propharma

VALIDATION CENTER

Statistical Intervals for Validation

VC v.23.11

1

Your ProPharma Group Facilitator

- Fred Wiles, ASQ CQE
- Principal Validation Consultant, Life Sciences Consulting
- 30 years experience specializing in process development and validation, statistics and statistical process control, design of experiments, sampling plans and measurement system analysis.
- Previous positions in the pharmaceutical, medical device, and biologics industries include metrology, risk management, computer system validation, equipment qualification, cleaning and process validation.
- Senior Member of American Society for Quality (ASQ), Chicago Section, Northeastern Illinois Section, and Food, Drug, and Cosmetic Division

2

Target Audience

Industries
- Pharmaceutical & Biologics
- Medical Device
- Clinical Studies
- Blood Products

Regions
- Global

Personnel
- Metrology
- Research and Development
- Validation
- Quality Control
- Quality Assurance

3

Webinar Outline

1	• **Introduction**
2	• **Estimating Population Parameters**
3	• **Confidence Intervals**
4	• **Prediction Intervals**
5	• **Tolerance Intervals**
5	• **Conclusion**

4

Why Should I Care?

Regulatory agencies around the world are putting more emphasis on scientific approaches to validation which requires applying more rigorous statistical methodologies to the analysis of validation data.

- Science based
- Make the right decisions
- Demonstrate conformance to specifications
- Avoid the FDA complete response letter

Better Science = Better Outcomes = Less Cost

5

Statistical Intervals Introduction

6

Statistical Nomenclature

Statistical Parameters

Statistical parameters describe the location and spread of a population.

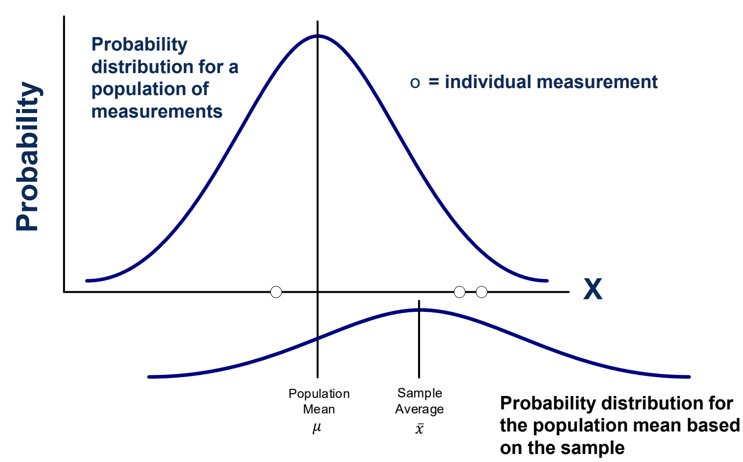
Statistical Parameters may be based on an entire population or estimated based on a sample from the entire population.

Examples:

- Population Mean (μ)
- Population Standard Deviation (σ)
- Sample Average ($\bar{x}$)
- Sample Standard Deviation (s)

7

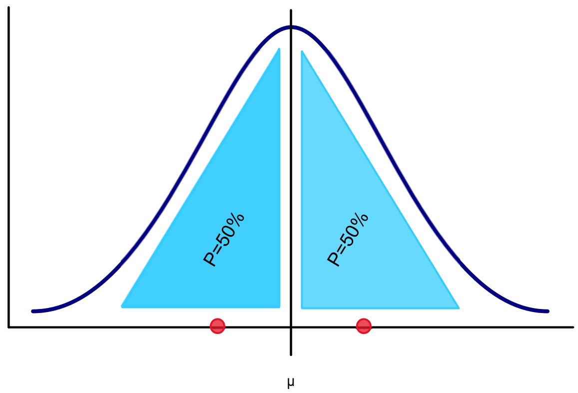
Estimating a Population Parameter Based on Sample Data



8

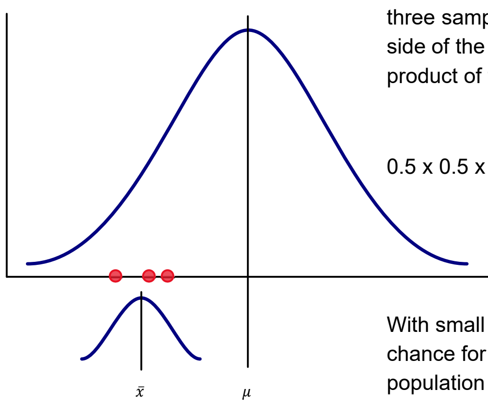
Estimating a Population Parameter Based on Sample Data

In random sampling, each sample has a 50/50 or 0.5 chance of being pulled from one side of the distribution or the other.



9

Estimating a Population Parameter Based on Sample Data



For a sample size of 3, the chance of all three samples being pulled from the same side of the distribution is calculated as the product of the individual probabilities; thus:

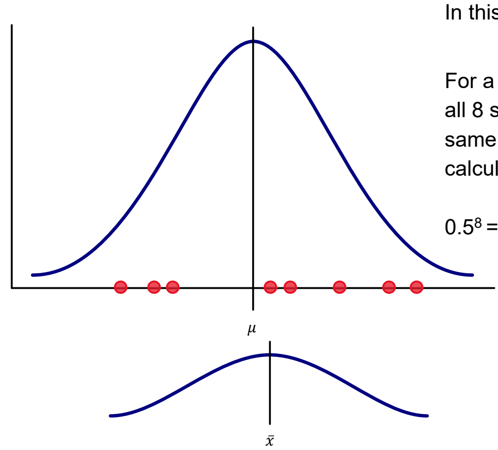
0.5 x 0.5 x 0.5 or 0.5^3 = 12.5%

With small sample sizes, there is a greater chance for clustering on one side of the population distribution.

This can lead to gross estimation errors (sampling errors).

10

Estimating a Population Parameter Based on Sample Data



In this example, the sample size is 8.

For a sample size of 8, The chance that all 8 samples will be pulled from the same side of the population distribution is calculated as:

$0.5^8 = 0.4\%$

Thus, we can decrease the chance for clustering on one side of the population mean by increasing the sample size.

…and improve the accuracy of the parameter estimates

11

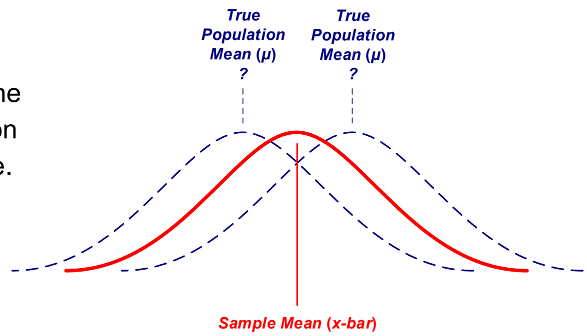
Confidence Intervals

12

Confidence Intervals

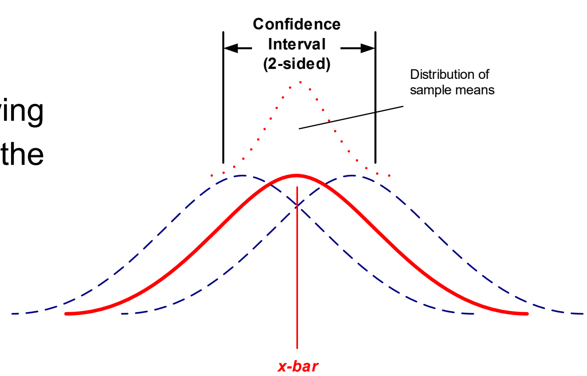
This diagram shows there is uncertainty of knowing the true location of a population parameter.

- The solid curve represents the inferred location of the population distribution based on the sample.
- The dashed curves represent other possible locations of the population distribution.



Confidence Intervals

Example of a two-sided confidence interval representing the uncertainty in knowing the true location of the population mean.



Confidence Intervals

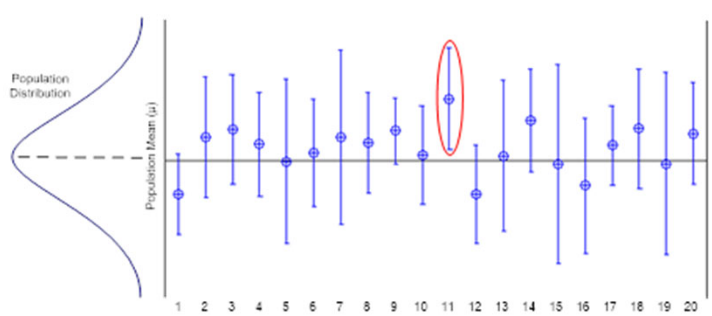
Confidence intervals are relatively straight forward to calculate. A two-sided confidence interval is defined as:

$$\bar{x} \pm t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

where:

$\bar{x}$	is the sample average,
s	is the sample standard deviation,
n	is the sample size,
$1-\alpha$	is the desired confidence level, and
t	is the $100(1-\alpha/2)$ percentile of the t distribution with $n-1$ degrees of freedom.
$\frac{s}{\sqrt{n}}$	is the standard error of the mean

Confidence Interval Interpretation



Interval plot for 20 simulated sample sets. Each interval was calculated providing 95% confidence. Note that one interval (circled in red) out of twenty does not cover the known mean representing 5% of the total number of intervals. This is in line with the expectation that 95% of the intervals will contain the true population mean.

Confidence Intervals (pH Example)

pH example:

Ten samples measured yield the following pH estimates:

$n = 10$, $\bar{x} = 6.52$, $s = 0.11$

The estimated the mean pH along with a statement of the uncertainty of this estimate is to be reported.

The analyst wants a two-sided confidence interval using α = 0.05 (confidence level = 95%).

The value for the *t* distribution is found using one-half of the value for α; that is, 0.05/2 = 0.025.

Sample #	pH
1	6.39
2	6.51
3	6.54
4	6.42
5	6.52
6	6.47
7	6.69
8	6.37
9	6.63
10	6.62
Mean ($\bar{x}$)	6.52
Sample SD (s)	0.11

Validation Center™ Copyright © 2022, ProPharma Group 17

17

Confidence Intervals (pH Example)

This value along with the degrees of freedom (*n-1* = *9* in this case) will be used to look up the value of the Student's *t* distribution.

The *t* value for this example is 2.262.

$$CI = \bar{x} \pm t_{1-\alpha/2,\,n-1}\frac{s}{\sqrt{n}} = 6.52 \pm 2.262\frac{0.11}{\sqrt{10}} = 6.52 \pm 0.08$$

The lower interval bound is 6.52 - 0.08 = 6.44;

the upper interval bound is 6.52 + 0.08 = 6.60.

Validation Center™ Copyright © 2022, ProPharma Group 18

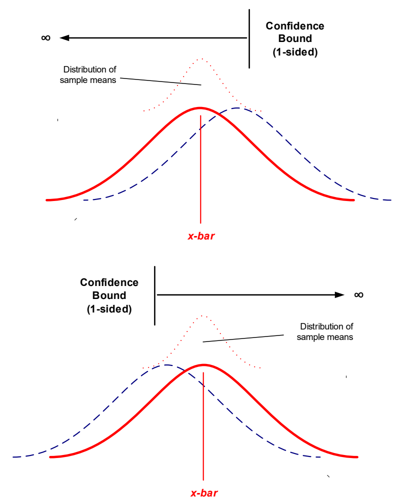
18

Confidence Intervals (One Sided)

One-sided confidence interval to indicate the uncertainty of knowing the true population mean relative to a single specification limit.

A single-sided confidence interval can be calculated simply by substituting α for α/2.

One-sided confidence intervals should not be used to show the uncertainty of the mean relative to a two-sided specification limit.



19

Prediction Intervals

FW0

20

Slide 20

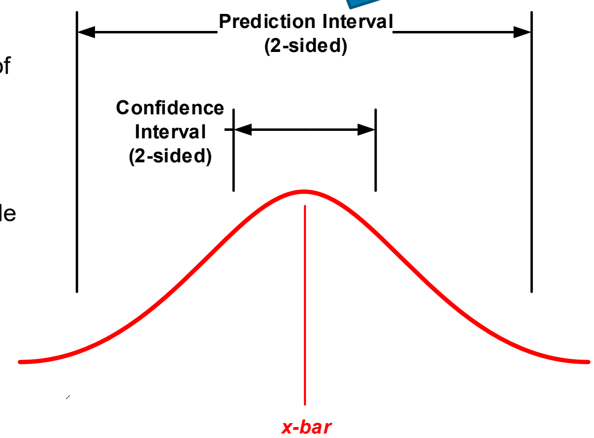
FW0 Add a slide to include how prediction intervals are used in validation
Frederick Wiles, 2022-02-21T18:21:52.398

Prediction Intervals

Prediction intervals are wider than confidence intervals

Prediction Intervals represent the uncertainty of predicting the value of

- a single future observation or
- a fixed number of multiple future observations



21

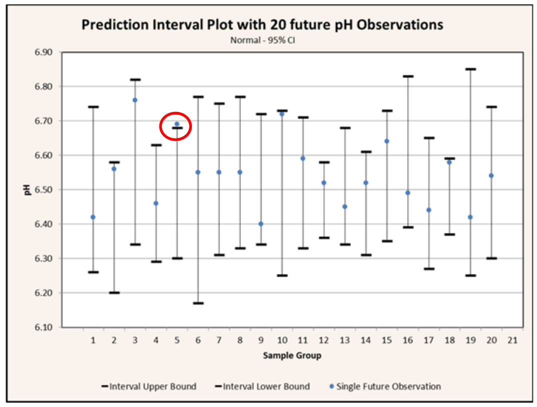
Prediction Interval Interpretation

✗	**Interval calculated from a single sample contains a specified percentage of future observations**
✓	Intervals calculated for several successive samples from the same population contain a future observation a specified percentage of the time.

22

Prediction Interval Interpretation

- Upper and lower prediction interval bounds for 20 samples of 10 pH measurements drawn from the same population.
- Includes 20 individual future observations.
- Notice that 1 out of the 20 single future pH readings (circled in red) is outside of the associated prediction interval region.



- This is in alignment with the 5% of intervals not expected to contain a single future observation for a confidence level of 95%.

Prediction Interval for Normal Data

The formula for a prediction interval is nearly identical to the formula used to calculate a confidence interval.

Recall that the formula for a two-sided confidence interval is:

$$\bar{x} \pm t_{1-\alpha/2,n-1} s\sqrt{1/n}$$

To calculate a prediction interval, add 1 to the $1/n$ term under the square root symbol to account for the variability of a single observation about the mean.

Doing so yields the prediction interval formula for normally distributed data:

$$\bar{x} \pm t_{1-\alpha/2,\, n-1} s\sqrt{1 + 1/n}$$

Prediction Interval for Normal Data (pH Example)

An analyst wants to know, based on a sample, the two-sided prediction interval within which a single future pH observation is likely to lie with some level of confidence.

The average pH, $\overline{x}$, is 6.52

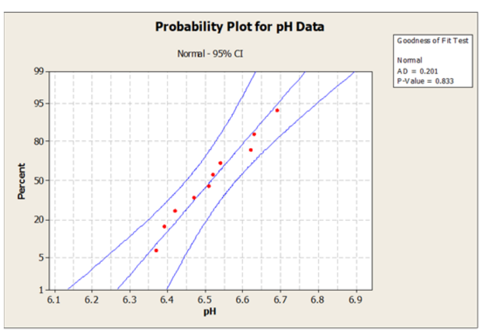
The sample standard deviation, s , is 0.11

The confidence level chosen is 95% (α=0.05)

Sample #	pH
1	6.39
2	6.51
3	6.54
4	6.42
5	6.52
6	6.47
7	6.69
8	6.37
9	6.63
10	6.62
Mean ($\overline{x}$)	6.52
Sample SD (s)	0.11

Prediction Interval for Normal Data

Normal Probability Plot for pH Data.



To check this assumption, a normal probability plot is constructed for the data.

- The sample pH data closely follow a straight line
- The P-value is much greater than 0.05.

There is no reason to reject the assumption that the assay data are normally distributed.

Prediction Interval for Normal Data

Prediction Interval for a single Future Observation:

Find the value for *t:*

$$t_{1-\alpha/2,n-1} = t_{1-0.05/2,\,10-1} = 2.262$$

Calculate the 95% prediction interval
(Exact for a single future observation).

$$\bar{x} \pm t_{1-\alpha/2,n-1} s\sqrt{1+1/n} = 6.52 \pm 2.262(0.11)\sqrt{1+1/10}$$

$$= 6.52 \pm 0.26$$

27

Prediction Interval for Normal Data

Prediction Interval for Multiple Future Observations:

Find the value for *t:*

$$t_{1-\alpha/(2k),n-1} = t_{1-0.05/(2\times 5),\,10-1} = 3.250$$

Calculate the 95% prediction interval
(approximation for multiple future observations)

$$\bar{x} \pm t_{1-\alpha/(2k),n-1} s\sqrt{1+1/n} = 6.52 \pm 3.250(0.11)\sqrt{1+1/10}$$

$$= 6.52 \pm 0.37$$

28

One-sided Prediction Interval for Normal Data

For a single future observation:

$$\bar{x} \pm t_{1-\alpha/2,\,n-1}\, s\sqrt{1+1/n} \quad \rightarrow \quad \bar{x} + t_{1-\alpha,\,n-1}\, s\sqrt{1+1/n}$$

and

$$\bar{x} - t_{1-\alpha,\,n-1}\, s\sqrt{1+1/n}$$

For multiple future observations:

$$\bar{x} \pm t_{1-\alpha/(2k),\,n-1}\, s\sqrt{1+1/n} \quad \rightarrow \quad \bar{x} + t_{1-\alpha/(k),\,n-1}\, s\sqrt{1+1/n}$$

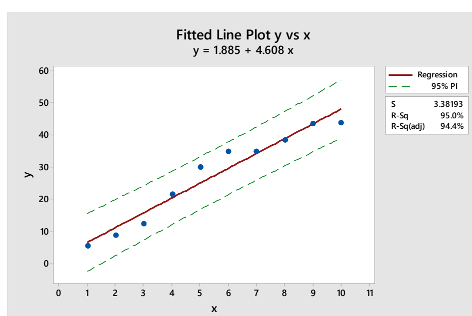
and

$$\bar{x} - t_{1-\alpha/(k),\,n-1}\, s\sqrt{1+1/n}$$

Prediction Interval for Regression

In regression statistics, a prediction interval defines a range of values within which a response is likely to fall given a specified value of a predictor.

Normally distributed data are statistically independent of one another



Regressed data are dependent on a predictor value and therefore not normally distributed.

Prediction intervals applied to regression statistics are considerably more involved to calculate then are prediction intervals for normally distributed data.

Prediction Interval for Simple Linear Regression

The uncertainty represented by a prediction interval includes uncertainties in:

- True location of the new observation
- True location of the population mean
- True location of the regression parameters

The uncertainty estimates must be combined using root-sum-of-squares to yield the total uncertainty, s_p.

- Variation contributed by the response is s
- Variation contributed by the estimate of the mean response is $s/\sqrt{n}$
- Variation contributed by the regression parameters is s_f

$$s_p = \sqrt{s^2 + \frac{s^2}{n} + s_f^2}$$

$$y = \beta_0 + \beta_1 x$$

$$s = \left(\frac{S_{yy} - \beta_1 S_{xy}}{n-2}\right)^2$$

31

Prediction Interval for Simple Linear Regression

Expressing $s_f{}^2$ in terms of the predictors:

$$s_f^2 = s^2 \frac{(x_{Predictor} - \bar{x})^2}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

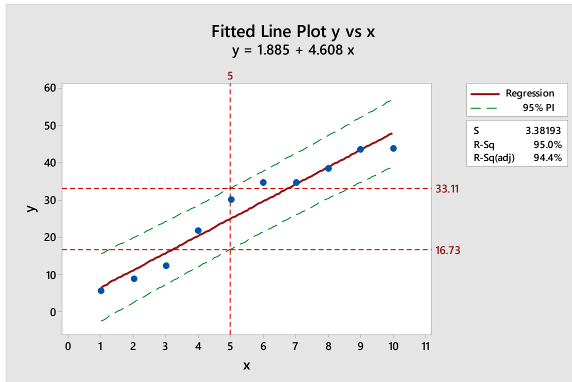
Adding the other two terms under the radical and taking the square root of s^2 yields the two-sided prediction interval formula for the linear regressed response variable:

$$y \pm t_{\alpha/2,n-2}\ s \sqrt{1 + \frac{1}{n} + \frac{(x_{Predictor} - \bar{x})^2}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}}$$

32

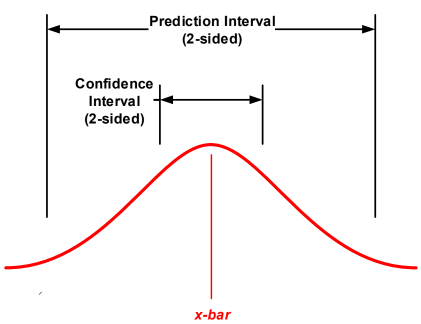
Prediction Interval for Simple Linear Regression

Fitted Line Plot plot with prediction interval bounds for the estimated response, *y* . The predictor value in this case is 5.



33

Prediction Intervals Summary



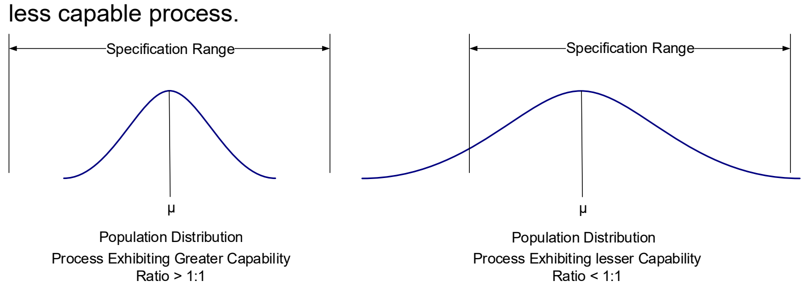
- Prediction intervals quantify the uncertainty of future observations from a population.
- Prediction intervals are wider than confidence intervals.
- Prediction intervals are best suited for regression statistics.
- Prediction intervals are more susceptible to the assumption of normality.

34

Tolerance Intervals

Tolerance Intervals

- Interval to cover a specified proportion of a population distribution with a given confidence.
- Tolerance intervals are closely related to measures of process capability, i.e., the ratio of the specification range for a measured characteristic to the observed variation associated with the characteristic.
- Larger ratios indicate a more capable process while smaller ratios indicate a less capable process.



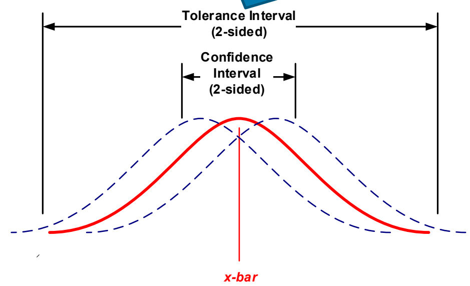
Tolerance Intervals

Tolerance intervals are wider than both confidence and prediction intervals

Graphical depiction of a tolerance interval.

The solid line represents the actual population distribution.

The dotted line distributions result from the uncertainty of knowing the true location of the population mean.



The uncertainty of knowing the true location of the mean defined by the confidence interval.

Tolerance intervals also account for the uncertainty associated with estimating the true population variance.

37

Tolerance Intervals

Two-sided Tolerance intervals are calculated as:

$$\bar{x} \pm k_2 s$$

$\bar{x}$ is the sample mean;

s is the sample standard deviation; and

k_2 is a factor for a two-sided tolerance interval defining the number of sample standard deviations required to cover the desired proportion of the population.

38

Tolerance Intervals

Exact values of k_2 are tabularized in Hahn & Meeker (1991) and ISO 16269-6.

Exact values of k_2 can be calculated iteratively using a numerical integration.

This calculation, being very complex, is beyond the scope of this presentation.

A reasonable approximation of k_2 can be obtained using a formula originally proposed by Howe (1969) and later improved by Guenther (1977):

$$k_2 = z_{(1-p)/2}\sqrt{\frac{(n-1)\left(1+\frac{1}{n}\right)}{\chi^2_{\gamma,\,n-1}}\left(1+\frac{(n-3)-\chi^2_{\gamma,\,n-1}}{2(n+1)^2}\right)}$$

Where:

n is the sample size

$Z_{1-p/2}$ is the standard normal variate corresponding to one minus the proportion of the population to be covered divided by two.

$\chi^2_{\gamma,n-1}$ is the critical value of the chi-square distribution with n-1 degrees of freedom surpassed with probability γ, the statistical confidence.

Tolerance Intervals

Example:

Suppose we have assay data for ten randomly-selected containers of a drug product and want to determine if the process is capable.

The assay specification range is

9.000 – 11.000 mg

Calculate a tolerance interval to cover 99% of the data population with 95% confidence.

Assume that the variation contributed by measurement method is insignificant with respect to the process variation.

Container #	Assay (mg)
1	9.925
2	9.681
3	10.061
4	10.319
5	10.300
6	10.433
7	9.454
8	9.941
9	10.274
10	9.728
Average ($\bar{x}$)	10.012
Sample SD (s)	0.3231

Assay data for ten randomly-selected containers of a drug product

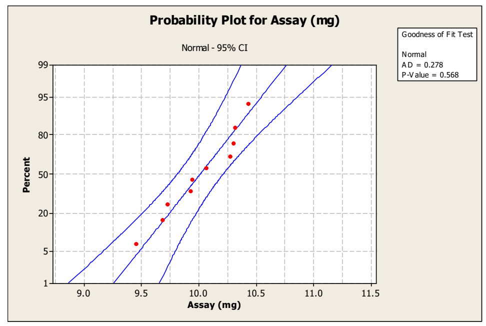
Tolerance Intervals

Tolerance intervals are calculated assuming the data are normally distributed

To check this assumption, a normal probability plot is constructed for the data.

- The sample assay data closely follow a straight line
- The P-value is much greater than 0.05.

There is no reason to reject the assumption that the assay data are normally distributed.



Tolerance Intervals

For a sample size of 10, k_2 to calculate a 95% tolerance interval to cover 99% of the population is 4.437.

Knowing the value for k_2, the tolerance interval is then calculated as:

$$Tolerance\ Interval = \bar{x} \pm k_2 s = 10.012 \pm 4.437 \times 0.3231 = 10.012 \pm 1.434$$

The target value of the process is 10.000 mg and the specification limits are 9.000 to 11.000, so the capability based on a 95/99 two-sided tolerance interval is:

$$Capability_{Process} = \frac{Specification\ Range}{Tolerance\ Interval\ Range} = \frac{2.000}{2 \times 1.434} = 0.697$$

One-sided Tolerance Intervals

One-sided tolerance interval

Use when the process capability is measured relative to a single-sided limit:

- Measured characteristic must be above a minimum specification limit
- Measured characteristic must be below a maximum specification limit

The calculation of an approximate k factor for a one-sided tolerance interval comes from a formula described by Natrella (1963):

$$k_1 = \frac{z_p + \sqrt{z_p^2 - ab}}{a} \qquad \text{Where: } \begin{cases} a = 1 - \dfrac{z_\gamma^2}{2(n-1)} \\ b = z_p^2 - \dfrac{z_\gamma^2}{n} \end{cases}$$

$z = z$ score, the number of standard deviations a value is away from the mean
$\gamma = 1 - \alpha$, the confidence level chosen
$p =$ the proportion of the population to be covered

One-sided Tolerance Intervals

For example, the maximum limit on the number of particles greater than a certain size within a liquid suspension drug product is 100.

10 containers are randomly pulled from the finished product batch and tested for particle counts.

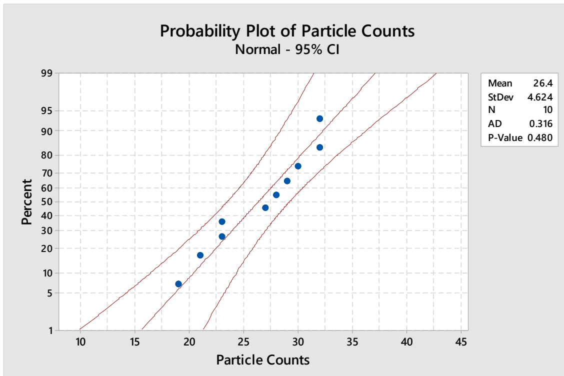
Calculate a single-sided upper tolerance bound to cover 99% of the particle count population with 95% confidence

Assume that the variation contributed by measurement method is insignificant with respect to the process variation.

Container #	Particle Counts
1	23
2	30
3	23
4	27
5	29
6	32
7	28
8	19
9	21
10	32
Mean ($\bar{x}$)	26
Sample SD (s)	5

Particle count data for ten randomly selected containers

One-sided Tolerance Intervals



A normal probability plot of the data shows there is no reason to reject the assumption that the data are normally distributed.

One-sided Tolerance Intervals

For a sample size of 10, k_1 to calculate a 95% upper tolerance bound to cover 99% of the population is 3.937.

With the value for k_1 in hand, the upper tolerance bound is:

$$Upper\ Tolerance\ Bound = \bar{x} + k_1 s = 26 + 3.937 \times 5 = 26 + 20 = 46$$

Thus, the upper 99% single-sided tolerance bound calculated with 95% confidence is 46. What this means is that for 95 out of 100 samples of size 10 taken from the same population, we would only expect 1% to have a particle count greater than 46.

The capability then is:

$$Capability_{Process} = \frac{Upper\ Specification\ Limit\ - Mean\ Count_{Sample}}{Upper\ Tolerance\ Bound - Mean\ Count_{Sample}} = \frac{100 - 26}{46\ - 26} = 3.7$$

One-sided Tolerance Intervals

If we were instead concerned with meeting a lower limit on particle counts rather than staying below a maximum value, we would simply subtract k1 from the mean instead of adding as we did in the example above.

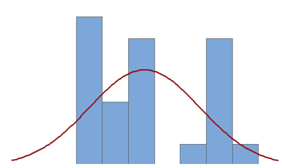
In this case, the lower 99% single-sided tolerance bound calculated with 95% confidence is 6. If the lower specification limit in this case is 10, then the capability is:

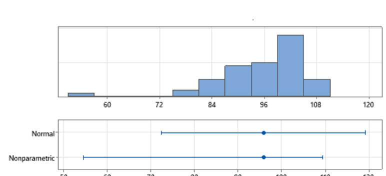
$$Capability_{Process} = \frac{Mean\ Count_{Sample} - Lower\ Specification\ Limit}{Mean\ Count_{Sample} - Lower\ Tolerance\ Bound} = \frac{26 - 10}{26\ - 6} = 0.8$$

Distribution Free Intervals

Data are not always normally distributed

- Discrete Data (e.g., Binary or Categorical)
- Data that cannot be fit to a specific distribution (e.g., bimodal distribution)





Intervals can be constructed based on order statistics

- Difficult to get desired confidence level with small sample sizes
- Usually much longer than those based on a specific distribution

Conclusion

The appropriate statistical interval to use depends on the question to be answered:

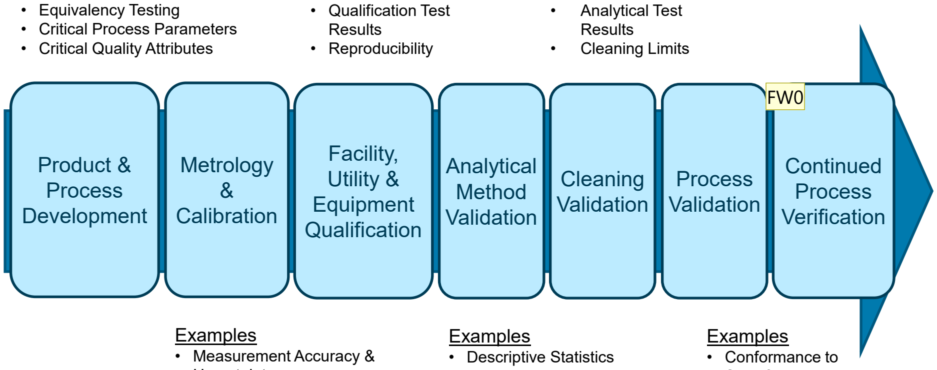
Interval	Contains	Uses
Confidence	• Population parameter (μ , σ)	• Uncertainty associated with a population estimate $(\bar{x}, s)$
Prediction	• A single observation or multiple future observations • Range of response values given a specified value of a predictor	• Probability of one or a small number of future events • Uncertainty of a predicted response
Tolerance	• Specified proportion of a population	• Population limits • Process capability

Interval Uses in a Validation Program

Examples
- Design of Experiments
- Analysis of Variance
- Equivalency Testing
- Critical Process Parameters
- Critical Quality Attributes

Examples
- Qualification Test Results
- Reproducibility

Examples
- Analytical Test Results
- Cleaning Limits



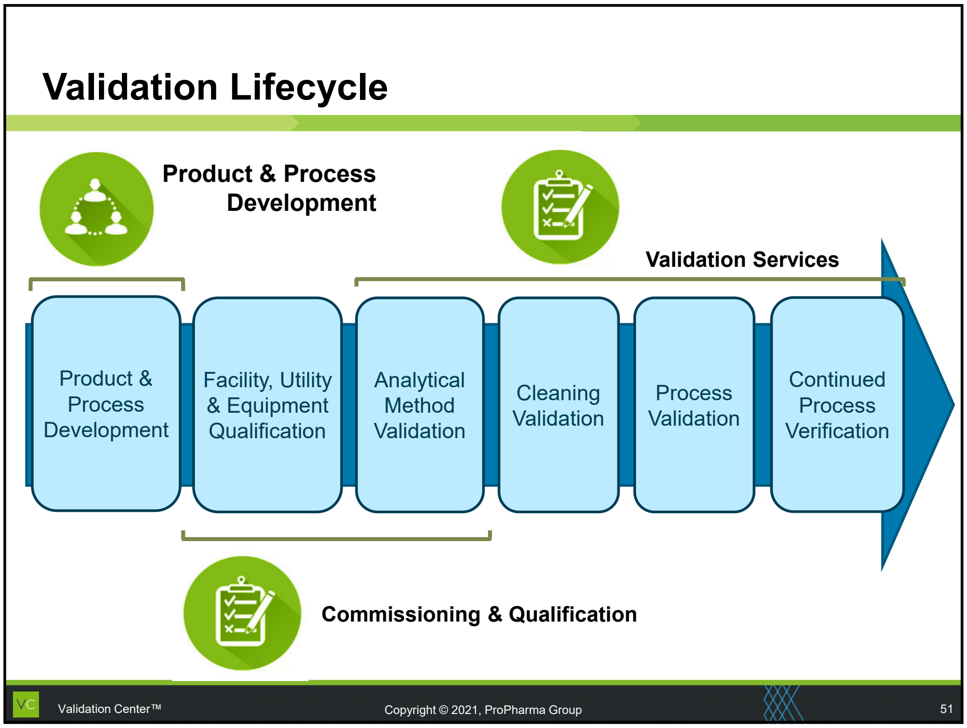
Examples
- Measurement Accuracy & Uncertainty
- Gage R&R

Examples
- Descriptive Statistics
- Repeatability
- Accuracy & Linearity
- Intermediate Precision
- Equivalency Testing

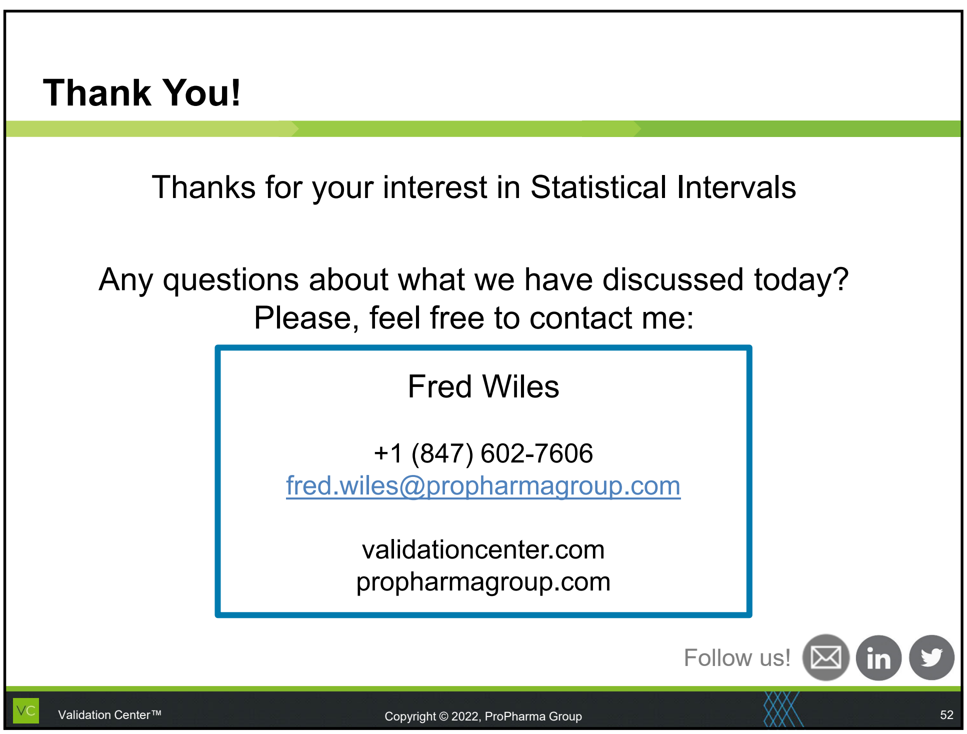
Examples
- Conformance to Specifications
- Process Capability
- Sampling Plans

Slide 50

FW0 Add CPV
Frederick Wiles, 2022-03-04T17:10:38.078


Validation Lifecycle
Product & Process Development
Validation Services
Product & Process Development
Facility, Utility & Equipment Qualification
Analytical Method Validation
Cleaning Validation
Process Validation
Continued Process Verification
Commissioning & Qualification
Validation Center™
Copyright © 2021, ProPharma Group
51

51


Thank You!
Thanks for your interest in Statistical Intervals
Any questions about what we have discussed today?
Please, feel free to contact me:
Fred Wiles
+1 (847) 602-7606
fred.wiles@propharmagroup.com
validationcenter.com
propharmagroup.com
Follow us!
Validation Center™
Copyright © 2022, ProPharma Group
52

52