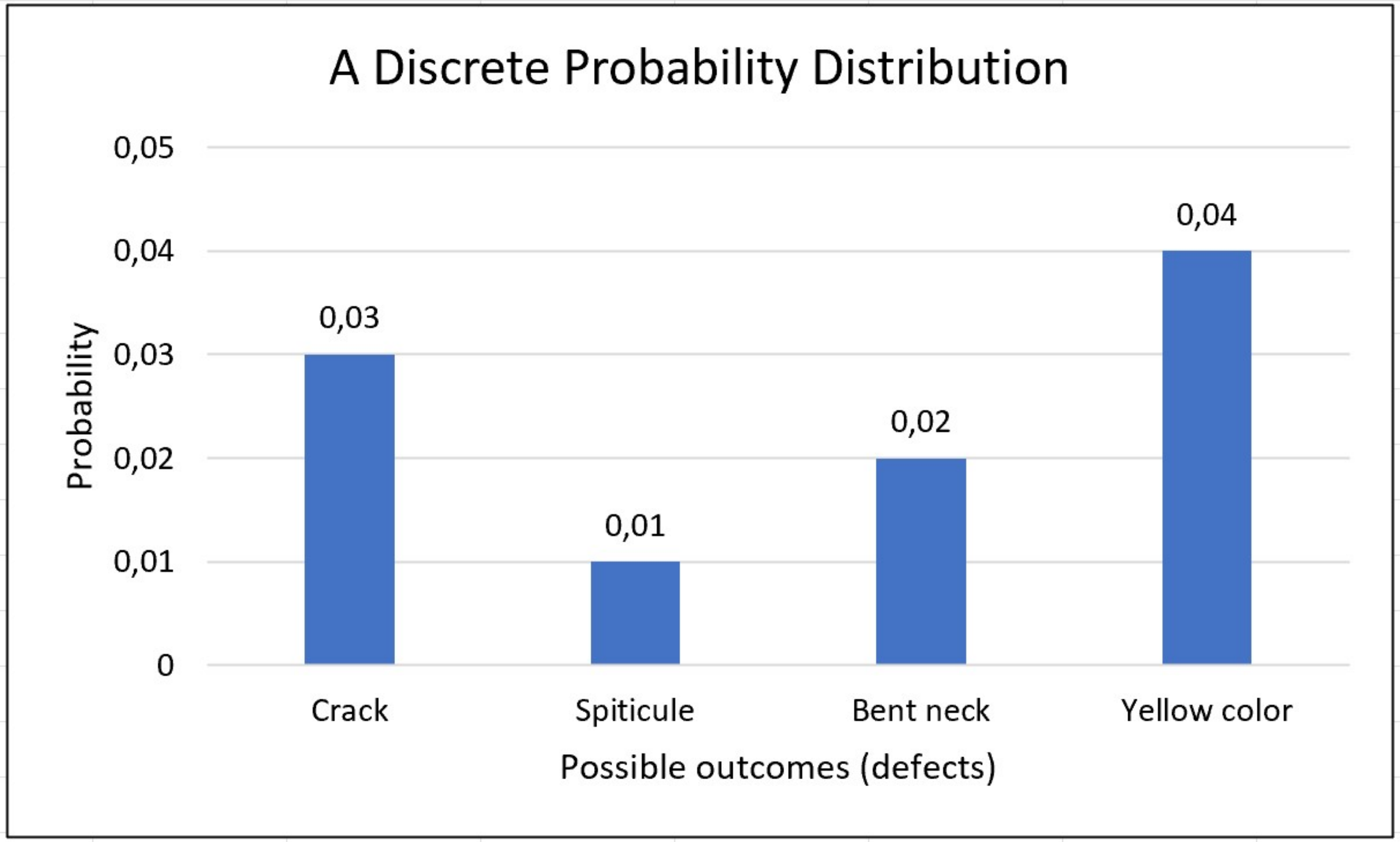
Consequently, the probability of choosing at random an acceptable vial is:

$$1 - 0.10 = 0.90 \text{ or } 90\%$$

The four outcomes listed in the table and their associate probability values form a *sample probability distribution* which can be graphically represented as:

INFERENTIAL STATISTICS

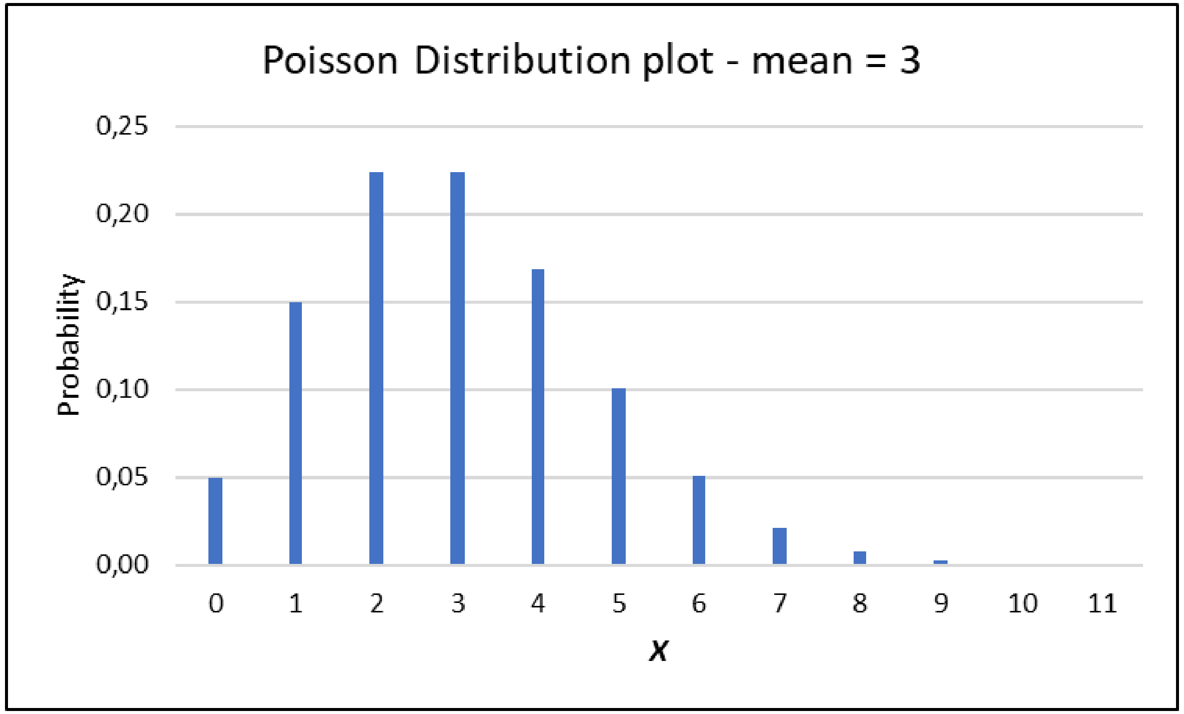
Possible Outcomes	Probability
Crack	0,03
Spiticule	0,01
Bent neck	0,02
Yellow color	0,04

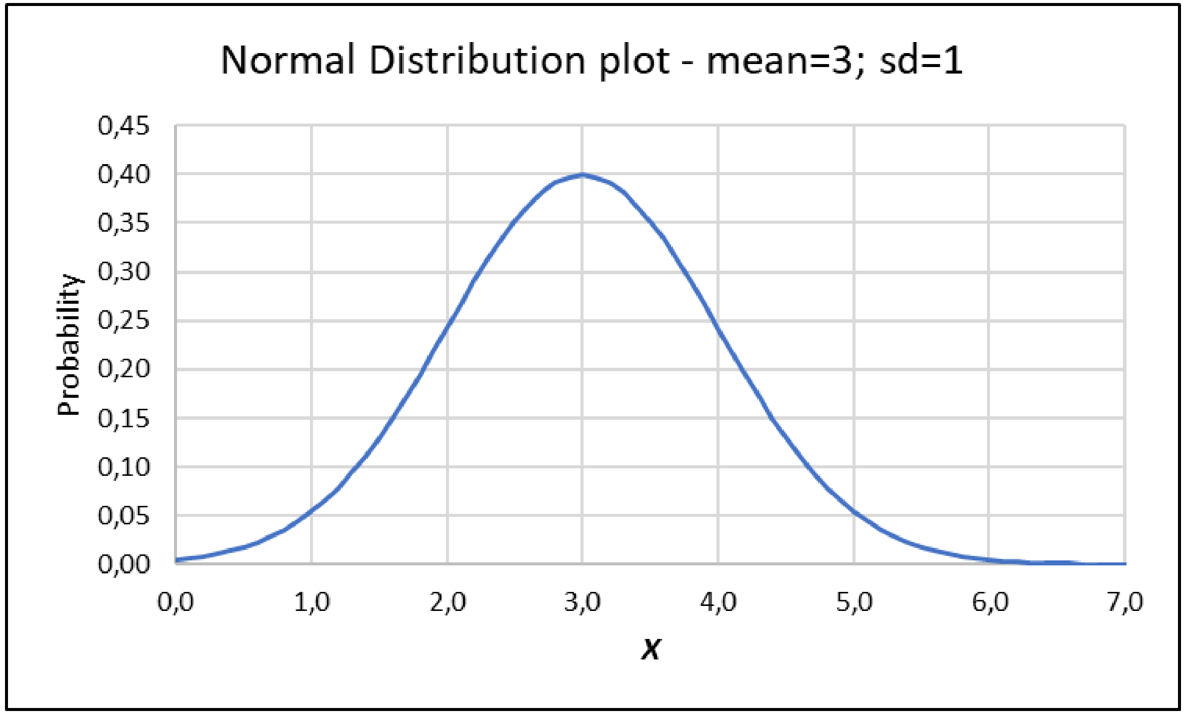


INFERENTIAL STATISTICS

- A ***distribution*** (or ***probability distribution***) is a set of values of a ***variable*** (in this case: *glass vials defects*), along with the associated probability of each value of the variable.
- Distributions are usually visualized plotting the variable on the x-axis and the probability on the y-axis
- In the example in the previous slide the distribution is ***discrete***, *i.e.*, it can assume a finite number of values.
- If, on the other hand, a random variable takes on all the values belonging to an interval (a, b) then it is called ***continuous.***

INFERENTIAL STATISTICS





Poisson Distribution

Discrete data and Discrete probability curve

Normal Distribution

Continuous data and Continuous probability curve

INFERENTIAL STATISTICS

- In general, distributions can be numerically described using three categories of parameters:
  - *central tendency (e.g., mean)*
  - *variation / spread (e.g., variance, standard deviation)*
  - *shape (e.g., skewness)*

- *The mathematical function that associates a probability value to each value assumed by the variable is called the* ***probability function (Discrete Distribution)*** *or* ***probability density function (Continuous Distribution).***

INFERENTIAL STATISTICS

The most important probability distributions belonging to these two categories are:

- ***Binomial*** and ***Poisson*** : *discrete*
- ***Normal (or Gaussian)*** : *continuous*
- ***Student's t-distribution*** : *continuous*

Let's start with Poisson's Distribution

INFERENTIAL STATISTICS

Introduced by Siméon Denis Poisson in a book he wrote regarding the application of probability theory to lawsuits (1837), it applies in diverse areas as:

- number of misprints on a page (or number of pages) in a book,
- number of people in a community living 100 years of age,
- number of wrong phone numbers dialed in a day,
- number of equipment failures in a given time period, *etc.*

Poisson's Distribution is known as the ***« distribution of rare events »***

S.M. Ross, A first course in probability – 9th Edition, Pearson College (2012)

INFERENTIAL STATISTICS

Beyond all these apparently abstract aspects, the Poisson Distribution represents a ***useful model*** for various phenomena in the pharmaceutical field such as, for example:

- ***Black particles in tablets or vials***
- ***Microbial counts***
- ***Acceptance sampling plans by attributes***
- ***etc.***

INFERENTIAL STATISTICS

Another area of application of the Poisson distribution is, for example, in the ***Acceptance Statistic Sampling.***

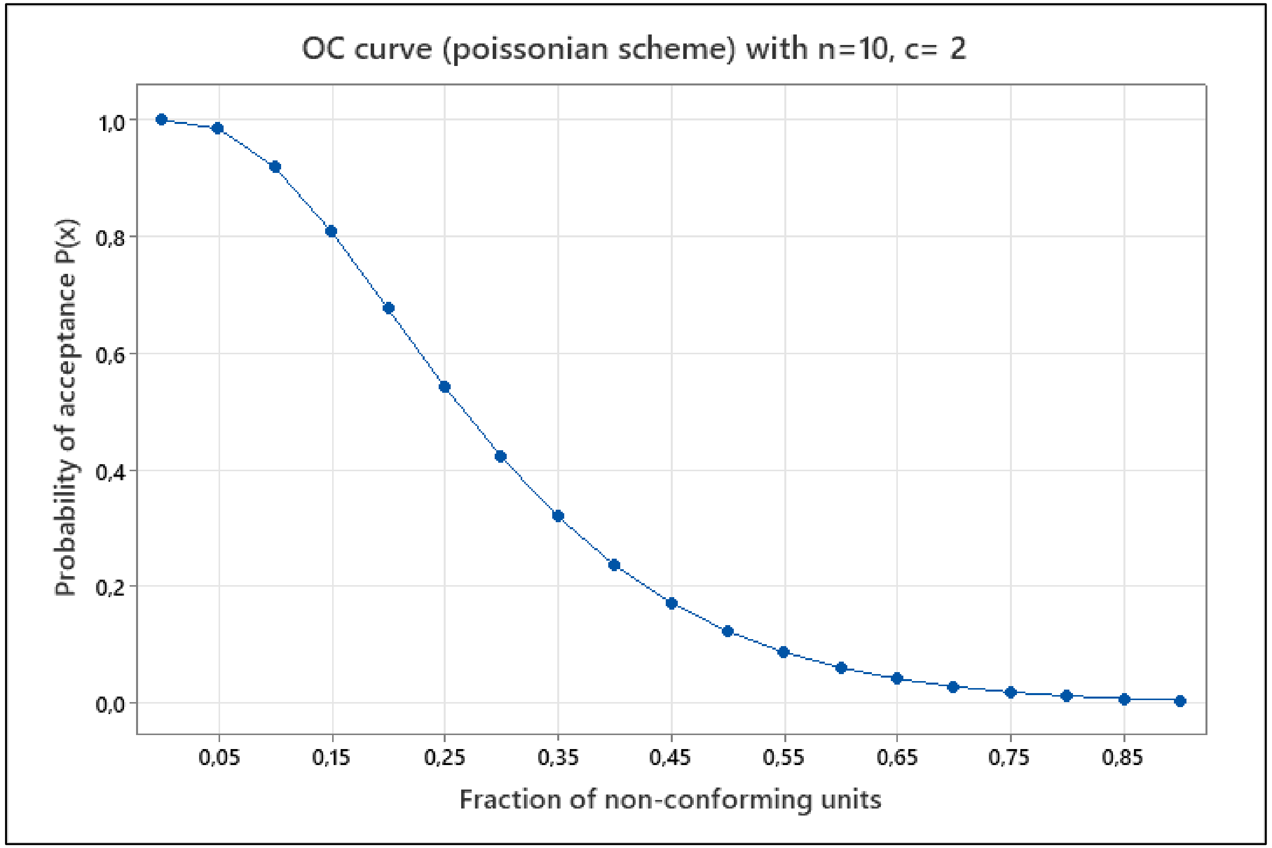
Here is an example of construction of the Characteristic Operating Curve in the Poissonian case:

$$\boldsymbol{N} = 100 \quad \boldsymbol{n} = 10 \quad \boldsymbol{c} = 2$$

$$P_a(x) = \sum_{x=0}^{2} \frac{e^{-10p} \times (10p)^x}{x!}$$

x	**0.00**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
Pa(x)	1	0.9197	0.6767	0.4232	0.2381	0.1247	0.0620	0.0296	0.0138	0.0062

INFERENTIAL STATISTICS



INFERENTIAL STATISTICS

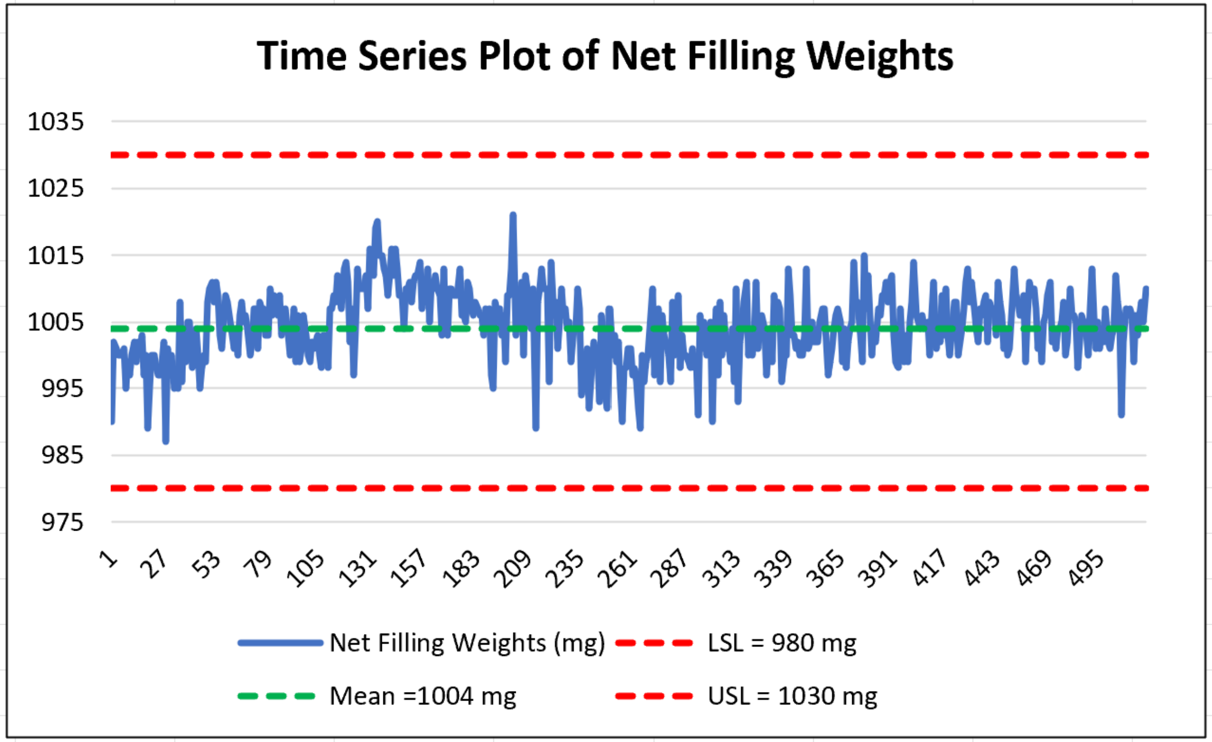
- The ***Normal Curve*** is due to the French mathematician Abraham De Moivre who mentioned it in a paper published on November 12, 1733.

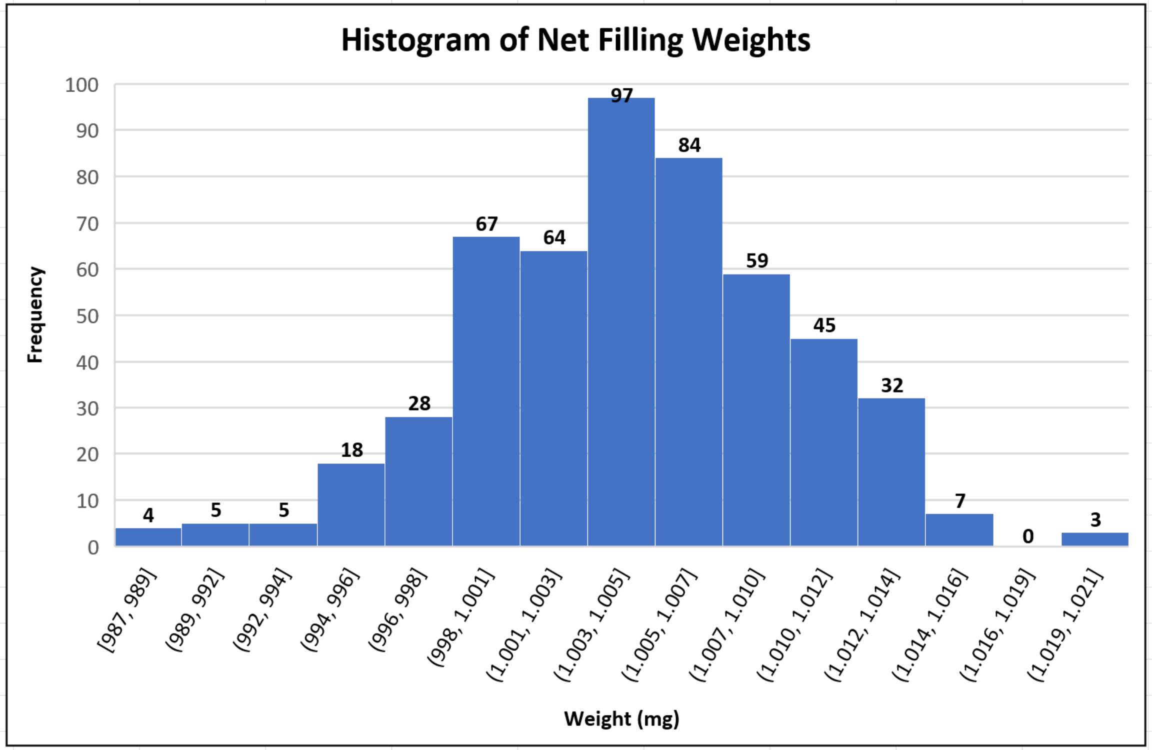
- The statistical use of the normal distribution began with Laplace and Gauss (distribution of errors) and Quételet made large use of it in Social Statistics (the *average man theory*: the individual person was synonymous with error, while the average person represented the true human being).

- This distribution was first called **normal distribution** by Sir Francis Galton in his lecture on *Typical Laws of Heredity* held at the Royal Institution on February 9, 1877.

- In the pharmaceutical field it occurs quite often. A typical example is shown in the next slide.

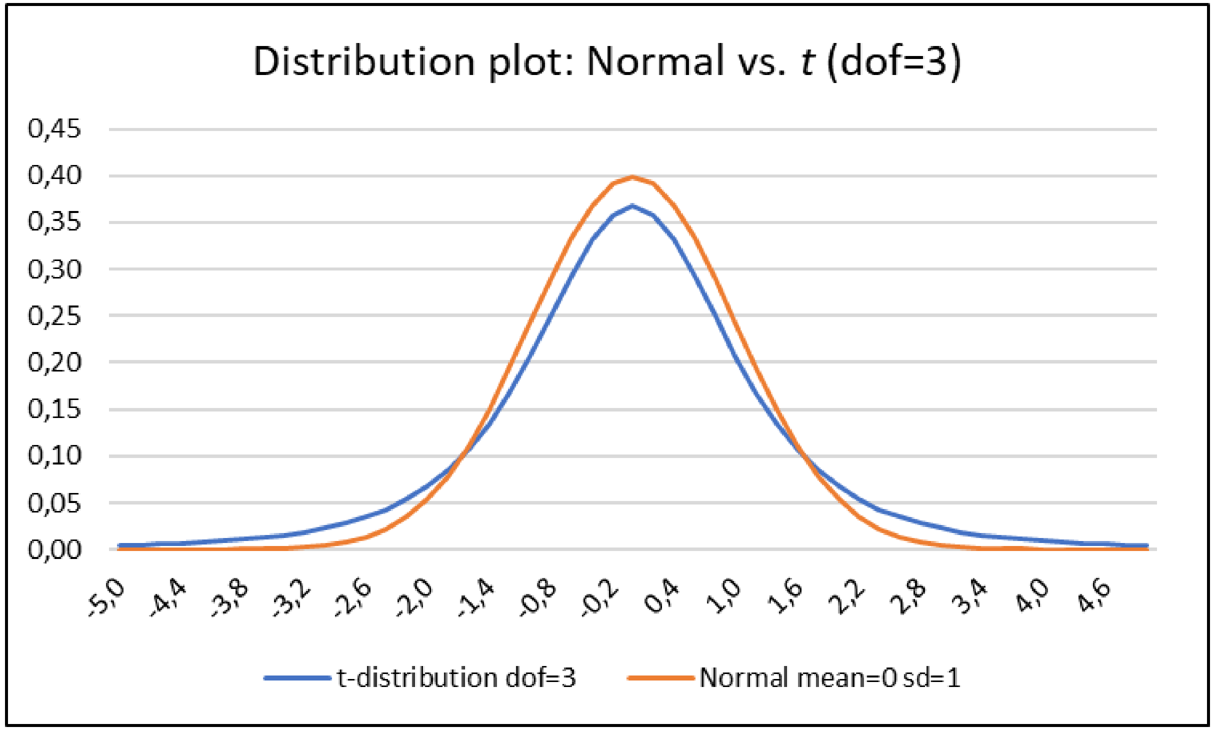
INFERENTIAL STATISTICS

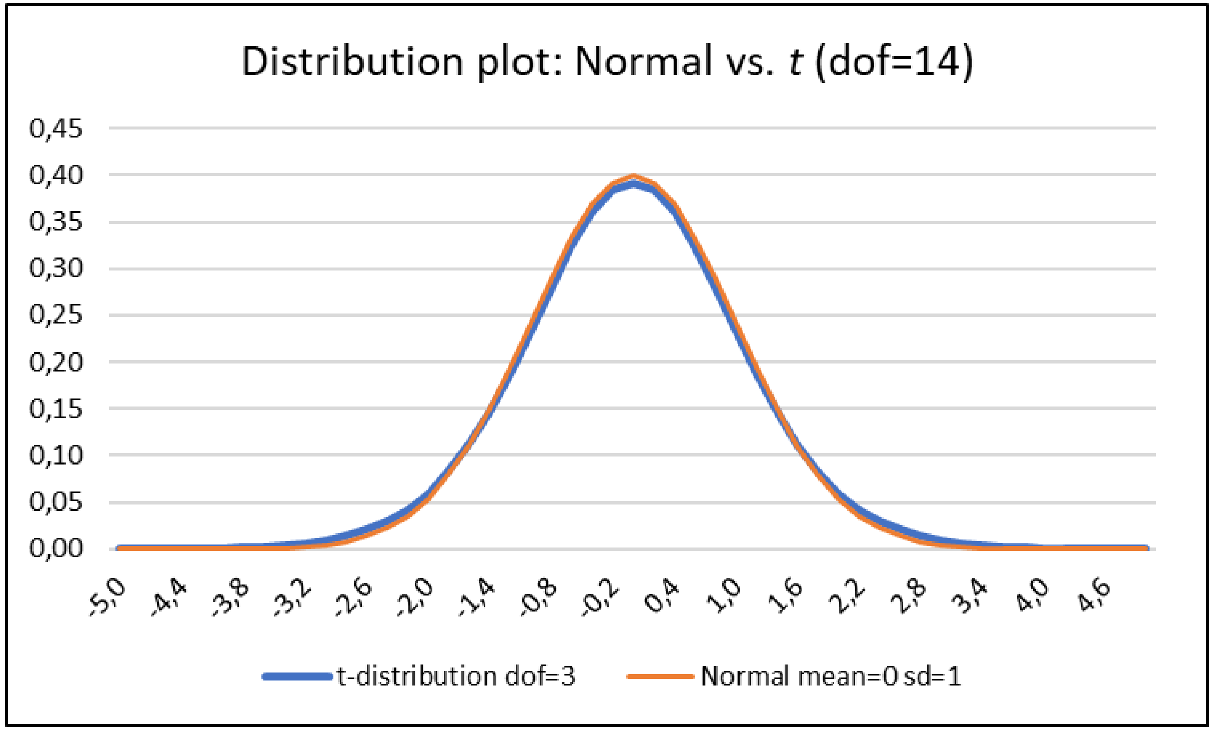




INFERENTIAL STATISTICS

Very similar to the Normal, and very useful, is the Student t-distribution or t-distribution.





INFERENTIAL STATISTICS

Normal Distribution vs. Student's t-Distribution

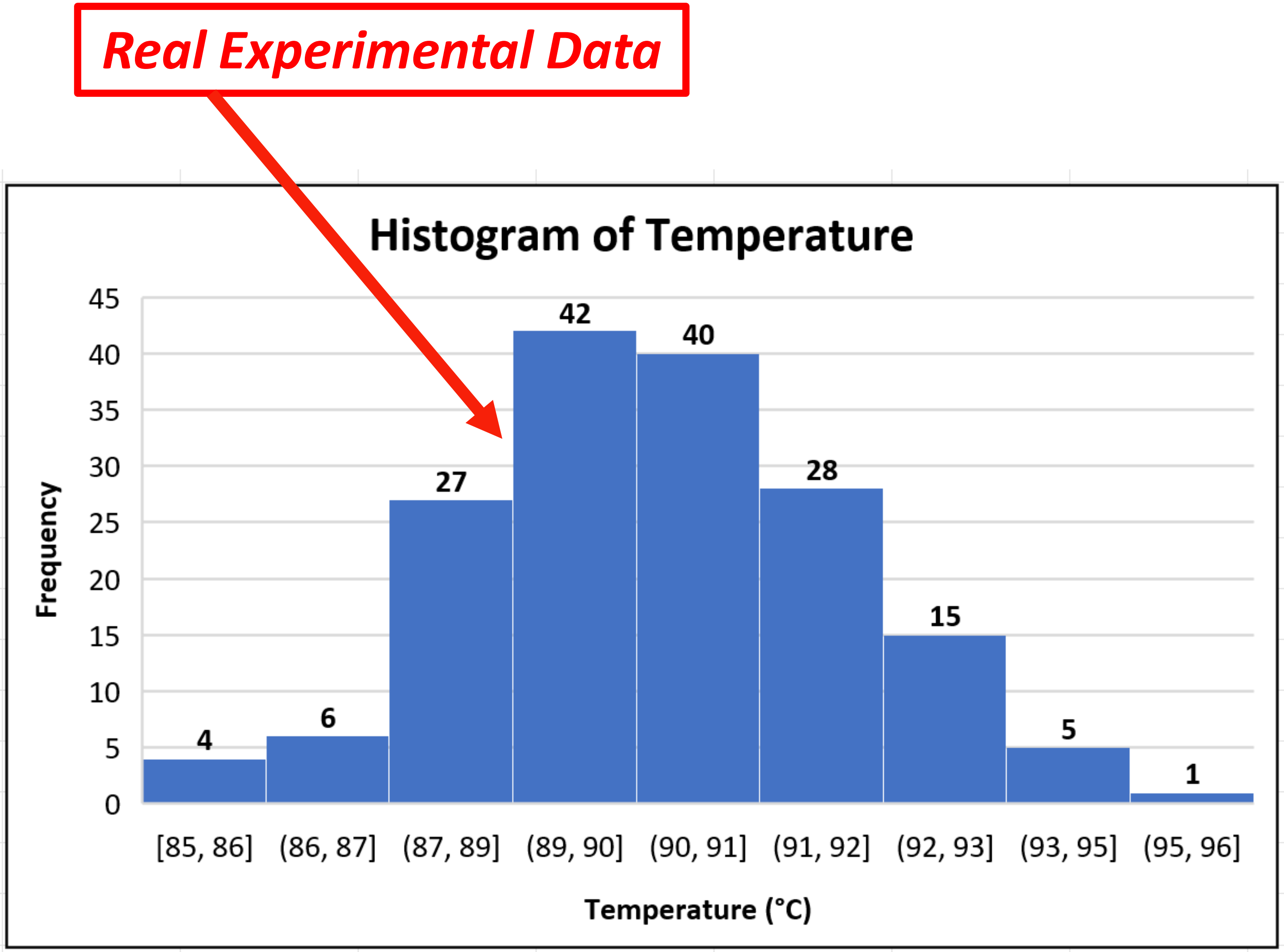
	Normal (aka Gaussian) distribution	Student's t-distribution
Type of distribution	continuous	
Shape	bell-shaped, symmetrical, the tails approach the horizontal axis but never touch it	
Mean = Median = Mode	Yes	
Test statistic	$z = \frac{(\bar{x} - \mu)}{\sigma}$	$t = \frac{(\bar{x} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)}$
Varies with sample size	**No**	**Yes**
To be used when	Population or process Standard Deviation is known or **Sample Size ≥ 30**	Population or process Standard Deviation is unknown or **Sample Size < 30**

What is the practical use of all this?

Let see a practical example !

INFERENTIAL STATISTICS

Let’s consider, for example, the 10-year data of a critical parameter (a *reaction critical temperature*) whose value must be between 85 °C and 95 °C otherwise the process leads to the formation of unwanted impurities.



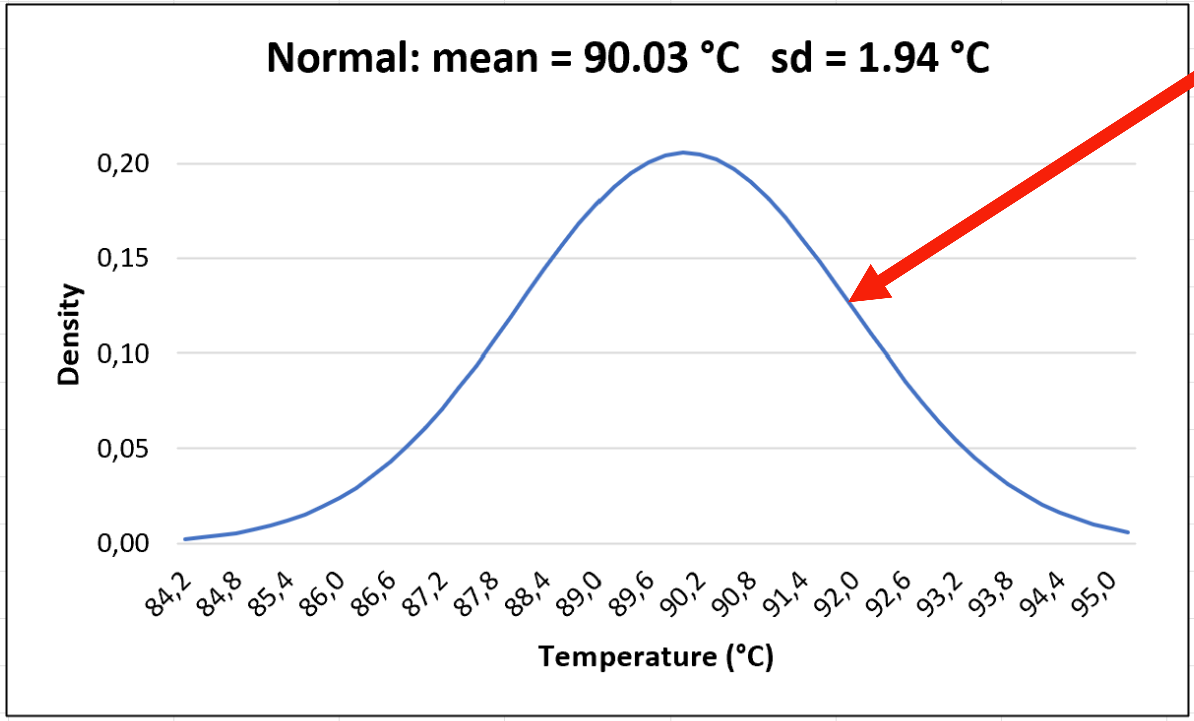
INFERENTIAL STATISTICS

Experimental data can be approximated using a Normal random variable X (the critical temperature) characterized by:

$\bar{x} = 90{,}03\ °C \quad s = 1.94\ °C$

Mathematical Model

Temperature: Descriptive Statistics	
Mean	**90,03**
Standard Error	0,15
Median	90
Mode	89
Standard Deviation	**1,94**
Sample Variance	3,77
Kurtosis	-0,24
Skewness	0,02
Range	10
Minimum	85
Maximum	95
Sum	15125,85
Count	168



$$f(x) = \frac{1}{\sqrt{2\pi}\, s} e^{-\frac{(x-\bar{x})^2}{2\, s^2}}$$

$$-\infty < x < \infty$$

The area under the curve equals 1 or 100%

INFERENTIAL STATISTICS

What is the probability that *P* (*X* < 85 °C and *X* > 95 °C) ?

or, in other words, what is the probability that the critical temperature exceeds the foreseen limits ?

$P\ (X \leq 95.0) =$	0,9948
$P\ (X < 85.0) =$	0,0048
$P(85.0 < X \leq 95.0) =$	0,9900

=NORM.DIST(95;90,03;1,94;TRUE)

=NORM.DIST(85;90,03;1,94;TRUE)

The NORM.DIST function returns the normal distribution for the specified mean and standard deviation. If TRUE, it returns the *cumulative distribution function*; if FALSE, it returns the *probability density function*.

What does this mean in practice?

There is about 1% probability that the critical reaction parameter exceeds the specification limits!

What does this mean in practice?

- Based on these data there is about 1% probability that the critical reaction parameter could exceed the limits

- ***OOS results may be observed !***

INFERENTIAL STATISTICS

- This can be considered a simple example of

Science based QA

since:

- The conformance (or criticality as in this case) to specifications can be demonstrated
- Any future actions can be taken correctly

Better Science = Better Outcomes = Less Costs

WARNING !

What we have just seen is none other than what,

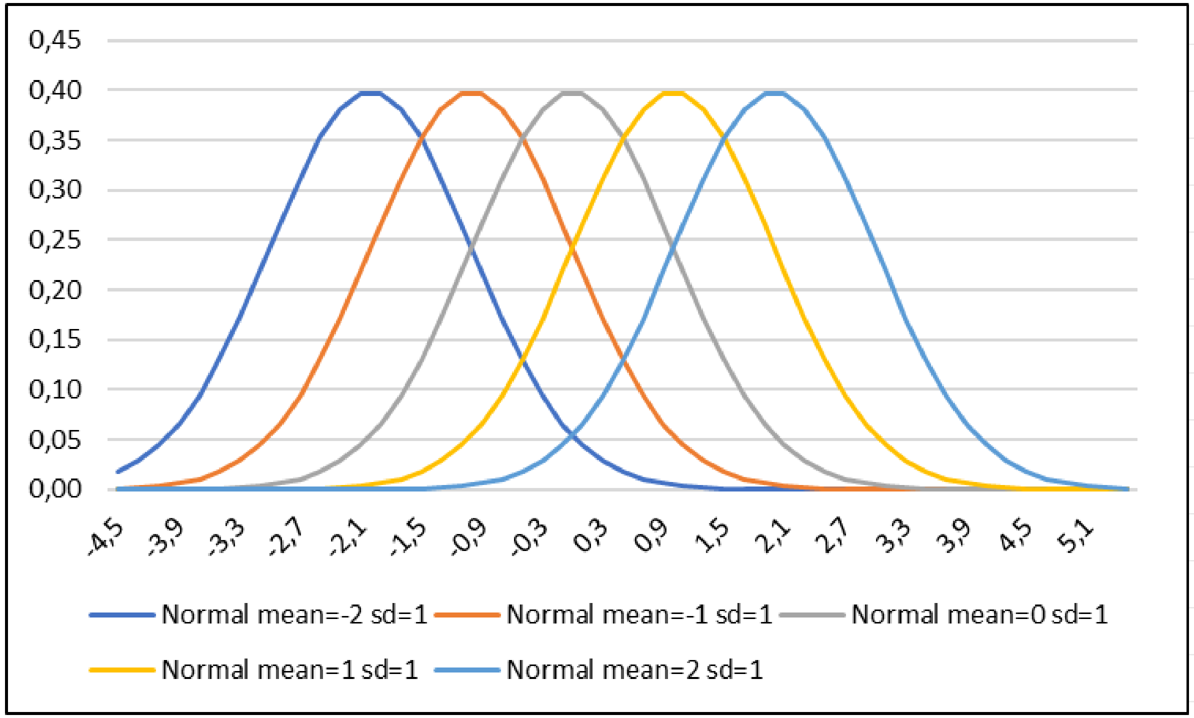
in the end, the Capability Analysis returns!

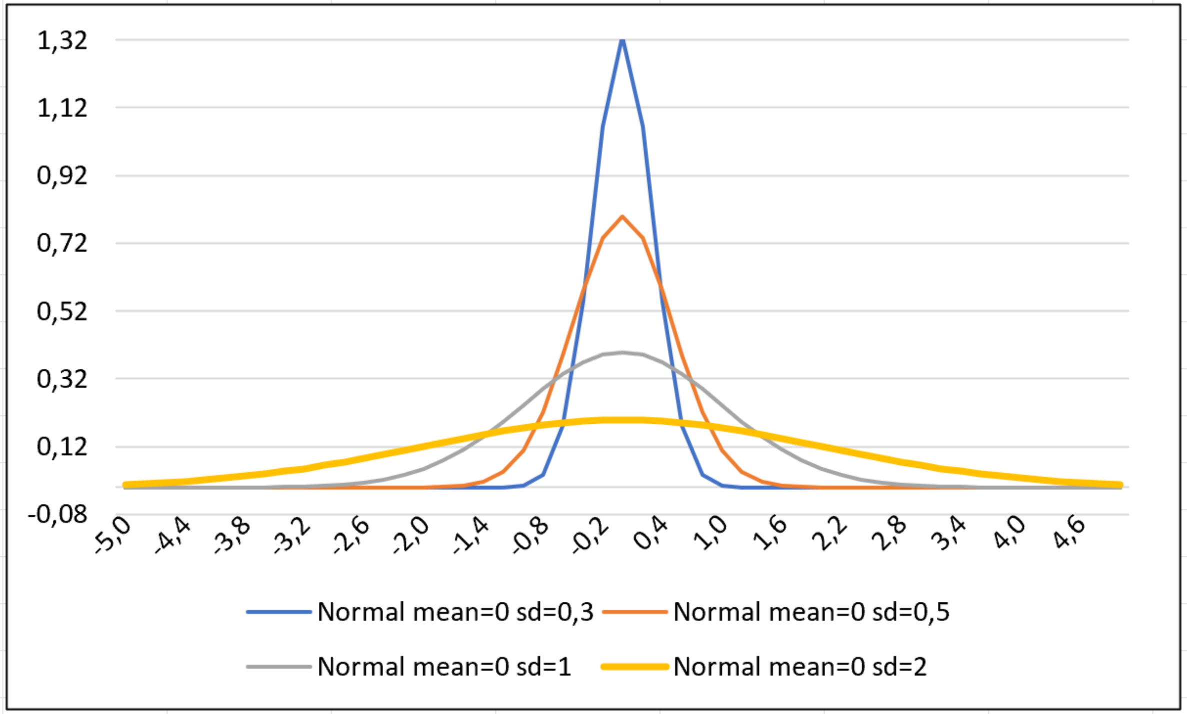
Let's now go back to the

Normal Distribution and its characteristics !

INFERENTIAL STATISTICS

Normal Distributions that can be generated by varying mean (μ) and standard deviation (σ) are infinite !





INFERENTIAL STATISTICS

To simplify :

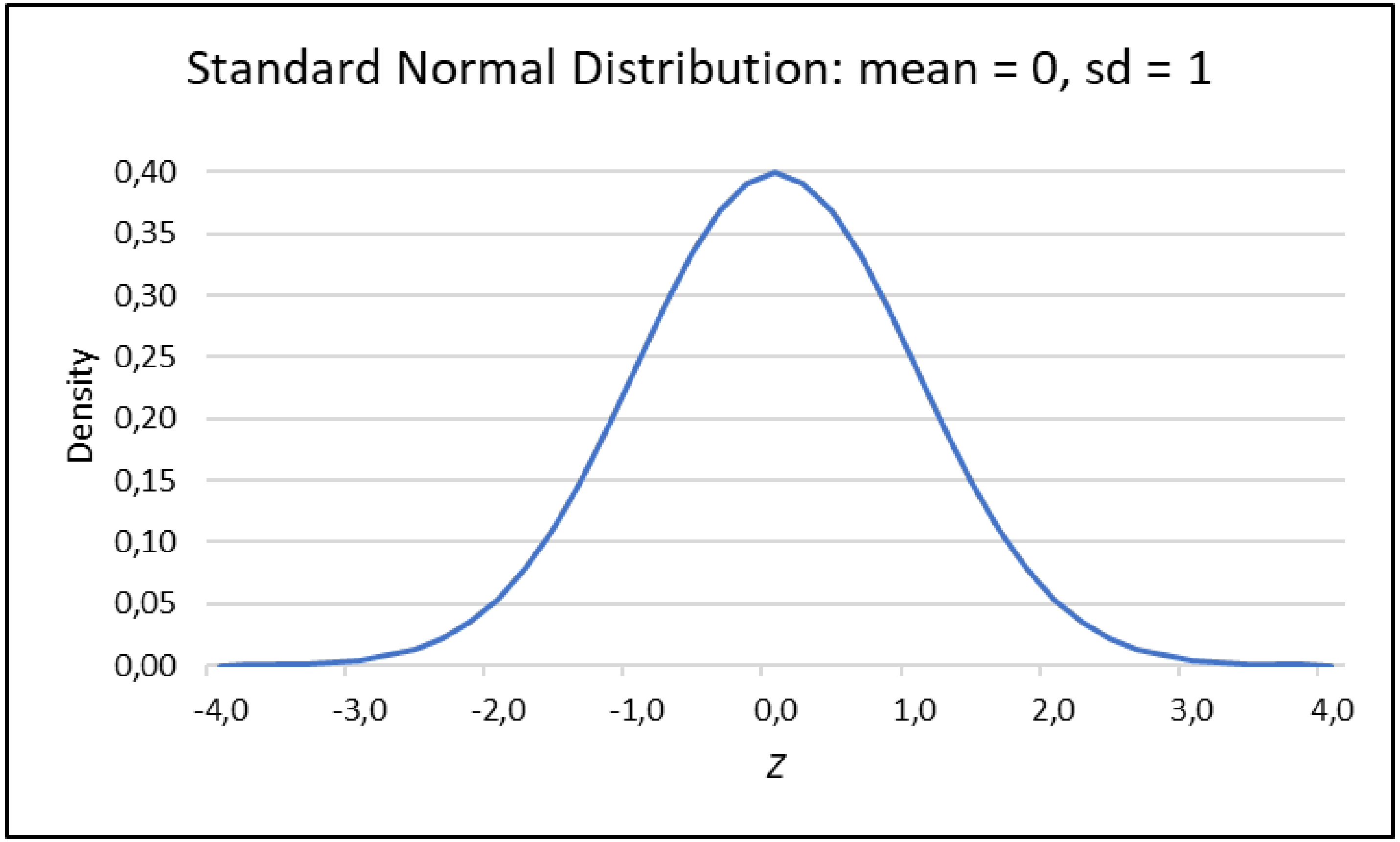
STANDARDIZATION

In other words:

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

The ***Standardized Normal Distribution*** is characterized by:

$$\bar{Z} = 0 \qquad \sigma_Z^2 = 1$$



S.M. Ross, A first course in probability– 9th Edition, Pearson College (2012)

INFERENTIAL STATISTICS

- The ***z transformation*** allows to transform ***any*** Normal Distribution into the Standard Normal Distribution.

- The values of the ***Z test statistic*** are plotted along the horizontal axis and correspond to standard deviations.

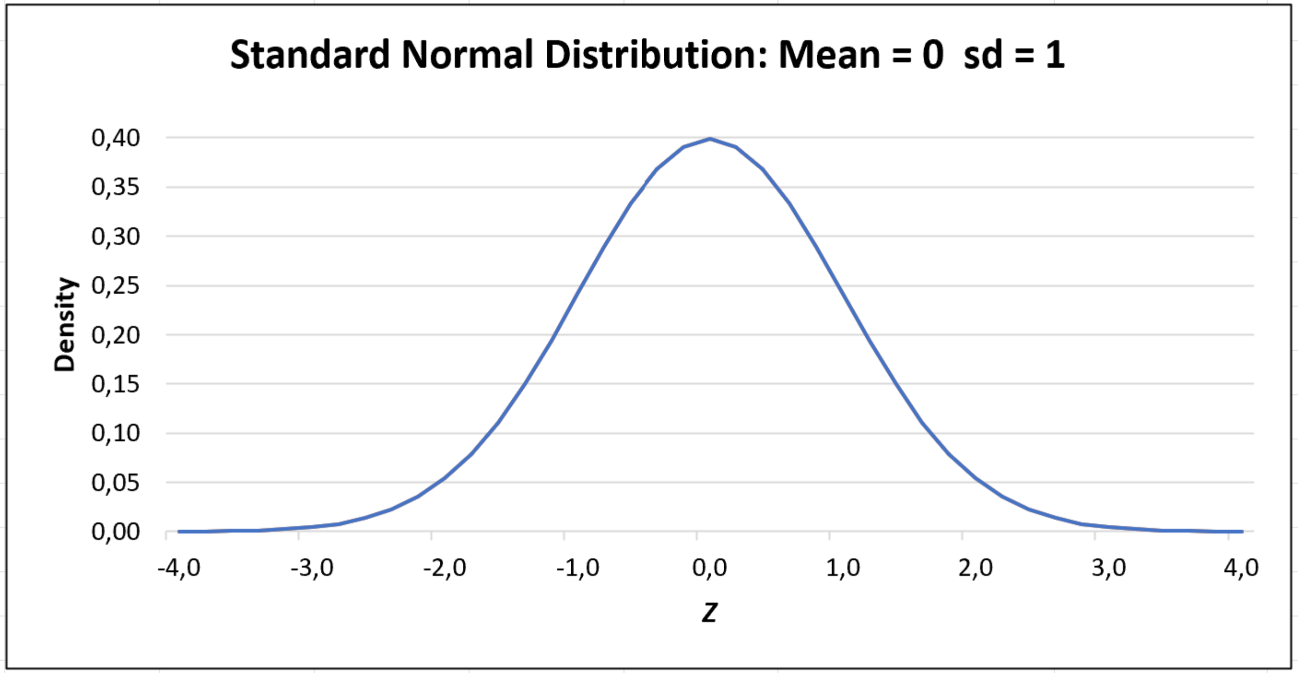
- As an exercise, let's try to calculate the probability values between +1 and -1 or between +2 and -2 or between +3 and -3 using the Excel function:

NORM.S.DIST

which returns the standard normal distribution. If TRUE, NORM.S.DIST returns the *cumulative distribution function*; if FALSE, it returns the *probability mass function*.

INFERENTIAL STATISTICS

- P (-1 < *Z* < +1) = =NORM.S.DIST(1;TRUE) - NORM.S.DIST(-1;TRUE) = 0,682689492 ➡ **~ 68,27%**
- P (-2 < *Z* < +2) = =NORM.S.DIST(2;TRUE) - NORM.S.DIST(-2;TRUE) = 0,954499736 ➡ **~ 95,45%**
- P (-3 < *Z* < +3) = =NORM.S.DIST(3;TRUE) - NORM.S.DIST(-3;TRUE) = 0,997300204 ➡ **~ 99,73%**



INFERENTIAL STATISTICS

HOWEVER, ALWAYS REMEMBER THAT:

- *In all cases, these are mathematical models with respect to which the distributions of real data are compared.*

- *the use of these models is convenient only because, by dealing with mathematical functions, the theory provides simple formulas for the calculation of practical parameters such as those just seen.*

INFERENTIAL STATISTICS

How can we practically and easily determine whether a given probability distribution is a reasonable model for the experimental data?

PROBABILITY PLOT or QQ-Plot

- It deals of *graphical methods* that are used to compare the distribution of a set of experimental data with a theoretical reference distribution, usually the Normal.
- If you want to statistically verify that the data follow a certain distribution, you have to use specific tests such as those of Kolmogorov-Smirnov or Anderson-Darling.

What are Quantiles ?

- A ***quantile*** is a value that divides a dataset into equal-sized groups.
- If you divide a dataset into four equal parts, each part is called a ***quartile***.
- The first quartile (Q1) represents the 25th ***percentile***, the second quartile (Q2) represents the 50th percentile (which is also the *median*), and the third quartile (Q3) represents the 75th percentile. These quartiles are examples of quantiles.

INFERENTIAL STATISTICS

The idea of a *QQ-plot* is straightforward: we want to form a scatterplot that relates our data values to the ideal values of the theoretical distribution.

How it works ? Simple:

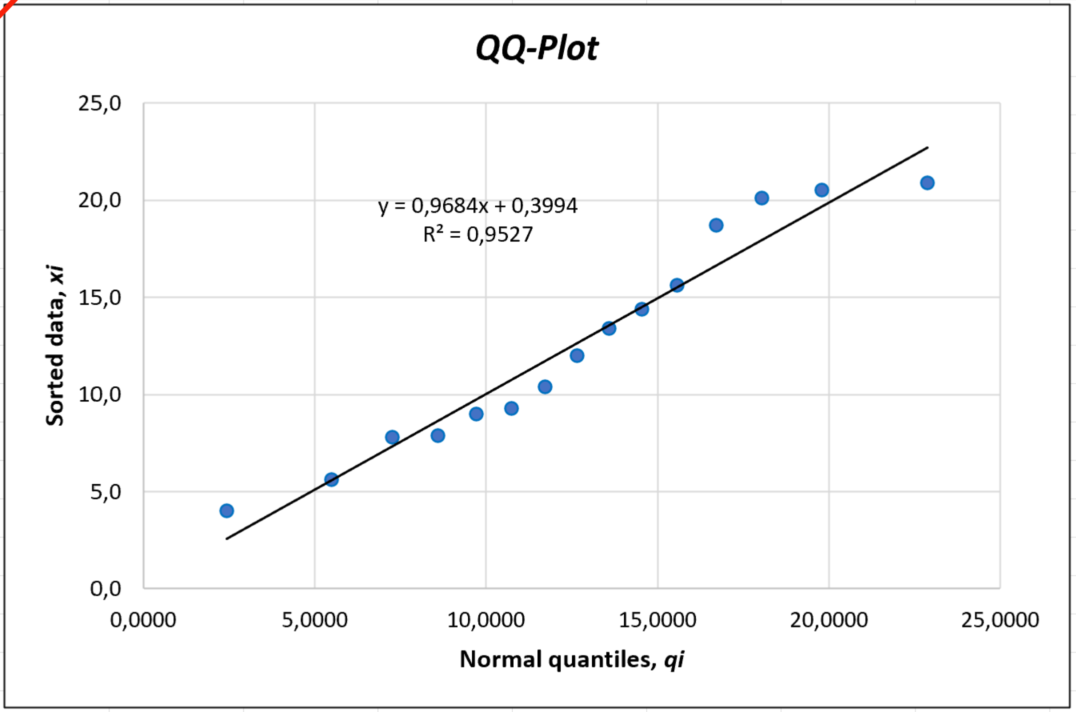
- Data values are first arranged in increasing order
- For each data value $\boldsymbol{x_i}$, we use the data to estimate the probability $\boldsymbol{p_i}$ that a random value in the distribution we are sampling from is less than $\boldsymbol{x_i}$
- Finally, the ideal values, or *theoretical quantiles*, $\boldsymbol{q_i}$, are chosen from our comparison distribution. That is, $\boldsymbol{x_i}$ is the same quantile in the data as in the comparison distribution (*e.g.*, Normal).

INFERENTIAL STATISTICS

=(E3-0,5)/15

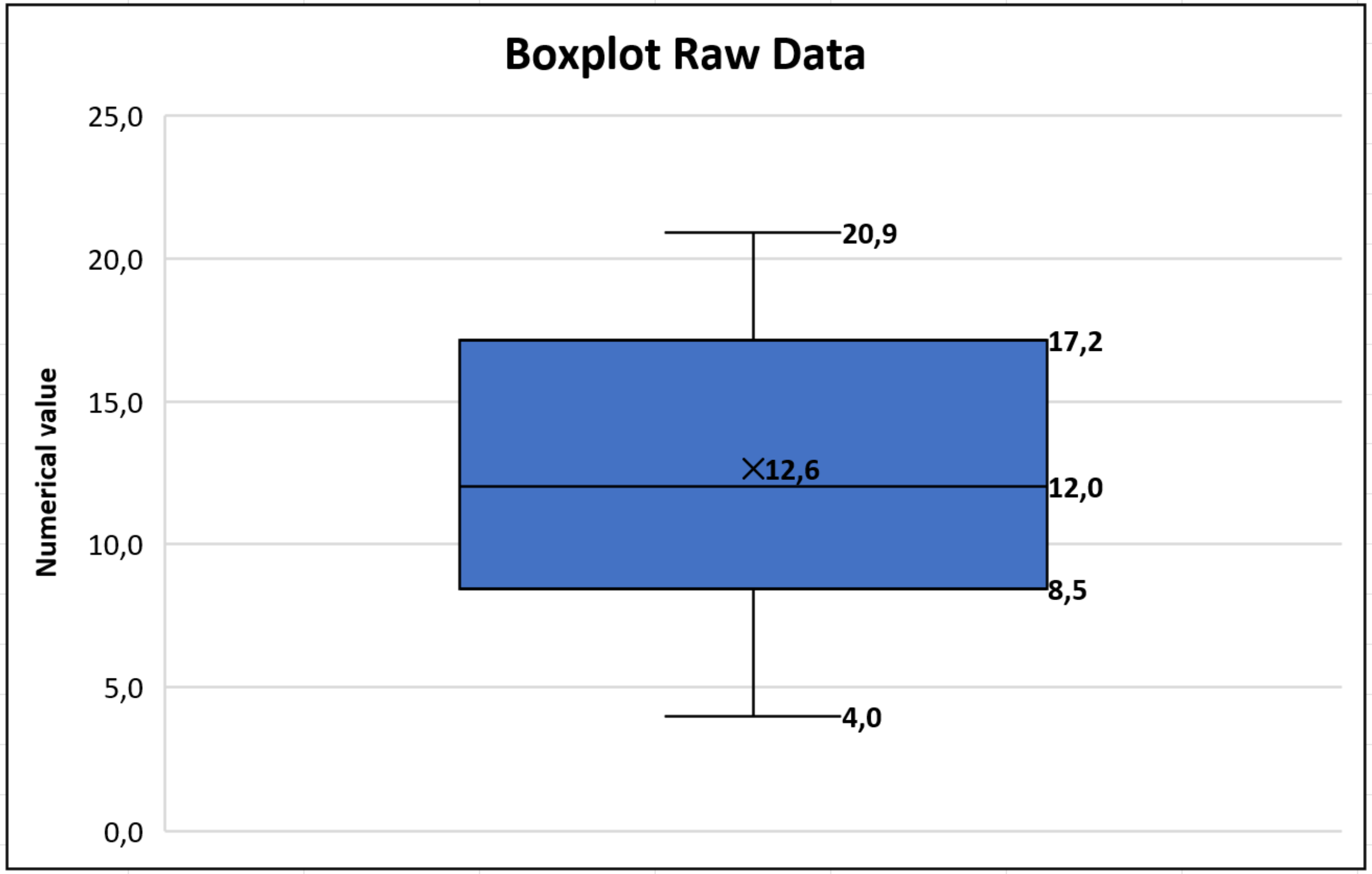
=NORM.INV(F3;D19;D20)

Raw Data	Sorted data, *xi*	Ranking	Probability, *pi*	Normal quantiles, *qi*
18,7	4,0	1	0,03	2,4203
12,0	5,6	2	0,10	5,4984
7,8	7,8	3	0,17	7,2489
7,9	7,9	4	0,23	8,5836
20,5	9,0	5	0,30	9,7177
5,6	9,3	6	0,37	10,7414
20,1	10,4	7	0,43	11,7044
9,0	12,0	8	0,50	12,6400
9,3	13,4	9	0,57	13,5756
10,4	14,4	10	0,63	14,5386
4,0	15,6	11	0,70	15,5623
15,6	18,7	12	0,77	16,6964
20,9	20,1	13	0,83	18,0311
14,4	20,5	14	0,90	19,7816
13,4	20,9	15	0,97	22,8597
Mean =	12,6			
Std Dev =	5,5726			



INFERENTIAL STATISTICS

The fact that the data appear almost normally distributed is also indicated by the boxplot shown here, which is fairly symmetric, *i.e.*, mean and median are very close values, the two halves of the box and whiskers are comparable.

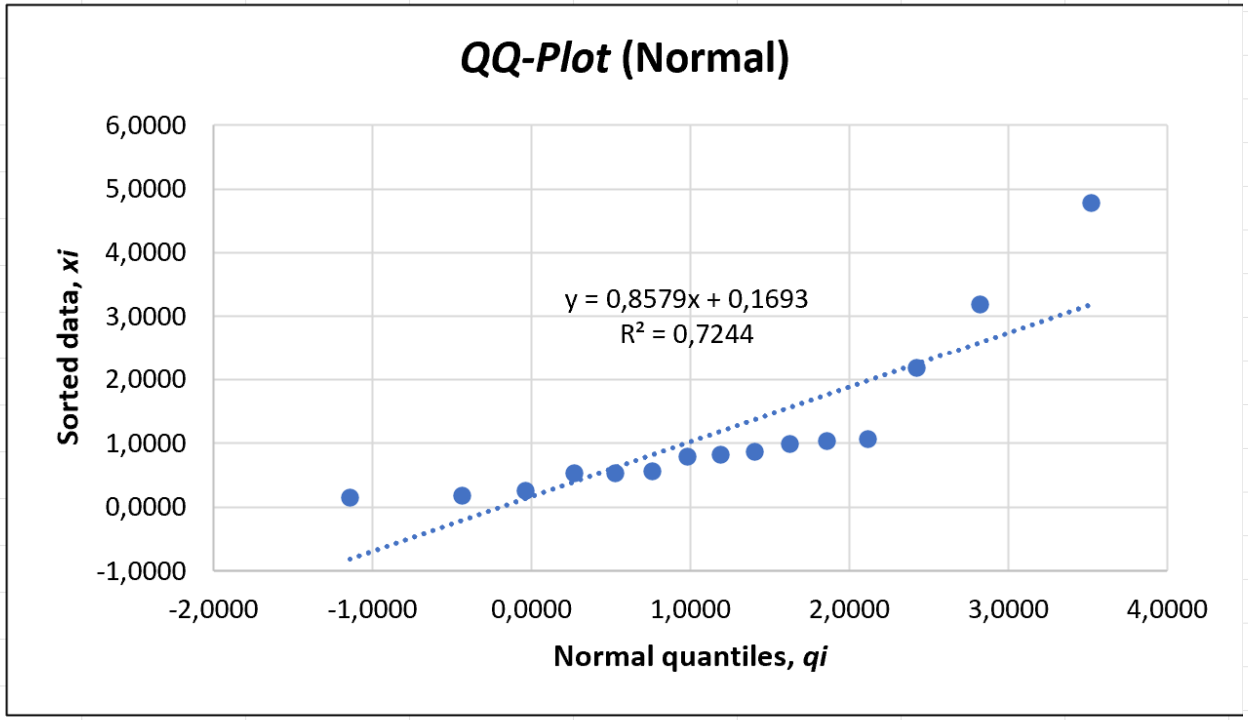


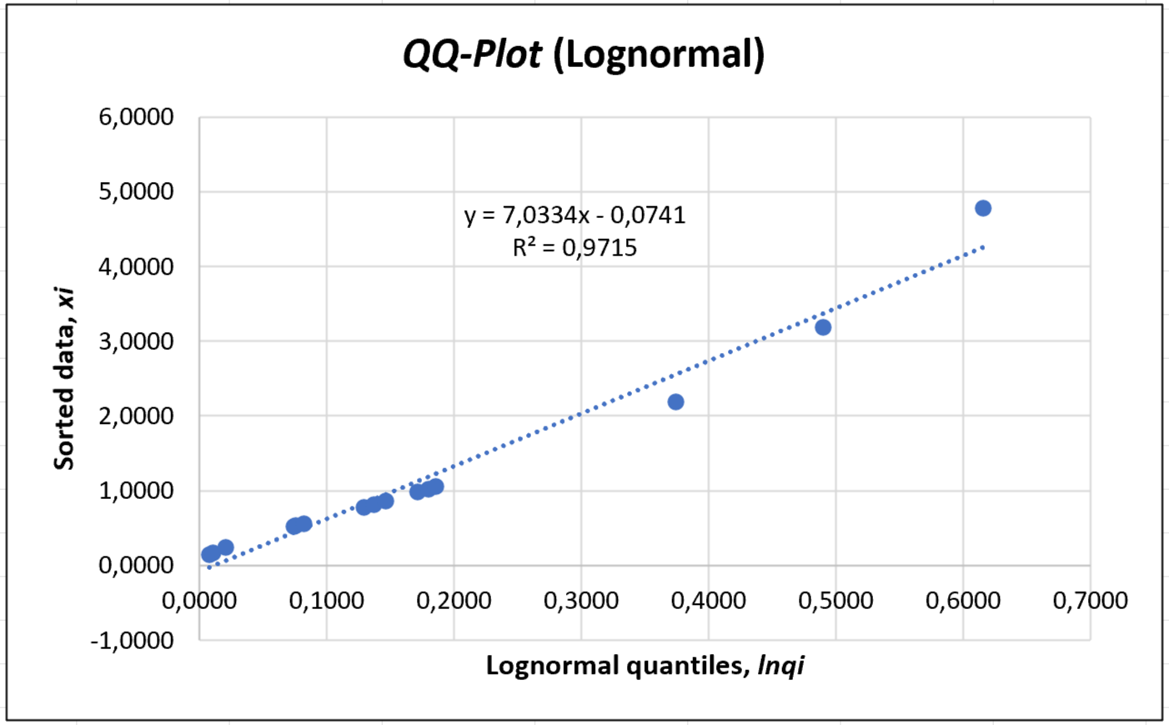
INFERENTIAL STATISTICS

Let us now consider the data that is certainly skewed as those distributed in a lognormal way and proceed as before.

Raw Data	Sorted data, *xi*	Ranking	Probability, *pi*	Normal quantiles, *qi*	Lognormal quantiles, *lnqi*
0,5305	0,1510	1	0,03	-1,1406	0,0077
0,7821	0,1737	2	0,10	-0,4382	0,0103
0,8641	0,2469	3	0,17	-0,0388	0,0208
0,9851	0,5224	4	0,23	0,2657	0,0739
1,0264	0,5305	5	0,30	0,5245	0,0756
0,1510	0,5591	6	0,37	0,7580	0,0816
2,1861	0,7821	7	0,43	0,9777	0,1292
0,1737	0,8186	8	0,50	1,1912	0,1369
0,8186	0,8641	9	0,57	1,4047	0,1465
0,5224	0,9851	10	0,63	1,6244	0,1714
1,0569	1,0264	11	0,70	1,8580	0,1797
0,2469	1,0569	12	0,77	2,1167	0,1858
4,7780	2,1861	13	0,83	2,4213	0,3738
0,5591	3,1874	14	0,90	2,8207	0,4900
3,1874	4,7780	15	0,97	3,5230	0,6153
Mean =	1,1912				
Std Dev =	1,2715				

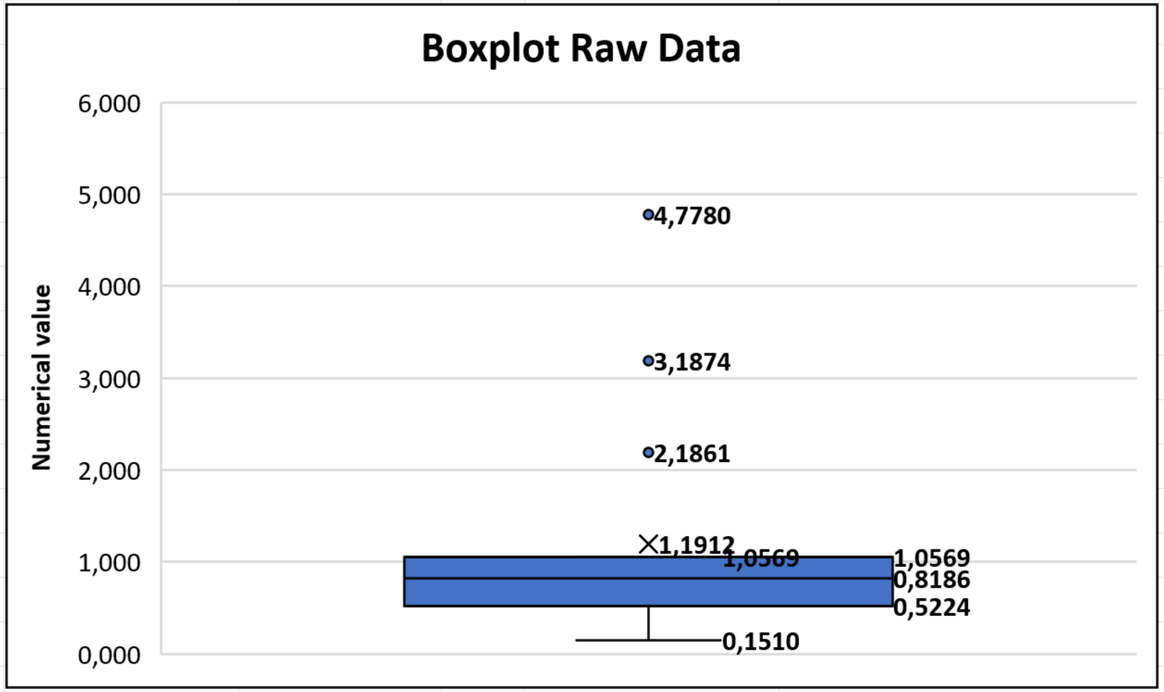
INFERENTIAL STATISTICS

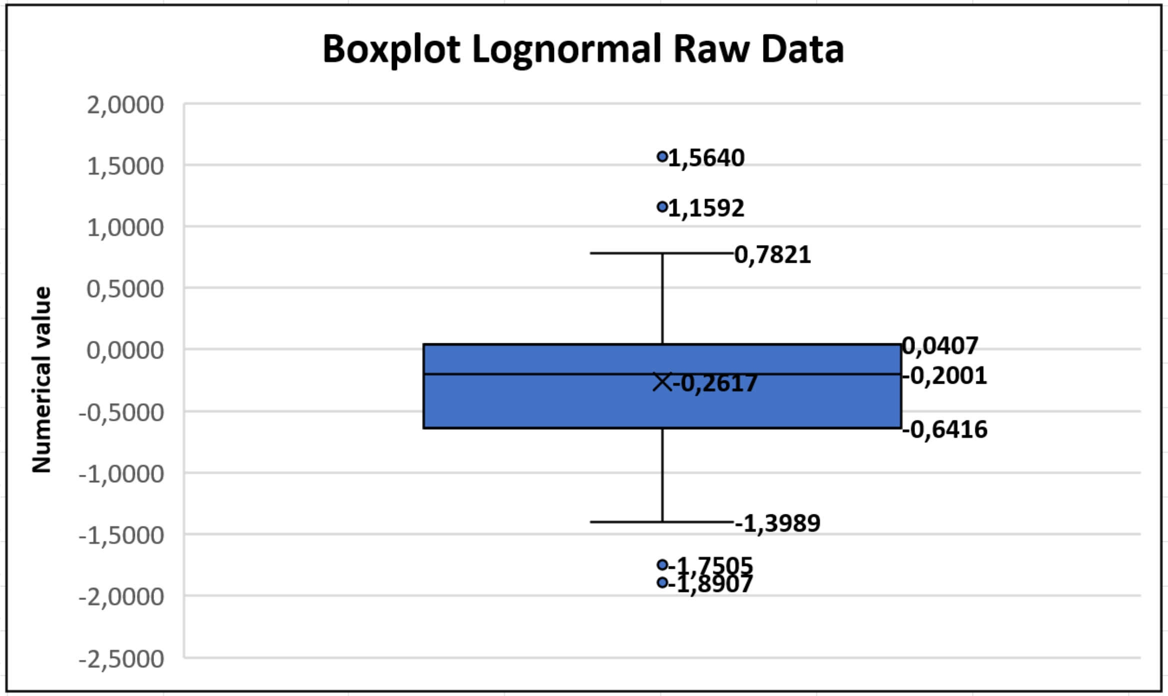




INFERENTIAL STATISTICS

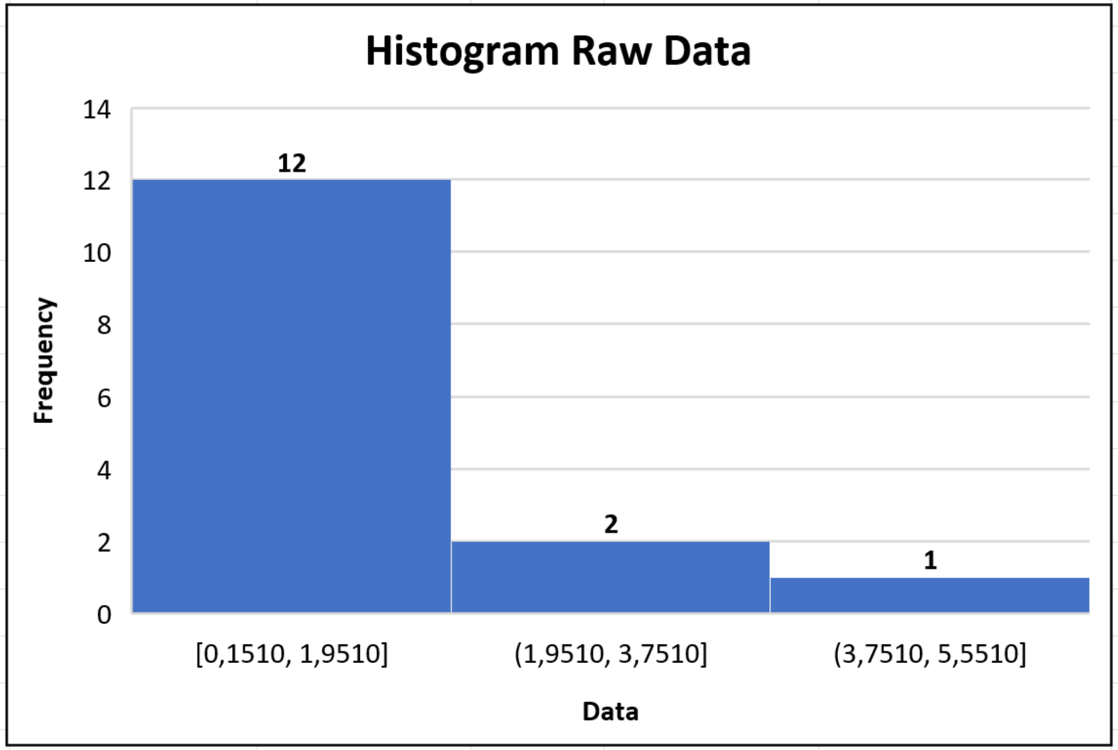
In this case the situation of "imbalance" in the data distribution is also well indicated by the boxplots: that of raw data *as is* looks visibly asymmetrical while that of natural logarithm of raw data looks symmetrical.

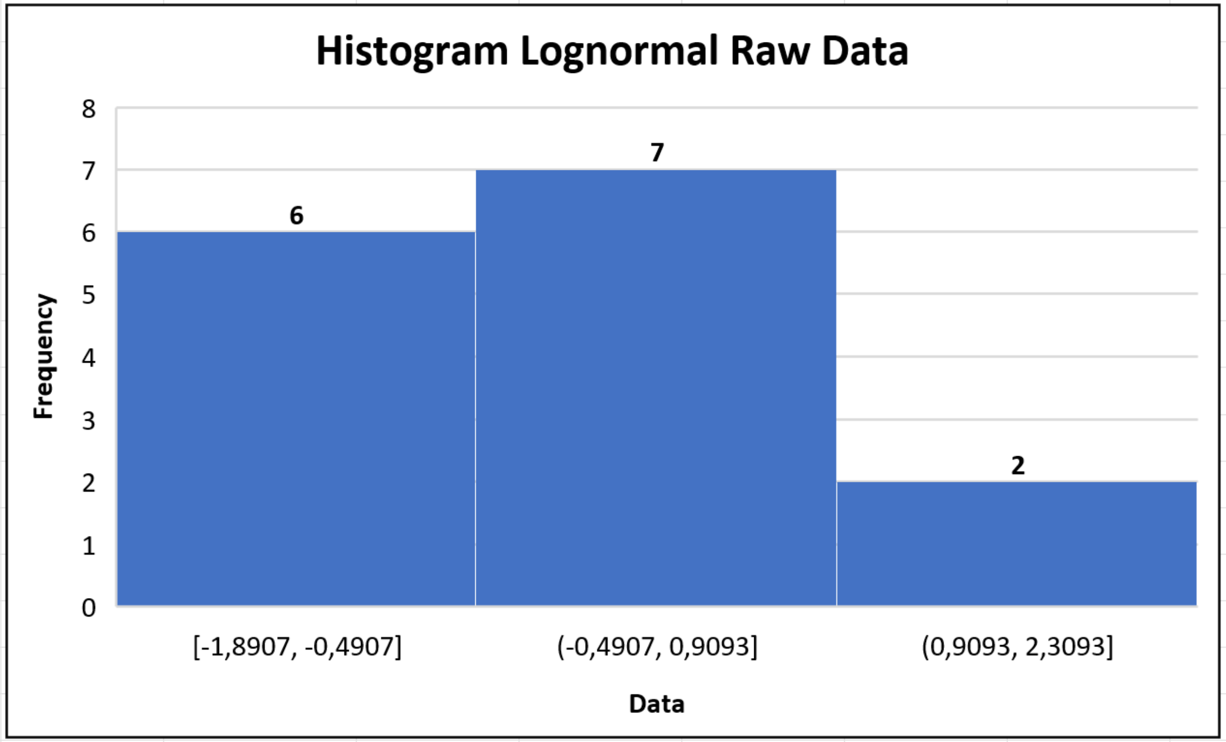




INFERENTIAL STATISTICS

In this case histograms are even more explicative.





INFERENTIAL STATISTICS

IF THE REAL DATA IS NOT NORMALLY DISTRIBUTED

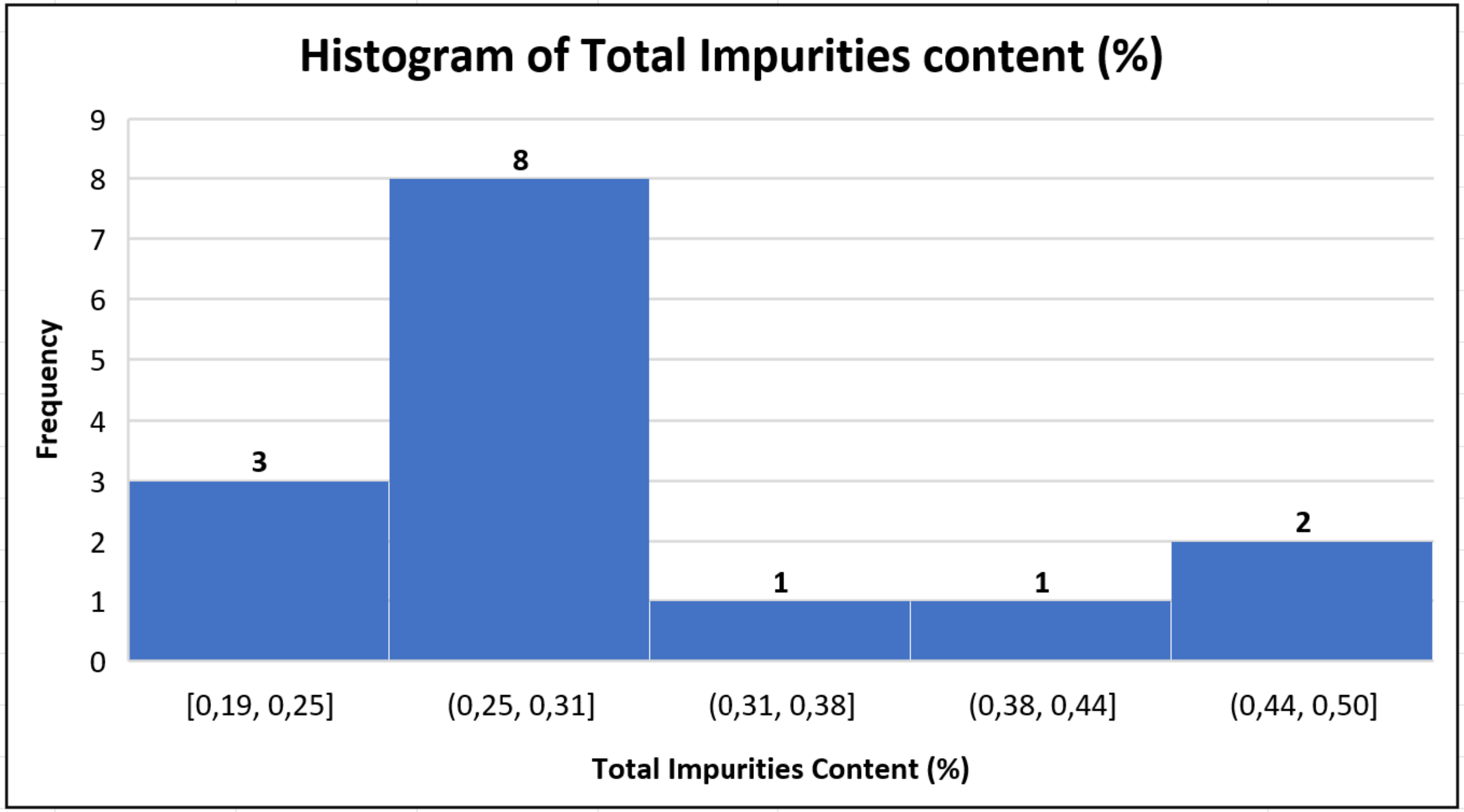
IT IS NOT THE END OF THE WORLD!

The data can be normalized by performing mathematical operations on them (*e.g.*, natural logarithm, square root, reciprocal, *etc.*) or different types of tests can be used, the so-called «non-parametric tests».

An example for all: the TOTAL IMPURITIES CONTENT for a series of batches

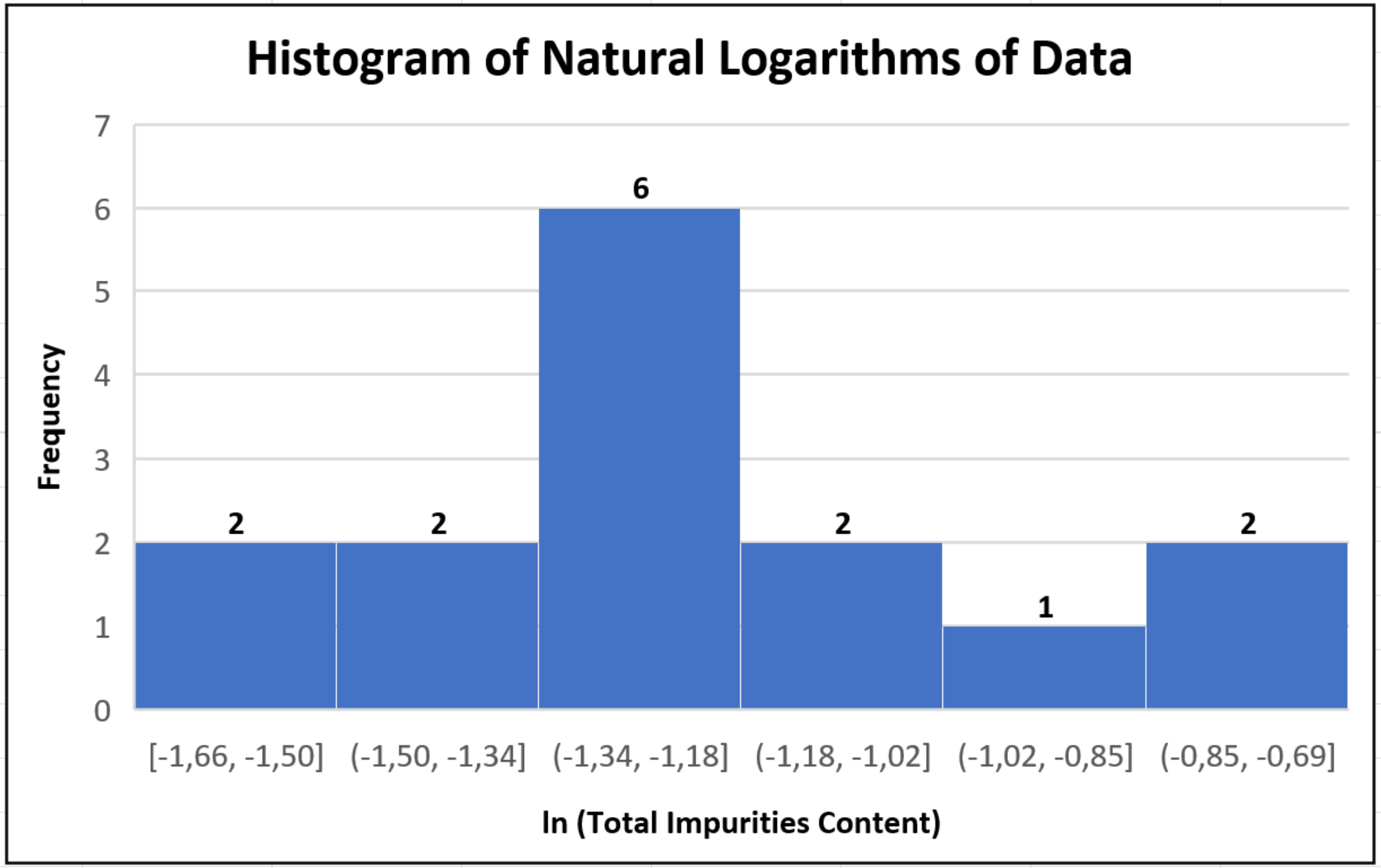
INFERENTIAL STATISTICS

Total Impurities Content (%)
0,19
0,22
0,45
0,30
0,30
0,40
0,50
0,32
0,28
0,30
0,30
0,31
0,26
0,25
0,27



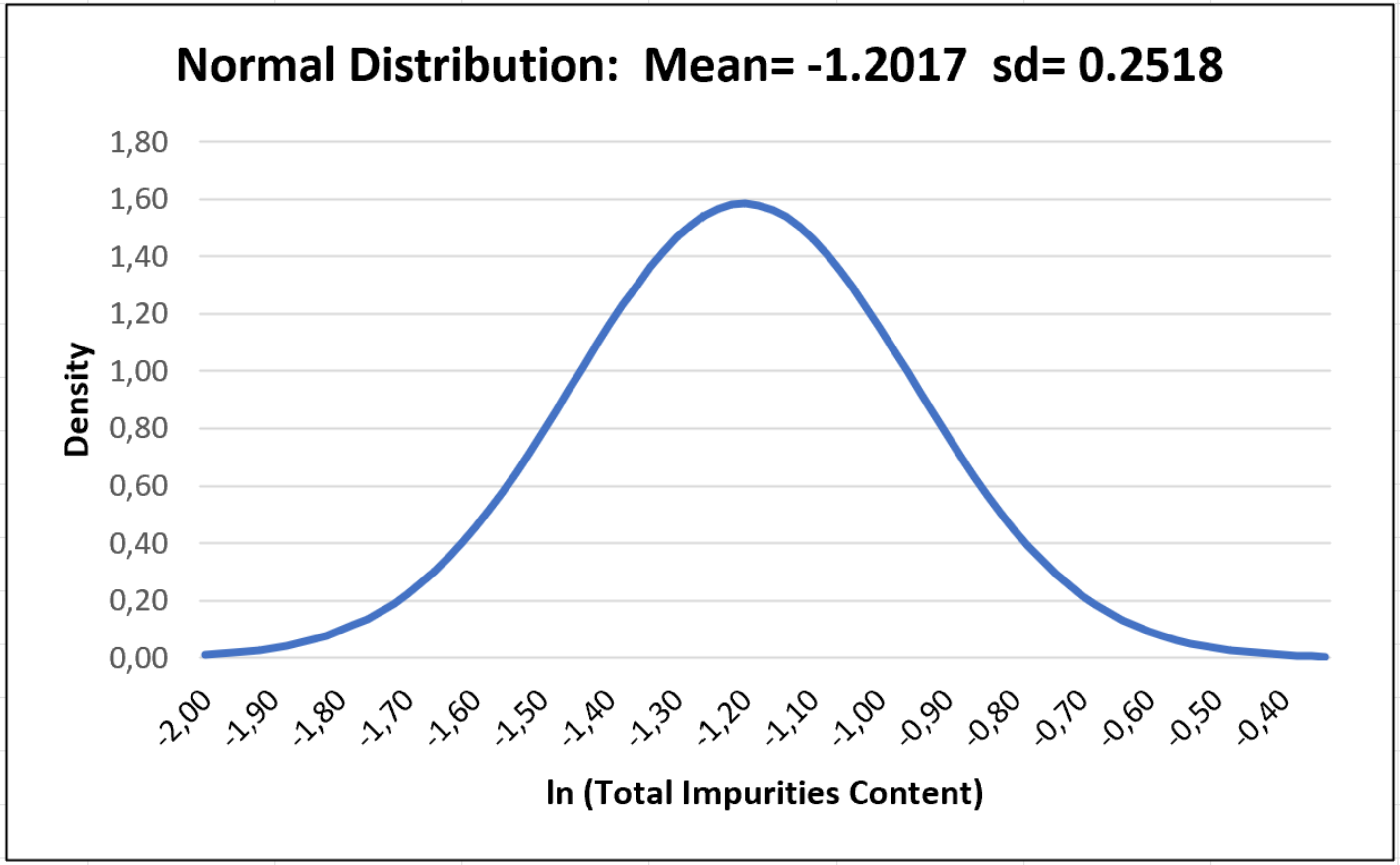
INFERENTIAL STATISTICS

Natural Logarithm of Total Impurities Content (%)
-1,6607
-1,5141
-0,7985
-1,2040
-1,2040
-0,9163
-0,6931
-1,1394
-1,2730
-1,2040
-1,2040
-1,1712
-1,3471
-1,3863
-1,3093



INFERENTIAL STATISTICS

Natural Logarithm of Total Impurities content: Descriptive Statistics	
Mean	**-1,2017**
Standard Error	0,0650
Median	-1,2040
Mode	-1,2040
Standard Deviation	**0,2518**
Sample Variance	0,0634
Kurtosis	0,4853
Skewness	0,4249
Range	0,9676
Minimum	-1,6607
Maximum	-0,6931
Sum	-18,0250
Count	15



INFERENTIAL STATISTICS

What is the probability that *P* (*X* > 0,50%) or *P* (ln *X* > - 0,6931) ?

or, in other words:

What is the probability that the Total Impurities Content could exceed the limit ?

P(X < 0,50 or lnX < -0,6931) =	0,9783
P(X > 0,50 or lnX > -0,6931) =	0,0217
in percent =	**2,17**

fx = NORM.DIST(-0,6931;-1,2017; 0,2518;TRUE)

fx = 1- 0,9783

fx = 0,0217 * 100

What does this mean in practice?

- Based on these data there is more than 2% probability that the Total Impurities Content could exceed the upper specification limit

- ***OOS may be observed !***

INFERENTIAL STATISTICS

- In the examples shown up to now (*i.e.*, critical temperature and total impurities content) the possibility of calculating the probability associated with a given range of values has been used.

- However, it is also possible to proceed "in the opposite direction" and this can be useful for practical cases such as the one in the next case study.

- For this purpose, Excel provides the **NORM.INV** *function which returns the inverse of the normal <u>cumulative</u> distribution for a specified mean and standard deviation.*

INFERENTIAL STATISTICS

- Let's suppose we want to estimate the mean and standard deviation of a compressing process to produce tablets whose weight must be 50 $\pm$ 2 mg.

- Let's say we want 99.7% of our tablets to fall within our specification limits (48mg to 52mg). This is equivalent to allowing a total of 0.3% defects, or 0.15% on each side of the distribution (assuming it's symmetric).

- The z-scores corresponding to these defect rates can be found using the NORM.S.INV function in Excel, *i.e.*:

  = ABS(NORM.S.INV(0.0015))

  The result will be approximately 2.9677. This is the number of standard deviations away from the mean that corresponds to the top and bottom 0.15% of the distribution.

- Now, let's estimate the mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) using the following formulas:

$$\hat{\mu} = \frac{(LTL \times z_{UTL}) - (UTL \times z_{LTL})}{z_{UTL} - z_{LTL}}$$

$$\hat{\sigma} = \frac{UTL - LTL}{z_{UTL} - z_{LTL}}$$

where:

- UTL and LTL represent the Upper and Lower Tolerance Limits (*i.e.*, 52 mg and 48 mg)
- z_{UTL} *and* z_{LTL} represent the standardized errors estimated using **NORM.INV** (*i.e.*, 2.9677 and - 2.9677)

INFERENTIAL STATISTICS

Substituting these values into the formulas:

$$\hat{\mu} = \frac{\left((48 \times 2.9677) - \left(52 \times (-2.9677)\right)\right)}{\left(2.9677 - (-2.9677)\right)} = 50\ mg$$

$$\hat{\sigma} = \frac{(52 - 48)}{\left(2.9677 - (2.9677)\right)} = 0.67\ mg$$

Therefore, the estimated mean ($\hat{\mu}$) will be 50mg and the estimated standard deviation ($\hat{\sigma}$) will be approximately 0.67mg. This is the standard deviation that we need in order to ensure that 99.7% of our tablets are within the specification limits of 48mg to 52mg.

INFERENTIAL STATISTICS

This last case study can also be considered a simple example of

Science based QA

since the outcome of the compressing process is “modeled” on a logical basis (*i.e.*, normally distributed weights) and it is not left to chance.

Better Science = Better Outcomes = Less Costs

PARAMETER ESTIMATION

INFERENTIAL STATISTICS

Back to the introduction to Inferential Statistics methods, two big topics were mentioned and the first was:

Parameter Estimation

which consists in the best evaluation of an unknown parameter of the population (for example, the mean μ or the standard deviation σ) using the sample data.

This evaluation can be of two types: ***punctual*** or ***by intervals***.

What does it mean ?

INFERENTIAL STATISTICS

- ***punctual estimation methods*** provide, for the estimated parameters, a single value and do not offer any information on the precision of this value.
  For this reason, it is often preferred to use ***interval estimates*** that provide a range of possible values.

- from a “punctual" point of view, for example, the sample mean, $\bar{x}$, is an "appropriate estimator" of the unknown population mean, μ, *but this in no way implies that the sample mean coincides exactly with that of the population from which that sample comes.*

INFERENTIAL STATISTICS

- the ***method of interval estimates***, due to *Neyman*, allows to determine, on the basis of sample observations, an interval called ***confidence interval***, within which lies, with a *prefixed probability* (usually 95% or 99% or 0.95, 0.99) *called level of confidence,* ***C***, the true and unknown parameter to be estimated (*e.g.*, μ or σ).

- The complement to 1 of ***C*** is the so-called *Level of Significance* and it is indicated with α (= 1- C) and it equal to 0.05 or 0.01.

- *Level of Confidence,* ***C***, and Level of Significance, α, measure the same thing: how sure we are that we are making the right decision or not !

What is the practical use of all this?

Let see two practical examples !

INFERENTIAL STATISTICS

20 tablets from a validated process are sampled in-process and weighed. We want to determine the 95% and 99% confidence intervals for the mean weight of all tablets produced.

using the Data Analysis tool

using CONFIDENCE.NORM(0,05; 0,83;20)

using CONFIDENCE.T(0,05;0,83;20)

using CONFIDENCE.NORM(0,01;0,83;20)

using CONFIDENCE.T(0,01;0,83;20)

Tablet weight (mg)
50,29
48,81
49,79
51,48
49,19
50,23
49,46
48,14
49,18
50,38
50,56
48,93
50,06
49,29
49,72
49,51
50,45
49,74
50,15
51,45

Tablet weight (mg) Summary Statistics	
Mean	**49,84**
Standard Error	0,19
Median	49,77
Standard Deviation	**0,83**
Sample Variance	0,70
Kurtosis	0,23
Skewness	0,19
Range	3,35
Minimum	48,14
Maximum	51,48
Sum	996,83
Count	20
Confidence Interval (95,0%)	**0,3906**

Confidence interval (95%) normal distribution	**0,3638**
Confidence interval (95%) t-distribution	**0,3885**
Confidence interval (99%) normal distribution	**0,4781**
Confidence interval (99%) t-distribution	**0,5310**

INFERENTIAL STATISTICS

A few remarks:

- The **CONFIDENCE.NORM** function returns the confidence interval for a population mean, using a Normal distribution while the **CONFIDENCE.T** function returns the confidence interval for a population mean, using a Student's t distribution.

- The **CONFIDENCE.NORM** function should be used with a sample “large enough” (*i.e.*, 30 or more observations) while for smaller samples it is better using the **CONFIDENCE.T** function.

- The Confidence Interval calculated using the “Data Analysis” tool is more similar to the one obtained using the **CONFIDENCE.T** function rather than the **CONFIDENCE.NORM** function. This makes sense since the sample consisted of only 20 observations.

INFERENTIAL STATISTICS

- Using the Confidence Level value (95%) calculated using the "Data Analysis" tool, which is slightly higher as it is calculated assuming an "unknown variance", it is possible to calculate the Confidence Interval as shown on the side.
- *Thus, with a 95% probability, our validated process will produce tablets having an average weight between 49.65 mg and 50.04 mg.*
- *Among other things, this measure can tell us quickly and above all in a serious way, if our process is working well or not !*

Calculations	
Confidence Interval (95,0%)	0,3906
Count	20
Mean	49,84
Standard Deviation	0,8345
Confidence Interval for the Mean	
Lower Limit	49,65
Upper Limit	50,04

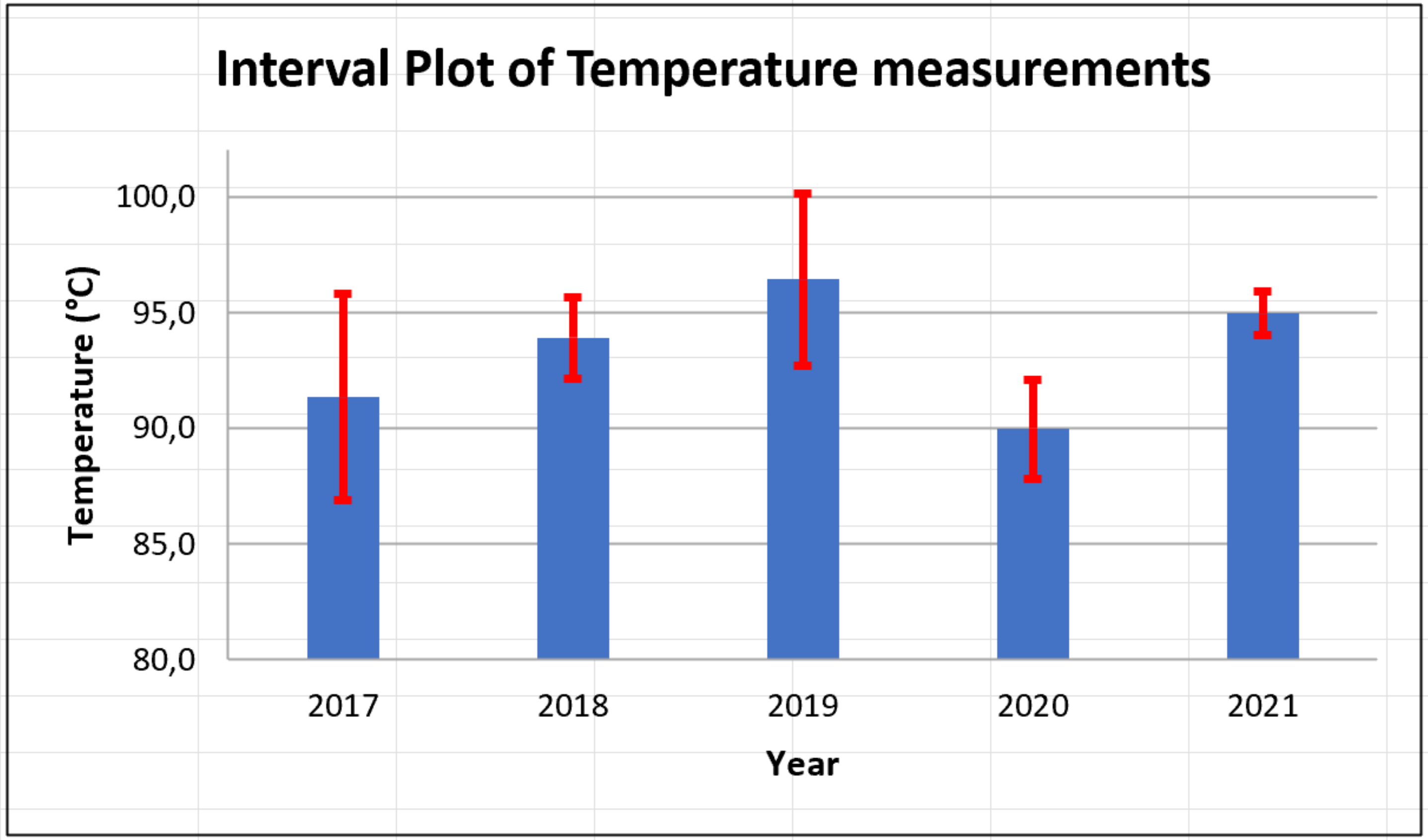
Now let's consider another case study that well illustrates the practical importance of using Confidence Intervals

INFERENTIAL STATISTICS

Let’s consider, for example, a retrospective analysis of temperature measurements (*e.g.*, for APQR) which should not exceed a limit of 100 °C.

Individually none of the values is equal or greater to 100°C but….

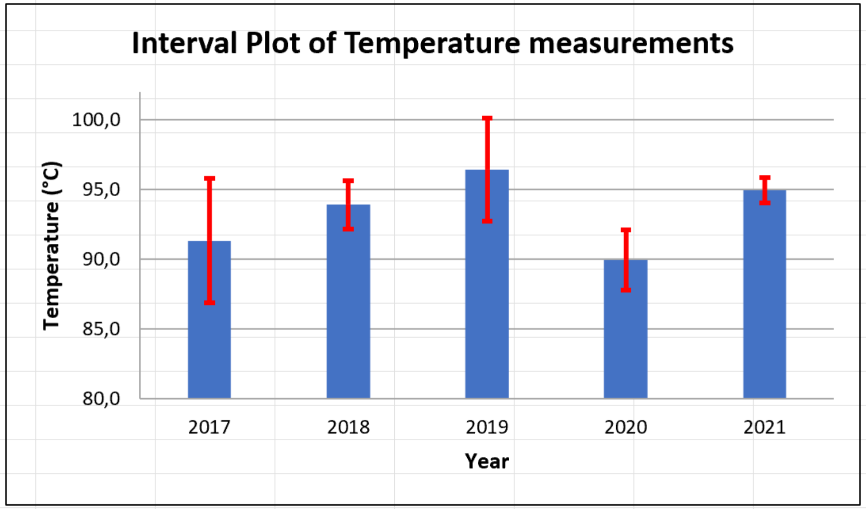
2017	2018	2019	2020	2021
91,0	97,0	98,8	93,2	95,0
93,8	90,8	99,4	91,0	95,7
97,4	91,8	98,0	87,1	94,2
95,4	96,7	89,5	88,5	
79,2	93,3			



INFERENTIAL STATISTICS

	2017	2018	2019	2020	2021
	91,0	97,0	98,8	93,2	95,0
	93,8	90,8	99,4	91,0	95,7
	97,4	91,8	98,0	87,1	94,2
	95,4	96,7	89,5	88,5	
	79,2	93,3			
Mean	91,4	93,9	96,4	90,0	95,0
Dev. Std	7,1894	2,8208	4,6522	2,7012	0,7506
CV%	**7,87**	3,00	4,82	3,00	0,79
Count	5	5	4	4	3
Confidence interval (95%)	8,9269	3,5025	7,4026	4,2983	1,8645

Year	Mean	Lower Conf. Interval Limit	Upper Conf. Interval Limit
2017	91,4	4,5	4,5
2018	93,9	1,8	1,8
2019	96,4	3,7	3,7
2020	90,0	2,1	2,1
2021	95,0	0,9	0,9



WARNING

- The example just shown does not apply only to a situation like the one described (*e.g.*, APQR) but also, for example, to the ***management of OOS***.
- An « anomalous data », in fact, is not so « anomalous » if the average of the population from which it derives is in an interval that exceeds a specific limit.

When investigating an OOS always look at the Confidence Interval !

WARNING !

Using Excel, it is also possible to calculate other statistical intervals such as those of Prediction and Tolerance.

However, their calculation has not been considered here as it is a bit more laborious, and this would have further burdened the presentation.

HYPOTHESIS TESTING

INFERENTIAL STATISTICS

Back to the introduction to Inferential Statistics methods, the second topic mentioned was:

Hypothesis Testing

The statistical verification of the hypotheses evaluates the degree of reliability that can be attributed to them in the face of the empirical evidence represented by the sample observations available.

We will see, once again, the practical utility of probability distributions!

INFERENTIAL STATISTICS

In practice:

- **Statistical hypothesis**: an ***assertion*** regarding the parameters of one or more populations that we want to test or investigate.

- **Hypothesis testing**: the ***procedure*** that leads to a decision concerning a particular hypothesis and is based on a random sample extracted from the population of interest.

INFERENTIAL STATISTICS

- **Null Hypothesis**: H_0, is the "*default hypothesis*", the "*thing that is accepted*", the currently accepted value for a certain parameter.

- **Alternative Hypothesis**: H_a or H_1 and also called, in some books, "*the research hypothesis*", involves the assertion to be tested.

Let's see a practical example

INFERENTIAL STATISTICS

Within a Company it is believed that, on the average, a given chemical process leads to 100 kg of API. A QA Officer claims that, after the last change to the equipment, the **average yield** is no longer 100 kg.

Statistical hypothesis: H_0: μ = 100 kg (**Null hypothesis**)

H_1: $\mu \neq$ 100 kg (**Alternative hypothesis**)

two-tails

Note :

- Hypotheses are always statements about the population or distribution being studied, NOT about the sample.
- ***H_0 and H_1 are mathematical opposites of one another and together they cover all possibilities !***

INFERENTIAL STATISTICS

- There are just two possible outcomes:
  - **Reject the Null Hypothesis**: we then believe H_1 to be the case
  - **Fail to reject the Null Hypothesis** : we basically keep H_0

How can we do the testing ?

How can we reject H_0 or not?

INFERENTIAL STATISTICS

With regard to our case study, let us first define some key points:

- it is a hypothesis test about a population mean, $\boldsymbol{\mu}$, that it is reasonable to assume is normally distributed
- we assume that the population variance, $\boldsymbol{\sigma}^2$, is unknown
- let's suppose we have a limited number of yield values, and this implies that the "test-statistic" to be used is the *t-statistic.*
- In practice we have only 15 yield values with an average yield $\bar{\boldsymbol{x}}$ = 101.2 Kg and a standard deviation $\boldsymbol{s}$ = 1.3 Kg.

INFERENTIAL STATISTICS

The ***test statistics t*** to be calculated is:

$$T = \frac{\bar{x} - \mu}{s \,/ \sqrt{n}} = \frac{101.2 - 100}{1.3 \,/ \sqrt{15}} = 3.575$$

While $-t_c$ and $+t_c$ are obtained using the **T.INV.2T function** which returns the two-tailed inverse of the Student's *t*-distribution.

Since *T* value falls outside the acceptance zone bounded by $-t_c$ and $+t_c$, there is evidence to reject the null hypothesis at $\alpha = 0.05$.

In other words:

the QA Officer was right !

	A	B	C	D
1				
2		**Terms of the problem**		
3				
4		μ_0 =	100,0	
5		$\mu_1 \neq$	100,0	
6		α =	0,05	
7				
8		**Experimental evidence**		
9				
10		n. of batches =	15	
11				
12		Average yield =	101,2	
13				
14		Standard deviation =	1,3	
15				
16		***tc* =**	**-2,145**	**2,145**
17				
18		**T =**	**3,575**	

INFERENTIAL STATISTICS

Let's remember the initial statistical hypothesis, *i.e.*:

H_0: μ = 100 kg (**Null hypothesis**)

H_1: $\mu \neq$ 100 kg (**Alternative hypothesis**)

two tails test

If, instead, the assumption of the QA Officer had been that the yield was greater than 100 Kg, how would have been H_0 and H_1? Simple:

H_0: $\mu \leq$ 100 kg (**Null hypothesis**)

H_1: $\mu >$ 100 kg (**Alternative hypothesis**)

one (right) tail test

and what would hypothesis testing be like?

INFERENTIAL STATISTICS

Again, the value of the *T*-test statistic would be the same as calculated before, *i.e.*, 3.575.

However, since in this case the test is "one side only", the $\boldsymbol{t_c}$ value will be calculated using the **T.INV function** which returns the inverse of the left tail Student's *t*-distribution.

Also, in this case the value of ***T*** falls beyond the limit corresponding to $\boldsymbol{t_c}$, and therefore there is evidence to reject the null hypothesis at α = 0.05.

In other words:

the QA Officer is still right !

H	I	J
Terms of the problem		
μ_0 =	100,0	
μ_1 >	100,0	
α =	0,05	
Experimental evidence		
n. of batches =	15	
Average yield =	101,2	
Standard deviation =	1,3	
tc =	**1,761**	
T =	**3,575**	

INFERENTIAL STATISTICS

- From what has just been shown, the power and usefulness of hypothesis testing for practical purposes clearly emerge.

- It is therefore worth seeing some other applications of practical use.

INFERENTIAL STATISTICS

Let's consider a validated tableting process that, under normal operating conditions, produces tablets with an average weight of 50.36 mg and a standard deviation of 2.235 mg.

During the production of a batch of tablets, 20 in-process samples are taken randomly, the weights of which are shown in the table on the side.

We want to test the hypothesis that the process is under control, namely that:

$$H_0: \mu = 50.36 \text{ mg} \quad vs. \quad H_1: \mu \neq 50.36 \text{ mg}$$

at a significance level of 5% ($\alpha = 0.05$) or, alternatively, at a confidence level of 95% or 0.95.

Tablet weight (mg)
47,98
51,85
48,53
49,69
50,46
50,90
53,14
57,32
48,90
53,72
51,16
49,91
53,42
46,08
49,41
51,24
47,00
53,16
52,69
48,17

INFERENTIAL STATISTICS

Let's consider a validated tableting process that, under normal operating conditions, produces tablets with an average weight of 50.36 mg and a standard deviation of 2.235 mg.

Since the standard deviation (or variance) of the population is known, the test statistic to use is:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

The sample mean can easily be obtained from the weight values using the Excel **AVERAGE** function while critical values for Z can be obtained using the Excel **INV.NORM.S** function.

INFERENTIAL STATISTICS

The ***test statistics t*** to be calculated is:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{50{,}74 - 50{,}36}{2{,}235 / \sqrt{20}} = 0{,}760$$

While $-z_c$ and $+ z_c$ are obtained using the **NORM.S.INV function** which returns the inverse of the standard normal cumulative distribution. The distribution has a mean of zero and a standard deviation of one.

H	I	J	K	L
Terms of the problem with known process variance				
μ_0 =	50,36			
$\mu_1 \neq$	50,36			
σ =	2,235			
α =	0,05			
Experimental evidence				
n. of batches =	20			
Average yield =	50,74			
- z(1-α/2)= - z(0.975) =	**-1,960**		**+ z(1-a/2)= + z(0.975) =**	**1,960**
Z =	**0,760**			

Since Z value falls <u>within</u> the acceptance zone bounded by $-z_c$ and $+z_c$, there is insufficient evidence to reject the null hypothesis at α= 0.05.

In other words:

based on the sample data the process is under control !

INFERENTIAL STATISTICS

Let's consider the example just seen assuming we don't know the standard deviation (or variance) of the process:

- it is a hypothesis test about a population mean, μ, that it is reasonable to assume is normally distributed
- the population variance, σ^2 , is unknown
- In practice we have 20 weight values with an average value of $\bar{x}$ = 50.74 mg and a standard deviation s = 2.6982 mg.
- Since we have a limited number of weight values, the "test-statistic" to be used is the *t-statistic.*

INFERENTIAL STATISTICS

The ***test statistics t*** to be calculated is:

$$T = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{50.74 - 50.36}{2.6982 / \sqrt{20}} = 0.6298$$

While $-t_c$ and $+t_c$ are obtained using the **T.INV.2T function** which returns the two-tailed inverse of the Student's *t*-distribution.

Since T value falls within the acceptance zone bounded by $-t_c$ and $+t_c$, there is insufficient evidence to reject the null hypothesis at α= 0.05.

In other words:

based on the sample data the process is under control !

O	P	Q	R
Terms of the problem with unknown process variance			
μ_0 =	50,36		
$\mu_1 \neq$	50,36		
α =	0,05		
Experimental evidence			
n. of batches =	20		
Sample Average weight =	50,74		
Sample Standard deviation =	2,698		
tc =	**-2,093**	**2,093**	
T =	**0,630**		

INFERENTIAL STATISTICS

Consider an automated manufacturing process that rejects tablets if they weigh less than 95 mg or more than 108 mg.

Out of 100 tablets we obtained: 3 tablets < 95 mg and 2 tablets > 108 mg.

with this information alone we can estimate the average and standard deviation of the production process that generated it!

In fact, assuming that the weights of the tablets are normally distributed, which is reasonable, then….

INFERENTIAL STATISTICS

$$\begin{cases} P\,(w < 95 \text{ mg}) = \Phi\left(\frac{95-\mu}{\sigma}\right) \\ P\,(w > 108 \text{ mg}) = 1 - \Phi\left(\frac{108-\mu}{\sigma}\right) \end{cases} \Rightarrow \begin{cases} \Phi\left(\frac{95-\mu}{\sigma}\right) = 0.03 \\ 1 - \Phi\left(\frac{108-\mu}{\sigma}\right) = 0.02 \end{cases}$$

from which it follows that:

$$\begin{cases} 95 - \mu = \sigma\ Z_{0.03} \\ 108 - \mu = \sigma\ Z_{0.98} \end{cases} \Rightarrow \begin{cases} 95 - \mu = \sigma\,(1.88) \\ 108 - \mu = \sigma\,(2.05) \end{cases} \Rightarrow \begin{matrix} \boldsymbol{\mu = 101.22 \text{ mg}} \\ \boldsymbol{\sigma = 3.31 \text{ mg}} \end{matrix}$$

where $Z_{0.03}$ hand $Z_{0.98}$ have been calculated using the Excel **NORM.S.INV** function

INFERENTIAL STATISTICS

Everything seen so far has shown how the **STATISTICAL HYPOTHESIS TEST** can be useful in many practical cases:

- ***"infer" from experimental data crucial information on the state of a process***
- ***check if a certain "parameter" lies within the confidence interval*** (typical application: determining if a result is an OOS)
- ***compare the mean values or the spreads of two or more datasets*** (typical applications of this are in: suppliers' validation, comparison of analytical data generated by different methods, *etc.*)

INFERENTIAL STATISTICS

1-Sample t test, 2-Sample t test and

2-Variances test

INFERENTIAL STATISTICS

Hypothesis tests, such as those just also allow to establish if:

- The mean of a sample differs significantly from a specified value → ***1-Sample t test***
- Two data group means are different → ***2-Sample t test***
- The variances, or the standard deviations of two data groups differ → ***2 Variances test***

1-Sample t test

Null hypothesis:

$H_0: \mu = \mu_0$ The population mean (μ) equals the hypothesized mean (μ_0)

Alternative hypothesis:

$H_1: \mu \neq \mu_0$ The population mean (μ) differs from the hypothesized mean (μ_0)

$H_1: \mu > \mu_0$ The population mean (μ) is greater than the hypothesized mean (μ_0)

$H_1: \mu < \mu_0$ The population mean (μ) is less than the hypothesized mean (μ_0)

INFERENTIAL STATISTICS

2-Sample t test

Null hypothesis

$H_0: \mu_1 - \mu_2 = 0$ The difference between the population means ($\mu_1 - \mu_2$) equals zero

Alternative hypothesis

$H_1: \mu_1 - \mu_2 \neq 0$ The difference between the population means ($\mu_1 - \mu_2$) does not equal zero

$H_1: \mu_1 - \mu_2 > 0$ The difference between the population means ($\mu_1 - \mu_2$) is greater than zero

$H_1: \mu_1 - \mu_2 < 0$ The difference between the population means ($\mu_1 - \mu_2$) is less than zero

2-Variances test

Null hypothesis

$H_0: \sigma_1 / \sigma_2 = 1$ The ratio between the first population standard deviation (σ_1) and the second population standard deviation (σ_2) is equal to 1.

Alternative hypothesis

$H_1: \sigma_1 / \sigma_2 \neq 1$ The ratio between the first population standard deviation (σ_1) and the second population standard deviation (σ_2) does not equal 1

$H_1: \sigma_1 / \sigma_2 > 1$ The ratio between the first population standard deviation (σ_1) and the second population standard deviation (σ_2) is greater than 1

$H_1: \sigma_1 / \sigma_2 < 1$ The ratio between the first population standard deviation (σ_1) and the second population standard deviation (σ_2) is less than 1

Let's see a few practical examples

INFERENTIAL STATISTICS

Let's consider six HPLC assay values within specs (NLT 100%) and one "borderline" value (99,85%).

Is this an OOS result, or does it belong to the same population of the other values ?

Assay values (%)			
100,05			
100,00	**Mean ($\bar{x}$)**	100,05	= AVERAGE(C5:C10)
100,07	**Std. Dev. (s)**	0,0362	= STDEV.S(C5:C10)
100,10	**Count**	6	= COUNT(C5:C10)
100,02	**Standard Error of Mean (SEM)**	0,0148	= F7/(SQRT(F8))
100,03	**Degrees of freedom (dof)**	5	= F8-1
	Hypothesized mean (μ)	99,85	
	t-statistic	13,19698	= (F6-F11)/F9
	P-value **(two-tail test)**	0,0000	= T.DIST.2T(F13;F10)

1 Sample t-test

Since *P-value* < 0.05 there is evidence enough to reject the Null Hypothesis, *i.e.,* H_0:μ = 99.85 or Mean Assay value = 99.85

INFERENTIAL STATISTICS

Let's consider two series of pH values, one determined in-house on real samples and the other reported on the corresponding CoAs provided by the supplier together with the samples.

	Sodium Acetate pH values	
	In-house	**Supplier's CoA**
Sample 1	8.1	8.1
Sample 2	8.3	8.1
Sample 3	8.2	8
Sample 4	8.5	8.4
Sample 5	8.5	8.4
Mean value	**8.32**	**8.20**

On the average are the two series of data here above reported, statistically different or not?

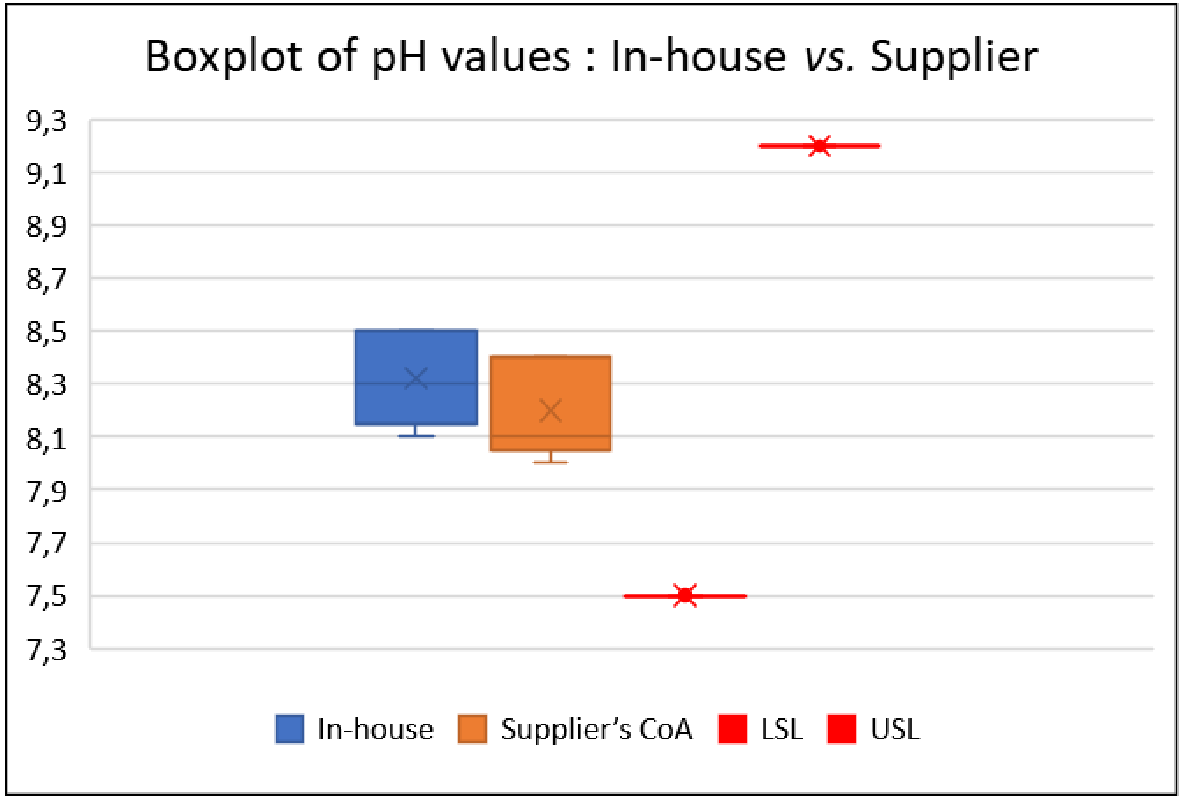
INFERENTIAL STATISTICS

Let's first look at data visualization using boxplots.

Both datasets are within specs and box widths look rather similar.

Apart from this, we cannot say much more.

*The **t-test** can tell us whether the two mean values are statistically different or not, but before applying it, it must be established whether the variances of the two populations significantly differ from each other or not. In fact, there are two possible types of t-tests !*



INFERENTIAL STATISTICS

B	C	D
	In-house	Supplier's CoA
	8,1	8,1
	8,3	8,1
	8,2	8,0
	8,5	8,4
	8,5	8,4
Mean =	8,3	8,2
Variance =	0,032	0,035

F-Test Two-Sample for Variances

	Supplier's CoA	*In-house*
Mean	8,20	8,32
Variance	0,035	0,032
Observations	5	5
df	4	4
F	**1,0938**	
P(F<=f) one-tail	0,4664	
F Critical one-tail	**6,3882**	

Examination of the variances in the two samples shows that one is numerically greater. The F-test is then performed using this as the first sample. **THIS IS VERY IMPORTANT IN EXCEL !!**

The outcome of the test does not show a significant difference in the variances of the two populations and therefore we will be able to apply the ***t-test assuming equal variances***.

INFERENTIAL STATISTICS

*Since the value of the **t-test statistic** (1.0366) is found to be **within** the **two-tailed critical t interval** (-2.3060, +2.3060), at the 5% significance level (or 95% confidence) we can say that there is **no significant difference** between the two mean values.*

t-Test: Two-Sample Assuming Equal Variances		
	In-house	*Supplier's CoA*
Mean	8,32	8,20
Variance	**0,032**	**0,035**
Observations	5	5
Pooled Variance	0,0335	
Hypothesized Mean Difference	0	
df	8	
t Stat	**1,0366**	
P(T<=t) one-tail	0,1651	
t Critical one-tail	1,8595	
P(T<=t) two-tail	0,3302	
t Critical two-tail	**2,3060**	

INFERENTIAL STATISTICS

Instead, let's now consider the data in the table on the side relating to a different supplier (Supplier 1).

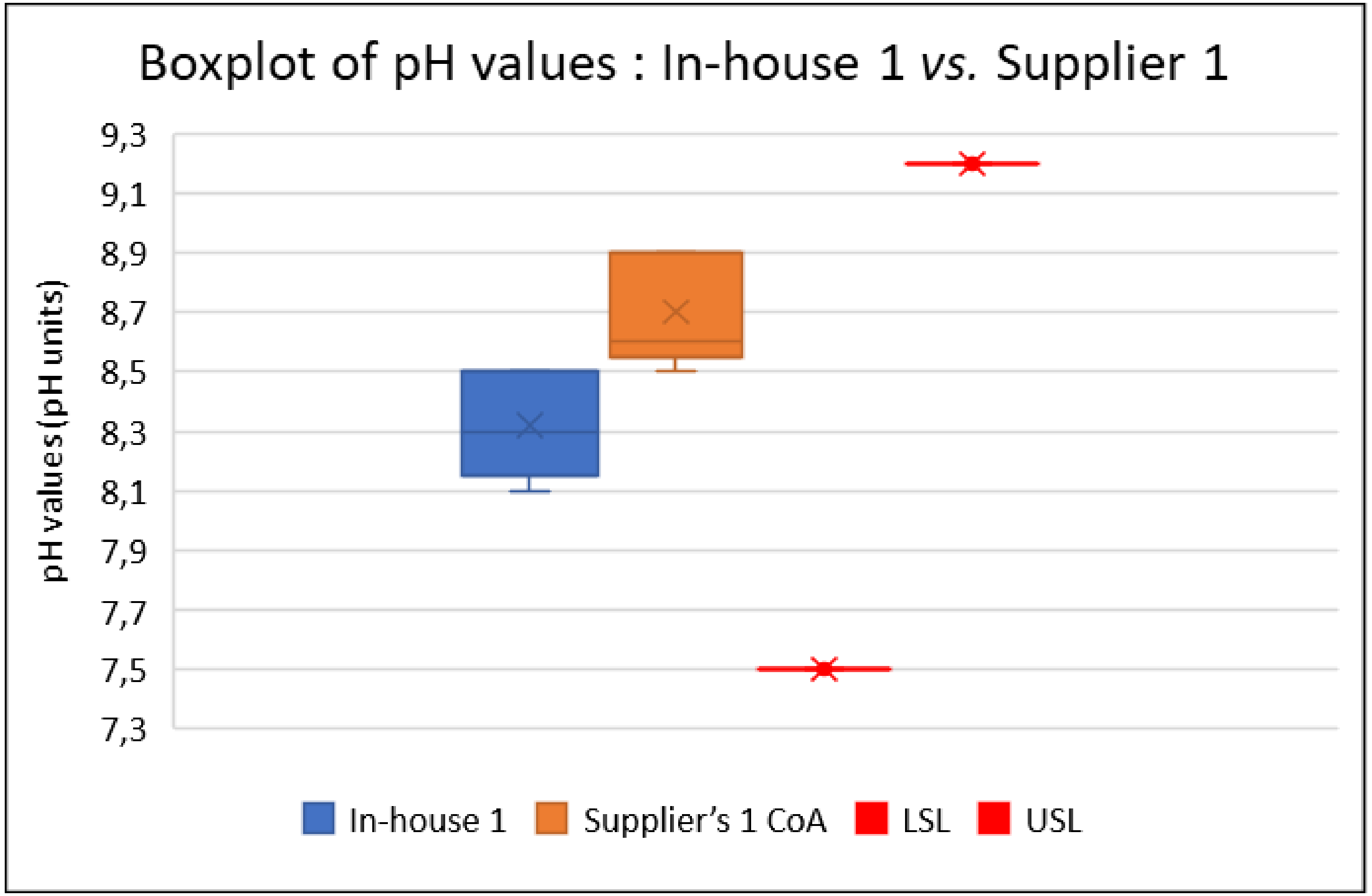
	Sodium Acetate pH values	
	In-house 1	**Supplier's 1 CoA**
Sample 1	8.1	8.6
Sample 2	8.3	8.6
Sample 3	8.2	8.5
Sample 4	8.5	8.9
Sample 5	8.5	8.9
Mean value	**8.32**	**8.70**

Again: are the two mean values here above reported, statistically different or not?

INFERENTIAL STATISTICS

In this case it is evident that the two pH data distributions are shifted from each other. However, box widths are still comparable ⇨ data spreads look similar.

Only the t-test can confirm whether the two average values are significantly different or not, but, once again, to apply the correct one, we must first establish whether the variances of the two populations can be considered equal or not.



INFERENTIAL STATISTICS

	In-house 1	Supplier's 1 CoA
	8,1	8,6
	8,3	8,6
	8,2	8,5
	8,5	8,9
	8,5	8,9
Mean =	8,3	8,7
Variance =	0,032	0,035

F-Test Two-Sample for Variances

	Supplier's 1 CoA	*In-house 1*
Mean	8,70	8,32
Variance	0,035	0,032
Observations	5	5
df	4	4
F	**1,0938**	
P(F<=f) one-tail	0,4664	
F Critical one-tail	**6,3882**	

As also for the previous case, the examination of the variances in the two samples shows that one is numerically greater. The F-test is then performed using this as the first sample.

The outcome of the test does not show a significant difference in the variances of the two populations and therefore we can apply the ***t-test assuming equal variances***.

INFERENTIAL STATISTICS

In this case, since the value of the ***t-test statistic*** (-3.2827) is **outside** the ***two-sided critical t interval*** (-2.3060, +2.3060), at the level of significance of 5% (or 95% confidence) it can be said that **there is a significant difference between the two mean values**.

t-Test: Two-Sample Assuming Equal Variances		
	In-house 1	*Supplier's 1 CoA*
Mean	8,32	8,70
Variance	**0,032**	**0,035**
Observations	5	5
Pooled Variance	0,0335	
Hypothesized Mean Difference	0	
df	8	
t Stat	**-3,2827**	
P(T<=t) one-tail	0,0056	
t Critical one-tail	1,8595	
P(T<=t) two-tail	0,0111	
t Critical two-tail	**2,3060**	

INFERENTIAL STATISTICS

Summing up:

- in both cases no significant difference was observed in the variances of the populations from which the samples under study were extracted and therefore the ***t-test for equal variances*** was always applied
- unlike the first case, in the second a significant difference was observed between the averages of the values measured at home and those reported on the CoA of Supplier 1.

A possible hypothesis could be that Supplier 1 uses a different method than the in-house one which systematically overestimates the values … *but this is a matter for another investigation* ☺

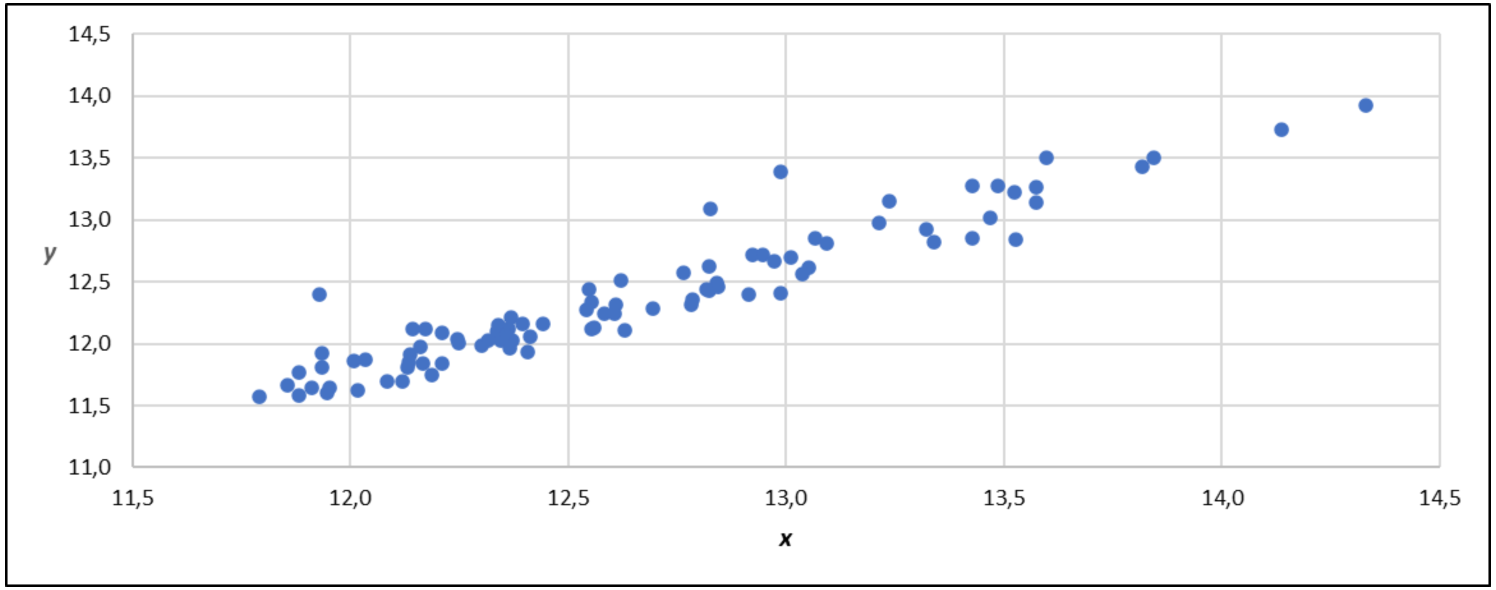
LINEAR REGRESSION

LINEAR REGRESSION

The objective of **Ordinary**, or simple, **Linear Regression** (OLR) is to mathematically describe the effect of an independent variable X (aka, *predictor, regressor* or *explanatory variable*) on a dependent variable Y (aka, *response, outcome*) using a formula which shows what happens to variable Y when the variable X changes.

LINEAR REGRESSION

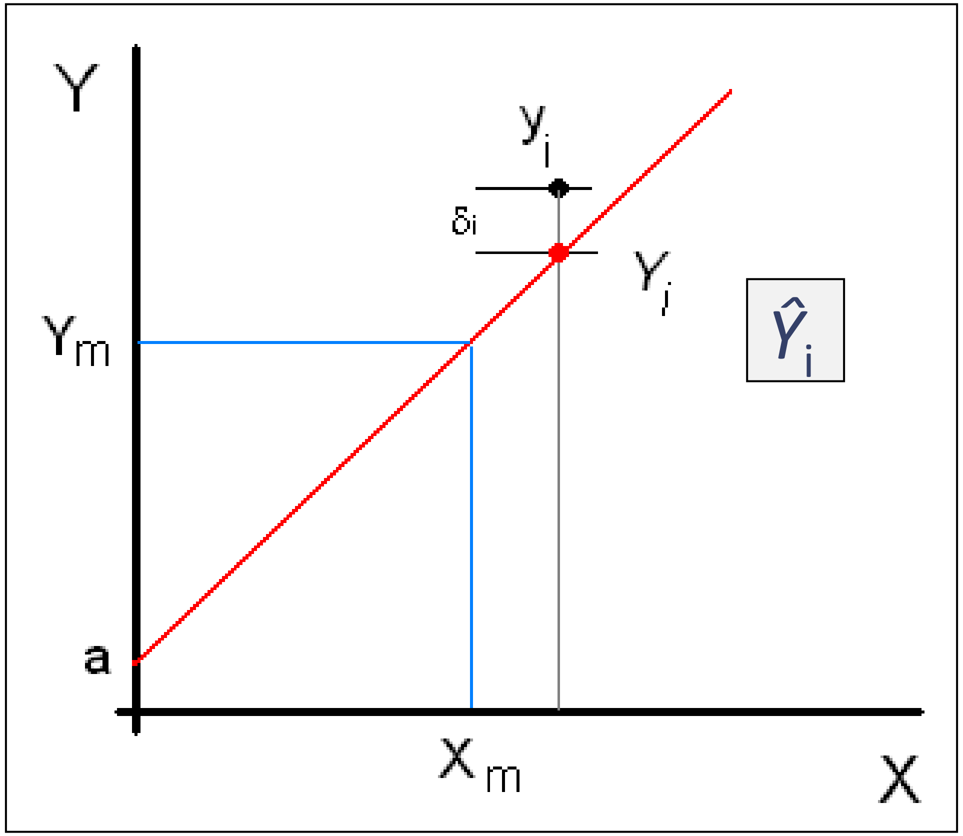
Since data pairs usually appear as a cloud of points like that shown here on the side, the problem is to find the so called *best-fit line* also known as *regression line*.



To obtain this line, OLR uses the so-called ***Least Squares Method*** which minimizes the distance between the experimentally measured data and the straight line we are looking for.

LINEAR REGRESSION

The classical regression line, or **Ordinary Least-Squares Regression** (**OLR** or **LSR**), is based on the minimization of the sum of the squares of the differences between the observed values of Y (y_i) and those estimated by the regression line ($\hat{y}_i$) relative to the variable Y only.



Residual or **Residual Error**

$$\delta_i = y_i - \hat{y}_i$$

y_i = experimental data
$\hat{y}_i$ = calculated value

LINEAR REGRESSION

Regression line equation

$$\boxed{\hat{\boldsymbol{y}} = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{x}}$$

Line intercept

$$\boxed{a = \bar{y} - b\bar{x}}$$

Line Slope or Regression coefficient

$$\boxed{b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{Cov(X,Y)}{\sigma_X^2}}$$

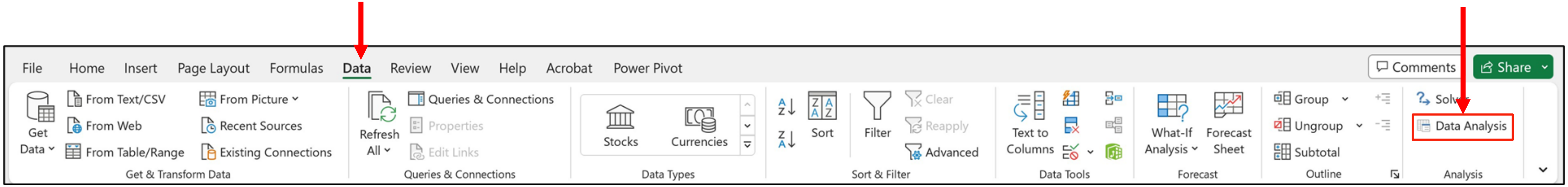
LINEAR REGRESSION

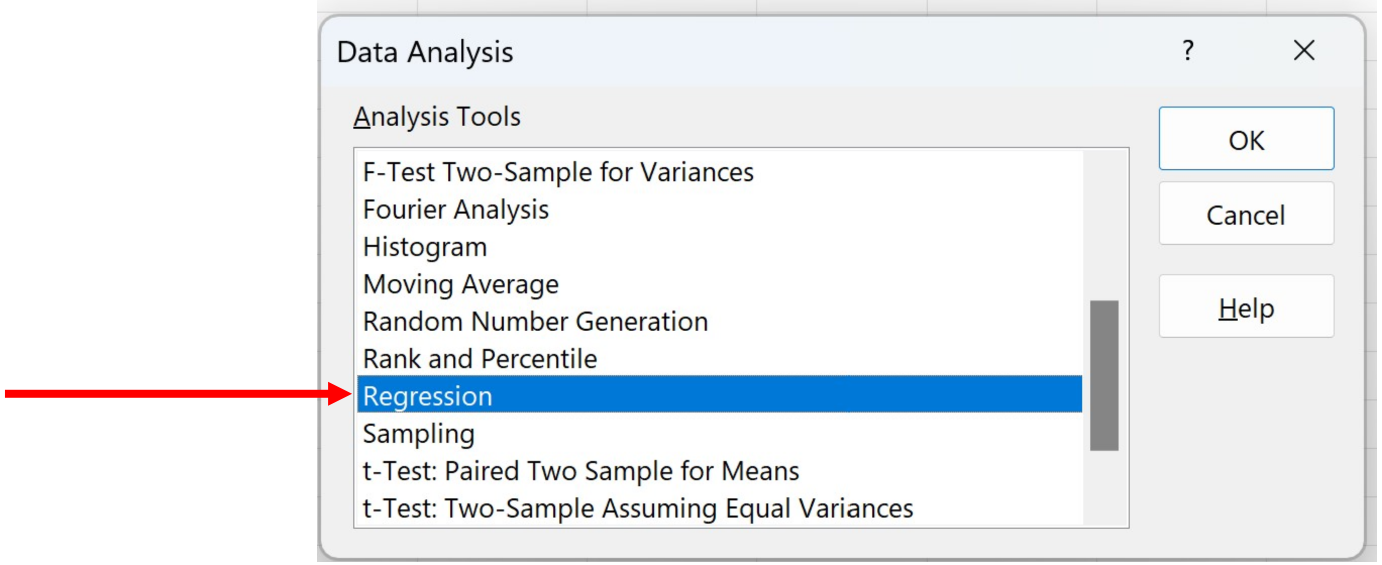
The fact that OLR is based on minimizing the sum of squared deviations, or "residuals“, only in the « *y* direction» has profound practical implications:

- **If we invert the two variables *x* and *y,* we obtain a different Least Squares Regression line.**
- **Understanding the properties of residuals is vital in determining whether the model is good or not.**
- **It is desirable that the residues be small and undistorted (or *unbiased*).**
- **The model is susceptible to *outliers* and *anomalous data.***

LINEAR REGRESSION

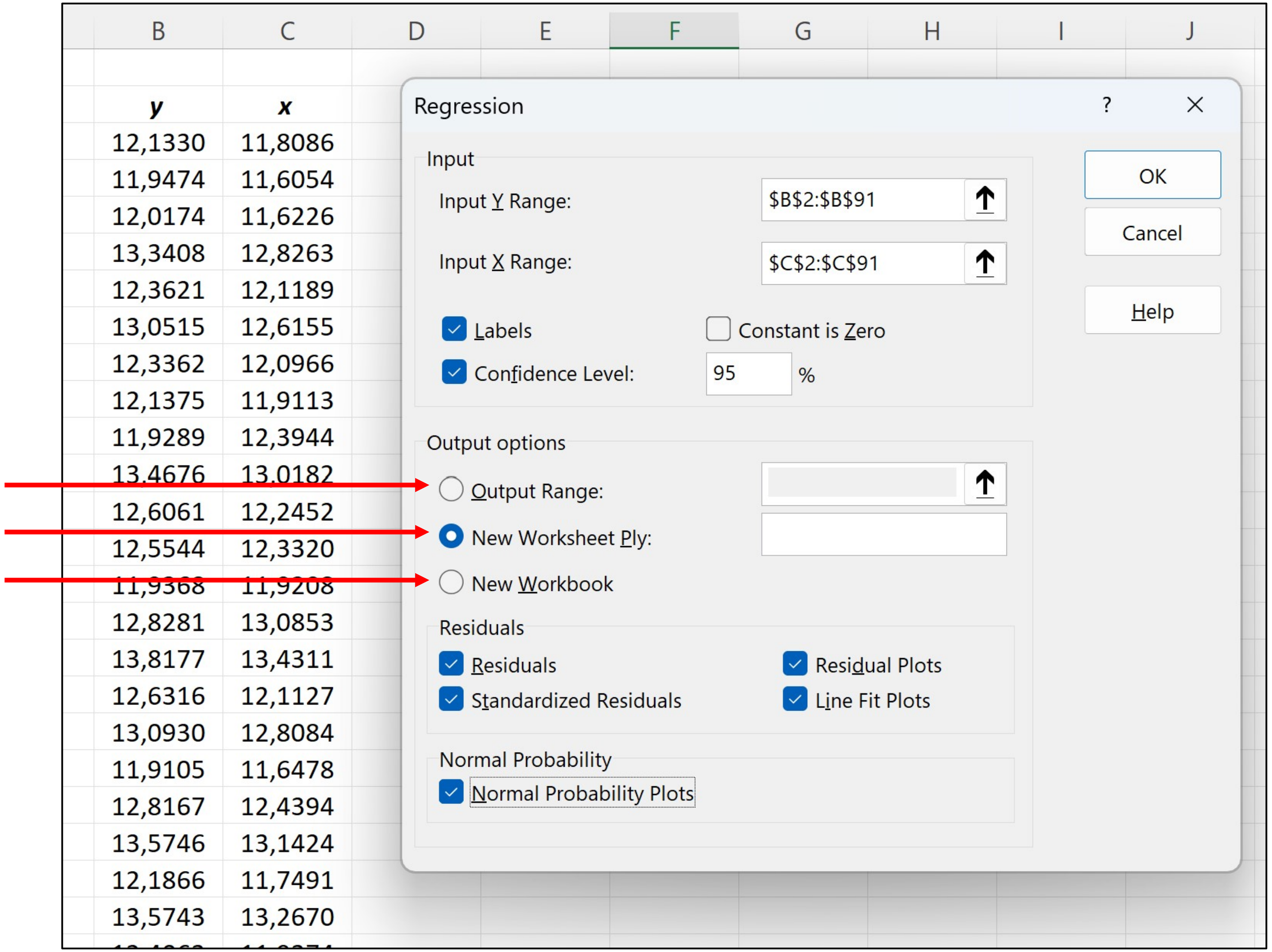
For *regression analysis* it must be used the "*regression tool*" accessible from "Data Analysis"





LINEAR REGRESSION

Regression analysis results can be obtained on the same worksheet, in a new worksheet or even in a new workbook selecting the appropriate output option.



LINEAR REGRESSION

y	x
12,1330	11,8086
11,9474	11,6054
12,0174	11,6226
13,3408	12,8263
12,3621	12,1189
13,0515	12,6155
12,3362	12,0966
12,1375	11,9113
11,9289	12,3944
13,4676	13,0182
12,6061	12,2452
12,5544	12,3320
11,9368	11,9208
12,8281	13,0853
13,8177	13,4311
12,6316	12,1127
13,0930	12,8084
11,9105	11,6478
12,8167	12,4394
13,5746	13,1424
12,1866	11,7491
13,5743	13,2670
12,4063	11,9374
13,4878	13,2710
12,6940	12,2805
12,1598	11,9745
12,7859	12,3615
12,1670	11,8395
12,1212	11,6923

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,949704
R Square	0,901938
Adjusted R	0,900811
Standard E	0,180967
Observatio	89

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	26,20543	26,20543	800,1923	1,21E-45
Residual	87	2,849155	0,032749		
Total	88	29,05458			

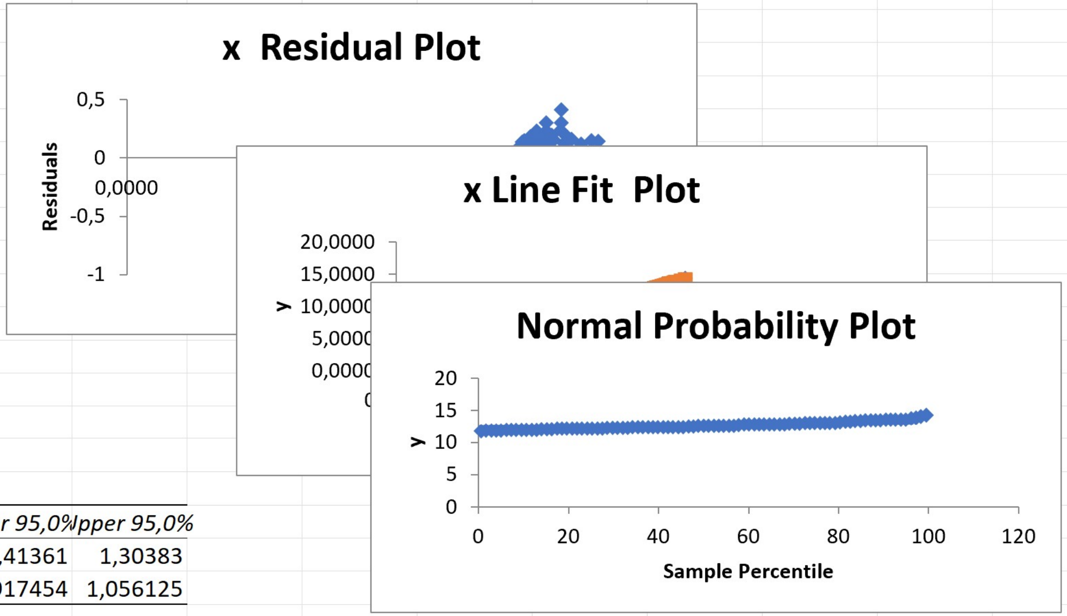
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*Ipper 95,0%*
Intercept	0,44511	0,432037	1,03026	0,305744	-0,41361	1,30383	-0,41361	1,30383
x	0,98679	0,034884	28,28767	1,21E-45	0,917454	1,056125	0,917454	1,056125

RESIDUAL OUTPUT

Observation	*Predicted y*	*Residuals*	*dard Residuals*
1	12,0977	0,035349	0,196452
2	11,89724	0,050149	0,278703
3	11,91419	0,103181	0,573431
4	13,10192	0,238877	1,327569
5	12,40396	-0,04188	-0,23274

PROBABILITY OUTPUT

Percentile	*y*
0,561798	11,79134
1,685393	11,85647
2,808989	11,88123
3,932584	11,88247
5,05618	11,91049



LINEAR REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,9497
R Square	0,9019
Adjusted R Square	0,9008
Standard Error	0,1810
Observations	89

This section contains *summary indices* such as ***R square*** which is used as an index of the goodness of the regression curve. ***Multiple R*** is the square root of R square and is a "sample correlation coefficient”. ***Adjusted R square*** is R square but adjusted for the number of terms in the model.

The ***Standard Error*** or Standard Error of Estimates (SEE) measures the variability (standard deviation) of the observed values (data) around the regression line. ***The higher it is, the further the experimental data are from the regression line !***

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	26,2054	26,2054	800,1923	0,0000
Residual	87	2,8492	0,0327		
Total	88	29,0546			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	0,4451	0,4320	1,0303	0,3057	-0,4136	1,3038	-0,4136	1,3038
x	0,9868	0,0349	28,2877	0,0000	0,9175	1,0561	0,9175	1,0561

This Standard Error is instead the ***Standard Error of the sampling distribution***

LINEAR REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,9497
R Square	0,9019
Adjusted R Square	0,9008
Standard Error	0,1810
Observations	89

Regression Sum of Squares represents the variability that the model explains. The bigger, the better.

Residual Sum of Squares represents the variability that the model does not explain. The smaller, the better.

Total Sum of Squares represents the total variability due to the dependent variable

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	26,2054	26,2054	800,1923	0,0000
Residual	87	2,8492	0,0327		
Total	88	29,0546			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	0,4451	0,4320	1,0303	0,3057	-0,4136	1,3038	-0,4136	1,3038
x	0,9868	0,0349	28,2877	0,0000	0,9175	1,0561	0,9175	1,0561

ESTIMATION OF THE GOODNESS OF FIT MODEL

ESTIMATION OF THE GOODNESS OF FIT MODEL

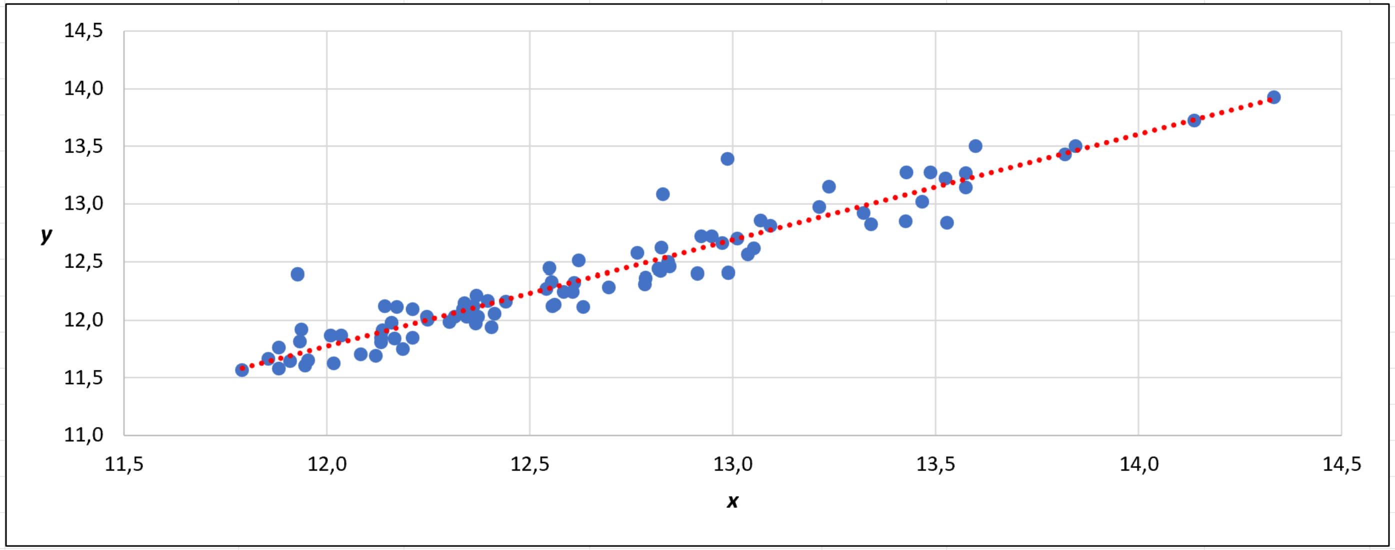
- Residuals represent the difference between the real value of the dependent variable (Y) and the model predicted value (predicted Y or **Ŷ**)

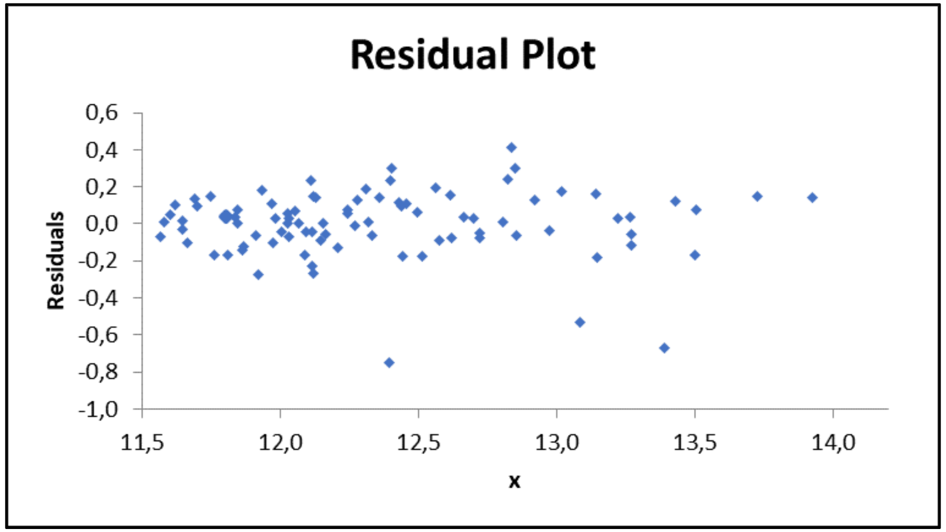
- Residues should have the following characteristics:
  - have an average value of zero
  - be independent and «normally distributed» (or, better, they do not display any patterns)

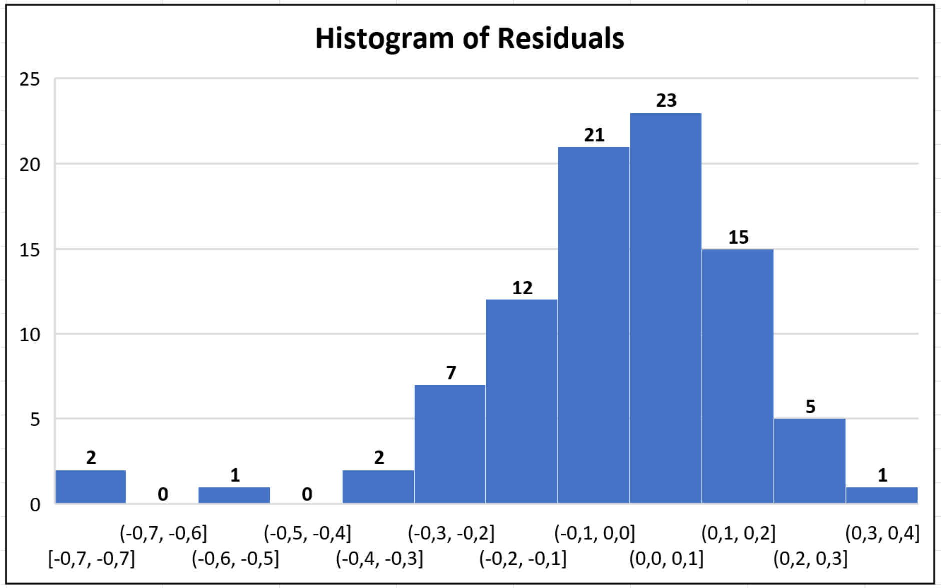
- In general. the value of $Residue \;=\; \boldsymbol{y_i} \;-\; \widehat{y}_i$ is plotted *vs.* $\widehat{y}_i$ or $\boldsymbol{x_i}$

  observed - **calculated**

ESTIMATION OF THE GOODNESS OF FIT MODEL



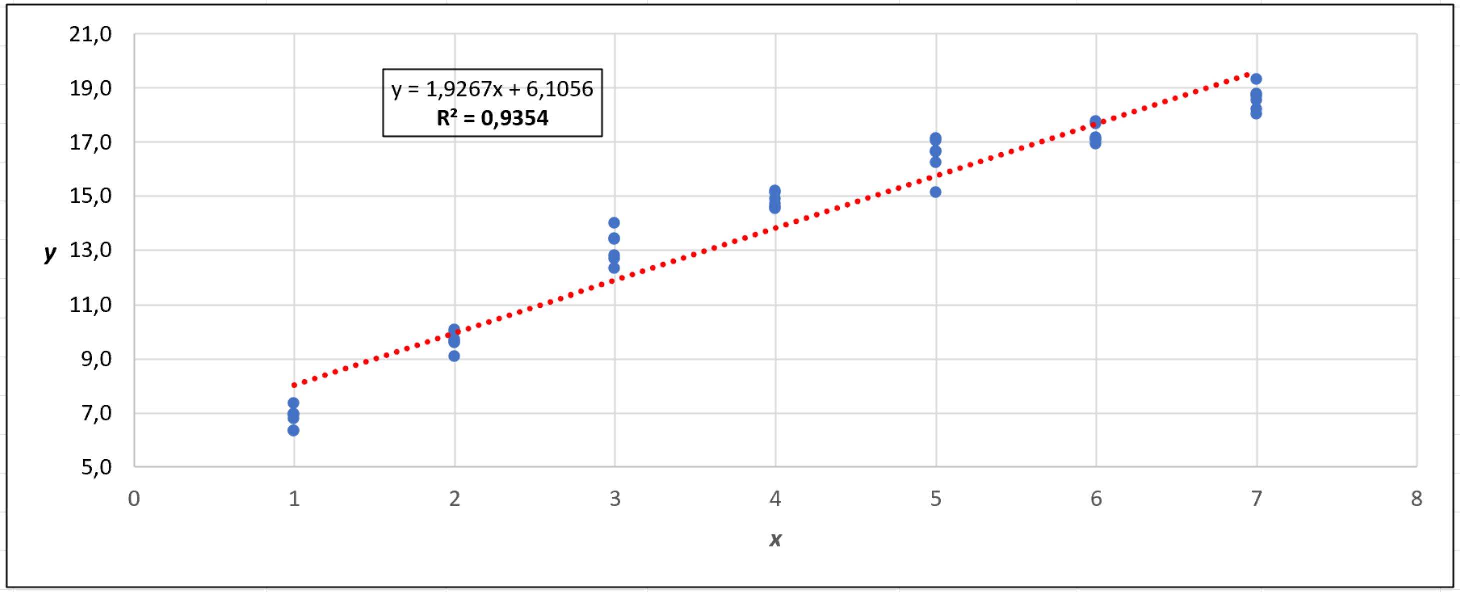


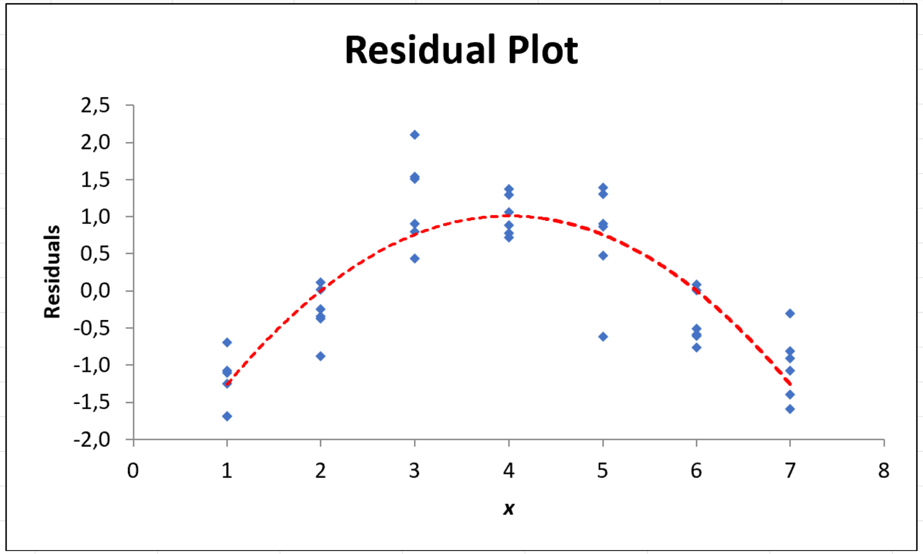


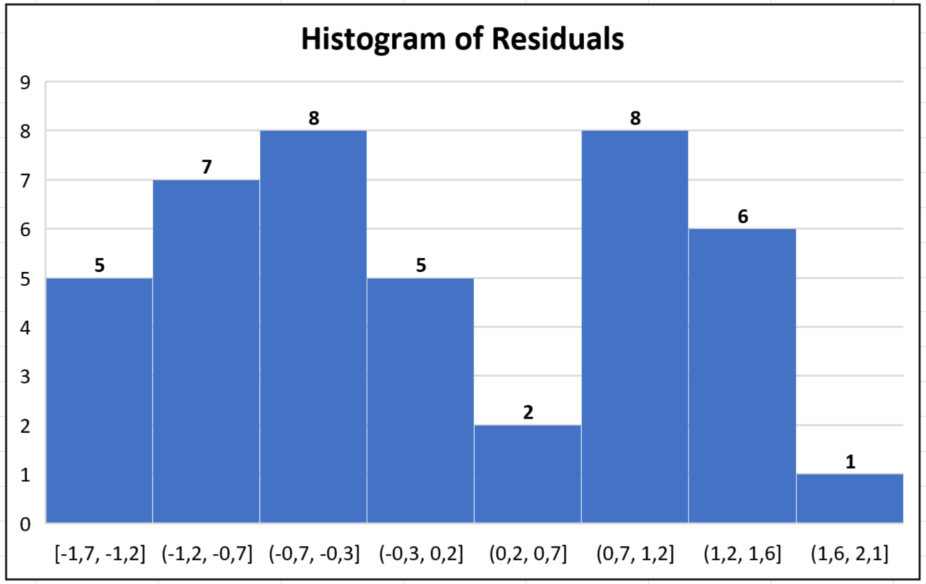
The residuals do not show any pattern!

NO lack-of-fit

ESTIMATION OF THE GOODNESS OF FIT MODEL







The residuals show a pattern!

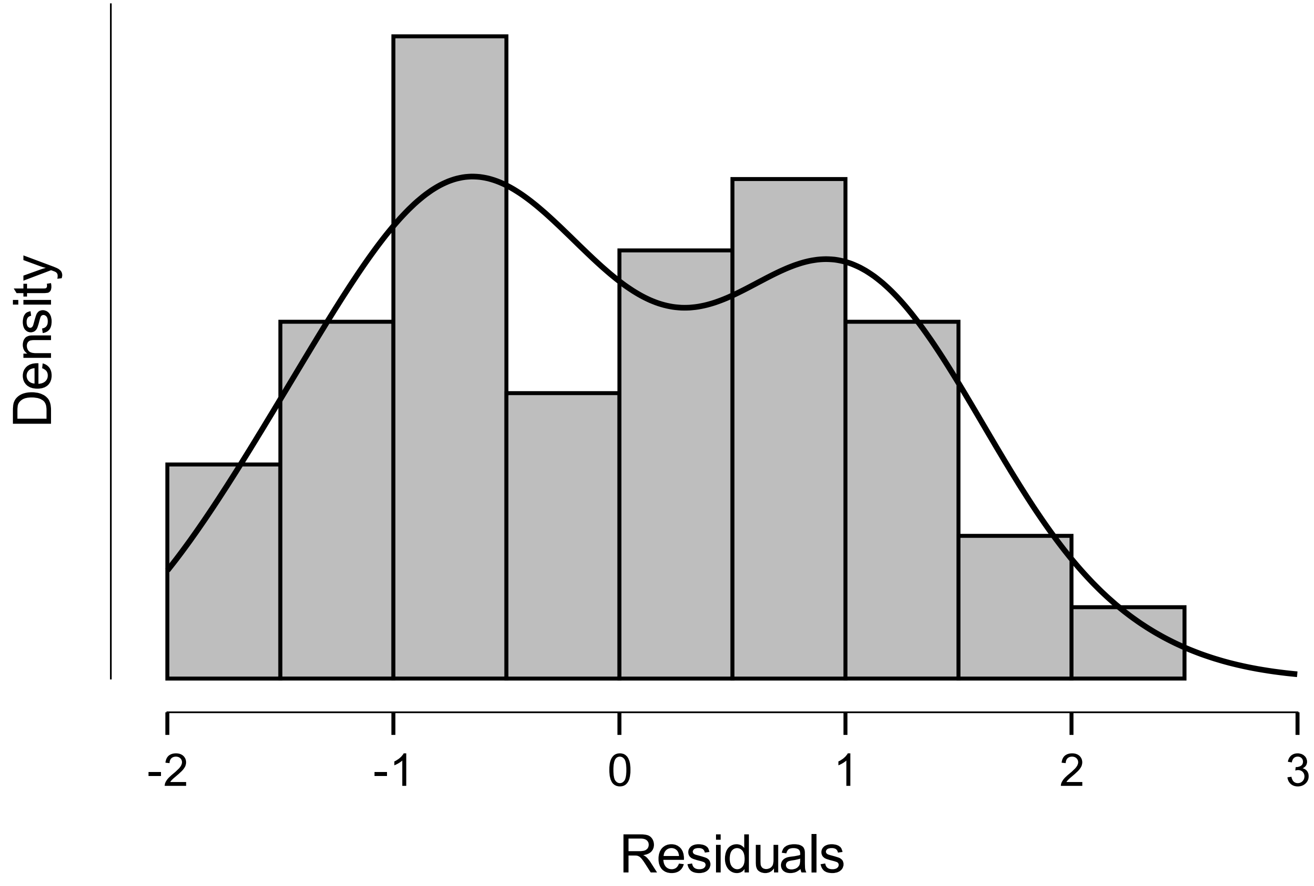
Lack-of-fit

ESTIMATION OF THE GOODNESS OF FIT MODEL

Residuals plot consisting of

Histogram + density curve

obtained using *JASP 0.17.2*



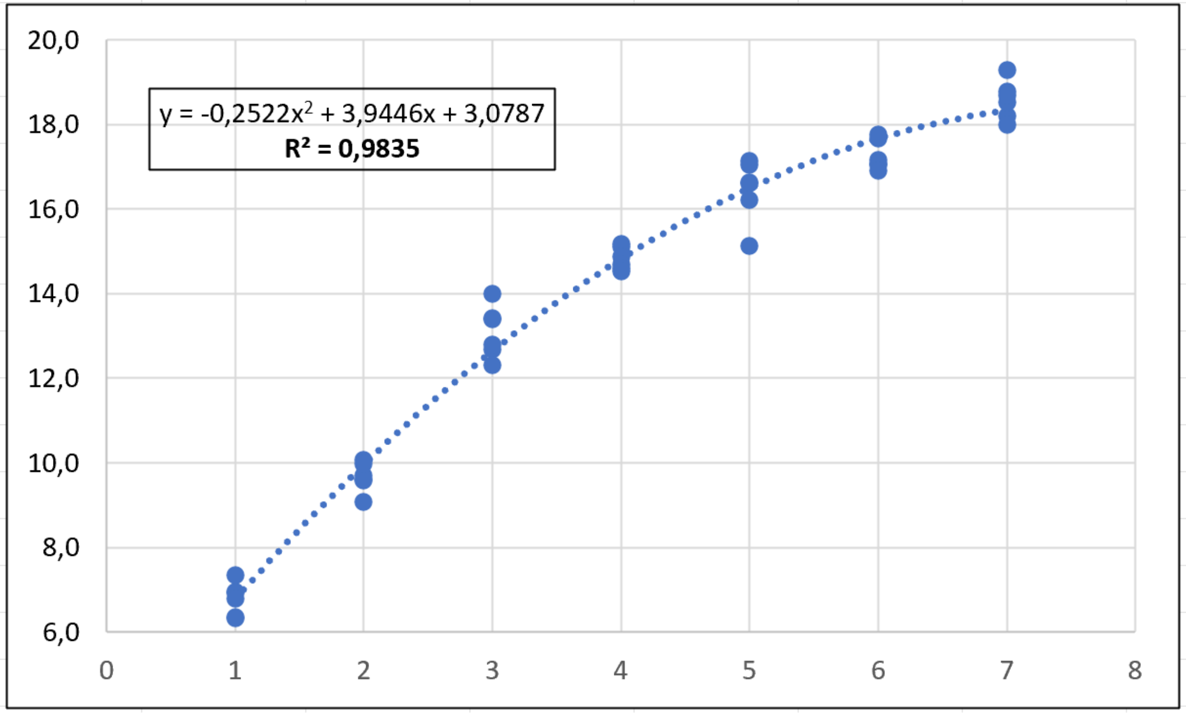
ESTIMATION OF THE GOODNESS OF FIT MODEL

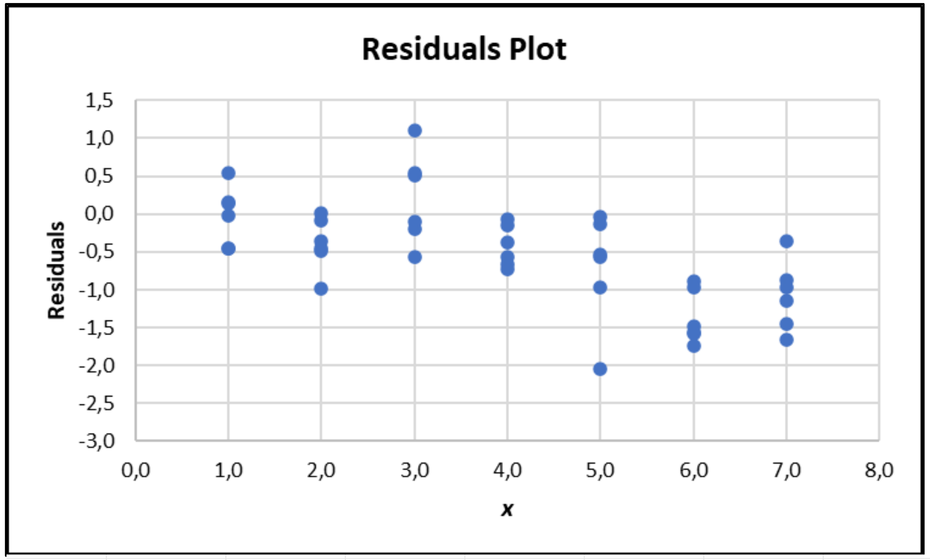
Lack-of-fit means curvature in data.

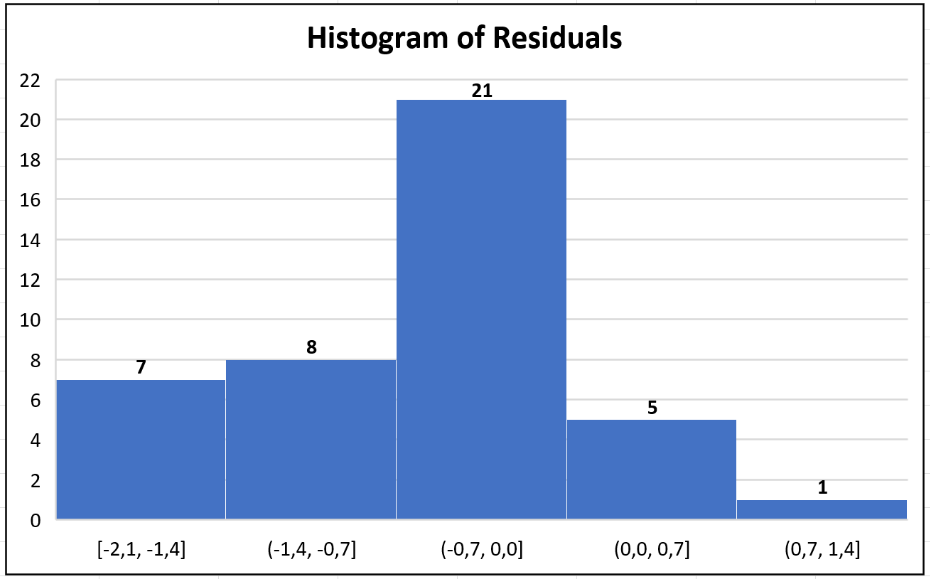
What to do ?

SIMPLE : add a quadratic term !

ESTIMATION OF THE GOODNESS OF FIT MODEL







NO lack-of-fit

CONCLUSIONS

CONCLUSIONS

- Microsoft Excel® is undoubtedly the simplest, most widespread and most used "data management" program in companies, including those in the chemical-pharmaceutical sector.

- Even if it is not a specific software for the statistical field, Excel allows you to do a lot and at "almost zero" cost.

- Although we have seen many applications, there are still many that we cannot cover here due to time constraints, ***but not only....***

CONCLUSIONS

- Excel has in fact numerous limitations precisely because it was originally developed for other purposes and only subsequently also adapted for statistical purposes. An example for all can be the control charts and, in particular, those divided by year.

- However, there is no doubt that its constant use would greatly increase the knowledge of the processes through the data they generate, would keep them better under control and would also find ideas for their improvement.

REFERENCES

J. Schmuller, *Statistical Analysis with Excel® for dummies*, 5th Edition, Wiley (2022)

M. Alexander, D. Kuleiska, *Microsoft ® Excel ® 365 Bible*, Wiley (2022)

D. Giuliani, M.M. Dickson, *Analisi Statistica con Excel ®*, Maggioli Editore (2015)

G. Bonollo, *Applicazioni Statistiche con Excel*, FrancoAngeli (2004)

F. Kronthaler, *Statistics Applied with Excel®*, Springer (2021)