QUALITY METRICS AND DATA CONSISTENCY

- Part 2 -

INDEX

- **CASE STUDY 3** : Capability Analysis: short term metrics (Cp and Cpk) and long term metrics (Pp and Ppk)

- **CASE STUDIES 4/5**: Probabilistic methods for a quick evaluation of the manufacturing process (Standardized Normal distribution, Poisson and Binomial distributions)

- **CASE STUDY 6**:
  - Processes non-normally distributed: impurities content, microbial counts, Particle Size Distributions (PSD), black particles (or *black specs*)
  - Normalization of non-normal data using mathematical transformations (logarithm, square root, reverse or reciprocal)
  - Johnson Transformations

INDEX *(cont.)*

- **CASE STUDY 7**: Multivariate methods : a different way to look at Quality Control data !

- **CONCLUSIONS**: Quality Metrics are ease of use quantitative indicators that allow to intercept the variability of products / processes, quantify it and therefore ensure Quality.

  Quality Metrics provide therefore a “quantitative knowledge” of the process that:
  - allows to manage anomalous events (OOT, OOS, deviations, *etc*.)
  - communicate awareness in what is done and reliability in the processes used.

  All this is summed up in two words: **ECONOMIC ADVANTAGE!**

CASE STUDY 3

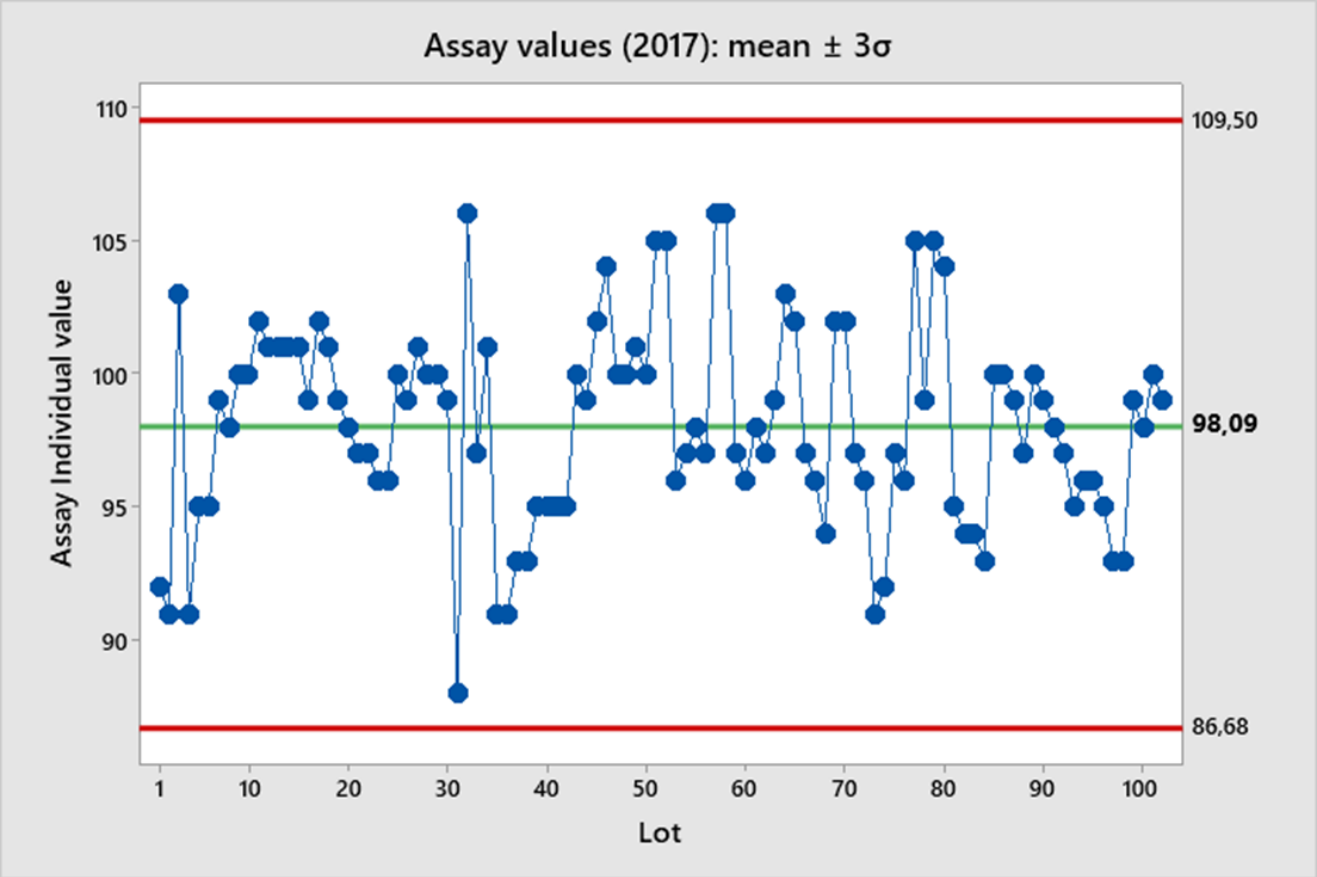
- Capability Analysis: metrics for stable /mature processes (Cp and Cpk) and metrics for new processes (Pp and Ppk)

CASE STUDY 3 – Capability Analysis

Let come back to the example seen in CASE STUDY 1.

For the sake of clarity, in this and the following two slides are summarized the key points.

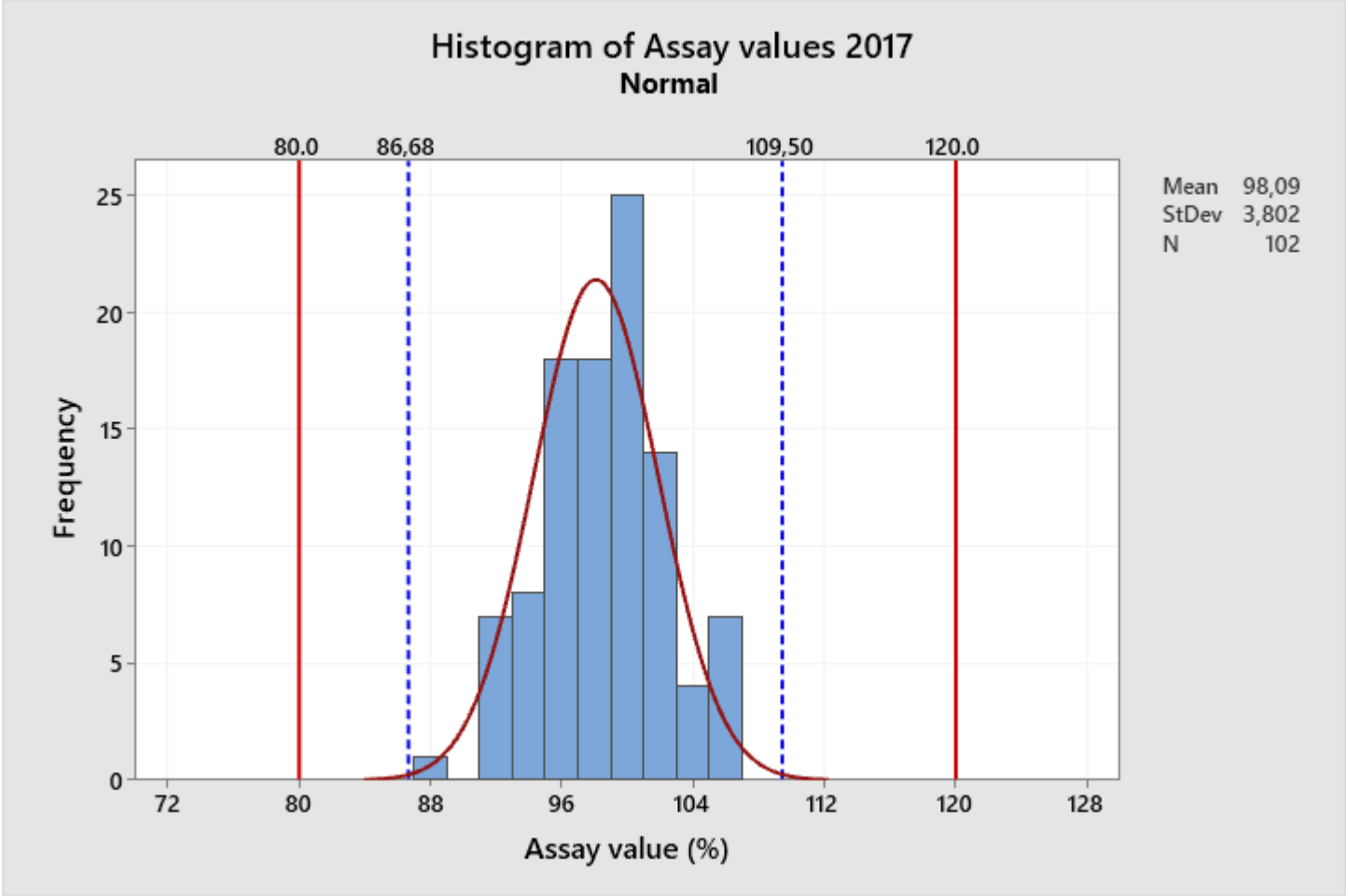
Here is the conventional plot reporting « average ± 3σ » for the HPLC assay values of the 102 lots of an API manufactured in 2017.



CASE STUDY 3 – Capability Analysis

Here, on the right, is the histogram that shows 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***.



CASE STUDY 3 – Capability Analysis

Here, on the side, is the *I-MR Chart* with a

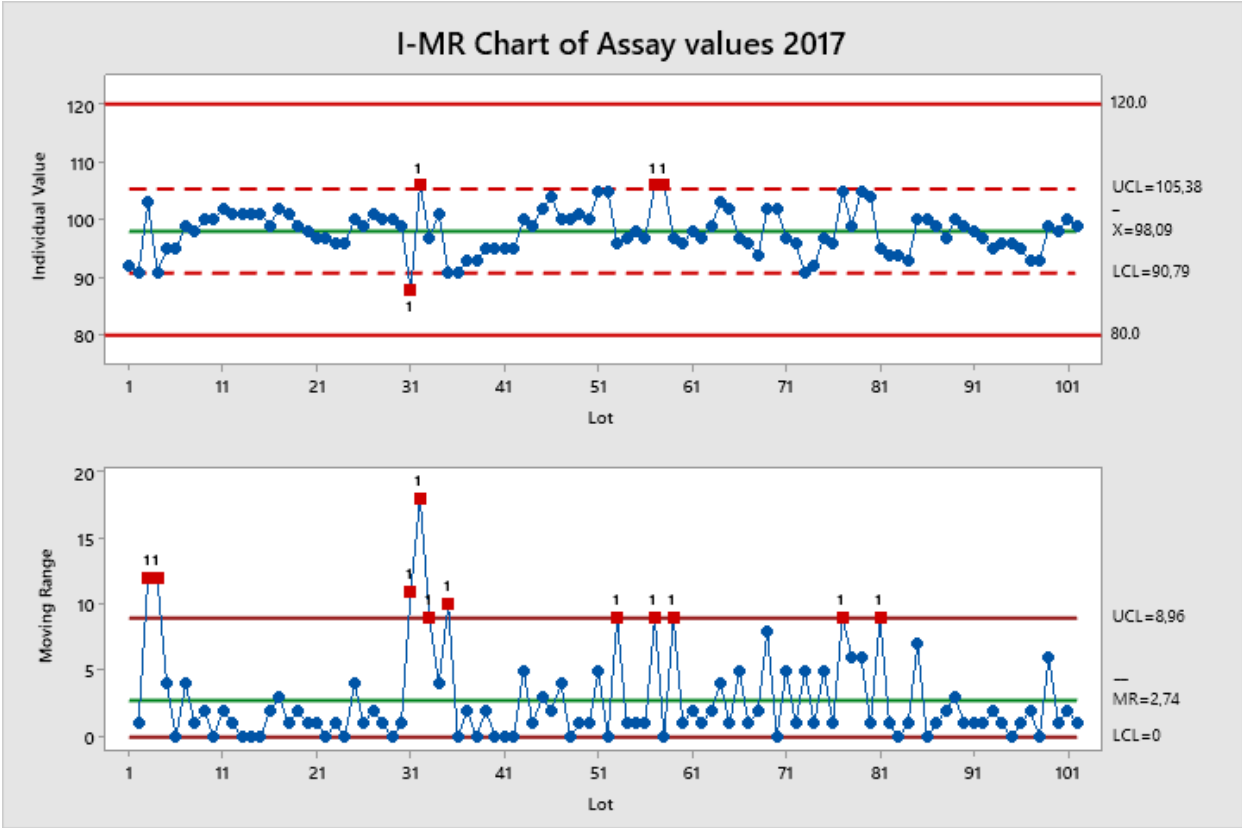
mR = 2.

This chart provides information on the:

- **variation inherent to the process** known as *process spread* or *voice of the process*

and

- **variation allowed by the Customer** known as *process specifications* or *voice of the Customer.*

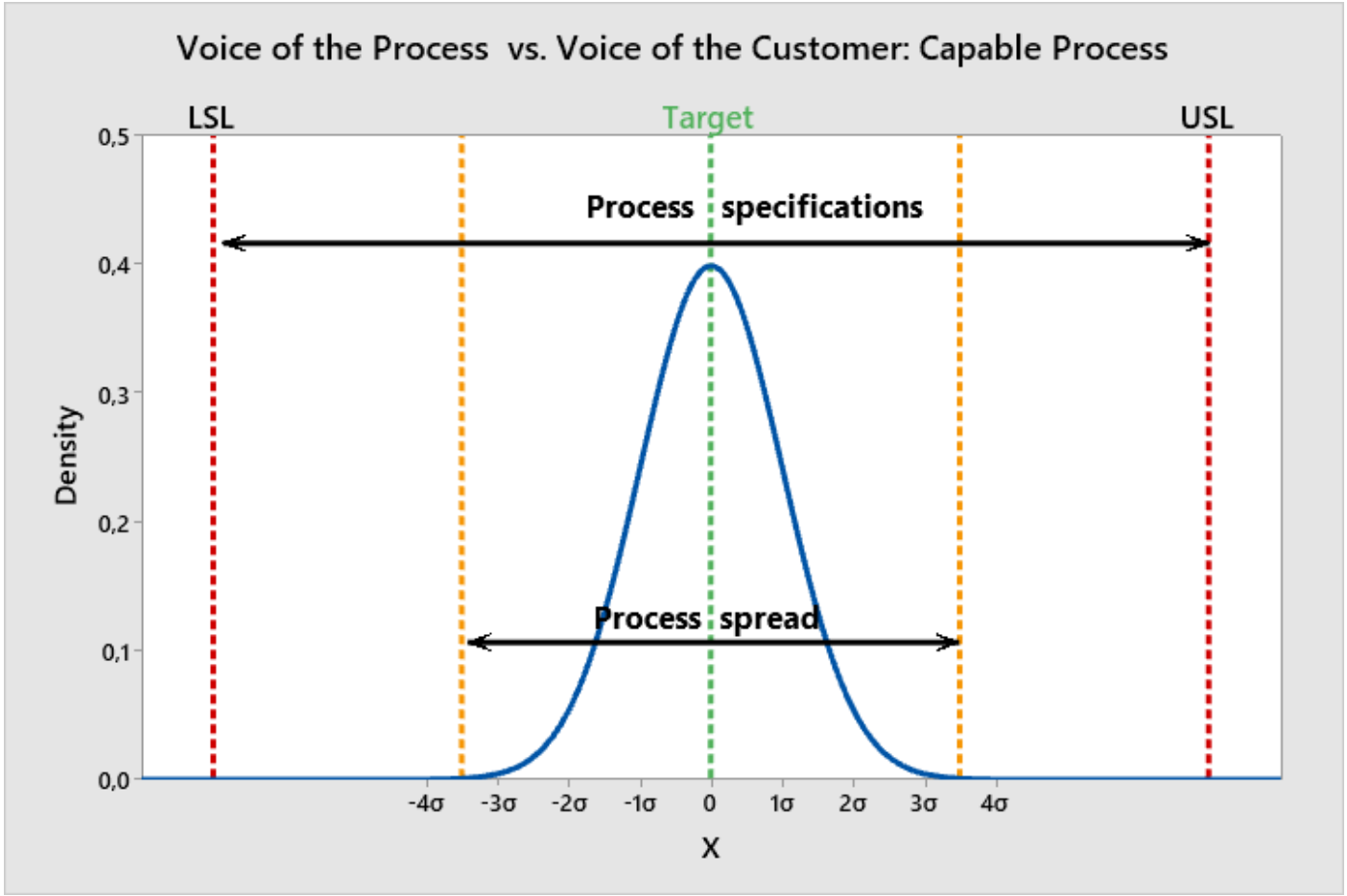


Note: *I-MR cards are generally used when it is difficult or impossible to measure in subgroups. This occurs when measurements are expensive or destructive, low production volumes of products or products have a very long or continuous cycle time.*

CASE STUDY 3 – Capability Analysis

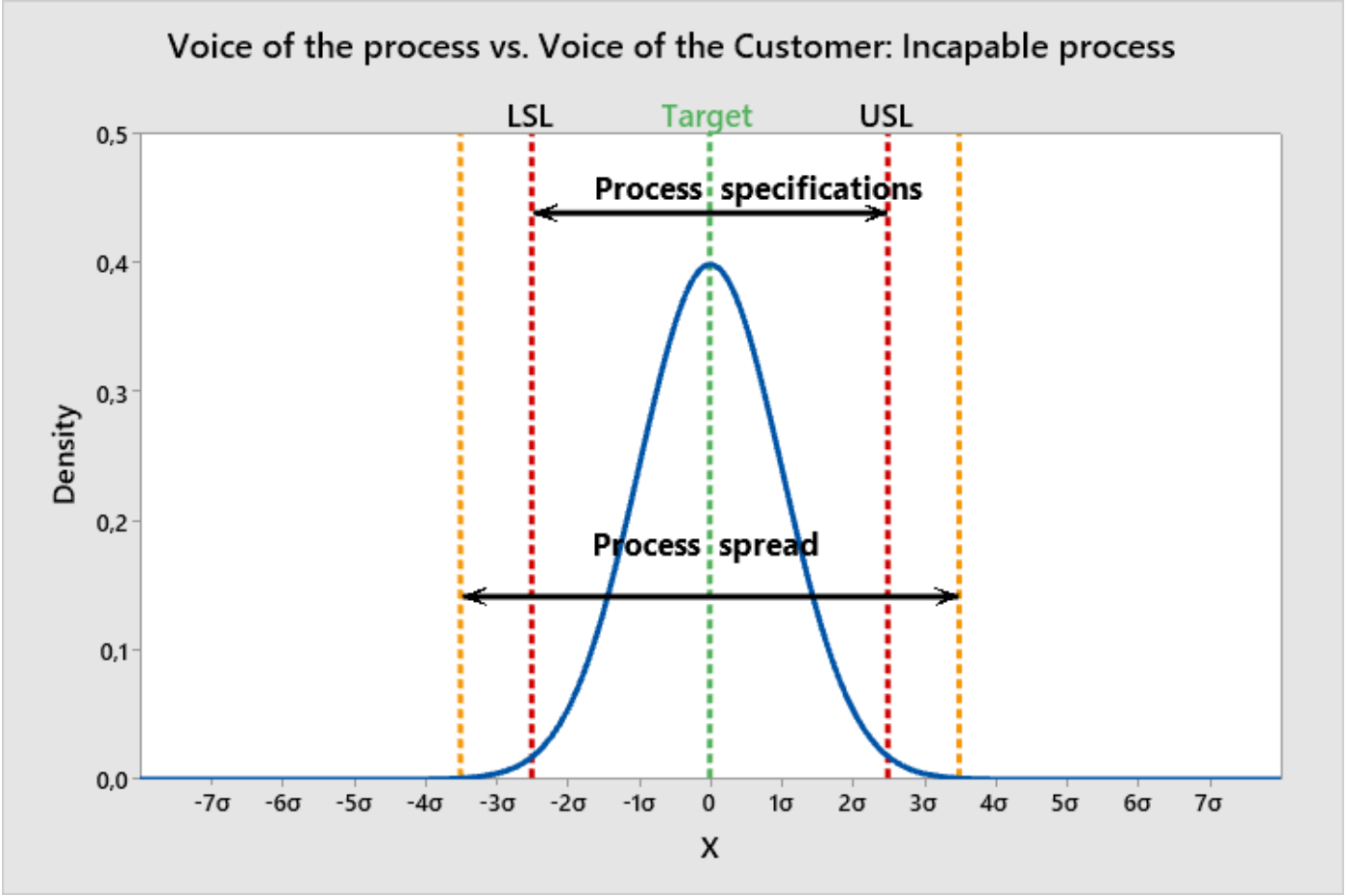
As long as the *process spread* (measured by the standard deviation, σ) lies within the process specifications, the process is said ***capable*** *of delivering the quality required by the Customer*.

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



CASE STUDY 3 – 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*.



CASE STUDY 3 – 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 view 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?

there is a need of more efficient indicators !

CASE STUDY 3 – Capability Analysis

Capability Indices

- ***Cp*** or ***potential capability*** is defined as $Cp = \frac{USL - LSL}{6\sigma}$ and it measures the ratio between the *admissible dispersion for the process* (difference between the tolerance limits) and the *natural tolerance* (6σ). 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 !

CASE STUDY 3 – Capability Analysis

Capability Indices

- if ***Cp* = 1 ⇨ 0.27%** of the observations do not conform to the specifications (± 3σ)
- if ***Cp* = 1.33 ⇨ 0.0064%** of the observations do not conform to the specifications (± 4σ)
- if ***Cp* = 1.67 ⇨ 0.000057%** of the observations do not conform to the specifications (± 5σ)

As general indication:

- if ***Cp* ≥ 1.33** the process can be considered ***satisfactory***
- if **1.00 ≤ *Cp* < 1.33** the process can be considered ***adequate***
- if ***Cp* < 1.00** the process is ***inadequate***

CASE STUDY 3 – Capability Analysis

Capability Indices

- ***Cpk*** or ***real capability*** is defined as: min {(USL - μ)/3σ ; (μ - LSL)/3σ } or min {CPU ; CPL}

- ***Cpk*** also considers the *position* of the process with respect to the tolerance limits.

- if ***Cpk* > 1** : data are within *tolerance limits*
- if **0 < *Cpk* < 1** : part of the observations lie beyond the tolerance limits
- if ***Cpk* < 0** : data, on the average, are out of specifications
- if ***Cpk* = 1** : 99.73% of the observations are within the tolerance limits (*i.e.*, only 3 observations on 1000 are rejected)

CASE STUDY 3 – Capability Analysis

Capability Indices

In terms typical of the Quality Control:

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

CASE STUDY 3 – Capability Analysis

Capability Indices

- 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 specification.

CASE STUDY 3
Capability Analysis

Cp and Cpk: A Summary

Mature / Stable processes

	Cp (Process Spread)	**Cpk (Process Centering)**
Cp *vs.* Cpk	**Cp** is an index that **predicts** either any **mature process** can meet the specifications or not. It is assumed that the process is already under statistical control (*i.e.*, stable)	**Cpk** is an index that **predicts** how close to the specification limits is the process mean of any **mature process.** It is assumed that the process is already under statistical control (*i.e.*, stable)
	Cp is used to predict the **capability with respect to process variation** of a mature process, already under statistical control (*i.e.*, stable)	**Cp** is used to predict the **capability with respect to process variation and centering** of a mature process, already under statistical control (*i.e.*, stable)
	Use Cp and Cpk only once the **PROCESS IS ALREADY MATURE AND STABLE ENOUGH** (*i.e.*, in a state of STATISTICAL CONTROL)	
	Cp = (USL – LSL) /6σ USL = Upper Specification Limit LSL = Lower Specification Limit σ = Process Standard Deviation = $\overline{R}$ / d_2 or $\overline{S}$ / C_4 = Process Spread = Voice of the Process	**Cpk = min (CPU, CPL)** CPU = (USL - μ) / 3σ CPL = (μ - LSL) / 3σ σ = **Process** Standard Deviation μ = Arithmetic mean
	Cp and Cpk apply to sub-grouped data referring to different shifts, machines, operators, *etc.*	

CASE STUDY 3
Capability Analysis

Pp and Ppk: A Summary

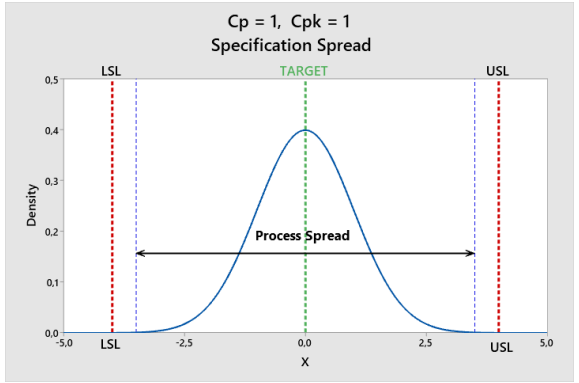
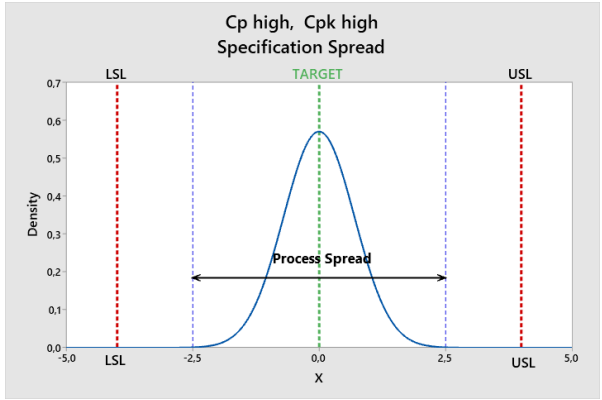
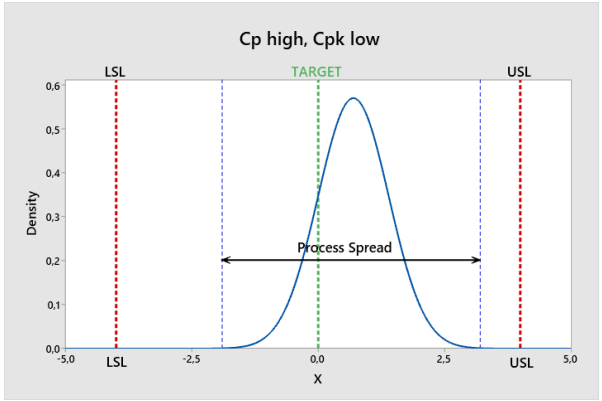
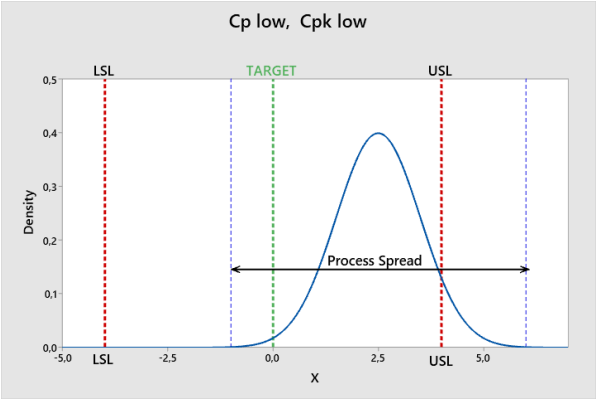
New processes

Pp vs. Ppk

	Pp (Process Spread)	**Ppk (Process Centering)**
	Pp is an index that **verifies** either any **new process** can meet the specifications or not. The process might not be under statistical control ⇨ **piloting**	**Ppk** is an index that **verifies** how close to the specification limits is the process mean of any **new process.** The process might not be under statistical control ⇨ **piloting**
	Pp is used to check the **capability with respect to process variation** of a **new process**	**Ppk** is used to check the **capability with respect to process variation and centering** of a new process
	Use Pp and Ppk only once you are **INITIALLY SETTING UP A NEW PROCESS**	
	Pp = (USL – LSL) /6σ USL = Upper Specification Limit LSL = Lower Specification Limit USL-LSL = Specification Spread Voice of the Customer σ = **Sample** Standard Deviation $= \frac{\sqrt{\sum(x_i - X)^2}}{\sqrt{n-1}}$	**Ppk = min (PPU, PPL)** PPU = (USL - μ) / 3σ PPL = (μ - LSL) / 3σ σ = **Sample** Standard Deviation μ = Arithmetic mean
	Pp and Ppk apply to the overall process and do not require any data grouping	

CASE STUDY 3
Capability Analysis

Graphical Summary

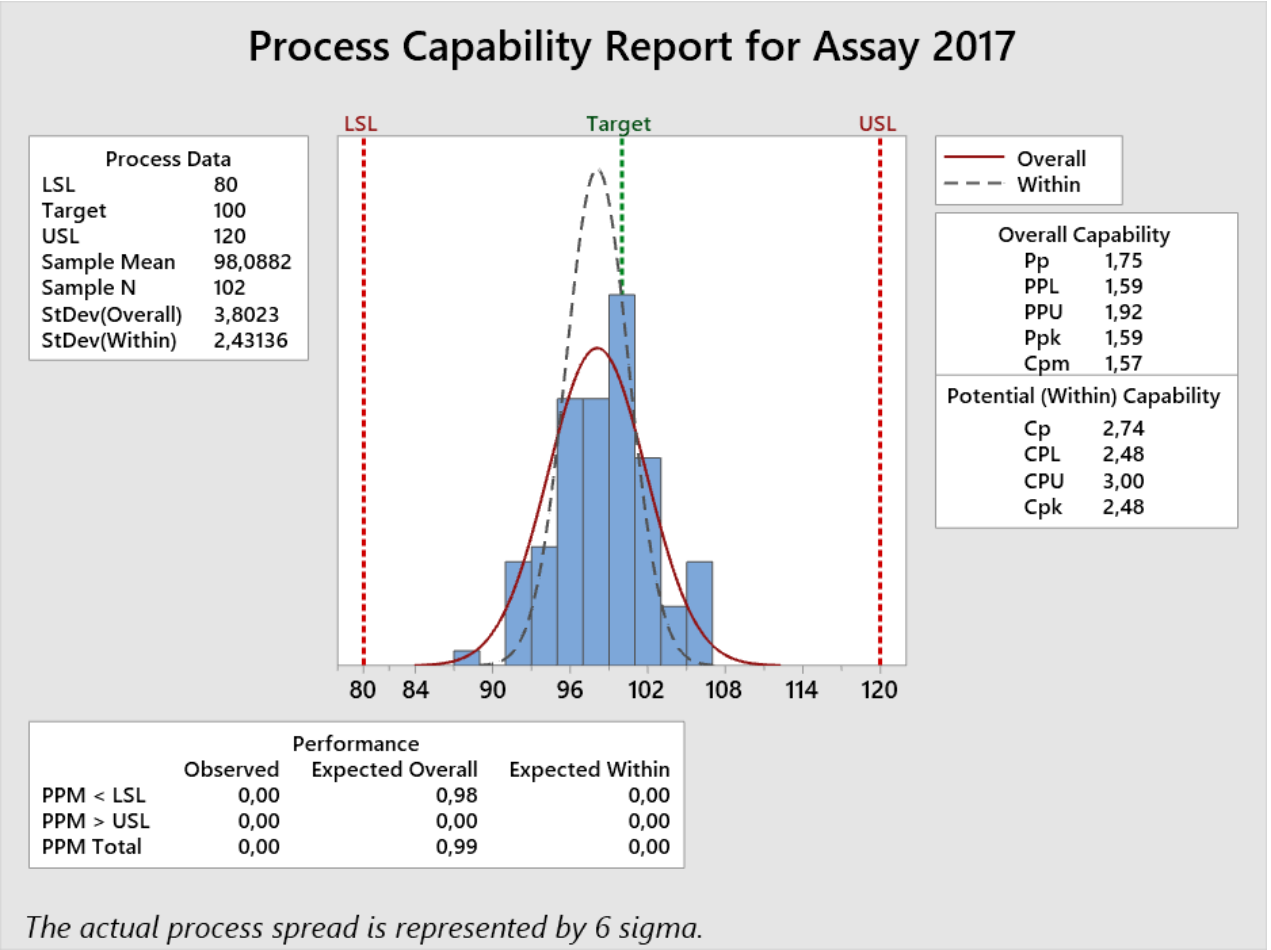
	Cp and Pp (Process Spread)	Cpk and Ppk (Process Centering)
Cp, Pp *vs.* Cpk, Ppk		
	Cp and Pp measure the constancy around the average performance	Cpk and Ppk measure the constancy around the average performance and the closeness to target
		

CASE STUDY 3 – Capability Analysis

Let's now consider our initial process.

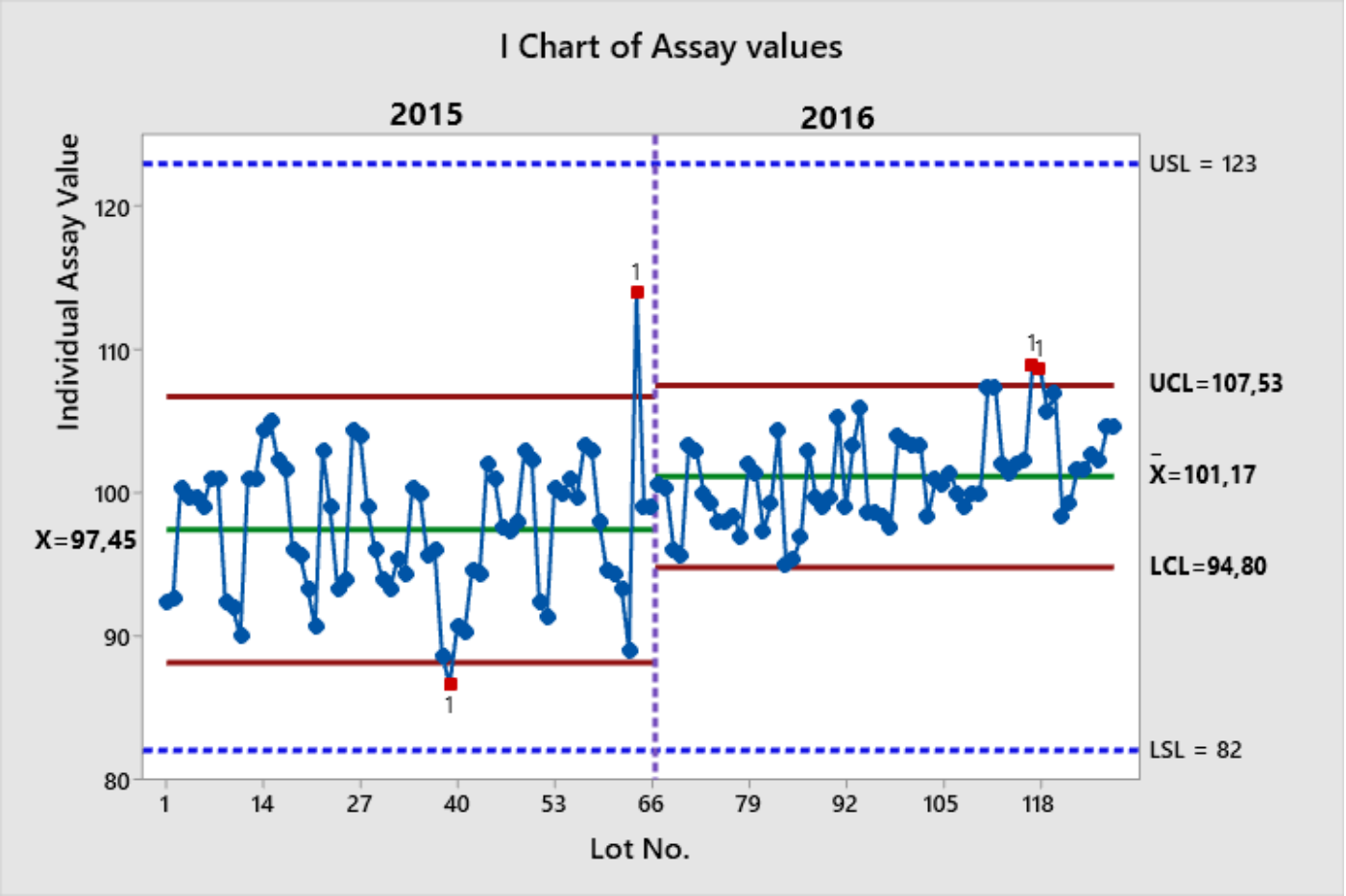
As expected, *Cp* > *Cpk* (in fact 2.74 > 2.48), but it deals of excellent values anyway. *The difference is due to the fact that the process is not well centered on target.*

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



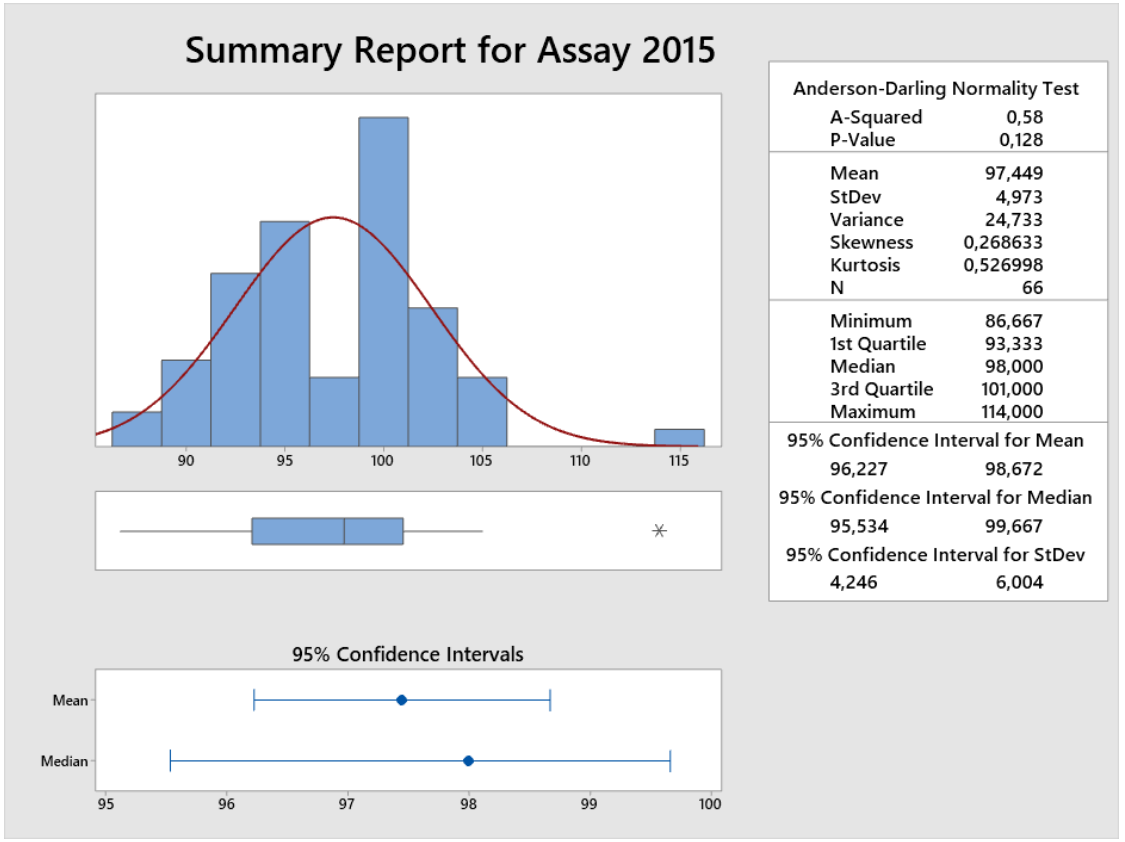
CASE STUDY 3 – Capability Analysis

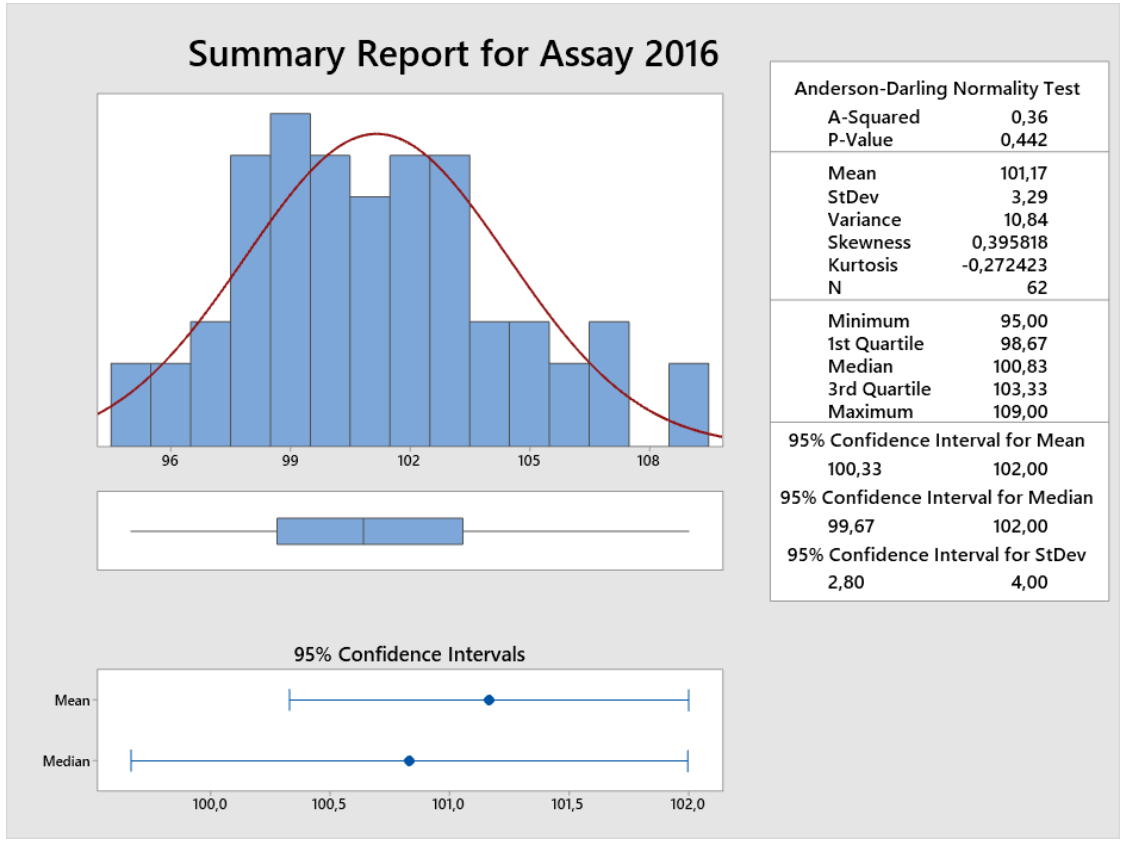
Here is an *I Chart* displaying the assay values pertinent to an API manufacturing process collected in two subsequent years.
Let's see how a Capability Analysis can be set up and what it reveals.



CASE STUDY 3 – Capability Analysis

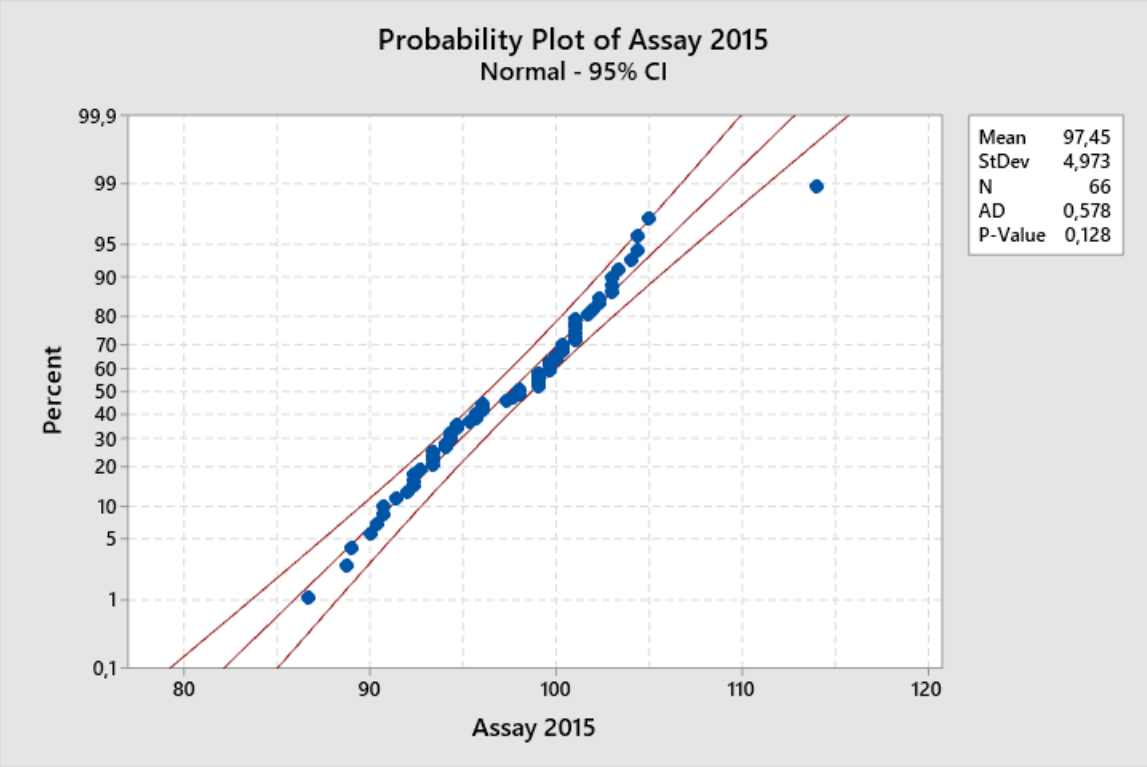
The first step is to investigate how data is distributed: *P-value* >0.05 ⇨ Normal distribution

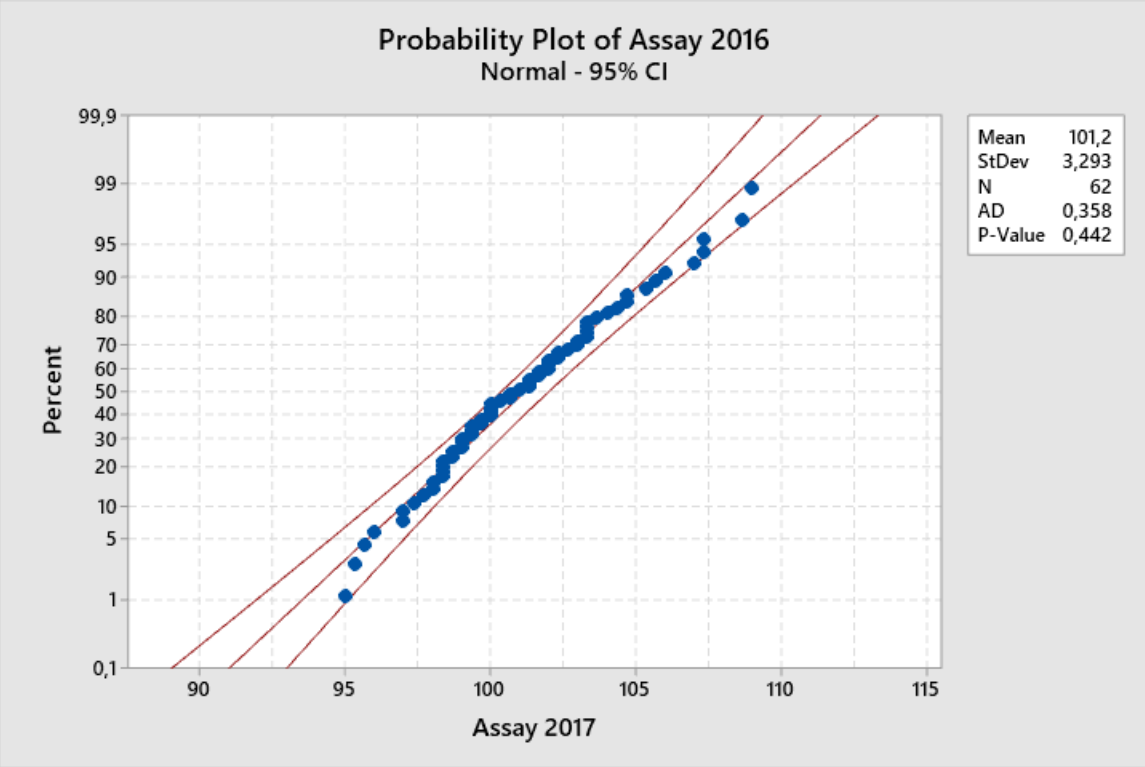




CASE STUDY 3 – Capability Analysis

The Probability plots here below that data is normally distributed in both cases.

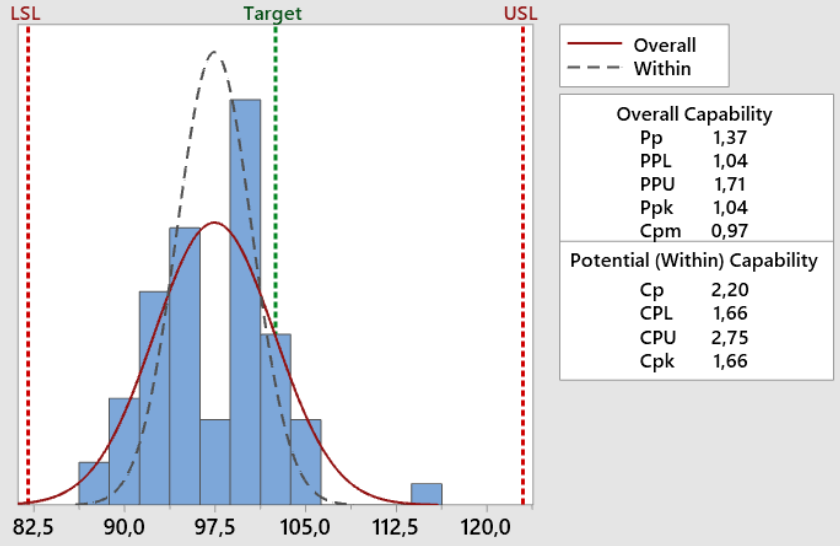




CASE STUDY 3 – Capability Analysis

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

Process Capability Report for Assay 2015



Process Data	
LSL	82
Target	102,5
USL	123
Sample Mean	97,4495
Sample N	66
StDev(Overall)	4,97326
StDev(Within)	3,10056

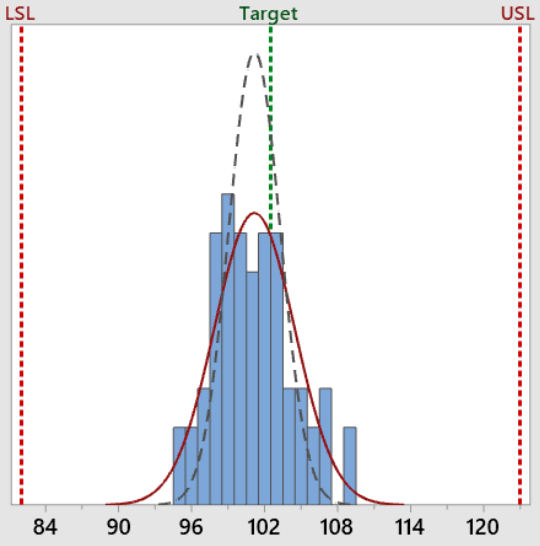
Overall Capability	
Pp	1,37
PPL	1,04
PPU	1,71
Ppk	1,04
Cpm	0,97

Potential (Within) Capability	
Cp	2,20
CPL	1,66
CPU	2,75
Cpk	1,66

Performance	Observed	Expected Overall	Expected Within
PPM < LSL	0,00	946,54	0,31
PPM > USL	0,00	0,14	0,00
PPM Total	0,00	946,68	0,31

The actual process spread is represented by 6 sigma.

Process Capability Report for Assay 2016



Process Data	
LSL	82
Target	102,5
USL	123
Sample Mean	101,167
Sample N	62
StDev(Overall)	3,29251
StDev(Within)	2,12185

Overall Capability	
Pp	2,08
PPL	1,94
PPU	2,21
Ppk	1,94
Cpm	1,94

Potential (Within) Capability	
Cp	3,22
CPL	3,01
CPU	3,43
Cpk	3,01

Performance	Observed	Expected Overall	Expected Within
PPM < LSL	0,00	0,00	0,00
PPM > USL	0,00	0,00	0,00
PPM Total	0,00	0,00	0,00

The actual process spread is represented by 6 sigma.

CASE STUDY 3 – CONCLUSION

- Capability Analysis allows to verify if a certain process, despite its variability, is able to respect the specified specification limits.
- Once a process is under 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 ***Cp*** and ***Cpk*** 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.

CASE STUDY 3 – CONCLUSION

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 based on 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)

CASE STUDY 3 – CONCLUSION

- ***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)
- In the next case studies we will instead see useful process metrics that are more "inferential in nature" than “descriptive”

CASE STUDIES 4 / 5

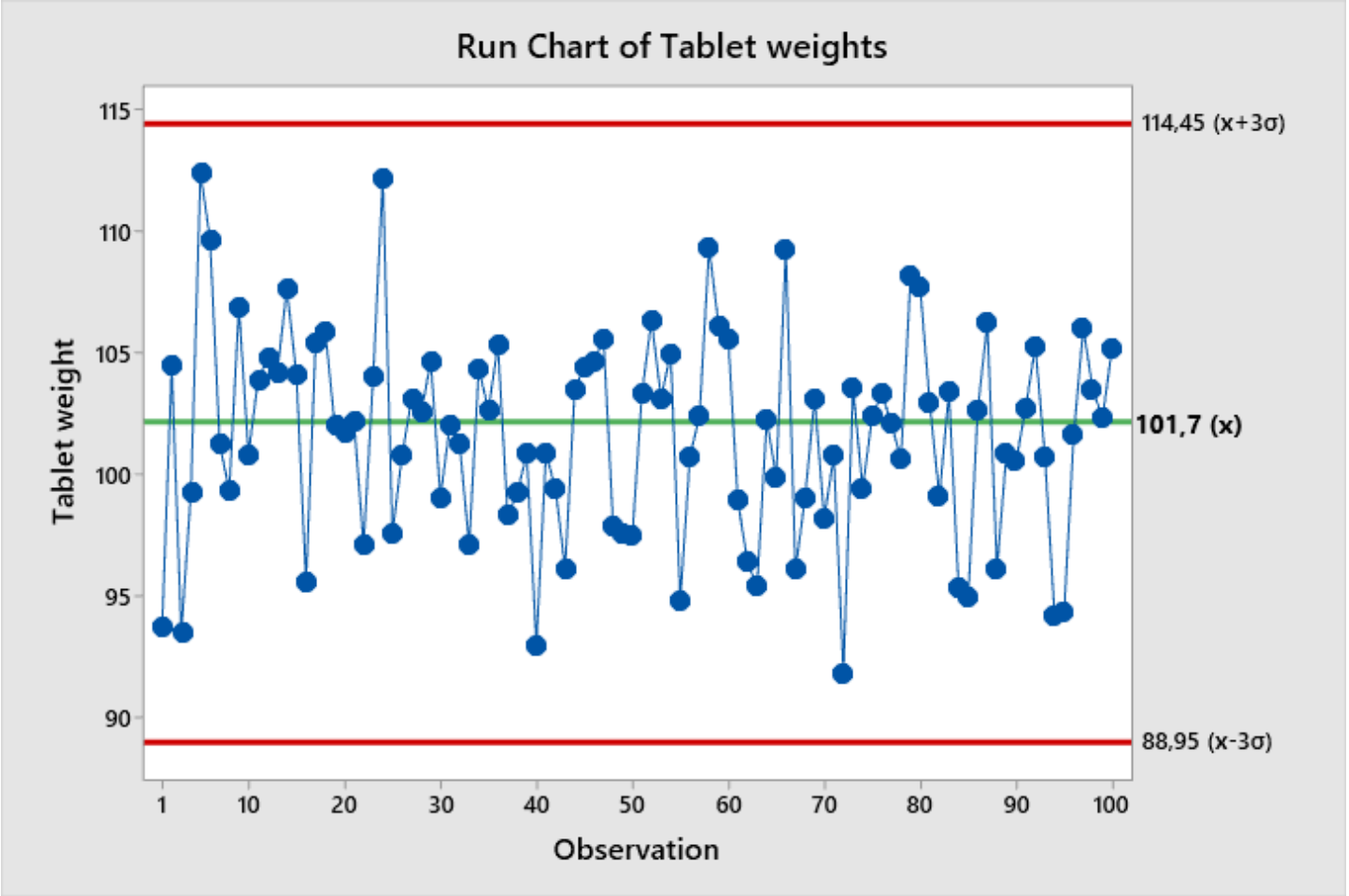
- Probabilistic methods for a quick evaluation of the manufacturing process (Standardized Normal distribution, Poisson and Binomial distributions)

CASE STUDY 4

During the production of a batch of tablets, 100 are sampled in-process, obtaining the weight trend shown here.

This indicates that the tablets have an average weight of 101.7 ± 4,249 mg.

Now, if the release test foresees that no more than 2 out of 20 tablets can exceed 10% of the average weight, can we already say *in-process* if the batch passes?

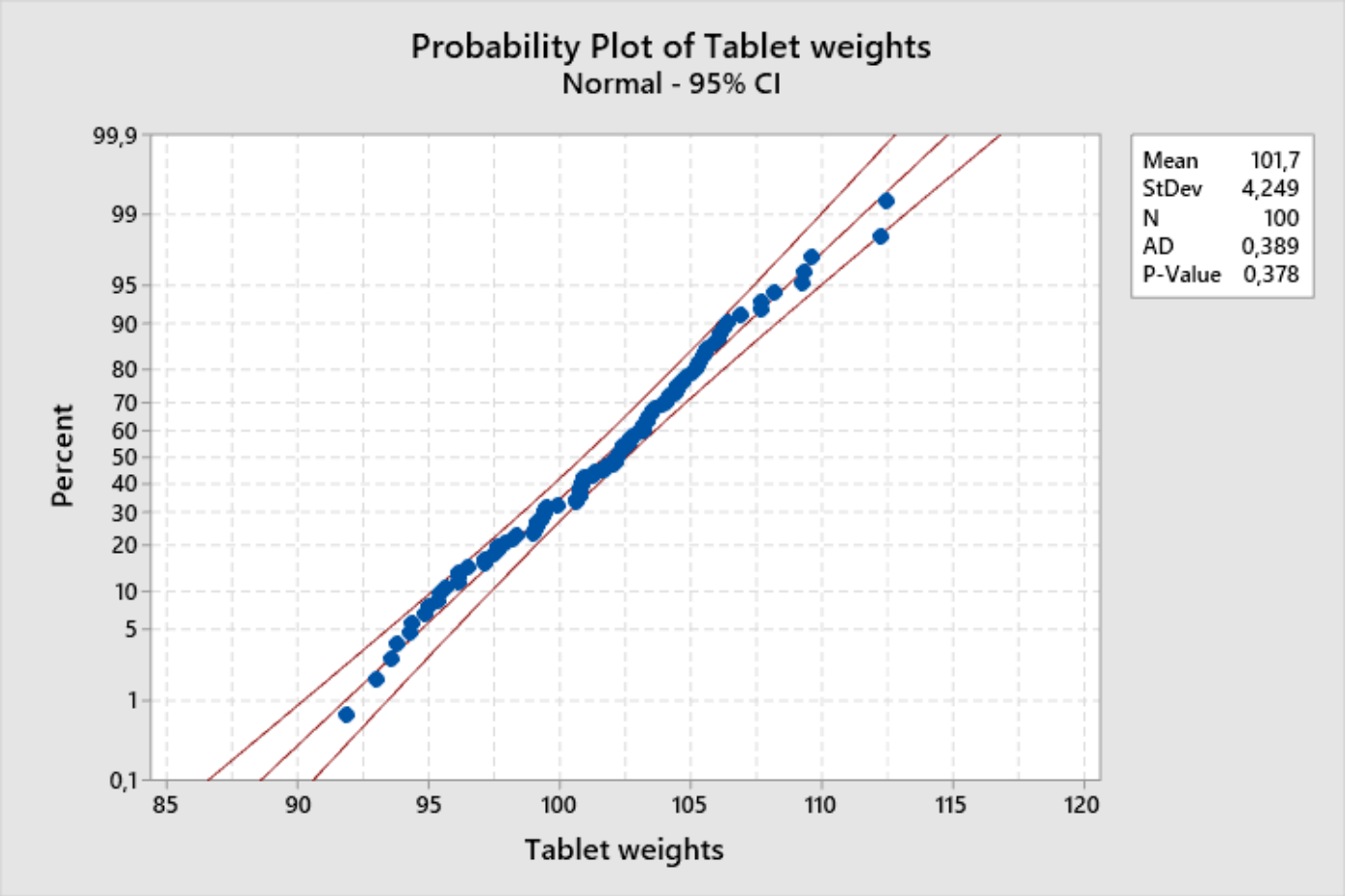


CASE STUDY 4

Given that **this example has general validity** (in fact it could equally apply to the weight of vials taken from a filling line or to the volumes of pre-filled syringes, *etc.*), the first thing to do is to look at how the weights obtained in Production are distributed.

As expected, the graph alongside shows that the data are normally distributed.

We can then proceed as shown in the next slide.

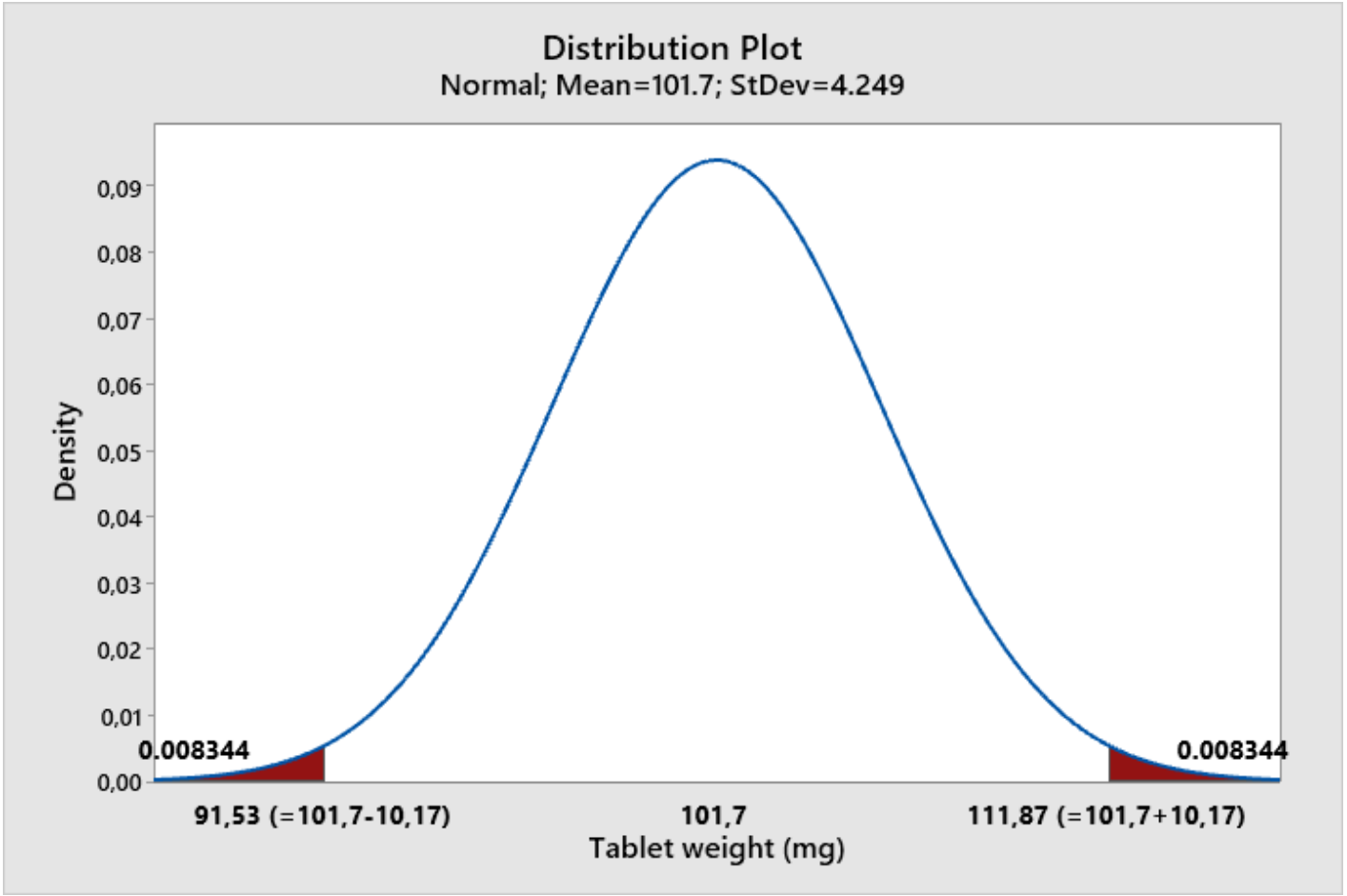


CASE STUDY 4

The initial question can be reworded as follows:

What percentage of tablets weigh between 91.53 mg and 111.87 mg? If this percentage is equal to or greater than 98%, the batch passes the test.

The percentage of tablets of interest can be estimated by calculating the area underlying the normal curve in the figure in the range between 91.53 and 111.87 mg as follows:



CASE STUDY 4

$$Z = \frac{X - \mu}{\sigma} = \frac{111.87 - 101.7}{4.249} = 2.39$$

Standard tables show that 2.39 corresponds to an area of 0.9916 and therefore the area greater than 111.87 is: 1- 0.9916 = 0.0084: this area corresponds to the probability of finding a tablet that weighs 111.87 mg or more.

Since the range considered is symmetric, the same probability also corresponds to the probability of finding a tablet that weighs 91.53 mg or less and therefore the total probability that a tablet weighs less than 91.53 mg or more than 111.87 is:

0.0084 + 0.0084 = 0.0168 or 1.68%.

From this it follows that 100 - 1.68 = 98.32% of the tablets will be included between ± 10% of the average weight.

CASE STUDY 4

Formally, the initial condition is satisfied even if a little at the limit (98.32% *vs.* 98%) and therefore the batch should pass the « average weight ± 10% » test also during the final analysis!

However, there is an overall probability of 1.68% that a tablet may exceed the weight limits and therefore a possibility, albeit small, that a random sample, for example of 30 tablets taken for the Content Uniformity test, may not pass it! If in fact the batch from which the 100 tablets were sampled was, for example 10000 units, 1.68% are still 168 tablets.

This result is important to be aware that, <u>given a certain process</u>, more often than is believed, the totality of the pieces produced does not meet all the set limits !

Hitting in an OOS or an OOT is therefore not so strange ! ☺

CASE STUDY 4

In any case, knowing the percentage of defectiveness of the tablets in a batch (*e.g.*, 1%, to simplify the calculations), Probability Theory allows you to estimate quite easily what is the probability of finding for example 3 defective units out of 30 sampled.

In fact, using the Poisson distribution (as an approximation of the binomial), that is:

$$p(x) = \frac{(np)^x}{x!} e^{-np} = \frac{\lambda^x}{x!} e^{-\lambda}$$

it can be estimated that:

$$p(3) = \frac{(30 \times 0.01)^3}{3!} e^{-(30 \times 0.01)} = 0.0033 \ (= 0.33\%)$$

There is therefore less than 1% probability that by randomly sampling 30 tablets, 3 of them are defective.

CASE STUDY 4

A random (or *aleatory* or *stochastic*) variable is distributed according to the Poisson law if its *probability mass function* (or *probability function*) is:

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

« … if *n* independent trials, each of which results in a success with probability *p* , are performed, then, when *n* is large and *p* small enough to make *np* moderate, the number of successes occurring is approximately a Poisson random variable with parameter $\lambda = np$ »

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

CASE STUDY 4

Introduced by Siméon Denis Poisson in a book he wrote regarding the application of probability theory to lawsuits (1837), it has a tremendous range of applications in diverse areas even rather common such 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,
- number of insects in a specified volume of soil, *etc.*

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

CASE STUDY 4

Using the above formula (obviously remaining within its validity field, that is: large size and small defectiveness), if the weights sampled in the process are a good estimate of the production batch, you can build a table of « estimated defects » such as the one shown below, again when taking a defect of 1% and a total sample of 30 tablets:

To demonstrate that the approximation provided in cases of this type by Poisson distribution is acceptable, it has been added the results that would have been obtained using the Binomial distribution. As expected, the agreement between the two data sets is good.

Number of defective tablets on a total of 30 sampled	Poisson	Binomial
0	0.7408	0.7397
1	0.2222	0.2242
2	0.0333	0.0328
3	0.0033	0.0031
…	…	…

CASE STUDY 4 - CONCLUSION

Since the considerations made regarding the tablets can be extended to other situations (*e.g.*, filling weights of vials containing sterile powders, « black specs » in tablets or dosed powders, *etc.*), it has been seen as the application of simple quantitative methods (or **quality metrics**) allow us to extract useful information from simple in-process weighing operations and therefore already have an idea of the fate of the lot before it is submitted to the release analysis.

CASE STUDY 5

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

CASE STUDY 5 - STATISTICAL HYPOTHESIS TESTING

The statement H_0: $\mu = 50.36$ mg is the **NULL HYPOTHESIS** and H_1: $\mu \neq 50.36$ is the **ALTERNATIVE HYPOTHESIS**.

In this case, being the sample size n = 20 (*i.e.*, small sample as $n < 30$) we can assume a *t-distribution* for the weights of tablets.

The *t-distribution* is bell-shaped like the Gaussian, but its exact shape depends on the « degrees of freedom» that in this case are 19 (*i.e.*, dof = n -1).

D.C. Montgomery, Statistical Quality Control: A Modern Introduction – J. Wiley (2013)

CASE STUDY 5

The sample mean is: $\bar{x} = 50.7375$ mg and its standard deviation is: $\sigma = 2.6982$

From tables it follows that : $t_{\alpha/2}$ at 19dof = 2.093 and therefore $-t_{\alpha/2}$ at 19dof = -2.093

At this point the «test statistic» can be calculated : $t_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{50.7375 - 50.36}{2.6982 / \sqrt{20}} = -0.6257$

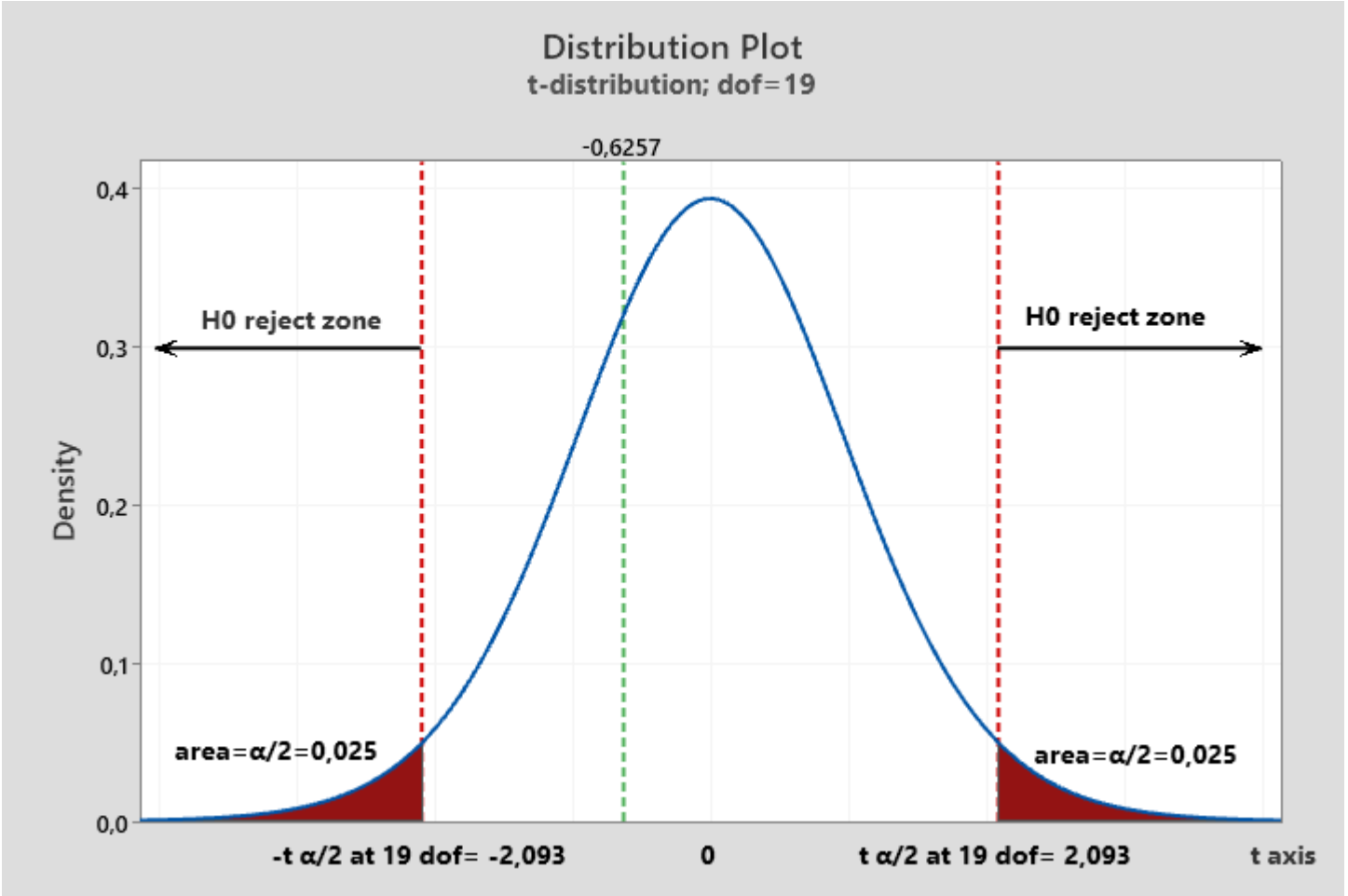
Since the value assumed by the test statistic does not fall within the so-called «reject zone»:

$$t_{\bar{x}} = -0.6257 < t_{0.025} = 2.093 \text{ e } t_{\bar{x}} = -0.6257 > -t_{0.025} = -2.093$$

there is no experimental evidence to reject the null hypothesis and therefore conclude that the process is « out of control».

CASE STUDY 5

As it can be seen from the graph here, H_0 would be rejected for values of the «test statistic» lying in the reject zone.



CASE STUDY 5-bis

Close to the example just seen, it could also be the following case study:

The tablets obtained from a given process are rejected if they weigh less than 95 mg or more than 108 mg. 100 tablets 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!

Being the sample size greater than 30 we can assume a Gaussian distribution for the weight of the tablets and therefore….

CASE STUDY 5-bis

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

from which it follows that:

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

CASE STUDY 5 - CONCLUSION

These two examples were intended to show how:

- taking random samples from a production line

or

- analyzing « processing waste »

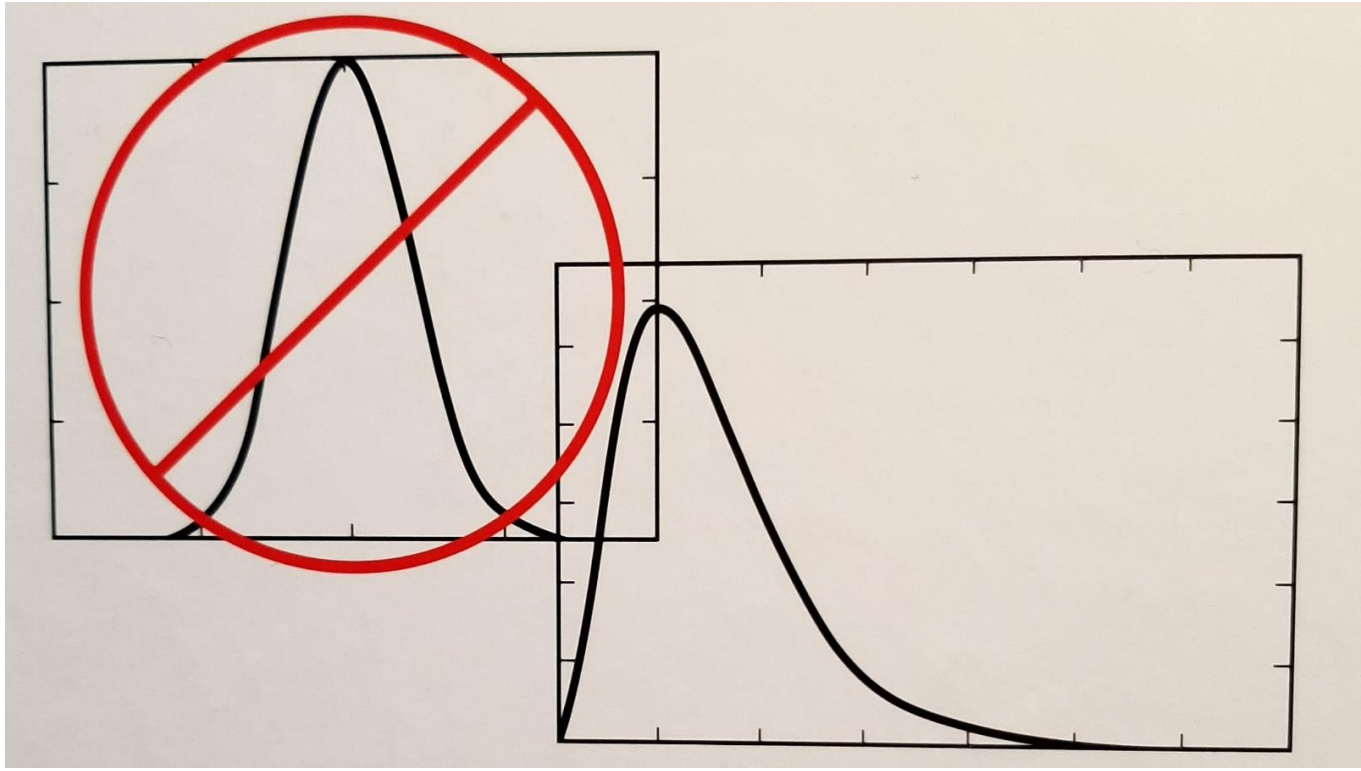
and using a pocket calculator and standard tables available everywhere it is possible to easily **determine quality metrics capable of giving crucial information on the « state of the process ».**

CASE STUDY 6

- Processes non-normally distributed: impurities content, microbial counts, Particle Size Distributions (PSD), black particles (or *black specs*)
- Normalization of non-normal data using mathematical transformations (logarithm, square root, reverse or reciprocal)
- Johnson Transformations

CASE STUDY 6

"… One of the major sources of frustration in the application of statistical process control (SPC) methods to a chemical process is the prevalence of variables whose values have distributions that are, by nature, distinctly non-normal. Typically, methods used to analyze these variables are based on the normal distribution and, as such, are unrealistic…"



W.A. Levinson, Statistical Process Control for Real-World Applications – CRC Press (2011)

D.C. Jacobs, Watch out for Non normal Distributions - Chemical Engineering Process (Nov. 1990)

CASE STUDY 6

In fact, practical experience shows that many of the variables encountered in typical industrial processes cannot be adequately described through normal distribution. The reasons for this are numerous, for example:

- ***Confinement of a given variable within predetermined limits*** (*e.g.*, the temperature of a process that must not exceed a predetermined limit)
- ***Measurement of a characteristic that has its natural limit in zero*** (*e.g.*, the ***moisture content*** or the ***impurities content***)
- ***Mathematical relations between the variables*** (*e.g.*, the speed of a reaction that depends exponentially on temperature or ***microbial counts*** that follow non-normal distributions).

D.C. Jacobs, Watch out for Nonnormal Distributions - Chemical Engineering Process (Nov. 1990)

CASE STUDY 6

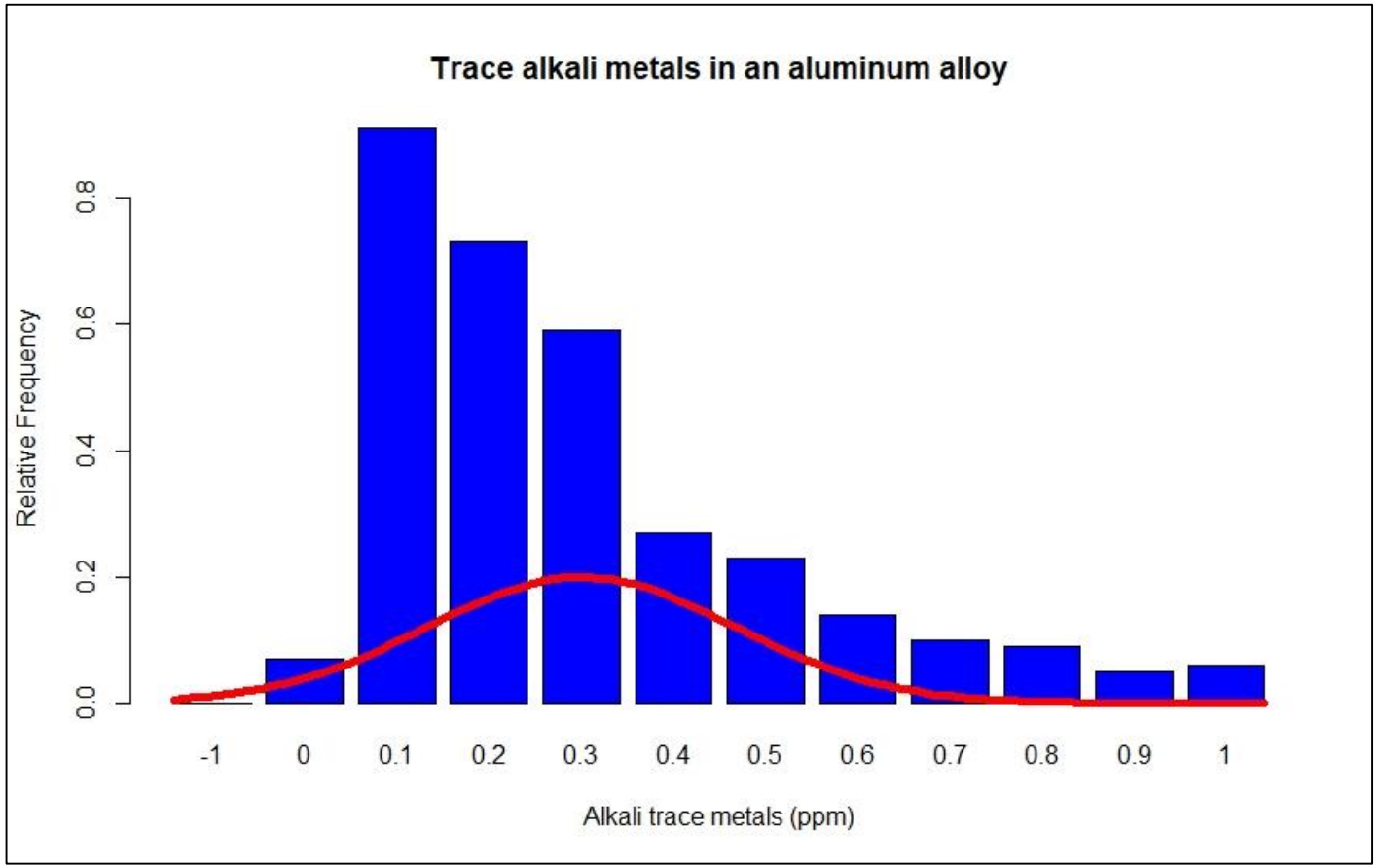
It is important to keep all this in mind because if it is true that assuming the normal distribution simplifies calculations, its use where it does not apply can lead to serious inconveniences!

In this regard, it is sufficient to think of the plot limits which are very often defined, as seen above, as « average ± 3σ ». In the case studies previously discussed, this was acceptable since the variables under study [*i.e.*, HPLC assays (CASE STUDY 1) and weights (CASE STUDY 2)] were distributed exactly as normal.

However, if the variable under study does not follow the normal distribution, the use of limits calculated precisely as « average ± 3σ » could highlight that they are « out of specifications data » (or OOS) which in fact are not!

CASE STUDY 6

For example, the figure here on the side shows a bar chart relating to the content of alkali metals in traces in traces in an aluminum alloy. Data refer to over 200 samples.



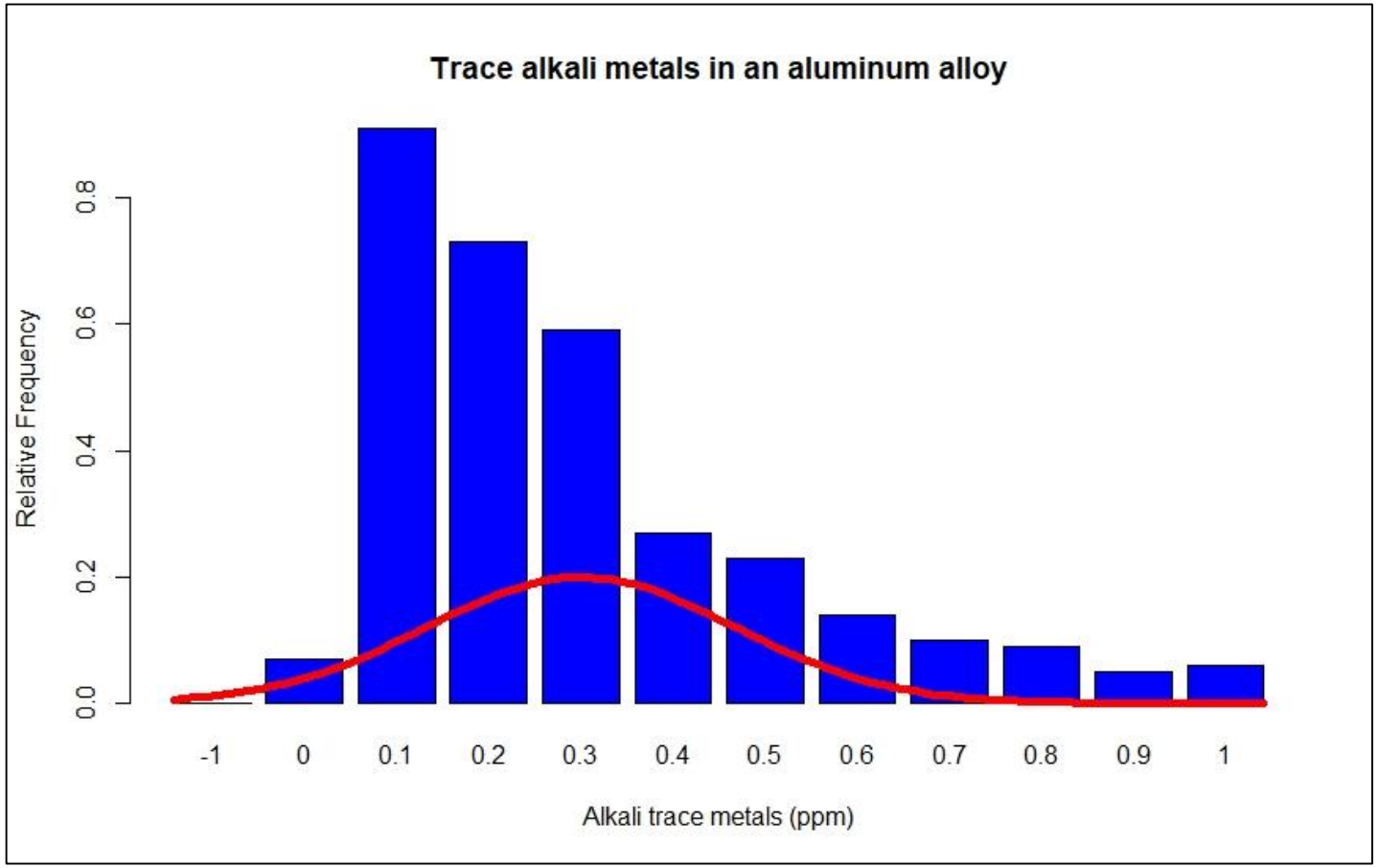
W.A. Levinson, Watch out for Non normal Distributions of Impurities – Chem. Eng. Proc. (May 1997)

CASE STUDY 6

The average is approximately 0.3 ppm and the standard deviation 0.33 ppm.

The superimposed red curve is a normal of average $\mu = 0.3$ ppm about and standard deviation $\sigma = 2$ ppm.

What looks « anomalous » in this graph?



CASE STUDY 6

What appears « anomalous » in the previous graph?

- The curve does not adapt well to the histogram
- The curve extends even beyond zero on the negative semi-axis!

This suggests the possibility of having even less than zero in terms of residual impurity content in the aluminum alloy under analysis!

NORMAL DISTRIBUTION IS THEREFORE NOT A REALISTIC MODEL FOR THIS DATA!

CASE STUDY 6

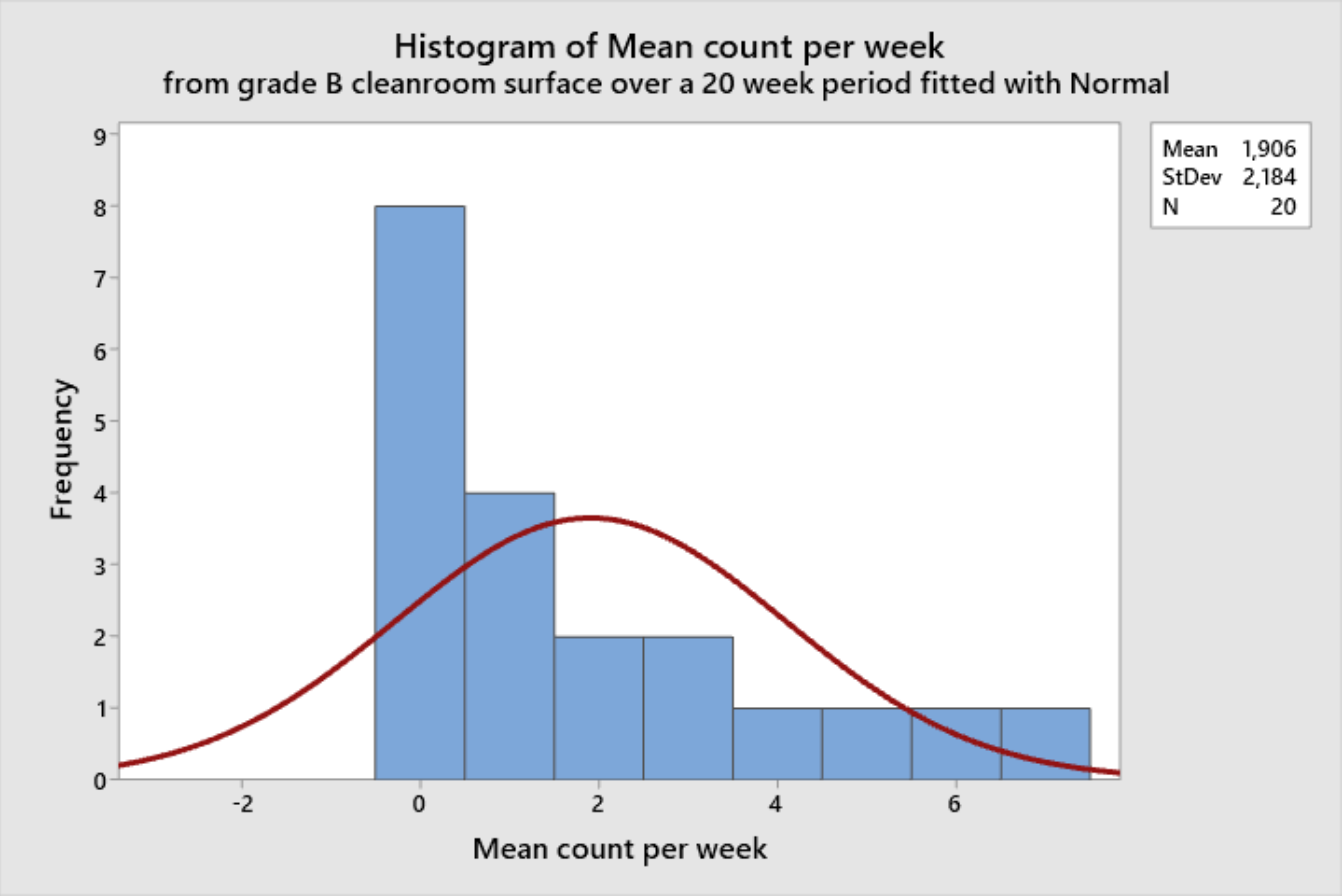
A similar situation occurs, for instance, when trying to interpolate a typical microbial distribution such as the one shown in the table on the side, with a Normal curve.

Week No.	Mean count *per* week
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)

CASE STUDY 6

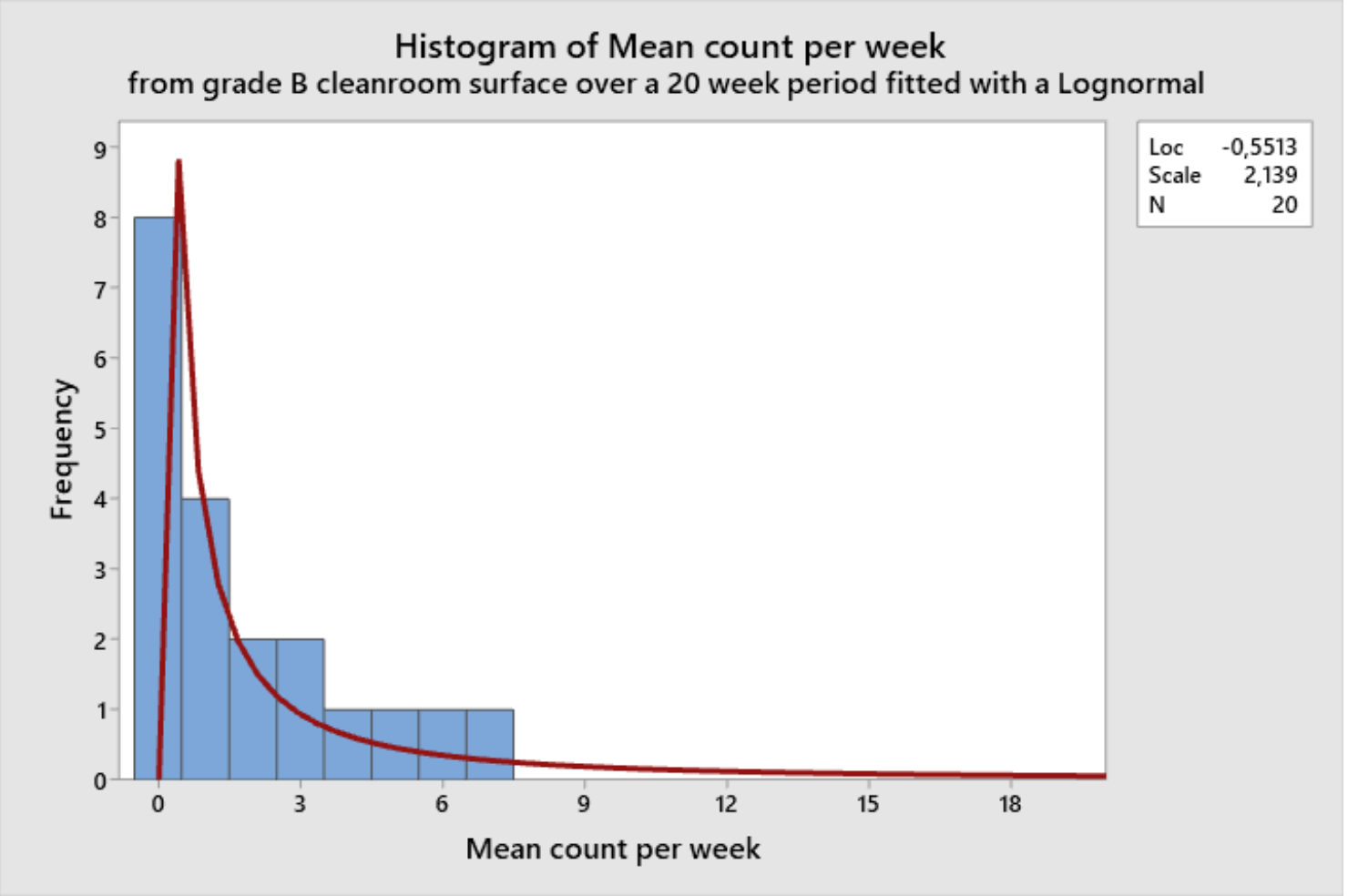
Since even « microbial count » data are not normally distributed (it deals, in fact of a characteristic that has its natural limit in zero), once again there is an unrealistic result !



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

CASE STUDY 6

In fact, these data are distributed much more correctly according to a so-called « lognormal » distribution, *i.e.*, a distribution in which the logarithm of the averages of the microbial counts follows the Normal*.

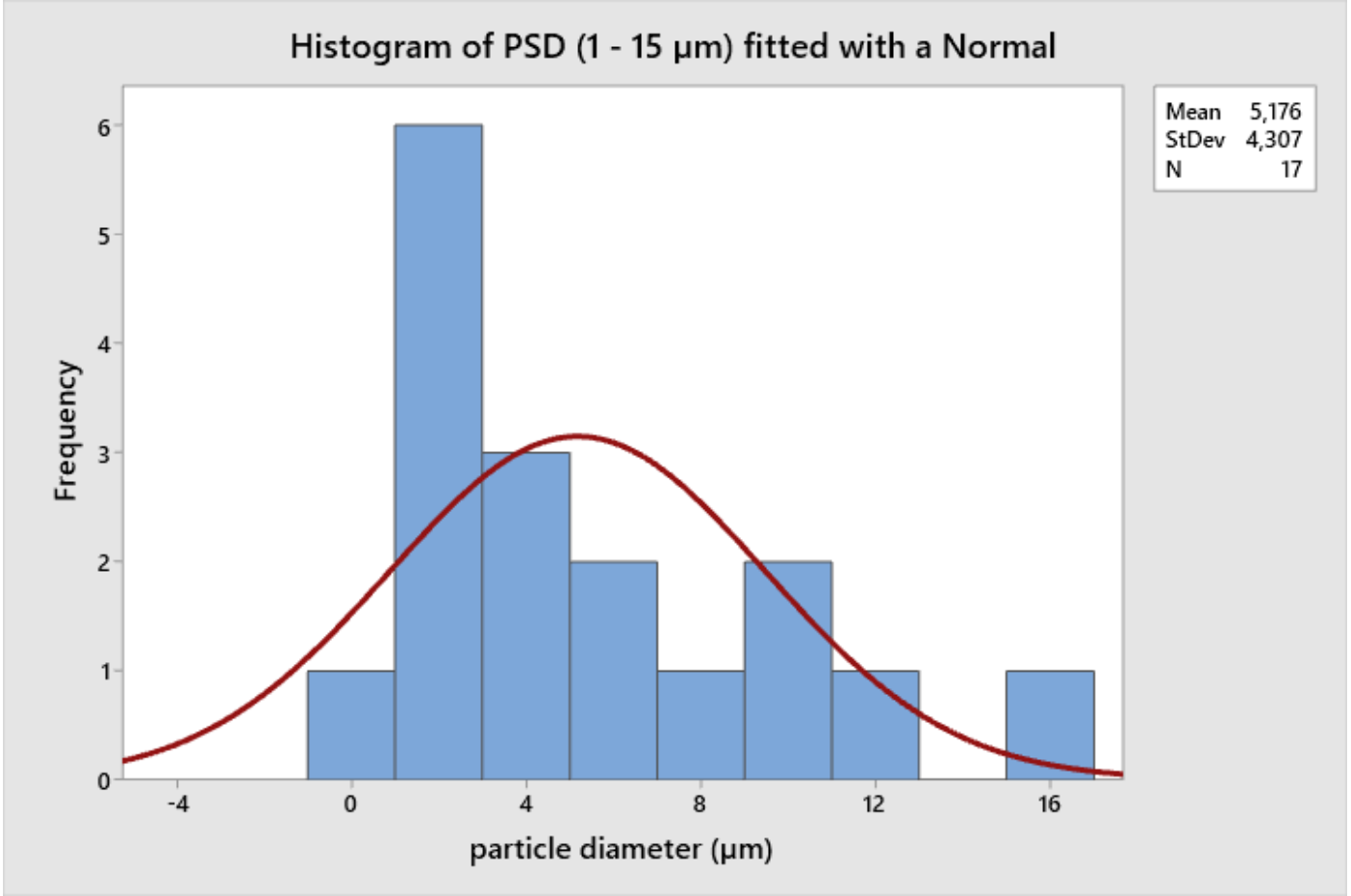


**Data are said to follow the Lognormal distribution when the logarithms of the measurements follow the Normal distribution.*

CASE STUDY 6

The same also occurs when trying to interpolate a typical particle distribution curve with a Normal.

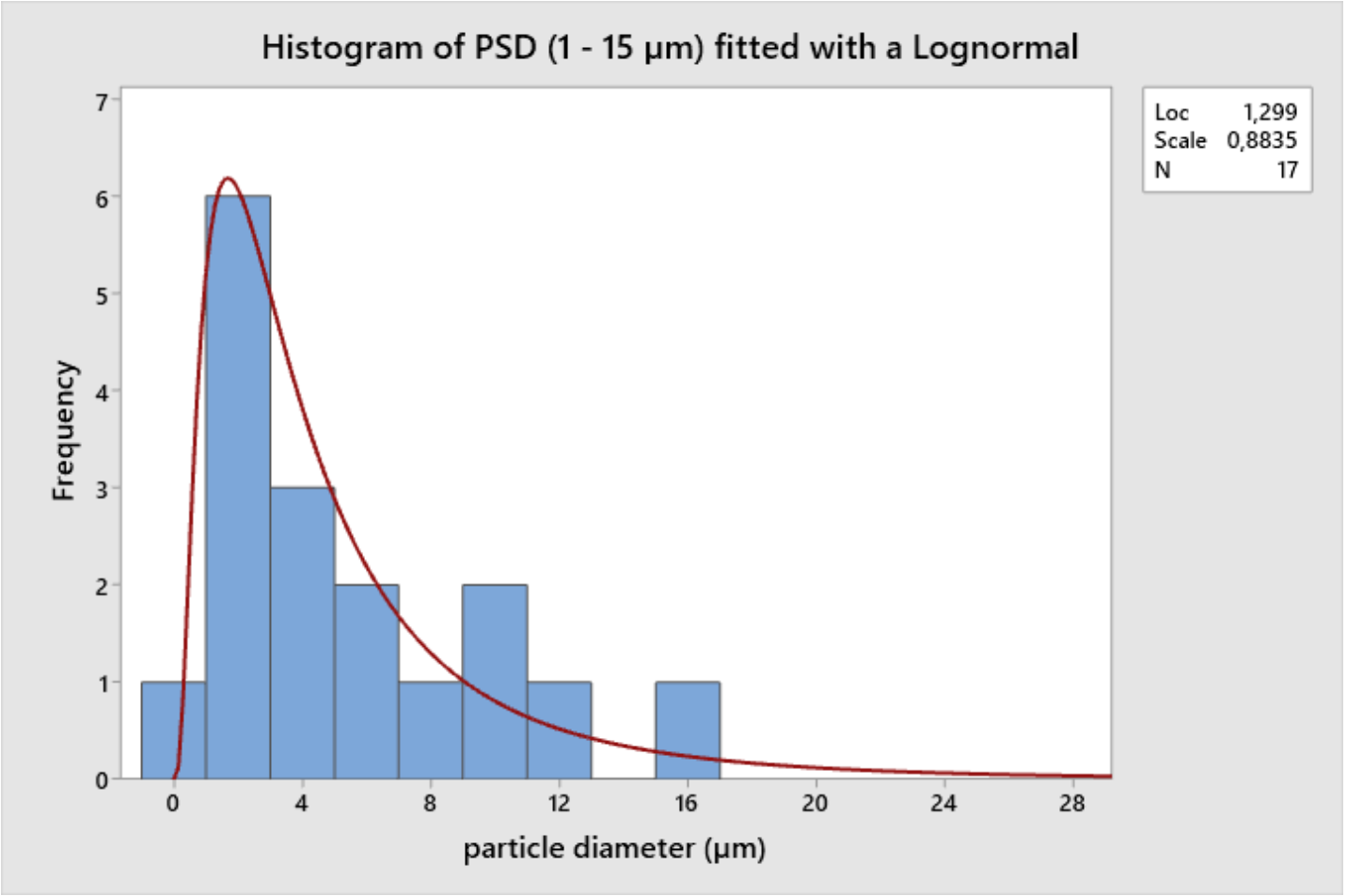
Since even particle size distribution (PSD) has its natural limit in zero, data interpolation using a Normal curve leads to an unrealistic result !



CASE STUDY 6

Using a lognormal instead, everything takes on meaning!

The advantage of this is, for example, being able to estimate in probabilistic terms the percentage of particles lower than a given threshold or included in a range of diameters. A practical example of this is shown in the next slide.

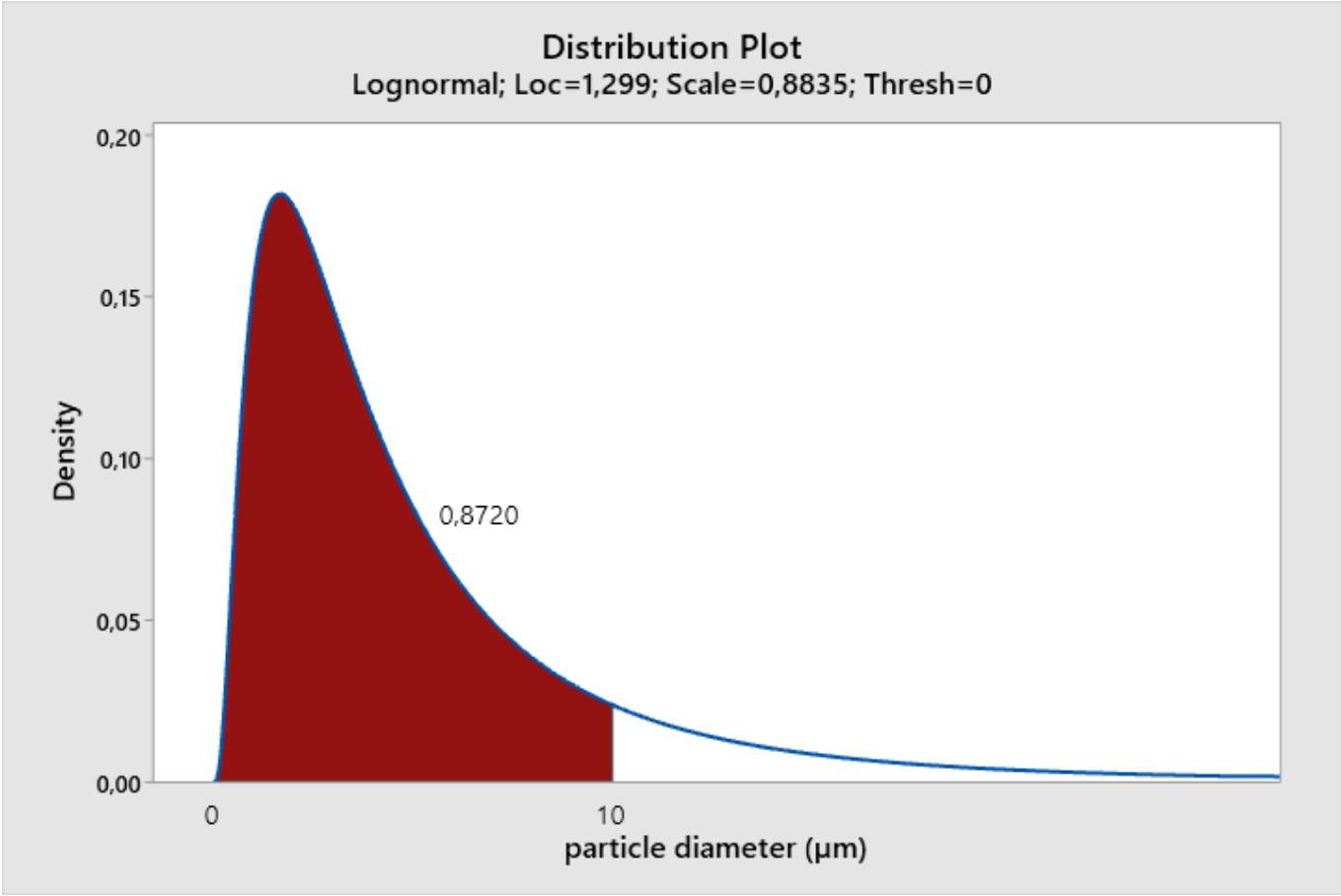


CASE STUDY 6

In the graph alongside, the red colored area corresponds to the probability that a particle has a diameter of less than 10 µm which, as seen, is 87.2%.

This information can be useful for a deepening of the data for purposes such as:

- process validation
- historical analysis
- handling of a complaint / OOS
- *etc*.



CASE STUDY 6

Non-normal distributions are helpful in many situations, an example?

« Black Specks » (or black particles) in tablets (or vials or APIs)

In general this is an occasional phenomenon that occurs randomly and with low frequency.

So one wonders: is this defect random or not? The answer can be given by the

POISSON DISTRIBUTION or « *of rare events* »

CASE STUDY 6

Let's consider, for instance, the case of black particles found by inspecting samples of 80 different and hypothetical batches of tablets (please, note that what said below also applies to black particles found in vials or in samples of APIs !)

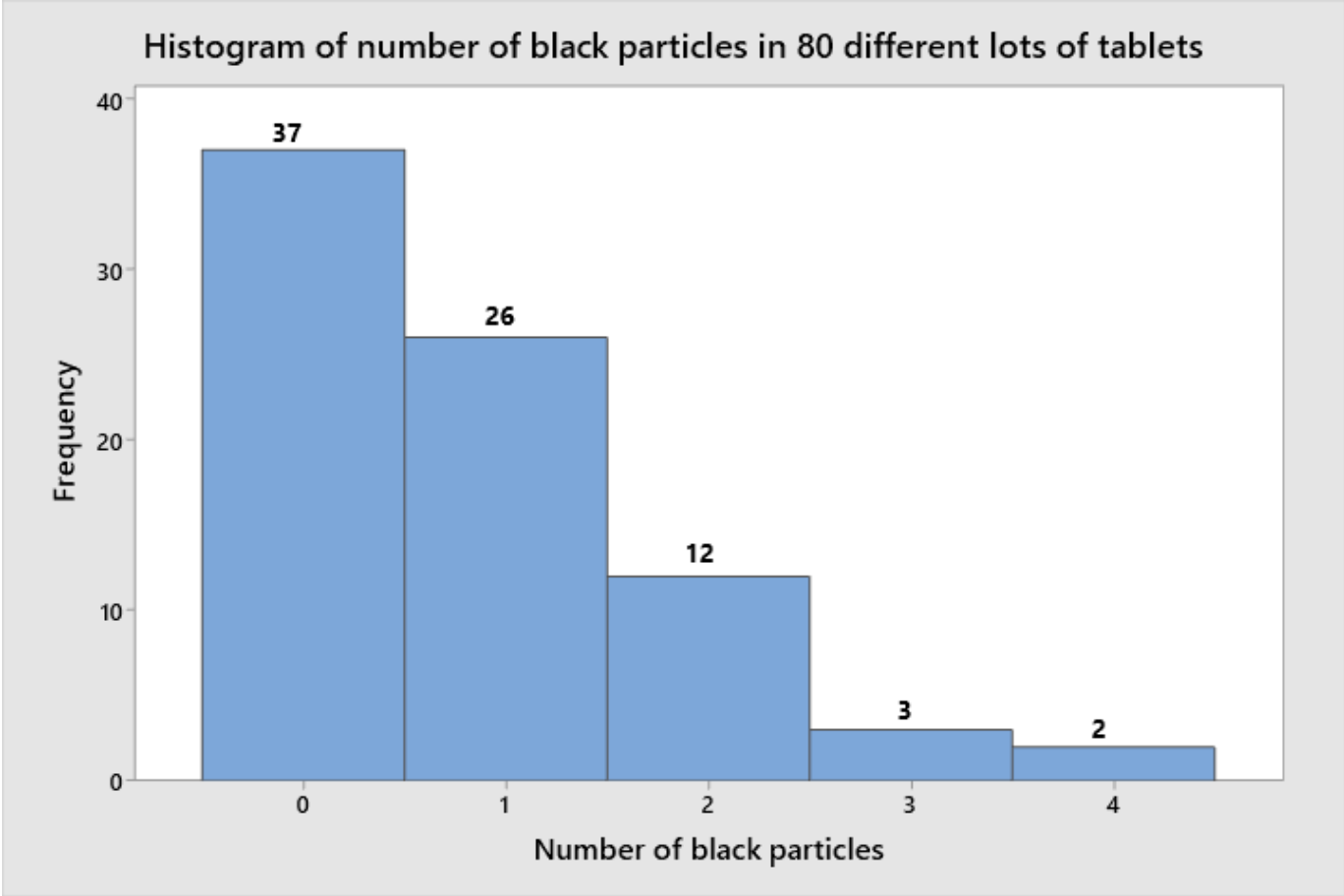
Lot	No.	Lot	No.	Lot	No.	Lot	No.	Lot	No.	Lot	No.	Lot	No.	Lot	No.
1	0	11	2	21	0	31	1	41	1	51	1	61	0	71	0
2	1	12	2	22	0	32	1	42	0	52	2	62	3	72	0
3	1	13	2	23	0	33	2	43	0	53	4	63	4	73	0
4	0	14	0	24	0	34	1	44	1	54	1	64	1	74	1
5	0	15	2	25	0	35	1	45	0	55	1	65	1	75	2
6	0	16	3	26	0	36	0	46	0	56	1	66	0	76	3
7	0	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

CASE STUDY 6

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

In summary:

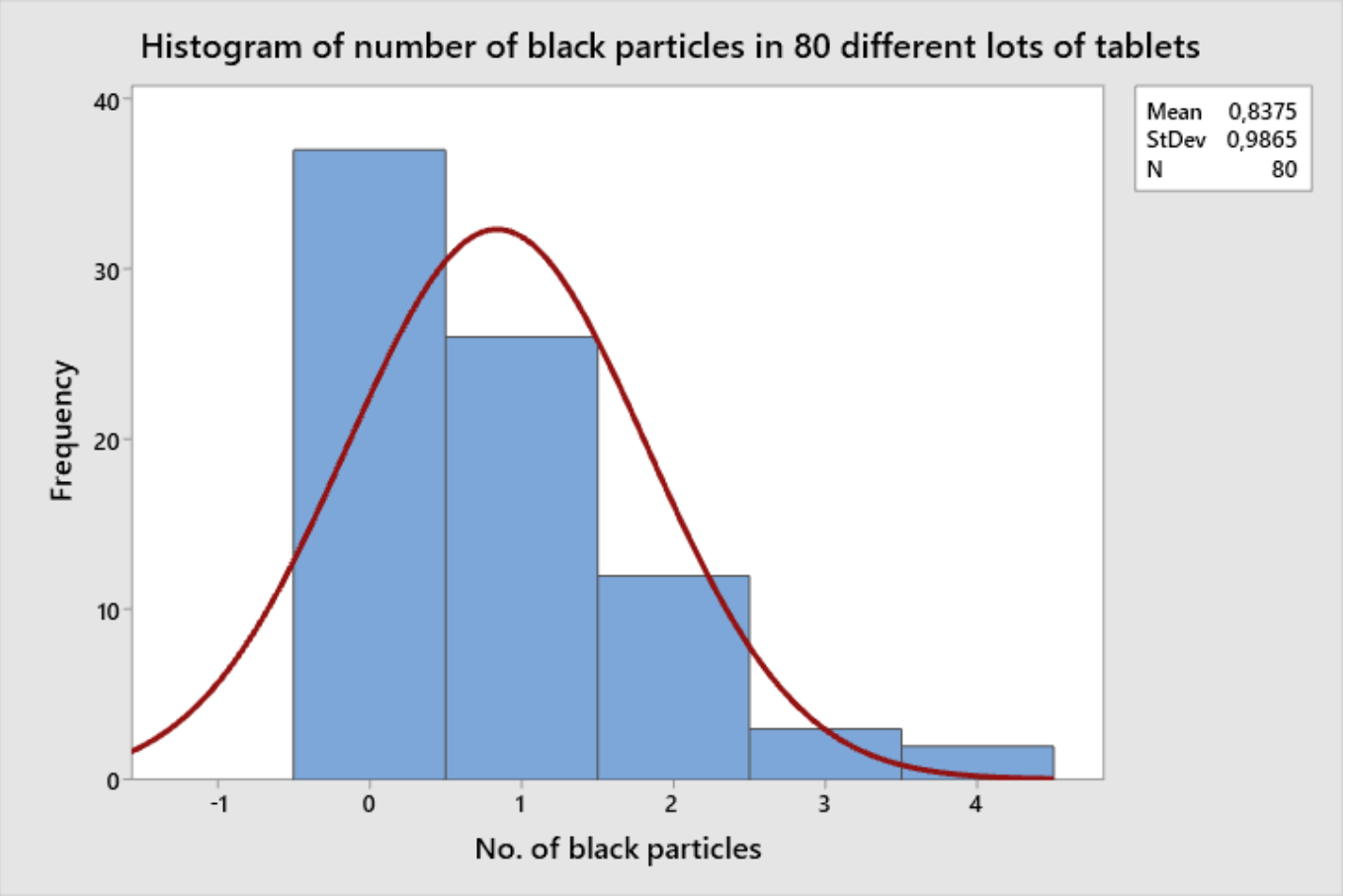
No. Black-specks	0	1	2	≥ 3
Frequencies	37	26	12	5



CASE STUDY 6

The shape of the histogram already indicates this, but the overlap with a normal curve confirms that this cannot be a good approximation for these data.

Let see if they are distributed according to the Poisson law.



CASE STUDY 6

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

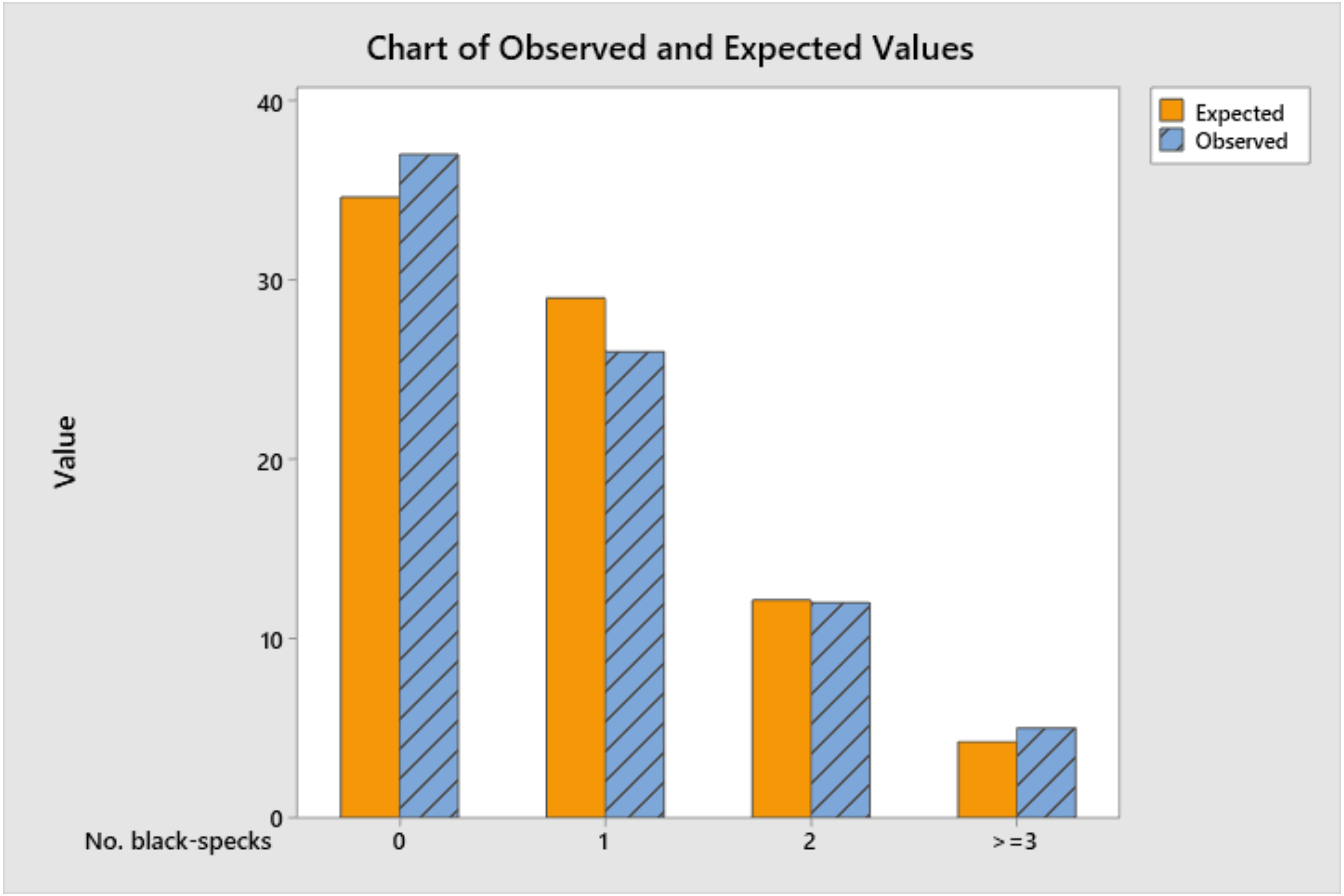
1 (25,00%) of the expected counts are less than 5.

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



CASE STUDY 6

- The good agreement between « expected frequencies » and « observed frequencies » shown in the previous slide indicates that the variable « number of black particles » (or black-specks) is distributed according to Poisson distribution.

- ***It is therefore reasonable to assume that the presence of these black particles in the analyzed lots is random.***

- This result, if on the one hand simplifies the situation because it excludes the presence of a specific cause ☺, on the other complicates it because it involves many possible causes and therefore its elimination could be difficult to solve ☹

CASE STUDY 6 - CONCLUSION

- This case study has shown that very often, in practice, experimental data are distributed in a non-normal way and this simply because they are manifestations of quantities (or variables) that are not distributed in a normal way.

  So, data distributed in a non-normal way does not necessarily imply anomalous behavior!

- It is therefore a conceptual error to want to force such data into a « normal dress » that is not theirs and this can be a source of inconveniences (*e.g.*, OOS that are not such).

- It therefore makes sense to establish whether an apparently anomalous datum is *out-of-trend* only after establishing the trend for the reference parameter (or variable).

- As seen, the use of appropriate statistic distributions allows us to extract information from these data that is very useful for the knowledge of the processes and their control.

CASE STUDY 6 – **ADDENDUM:** DATA TRANSFORMATION

- It must be said that in the presence of « non-normal » data (*i.e.*, data not distributed according to a Normal), a ***« mathematical transformation »*** that normalizes them is often used.

- In practice, there are three types of transformations that can be used to normalize « positively tailed » (or « right tailed ») data, namely:

  - logarithmic

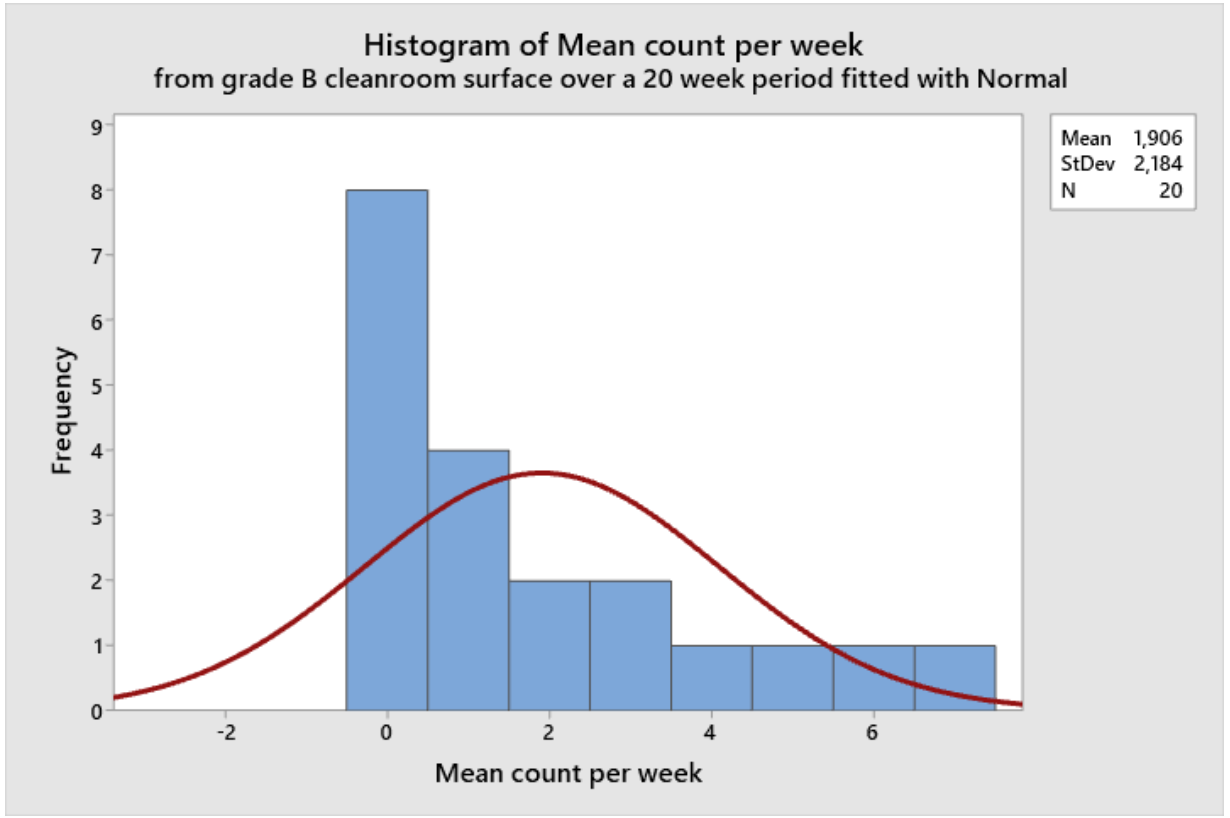
  - square root

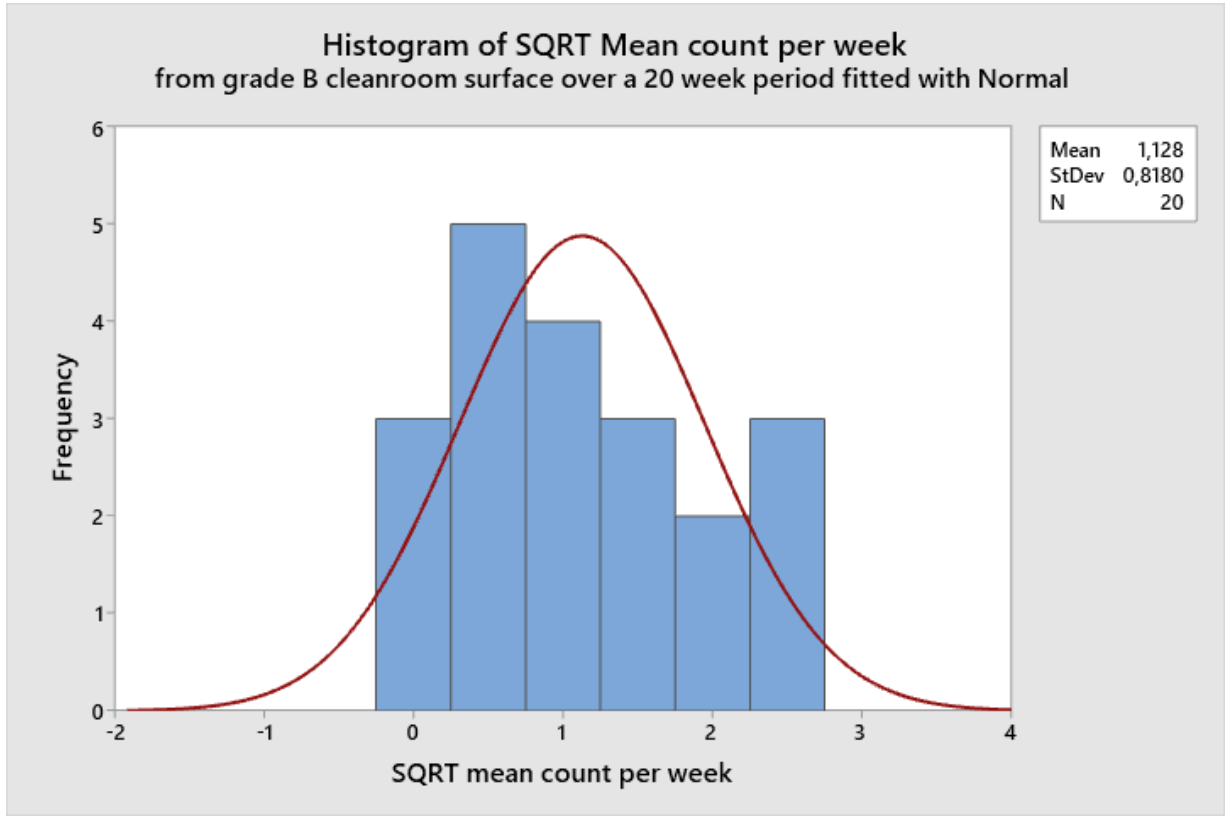
  - reverse (or reciprocal)

The « reverse transformation » is usually used for the more extreme cases of « positive tailing ». For the less extreme ones, « logarithmic transformation » is usually used, while in the presence of only slightly « tailed to the right » data, the « square root» is used.

CASE STUDY 6 – ADDENDUM: DATA TRANSFORMATION

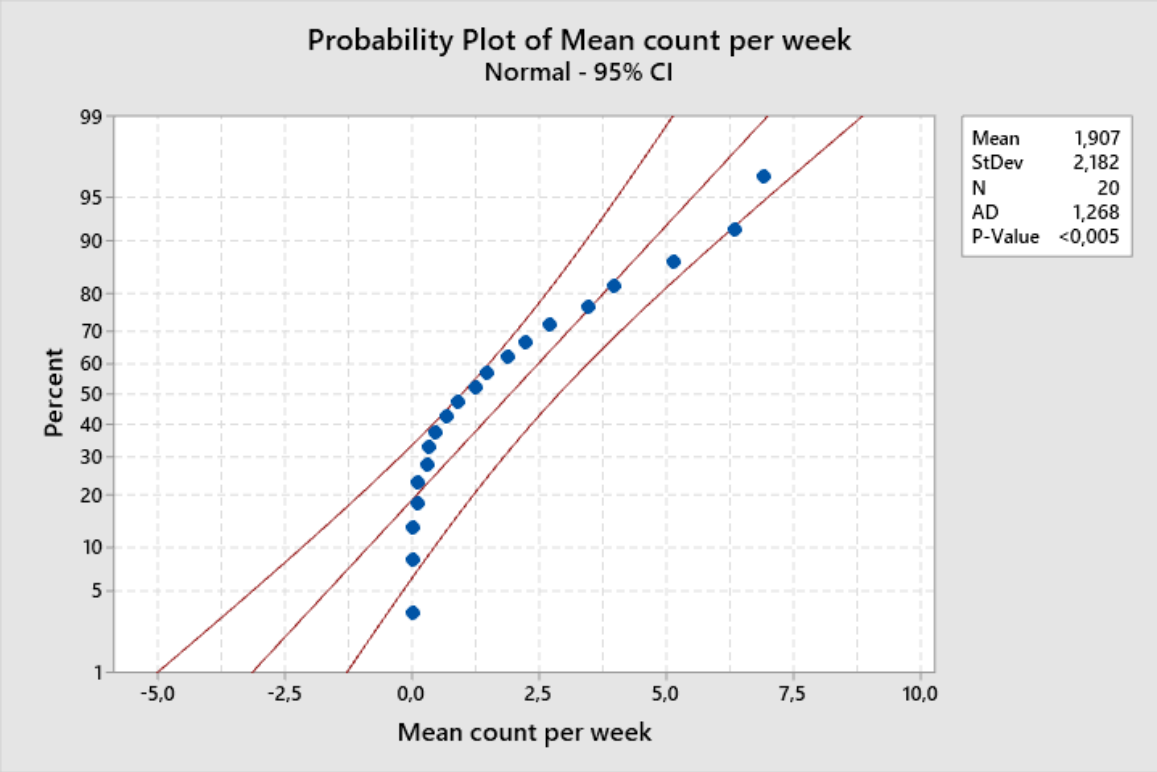
Let's consider the case of the mean microbial count *per* week discussed before. Here below are the histogram before and after data transformation (*i.e.*, square root)

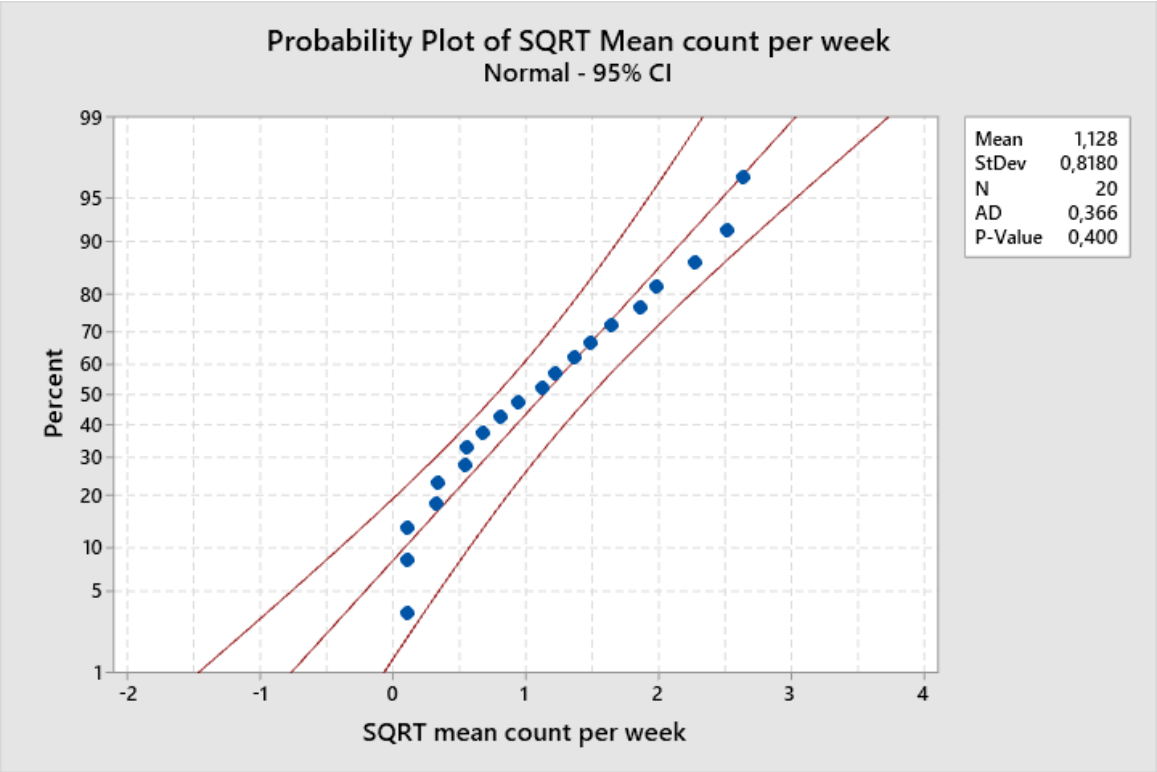




CASE STUDY 6 – ADDENDUM: DATA TRANSFORMATION

Mean microbial count *per* week: here below are the probability plot before and after data transformation (square root). After transformation data are normally distributed (*P-value* > 0.05).





CASE STUDY 6 – ADDENDUM: JOHNSON TRANSFORMATIONS

The system of transformations developed by Norman L. Johnson in 1949, computes an optimal transformation function from three flexible distribution families and, in particular:

S_B or *Bounded* : $Z = \gamma + \eta \ln\left(\frac{x - \varepsilon}{(\lambda + \varepsilon - x)}\right)$

S_L or *Lognormal* : $Z = \gamma + \eta \ln(x - \varepsilon)$

S_U or *Unbounded* : $Z = \gamma + \eta \, asinh\left(\frac{(x - \varepsilon)}{\lambda}\right)$

in which Z is the standard normal variable, and x is the non-normal original data, all the necessary parameters will be returned.

CASE STUDY 6 – ADDENDUM: JOHNSON TRANSFORMATIONS

This system has the practical and theoretical advantages of covering a wide variety of shapes.

The Johnson system can closely approximate many of the standard continuous distributions through one of the three functional forms and is thus highly flexible.

CASE STUDY 7

- Multivariate methods: a different way to look at Quality Control data !

CASE STUDY 7

- The metrics seen so far are highly informative, but they provide a so-called ***univariate*** information that is referred to the single parameter (or variable) considered.

- Experimental data, however, are now available in large quantities and provide multiple information for a given study subject. A typical example are process data or Quality Control data which are usually organized in datasets containing different types of measurements (chemical and microbiological) on different samples, each representative of a given production batch.

- In today's common practice, despite all this mass of available data, they are considered individually and not globally.

CASE STUDY 7

Analytical parameter (or variable)	Unit	Range	Analytical technique	Abbreviation
pH	pH units	5.0 – 8.0	pH-meter	ph
Water content	%	1.0 – 5.0	Karl-Fisher titration	h2o
Assay	%	80 - 92	HPLC	assay
Starting material residual content	%	≤ 0.20	HPLC	sm
Largest known impurity	%	≤ 0.20	HPLC	known
Largest unknown impurity	%	≤ 0.20	HPLC	unk
Total impurities	%	≤ 1.0	HPLC	total
Residual content solvent 1	%	≤ 5.0%	Gas-chromatography	solv1
Residual content solvent 2	%	≤ 5.0%	Gas-chromatography	solv2
Residual content solvent 3	%	≤ 1.0%	Gas-chromatography	solv3

Usually, despite this multiplicity of available data, they are considered individually and not globally.

CASE STUDY 7

- An alternative approach is that offered by the so-called ***Multivariate Analysis* (*MVA*)** which allows the simultaneous analysis of a set of parameters (or variables) offering, compared to the separate analysis of each variable, the information content resulting from the relationships existing between the variables (*)

- The combined use of *MVA* and *Data Visualization* allows you to quickly extract the information contained in the dataset and convert it into « ready-to-use knowledge ».

(*) B. Everitt, T. Hothorn, An Introduction to Multivariate Analysis with R, Springer, 2011

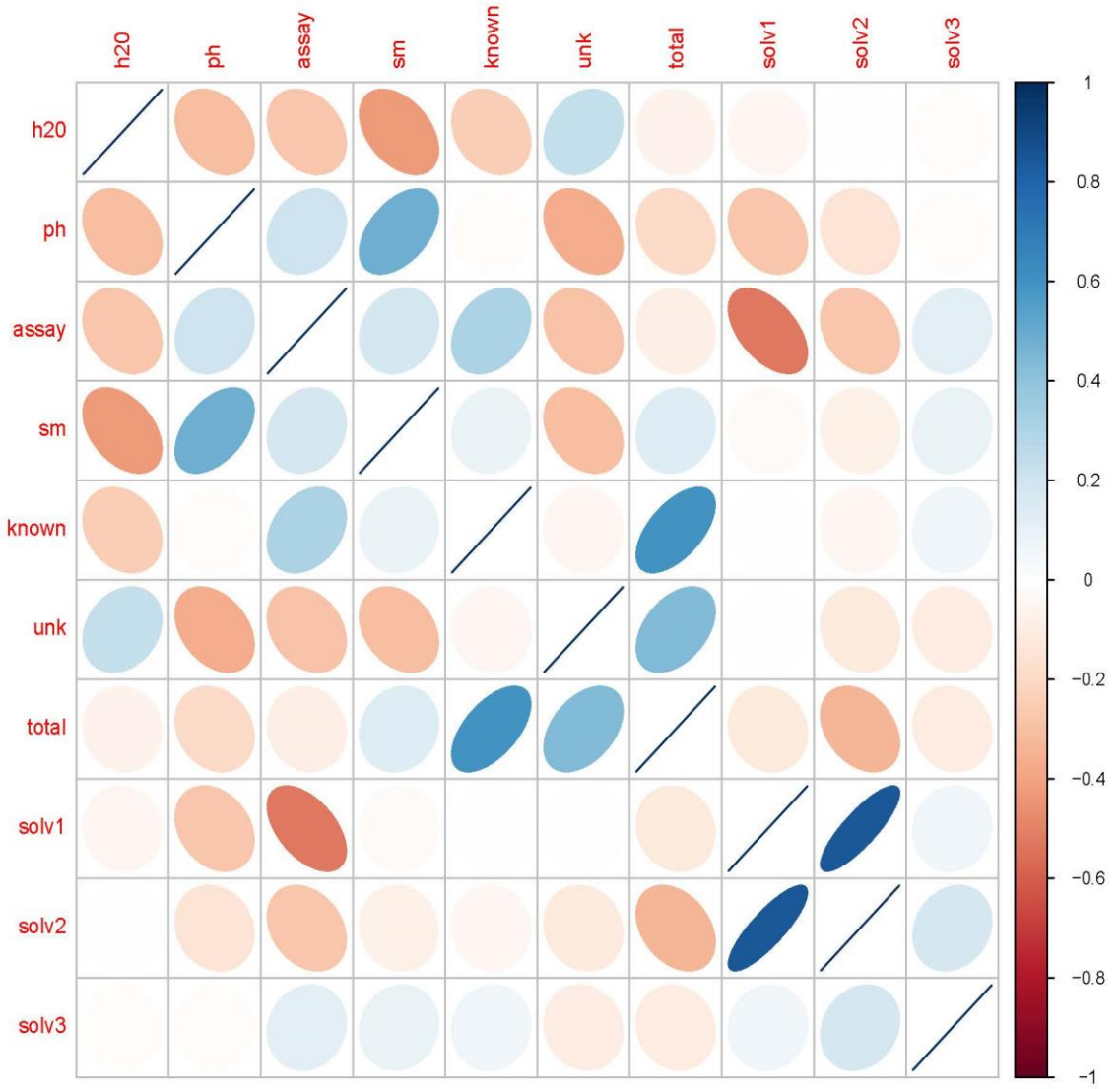
CASE STUDY 7

- As a « case study » let's consider a set of chemical QC data (*) of the type shown in the table above and relating to thirty-one (31) samples each representative of a different batch of active ingredient produced.
- Obviously, all batches are assumed to be produced using the same production method.
- In the following slide the first type of « graph » that can be obtained using the *MVA* methods is reported, *i.e.*, the so-called « *correlogram* » (or correlation diagram).

(*) *R. Bonfichi, A different way to look at pharmaceutical Quality Control data: multivariate instead of univariate, www.riccardobonfichi.it, 2018*

CASE STUDY 7

- Each element is a geometric figure that becomes the more elliptical and intensely colored the more the two variables are related to each other. On the main diagonal, where the correlation is greatest, the ellipses become segments.
- Ellipses oriented to the right and blue colored indicate that the two variables are ***positively correlated*** to each other (*i.e.*, as the one grows the other also grows), while if they are oriented to the left and brick-colored the variables are ***negatively correlated***.



CASE STUDY 7

The correlogram shows ***strong correlations*** between:

- Residual quantity solvent 1 and solvent 2 residual quantity: positive
- Residual quantity solvent 1 (solv1) and assay (assay): negative
- Amount major impurity known (known) and total impurities (total): positive

The correlogram also shows ***weak correlations*** between:

- Residual amount starting material (sm) and pH value (ph): positive
- Amount largest unknown impurity (unk) and total impurities (total): positive
- Residual amount moisture (h2o) and largest unknown impurity amount (unk): positive
- Residual amount starting material (sm) and residual moisture amount (h2o): negative

CASE STUDY 7

It is important to underline that:

- these correlations between the parameters, especially the stronger ones, highlight some aspects of the production process worthy of further study (for example, the strong correlation between *solvent 1* and *solvent 2*, *etc.*)

- obviously, the correlation pattern becomes more informative the better the production process is known in detail.

CASE STUDY 7

A powerful *MVA* technique is the so-called *Principal Component Analysis* (*PCA*) which, in general, allows to reduce the number of variables in play to just two / three « main variables », below indicated with « Dim1 », « Dim2 » and « Dim3 ».

Using pairs of these « main variables » (*e.g.*, Dim1 and Dim2 or Dim1 and Dim3, *etc.*) it is possible to build up real « maps », such as those shown in the following slides, within which each lot is identified as a point.
This gives a graphic representation of the lots under study.

I.T. Jolliffe, Principal Component Analysis, 2nd Edition, Springer, 2002

CASE STUDY 7

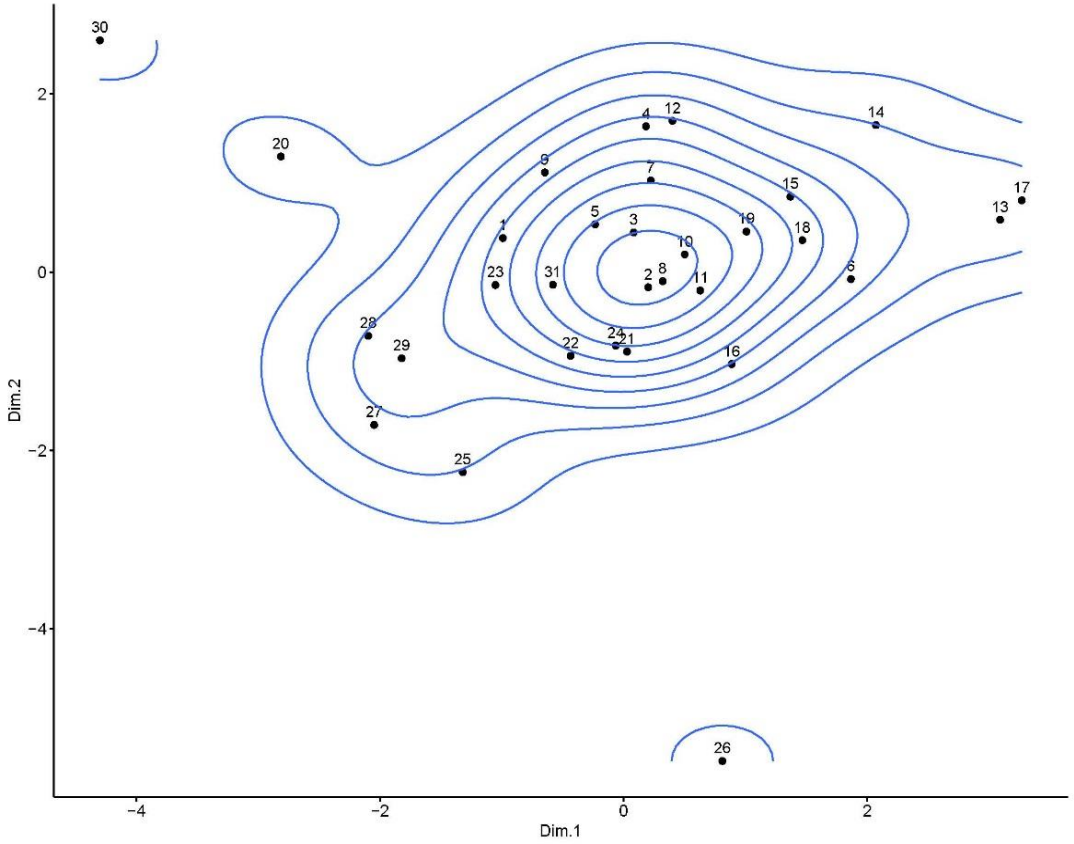


Figure 1

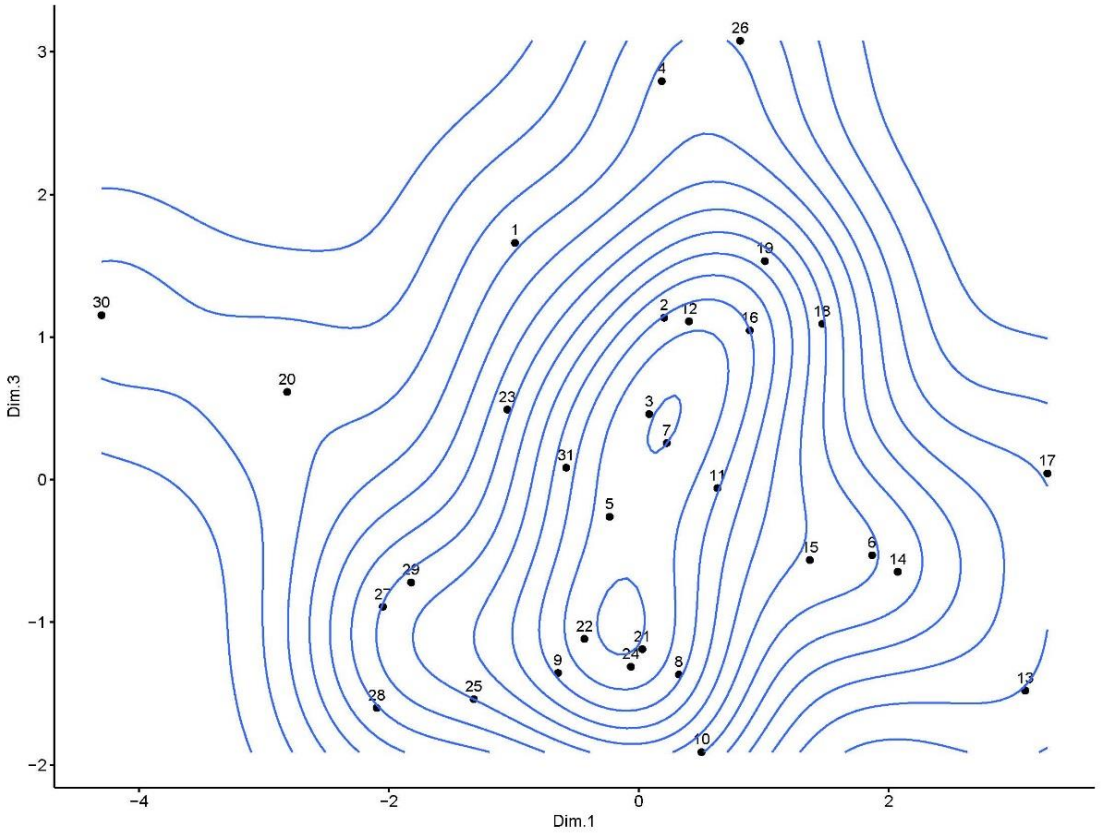


Figure 2

CASE STUDY 7

The examination in Figure 1 shows that:

- Most of the 31 lots considered appear to be centered around a central core defined by data points 2, 8 and 10).
- some lots form a separate group on the left of the diagram (25, 27, 28, 29)
- three lots appear evidently unrelated to the rest of the production (30, 26 and 20).

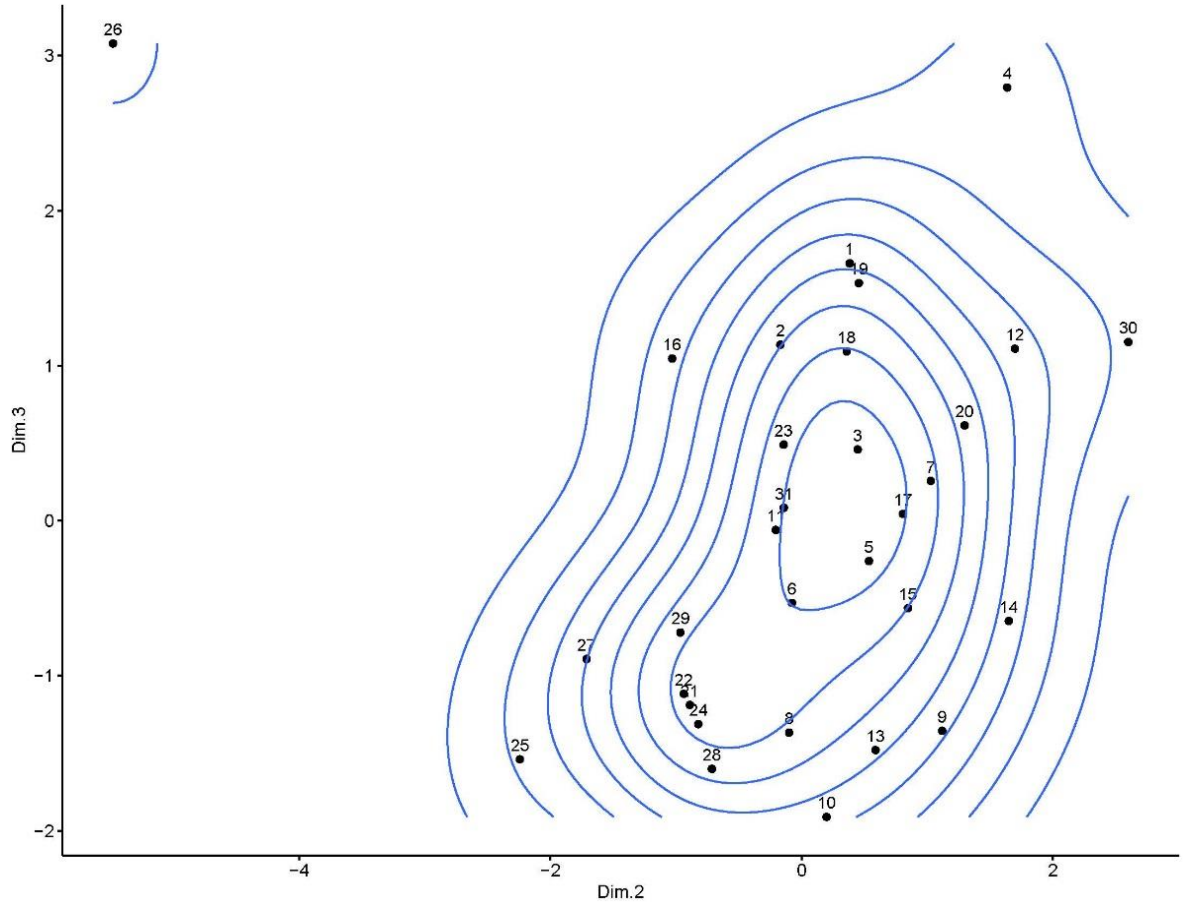


Figure 3

CASE STUDY 7

- Figures 2 and 3 also capture the anomaly represented by these three lots and suggest a possible arrangement of the aggregated data points around two centers.

- In other words, all this means that the set of 31 lots considered is not homogeneous, but that next to a main population there is a sub-population of three lots that stand out from the others and therefore should be further investigated.

- It is evident from these results that the « metrics » made available by *MVA* allow for overall analyzes not otherwise obtainable by studying one parameter at a time !

CASE STUDY 7 – CONCLUSION

- The chosen example has shown how **the *MVA* « metrics»**, which **analyze the data « all together» instead of « one at a time »**, reveal aspects that cannot be captured by studying each parameter individually.

- These too, like the « quality metrics » seen above, are not exhaustive, but provide information which can subsequently be further investigated.

- Although upstream there is a very complex theory, the use of *MVA* methods is, in practice, not so complicated and the use of graphic representation makes everything rather intuitive.

CONCLUSIVE SUMMARY

CONCLUSIVE SUMMARY

- Starting from the FDA stimulus to routinely use quality metrics that go beyond those described in the November 2016 Guidance, but that « manufacturers believe are useful in establishing the quality status of their products and processes », we have taken note of the ever-increasing attention of the Authorities towards quantitative methods of quality measurement.

- The attention of the Authorities follows from the awareness that only through the use of quantitative methods it is possible to intercept the variability of products / processes, control it and therefore ensure quality.

CONCLUSIVE SUMMARY

- Using seven case studies a quick overview was made aimed at evaluating Quality Metrics that are different from one another: from those based on Graphical Methods (DESCRIPTIVE STATISTICS) to those that use Probability Methods (INFERENTIAL STATISTICS) to get to those based on Multivariate Analysis Methods.

- Common features of all Quality Metrics are:
  - ease of use
  - ability to immediately return important information on the product / process and therefore on the quality status of the manufacturer.

CONCLUSIVE SUMMARY

- The analysis of the proposed "case studies" revealed the advantages associated with an increasing use of Quality Metrics, namely:
  - greater knowledge of the process and therefore ability to establish and control its status
  - possibility to use this competence to manage (preventing or justifying them) anomalous events (OOT, OOS, deviations, *etc.*)
  - possibility to document one's knowledge with quantitative and therefore measurable and verifiable topics.

All this is summed up in two words: **ECONOMIC ADVANTAGE!**

Before concluding, please, always remember that

Statistics is the art of learning from data

S.M. Ross, Probability and Statistics, 3rd Italian Edition, Maggioli (2015)

« Quelli che s'innamoran di pratica sanza scienzia son come 'l nocchier ch'entra in navilio sanza timone o bussola, che mai ha certezza dove si vada »

L. da Vinci, Treatise on Painting, Second Part (1540 ca.)