

**APPLIED STATISTICS
FOR QA & QC
IN A GMP ENVIRONMENT**

TABLE OF CONTENTS

- INTRODUCTION
- DESCRIPTIVE STATISTICS : DATA AND THEIR SYNTHESIS
- INFERENTIAL STATISTICS : FROM DATA TO THEIR GENERATING MODEL
- CONTROL CHARTS
- CAPABILITY ANALYSIS
- CONCLUSIONS

INTRODUCTION

INTRODUCTION

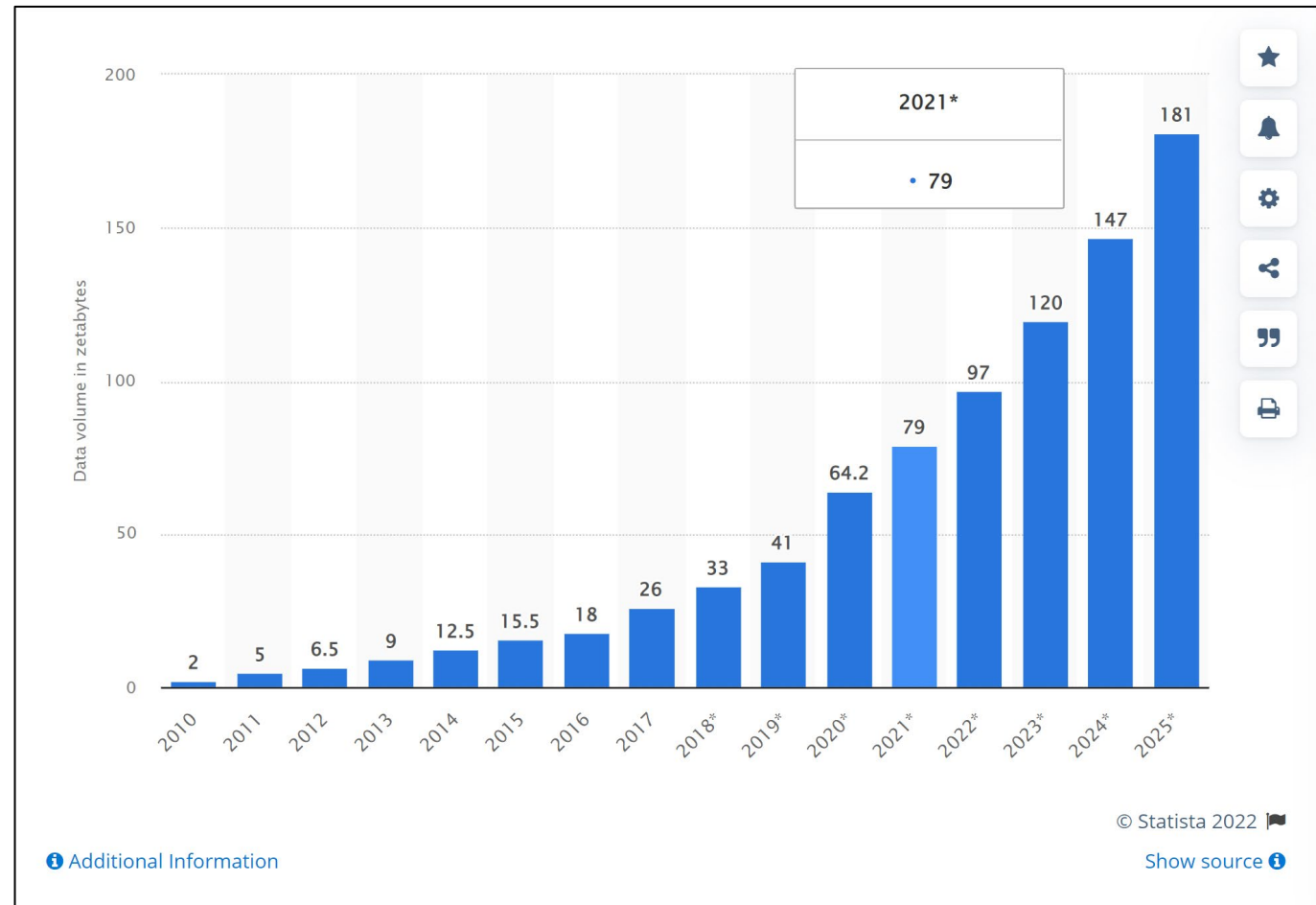
- Since mid-20th century we all live in the so-called **INFORMATION AGE** (aka **COMPUTER AGE** or **DIGITAL AGE**) which is characterized by a huge amount of data.
- According to market intelligence companies the total amount of data created, captured, copied, and consumed in 2021 reached 79 zettabytes.

79 zettabytes = $79 \cdot 10^{21}$ bytes

Name	Symbol	Multiple
<u>chilobyte</u>	kB	<u>10^3</u>
<u>megabyte</u>	MB	<u>10^6</u>
<u>gigabyte</u>	GB	<u>10^9</u>
<u>terabyte</u>	TB	<u>10^{12}</u>
<u>petabyte</u>	PB	<u>10^{15}</u>
<u>exabyte</u>	EB	<u>10^{18}</u>
<u>zettabyte</u>	ZB	<u>10^{21}</u>
<u>yottabyte</u>	YB	<u>10^{24}</u>

INTRODUCTION

Amount of data
created, consumed,
and stored
2010-2025



INTRODUCTION

Where is all this data coming from?

As an example, let's look at social media usage in 2018. **In just one minute:**

Twitter users sent 473,400 tweets

Snapchat users shared 2 million photos

Instagram users posted 49,380 pictures

LinkedIn gained 120 new users



forecast for 2022 : 97 zettabytes

INTRODUCTION

So, 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

Furthermore, we must also bear in mind that:

MEASUREMENT IS AT THE HEART OF MODERN SCIENCE

and that the ever-increasing importance of measurements of the utmost precision has created, or rather reaffirmed, the need for a systematic science of data analysis

INTRODUCTION

ATTENTION !!!

- **SYNTHESIS ONLY MEANS SYNTHESIS !**
 - STATISTICAL ANALYSIS OF THE DATA NEITHER IMPROVES NOR WORSENS THEM !
- **ANY SYNTHESIS INVOLVES LOSS OF INFORMATION !**
 - IT IS THEREFORE NECESSARY TO HAVE MORE INDICES TO RECONSTRUCT THE INITIAL INFORMATION !

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

« The measure of quality, no matter what the definition of quality may be, is a variable. We shall usually represent this variable by the symbol X »

W. A. Shewhart, Economic Control of Quality of Manufactured Product, Van Nostrand, New York, 1931

INTRODUCTION

« In every manufacturing process there is variability. The variability becomes evident whenever a quality characteristic of the product is measured »

Ellis R. Ott, Process Quality Control, McGraw-Hill, New York, 1975

THIS VALID NO MATTER WHICH TYPE OF PROCESS IS UNDER CONSIDERATION

tomato cans, pencils, soap bars, automobiles, drugs manufacturing, analytical controls or else.

INTRODUCTION

Furthermore:

ALL PRODUCTION PROCESSES TEND TO DEVIATE FROM THEIR INITIAL CONDITIONS !

This happens for the most diverse reasons:

- *changes in materials, personnel, environment,*
- *technological improvements,*
- *acquisition of production experience, etc.*

INTRODUCTION



Do not trust data which look too constant or too “perfect” !

Keep in mind that:

Round numbers are always false!

Samuel Johnson (1709-1784)

INTRODUCTION

FDA is so aware of this that in its Guidance on Process Validation encourages manufacturers to:

- **Understand the source of variation**
- **Detect the presence and degree of variation**
- **Understand the impact of variation on the process** and ultimately on product attributes
- **Control the variation** in a manner commensurate with the risk it represents to the process and the product.

FDA Guidance for Industry – Process Validation: General Principles and Practices (January 2011)

INTRODUCTION

« Manufacturers should use ongoing quality programs to collect and analyze product and process information to evaluate the state of control of the process. These programs must be capable of identifying process or product problems and opportunities for manufacturing improvements that can be evaluated and implemented throughout the lifecycle. »

This is the essence of:

- **Continued Process Verification** - FDA Guidance on Process Validation (2011) or
- **Ongoing Process Verification during Lifecycle** - Annex 15 (EudraLex - Volume 4), ICH Q10, and ICH Q12.

FDA Guidance for Industry (Draft) – Request for Quality Metrics (2015)

INTRODUCTION

« *Quality is inversely proportional to variability* »

D. C. Montgomery, Statistical Quality Control: A Modern Introduction, 7th Edition, Wiley (2013)

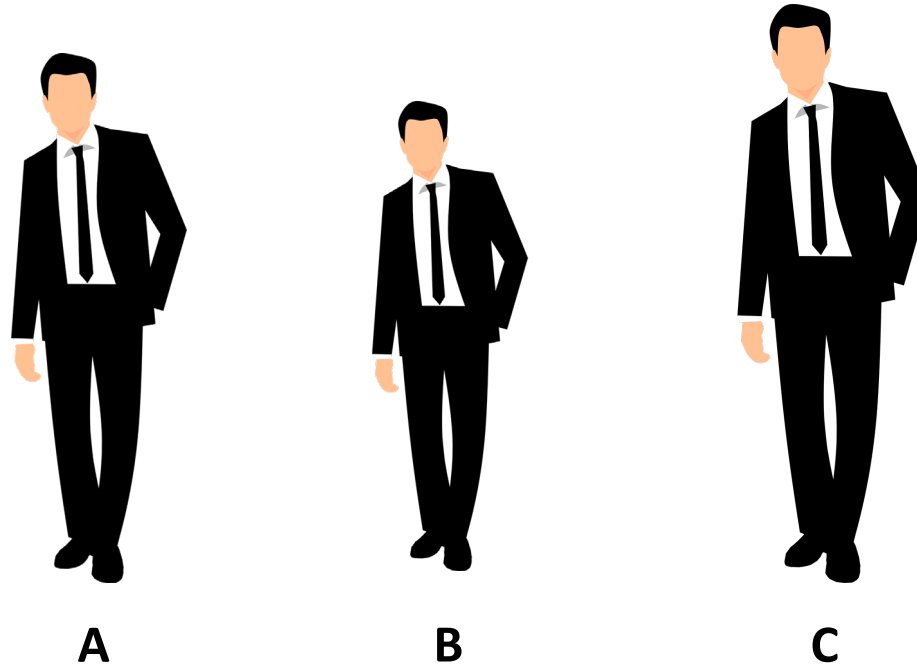
VARIABILITY IS THE "ENEMY OF QUALITY"

BUT IT IS ALSO ITS "ALLY" BECAUSE IT SENDS SIGNALS !!!

INTRODUCTION

NEVER FORGET THAT VARIATION IS SYNONYM OF INFORMATION !!!

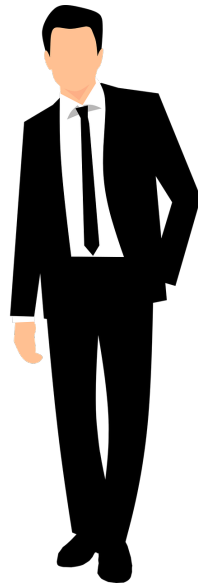
CAN YOU GUESS WHO'S WHO? NOW IT IS EASY !



Person	Height (cm)
Jonathan	145
Robert	160
Alex	185

INTRODUCTION

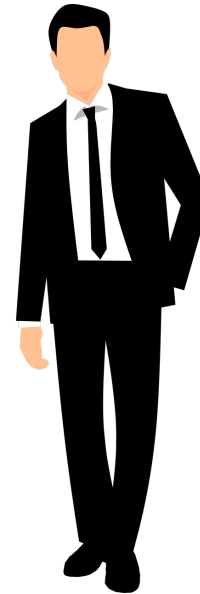
CAN YOU GUESS WHO'S WHO? NOW IT IS TOUGH!



A



B

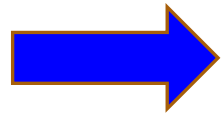


C

Person	Height (cm)
Marc	173
John	172
Frank	174

INTRODUCTION

TO CAPTURE THESE SIGNALS, TOOLS ARE NEEDED, NAMELY THE
"QUALITY METRICS"



Quality Metrics = Quantitative Indicators of Quality
= Key Process Indicators

INTRODUCTION

« Quality metrics are used throughout the drugs and biologics industry to monitor quality control systems and processes. Modern manufacturing includes robust quality metrics programs as a foundation for continual improvement of product and process quality.

Quality metrics are one element of companies' commitment to quality culture. »

Quality Metrics are tools provided by STATISTICS !

<https://www.fda.gov/drugs/pharmaceutical-quality-resources/quality-metrics-drug-manufacturing>

ELEMENTS OF STATISTICS

ELEMENTS OF STATISTICS

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* !

ELEMENTS OF DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

- Is the part of Statistics that quantitatively describes or summarizes features from a collection of information or data
- It *examines the results of real experiments, already occurred and definitive*, of which retrospectively studies the distribution of the character (or variable) X
- It represents the « *exploratory moment of reality* »
- To achieve this target, Descriptive Statistics essentially makes use of two basic tools: **PLOTS** and **SUMMARY INDICES**
- **DATA** are the values (or modalities) assumed by **VARIABLES** (or **CHARACTERS**)

DESCRIPTIVE STATISTICS

Data is the result of an experiment, measurement, observation and investigation, *etc.*

Data is the basis for
any scientific decision !

Most of world's data are obtained
as byproducts of operations !



DESCRIPTIVE STATISTICS

Qualitative
Data

Attributes
Data

There can be an **order relationship** (e.g., educational qualification: elementary, middle school, university) **or not** (e.g., city of residence: Milan, Rome, etc.)

Quantitative
Data

Variables
Data

Occur by *measures* and can be **discrete** (e.g., football championship scores: Inter 46, Juventus 42, etc.) or **continuous** (e.g., length, weight, purity, etc.)

DESCRIPTIVE STATISTICS

Typical examples of *ordered qualitative characters* (ordinal scale) are for example those of:

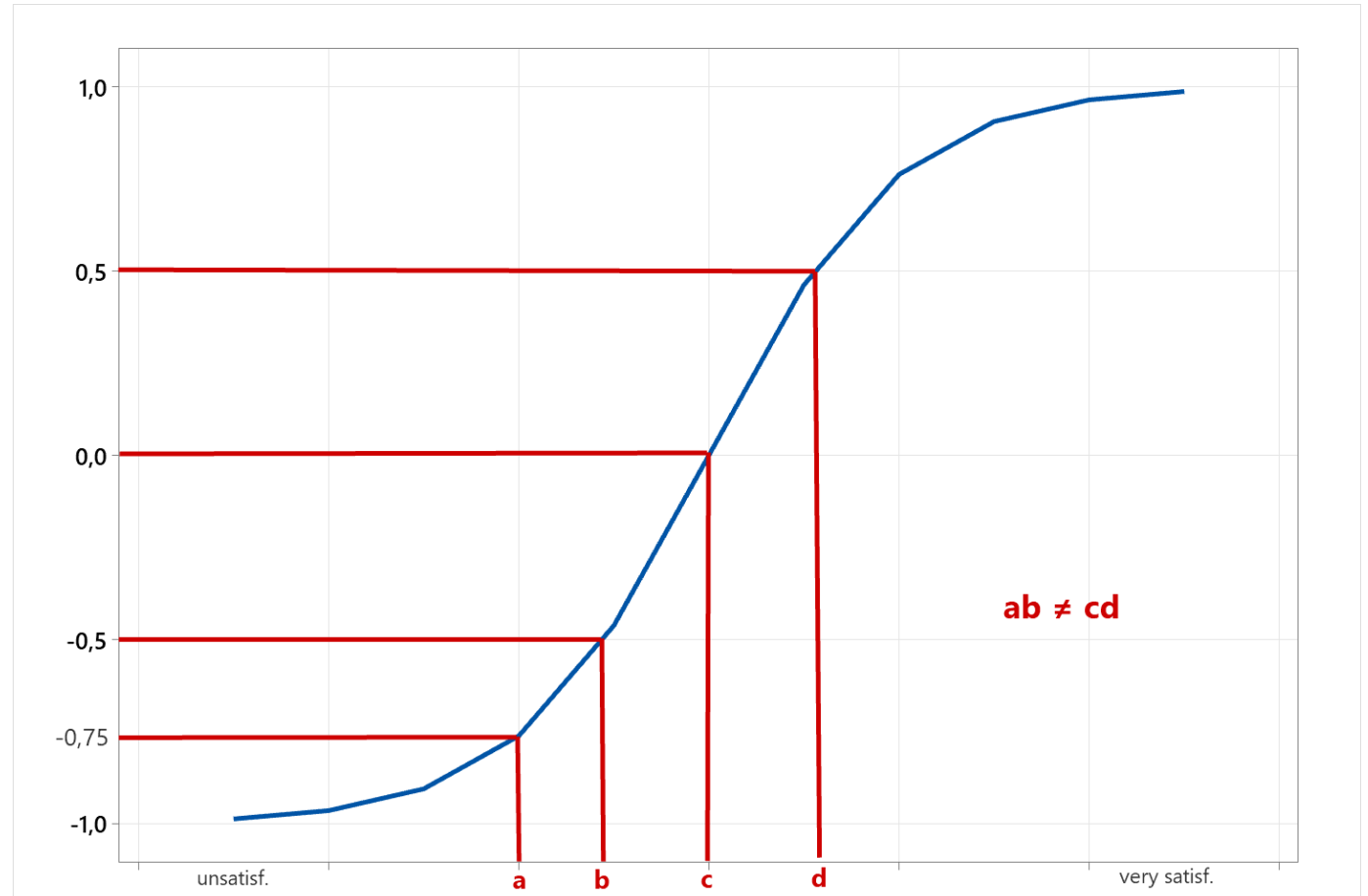
- customer satisfaction surveys: *dissatisfied, indifferent, satisfied, very satisfied*
- risk rating: *minor, moderate, substantial, severe*

ATTENTION !

the recoding of qualitative characters ordered on an ordinal scale is not necessarily linear!

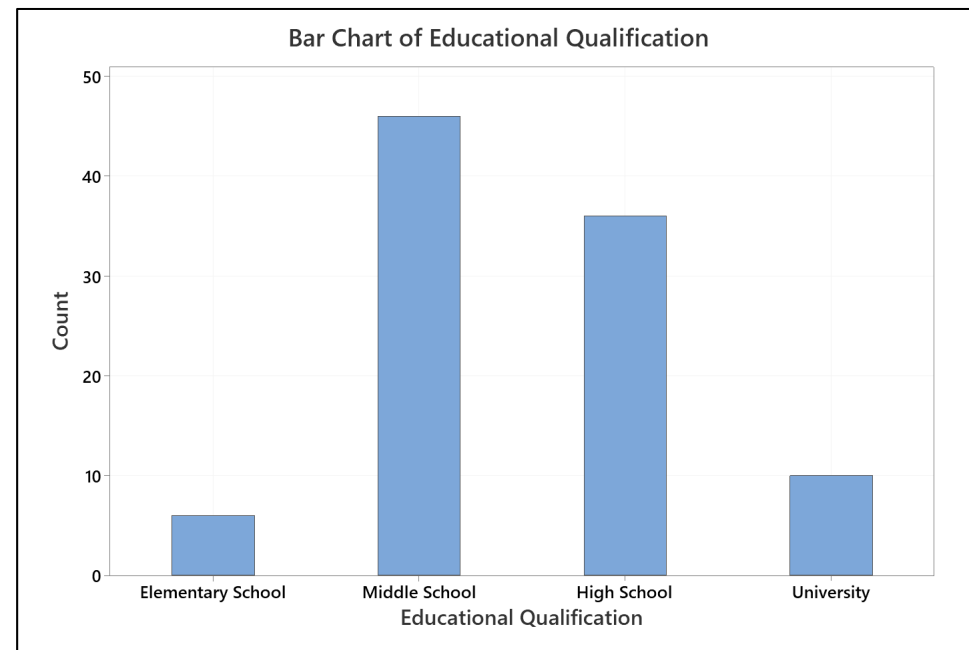
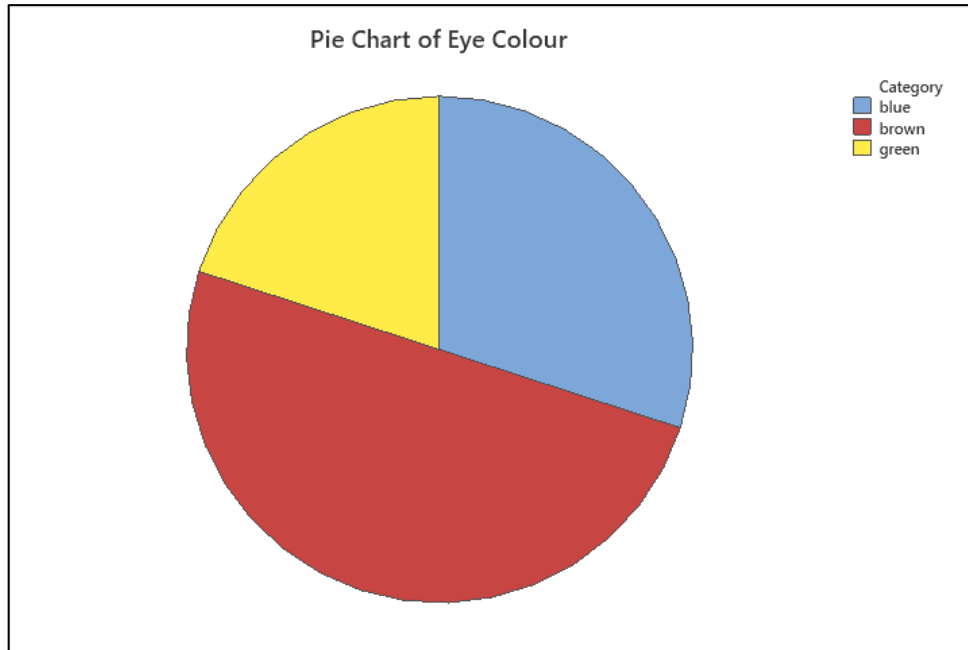
DESCRIPTIVE STATISTICS

- In general, it is of the *sigmoid type* as shown on the side and therefore it makes no sense to compare the distances between categories even if coded by numerical values !
- The categories of the qualitative character, even if expressed through numerical values, always remain ordered codes!



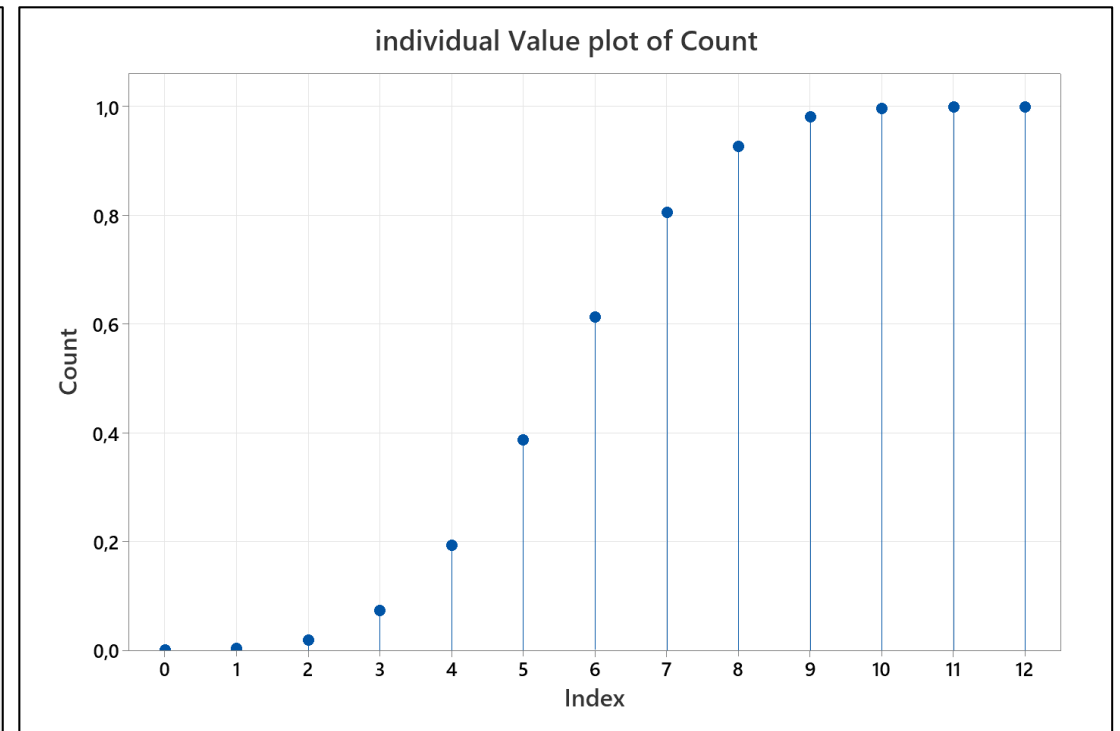
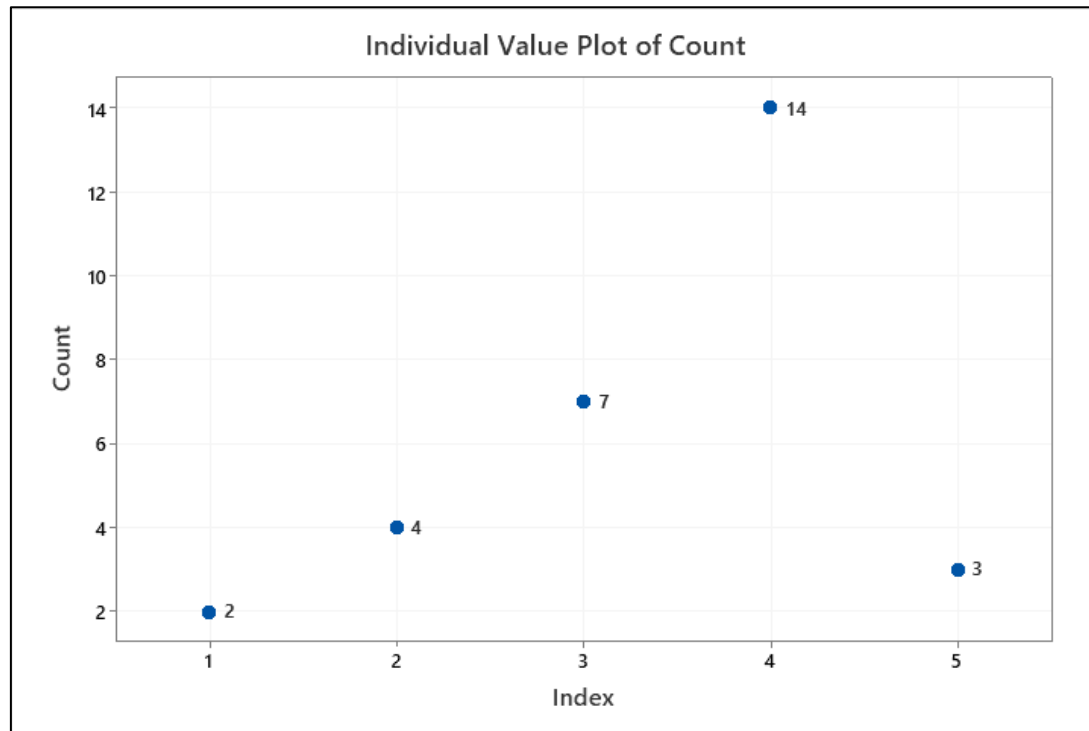
DESCRIPTIVE STATISTICS

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



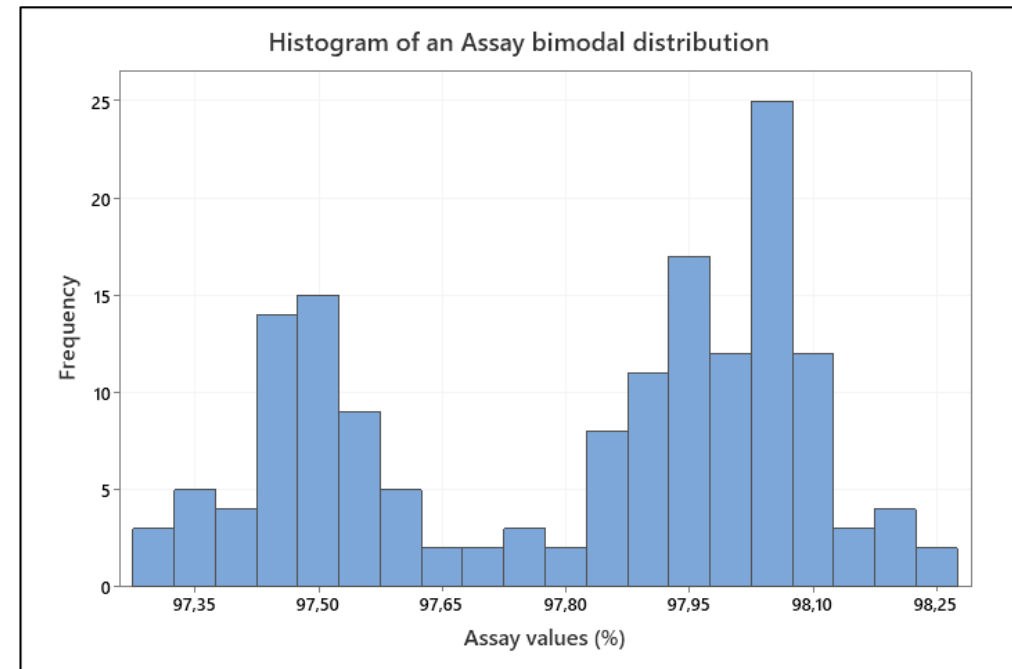
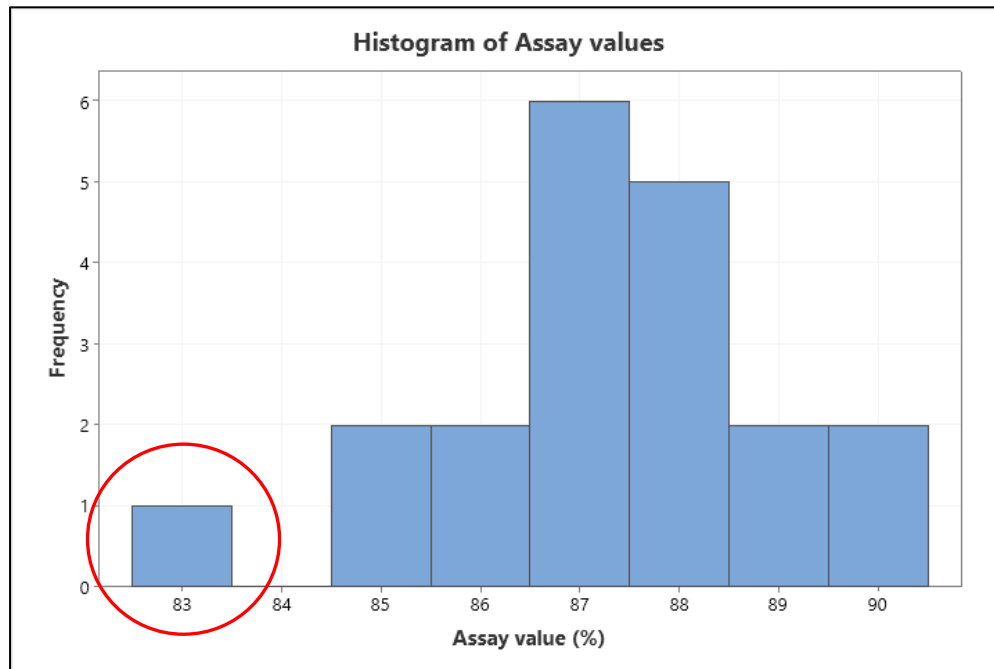
DESCRIPTIVE STATISTICS

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



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, but also to reveal *multimodal distributions*.



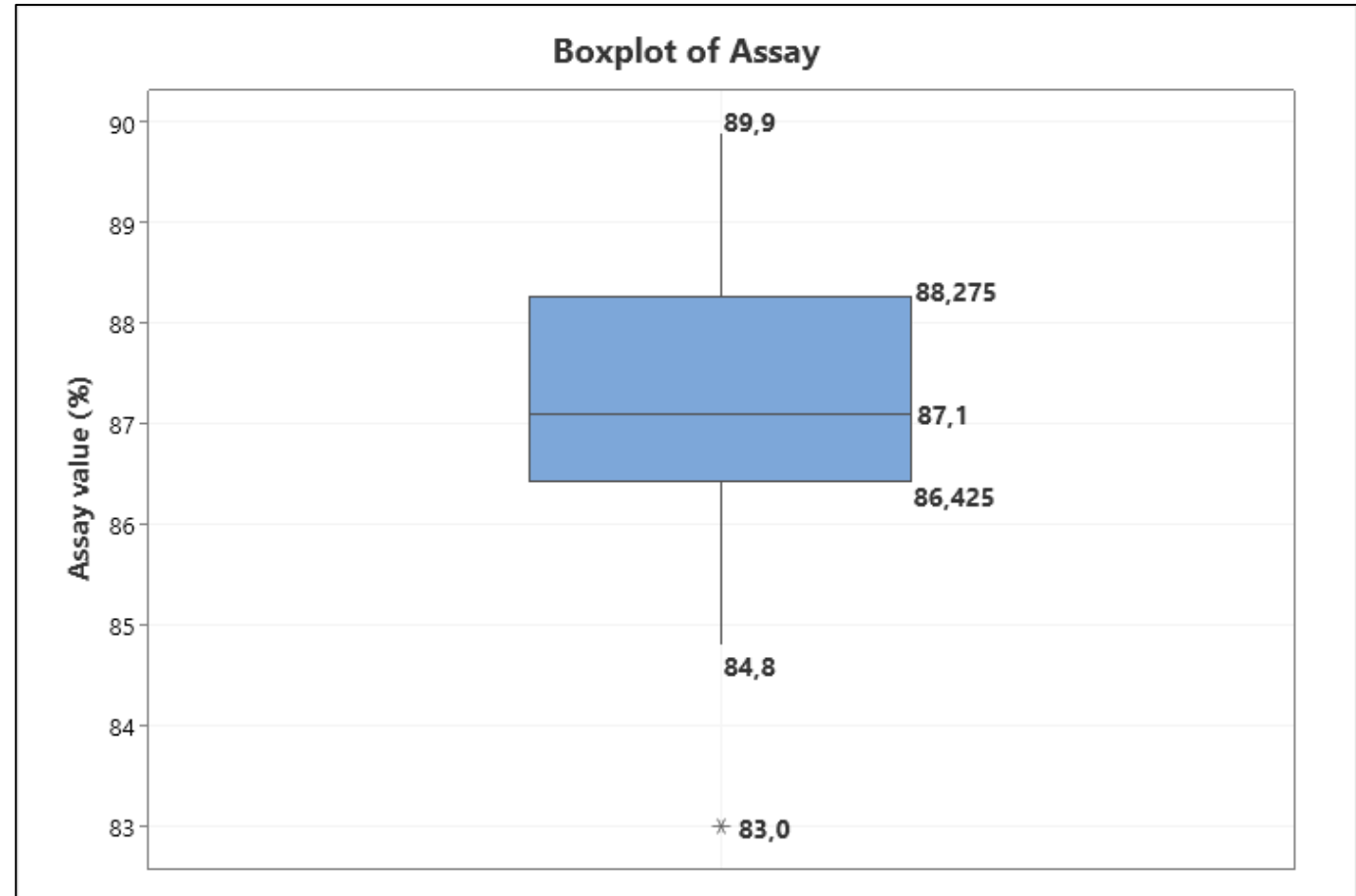
DESCRIPTIVE STATISTICS

CONTINUOUS

QUANTITATIVE DATA

can also be
effectively
represented even
using **BOX PLOTS**

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



DESCRIPTIVE STATISTICS

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

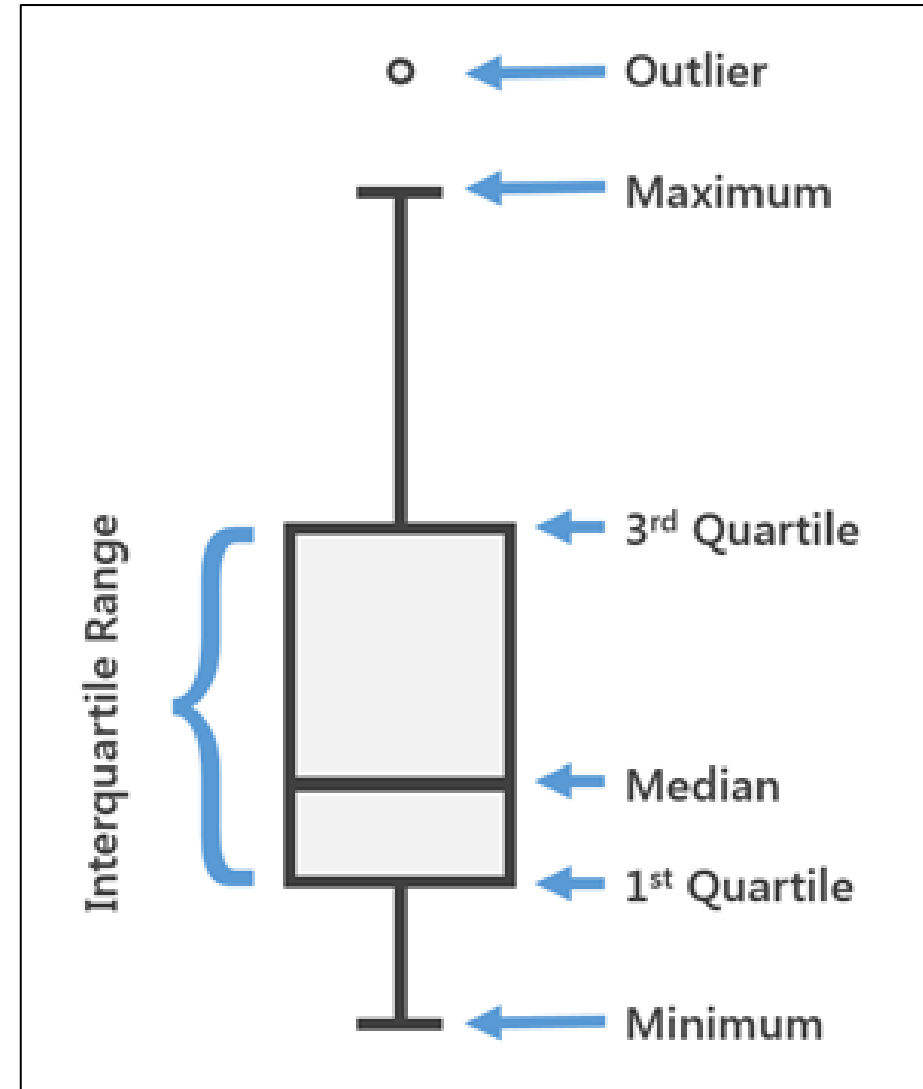
Median, Q2: 50% of the data \leq this value

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

Interquartile range: 50% of the data

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

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



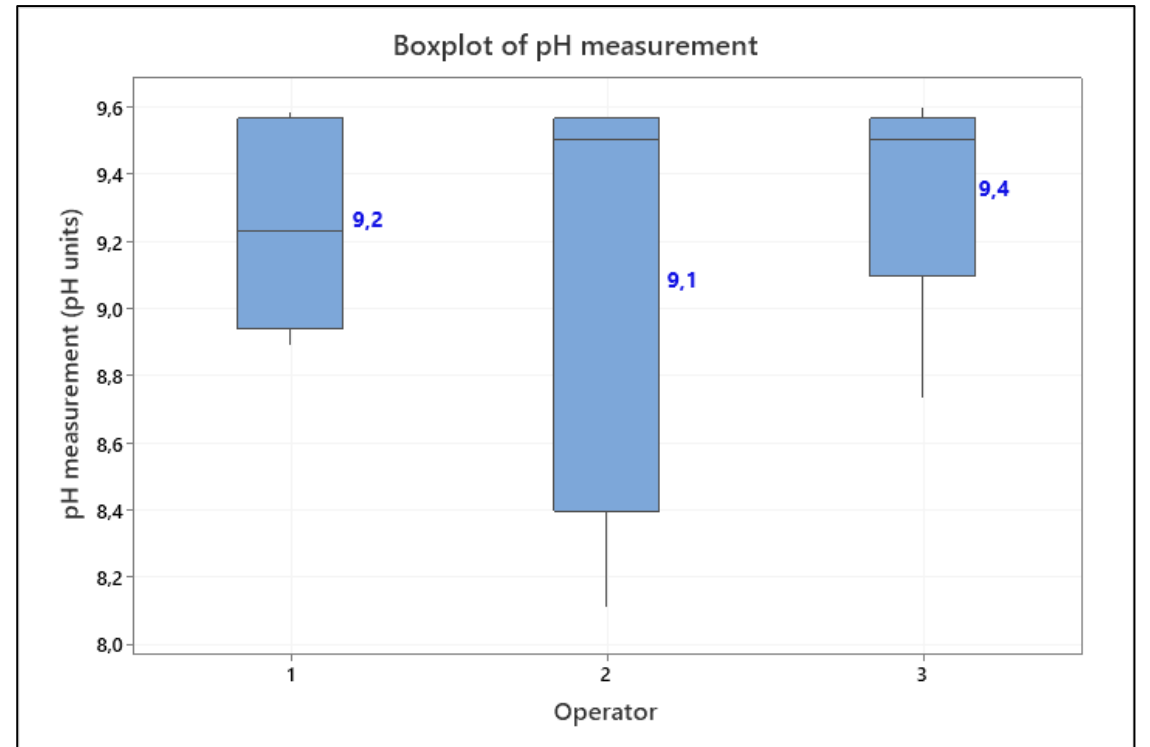
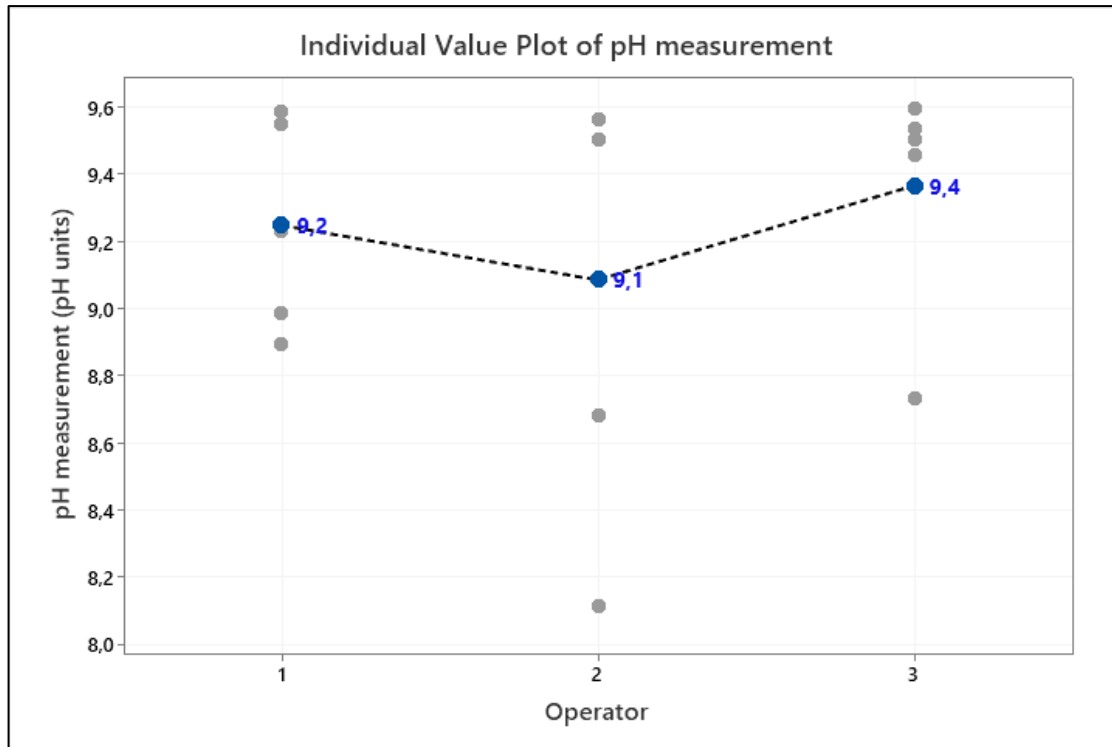
DESCRIPTIVE STATISTICS

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:

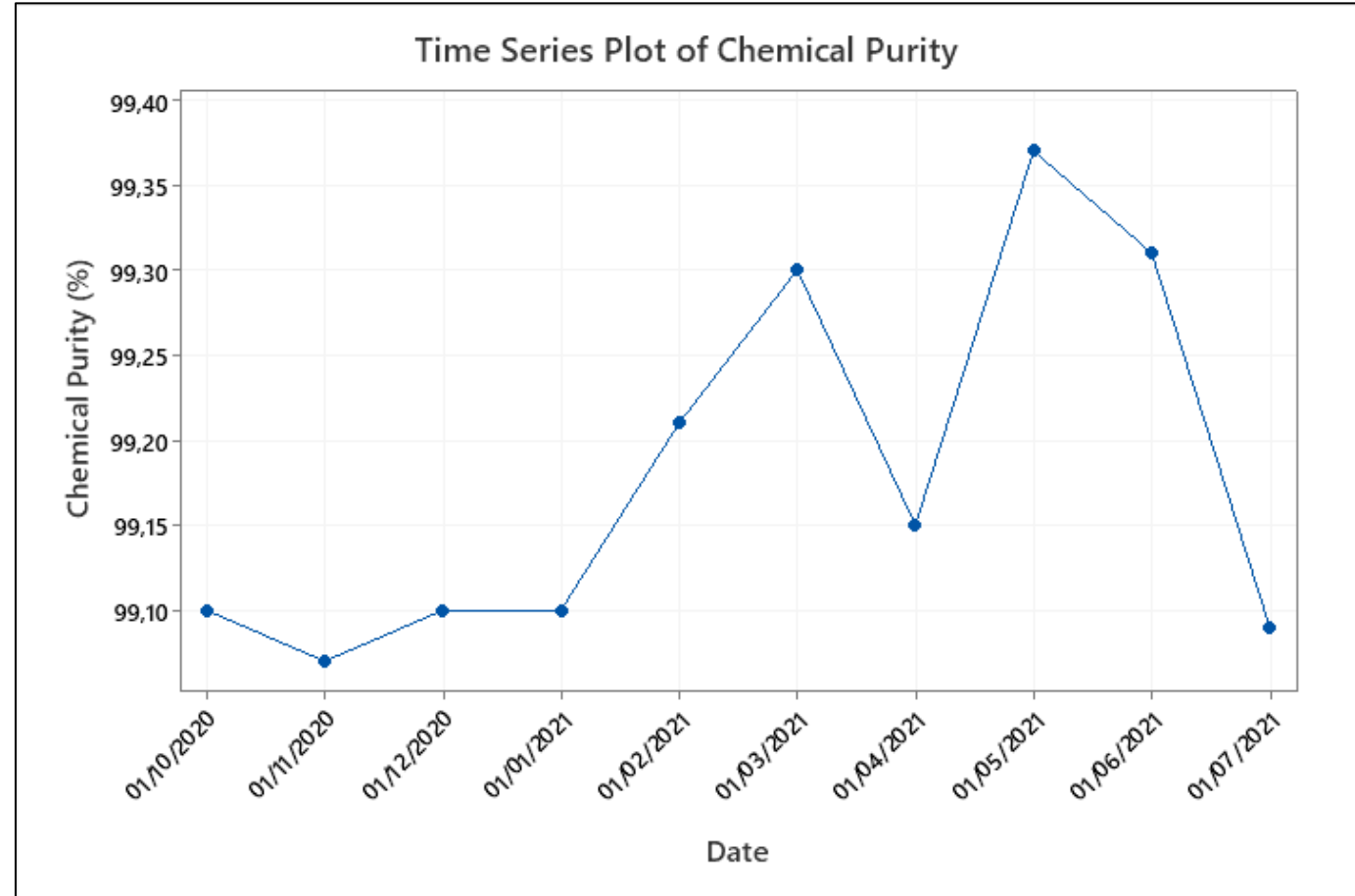


DESCRIPTIVE STATISTICS

TIME SERIES is a sequence of data points listed (or graphed) in time order.

In general, data is taken at successive equally spaced points in time (*e.g.*, process controls, APQR, stability studies, *etc.*)

This type of graphs are also known as **Line Graphs**.



DESCRIPTIVE STATISTICS

Please, duly consider the following quote:

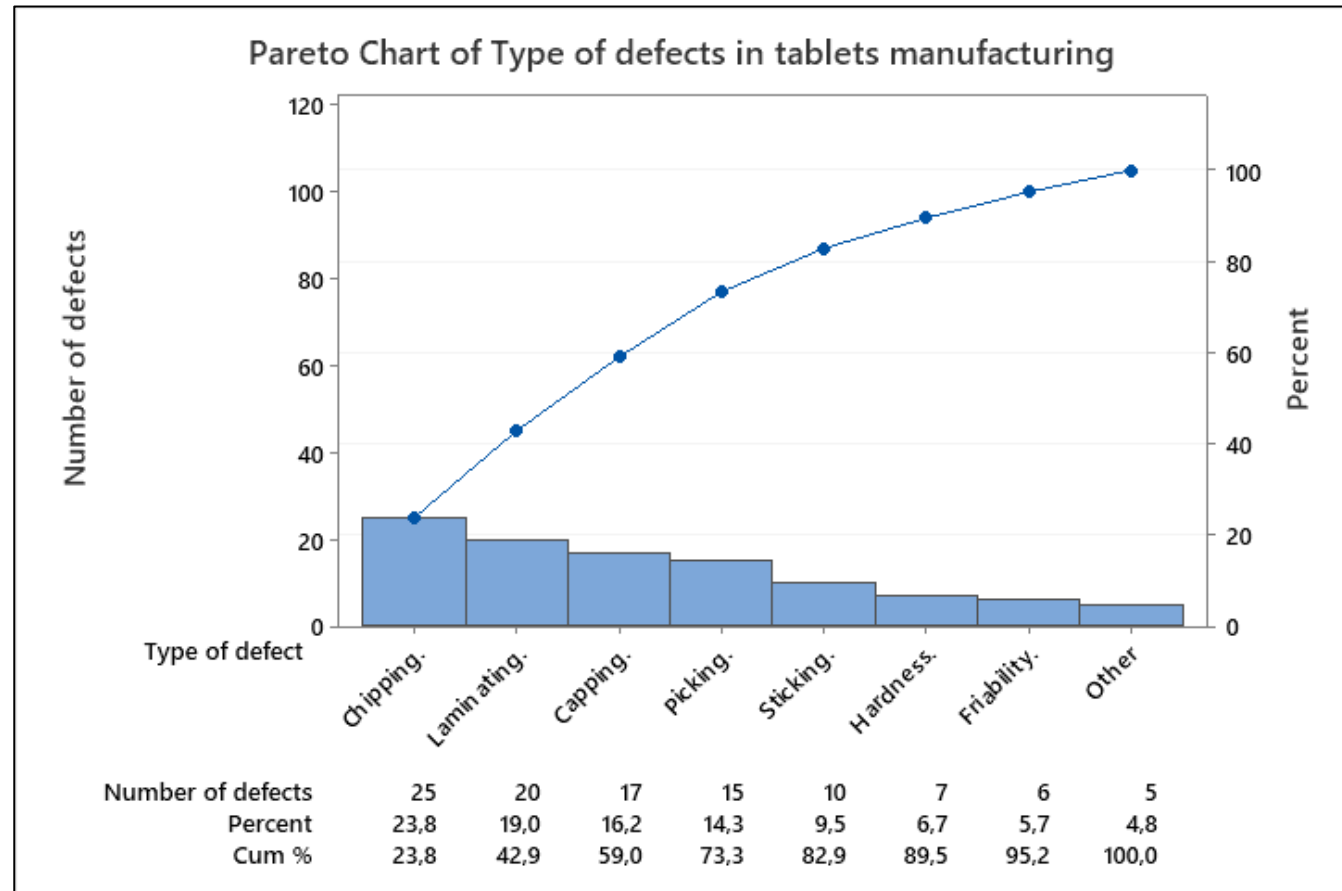
« ...**TIME SERIES PLOTS and HISTOGRAMS can be thought as COMPLEMENTARY TO EACH OTHER.**

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.

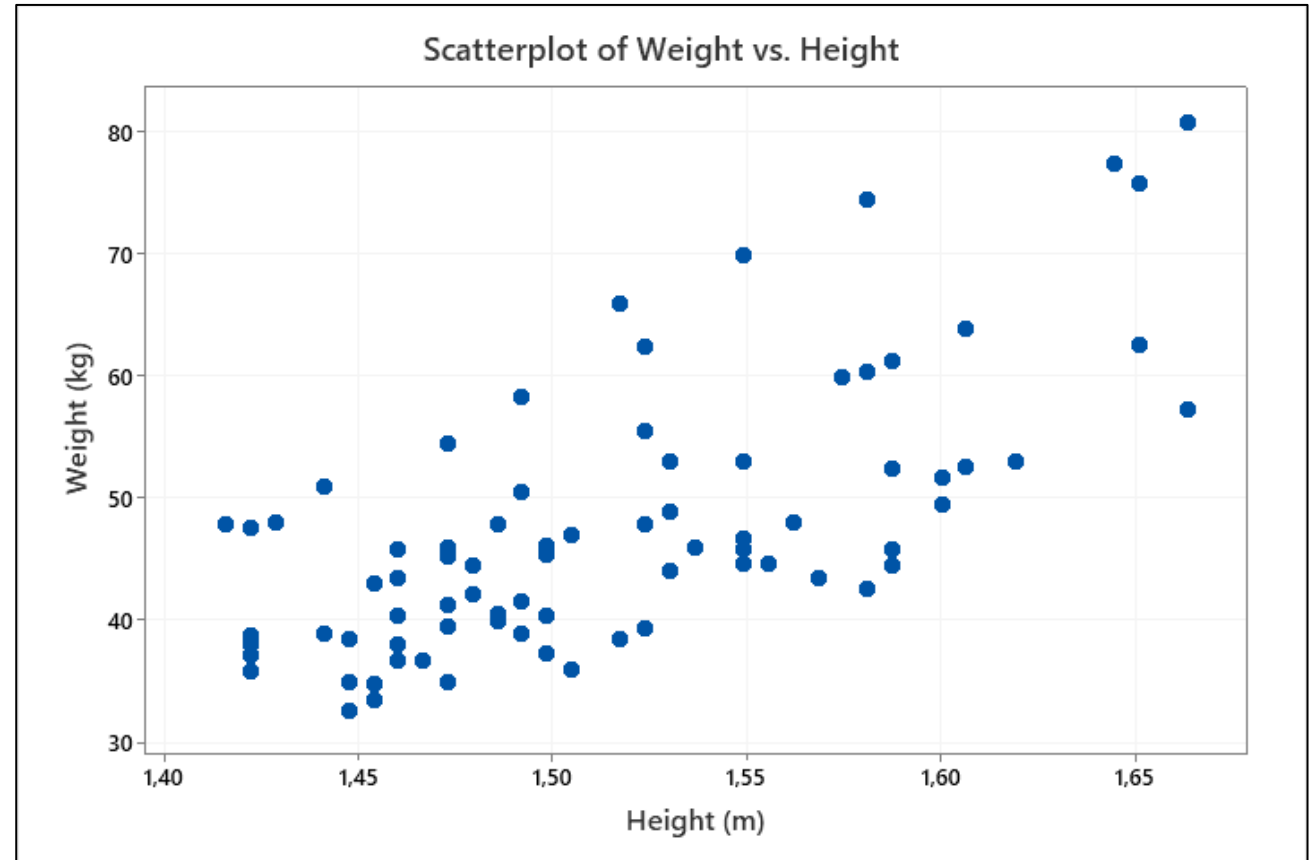


DESCRIPTIVE STATISTICS

Until now we have always considered only *one variable* (**UNIVARIATE**). Let now assume we have two continuous variables such as those shown here on side.

What does this graph tell us?

As height increases, also weight tends to increase !



DESCRIPTIVE STATISTICS

- *the scatterplot shows the existence of a direct and approximately linear relationship between height and weight, but it does not give any quantitative measure of the magnitude of this relationship !*
- « A **correlation** between variables indicates that as one variable changes in value, the other tends to change in a specific direction »

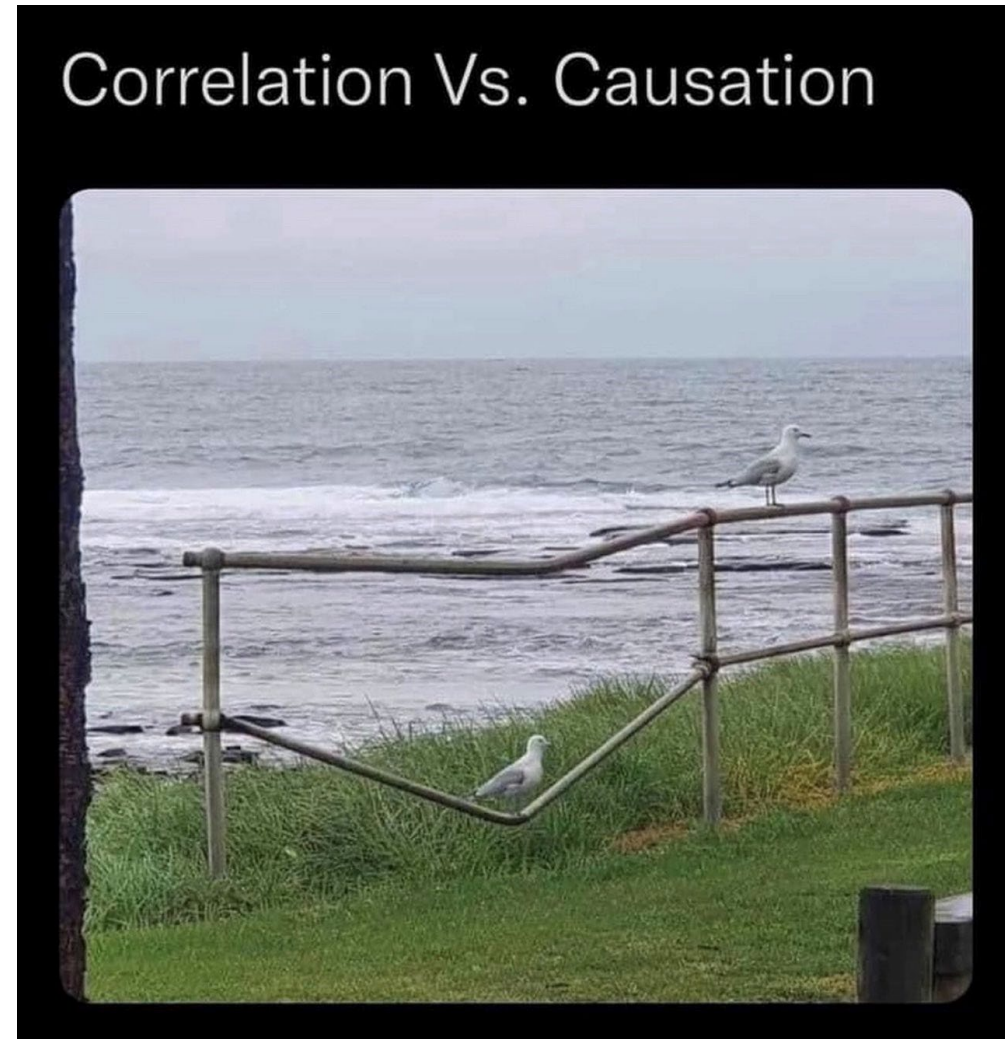
J. Frost, Introduction to Statistics, Statistics by Jim Publishing (2019)

- *Correlation methods measures the strength of association between two or more variables*
- *Correlation does not imply causation !*

DESCRIPTIVE STATISTICS

Correlation does not imply causation !

<https://www.tylervigen.com/spurious-correlations>



DESCRIPTIVE STATISTICS

Descriptive Statistics provides a powerful index that summarizes the magnitude of the *linear* link between two variables: the *Bravais-Pearson linear correlation coefficient* or, more appropriately, the *Galton-Pearson linear correlation coefficient**

$$\rho_{ij} = \rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i) \text{Var}(X_j)}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \quad \forall i \neq j$$

where:

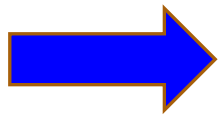
X_i = i -row of the data matrix

X_j = j -column of the data matrix

*J. L. Rodgers, A. Nicewander, *Thirteen Ways to look at the Correlation Coefficient*, *The American Statistician*, 42(1), 1988, 59-66

DESCRIPTIVE STATISTICS

- $Cov(X_i, X_j)$ is the **COVARIANCE** between variable X_i and variable X_j . Is a symmetrical index that measures the tendency of two variables to vary together.
- the **LINEAR CORRELATION COEFFICIENT** takes on values in the range $\rho_{ij} \in [-1, +1] \quad \forall i \neq j$
 - $\rho_{xy} = -1$ in the case of *perfect inverse linear relationship* between X and Y
 - $\rho_{xy} = +1$ in the case of *perfect direct linear relationship* between X and Y
 - $\rho_{xy} = 0$ in the case of *no correlation* between X and Y: no linear relationship



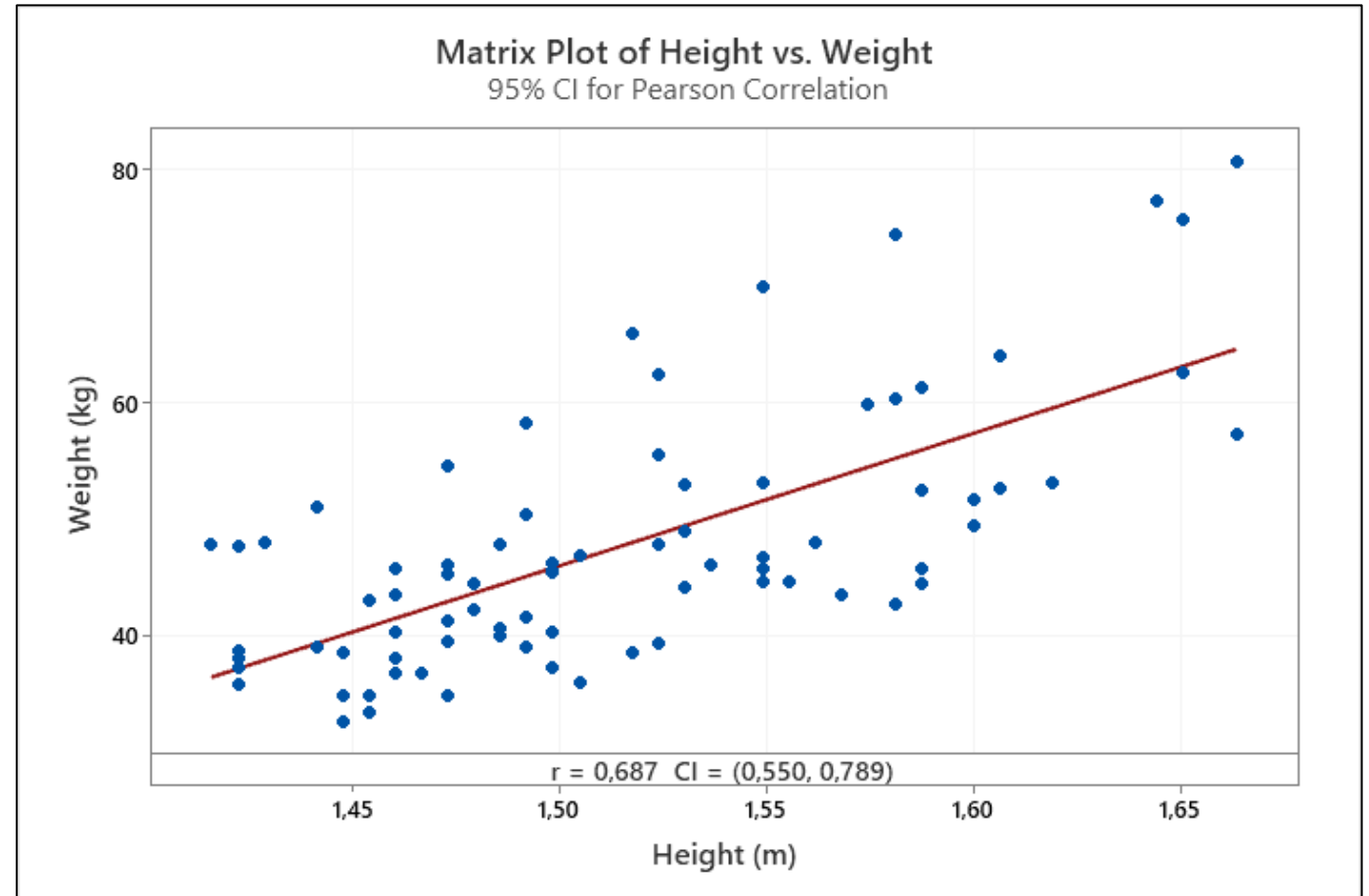
the linear correlation coefficient is the fundamental measure for studying the relationships between two quantitative variables

DESCRIPTIVE STATISTICS

In the case of the height-weight data shown above, the linear correlation coefficient holds:

Correlations

	Height (m)
Weight (kg)	0,687



DESCRIPTIVE STATISTICS

ATTENTION!

The *linear correlation coefficient* **only** measures linear relationships!

Other types of data relationships (e.g., curvilinear) would therefore not be detected. However, in that case, the *scatterplot would help us !*

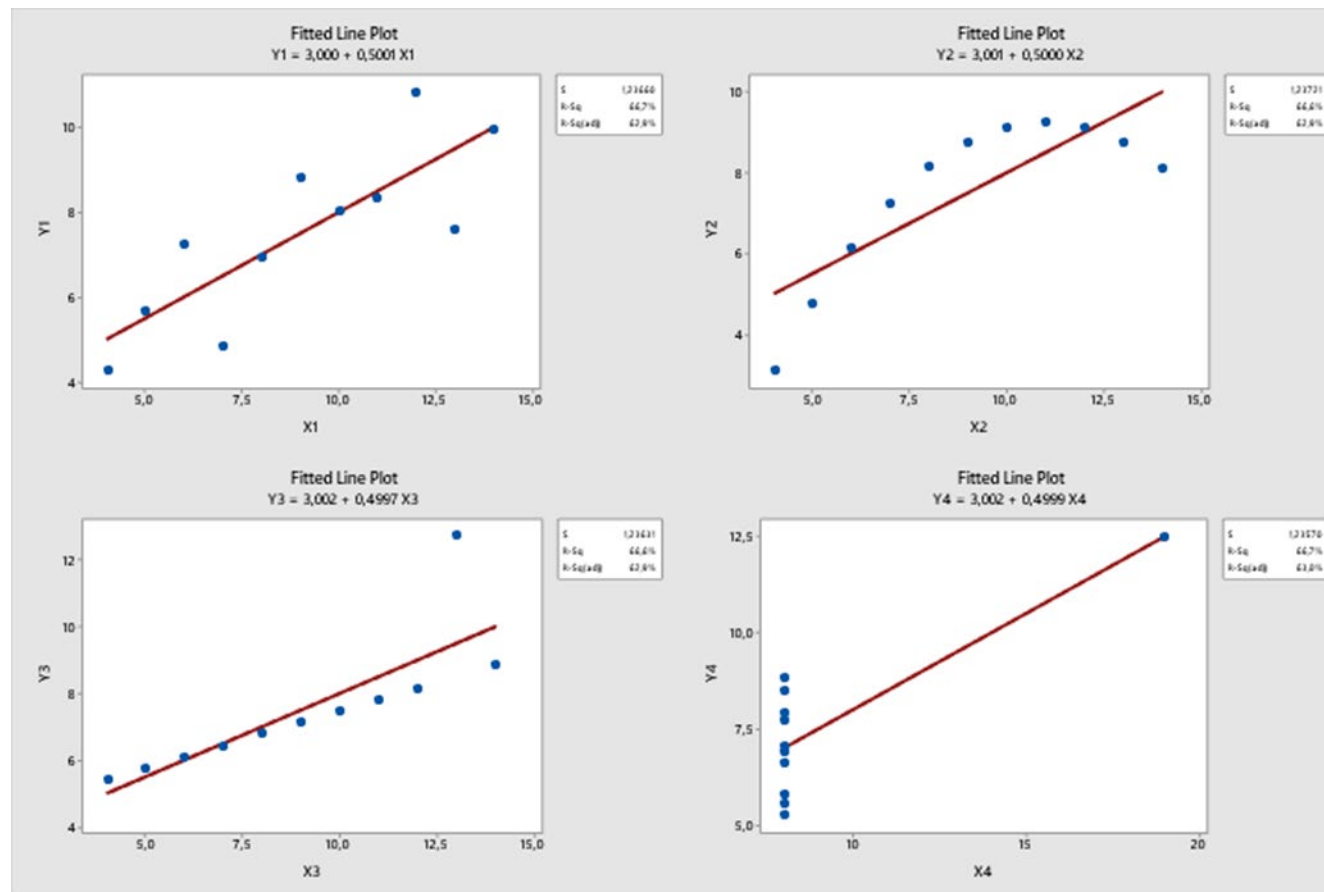
The *linear correlation coefficient* represents an example of a ***synthetic indicator***, or ***index***, which summarizes the important aspects of the variables under study.

DESCRIPTIVE STATISTICS

IN CONCLUSION:

**Why is it so important
to always graph data?**

**Because of
Anscombe's Quartet !**

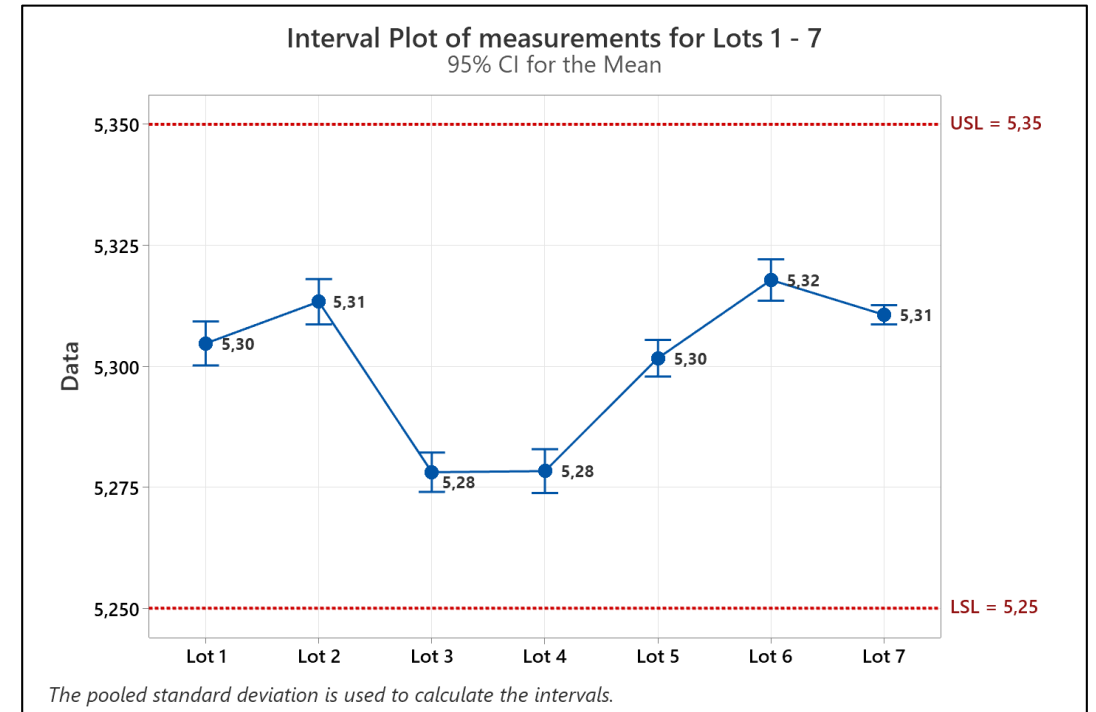
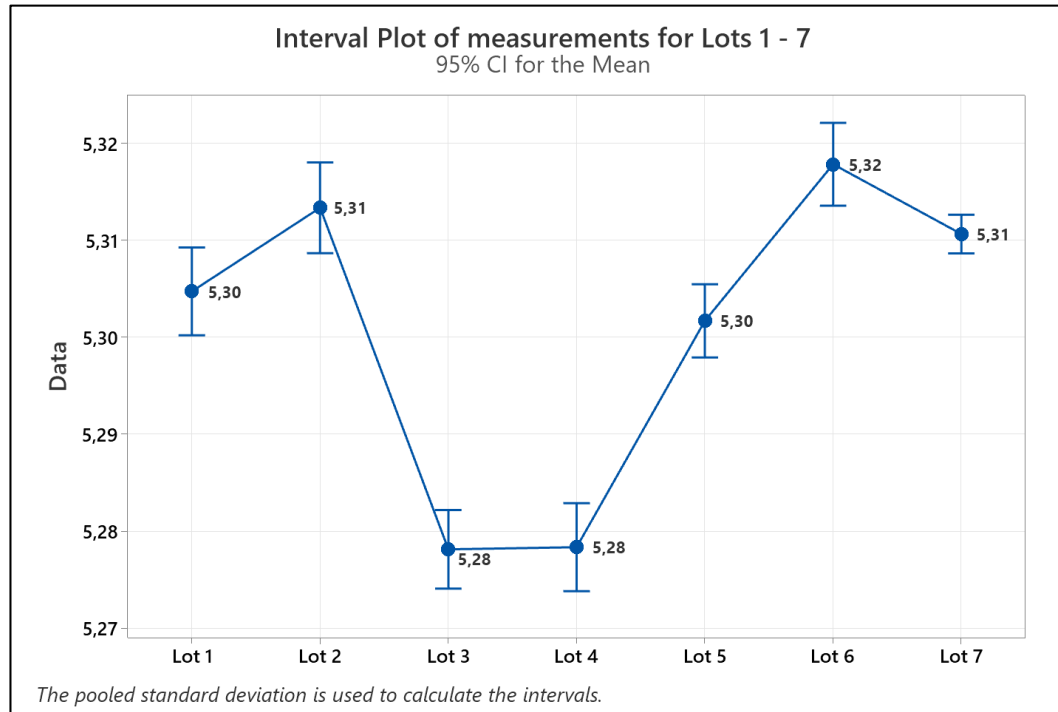


F.J. Anscombe, Graphs in Statistical Analysis, American Statistician, Vol. 27, No. 1 (1973)

DESCRIPTIVE STATISTICS

ONE LAST RECOMMENDATION

When conducting data studies, never forget to contextualize them (e.g., by reporting specification limits or else)



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

Let start with the **POSITION INDICES** :

- **MODE**: is the value that appears most often in a data set.

e.g.: 3, 3, 5, 6, **7, 7, 7**, 8, 8, 10 \rightarrow Mode = 7

Data distributions can have only one mode, more than one mode, or even no mode. In fact, if the values constituting the data set are all different from each other, that distribution will have no mode.

N.B.: Mode is a broad-sense average since the property of monotonicity is not valid.

DESCRIPTIVE STATISTICS

- **MEDIAN**: is the middle point in a dataset, half of the data points are smaller than the median and half of the data points are larger.

e.g.: 0, 0, 1, 1, 2, 3, 3, 3, 4 → Median = 2 (Mean = 1.89)

The median is not affected by the extreme values of the data distribution !

It is precisely for this reason that the median is said to be a "robust" central trend indicator!

e.g.: 0, 0, 1, 1, 2, 3, 3, 3, 4, 45, 50 → Median = 3 (Mean = 10.18)

In general, a summary indicator of a data distribution is said to be "robust" if it is not particularly influenced by extreme values, *i.e.*, by very large or very small ones.

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: $\mu = \frac{1}{3} (3 \times 1 + 5 \times 1 + 10 \times 1) = \frac{1}{3} (18) = 6$

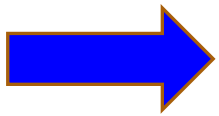
DESCRIPTIVE STATISTICS

Why this formula?

First of all, to say that *there is no single algebraic average* (i.e., arithmetic) as we are often led to believe.

In fact, there are also: *harmonic mean* ($r = -1$), *geometric mean* ($r = 0$), *quadratic mean* ($r = 2$), etc. and

$$\mu^{(-1)} \leq \mu^{(0)} \leq \mu^{(1)} \leq \mu^{(2)}$$



harmonic mean \leq geometric mean \leq arithmetic mean \leq quadratic mean

The Arithmetic Mean is the most used position index !

DESCRIPTIVE STATISTICS

The arithmetic mean has some properties, two of which are extremely important:

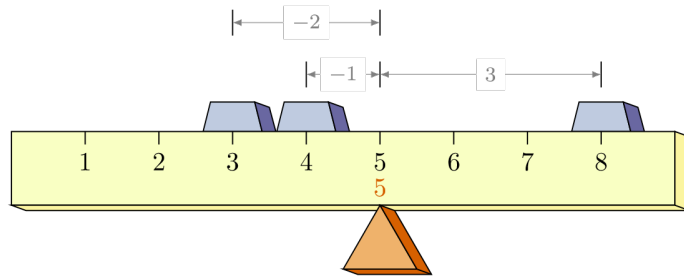
1. *The arithmetic mean zeroes the sum (or the average) of the differences between each value assumed with its sign.*

This property is also known as: **BARYCENTRIC PROPERTY OF THE MEAN** as the arithmetic mean can be considered the **center of gravity** of the dataset where the differences are balanced.

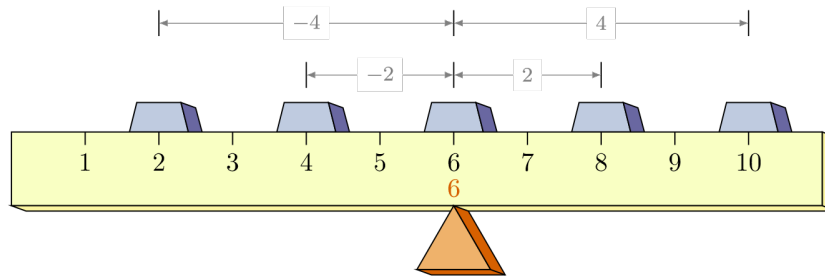
The arithmetic mean, in fact, is the only measure in which all values have the same weight !

$$\text{e.g.: } 3, 4, 8 \rightarrow \text{mean} = 5 \rightarrow \Sigma (3-5) + (4-5) + (8-5) = -2 - 1 + 3 = 0$$

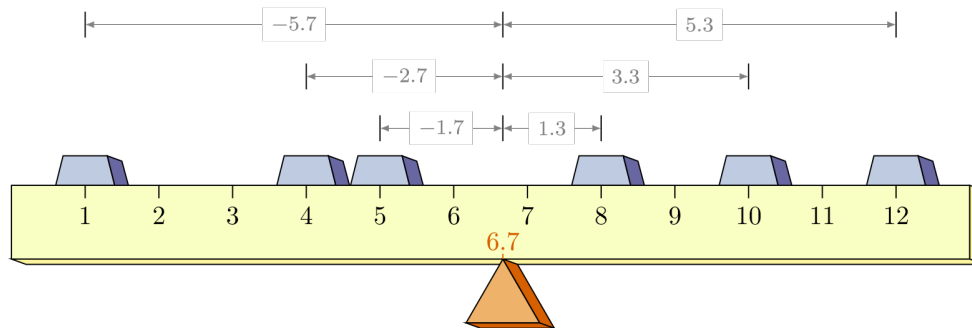
DESCRIPTIVE STATISTICS



$3, 4, 8 \rightarrow \text{mean} = 5$



$2, 4, 6, 8, 10 \rightarrow \text{mean} = 6$



$1, 4, 5, 8, 10, 12 \rightarrow \text{mean} \cong 6,7$

DESCRIPTIVE STATISTICS

2. *The arithmetic mean minimizes the sum of the squared deviations, i.e.:*

$$\mu = \sum_{i=1}^k (x_i - \alpha)^2 n_i = \min$$

e.g.: given: 10, 8, 15, 7 the arithmetic mean is: $\mu = 1/4 [(10 \times 1) + (8 \times 1) + (15 \times 1) + (7 \times 1)] = 1/4 (40) = \mathbf{10}$

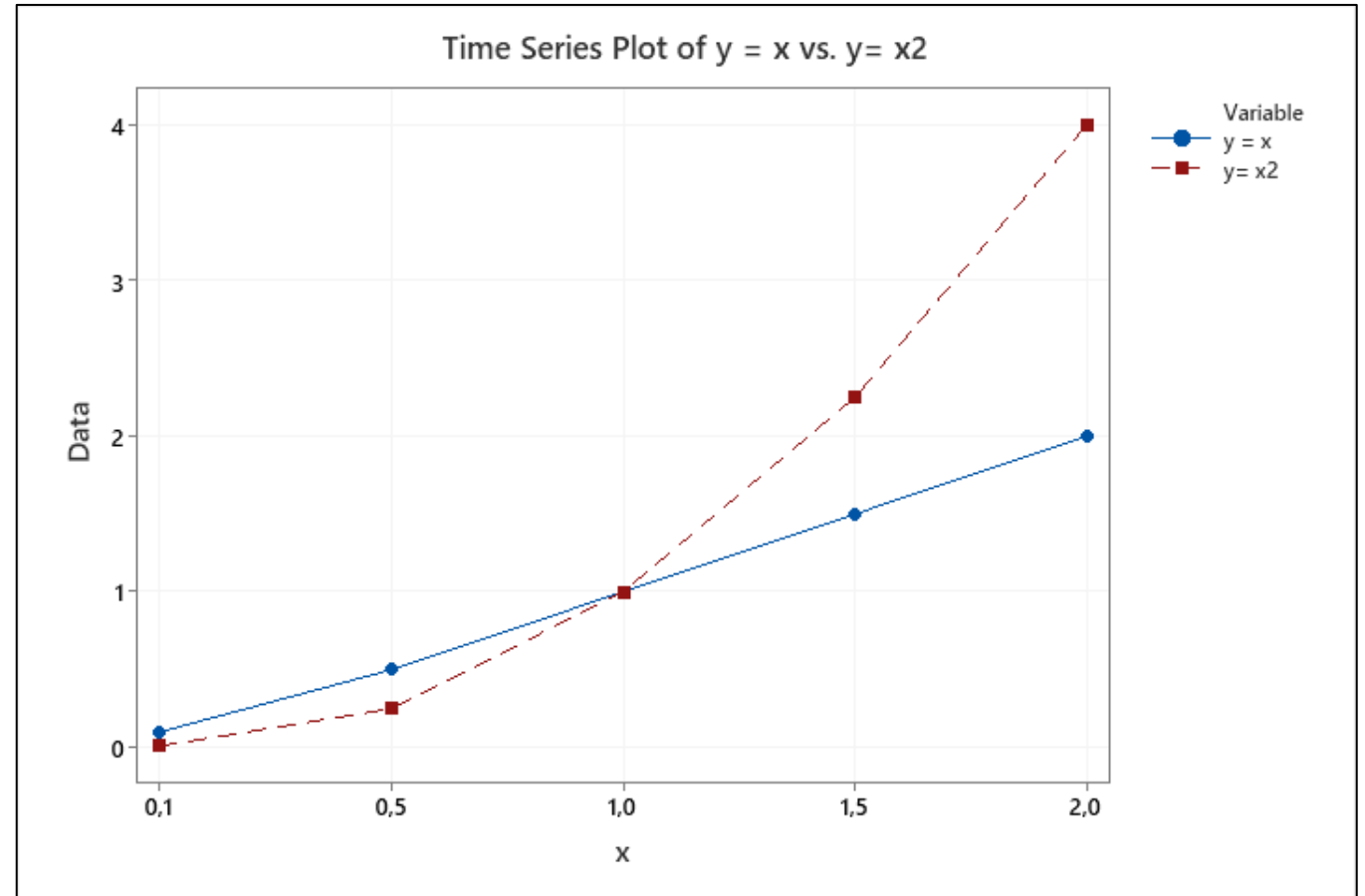
$$\rightarrow (10 - \mathbf{10})^2 + (8 - \mathbf{10})^2 + (15 - \mathbf{10})^2 + (7 - \mathbf{10})^2 = (0)^2 + (-2)^2 + (5)^2 + (-3)^2 = 0 + 4 + 25 + 9 = 38$$

using any other value, e.g., **8**

$$\rightarrow (10 - \mathbf{8})^2 + (8 - \mathbf{8})^2 + (15 - \mathbf{8})^2 + (7 - \mathbf{8})^2 = (-2)^2 + (0)^2 + (7)^2 + (-1)^2 = 4 + 0 + 49 + 1 = \mathbf{54} > \mathbf{38} \text{ c.v.d.}$$

DESCRIPTIVE STATISTICS

- The *1st and the 2nd properties of the arithmetic mean* are the reasons that explain why the differences (or deviations) always occur squared in statistical indices such as: variance, standard deviation, *etc.*
- Moreover, by squaring small differences are "rewarded" and large ones "penalized".



DESCRIPTIVE STATISTICS

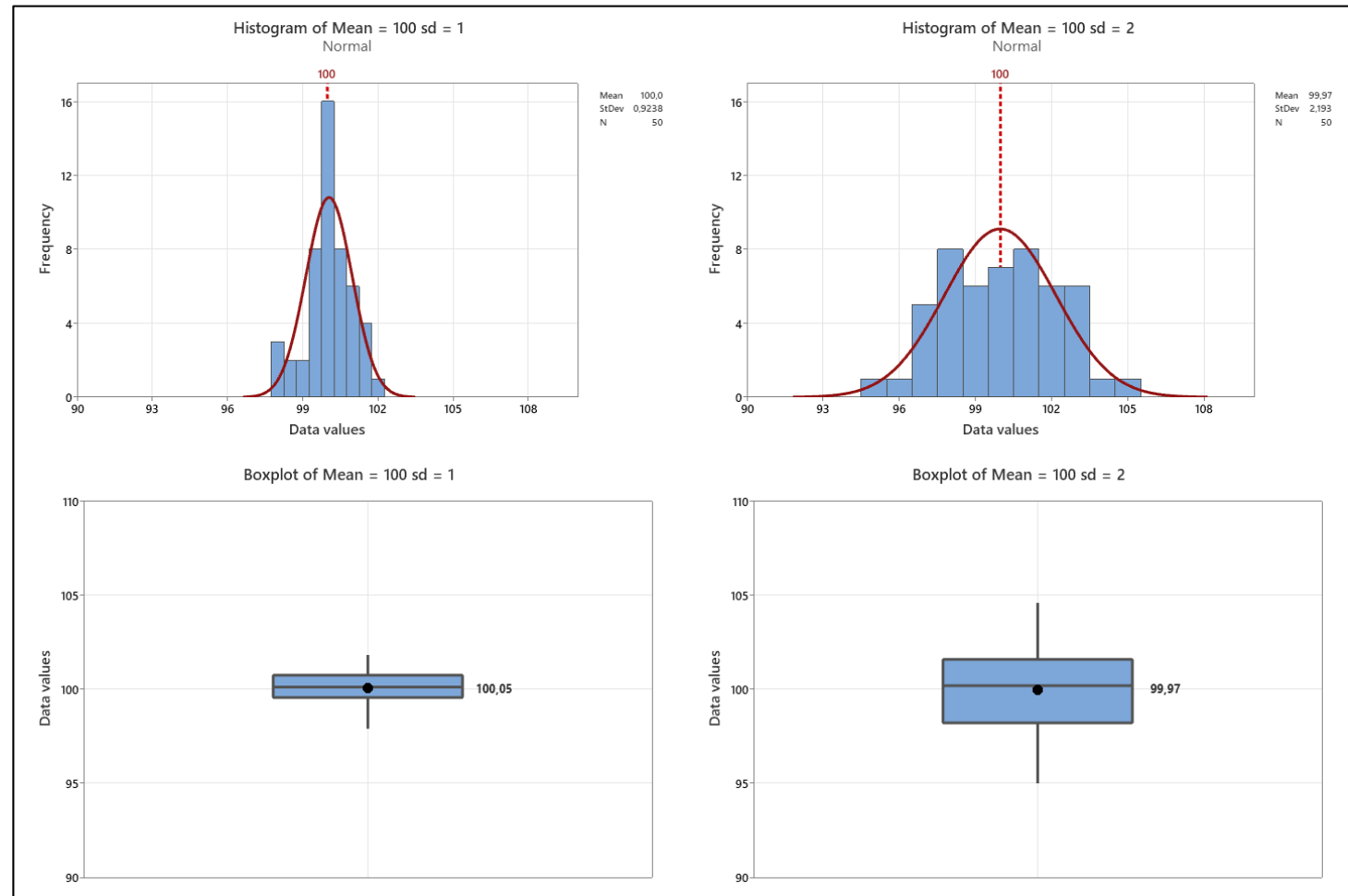
- The *position indices* are summary indices that replace the values of a variable with a single value that can be considered "representative of all the others".
- ***It is evident that, by itself, the position index is insufficient to describe a distribution of data !***
Synthesis, in fact, involves loss of information and therefore two data distributions, for example, can have the same average but be profoundly different from each other.

e.g.: 3, 4, 5, 6, 7 \rightarrow Mean = 5 vs. 0, 0, 0, 1, 24 \rightarrow Mean = 5

DESCRIPTIVE STATISTICS

As anticipated:

- *the Position Indices alone are insufficient to describe a given distribution of data!*
- *therefore, we need other summary indices to complete the information!*



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!

ALWAYS REMEMBER THAT:

VARIABILITY IS THE VERY REASON FOR THE EXISTENCE OF STATISTICS !!!

IF THERE WERE NO VARIABILITY, THERE WOULD BE NO STATISTICS!!!

DESCRIPTIVE STATISTICS

- The **Variability Indices** are essentially of two types:
 - **GLOBAL INDICES:** they are the ones who measure the distances of each modality from all the others
 - **DISPERSION INDICES:** are those who measure the distances of each modality from a particular ad hoc choice as a reference (e.g., the arithmetic mean) and
they are the ones we will consider !
- *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.



27



27



33



46



57

$$\text{Range} = \text{Maximum age} - \text{Minimum age} = 57 - 27 = 30$$

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

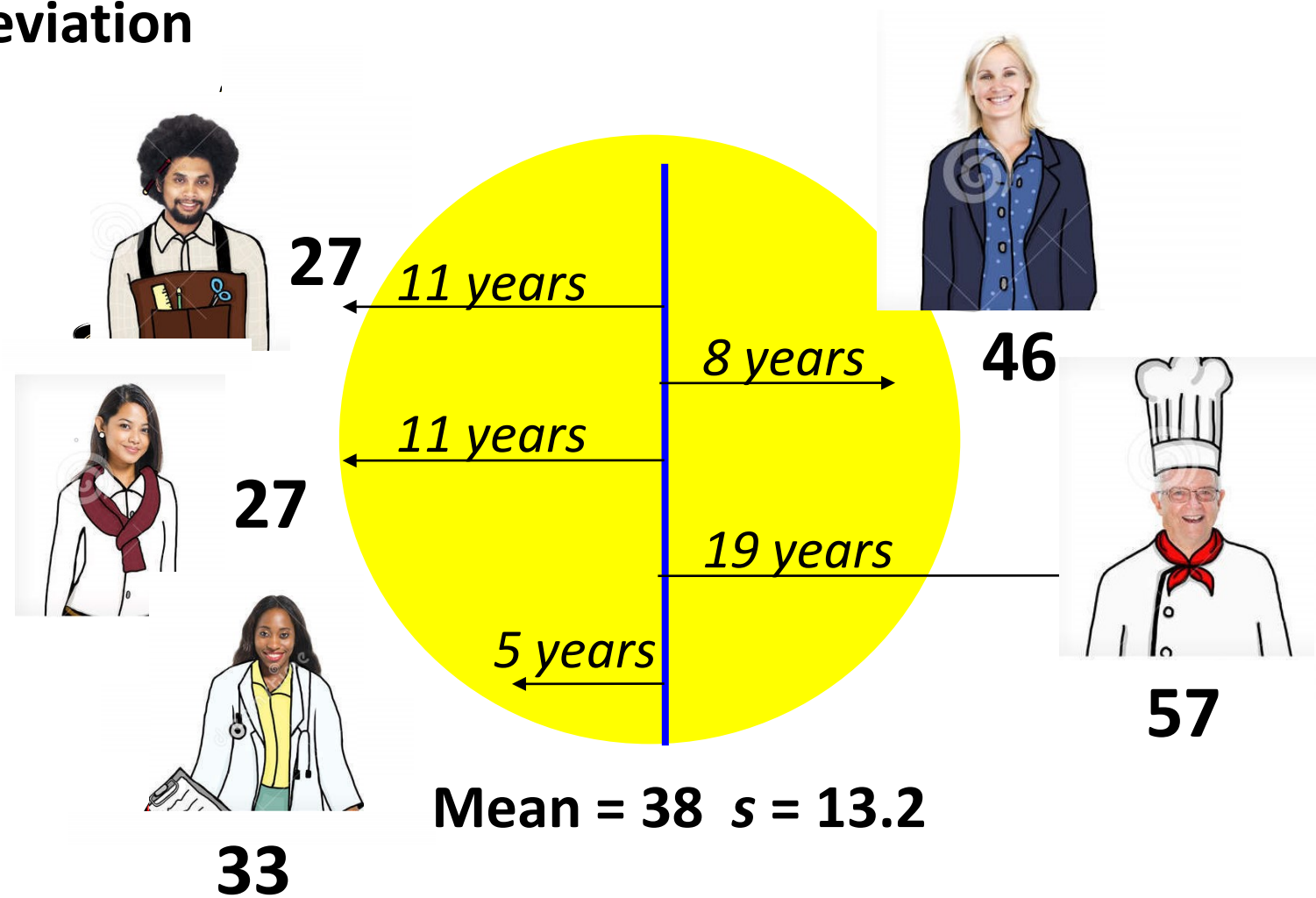
“ X_i ” is the value of each observation in the dataset

“ \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



DESCRIPTIVE STATISTICS

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

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

- 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* ».

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

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

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

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":

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

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

Analysis of Variance or ANOVA

DESCRIPTIVE STATISTICS

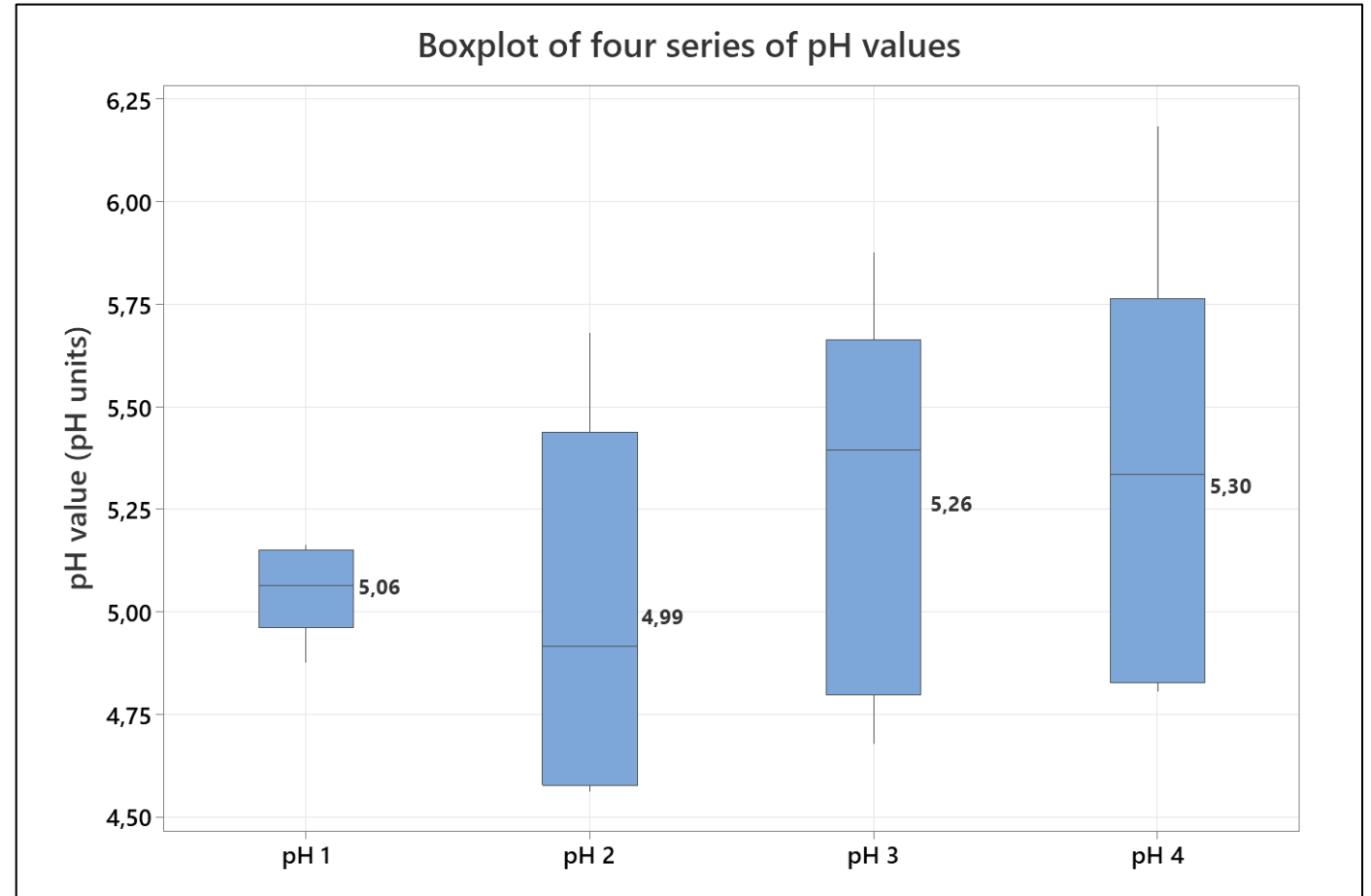
- The « **between variance** », $\sigma_{Between}^2$, or « variance of group means », measures how different the group means are from each other.
- The « **within variance** », σ_{Within}^2 , or « mean of group variances », provides a summary of the level of variability present within each data group.
- In applying these criteria to regression analysis using the least squares method, the $\sigma_{Between}^2$ is called the **explained variance** while the σ_{Within}^2 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
5,05	5,68	5,39	4,85
4,88	4,92	4,92	5,34
5,16	5,20	5,88	5,34
5,14	4,59	4,68	4,81
5,07	4,56	5,45	6,18



DESCRIPTIVE STATISTICS

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

Groups	Count	Sum	Mean	Variance
pH 1	5	25,2960	5,0592	0,0127
pH 2	5	24,9504	4,9901	0,2171
pH 3	5	26,3193	5,2639	0,2225
pH 4	5	26,5228	5,3046	0,3079

ANOVA						
Source of variation	Sum of Squares	dof	Mean of Squares	F calculated	Significance value	F crit
<i>Between groups</i>	0,3530	3	0,1177	0,6191	0,6127	3,2389
<i>Within groups</i>	3,0406	16	0,1900			
Total	3,3936	19				

DESCRIPTIVE STATISTICS

What does ANOVA One-Way tell us?

- The means of squares (or variances) are greater within individual data groups than between them. In other words:
variability (measured by the deviation from the mean) *is higher within the groups than between them!*
Measurement problems? May be, but it has to be ascertained 😊
- $F_{calculated} < F_{tabulated}$: average values of the data groups are not significantly different from each other 😊

DESCRIPTIVE STATISTICS

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 the same 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:

$$CV = \frac{\sigma}{\mu} \times 100$$

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

Example:

Yield Process A (%):	99.8	100.1	100.0	100.7	99.7	100.0	100.2	100.7	98.8
	Mean: 100.0 s = 0.52 RSD = 0.52%								
Yield Process B (%):	97.4	99.2	101.0	101.6	99.0	100.2	100.6	100.7	100.0
	100.5 Mean: 100.0 s = 1.20 RSD = 1.20%								

Conclusion: The yield of both processes is, on the average, equal to 100.0%, but process B is more variable.

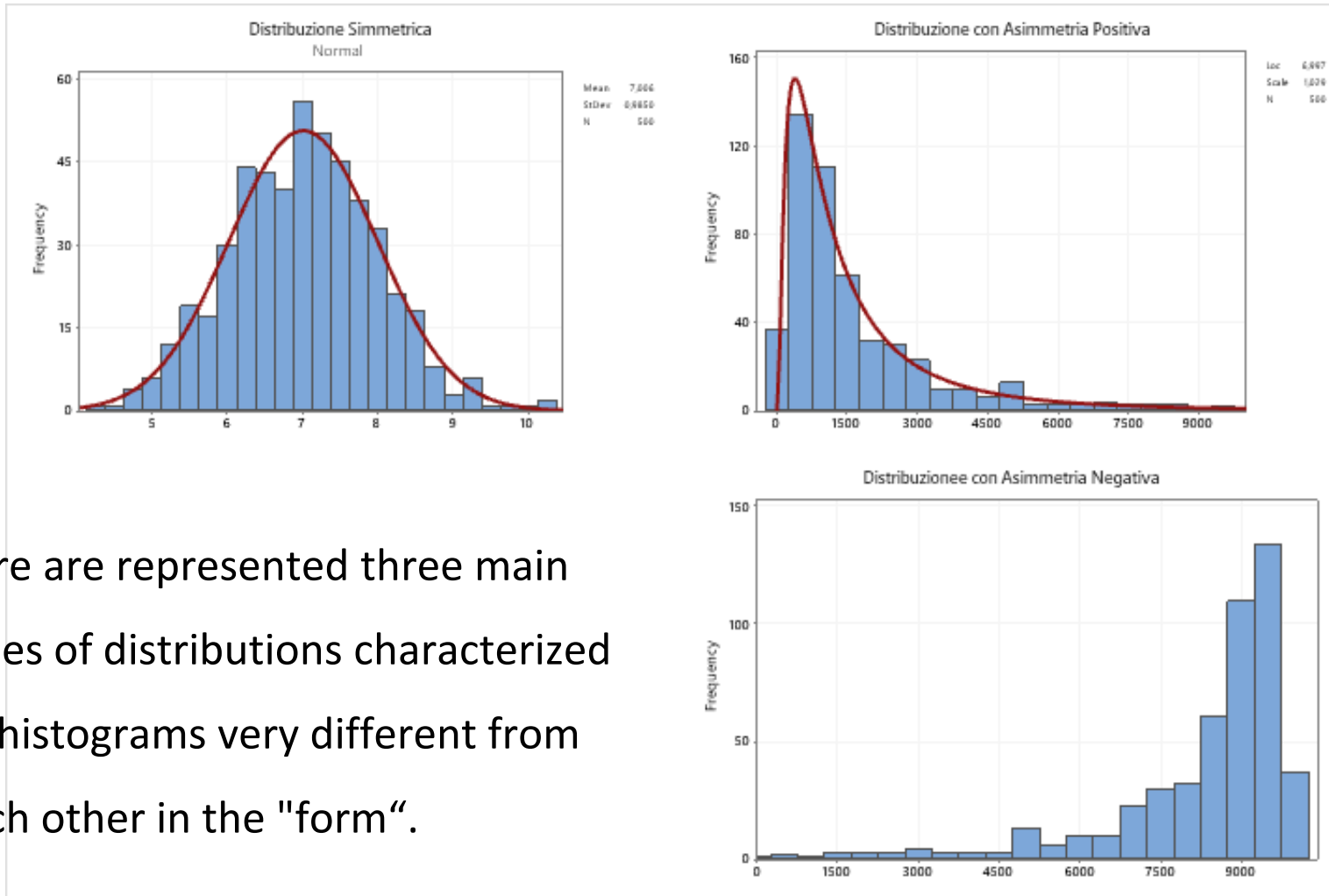
DESCRIPTIVE STATISTICS

- The third type of indices are the:

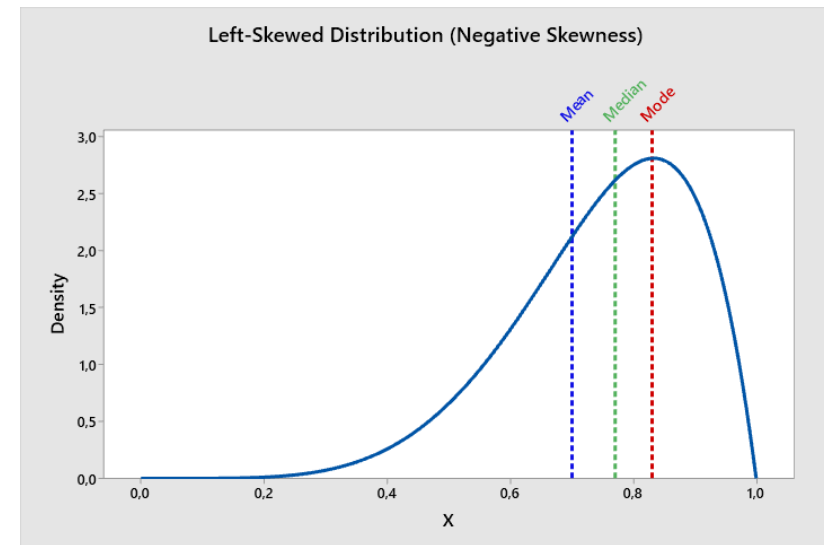
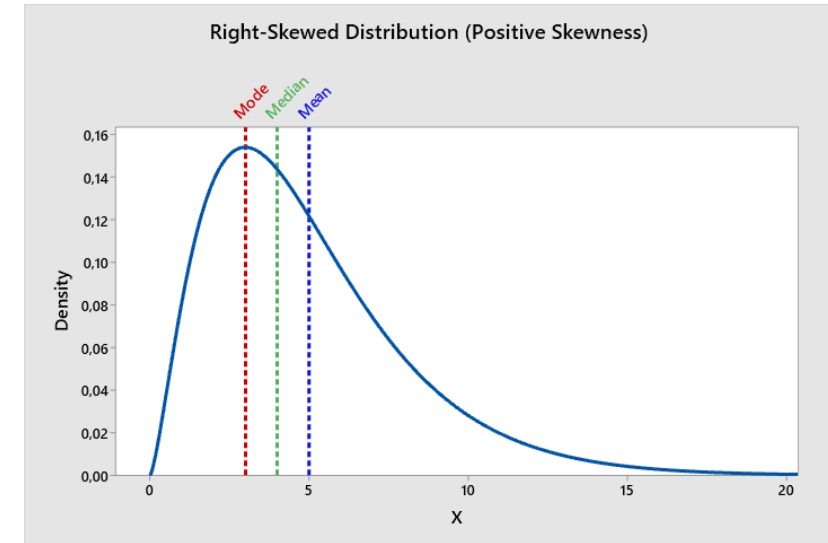
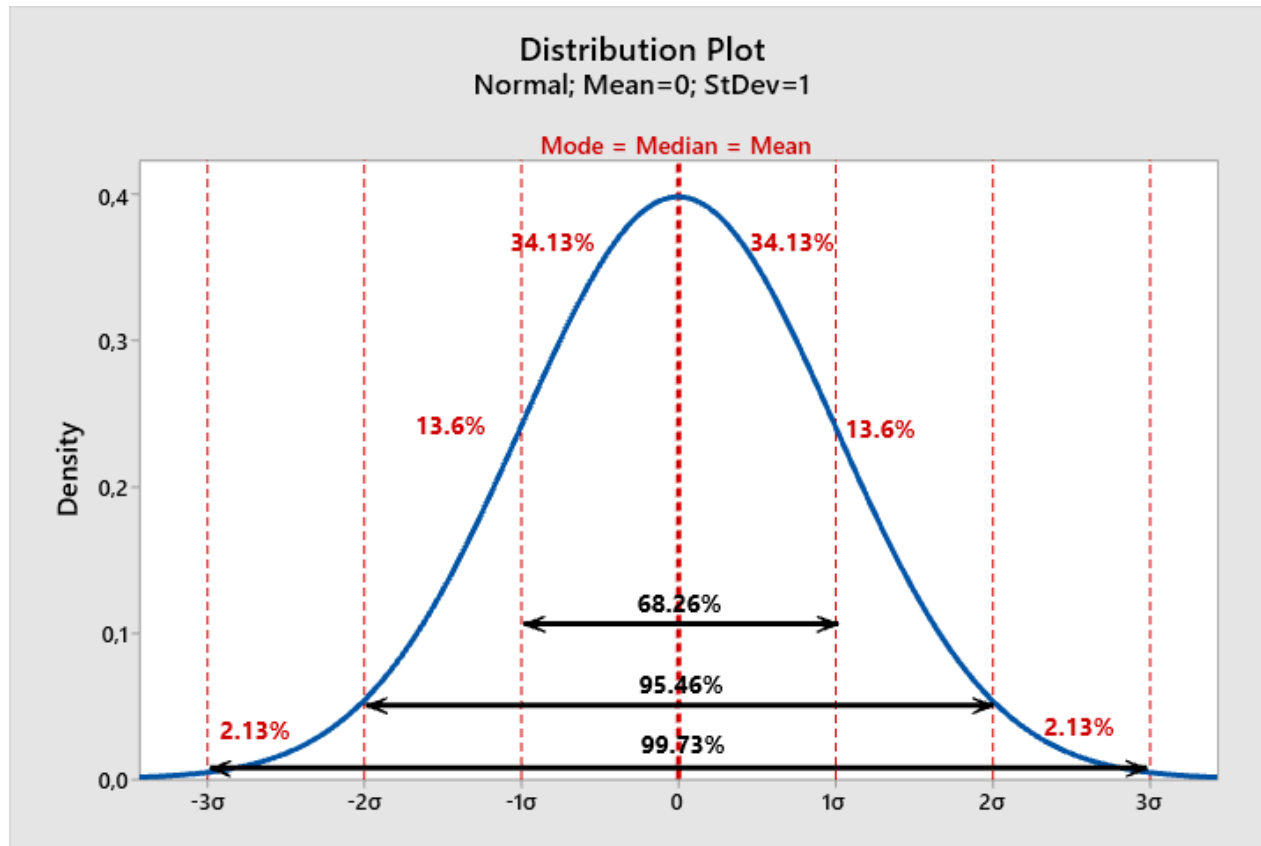
SHAPE INDICES

- 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



DESCRIPTIVE STATISTICS



DESCRIPTIVE STATISTICS

- FISHER or SKEWNESS ASYMMETRY INDEX:

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

- 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)

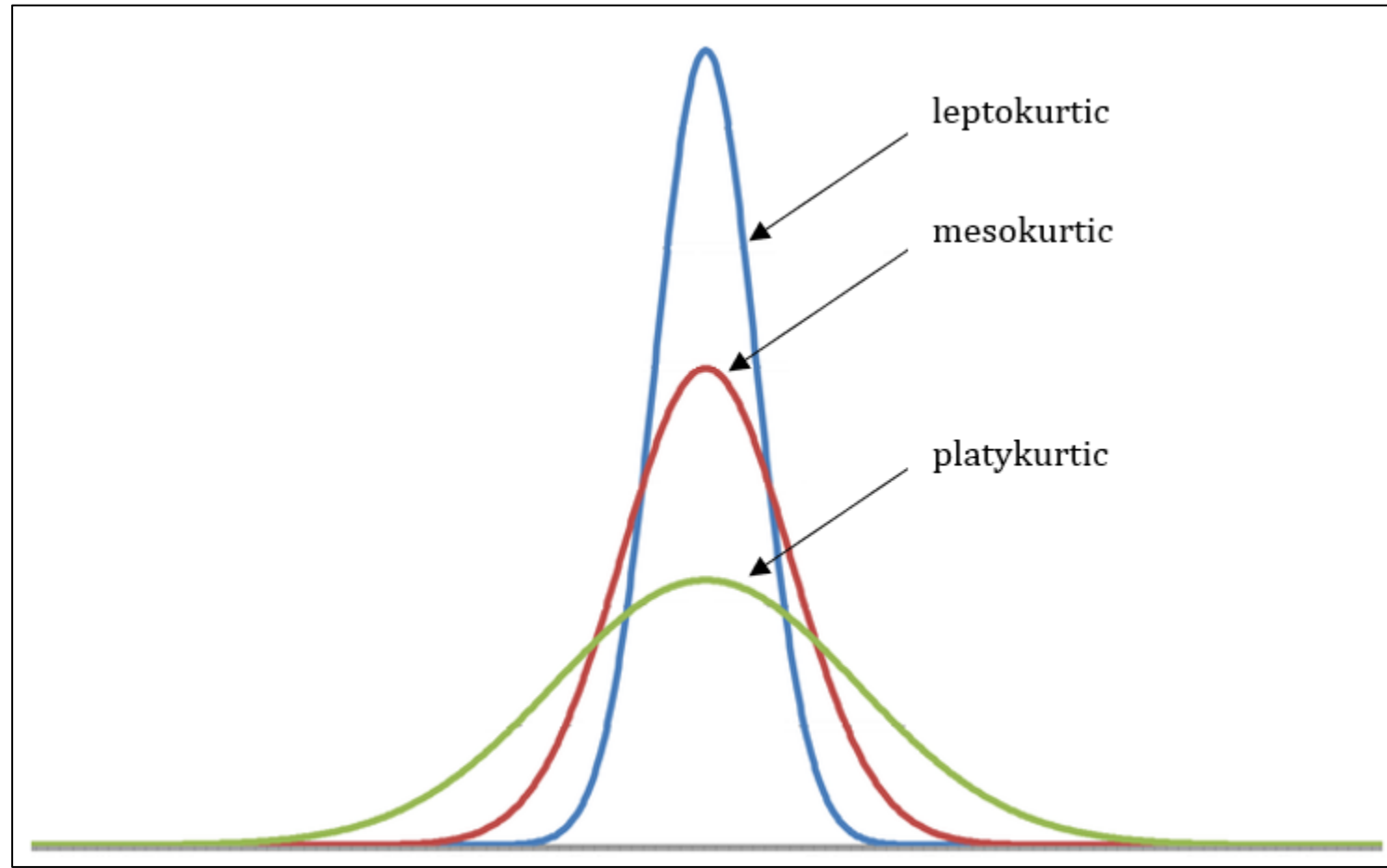
DESCRIPTIVE STATISTICS

■ KURTOSIS:

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

- 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



INFERENTIAL STATISTICS

INFERENCEAL STATISTICS

- Is that part of the Statistics that aims to make operational decisions and choices on the basis of limited and provisional information.
- It represents the « *confirmation moment of reality* »
- **INFERENCE:**
 - is the process of reaching a conclusion from a given set of statements (or *premises*)
 - it is of two types: **deductive** and **inductive**

INFERENCE 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!

INFERENCE STATISTICS

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

INFERENCEAL STATISTICS

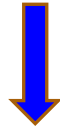
- **Statistical hypothesis**: an **assertion** regarding the parameters of one or more populations that we want to test or investigate.
- **Hypothesis testing** : the **procedure** that leads to a decision concerning a particular hypothesis and is based on a random sample extracted from the population of interest (survey).

INFERENCEAL 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.

INFERENCEAL STATISTICS

- **Example:** 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: $H_0: \mu = 100 \text{ kg}$ (Null hypothesis)

$H_1: \mu \neq 100 \text{ kg}$ (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

To do this work we need a few concepts that are the basis of the Inferential Statistics and precisely:

- *Probability* and *Probability Distribution*
- *alpha* (level of significance) or *level of confidence*
- *P-value*
- *Test statistics*

Let's open a parenthesis to introduce these concepts!

INFERENCEAL 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.

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

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).

INFERENCEAL 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 $\frac{1}{6}$ since there are 6 possible events (as many as there are faces of the die) and the favorable event is only one.

INFERENCEAL STATISTICS

If facing the occurrence of two or more events, it must be first considered if they are :

- ***Compatible*** (or *not mutually exclusive*) or ***Non-Compatible*** (or *mutually exclusive*)

After that, ***only compatible events*** can additionally be:

- ***Dependent*** or ***Independent***

INFERENCEAL STATISTICS

Two or more events are **compatible** (or **not mutually exclusive**) if they can occur at the same time and **incompatible** (or **mutually exclusive**) if they can't.

For instance, if the event consists in assessing if a tablet is *defective* or *flawless*, one possibility excludes the other. A sampled tablet can only be flawless or defective.

It is different if the event consists in assessing the possible defects affecting a tablet (*i.e.*, capping, chipping, *etc.*). In this case the several possibilities are not mutually exclusive among them.



Capping



Chipping

INFERENCEAL STATISTICS

Two *incompatible events* are those in which if one event occurs the other cannot occur.

(Note: In this case concepts of dependent or independent events do not apply!)

In this case, the probability of occurrence of two or more *incompatible (or mutually exclusive)* events is just the sum of the individual probabilities associated to each event.

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

*Addition rule for probabilities
for two mutually exclusive
events*

The logical connector \cup stands for **or**.

INFERENCEAL STATISTICS

Two or more *compatible and dependent events* are those that can occur at the same time.

In this case, the probability of occurrence is the sum of the individual probabilities associated to each event diminished by the overlap:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

*Addition rule for probabilities
for two non-mutually
exclusive events*

The logical connector \cup stands for *or*, while the connector \cap stands for *and*.

INFERENCEAL STATISTICS

Two *compatible and independent events* are those in which the occurrence of one event does not affect the occurrence of the other.

In this case, the probability of occurrence of two independent events is the product of the individual probabilities associated to each event.

$$P(E_1 \cap E_2) = P(E_1) * P(E_2)$$

*Multiplication rule for probabilities
for two independent events*

The logical connector \cap stands for *and*.

INFERENCEAL STATISTICS

What the above formula should remind you ?

$$\textit{Risk Priority Index (RPI)} = S \times O \times D$$

S = SEVERITY

O = OCCURRENCE

D = DETECTABILITY

INFERENCEAL STATISTICS

	<i>COMPATIBLE events</i>		<i>INCOMPATIBLE events</i>
$E_1 \cup E_2$	$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$		$P(E_1 \cup E_2) = P(E_1) + P(E_2)$
$E_1 \cap E_2$	<i>DEPENDENT events</i>	<i>INDEPENDENT events</i>	$P(E_1 \cap E_2) = 0$
	$P(E_1 \cap E_2) = P(E_1) * P(E_2 E_1)$	$P(E_1 \cap E_2) = P(E_1) * P(E_2)$	

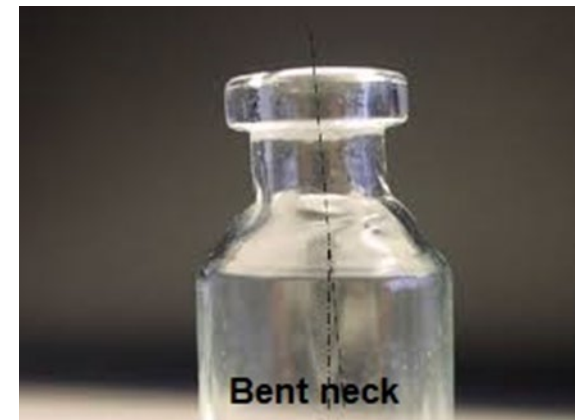
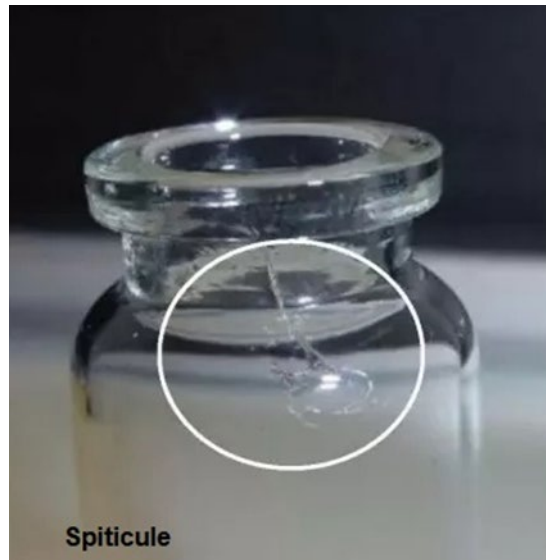
INFERENCEAL STATISTICS

Let's consider 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*.



INFERENCEAL 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

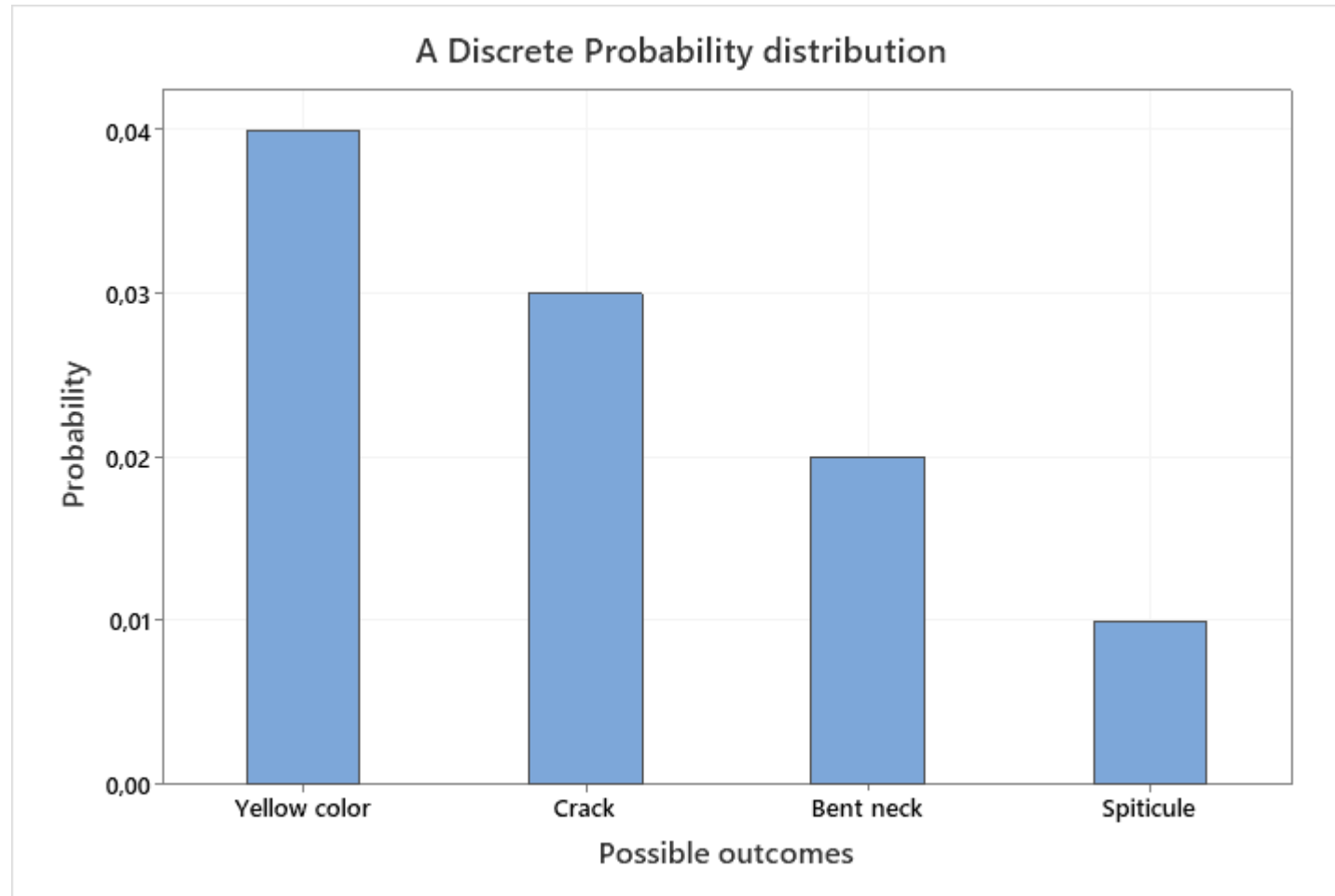
INFERENCEAL STATISTICS

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: $1 - 0.10 = 0.90$ 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:

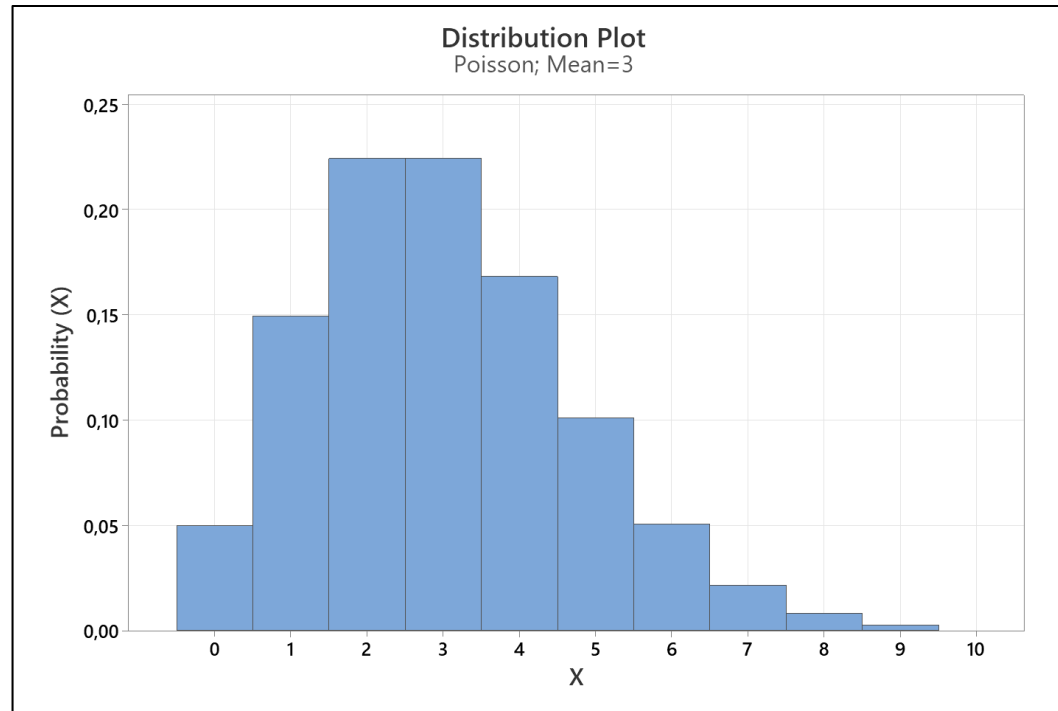
INFERENCEAL STATISTICS



INFERENCEAL STATISTICS

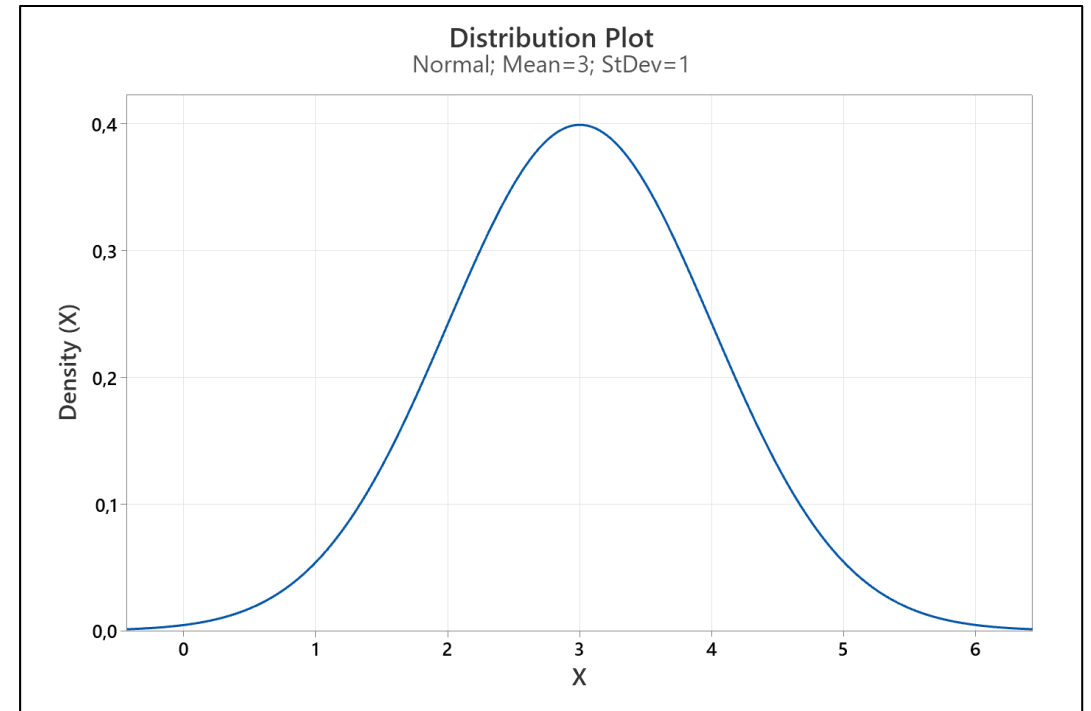
- A **distribution** (or **probability distribution**) is a set of values of a **variable** (in this case: *glass vials defects*), along with the associated probability of each value of the variable.
- Distributions are usually visualized plotting the variable on the x-axis and the probability on the 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**.

INFERENCE STATISTICS



Poisson Distribution

Discrete data and Discrete probability curve



Normal Distribution

Continuous data and Continuous probability curve

INFERENCEAL 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).*

INFERENCEAL 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

INFERENCEAL 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* »

INFERENCE STATISTICS

Mathematically the Poisson law it is defined as:

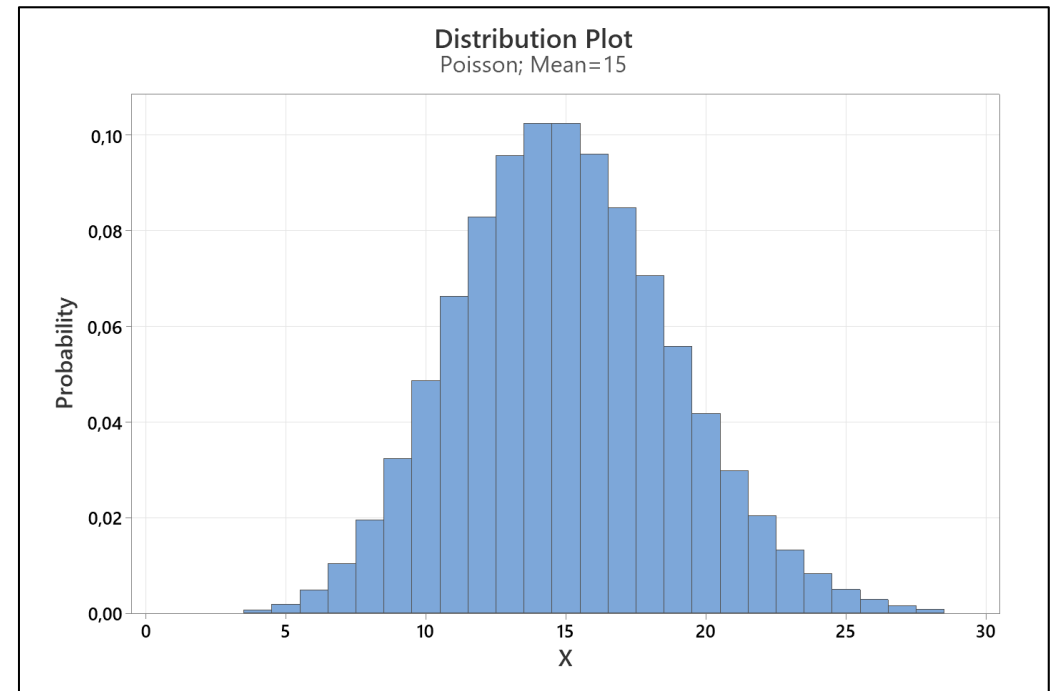
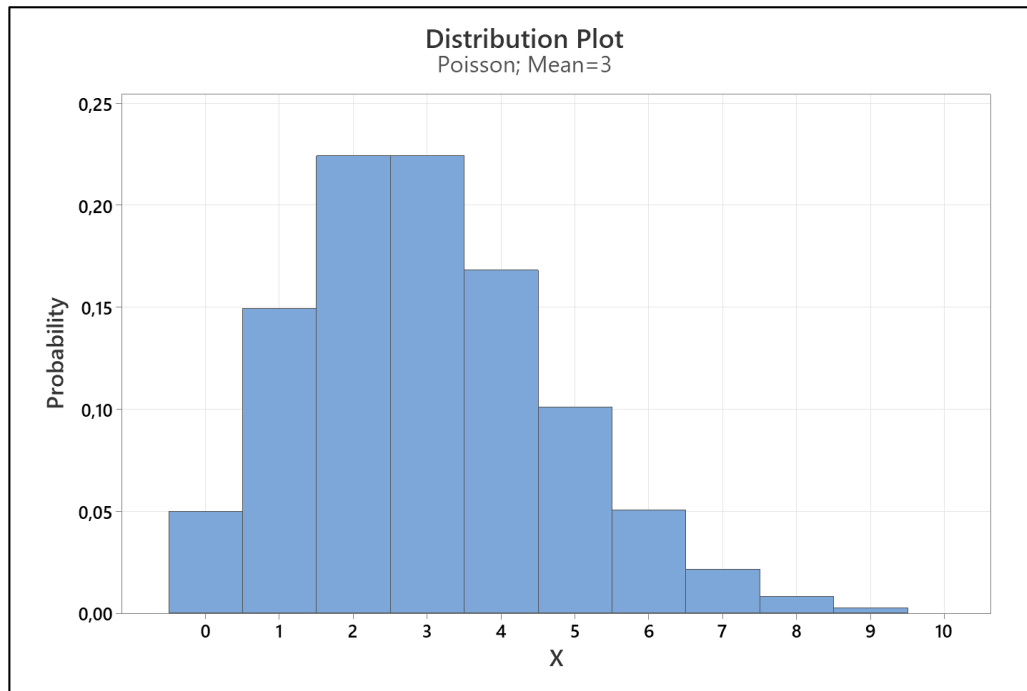
$$p(x) \begin{cases} = \frac{(np)^x}{x!} e^{-np} = \frac{\lambda^x}{x!} e^{-\lambda} & x = 0, 1, 2, \dots \\ = 0 & \text{elsewhere} \end{cases}$$

and its variance is equal to the mean and the parameter λ :

$$\sigma^2 = \mu = \lambda$$

Because of this there are «different» Poisson Distributions for different values of the mean, μ .

INFERENCE STATISTICS



INFERENCEAL 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.*

INFERENCEAL STATISTICS

Consider the case, for example, of black particles found by inspecting the samples of 80 different batches of tablets (please, note that it would be the same even in case of vials or lots of APIs).

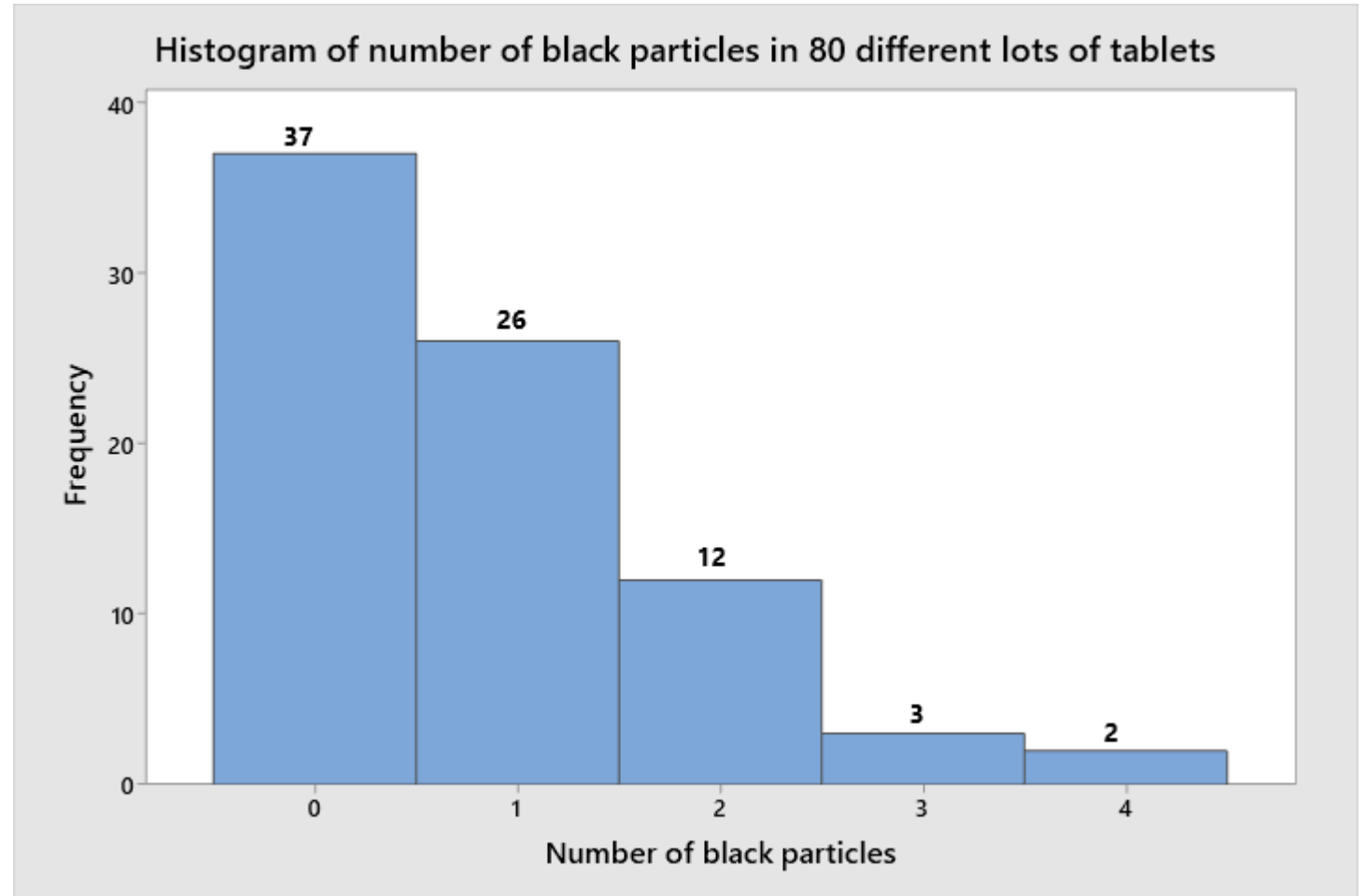
Lot	No.	Lot	No.	Lot	No.	Lot	No.	Lot	No.	Lot	No.	Lot	No.	Lot	No.
1	0	11	2	21	0	31	1	41	1	51	1	61	0	71	0
2	1	12	2	22	0	32	1	42	0	52	2	62	3	72	0
3	1	13	2	23	0	33	2	43	0	53	4	63	4	73	0
4	0	14	0	24	0	34	1	44	1	54	1	64	1	74	1
5	0	15	2	25	0	35	1	45	0	55	1	65	1	75	2
6	0	16	3	26	0	36	0	46	0	56	1	66	0	76	3
7	0	17	0	27	0	37	0	47	2	57	0	67	0	77	0
8	0	18	0	28	2	38	0	48	0	58	0	68	0	78	1
9	1	19	0	29	1	39	2	49	2	59	1	69	1	79	1
10	1	20	0	30	1	40	1	50	2	60	1	70	1	80	0

INFERENCEAL STATISTICS

The histogram here on the side shows the different numbers of black particles in the previous table, each with its own frequency.

In summary:

No. Black-specks	0	1	2	≥ 3
Frequencies	37	26	12	5



INFERENCEAL STATISTICS

Descriptive Statistics

N Mean

80 0,8375

Observed and Expected Counts for No. black-specks

No. black-specks	Poisson Probability	Observed Count	Expected Count	Contribution to Chi-Square
0	0,432791	37	34,6233	0,163149
1	0,362463	26	28,9970	0,309758
2	0,151781	12	12,1425	0,001672
>=3	0,052965	5	4,2372	0,137321

1 (25,00%) of the expected counts are less than 5.

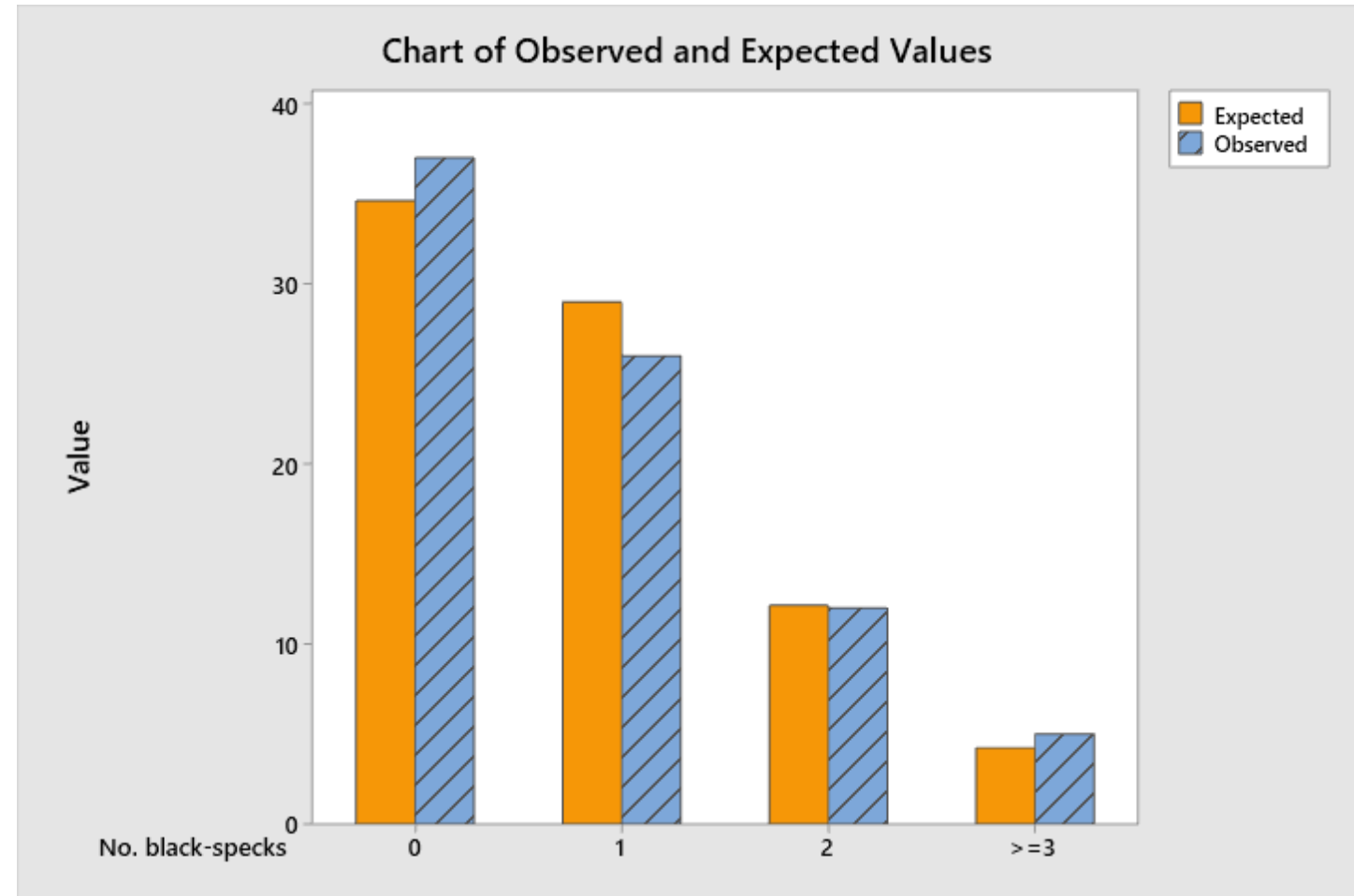
Chi-Square Test

Null hypothesis H_0 : Data follow a Poisson distribution

Alternative hypothesis H_1 : Data do not follow a Poisson distribution

DF Chi-Square P-Value

2 0,611900 0,736



INFERENCEAL STATISTICS

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

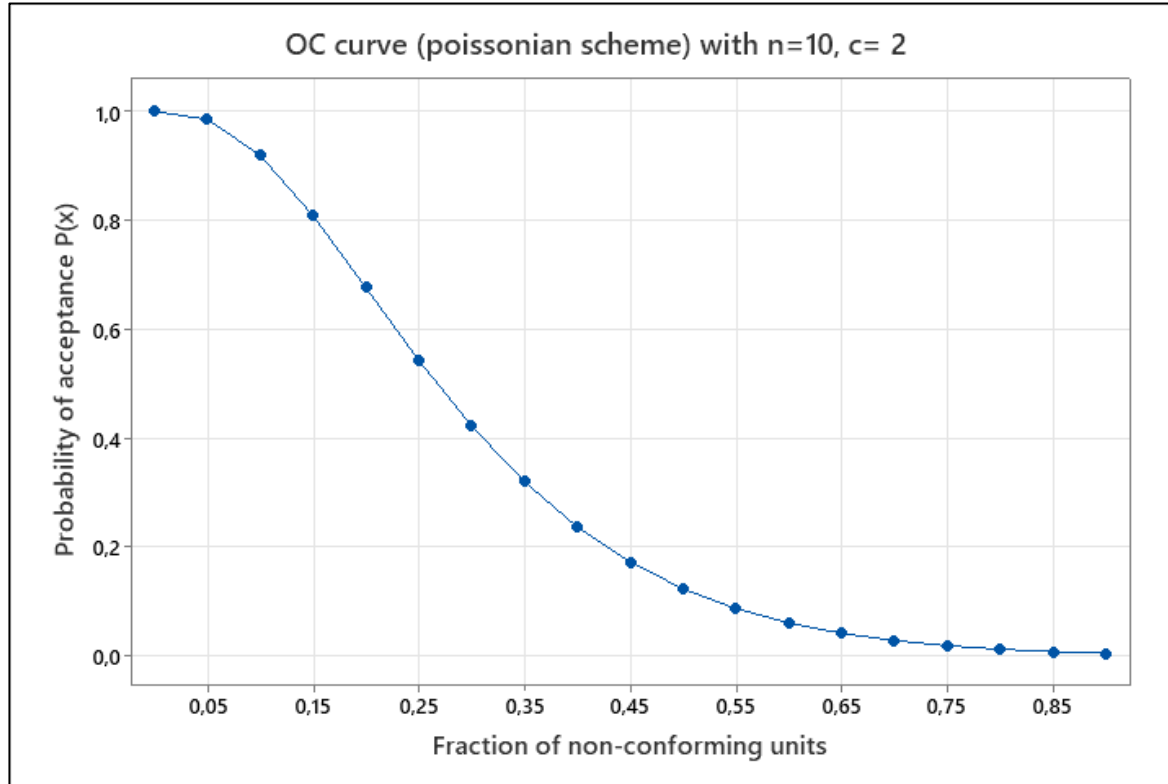
Here is an example of construction of the Characteristic Operating Curve in the Poissonian case:

$$N = 100 \quad n = 10 \quad c = 2$$

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

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

INFERENCE STATISTICS



INFERENTIAL STATISTICS

In regard to the **Normal Curve**, it is due to the French mathematician **Abraham De Moivre** who mentioned it first in a paper published on November 12, 1733 and shared only to friends.

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).

However, 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.

Karl Pearson started using the term only in 1893.

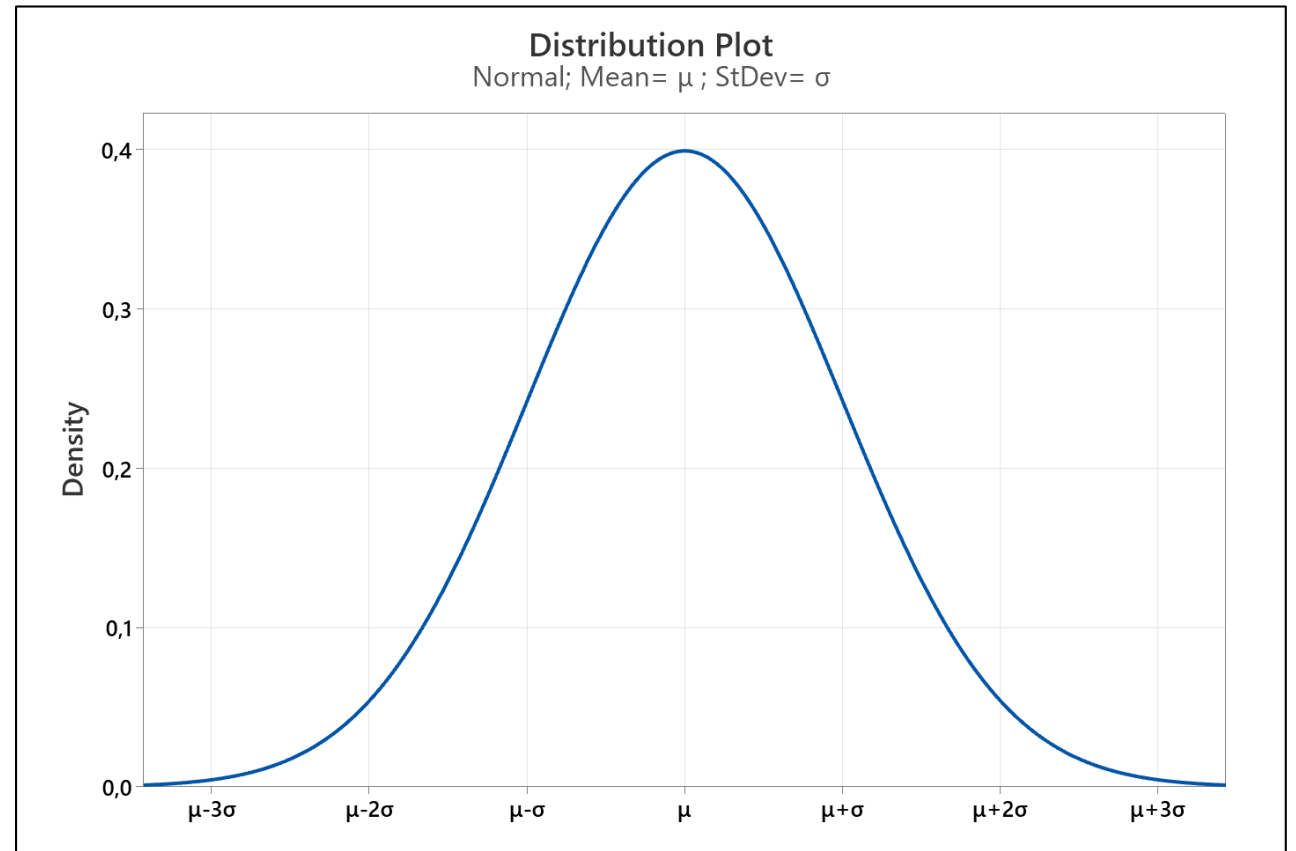
INFERENCEAL STATISTICS

Mathematically the Normal Distribution is defined as follows:

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

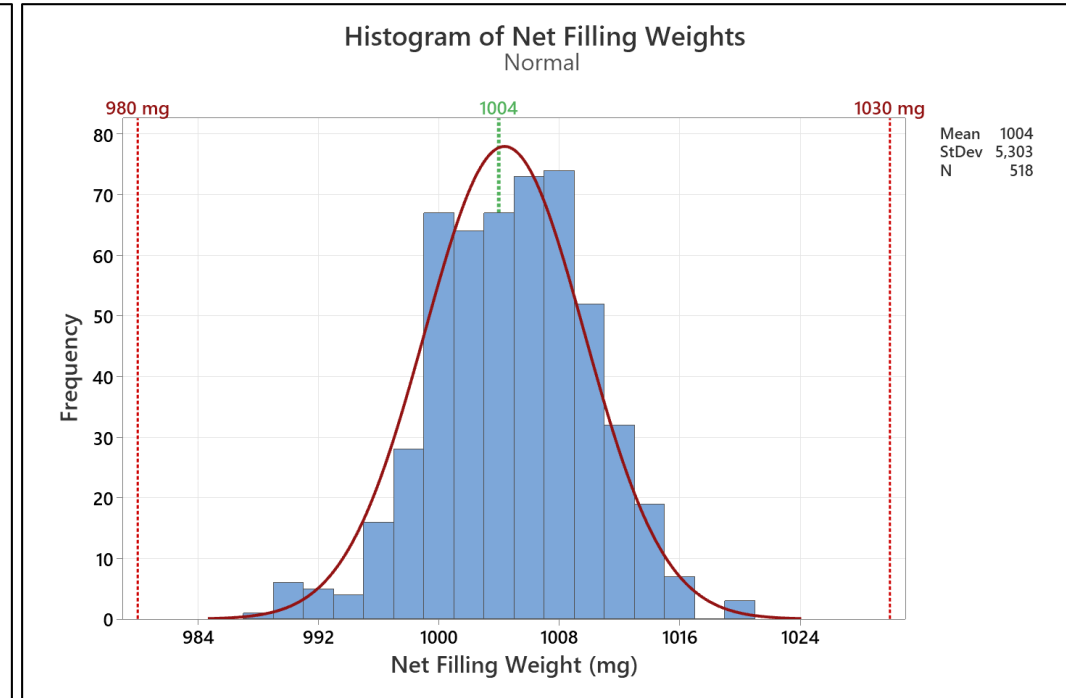
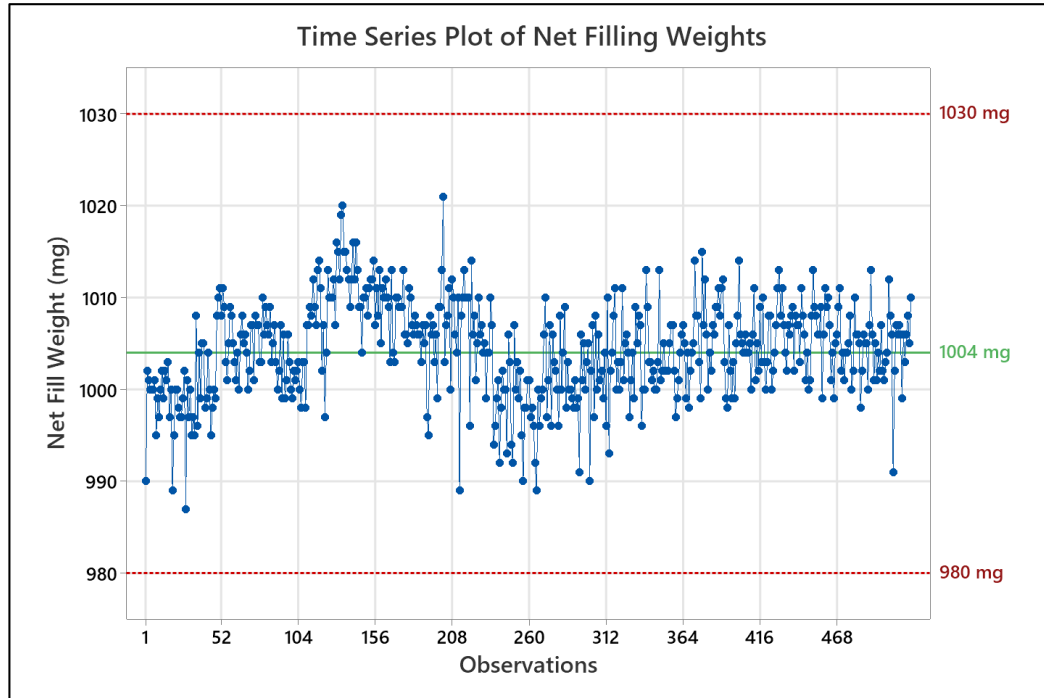
and its graphical aspect is that of a bell curve symmetrical with respect to μ .

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



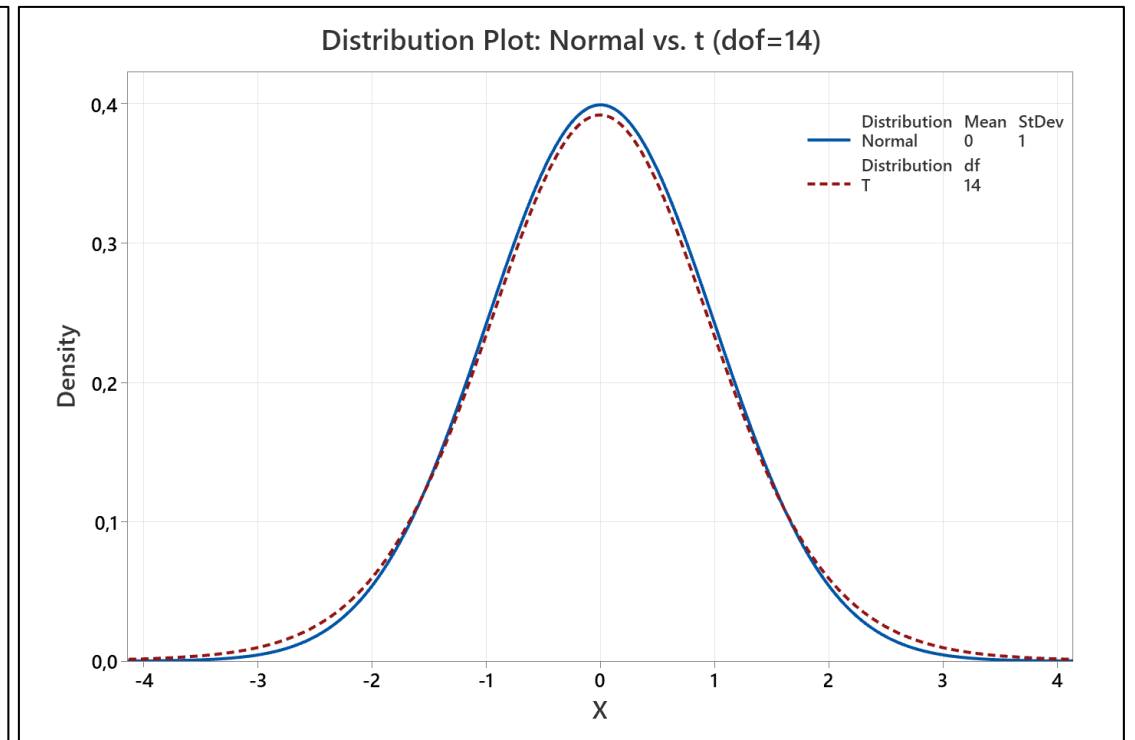
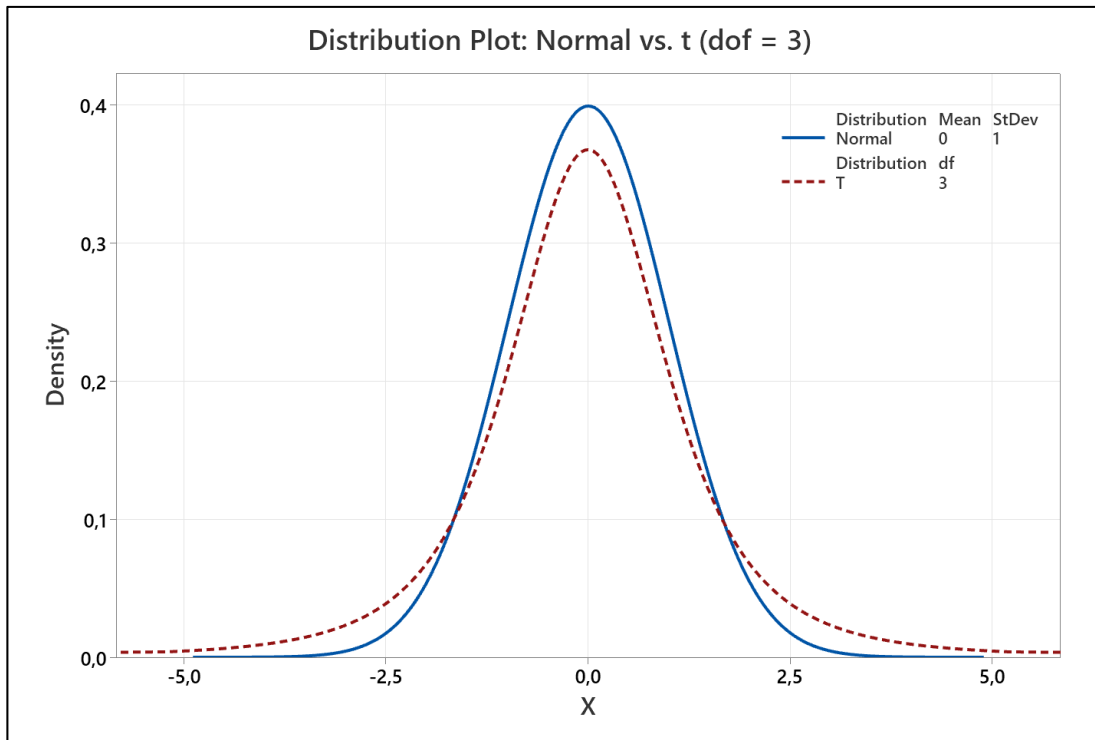
S.M. Ross, A first course in probability– 9th Edition, Pearson College (2012)

INFERENCEAL STATISTICS



INFERENCEAL STATISTICS

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



INFERENCEAL 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

INFERENCEAL STATISTICS

What is the practical use of all this?

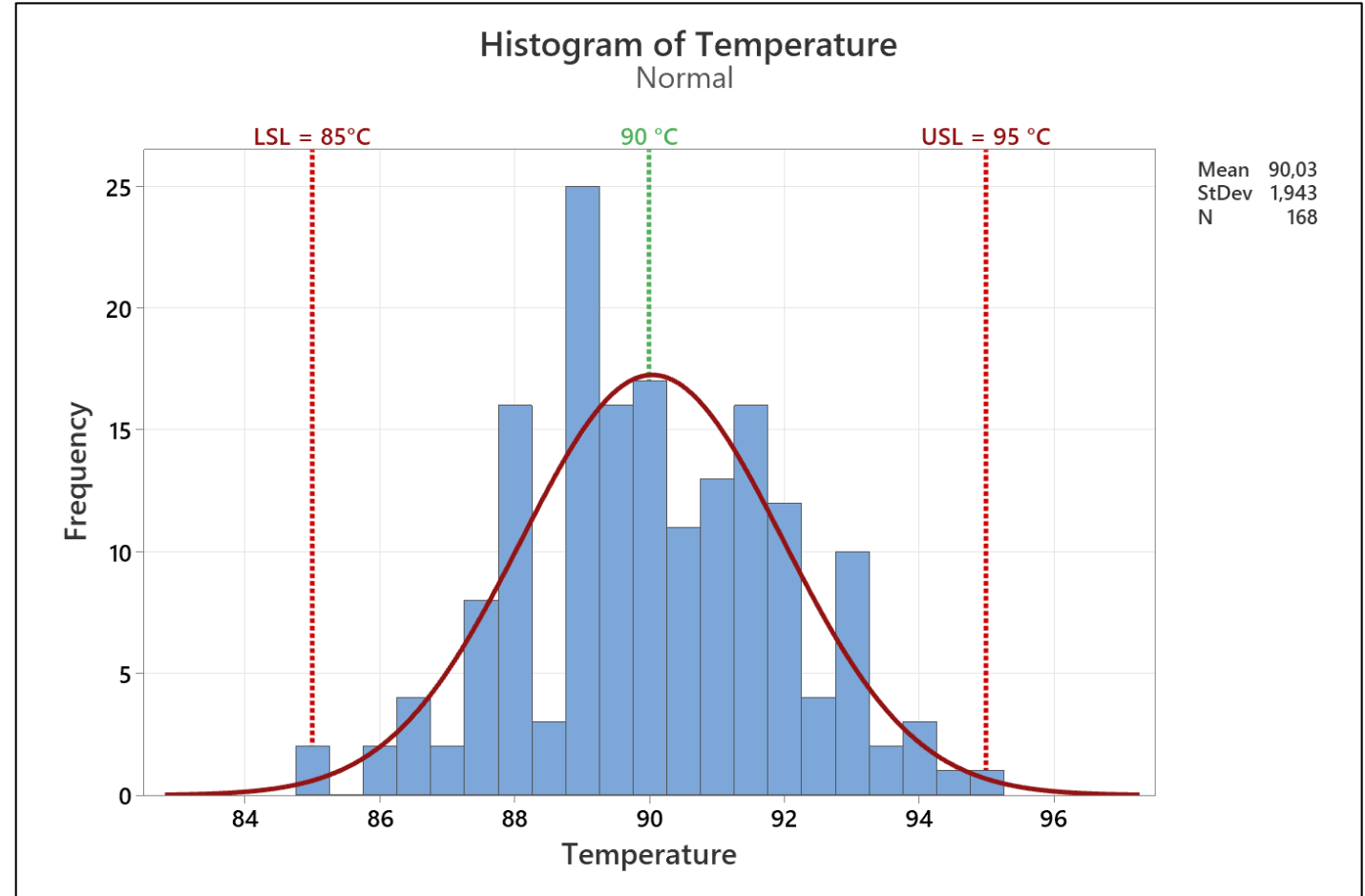
Let see a practical example !

INFERENCE STATISTICS

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

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

$$\bar{x} = 90\text{ }^{\circ}\text{C} \quad s = 1.9\text{ }^{\circ}\text{C}$$

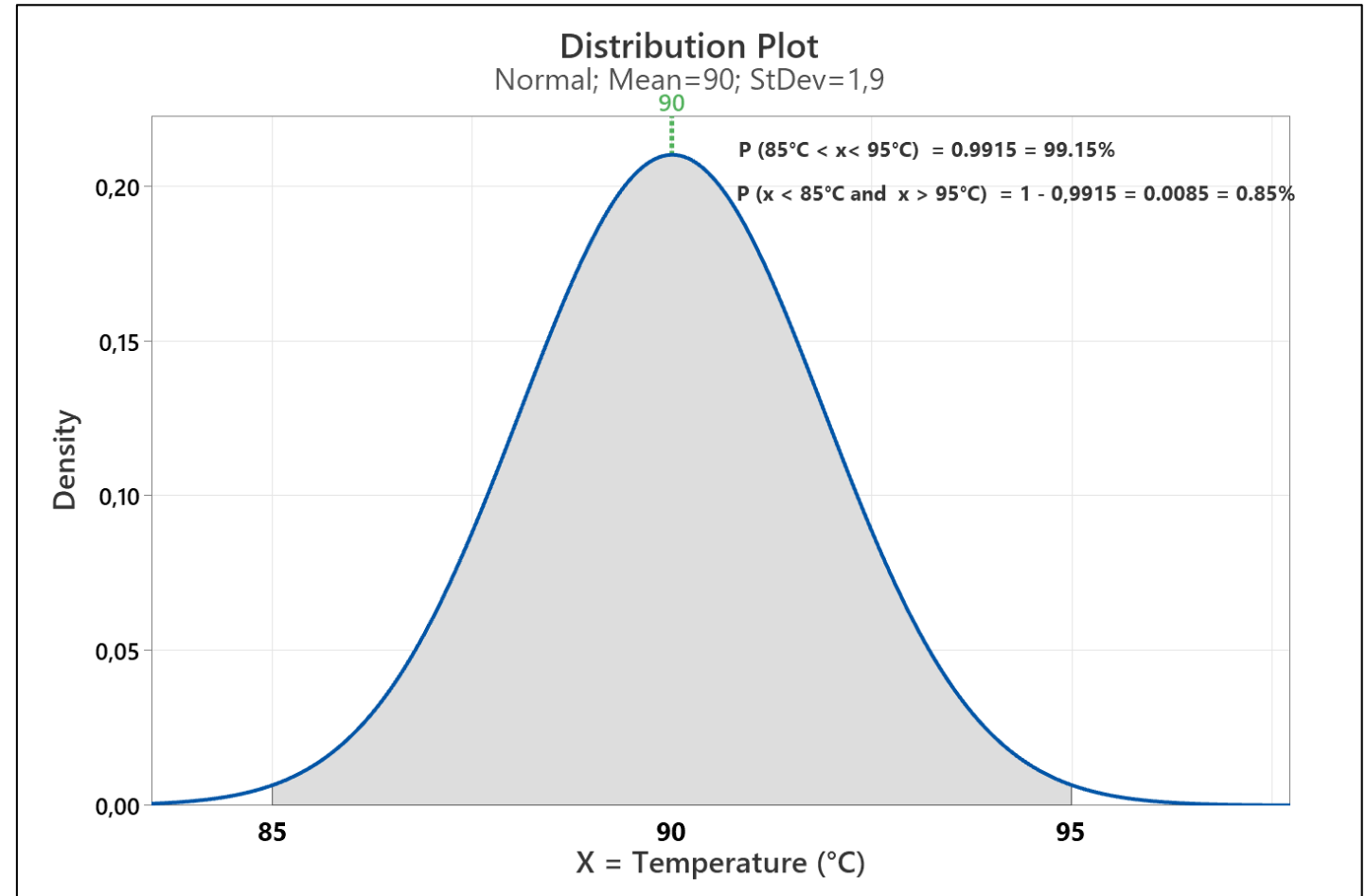


INFERENCE STATISTICS

What is the probability

$$P(X < 85^{\circ}\text{C and } X > 95^{\circ}\text{C}) ?$$

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



INFERENCEAL STATISTICS

The same result can also be achieved by performing the calculations by hand and making use of the standard tables.

It is just a matter of calculating the Z-test statistic at the two limits of the interval (85 and 95) around the average value of 90 °C or, much more simply, only in one of the two limits (e.g., 95°C) since the interval is symmetrical:

$$Z = \frac{\bar{x} - \mu}{\sigma} = \frac{95 - 90}{1.9} = 2.63$$

Since from table of Normal Standardized Distribution it can be found that the area under the curve between 0 and z (=2.63) is 0.4957 it is straightforward to obtain the whole area under the curve between 85°C and 95°C as twice that value. Therefore:

$$P(85 < x < 95) = 0.4957 \times 2 = 0.9914 \text{ or } 99.14\% \quad \text{c.v.d.}$$

INFERENTIAL STATISTICS

What does this mean?

There is less than 1% probability (0.85% to be precise) that the critical reaction parameter exceeds the limits!

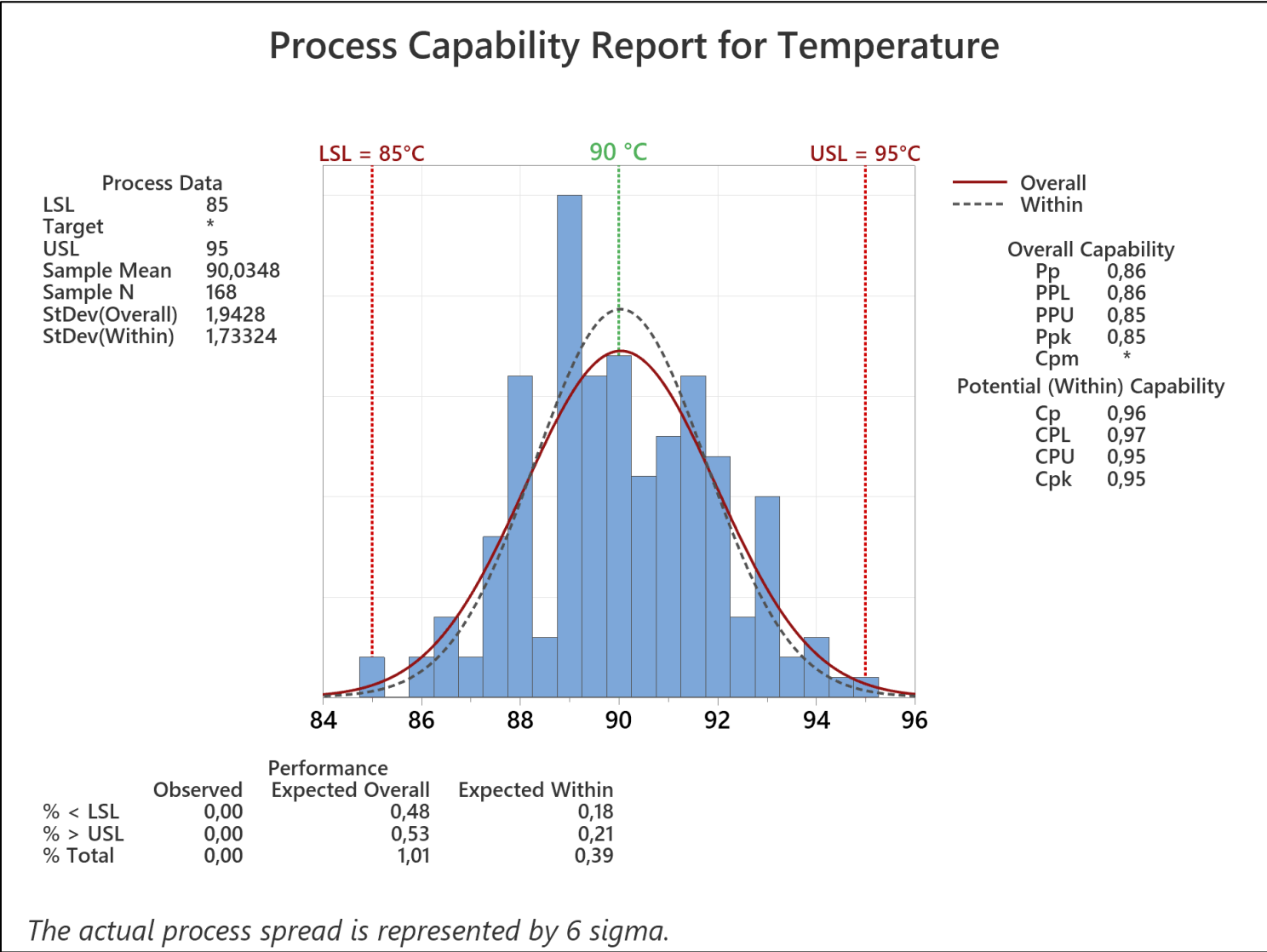
INFERENCEAL STATISTICS

But there is also much more....

In this case, as well as in many others that occur daily,
the possibility of OOS cannot be excluded *a priori* 😊

and, last but not least....

INFERENTIAL STATISTICS



INFERENCEAL STATISTICS

- This can be considered a simple example of

Science based QA

- The conformance to specifications can be demonstrated
- Any future actions can be taken correctly

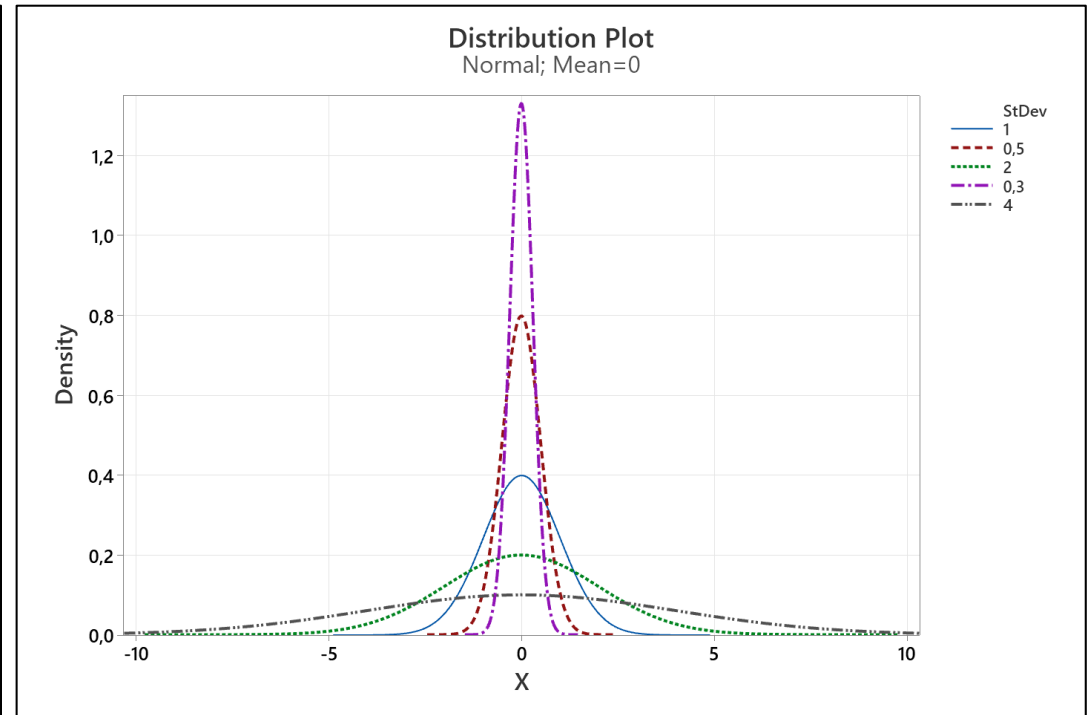
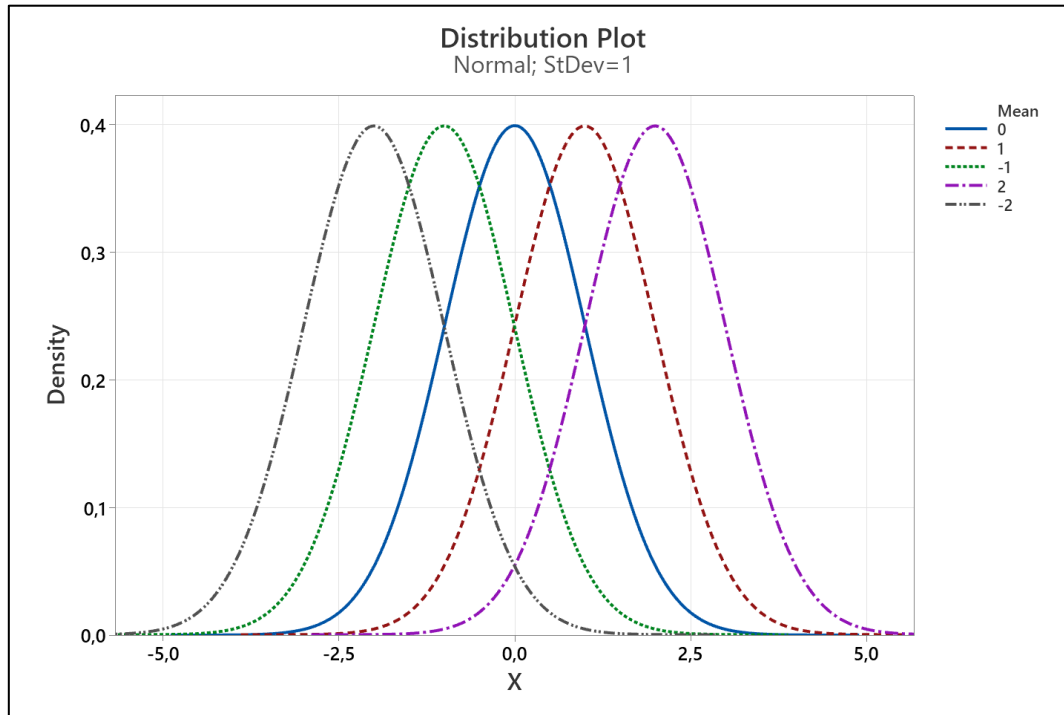
Better Science = Better Outcomes = Less Costs

INFERENCEAL STATISTICS

*Let's now go back to the
Normal Distribution and its characteristics !*

INFERENCEAL STATISTICS

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



INFERENCE STATISTICS

To simplify :

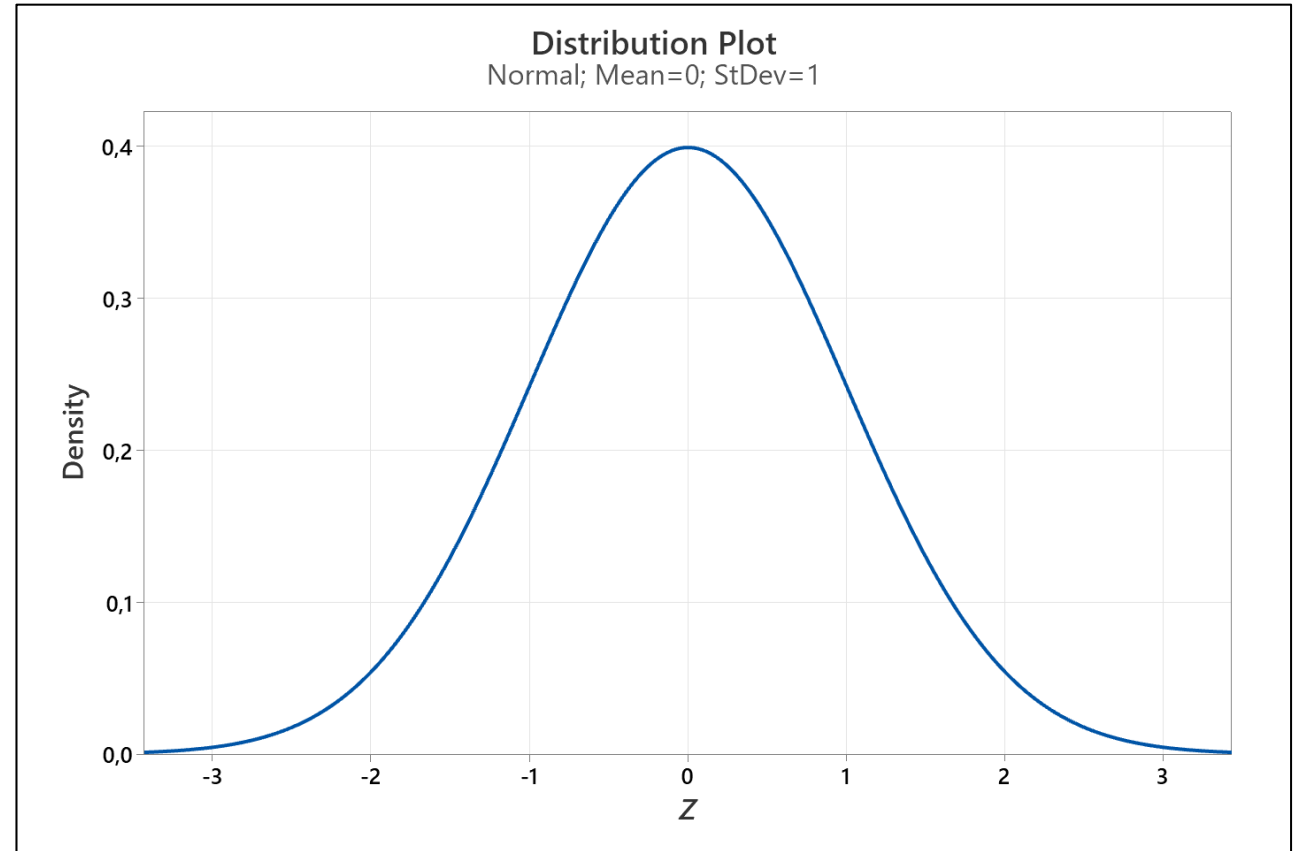
STANDARDIZATION

In other words:

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

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

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

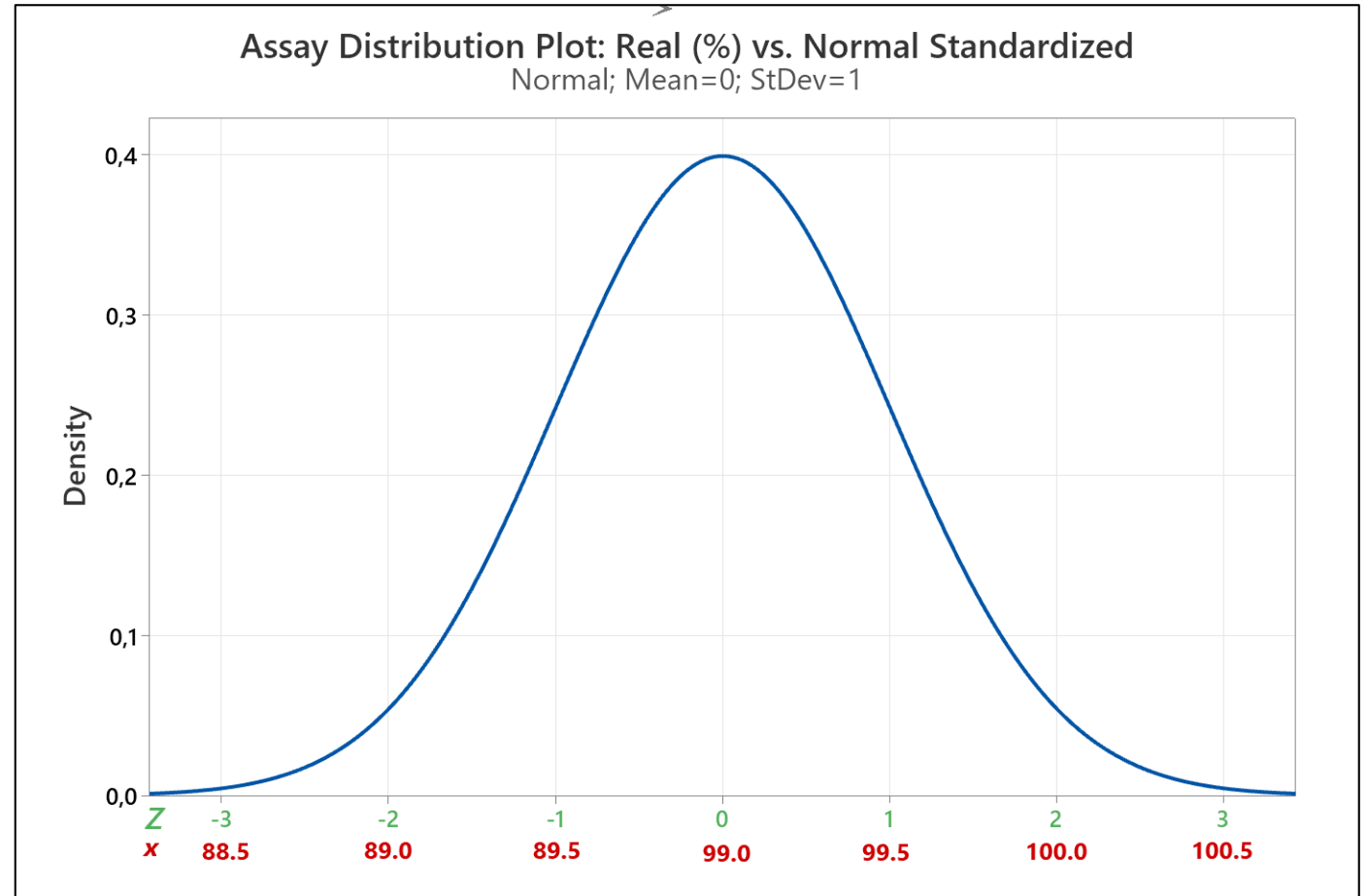


S.M. Ross, A first course in probability – 9th Edition, Pearson College (2012)

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.

Example: The typical assay value for an API is 99.0% with a standard deviation of 0.5%



INFERENCE STATISTICS

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, dealing with mathematical functions, the theory provides simple formulas for calculating the average and other parameters of practical use (e.g., variance, etc.)*

INFERENCEAL STATISTICS

THEREFORE:

***IF THE REAL DATA IS NOT NORMALLY DISTRIBUTED
IT IS NOT THE END OF THE WORLD!***

If the data are not normal, they can be normalized by performing mathematical operations on them (*e.g.*, natural logarithm, square root, reciprocal, *etc.*) or by using tests of a different type, the so-called « non-parametric tests ».

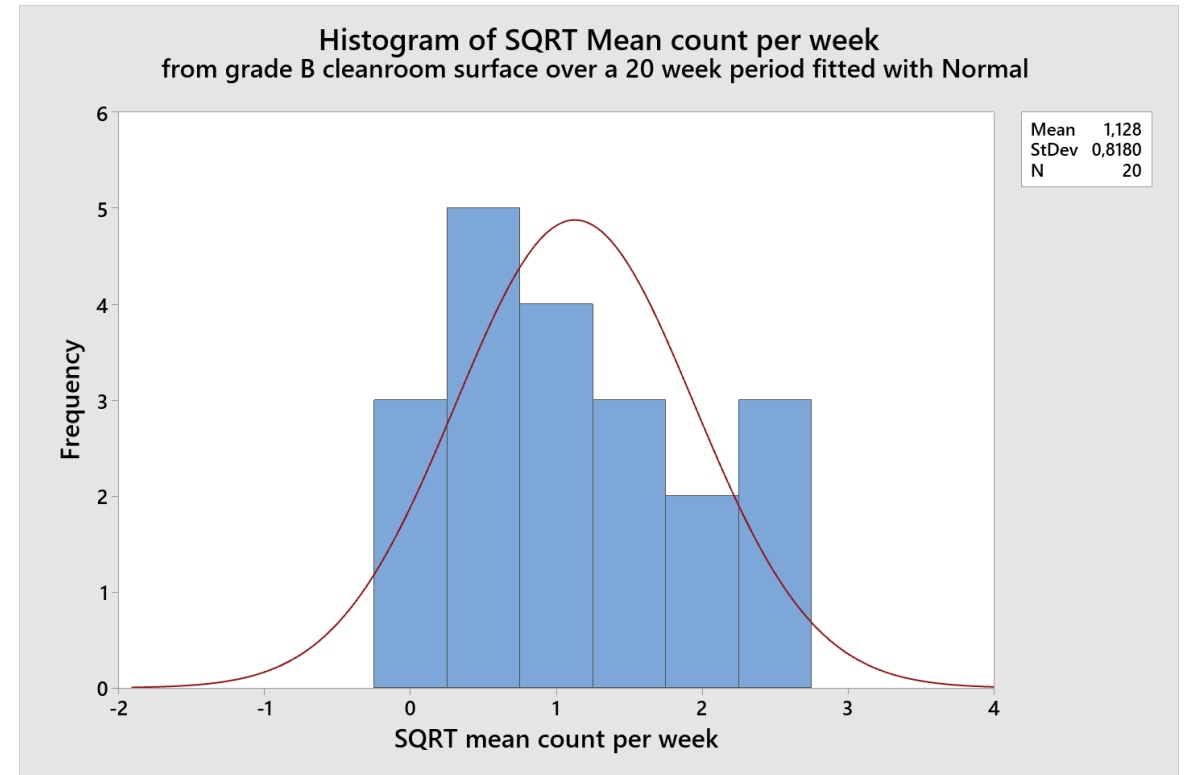
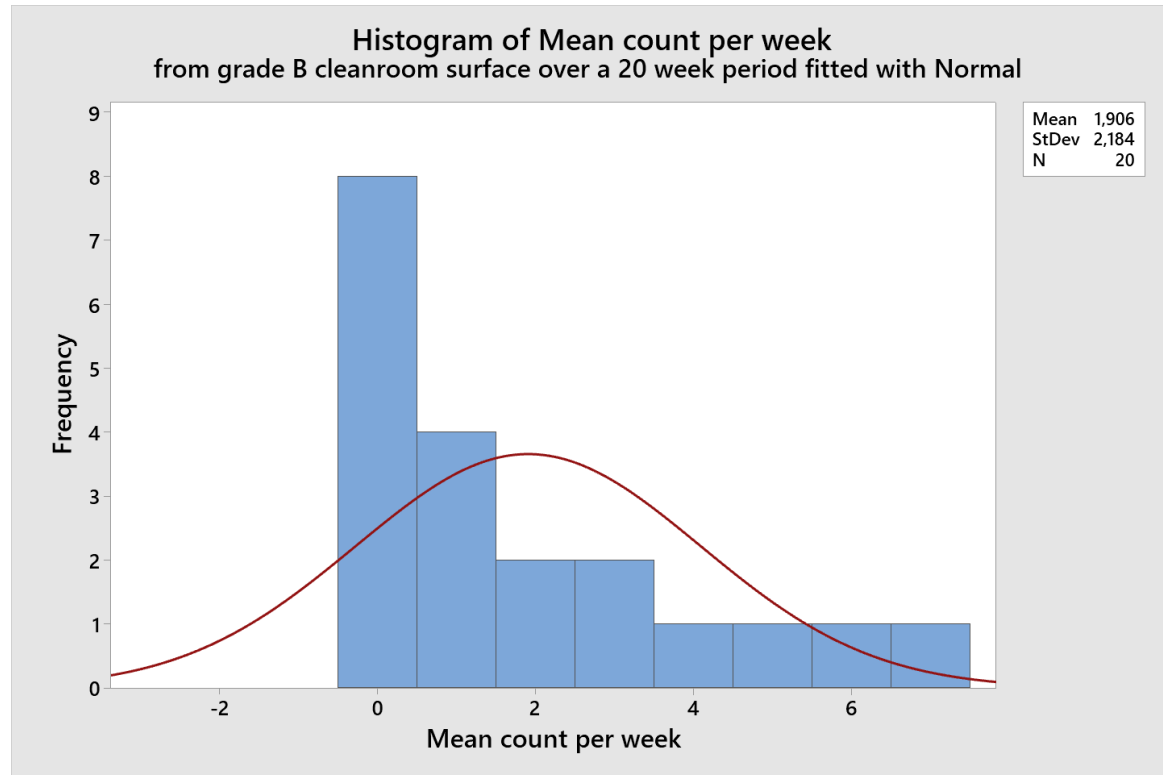
INFERENCEAL STATISTICS

Let's consider, for instance, the microbial distribution curve shown in the table on the side.

Week No.	Mean count <i>per week</i>
1	0.00
2	5.15
3	0.29
4	6.93
5	1.86
6	1.47
7	0.10
8	0.00
9	2.22
10	3.95
11	0.11
12	1.25
13	0.00
14	6.34
15	0.31
16	0.45
17	2.70
18	0.89
19	0.65
20	3.45

T. Sandle, Data Review and Analysis for Pharmaceutical Microbiology – Microbiology Solutions, 1st Ed., (Jan. 2014)

INFERENCEAL STATISTICS



INFERENCEAL STATISTICS

Back to the basic concepts of Inferential Statistics essential for taking a decision (*i.e.*, to reject H_0 or fail to reject H_0) there is that of *Level of Confidence* (indicated with C and typically: 95% or 99% or 0.95 and 0.99) *which tells us how sure we are that we have made the right decision or choice.*

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

Practically, *Level of Confidence* and *Level of Significance* give us a measure of the same thing:

how sure we are that we are making the right decision or not !

INFERENCEAL STATISTICS

Two types of errors can be made when testing hypothesis:

- **Type I error (or risk α)** : the null hypothesis is rejected when it is true
FALSE POSITIVE
- **Type II error (or risk β)** : the null hypothesis is not rejected when it is false
FALSE NEGATIVE

INFERENTIAL STATISTICS



INFERENCEAL STATISTICS

In summary:

	We Reject H_0 . (accept H_a)	We Fail to Reject H_0 (not enough evidence to accept H_a)
H_0 is true.	Type I Error	Correct Decision
H_0 is false. (H_a is true)	Correct Decision	Type II Error

What just seen represents the basis of the so-called: **STATISTIC RISK ANALYSIS**

INFERENCE STATISTICS

In Quality Control



- **risk α** , or just α , is called **PRODUCER'S RISK** because it *denotes the probability that a good lot will be rejected*, or the probability that a process producing acceptable values of a particular quality characteristic will be rejected as performing unsatisfactorily.
- **risk β** , or just β , is called **CONSUMER'S RISK** because it *denotes the probability of accepting a lot of poor quality* or allowing a process that is operating in an unsatisfactory manner relative to some quality characteristic to continue in operation.

INFERENCEAL STATISTICS

- *risks α and β risks are related to each other !*
- *risk α is generally regarded as the worst!*
- *P-value is the probability of making a “type α error”*
- *α is the highest value of p we are willing to tolerate and still say that a difference is « statistically significant »*
- *if $P\text{-value} \leq \alpha$ the observed difference is said to be « statistically significant »*
- *If $P\text{-value} > \alpha$ the observed difference is said to be « not significant »*

INFERENTIAL STATISTICS

In Hypothesis Testing if:

- *if $P\text{-value} \leq \alpha$*  *Reject the Null Hypothesis*
- *if $P\text{-value} > \alpha$*  *Fail to Reject (Accept) the Null Hypothesis*

INFERENCEAL STATISTICS

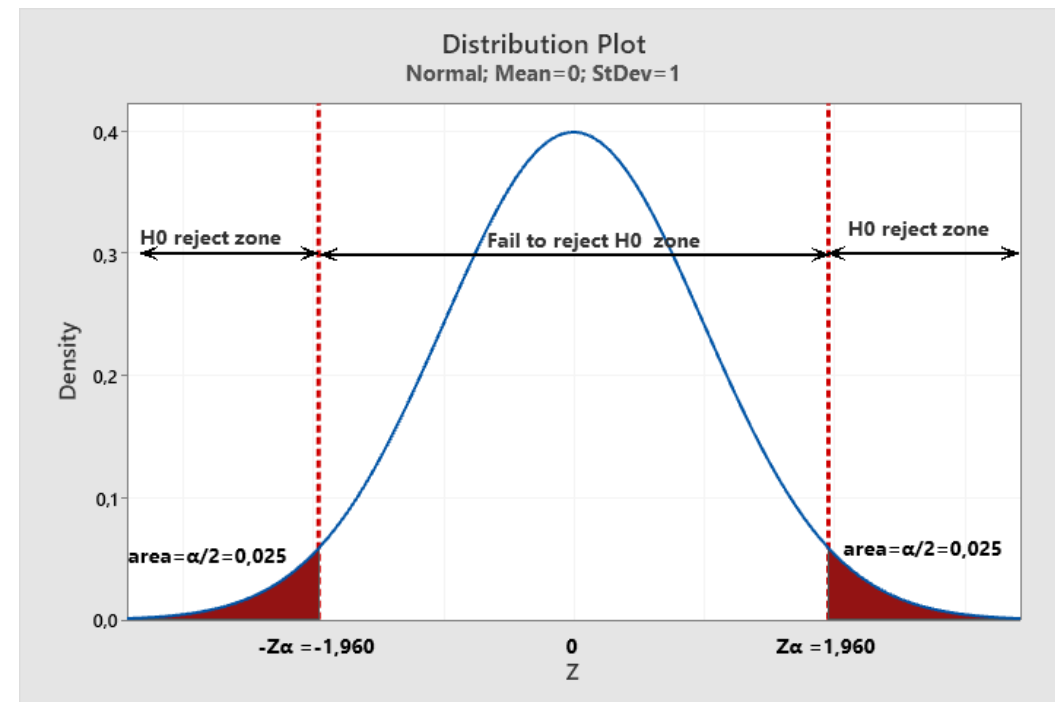
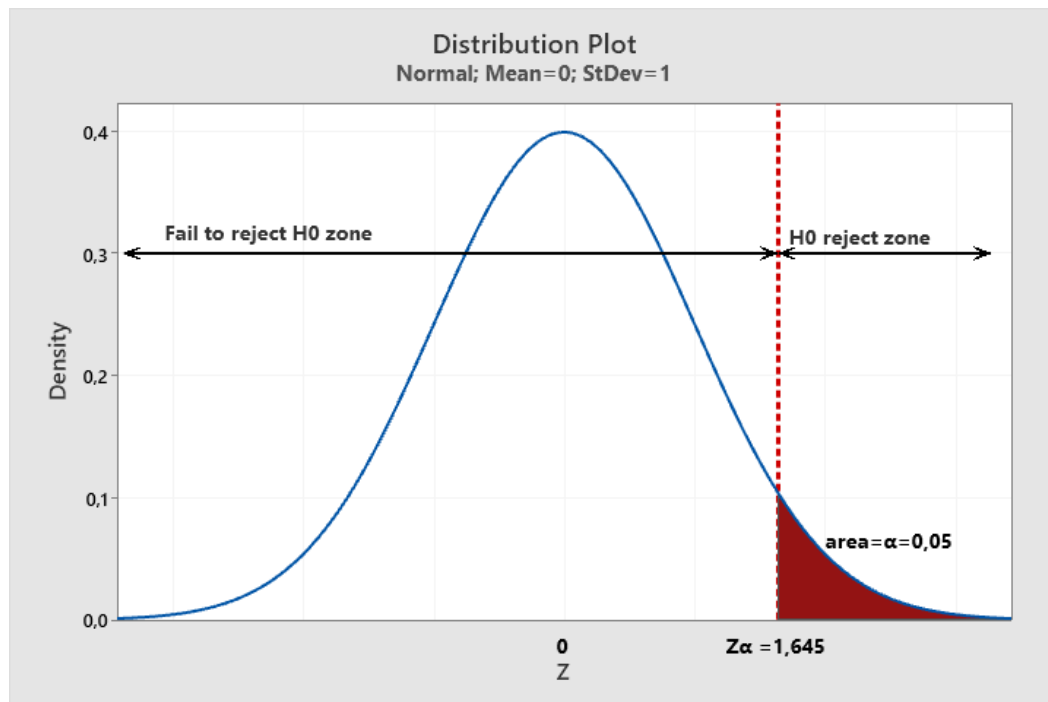
TO MAKE IT SIMPLE:

Since a probability of making a mistake of less than or equal to 5% ($\alpha = 0.05$) is *normally accepted*, in general:

- ***P-value* > 0.05** \Rightarrow the differences between samples are not statistically significant:
the Null hypothesis fails to be rejected
- ***P-value* \leq 0.05** \Rightarrow the differences between samples are statistically significant:
the Null hypothesis can be rejected

INFERENCE STATISTICS

Practically, the *level of significance*, α , is an area defined by a Z_α value that represent the corresponding *test statistic* value printout in standard tables.



INFERENCE STATISTICS

Once again: but what is the practical usefulness of all this?

Let's go back to our QA Officer who was convinced that the yield of the process was no longer 100 Kg / lot

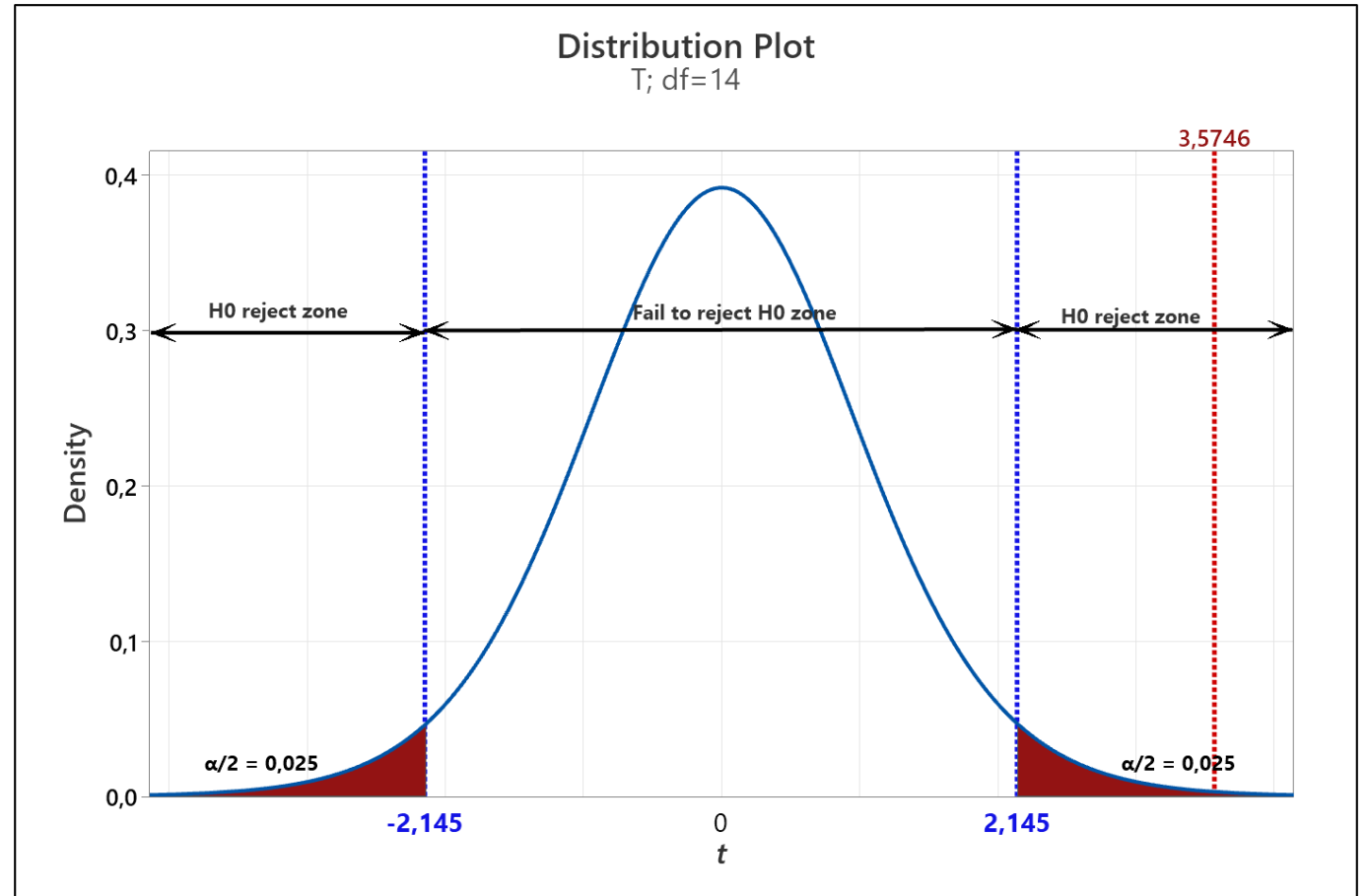
INFERENCEAL STATISTICS

He considers the last 15 batches manufactured after the intervention and they show an average yield of $\bar{x} = 101.2$ Kg, and a standard deviation of $s = 1.3$ Kg. He tests his claim at the 0.05 significance level (or 95% confidence level).

He calculates the **test statistics t** as follows

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{101.2 - 100}{1.3 / \sqrt{15}} = 3.5746$$

Since the value falls within the reject zone there is evidence to reject the null hypothesis at $\alpha = 0.05$.
In other words: the QA Officer was right !



INFERENCEAL STATISTICS

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

$$\begin{array}{l} H_0: \mu = 100 \text{ kg (Null hypothesis)} \\ H_1: \mu \neq 100 \text{ kg (Alternative hypothesis)} \end{array} \left. \vphantom{\begin{array}{l} H_0 \\ H_1 \end{array}} \right\} \text{two tails test}$$

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:

$$\begin{array}{l} H_0: \mu \leq 100 \text{ kg (Null hypothesis)} \\ H_1: \mu > 100 \text{ kg (Alternative hypothesis)} \end{array} \left. \vphantom{\begin{array}{l} H_0 \\ H_1 \end{array}} \right\} \text{one (right) tail test}$$

and what would hypothesis testing be like?

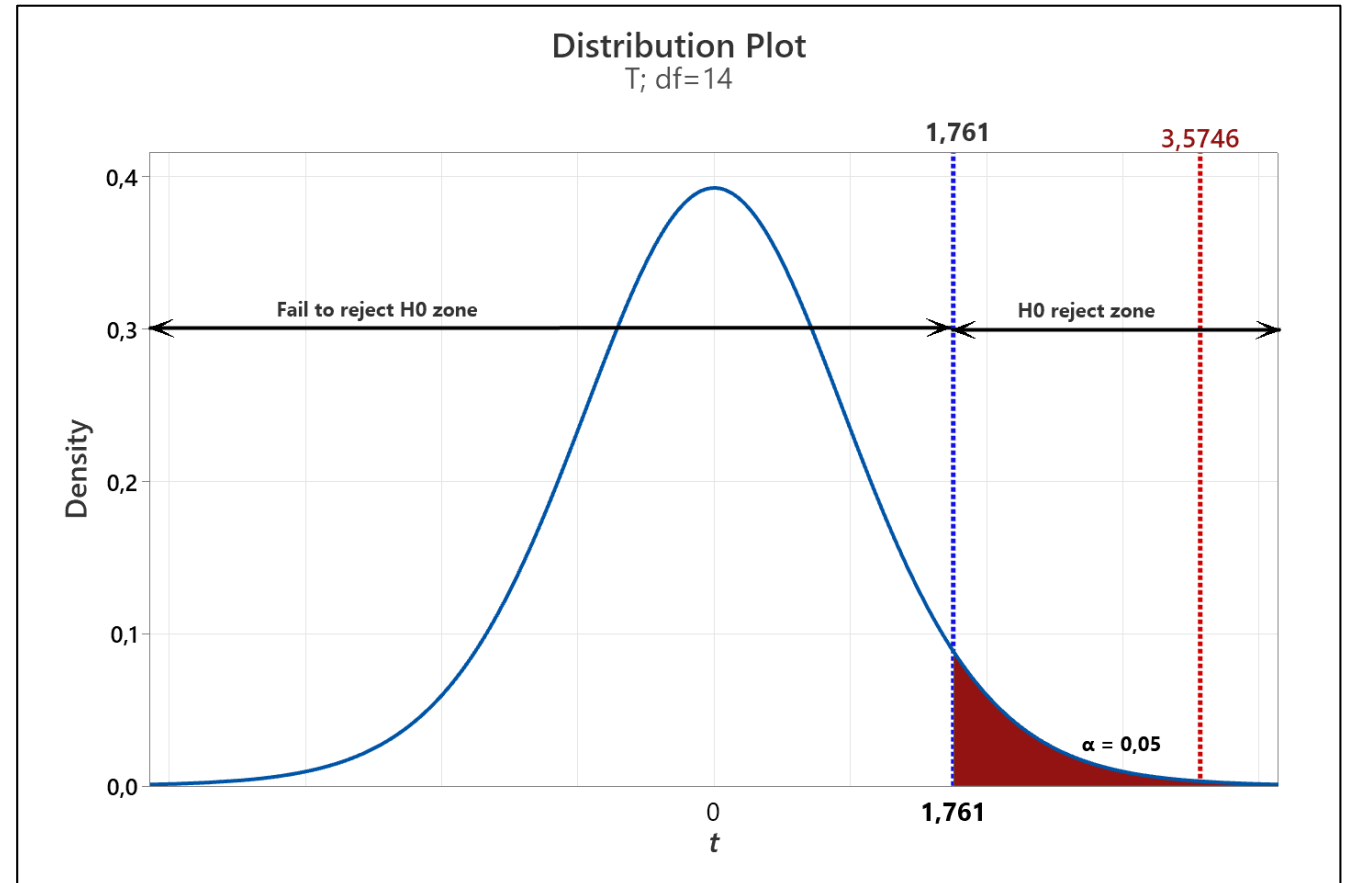
INFERENTIAL STATISTICS

Once again, the test statistic would be calculated as before, *i.e.*:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{101.2 - 100}{1.3 / \sqrt{15}} = 3.5746$$

This time, however, the test would have been all "one side only".

Even in this case, the QA Officer would have been right at $\alpha = 0.05$!



INFERENCEAL STATISTICS

STATISTICAL HYPOTHESIS TESTING is useful in many cases:

- *check if a certain value lies within the confidence interval*
(typical application: determining if a result is an OOS)
- *compare two datasets to see if they are really different or belong to the same population* (typical applications of this are in: suppliers' validation, comparison of analytical data generated by different methods, *etc.*)
- *check the strength of the correlation between one or more causes and the undesirable effect*
- *etc.*

INFERENCEAL STATISTICS

Let see two other practical examples !

INFERENCEAL STATISTICS

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

Tablets weights (mg)				
47.9842	50.4625	48.9013	53.4198	47.0006
51.8503	50.9037	53.7210	46.0764	53.1639
48.5344	53.1428	51.1559	49.4118	52,6852
49.6923	57.3226	49.9143	51.2395	48.1680

It is known that the process, in conditions of normal operation, produces tablets whose average weight is 50.36 mg and standard deviation 2.235 mg.

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

$H_0: \mu = 50.36 \text{ mg}$ vs. $H_1: \mu \neq 50.36 \text{ mg}$ at a significance level of 5% ($\alpha = 0.05$)

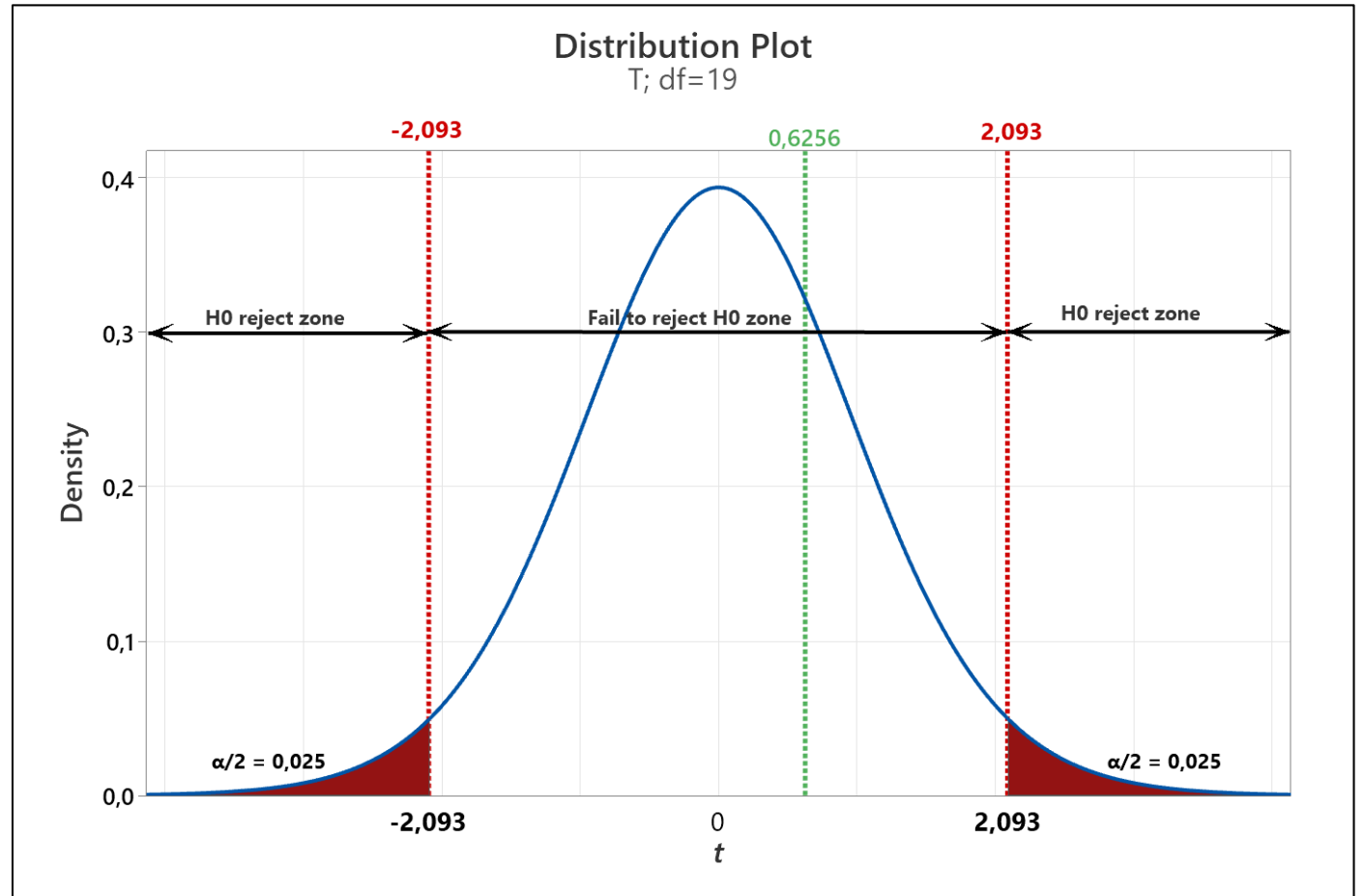
INFERENCEAL STATISTICS

- Sample mean is $\bar{x} = 50.7375$ mg and $s = 2.6982$
- Being the sample size < 30 it must be used the **Student Distribution** and the **test statistics t** can be calculated as follows :

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{50.7375 - 50.36}{2.6982 / \sqrt{20}} = 0.6256$$

- $t_{1-\alpha/2} = t_{0.975} = 2.093$ and $-t_{1-\alpha/2} = -t_{0.975} = -2.093$

there is no experimental evidence to reject the null hypothesis and, therefore, to exclude that the process is « under control » at $\alpha = 0.05$ (i.e., Level of Confidence = 95%).



INFERENCEAL STATISTICS

The approach just seen can also be used in reverse as, for example, in this case:

The tablets obtained from a given process are rejected if they weigh less than 95 mg or more than 108 mg.

100 are checked and there are: 3 tablets < 95 mg and 5 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 the Gaussian model for the weight of the tablets, as also logical in the absence of specific perturbations, then....

INFERENCEAL STATISTICS

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



$$\begin{cases} \Phi\left(\frac{95 - \mu}{\sigma}\right) = 0.03 \\ 1 - \Phi\left(\frac{108 - \mu}{\sigma}\right) = 0.02 \end{cases}$$

from which it follows that:

$$\begin{cases} 95 - \mu = \sigma Z_{0.03} \\ 108 - \mu = \sigma Z_{0.98} \end{cases}$$



$$\begin{cases} 95 - \mu = \sigma (-1.88) \\ 108 - \mu = \sigma (2.05) \end{cases}$$



$$\mu = 101.22 \text{ mg}$$

$$\sigma = 3.31 \text{ mg}$$

INFERENCEAL STATISTICS

This example was intended to show how, using simple notions of Inferential Statistics and:

- taking random samples from a production line

or

- analyzing « processing waste »

it is possible to « infer » from experimental data crucial information on the state of the process.

INFERENCEAL STATISTICS

***1-Sample t test. 2-Sample t test and
2-Variances test***

INFERENCEAL STATISTICS

Hypothesis tests, such as that seen in practice applied to the case of tablets, allow you to determine, starting from sample data, 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*

INFERENCEAL STATISTICS

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)

INFERENCEAL 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

INFERENCEAL STATISTICS

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

INFERENCEAL STATISTICS

Let's see a few practical examples

INFERENCEAL 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	8.2

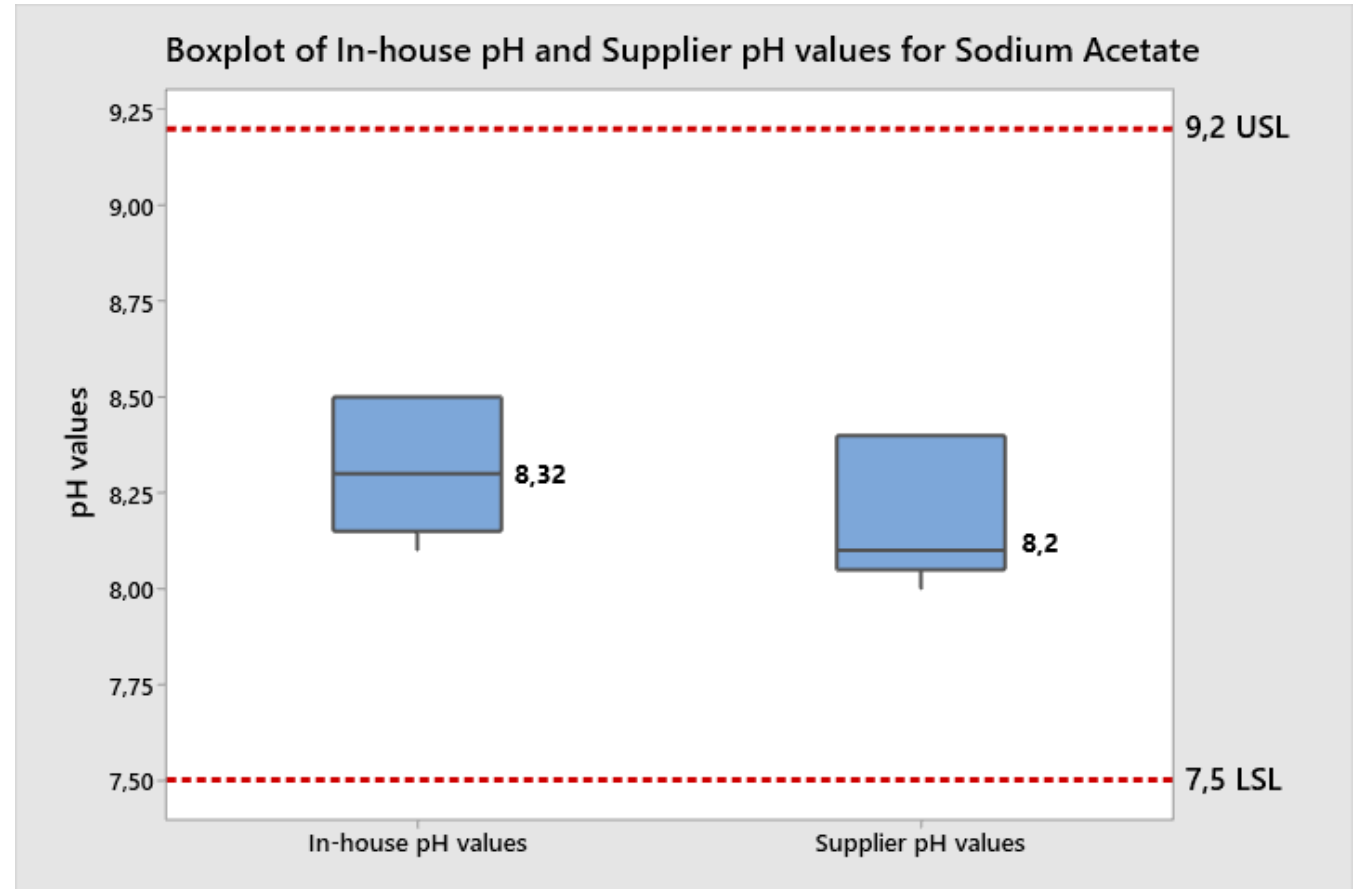
On the average are the two series of data here above reported, statistically different or not?

INFERENCEAL STATISTICS

Let's first look at data visualization using boxplots.

Box widths look rather similar, but, apart from this, we cannot say much more.

*The **t-test** can tell us if the two mean values are statistically different or not.*



INFERENCEAL STATISTICS

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
In-house pH values	5	8,320	0,179	0,080
Supplier pH values	5	8,200	0,187	0,084

Estimation for Difference

Difference	95% CI for Difference
0,120	(-0,154; 0,394)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value DF P-Value

1,04 7 0,334  As *P-value* > 0.05, we fail to reject H_0  No difference !

INFERENCEAL STATISTICS

The fact that there is no statistically significant difference between the average values of the two data groups suggests that, reasonably, there is no difference between the two methods of determining pH.

Instead, consider the data in the table here on the side. In this case, Sodium Acetate is provided by a different supplier.

	Sodium Acetate pH values	
	In-house	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	8.7

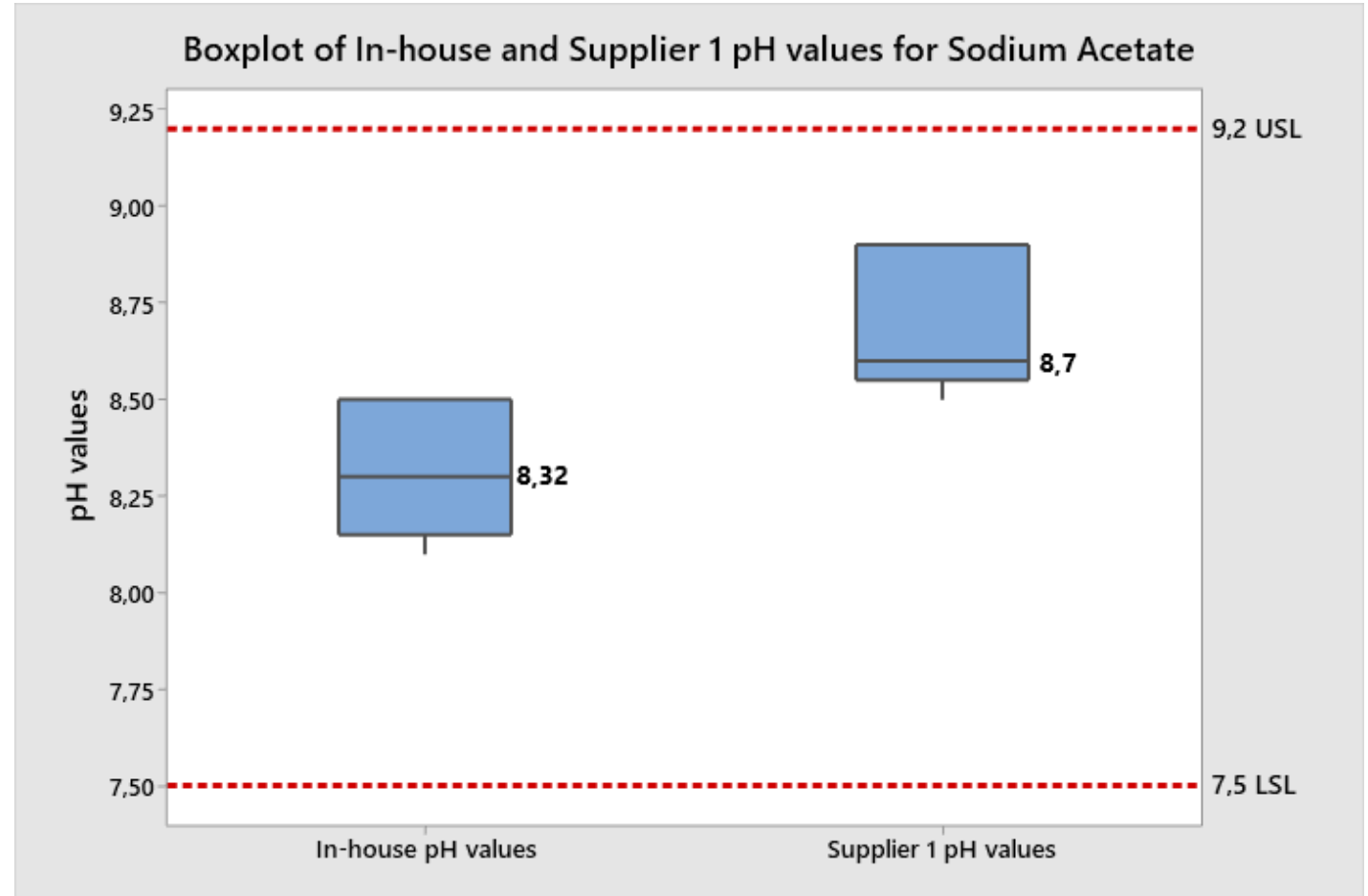
Are the two mean values here above reported, statistically different or not?

INFERENCEAL STATISTICS

In this case it is evident that the two pH data distributions are shifted from each other.

However, box widths are comparable \Rightarrow data spreads are similar.

The t-test can tell us if the two mean values are statistically different or not while the F-test can tell us if data spreads are really comparable or not.



INFERENCEAL STATISTICS

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
in-house pH values	5	8,320	0,179	0,080
supplier 1 pH values	5	8,700	0,187	0,084

Estimation for Difference

Difference	95% CI for Difference
-0,380	(-0,654; -0,106)

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-3,28	7	0,013



As *P-value* < 0.05, there is evidence to reject H_0



There is a difference !

INFERENCEAL STATISTICS

The fact that there is a statistically significant difference between the average values of the two data groups suggests that, reasonably, there is difference between the two methods of determining pH.

This finding is not so unusual if comparing data from different laboratories !

In such a case, even if the analytical techniques are different from each other, they should be of comparable precision and accuracy and therefore



2-Variances test :

Determine whether the variances or standard deviations of two groups differ. You can use this test to compare the process variance before and after you implement a quality improvement program.

INFERENCEAL STATISTICS

Descriptive Statistics

Variable	N	StDev	Variance	95% CI for σ	
In-house pH values	5	0,179	0,032	(0,102; 0,518)	
Supplier 1 pH values	5	0,187	0,035	(0,113; 0,510)	

Test

Null hypothesis	$H_0: \sigma_1 / \sigma_2 = 1$
Alternative hypothesis	$H_1: \sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0,05$

Method Test Statistic		DF1	DF2	P-Value
Bonett	0,02	1		0,896
Levene	0,00	1	8	1,000



As P-value > 0.05 there is no evidence to reject H_0 !

INFERENCEAL STATISTICS

STATISTICAL INTERVALS

INFERENCEAL STATISTICS

- *Different practical problems call for different types of intervals !*
- There are three main types of statistical intervals that may be calculated from sample data:
 - *Confidence intervals*
 - *Prediction Intervals*
 - *Tolerance Intervals*

G.J. Hahn, W.Q. Meeker, Statistical Intervals: A Guide for Practitioners – J. Wiley & Sons (1991)

INFERENCEAL STATISTICS

■ Confidence Interval : concept

- a range that contains, with know probability, the true value of a population parameter (*e.g.*, the mean μ or the standard deviation σ)
- they quantify our knowledge, or lack thereof, about a parameter or some other characteristic of a population, based upon a random sample.

INFERENCEAL STATISTICS

■ Confidence Interval : example

Let's consider the pH values measured on five different lots of Sodium Acetate provided by a chemical manufacturer: 8.1, 8.3, 8.2, 8.5, 8.5.

Sample mean (\bar{x}) and standard deviation (s) are $\bar{x} = 8.32$ and $s = 0.18$ pH units.

A *two-sided 95% confidence interval* for the mean μ of the population of sampled Sodium Acetate lots is:

$$[\mu^l, \mu^u] = 8.32 \pm 1.24 (0.18) = [8.10, 8.54]$$

This means that we are 95% confident that the interval 8.10 – 8.54 pH units contains the unknown pH mean value (μ) of the population of Sodium Acetate lots provided by the chemical manufacturer.

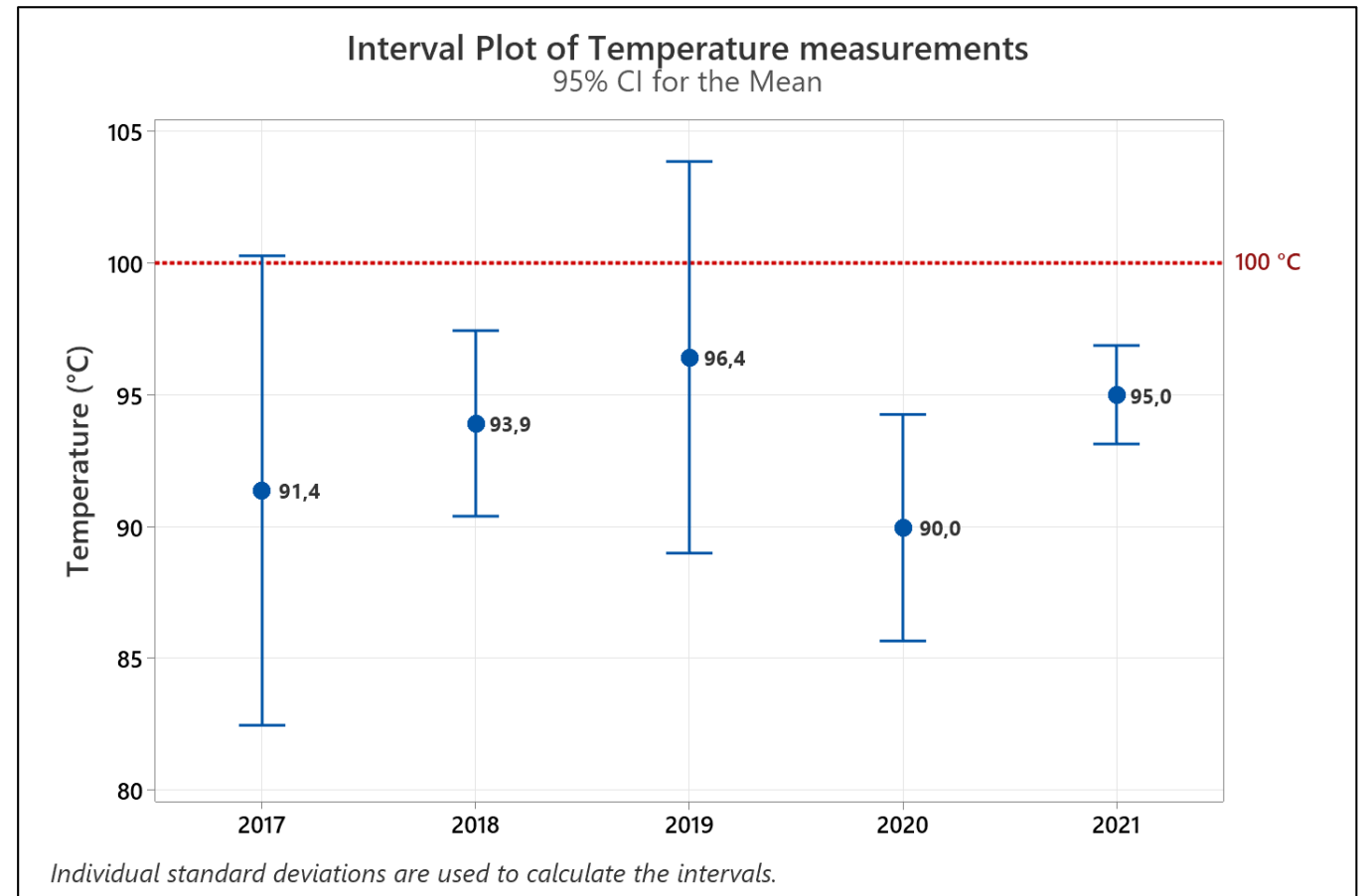
INFERENCEAL STATISTICS

■ Confidence Interval: example

Let's consider 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			



INFERENCE STATISTICS

ATTENTION

- 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 !

INFERENCEAL STATISTICS

■ Prediction Interval : concept

- a range that contains, with a specified degree of confidence, one or more future observations randomly selected from a population.
- interests a manufacturer (or user) who wishes to predict the performance of one or more future units.
- due to its "predictive" nature this type of interval is wider than the confidence interval.

INFERENCEAL STATISTICS

■ Prediction Interval : example

Let's consider the five pH values of five different lots of Sodium Acetate seen earlier.

A *two-sided 95% prediction interval* to contain the pH values of all of 10 additional Sodium Acetate lots randomly sampled from the same population is:

$$[y_{10}^l, y_{10}^u] = 8.32 \pm 5.23 (0.18) = [7.38, 9.26]$$

This means that we are 95% confident that the pH values of all 10 additional lots of Sodium Acetate manufactured by the chemical manufacturer will be contained within the interval 7.38 – 9.26 pH units.

INFERENCEAL STATISTICS

■ Tolerance Interval : concept

- a range expected to contain, with a specified degree of confidence, at least a specified proportion of the units from the sampled population
- within Six Sigma: tolerance limits reflect *customer requirements* and should be established before a product is designed
- this interval would therefore be of particular interest in setting limits on *process capability*
- tolerance intervals reflect the variation produced by a particular part, process and design

INFERENCEAL STATISTICS

■ Tolerance Interval : concept

- the concept of statistical **tolerance interval** it can be seen as an **extension of** that of **prediction interval** in case one wishes to draw conclusions about the performance of a relatively large number of future units (*e.g.*, 100, 1000, or any number m), based upon the data from a random sample from the population of interest.

INFERENCEAL STATISTICS

■ Tolerance Interval : concept

- Assume, for instance, that measurements of tablets weights have been obtained on a random sample of 20 units taken out from a production process.
- A tolerance interval calculated for such data provides limits that one can claim, with a specified degree of confidence (*e.g.*, 95%), contains the (measured) weights of at least a specified proportion (*e.g.*, 90%) of units from the sampled population.

The two percentages are well distinct: one (*i.e.*, 90%) refers to the percentage of the population while the other (*i.e.*, 95%) deals with the degree of confidence associated with the claim (*i.e.*, that the interval encloses at least 90% of the population).

INFERENCEAL STATISTICS

■ Tolerance Interval : example

Let's consider the five pH values of five different lots of Sodium Acetate seen earlier.

A *two-sided 95% tolerance interval* to contain at least 99% of the sampled population of Sodium Acetate is:

$$[T_{0.99}^l, T_{0.99}^u] = 8.32 \pm 6.60 (0.18) = [7.13, 9.51]$$

This means that we are 95% confident that the interval 7.13 – 9.51 pH units contains at least 99% of the population of Sodium Acetate lots provided by the chemical manufacturer.

INFERENCEAL STATISTICS

The use of Tolerance Intervals may be extremely important in OOS management !

- Consider the case of a CU test in which 10 tablets are tested, of which 9 are perfectly in specification with values between 99.5% and 101.5% of the label claim while one is 70%.
- Since all tablets have been destroyed in the analytical process and there is no way to establish if the anomalous result was due to the specific tablet or to some analytical error

What do you do?

Reject the batch although there is no historical or specific evidence of any issue ?

INFERENCEAL STATISTICS

Obviously no !

It can be performed an “extensive” CU testing based on the Tolerance Interval concept to get, for instance, 99% confidence that 99% of the tablets are between 99.5% and 101.5% of the label claim.

In this way the size of the new test will be at least established on a scientific and non-questionable basis !

CONTROL CHARTS

CONTROL CHARTS

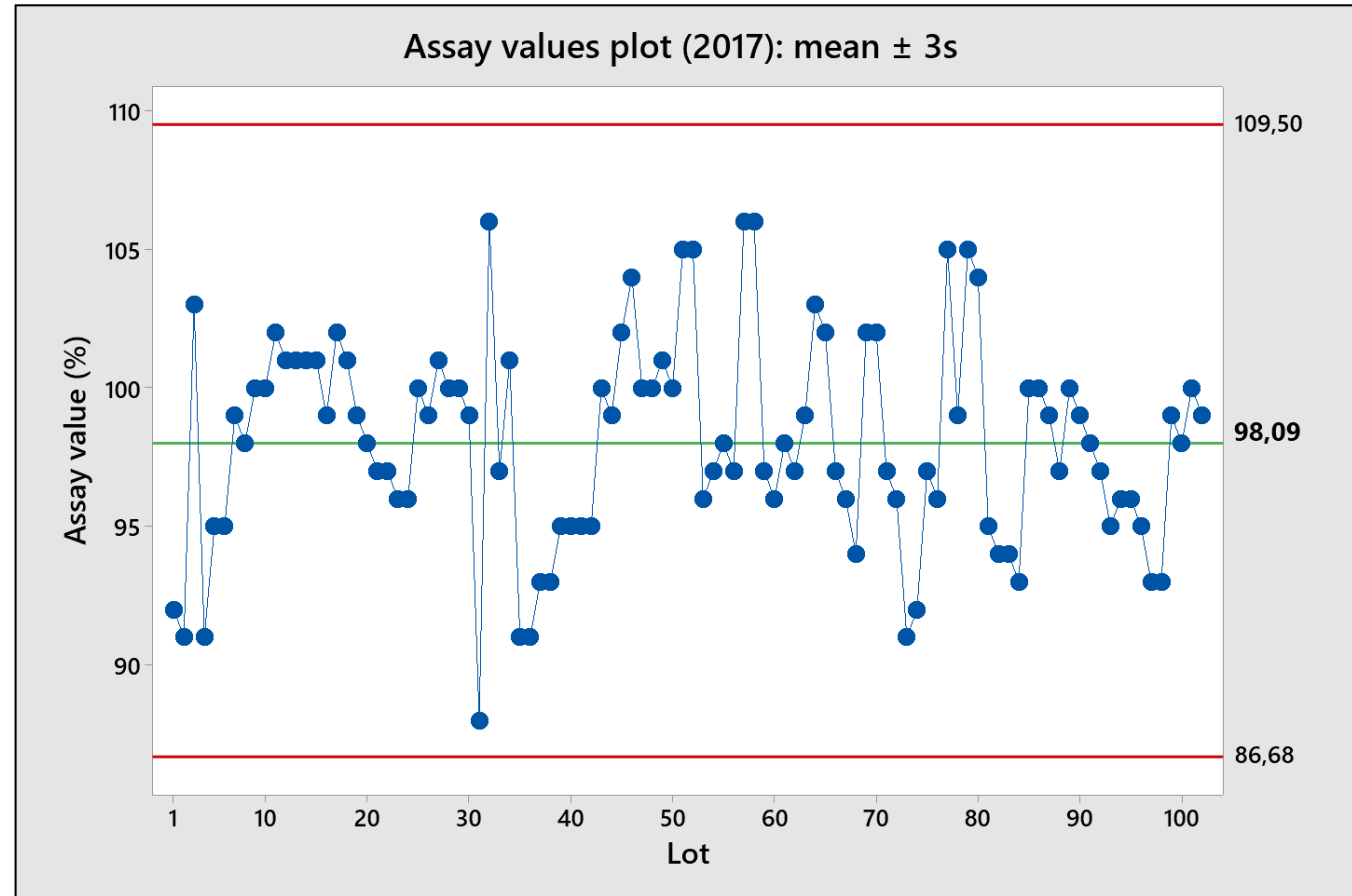
Before starting, let's immediately dispel a myth.

A conventional plot « average $\pm 3s$ » such as that shown here on the side

IS NOT A CONTROL CHART !

Why?

simply because it doesn't control anything especially if used as it is often done !



CONTROL CHARTS

The interval defined by « average $\pm 3s$ » brackets virtually all of the process outcomes regardless of whether or not the process is operated predictably !

CONTROL CHARTS

*Always remember what we said at the beginning:
the average is not a robust index because it is sensitive to
outliers, whether they are too small or large!*

CONTROL CHARTS

■ Shewhart' Control Charts

« The purpose of Shewhart's Control Charts is to detect a lack of control when it exists, and it should be able to do so, at least most of the time, even when the out-of-control data are used to compute the limits. Otherwise, the technique would not be of much use »

D.J. Wheeler, D.S. Chambers, Understanding Statistical process Control, 2nd Ed., SPC Press (1992)

CONTROL CHARTS

■ Shewhart' Control Charts

« Notice that there is no requirement of normality (or even approximate normality) ...

Control charts work well even if the data are not normally distributed. This issue was addressed by Shewhart in his first book, and it should never have been an issue »

D.J. Wheeler, D.S. Chambers, Understanding Statistical process Control, 2nd Ed., SPC Press (1992)

W.A. Shewhart, Economic Control of manufactured product, Van Nostrand (1931)

CONTROL CHARTS

■ Shewhart' Control Charts

Shewhart's Control Charts consist of the following main elements:

- A **centerline** (average or median) which represents the average (or median) of the values plotted in that panel
- **UCL** and **LCL** (upper and lower control limits) are the limits that a stable process is very unlikely to cross. **These limits are not specification limits.**
- A **plotted parameter** (*i.e.*, individual values, means, ranges, *etc.*)

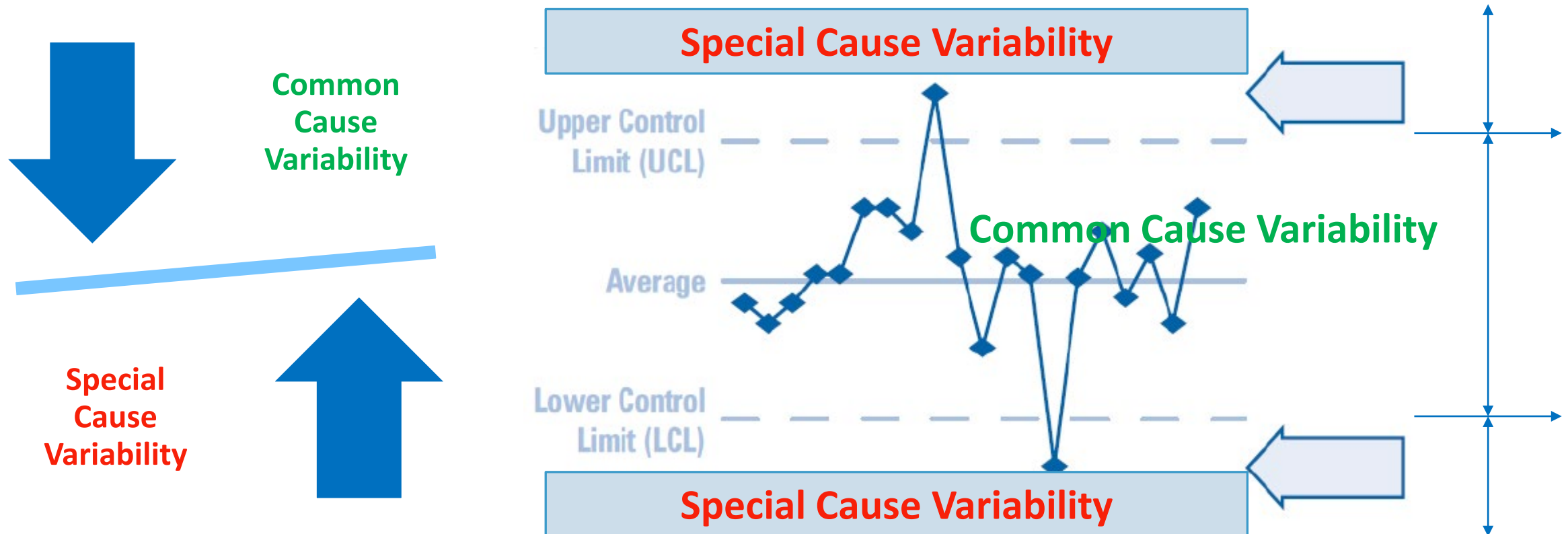
CONTROL CHARTS

■ Shewhart' Control Charts

- There is a close connection between control charts and hypothesis testing.
- The control chart is a test of the hypothesis that the process is in a state of statistical control:
$$H_0 : \text{process mean} = \mu_0$$
$$H_1 : \text{process mean} \neq \mu_0$$
- Control charts are used to detect departures from an assumed state of statistical control.

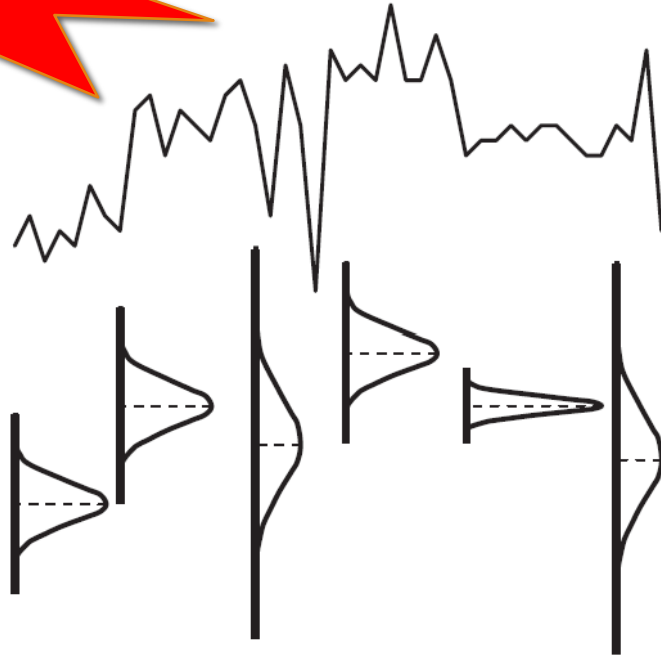
CONTROL CHARTS

The Shewhart Concept of Variation

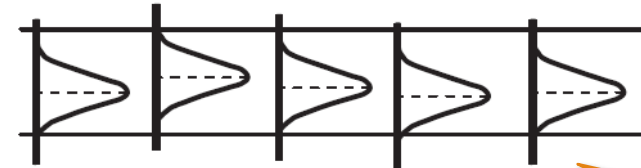
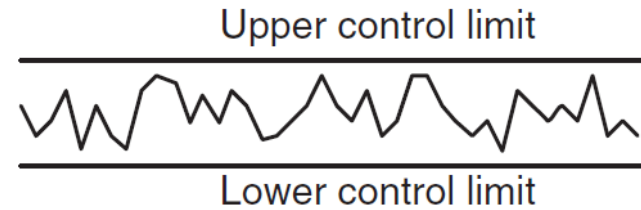


CONTROL CHARTS

Out of Control



Unstable process




Stable process

In Control

A process in (statistical) control is a predictable process !

CONTROL CHARTS

- the *causes that contribute to the variability* of a production process are essentially of two types: *common causes* and *special (W.E. Deming) or assignable (W.A. Shewhart) causes*.
- a process is said to be *under statistical control* when *its variability is due only to common causes*.
- « ... a phenomenon will be said to be controlled when, through the use of past experience, we can predict, at least within limits, how the phenomenon may be expected to vary in the future. »  *predictability*

W.A. Shewhart, *Economic Control of Quality of Manufactured Product*, Van Nostrand (1931) p. 6

CONTROL CHARTS

« ... we must also accept as axiomatic that **a controlled quality will not be a constant quality**. Instead, **a controlled quality must be a variable quality**. This is the first characteristics.»

W.A. Shewhart, Economic Control of Quality of Manufactured Product, Van Nostrand (1931) p. 6

« **Stability**, or the existence of a system, **is seldom a natural state**. **It is an achievement, the result of eliminating special causes one by one on statistical signal, leaving only the random variation of a stable process.** »

W.E. Deming, Out of the Crisis, MIT Press (2000) p. 322

CONTROL CHARTS

Moreover,

« *Statistical control does not imply absence of defective items. Statistical control is a state of random variation, stable in the sense that the limits of variation are predictable. »*

« *Statistical control of a process is not an end in itself. »*

W.E. Deming, Out of the Crisis, MIT Press (2000) p. 354

CONTROL CHARTS

SUMMARIZING

- While perfect stability is an unreachable goal, a relative stability is certainly attainable
- Control Charts are used to monitor the process and to be sure that it remain stable

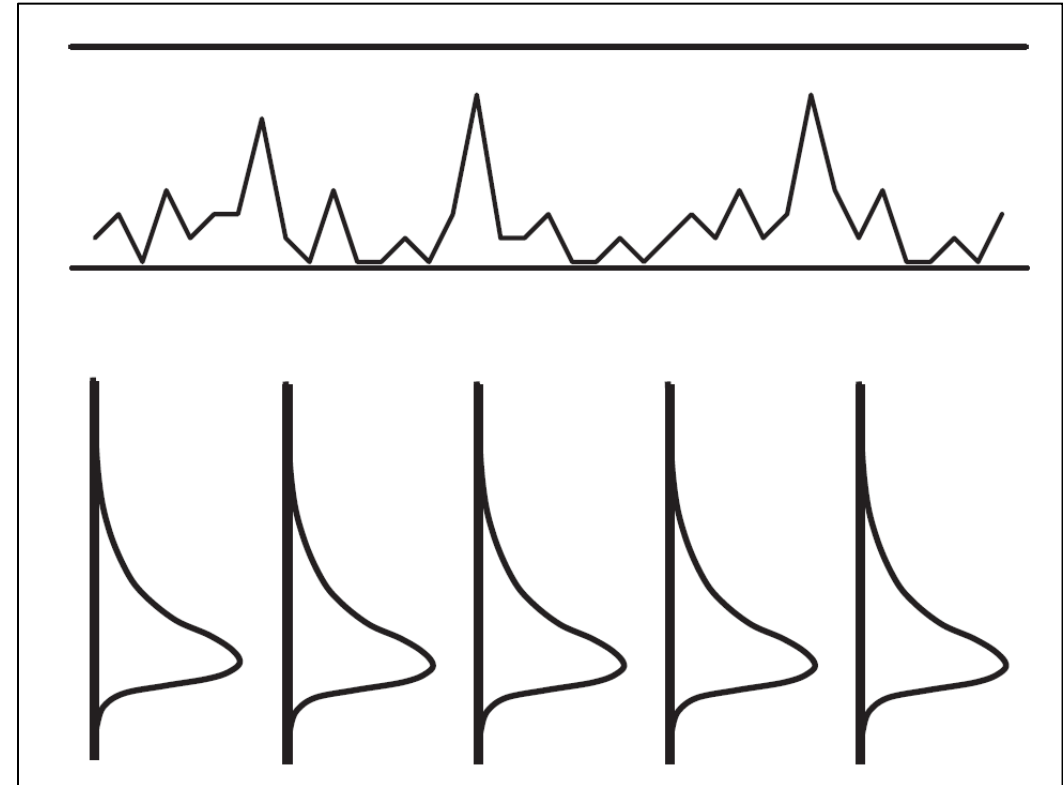
CONTROL CHARTS

REMEMBER !

A stable process does not have to be normally distributed !

There are many processes which naturally produce skewed distributions !

*An example for all:
Related Substances content*



CONTROL CHARTS

Let's start with the

Control Charts by Variables

*but, first of all, let's make a premise: all charts are
"double"*

CONTROL CHARTS

Why do we need two charts?

❖ The top chart (individual or bar-type)

- shows changes in the individual or average values of the process
- visualize long-term variability

❖ The bottom chart (associated to variability)

- shows short-term variability
- contains the elements for calculating the control limits in the upper chart

A process to be "in control", the datapoints must lie within the control limits in both charts.


CONTROL CHARTS

Control Charts by Variables

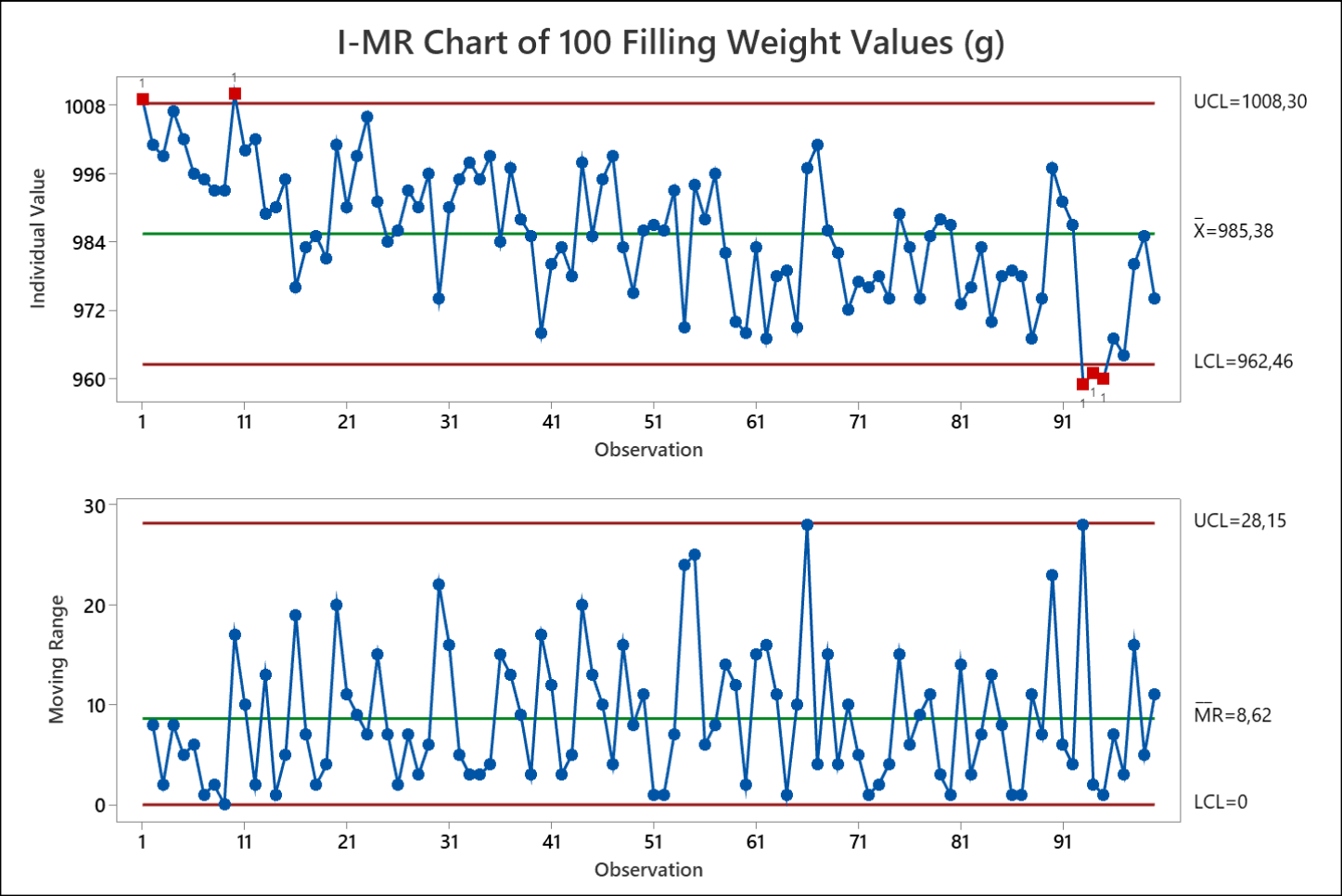
Control Chart Type	Control Chart Name	Central Line	Upper and Lower Control Limits
$\bar{X}-R$	Average and Range Chart	$\bar{\bar{X}} \quad \bar{R}$	$\bar{\bar{X}} \pm A_2\bar{R}$ $D_4\bar{R}, D_3\bar{R}$
$\bar{X}-s$	Average and Standard Deviation Chart	$\bar{\bar{X}} \quad \bar{s}$	$\bar{\bar{X}} \pm A_2\bar{R}$ $B_4\bar{s}, B_3\bar{s}$
$\overline{Me}-R$	Average of Medians and Range Chart	$\bar{M}_e \quad \bar{R}$	$\bar{M}_e \pm A_4\bar{R}$ $D_4\bar{R}, D_3\bar{R}$
$I-MR$	Individual Value and Moving Range Chart	$\bar{x} \quad R_m$	$\bar{x} \pm 2.660\bar{R}_m$ $3.267\bar{R}_m$

CONTROL CHARTS

■ *I-MR* Control Charts

- I-MR charts are generally used when it is difficult or impossible to measure in subgroups. This occurs when measurements are expensive or destructive, low production volumes of products or products have a very long or continuous cycle time.
- Typical applications of *I-MR* Charts are for limited series of individual measurements like *assay values, yields, etc.*  ***Annual Product Quality Reviews (APQR)***

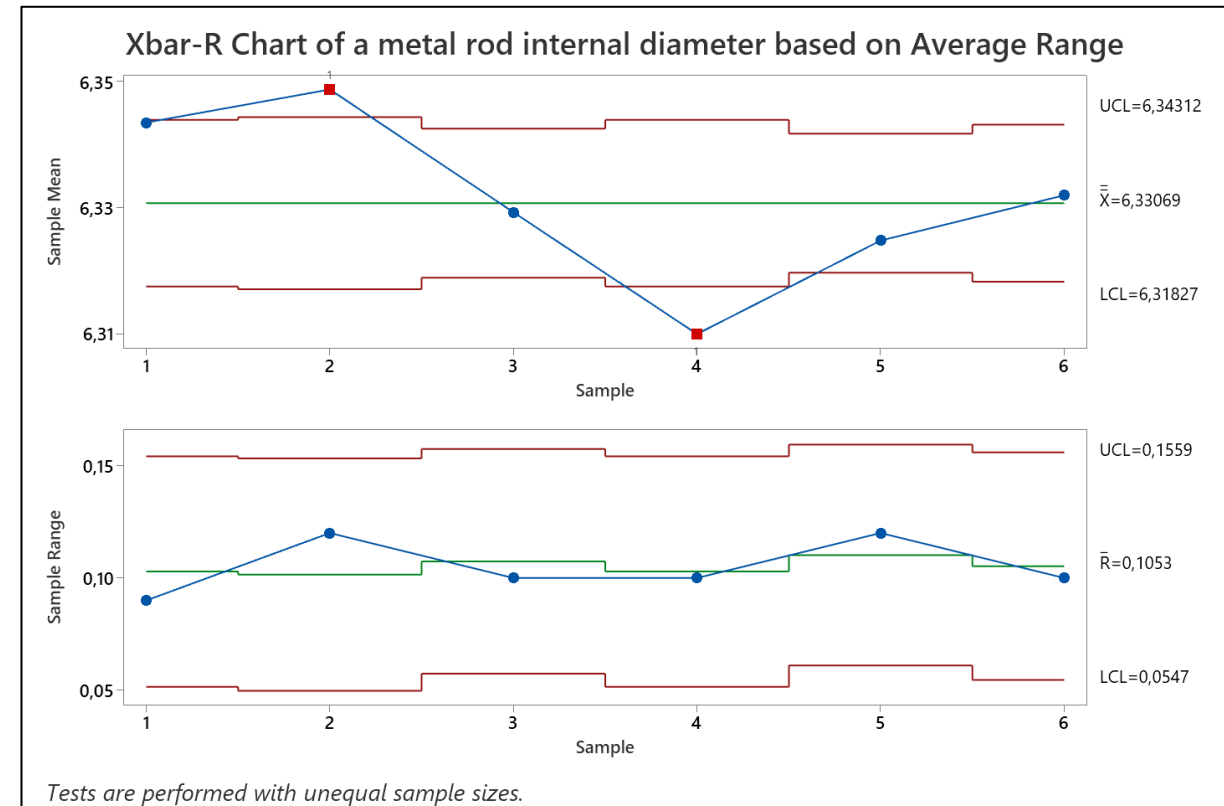
CONTROL CHARTS



CONTROL CHARTS

■ *Xbar-R Charts*

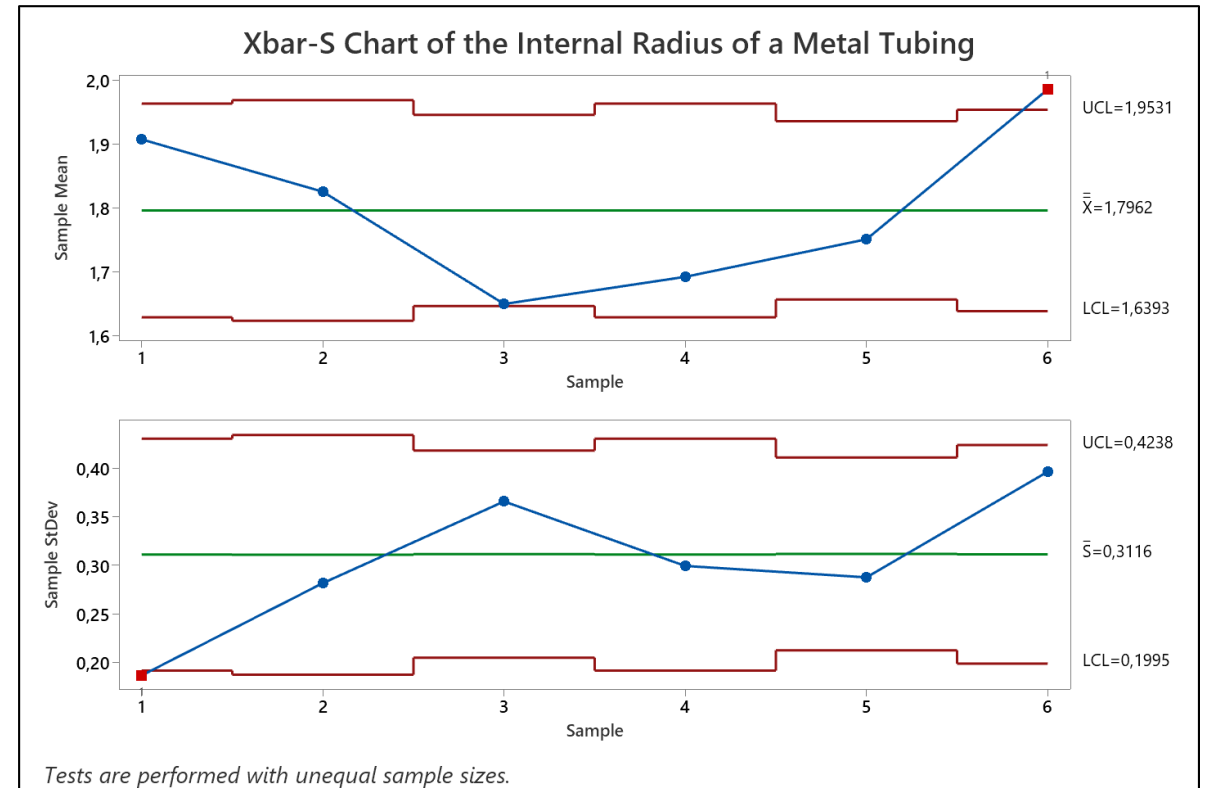
- The range of results within each data set is used to estimate overall variability.
- ***Xbar-R Charts*** are generally used to monitor the mean and variation of a process when you have continuous data and subgroup sizes of 8 or less.
- For subgroups that contains 9 or more values, it is in general recommended of using ***Xbar-S Charts***.



CONTROL CHARTS

■ *Xbar-S Charts*

- The standard deviation of results within each data set is used to estimate overall variability.
- *Xbar-S Charts* are generally used to monitor the mean and variation of a process when you have continuous data and subgroup sizes of 9 or more.



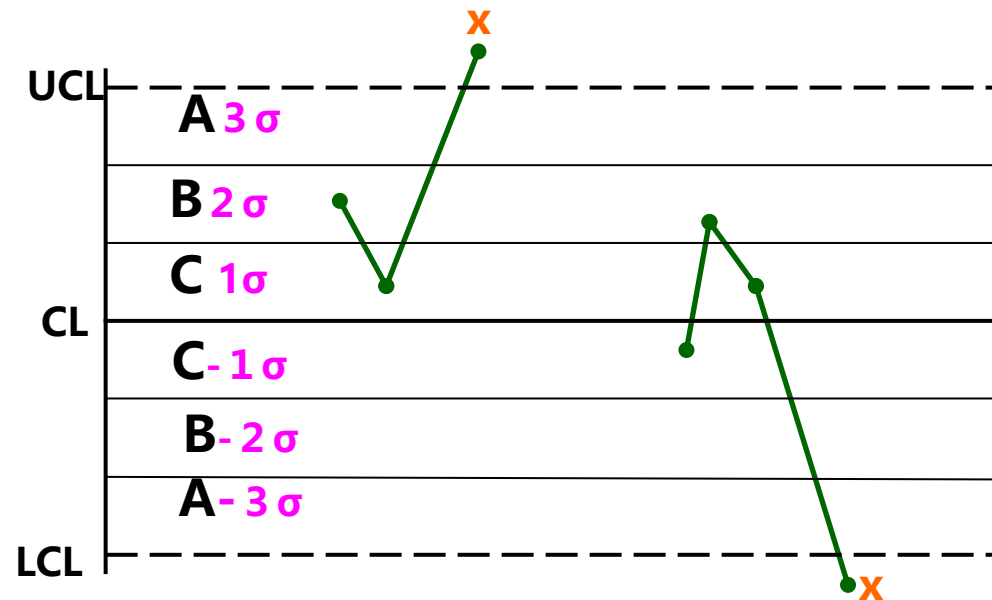
CONTROL CHARTS

- In the following four slides, eight additional tests are summarized which are used to interpret the trends in Shewhart's charts.
- The occurrence of any of the conditions predicted by these tests is an indication of the presence of possible identifiable causes of variability that should be investigated and, if confirmed, corrected.

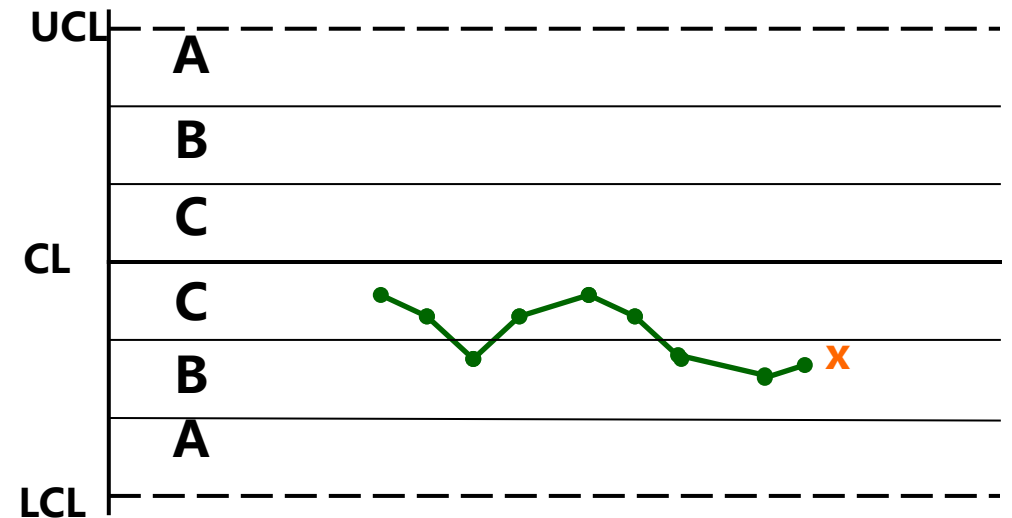
CONTROL CHARTS

■ Identification of assignable causes

Test 1. One point outside zone A



Test 2. Nine points in a row (in zone C or other) on the same side of center line

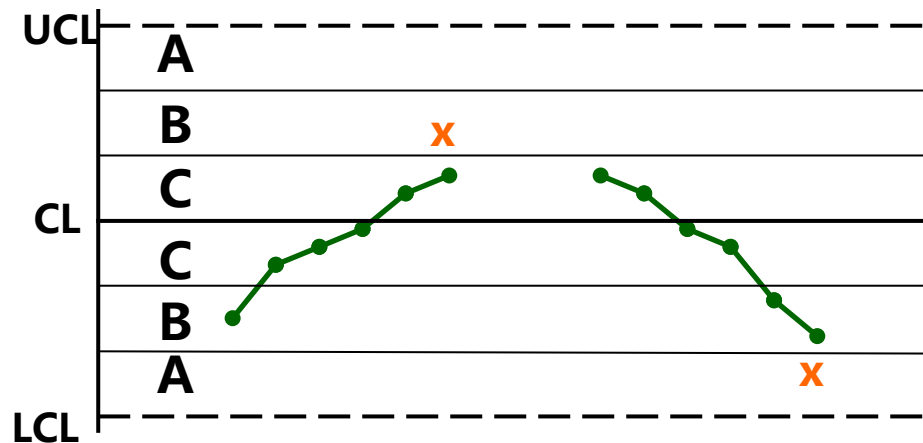


ISO 8258:2004 Shewhart control charts

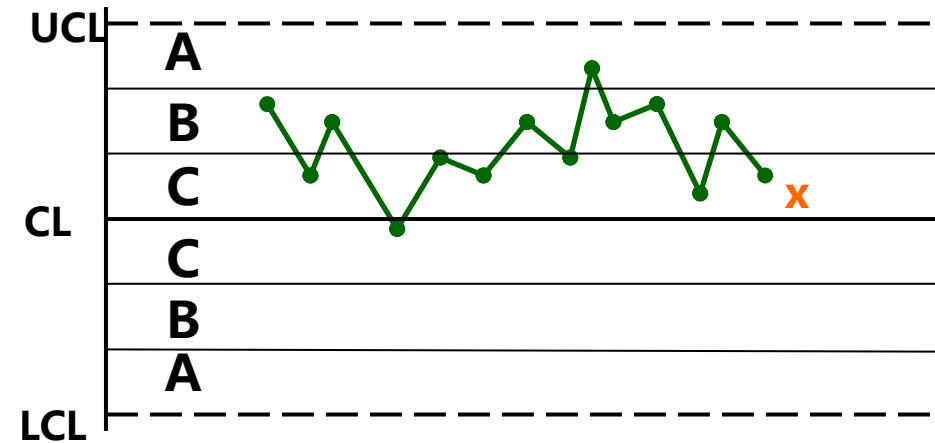
CONTROL CHARTS

■ Identification of assignable causes

Test 3. Six points in a row systematically increasing or decreasing



Test 4. Fourteen points in a row alternating up and down

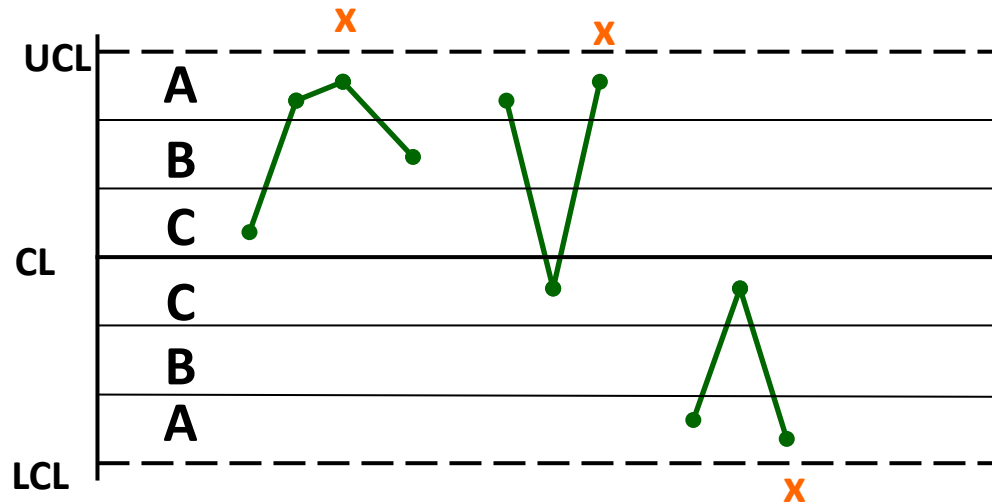


ISO 8258:2004 Shewhart control charts

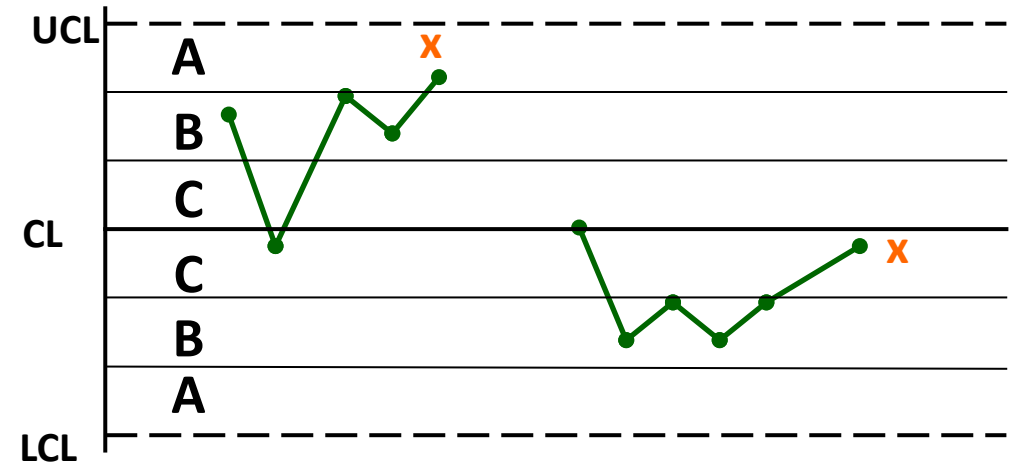
CONTROL CHARTS

■ Identification of assignable causes

Test 5. Two out of three points in a row in zone A or beyond



Test 6. Four out of five points in a row in zone B or beyond

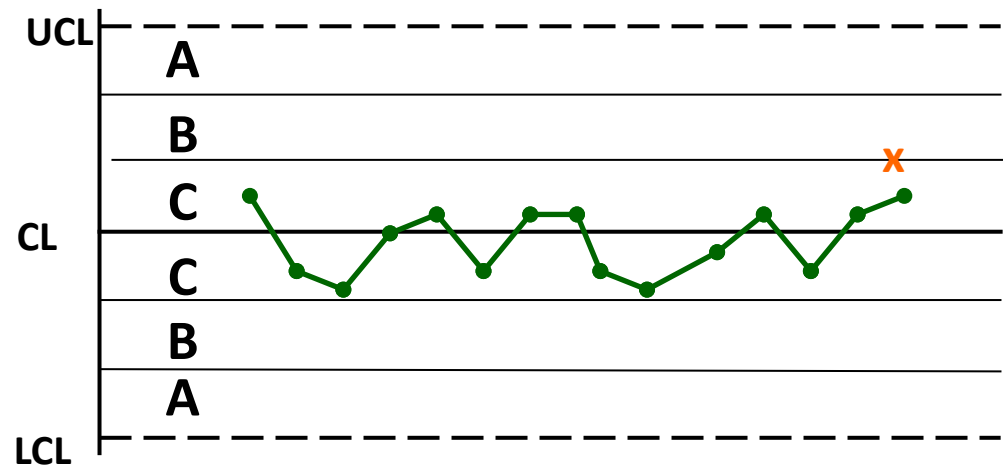


ISO 8258:2004 Shewhart control charts

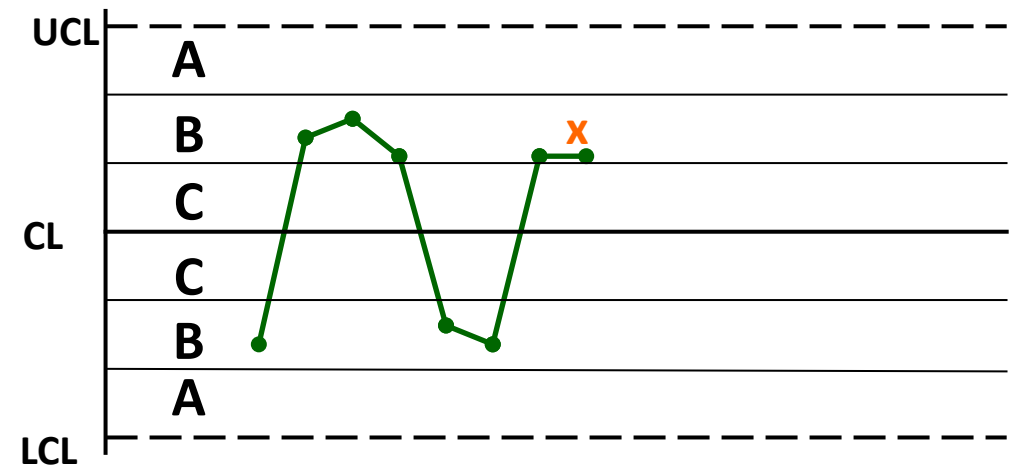
CONTROL CHARTS

■ Identification of assignable causes

Test 7. Fifteen points in a row in zone C above and below the central line



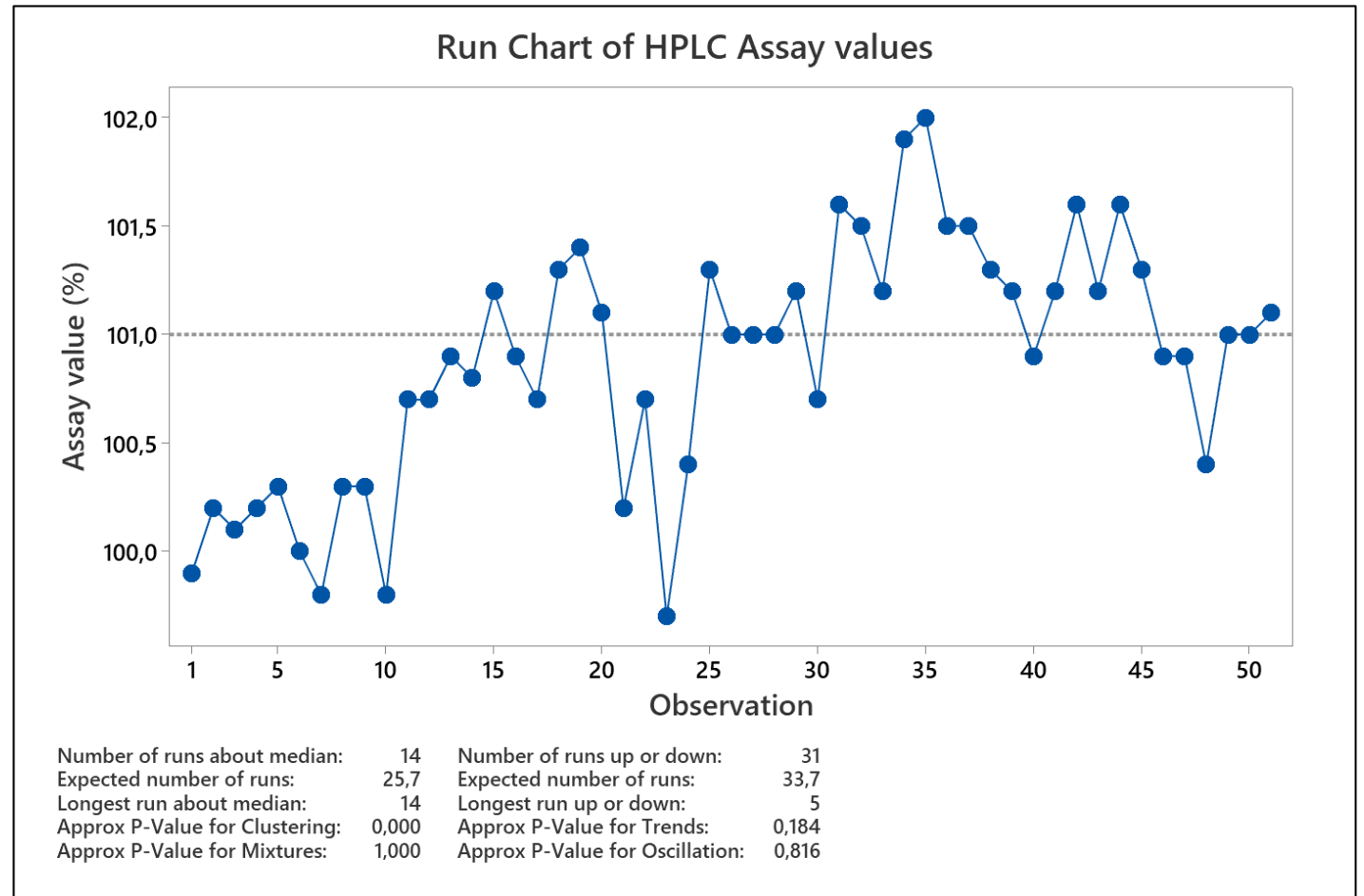
Test 8. Eight points in a row on both sides of the central line with none in zone C



CONTROL CHARTS

Similar to Shewhart's Control Charts is the *Run Chart*, which shows a measurement on the y-axis plotted over time (on the x-axis). A center line (CL) is drawn at the Median.

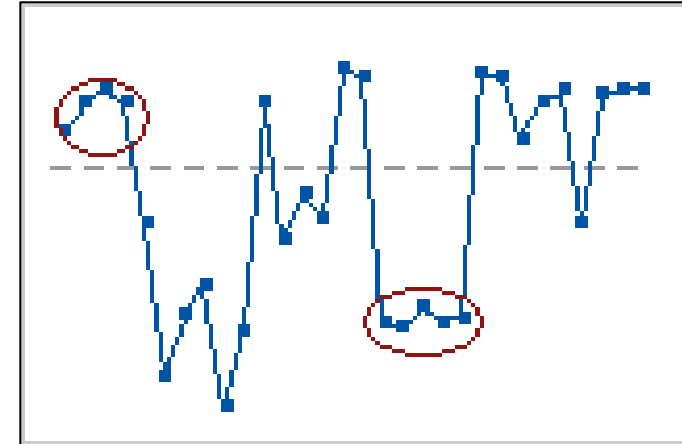
Run Charts are not just Line Graphs !



CONTROL CHARTS

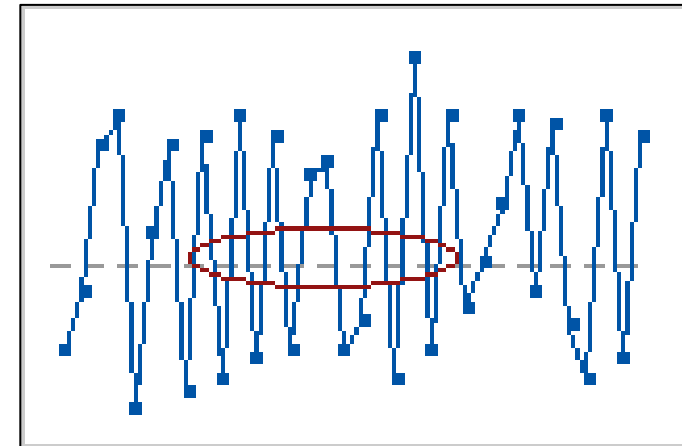
Clusters

- groups of datapoints in one area of the chart
- may indicate special-cause variation, such as measurement problems, lot-to-lot or set-up variability, or sampling from a group of defective parts
- If the *p-value* for clustering < 0.05 , possible clusters in data.



Mixtures

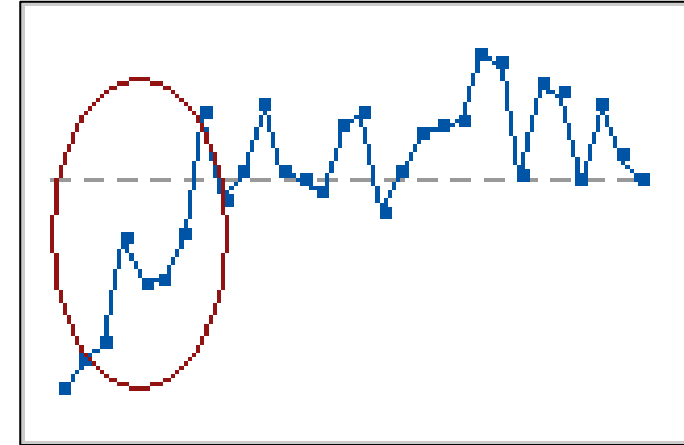
- data pattern characterized by frequent crossing of the center line
- Mixtures often indicate combined data from two populations, or two processes operating at different levels
- If the *p-value* for mixtures < 0.05 , possible mixtures in data.



CONTROL CHARTS

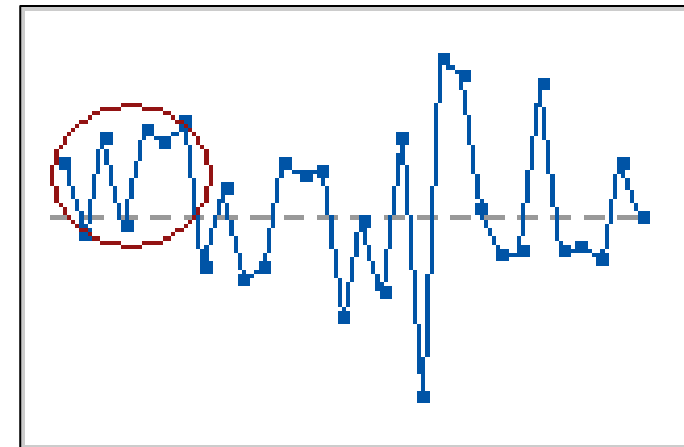
Trend

- sustained drift in the data, either up or down
- trends may warn that a process will soon go out of control
- A trend can be caused by factors such as worn tools, a machine that does not hold a setting, or periodic rotation of operators.
- if the p-value for trends < 0.05 , possible trend in data.

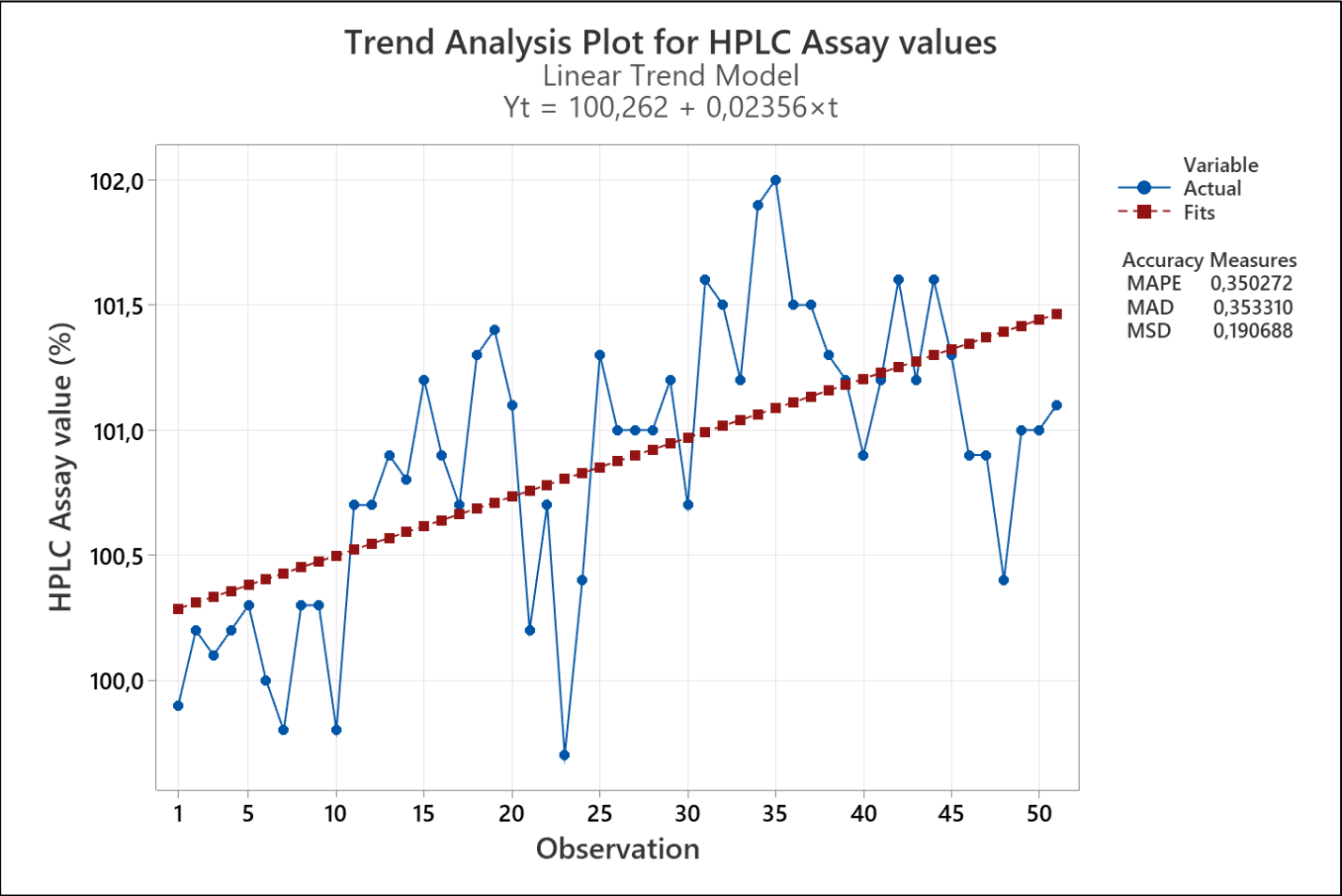


Oscillations

- occurs when the data fluctuates up and down, which indicates that the process is not steady.
- if the p-value for oscillation < 0.05 , possible oscillations in data



CONTROL CHARTS



CONTROL CHARTS

Control Charts by Attributes

Control Chart type	Use	Central Line	Upper and Lower Control Limits
p	to monitor the <u>proportion of defective items</u> that can be classified into one of two categories, such as pass or fail	\bar{p}	$\bar{p} \pm \sqrt{\bar{p}(1 - \bar{p})/n}$
np	to monitor the <u>number of defective items</u> that can be classified into one of two categories, such as pass or fail	$n \bar{p}$	$n \bar{p} \pm \sqrt{n \bar{p}(1 - \bar{p})}$
c	to monitor the <u>number of defects</u> where each item can have multiple defects	\bar{c}	$\bar{c} \pm 3\sqrt{\bar{c}}$
u	to monitor the <u>number of defects per unit</u> , where each item can have multiple defects.	\bar{u}	$u(-) \pm 3\sqrt{\bar{u}/n}$

CONTROL CHARTS

Alongside the control charts just discussed, which represent the more conventional and widely used ones (*e.g.*, SPC, *etc.*), there are also other types for "special applications", *e.g.*:

- *Short Run (Z-MR)* charts: using standardization (Z) allow to combine data from different (and short) runs in a single control chart.
- *Rare Events (G and T)* charts: monitor the number of days between rare events, *e.g.*, microbiological growth in highly controlled environments, *etc.*
- *Multivariate charts (T^2 and T)* charts: monitor whether the process location and the process variability of two or more related variables are in control. They are the multivariate counterpart to the \bar{X} -R, \bar{X} -S, and I-MR charts.

etc.

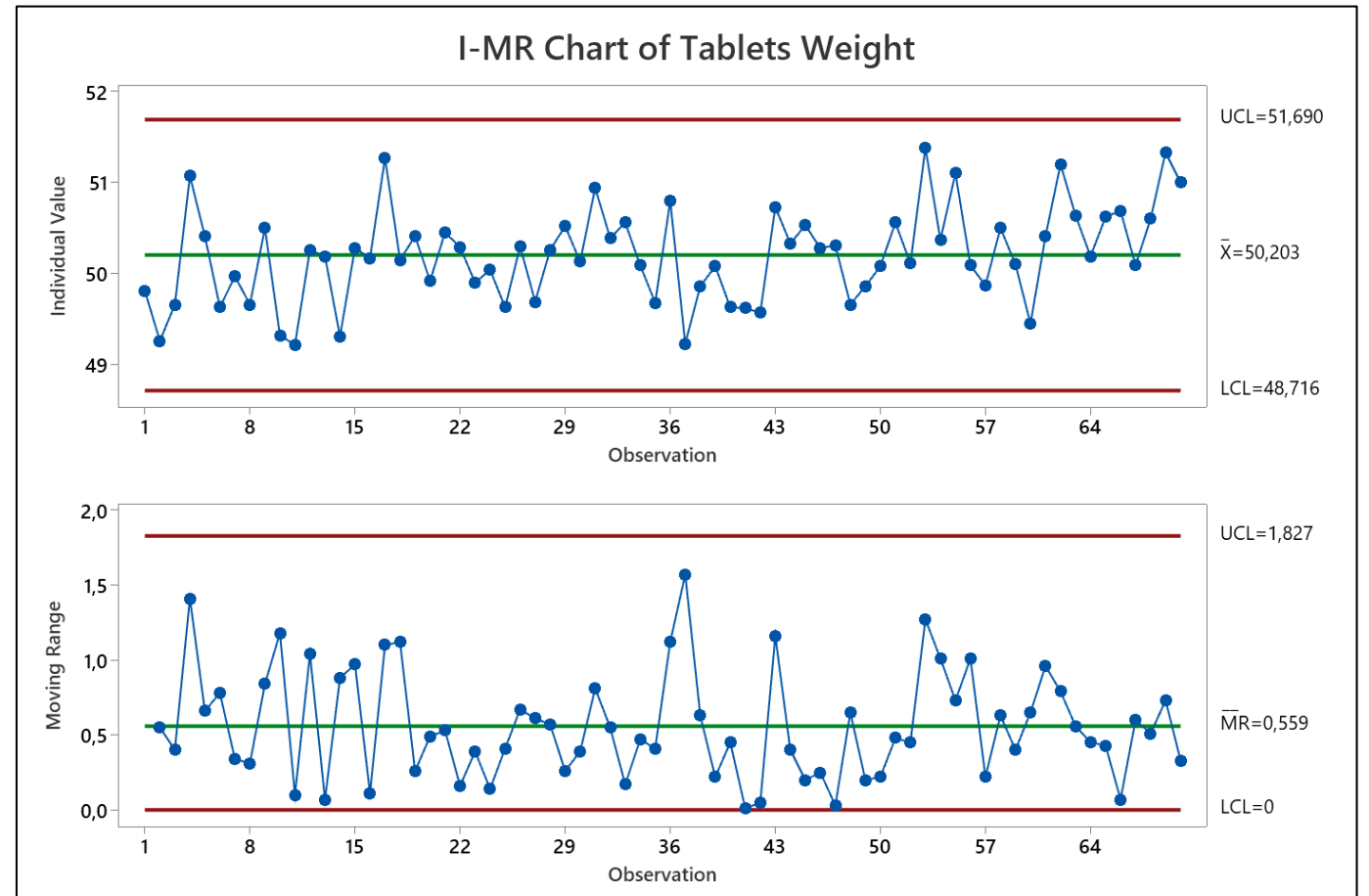
CONTROL CHARTS

- It is generally believed that Shewhart's control charts are not enough sensitive for quick identification of slight deviations of the parameters under control
- For this reason, other types of control charts have been introduced, among which the most used are:
 - *Exponentially Weighed Moving Average (EWMA)* charts
 - *Cumulative Sum (CUSUM)* control charts

CONTROL CHARTS

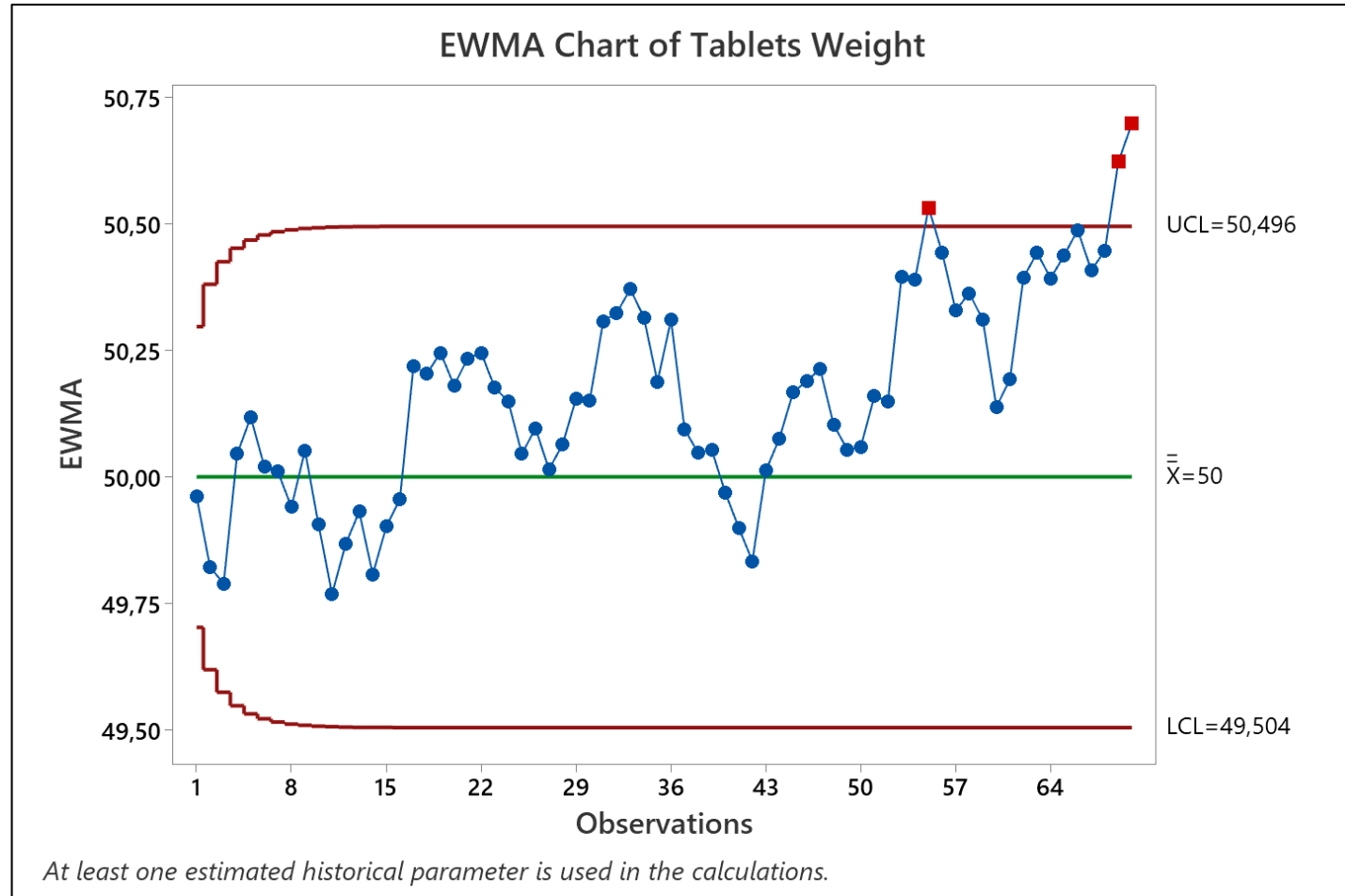
To check whether a particular tablet press can maintain a target weight of 50 mg, one tablet is sampled every 3 minutes and weighed. The data of 3.5 hours of production are examined.

An I-MR card is used to control the process. Apart from a slight increase in the process average at the end, the chart does not reveal anything else.



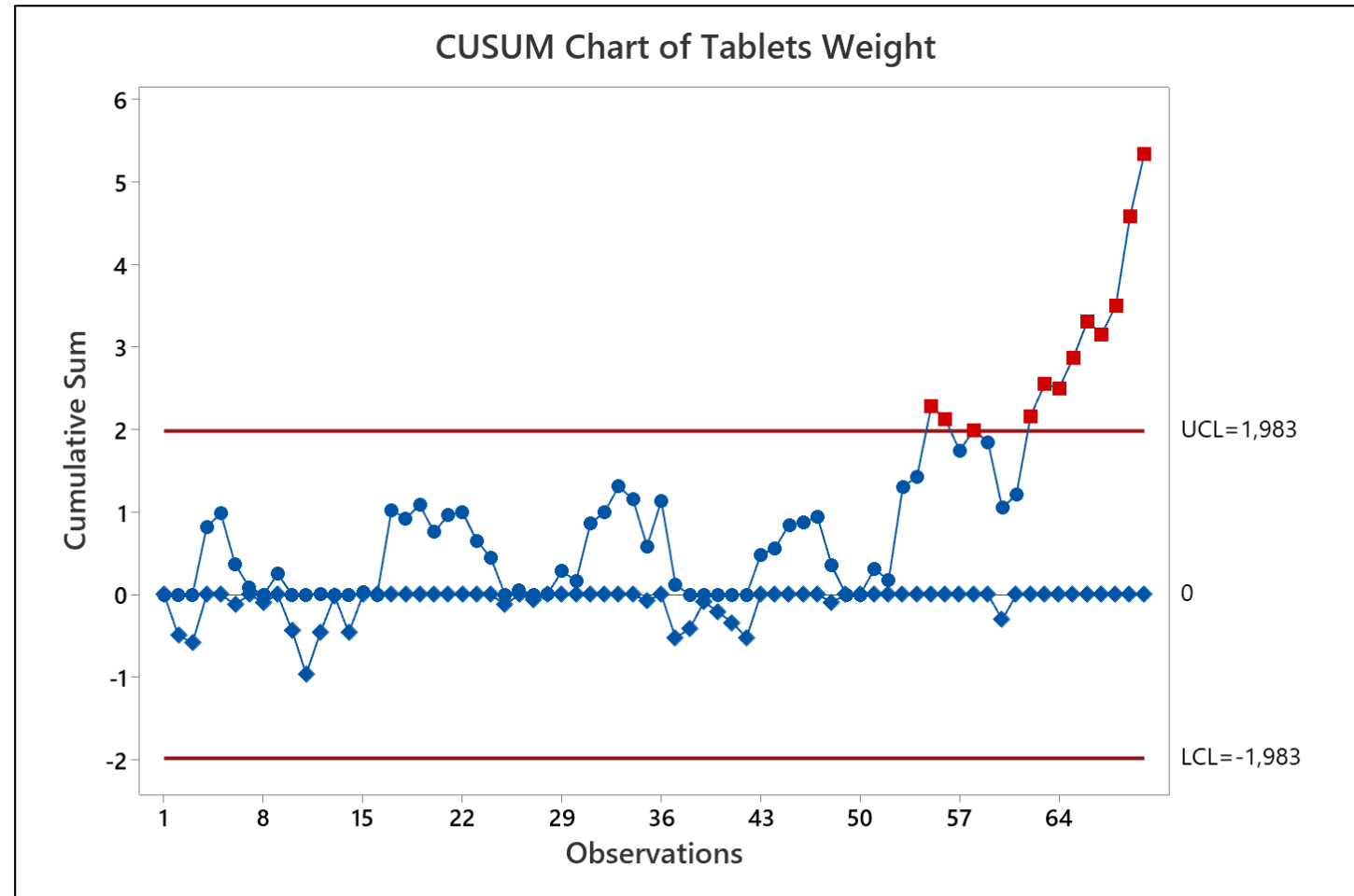
CONTROL CHARTS

Conversely, the use of an EWMA chart reveals a significant upward trend in the weight of the tablets as the process continues.



CONTROL CHARTS

While the EWMA chart reveals the presence of an upward trend (difficult to detect with more conventional means, *e.g.*, I-MR chart), the CUSUM control chart allows you to estimate exactly when this process of increasing the weight of the tablets begins.



CONTROL CHARTS

Summarizing:

A Control Chart can:

- demonstrate whether the process is stable and consistent over time. A stable process is one that includes only common-cause variation and does not have any out-of-control points, *i.e.*, ***is a predictable process.***
- verify that the process is stable before you perform a capability analysis. ***A capability analysis is only valid when performed on a stable process.***
- assess the effectiveness of a process change as it easily allow to compare shifts in the process mean and changes in the process variation.
- communicate the performance of the process during a specific period of time.

CAPABILITY ANALYSIS

CAPABILITY ANALYSIS

ICH guideline Q10 on Pharmaceutical Quality System (2008) : 4 times !

« To develop and use effective monitoring and control systems for process performance and product quality, thereby providing assurance of continued suitability and **capability of processes** » (page 3)

« Pharmaceutical companies should plan and execute a system for the monitoring of process performance and product quality to ensure a state of control is maintained. An effective monitoring system provides **assurance of the continued capability of processes** and controls to produce a product of desired quality and to identify areas for continual improvement » (page 8)

etc.

CAPABILITY ANALYSIS

FDA Guidance for Industry on Process Validation (2011) : 8 times !

« We recommend that a statistician or person with adequate training in statistical process control techniques develop the data collection plan and statistical methods and procedures used in measuring and evaluating process stability and **process capability** (page 14) »

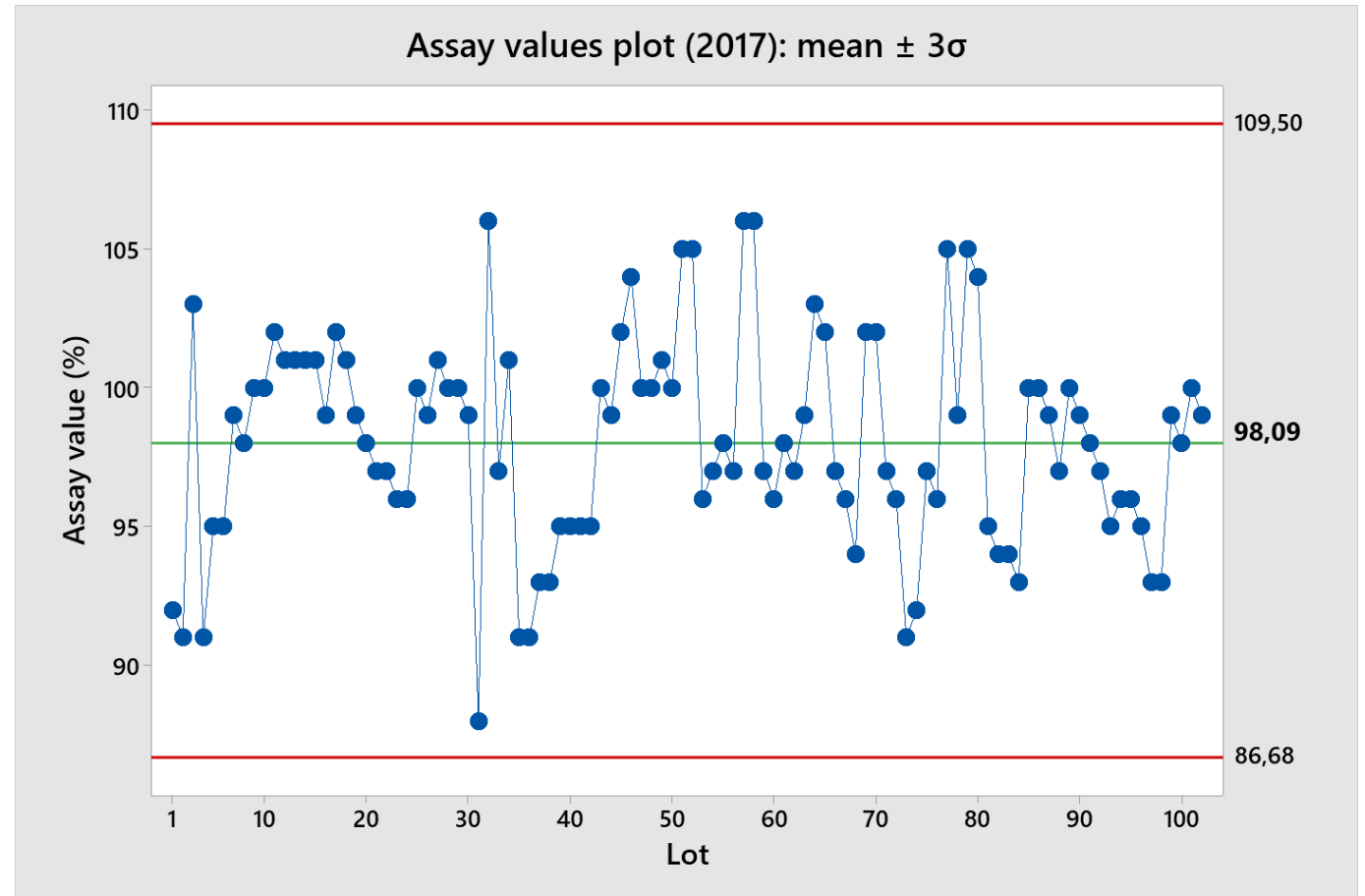
« Production data should be collected to evaluate process stability and **capability**. **The quality unit should review this information**. If properly carried out, these efforts can identify variability in the process and/or signal potential process improvements. (page 15) »

etc.

CAPABILITY ANALYSIS

For illustrative purposes, let's go back to an example we saw earlier.

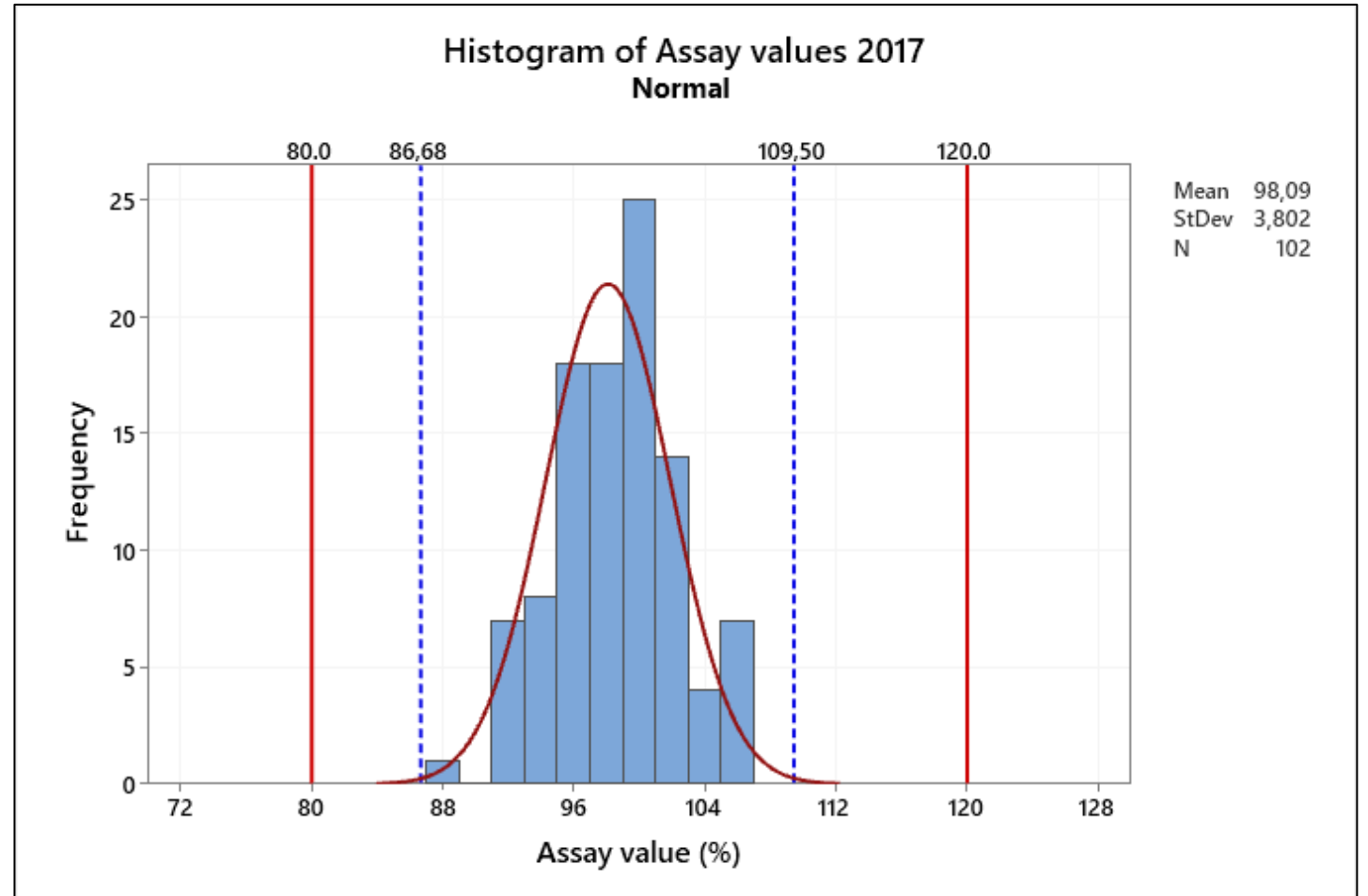
Here is the conventional plot showing « average $\pm 3\sigma$ » for the HPLC assay values of 102 lots of an API manufactured in 2017.



CAPABILITY ANALYSIS

Here, on the side, is the histogram that shows *the same data distribution* by representing for each assay value its frequency.

Using histograms is very easy to graphically identify the *central tendency* of the data as well as the *shape of the distribution*.



CAPABILITY ANALYSIS

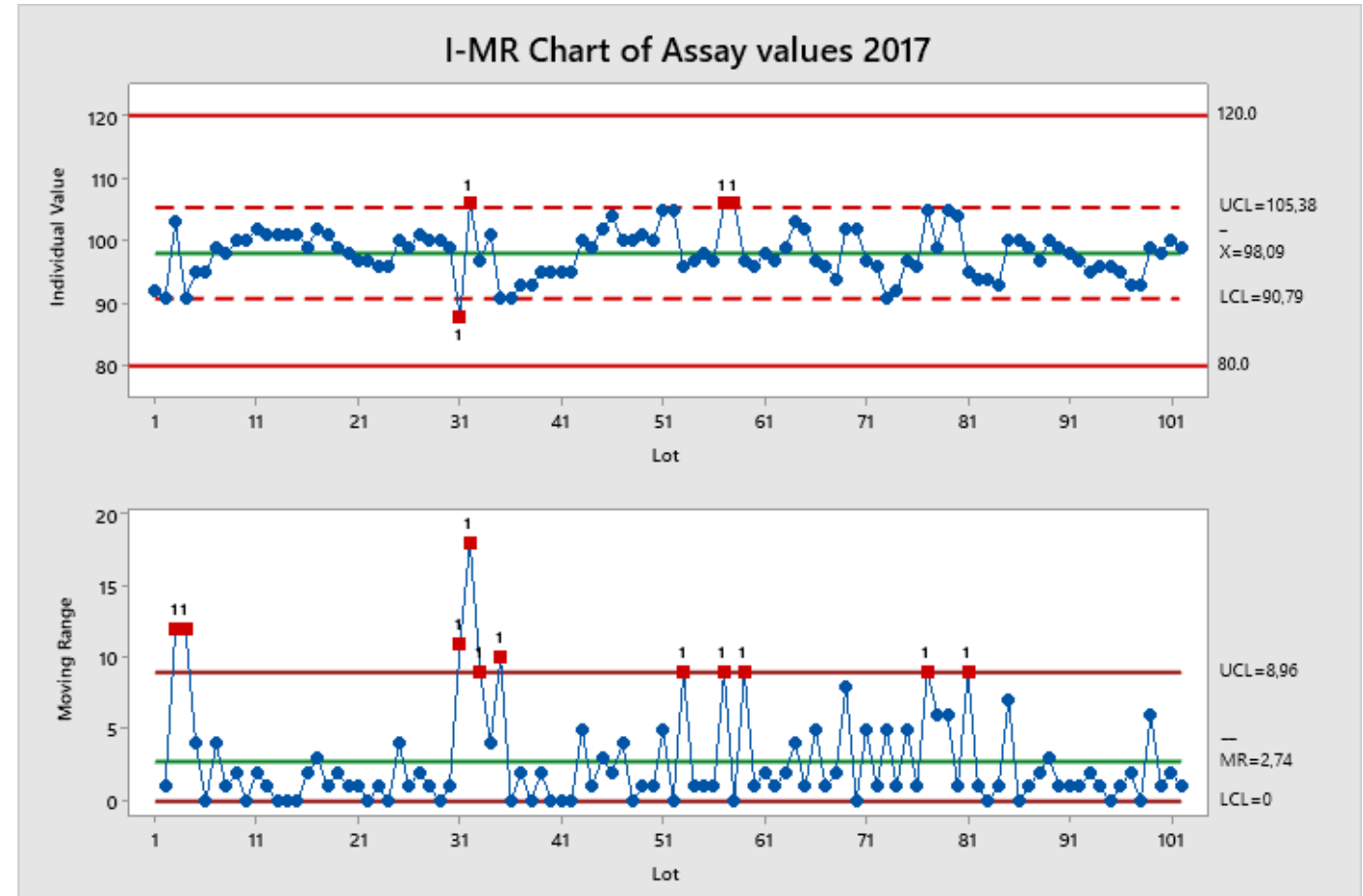
Here, on the side, is the *I-MR Chart* (mR = 2) of the same data distribution.

This chart provides information on the:

- **variation inherent to the process** known as *process spread* or *VOICE OF THE PROCESS, VOP*)

and

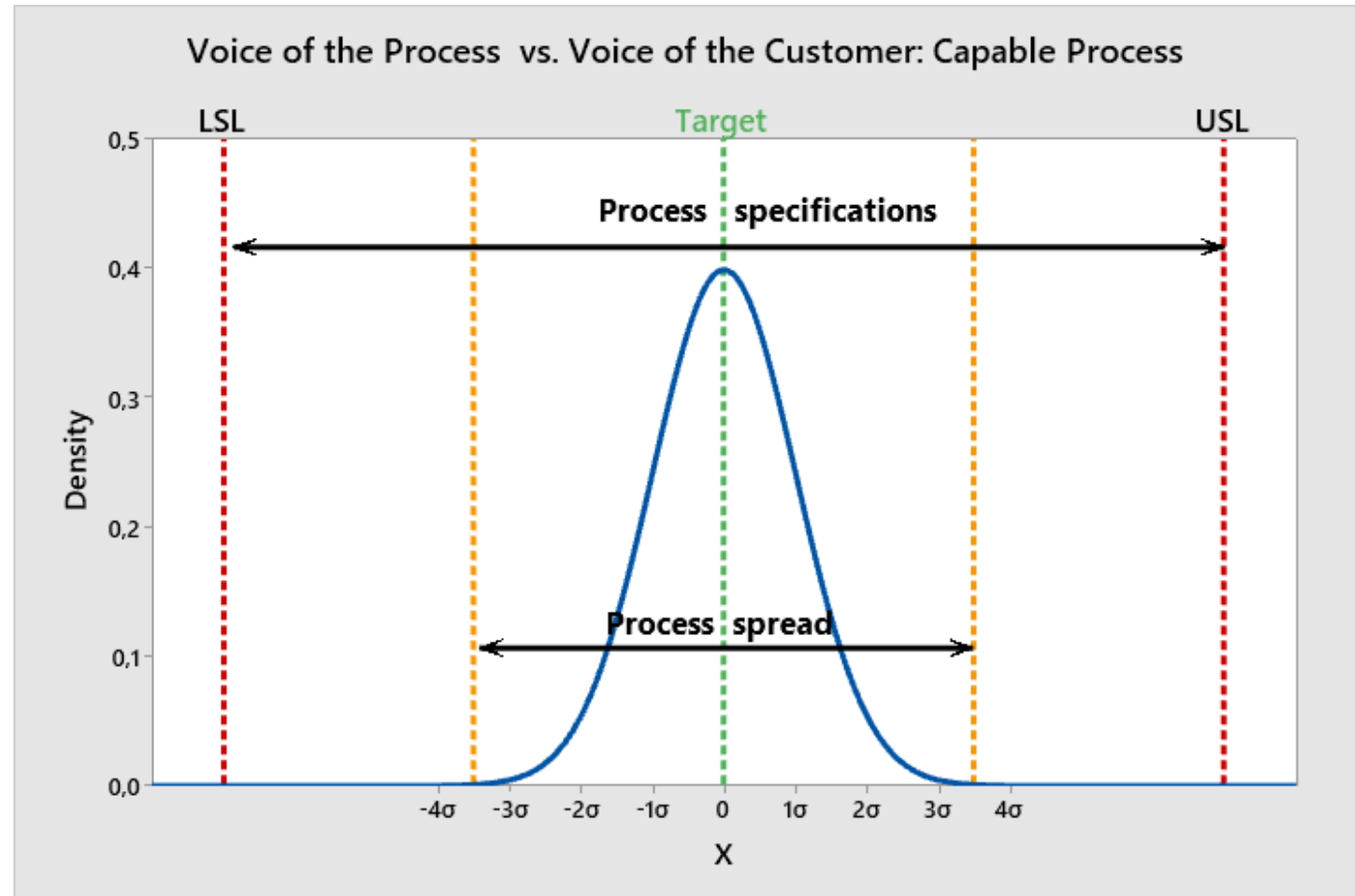
- **variation allowed by the Customer** known as *process specifications* or *VOICE OF THE CUSTOMER, VOC* (i.e., another department, a colleague, etc., **not necessarily the end user!**)



CAPABILITY ANALYSIS

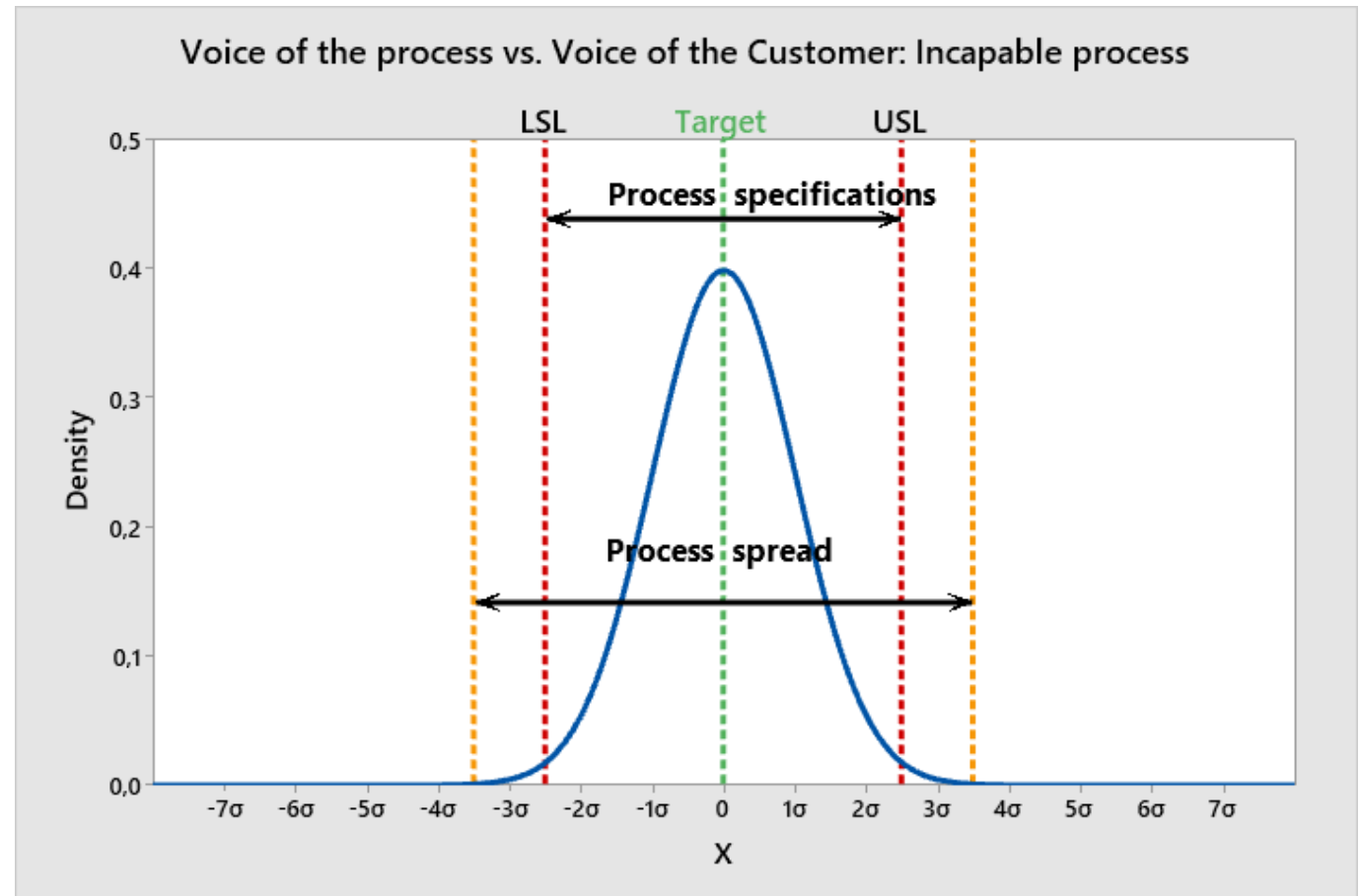
As long as the *process spread* (measured by the standard deviation, σ) lies within the process specifications, the process is said **capable** of delivering the quality required by the Customer.

The narrower is the process spread, the more capable is the process !



CAPABILITY ANALYSIS

Consequently, when the *process spread* is wider than the process specifications, the process is said **incapable** of delivering the quality required by the Customer.



CAPABILITY ANALYSIS

Quality is usually measured using the following indicators:

- defective units per million (ppm)
- defects per unit (dpu)
- defects per million opportunities (DPMO)
- defect yield

BUT

DEFECT YIELD is an indicator not informative in light of a process improvement as it cannot answer questions like:

- Is defectiveness a problem caused by the positioning of the mean or by excessive variability?
- To improve, should we then move the average or reduce process variability?



need of more efficient indicators !

CAPABILITY ANALYSIS



Capability Indices

C_p or *Capability Ratio* is defined as:

$$C_p = \frac{USL - LSL}{6 \text{ Sigma } (X)} = \frac{\text{Voice of the Customer}}{\text{Voice of the Process}} = \frac{\text{space available within specifications}}{\text{space required by the process}}$$

and it measures the ratio between the *admissible dispersion for the process* (difference between the specification limits set by the “customer”) and its *natural tolerance* (6σ).

*With a **predictable process** the Capability Ratio defines its ability to operate within the specifications.*

CAPABILITY ANALYSIS

*With a **predictable process** the Capability Ratio defines its ability or “elbow room” to operate within the specifications.*

*With an **unpredictable process** the Capability Ratio just defines its hypothetical “elbow room” to operate within the specifications.*

CAPABILITY ANALYSIS

6σ is used because in a normal distribution, such as the one under consideration, 99.73% of the observations is comprised of 6 times the standard deviation.

Because of this, *Cp can be calculated only if the process is stable and distributed normally.*

Cp is a good process indicator, but alone it is not enough because it only controls the *process dispersion*, but not its *centering*.

Cp indicates how capable a process is but only if it is centered !

CAPABILITY ANALYSIS

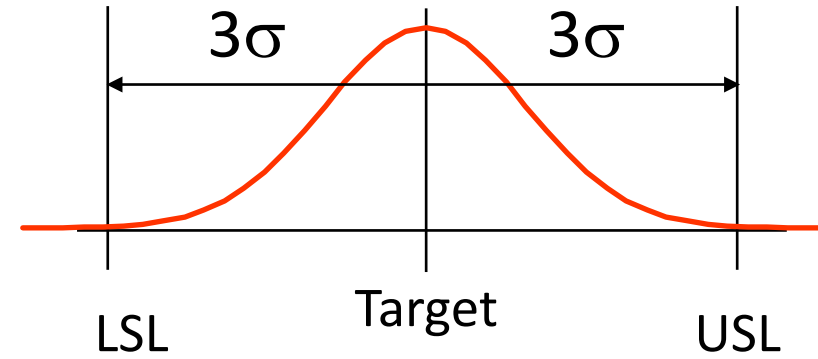
- if $C_p = 1 \Rightarrow 0.27\%$ of the observations do not conform to the specifications ($\pm 3\sigma$)
- if $C_p = 1.33 \Rightarrow 0.0064\%$ of the observations do not conform to the specifications ($\pm 4\sigma$)
- if $C_p = 1.67 \Rightarrow 0.000057\%$ of the observations do not conform to the specifications ($\pm 5\sigma$)

As general indication:

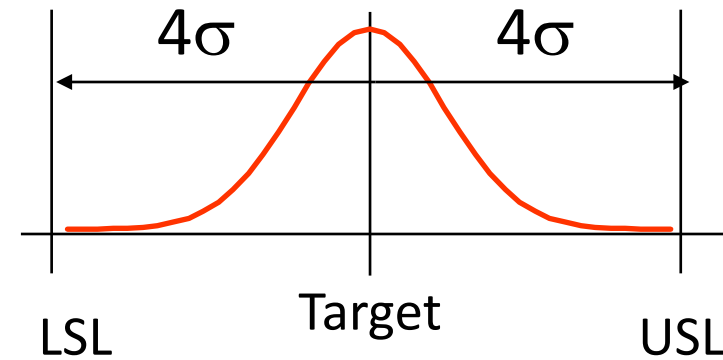
- if $C_p \geq 1.33$ the process can be considered *satisfactory*
- if $1.00 \leq C_p < 1.33$ the process can be considered *adequate*
- if $C_p < 1.00$ the process is *inadequate*

CAPABILITY ANALYSIS

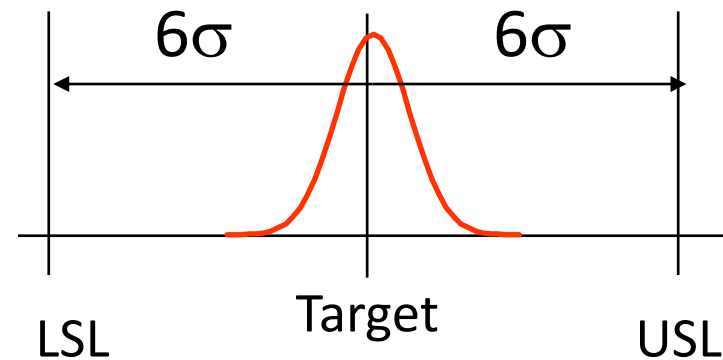
$$C_p = 1$$



$$C_p = 1.33$$



$$C_p = 2$$



CAPABILITY ANALYSIS

- ***Cpk*** or ***Centered Capability Ratio*** is defined as: $\min \{ (USL - \mu)/3\sigma ; (\mu - LSL)/3\sigma \}$ or $\min \{ CPU ; CPL \}$
- ***Cpk***, beside dispersion, also considers the *position* of the process with respect to the specification limits.
- if ***Cpk* > 1** : data are within *specification limits*
- if ***0 < Cpk < 1*** : part of the observations lie beyond the specification limits
- if ***Cpk < 0*** : data, on the average, are out of specifications
- if ***Cpk = 1*** : 99.73% of the observations are within the specification limits (*i.e.*, only 3 observations on 1000 are rejected)

CAPABILITY ANALYSIS

In terms typical of Quality Control:

- $Cpk > 1$: the process works well
- $Cpk = 1$: we are at the limit of the processing of non-conformed items
- $0 < Cpk < 1$: non-compliant items are processed
- $Cpk = 0$: half of the items are out of specification
- $-1 < Cpk < 0$: more than 50% of the items are out of specification
- $Cpk < -1$: nearly all items are out of specifications

CAPABILITY ANALYSIS

- In the manufacturing industry many Companies require their suppliers **Cpk** values of 1.33 or even 2.
Cpk = 1.33 means that the difference between the average value μ and the tolerance limit is 4σ ,
i.e., 99.994% of the product is within specification.
- An improvement from 1.33 to 2 is not always justified! It is a matter of a cost-benefit assessment.
- **Cpk** can never be greater than **Cp**, in the best case the two coincide.
- **Cpk = Cp** if the average value corresponds with the average value of the specification. **Cp** can therefore indicate how much better **Cpk** would be if the process was such that the distribution center was close to the midpoint of the specifications.

CAPABILITY ANALYSIS

- Beside the *Capability Ratio (Cp)* and the *Centered Capability Ratio (Cpk)* we can define two other performance indices, *i.e.*, *Performance Ratio (Pp)* and the *Centered Performance Ratio (Ppk)* which will have the same numerators as the capability indices, but whose denominators are based upon descriptive statistics such as the interval defined by « average $\pm 3\sigma$ » considered earlier.
- When the process is *operated predictably* the performance indexes will characterize the same things that are characterized by the capability indexes. However, when the process is not operated predictably the performance indices will describe the past.

CAPABILITY ANALYSIS

Pp or ***Performance Ratio*** is defined as:

$$Pp = \frac{USL - LSL}{6s} = \frac{\text{space available to the process within specifications}}{\text{space used by the process in the past}}$$

and it is important to remember that :

The extent to which the Capability Ratio (Cp) exceeds the Performance Ratio (Pp)
defines the degree of unpredictability for a process and
the opportunity that exists for improving that process !

CAPABILITY ANALYSIS

<div>Predictable process</div>	<div>Threshold State</div> <div>$Cpk < 1$</div> <div>Product trouble</div> <div>$Cp > 1$ Center the process</div> <div>$Cp < 1$ Reengineering of the process</div>	<div>Ideal State</div> <div>$Cpk > 1$</div> <div>No trouble</div>
	<div>State of Chaos</div> <div>$Ppk < 1$</div> <div>Double Trouble</div> <div>$Pp > 1$ Centering the process may help, but full process potential requires predictable operation.</div>	<div>Brink of Chaos</div> <div>$Ppk > 1$</div> <div>Process Trouble</div> <div>Full process potential requires predictable operation.</div>
	<div>Some Nonconforming Product Produced</div>	<div>100% Conforming Product Produced</div>

D.J. Wheeler, *The Six Sigma Practitioner's Guide to Data Analysis*, 2nd Ed., SPC Press(2010)

CAPABILITY ANALYSIS

SHORT-TERM CAPABILITY METRICS

- *C_p* and *C_{pk}* belong to the so-called *short-term capability metrics*, i.e., they are based on the short-term standard deviation of the process σ_{ST}
- If only a limited sample is used to estimate the process capability, then the sample only represents short-term capability regardless of which formulas are used.
- « *short-term variation* » is the variation observed when a sample of data is collected in a short period of time under essentially the same conditions. Such a sample is often called «*rational subgroup*»

CAPABILITY ANALYSIS

LONG-TERM CAPABILITY METRICS

- For each **Capability Index** (i.e., C_p , C_{pk}) there is a corresponding **Performance Index** (i.e., P_p , P_{pk}) which measures *how well the process performs over the long term*.
- Long-term capability metrics are based on the long-term standard deviation of the process σ_{LT} .
- « Long-term capability metrics » measure process variation over a period of time long enough to include all expected sources of variation.

CAPABILITY ANALYSIS

POTENTIAL and ACTUAL CAPABILITY METRICS

- **Potential Metrics** (i.e., short-term C_p , and long-term P_p) consider only the standard deviation of the process. They have the same values regardless from process centering. They assume that centering a process is easier than reducing its variation.

C_p and P_p describe how good a process could potentially be if centered.

- **Actual Metrics** (i.e., short-term C_{pk} , and long-term P_{pk}) consider both the average and the standard deviation of the process.

If a process is centered: $C_{pk} = C_p$ and $P_{pk} = P_p$

If a process is off-target: $C_{pk} < C_p$ and $P_{pk} < P_p$

C_{pk} and P_{pk} penalize processes that are off-target.

CAPABILITY ANALYSIS

	Capability Index	Performance Index
Type of indices	<i>Cp, Cpk</i>	<i>Pp, Ppk</i>
Data structuring	in subgroups	ungrouped
What it is measured?	Short-term variation	Long-term variation
	Within-group variation	Variation across all data
	What the process is capable of at its best	How does the process perform over the long term

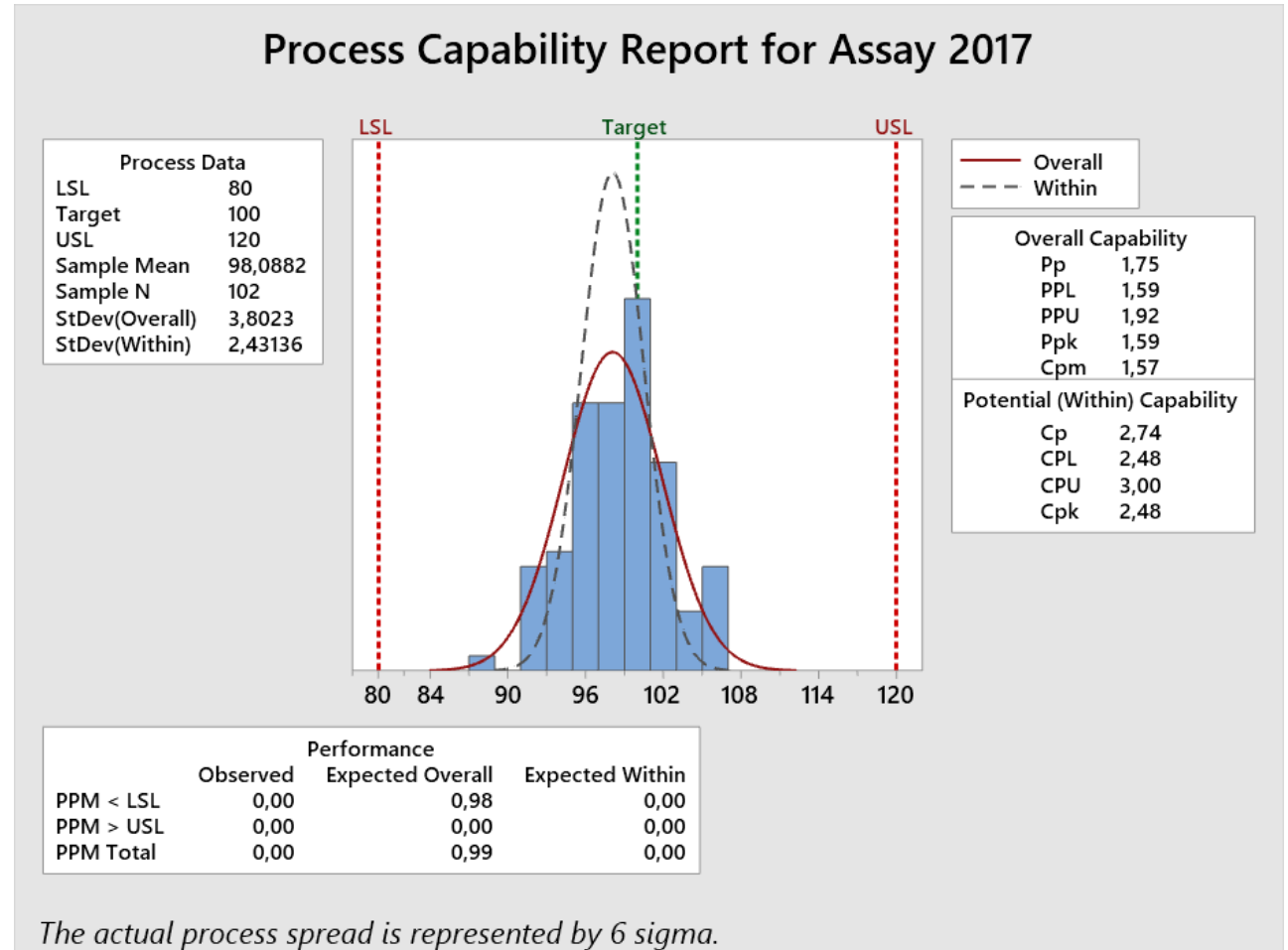
CAPABILITY ANALYSIS

Let's now consider our initial process.

As expected, $C_p > C_{pk}$ (in fact $2.74 > 2.48$), but it deals of very high values anyway. **The difference is due to the process which is not well centered on target.**

Overall Capability Metrics: are based on all variation seen in the analysis and reflect the current performance of the process.

Potential Capability Metrics: are based on short variation and reflect how good the process could be.

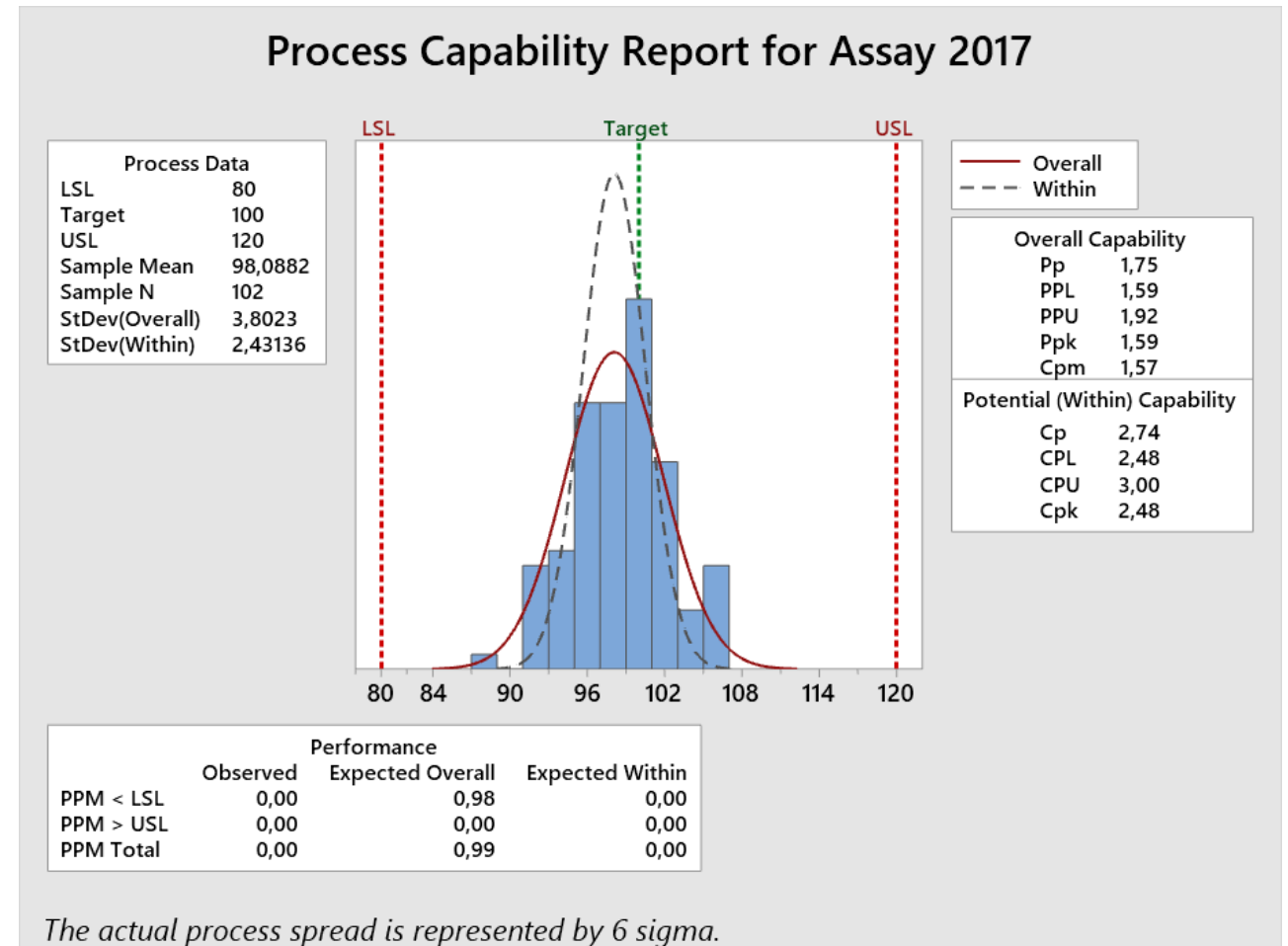


CAPABILITY ANALYSIS

Let's now consider our initial process.

As expected, $C_p > C_{pk}$ (in fact $2.74 > 2.48$), but it deals of very high values anyway. *The difference is due to the process which is not well centered on target.*

As PPM indicates the number of nonconforming parts in the process, expressed in parts per million, the Total PPM of Expected Overall Performance tells us that 1 lot on 1 million will be out of specs... but this is acceptable 😊

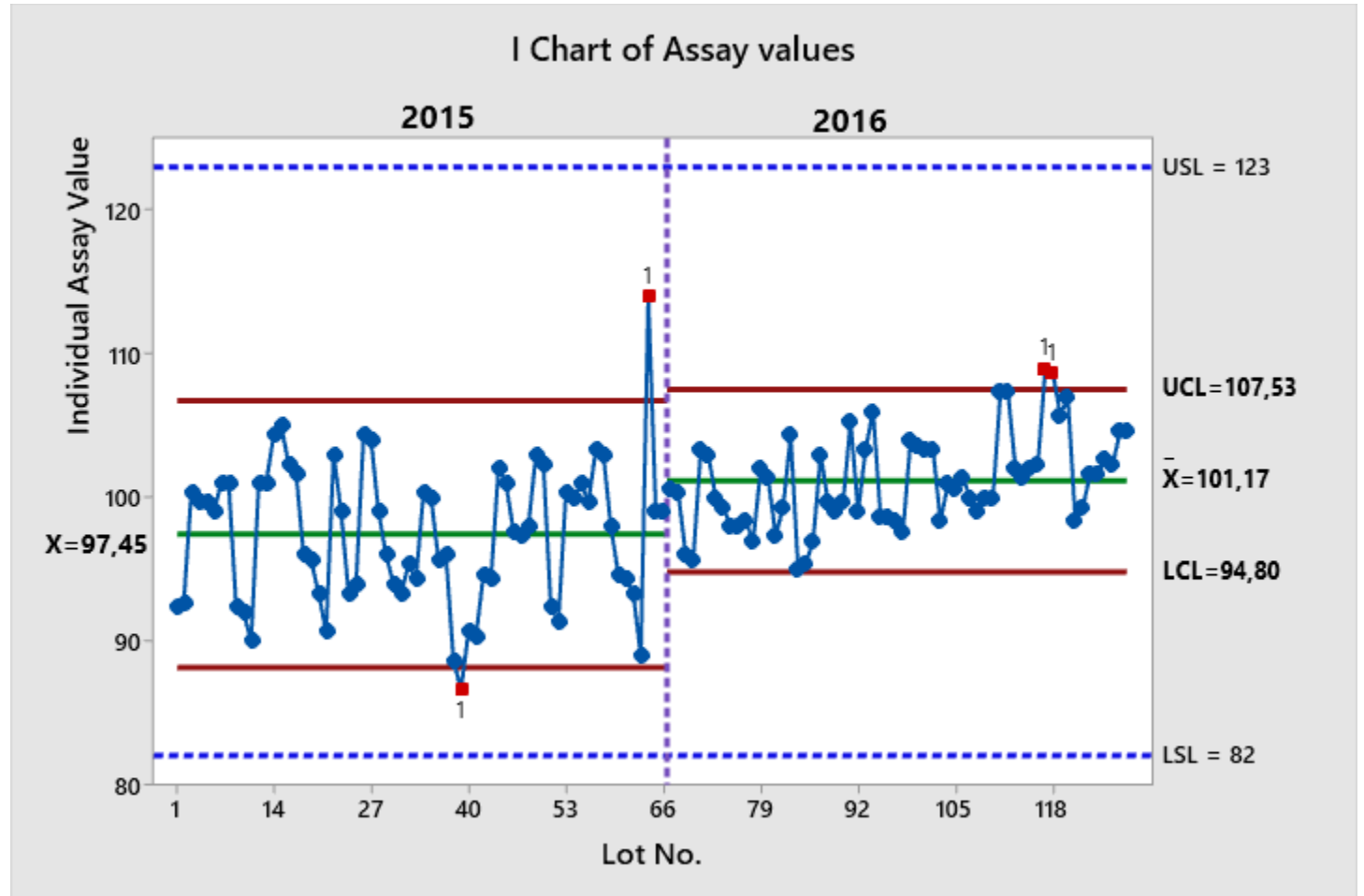


CAPABILITY ANALYSIS

Example 1

Here is an *I Chart* (or *X Chart*) displaying the assay values pertinent to an API manufacturing process collected in two subsequent years.

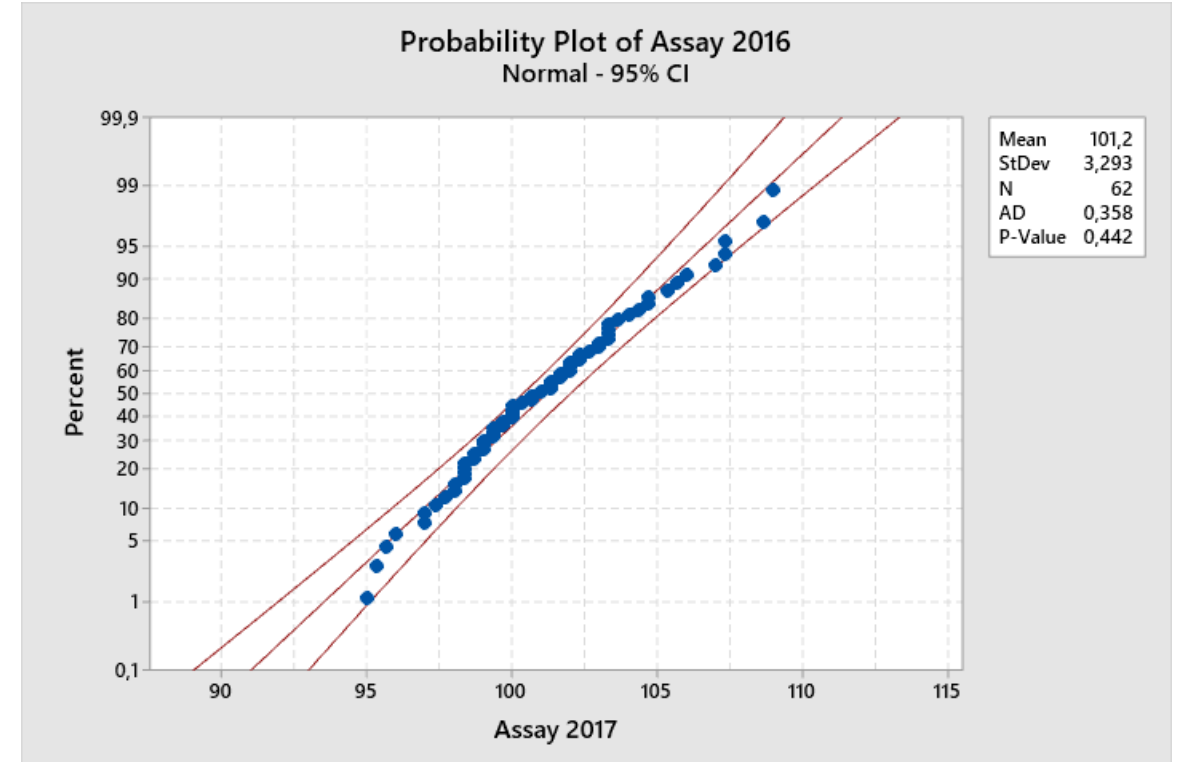
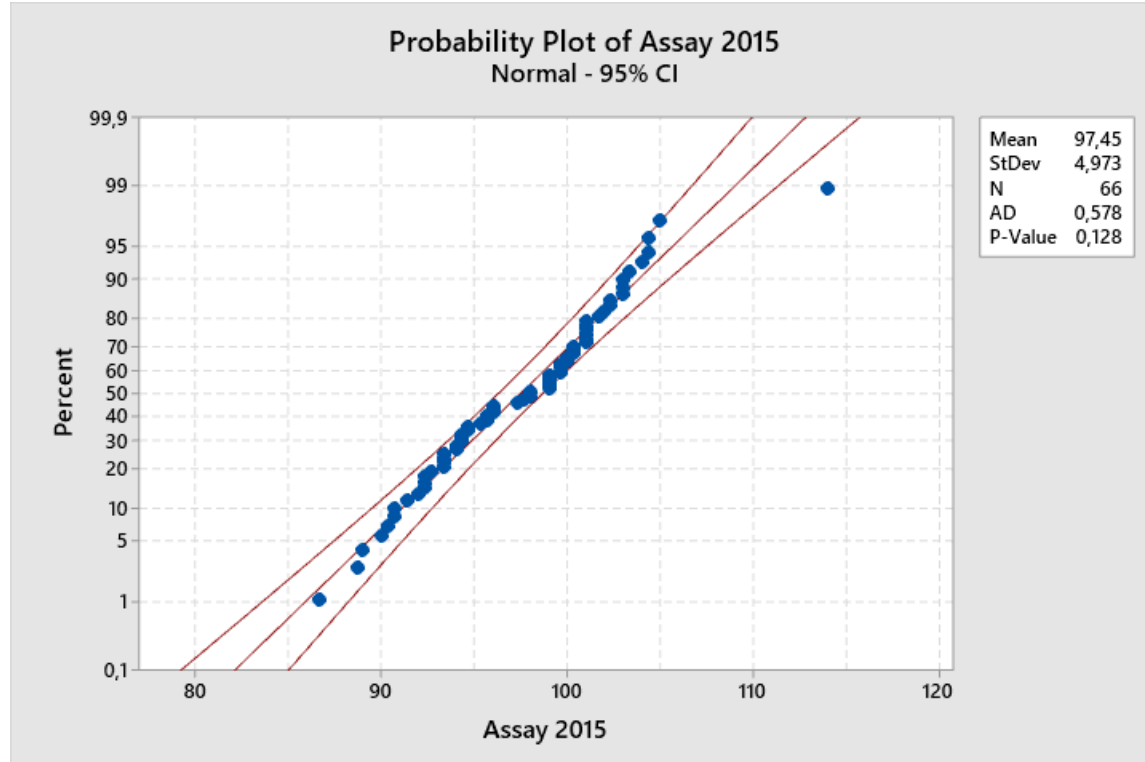
Let's see quickly how a Capability Analysis can be set up and what it reveals.



CAPABILITY ANALYSIS

Example 1 (cont.)

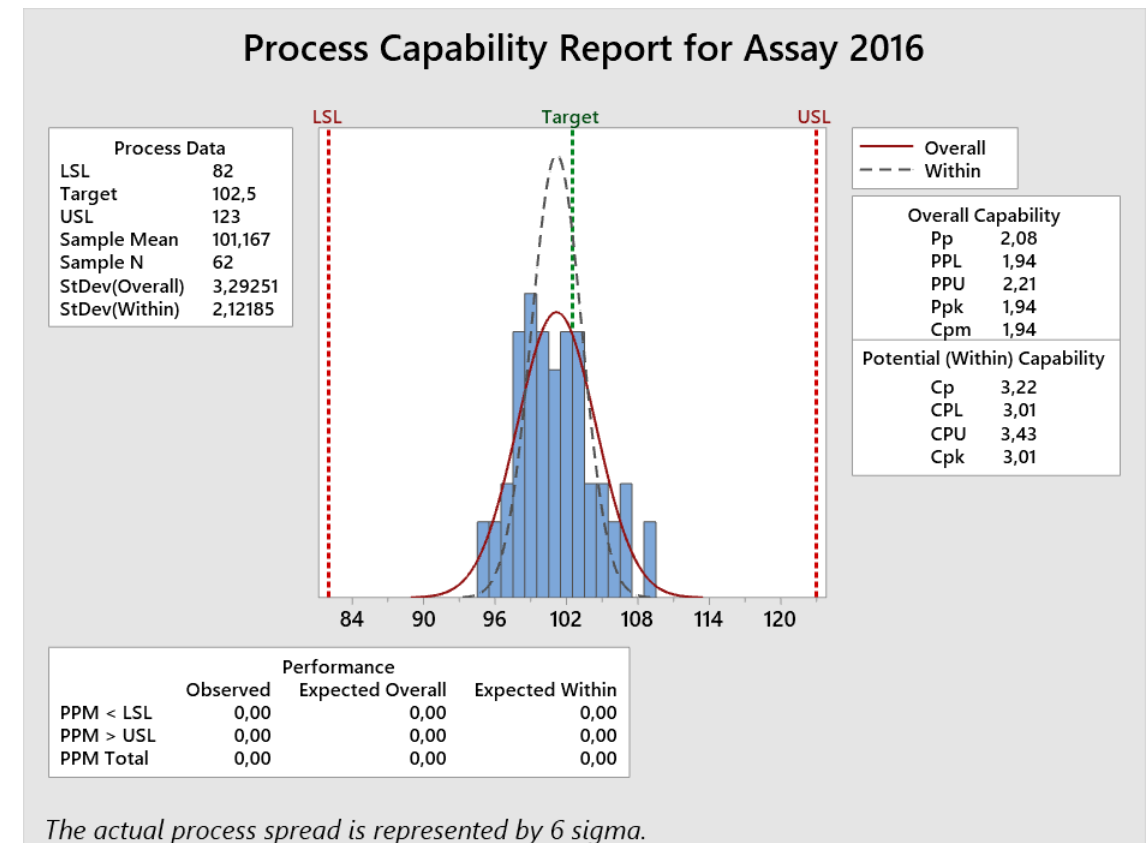
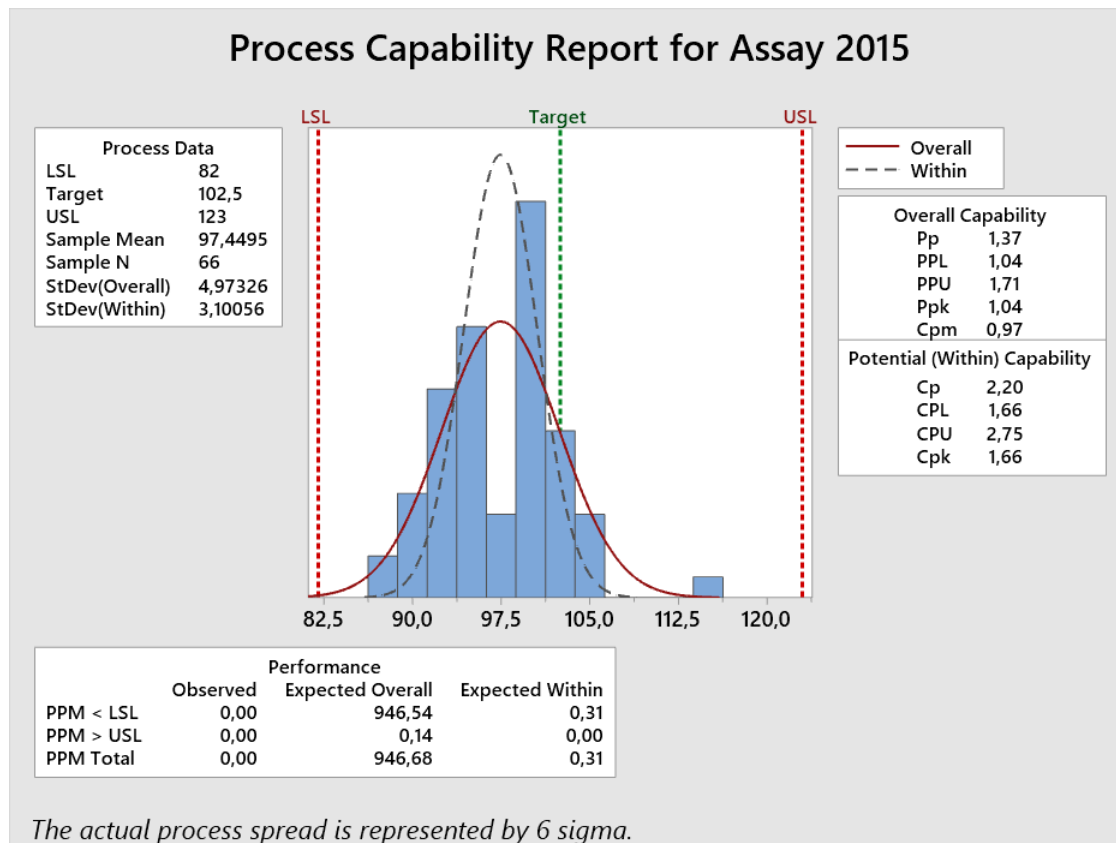
The first step is to investigate how data is distributed: $P\text{-value} > 0.05 \Rightarrow$ Normal distribution



CAPABILITY ANALYSIS

Example 1 (cont.)

Capability Analysis shows the overall process improvement resulting from spread reduction and centering.



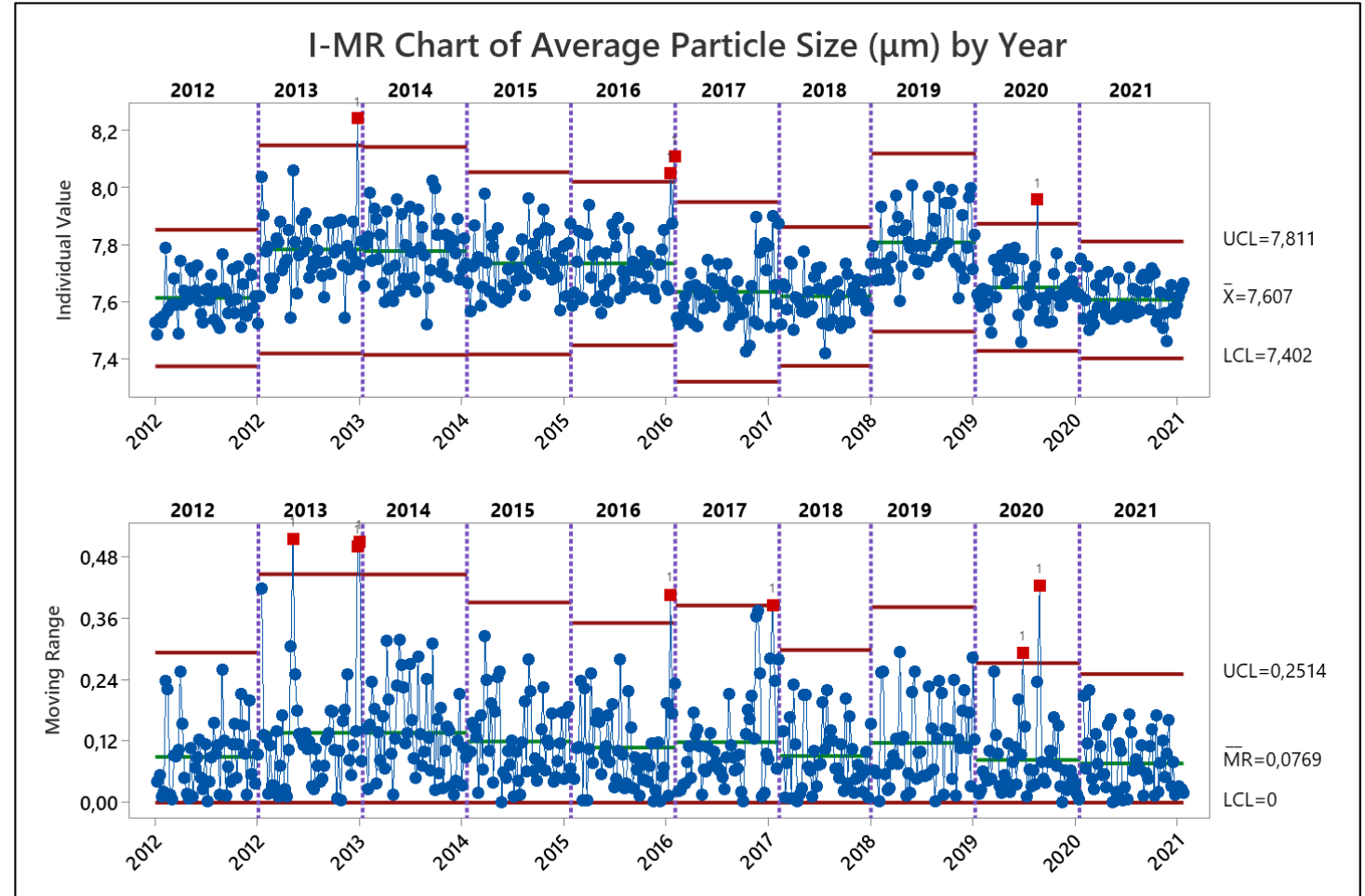
CAPABILITY ANALYSIS

Example 2

Here is an *I-MR Chart* (or *X-MR Chart*) displaying the average particle size values pertinent to an API collected in ten subsequent years.

The specification limit is :

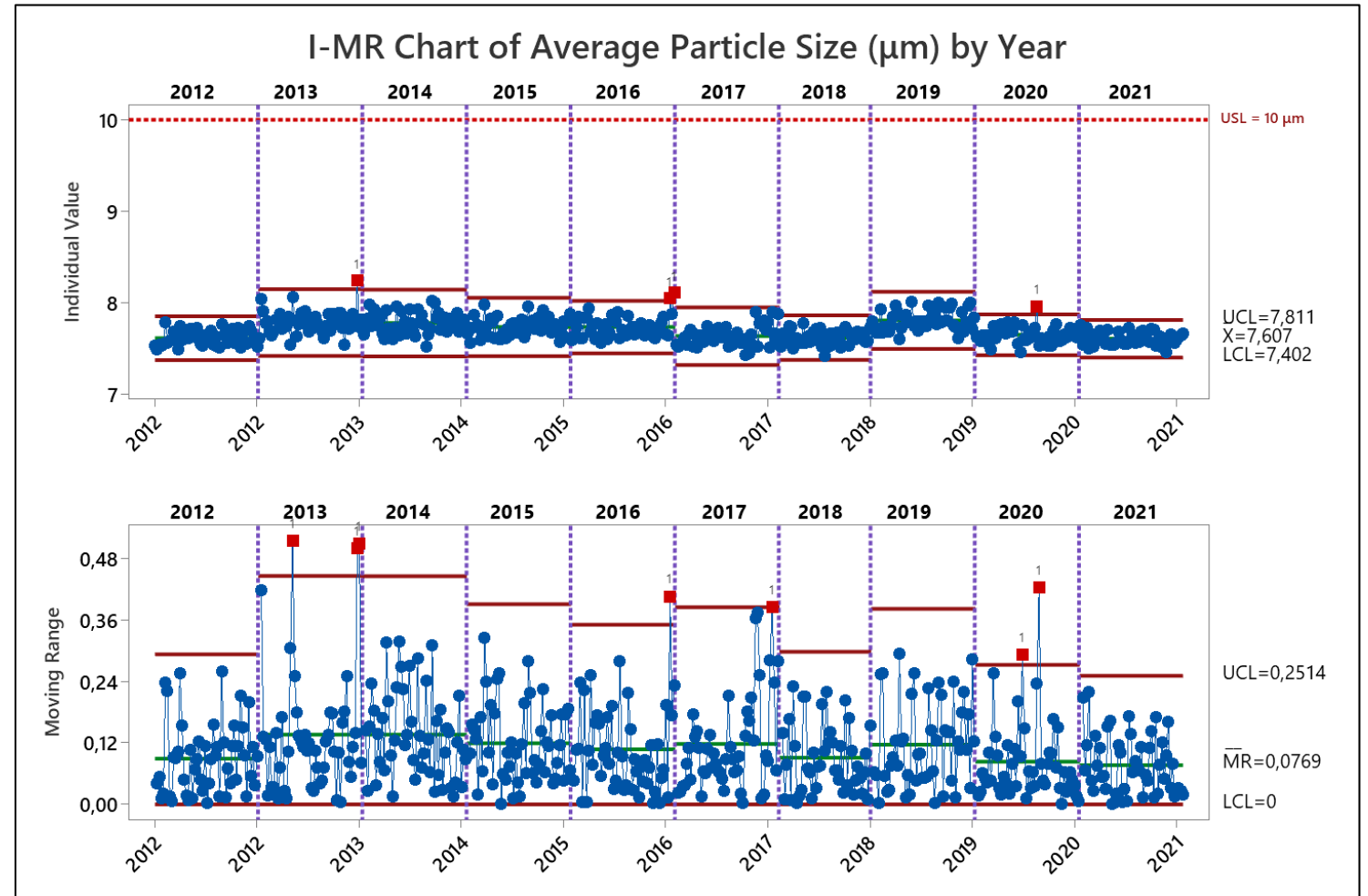
$$d_{100} < 10 \mu\text{m}$$



CAPABILITY ANALYSIS

Example 2 (cont.)

The control chart indicates a process that is overall within the control limits but appears too rich in data points and difficult to read. The specification limit is missing. Let's first add it!

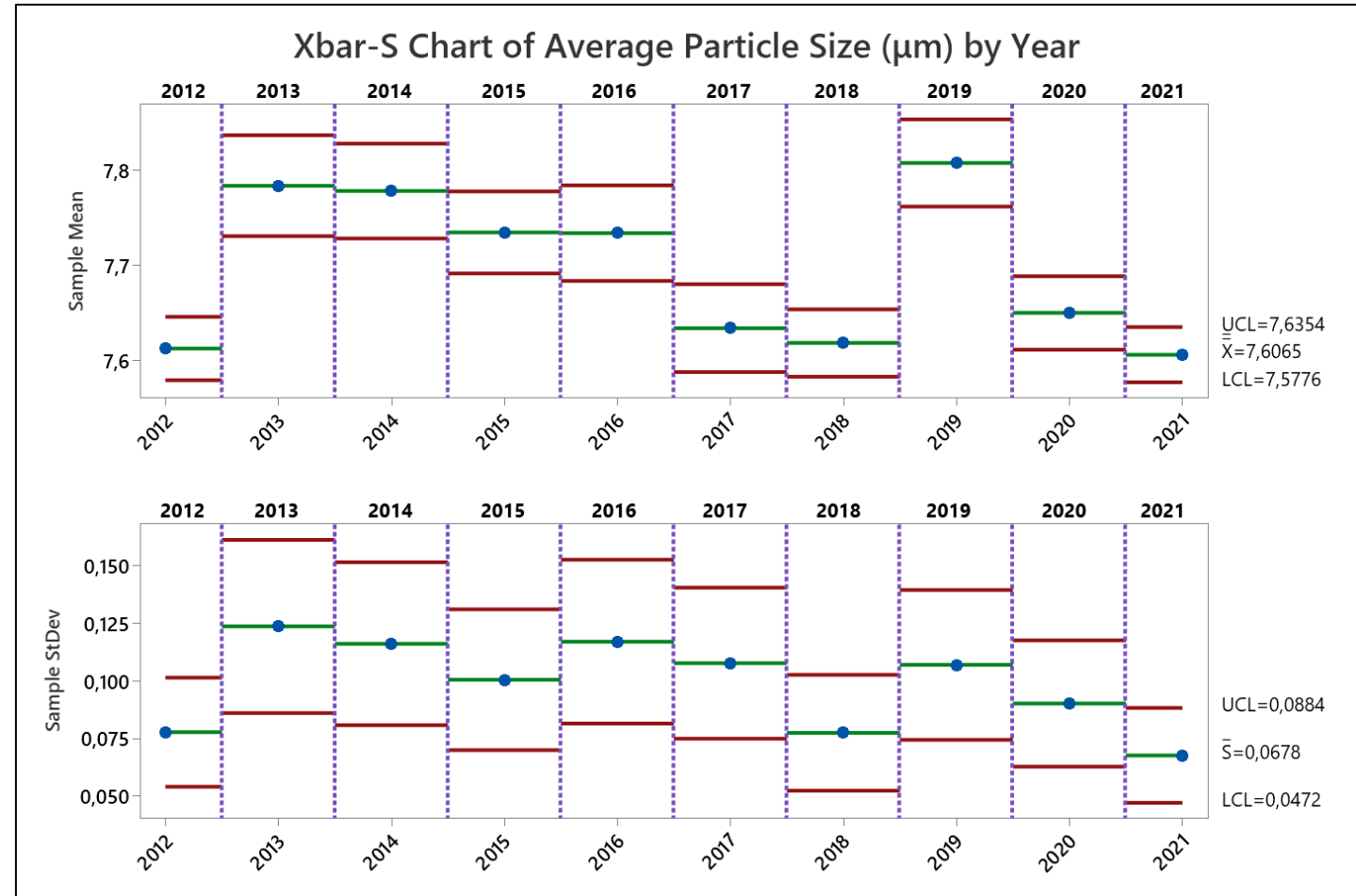


CAPABILITY ANALYSIS

Example 2 (cont.)

Now it is much better at least as regards the «voice of the process» compared to the «voice of the customer», but the graph is still difficult to read!

Let consider each year as a subgroup and use an Xbar-S Control Chart !

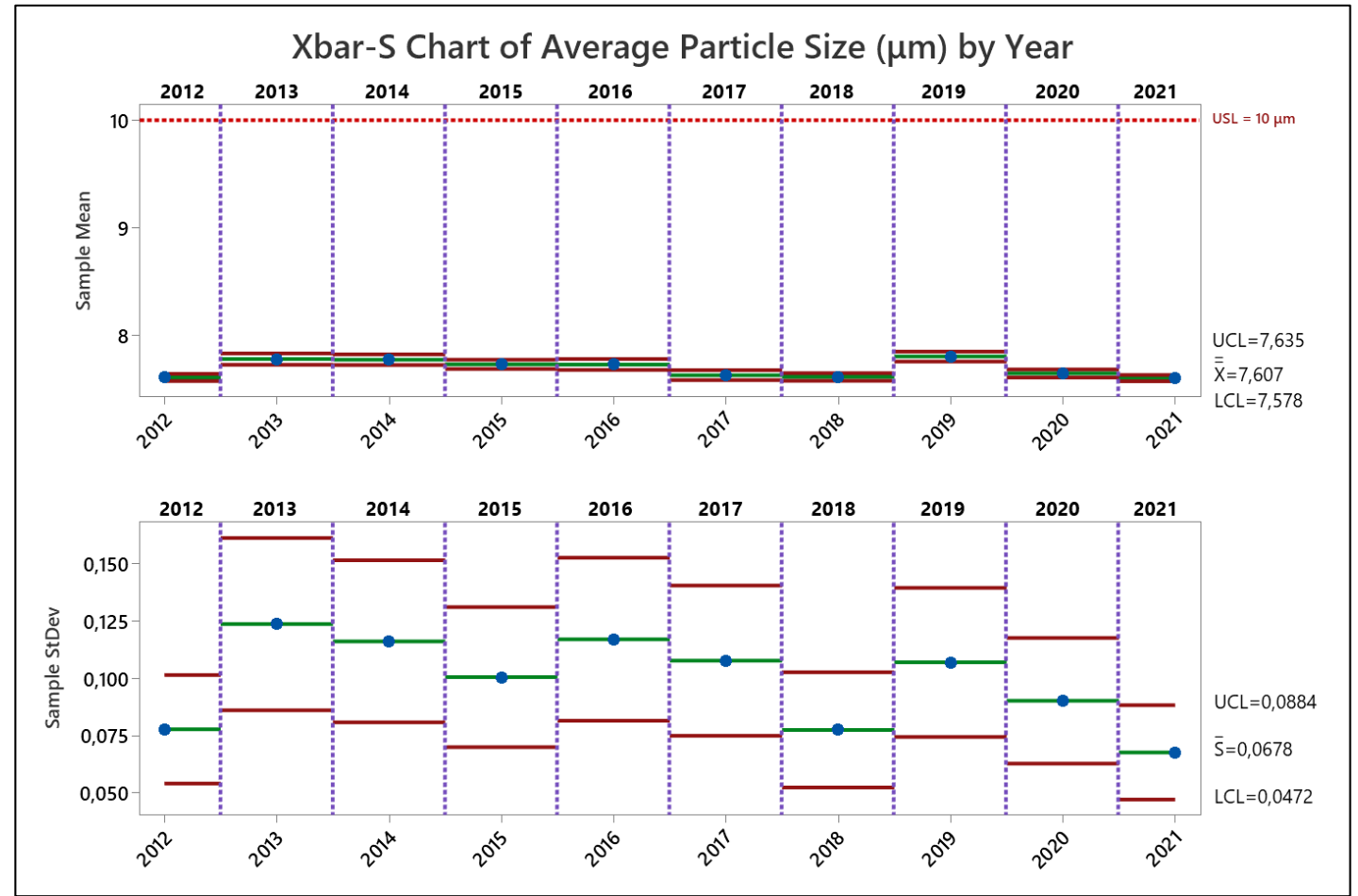


CAPABILITY ANALYSIS

Example 2 (cont.)

The control chart shows no variability outside the control limits calculated on the Average Standard Deviation and the whole process is well below the specification limit.

Let's now proceed with the Capability Analysis



CAPABILITY ANALYSIS

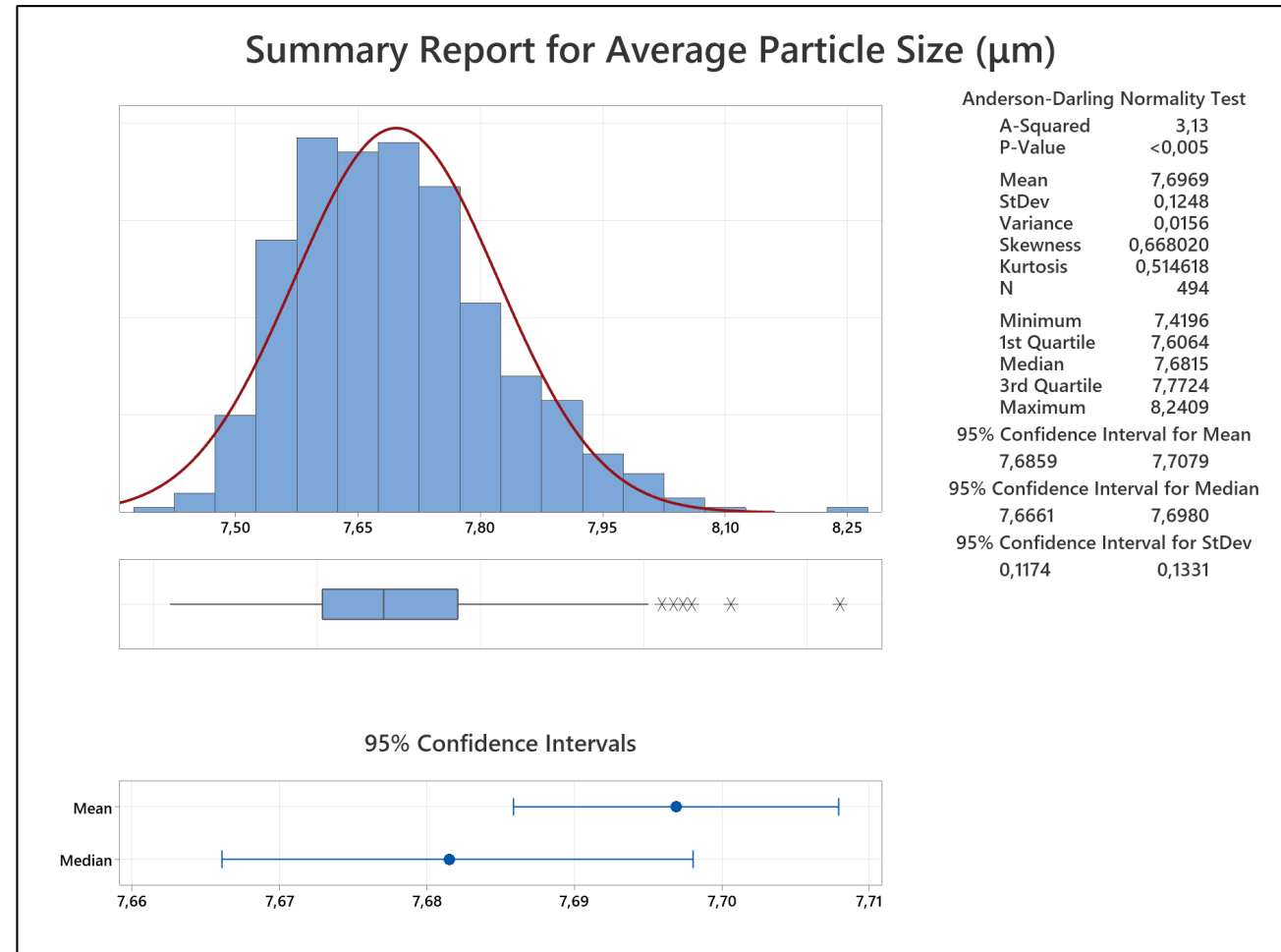
Example 2 (cont.)

Let's first verify the normality of the data.

Data are not normally distributed !

What can be done? There are two possibilities:

1. transform the data by normalizing them
2. evaluate the capability based on a Non-normal distribution



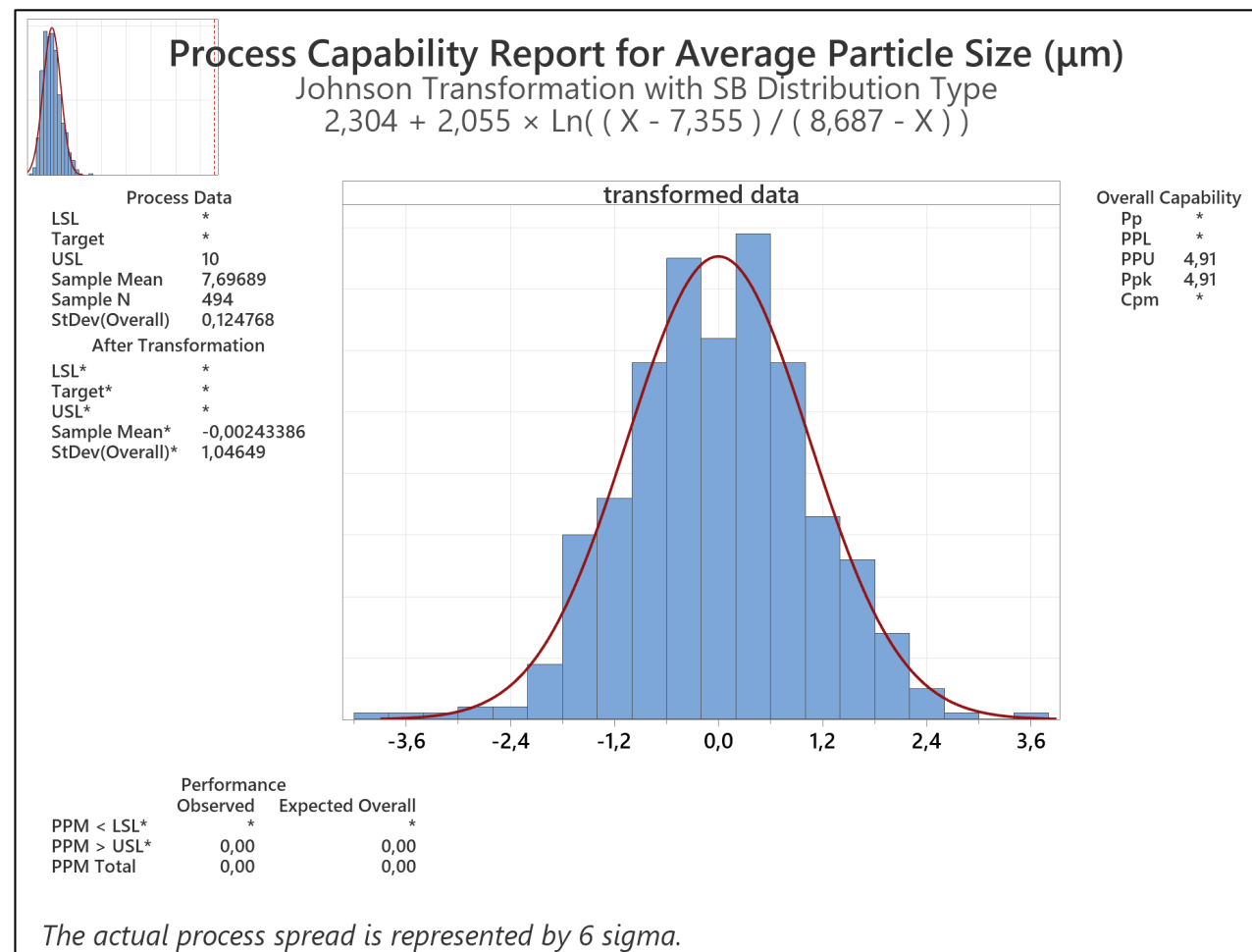
CAPABILITY ANALYSIS

Example 2 (cont.)

Transform the data by normalizing them

* NOTE *

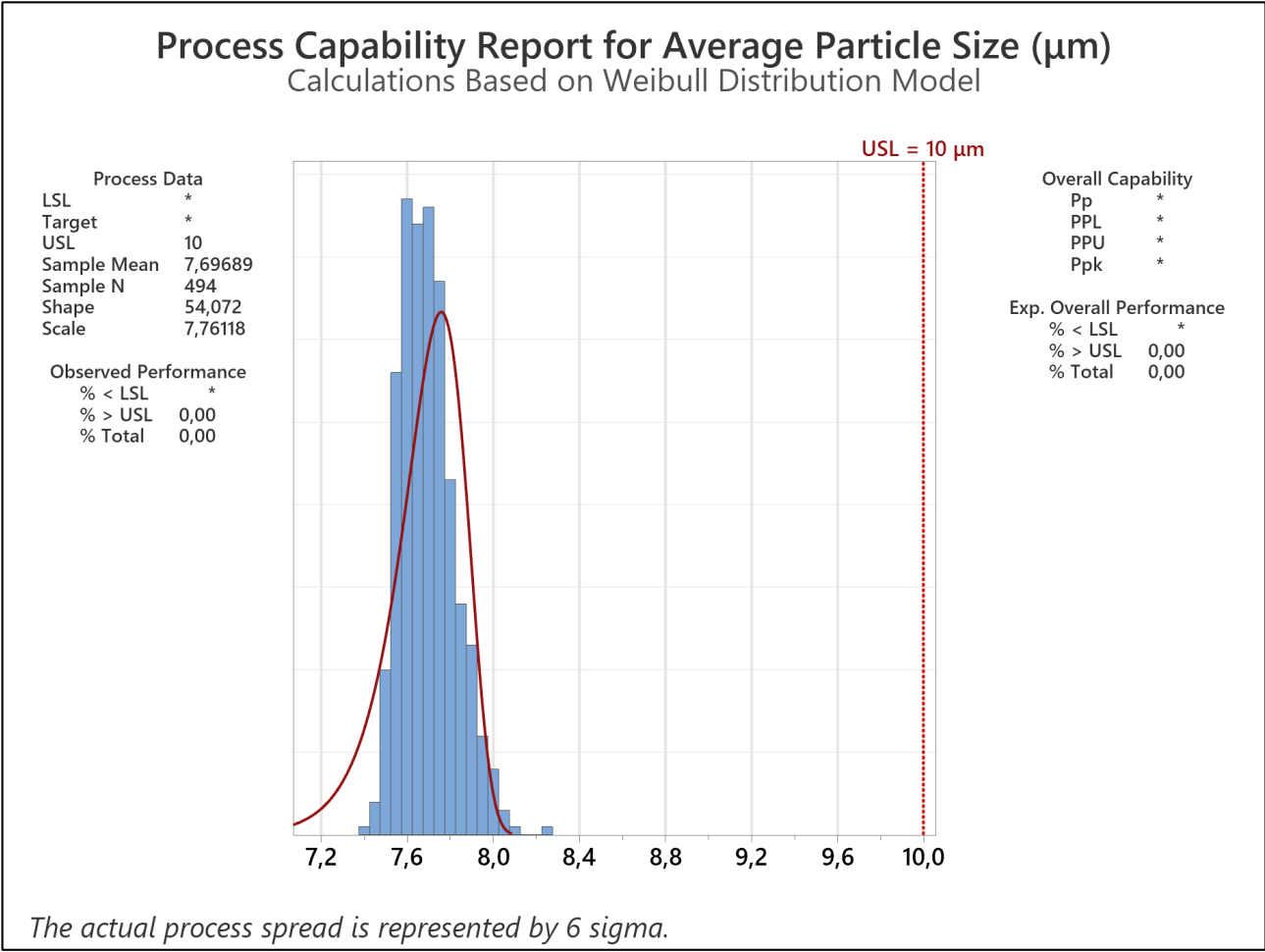
In this case the specification limit or target for Average Particle Size (μm) outside range of transformation function.



CAPABILITY ANALYSIS

Example 2 (cont.)

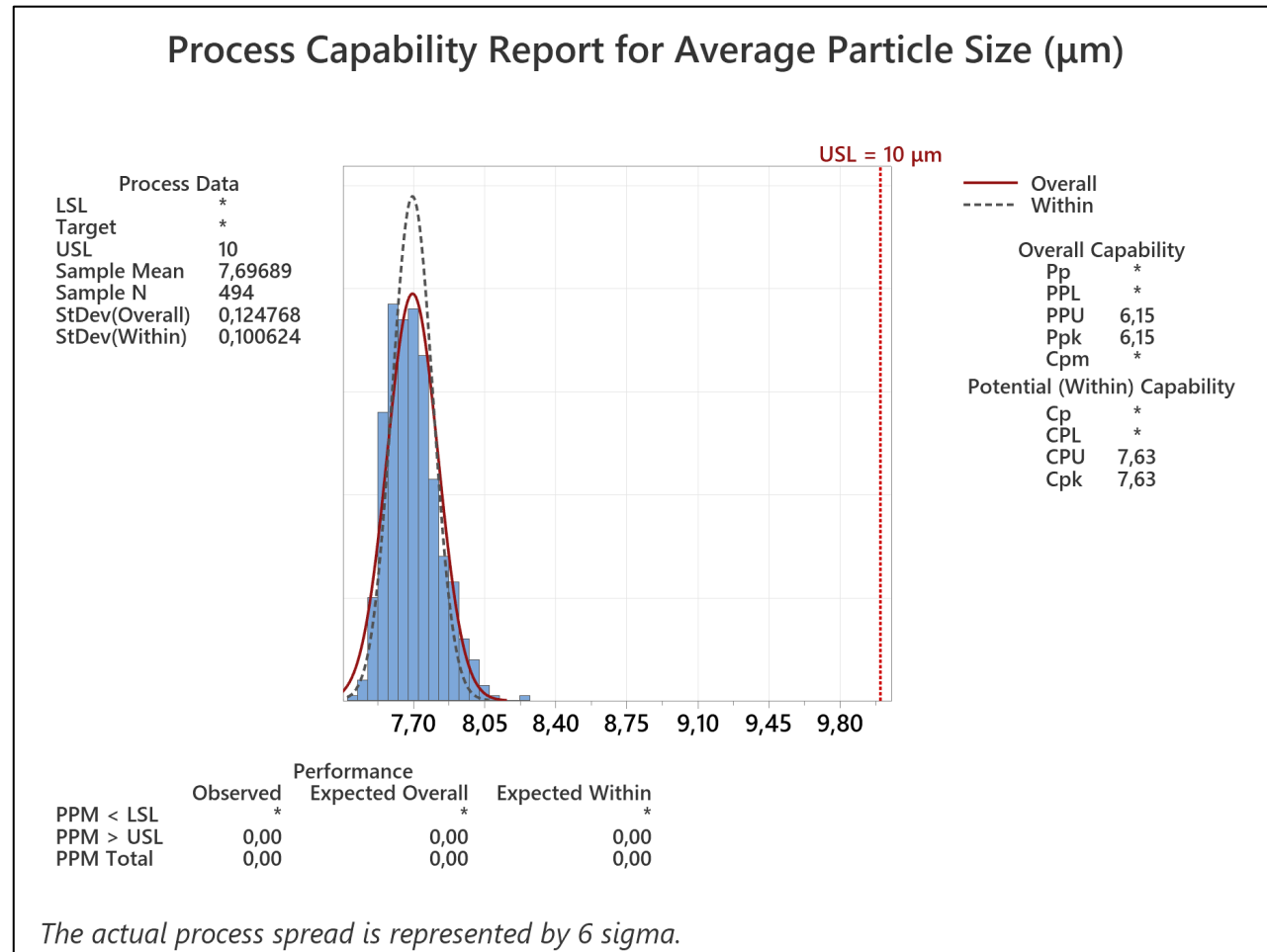
*Evaluate the capability
based on a Non-normal
distribution*



CAPABILITY ANALYSIS

Example 2 (cont.)

And what would have happened if we had assumed the data as normally distributed?



CAPABILITY ANALYSIS

Summarizing :

- *Capability Analysis allows to verify if a certain process, despite its variability, is able to respect the specified specification limits.*
- Once a *process is in statistical control* (remember *there is no capability without stability !*), the measure of quality (or metric) can be usefully expressed with the capability indices.
- The capability indices *C_p* and *C_{pk}* are dimensionless indices and therefore can be used to compare the capabilities of two processes with each other.
- The Cost of Poor Quality (COPQ) can be estimated from the ppm resulting from capability analysis.

CAPABILITY ANALYSIS

Process Capability Analysis is:

- performed on existing machines to assign them to the activities for which they are most suitable
- performed on new machines on the market to select them on the basis of a specific level of performance
- performed on new equipment as part of the qualification and approval process
- performed on existing processes to establish a baseline of current operations
- done periodically to monitor “wear and tear” on equipment and deterioration/drift of a process for whatever reason (material, personnel, environment, *etc.*)

M.L. George et al., The Lean Six Sigma Pocket Toolbook – McGraw-Hill (2005)

CAPABILITY ANALYSIS

- *Capability Indices are useful process metrics !*
- Given their nature of *summary indices* they have similarities with the classic *summary indices* of Descriptive Statistics (position, variability, shape)

SYNOPTIC TABLE

If you want to....	Then you can use...
Visualize shape, central tendency, and dispersion of continuous data	Histogram
Compare the central tendency and dispersion of various data sets	Box plot
Set priorities in the order in which you want to deal with topics	Pareto's chart
Graph the relationship between two variables	Scatterplot or Dispersion Diagram
Identify the presence of common and / or special causes in a process	Control Chart
Look for patterns in data	Run Chart
Understand about measures of central tendency, dispersion and shape of the data	Descriptive statistics
Evaluate the normality of the data	Anderson-Darling's normality test
Identify the variation caused by measuring process, measuring device and operators	Measurement Systems Analysis
Compare the variability associated with the process with the specification limits	Process Capability Analysis
Compare an average value to a single one (target)	One-sample t-test
Compare the means of two data sets	Two-sample t-test
Compare the means of more than two data sets	ANOVA
Compare a median to a single value (target)	One-sample sign test
Compare the medians of two data sets	Two-sample Mann-Whitney test
Compare the medians of two or more data sets	Kruskal-Wallis test
Compare the variances of two data sets	F-test
Compare the variances of three or more data sets	Bartlett test
Compare the variances of non-normal data	Levene test

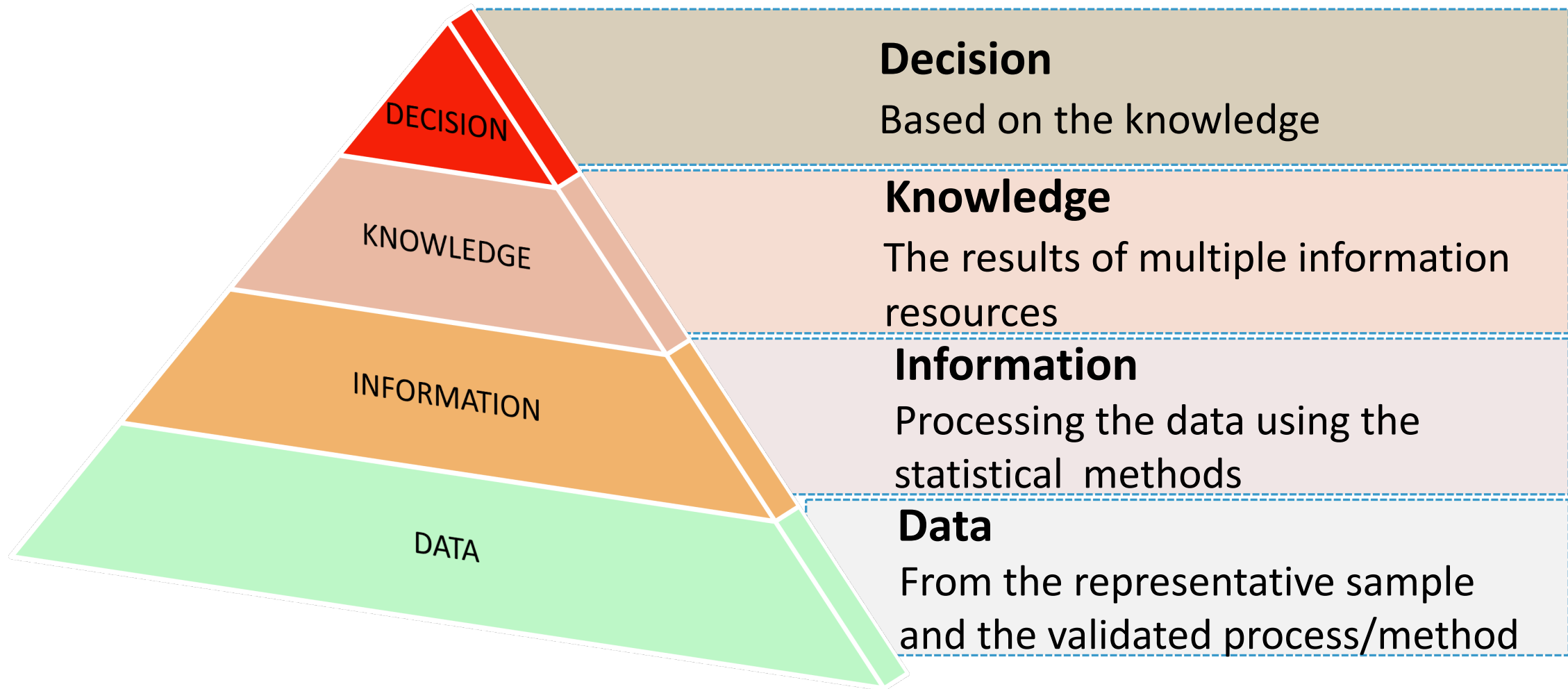
If you want to....	Then you can use...
Evaluate the relationship between an independent variable (x) and a dependent variable (y)	Simple Linear Regression
Evaluate the relationship between two or more independent variables (x1, x2, x3, ...) and a dependent variable (y)	Multiple Linear Regression
Identify the input variables that impact the output variables	Design of Experiment (DoE)
Identify sources of process variability for variable data (subgroup size = 1)	<i>IM-R Chart</i>
Identify sources of process variability for variable data (subgroup dimensions: 2 to 6)	<i>Xbar-R Chart</i>
Identify sources of process variability for variable data (subgroup size: greater than 6)	<i>Xbar-S Chart</i>
Identify the sources of variability for "attributes" data as percentage defects	<i>p Chart</i>
Identify the sources of variability for "attributes" data as number of defects	<i>np Chart</i>
Identify the sources of variability for "attributes" data as number of defects per subgroup	<i>c Chart</i>
Identify the sources of variability for "attributes" data as number of defects per unit	<i>u Chart</i>

CONCLUSIONS

FROM DATA TO AN INFORMED DECISION

Clearly, the purpose of everything seen so far is not the data itself, but the *data as a means of reaching an informed and conscious decision* not based on speculation or conjecture.

FROM DATA TO AN INFORMED DECISION



STATISTICS & GMP

January 2011

*FDA Guidance for Industry on Process Validation:
General Principles and Practices*

STATISTICS & GMP

- ❖ *The term “statistical” occurs 13 times !*
- ❖ « Criteria and process performance indicators should include a description of the *statistical methods to be used in analyzing all collected data* (e.g., *statistical metrics defining both intra-batch and inter-batch variability*) »
- ❖ « An ongoing program to collect and analyze product and process data that relate to product quality must be established (§ 211.180(e)) »

STATISTICS & GMP

- ❖ « The *data should be statistically trended and reviewed by trained personnel*. The information collected should verify that the quality attributes are being appropriately controlled throughout the process »
- ❖ « *We recommend that a statistician or person with adequate training in statistical process control techniques develop the data collection plan and statistical methods and procedures used in measuring and evaluating process stability and process capability* »

FDA Guidance for Industry – Process Validation: General Principles and Practices (January 2011)

STATISTICS & GMP

- ❖ « Many tools and techniques, some statistical and others more qualitative, can be used to detect variation, characterize it, and determine the root cause. *We recommend that the manufacturer use quantitative, statistical methods whenever appropriate and feasible.* »
- ❖ « We recommend continued monitoring Monitoring can then be adjusted to a statistically appropriate and representative level. »
- ❖ « Capability of a process: Ability of a process to produce a product that will fulfill the requirements of that product. *The concept of process capability can also be defined in statistical terms.* »

FDA Guidance for Industry – Process Validation: General Principles and Practices (January 2011)

STATISTICS & GMP

TEN YEARS LATER July 2021

EMA Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development

STATISTICS & GMP

- ❖ « This reflection paper identifies specific areas where the quantitative comparative evaluation of drug product quality characteristics plays an important role from the regulatory perspective »
- ❖ « The document focusses on methodological aspects in relation to [statistical data comparison](#) approaches [for pre- and post-manufacturing changes](#) »
- ❖ « The reflection paper ... addresses questions related to comparison objectives, sampling strategies, sources of variability and options (or limitations) for statistical inference »

EMA Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development (July 2021)

STATISTICS & GMP

- ❖ « the goal of this paper is to reflect under which circumstances, and to what extent, the implementation of inferential statistical methodology can assist or even facilitate comparative evaluation of QA data »
- ❖ etc.

The « heroic age » of the classic « average $\pm 3\sigma$ » diagram and similar old stuff are coming to an end !

We must be ready for new challenges ! We need new eyes !

Thank you for the attention !

<http://riccardobonfichi.it>

REFERENCES (*just a few*)

D. C. Montgomery, *Statistical Quality Control: A Modern Introduction*, 7th Edition, Wiley (2013)

W.A. Shewhart, *Economic Control of Quality of Manufactured Product*, Van Nostrand (1931)

W.E. Deming, *Out of the Crisis*, MIT Press (2000)

Q. Brook, *Lean Six Sigma & Minitab*, 6th Ed., OPEX Resources (2020)

Technical Report No. 59, *Utilization of Statistical Method for Production Monitoring*, PDA (2012)

Technical Report No. 60, *Process Validation: A Lifecycle Approach*, PDA (2013)

Guidance for Industry, *Process Validation: General Principles and Practices*, FDA (2011)

ICH Q10, *Pharmaceutical Quality System*, Step 4 (2008)