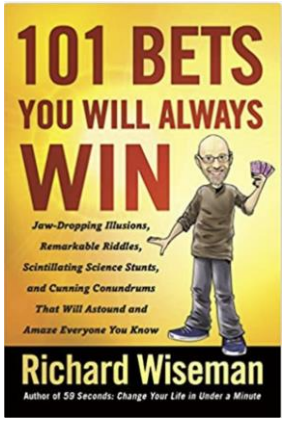
Comparability and Equivalence

Navy Additive Manufacturing

- Initiative with Submarine Industrial Base to rapidly integrate Additive Manufacturing qualified parts into supply and manufacture system
- “Qualified” material requires testing and objectives are often to be “as good as” the traditional or wrought version
- Many methods used across AM community to establish equivalency but not codified

Equivalence Testing



Equivalence Introduction

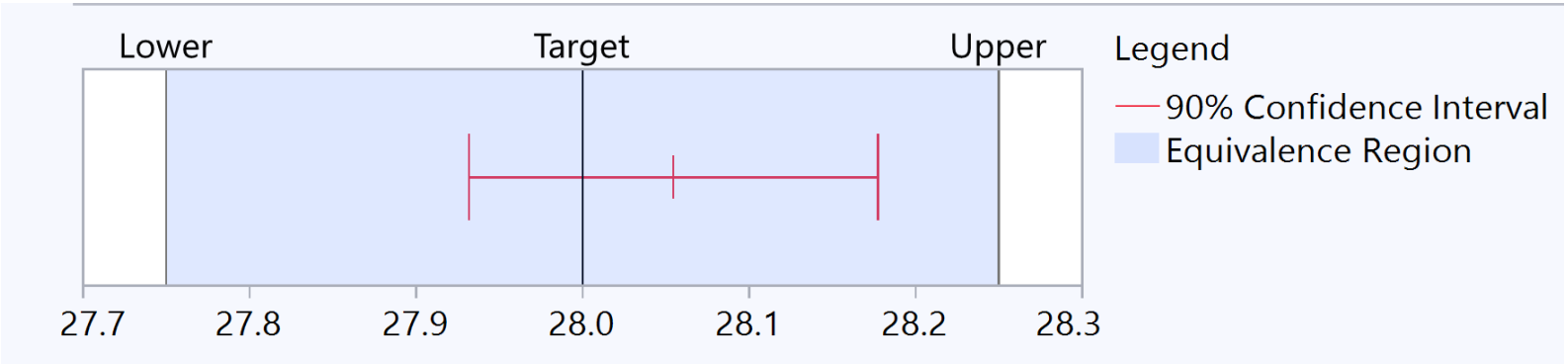
- Problem: Some batches are experiencing tensile that differs from 28 ksi . A consultant has recently implemented design changes and SPC methods.
- Data: We sample n= 20 coupons and they have a mean of $\bar{y}$ and standard deviation of *s*.
- Method: Two-sided one-sample t-test
- Conclusion: Consultant …"with a t-statistic of 1.67 and p-value of 0.15 we have proven our mean is equal to 28 ksi"

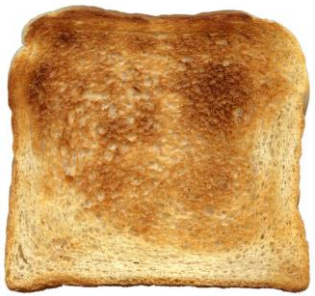
- Question: Have we really established the equivalency of the mean of 28?
- Practitioners may criticize us for the stats term ***fail to reject the null*** believing you either accept the null or accept the alternate hypothesis; but it is quite descriptive.

Equivalence and Comparability

- Failing to reject is a good start! Target value should fall within Confidence Intervals too.
- Need to determine what difference Δ from 28 is practically significant. Is it 0.000001, 0.1, 1, or 10 ksi?
- Conduct two one-sided tests (TOST) by adding and subtracting this delta value to the desired target (28 ksi).

- Test of equivalence for a delta of 0.25

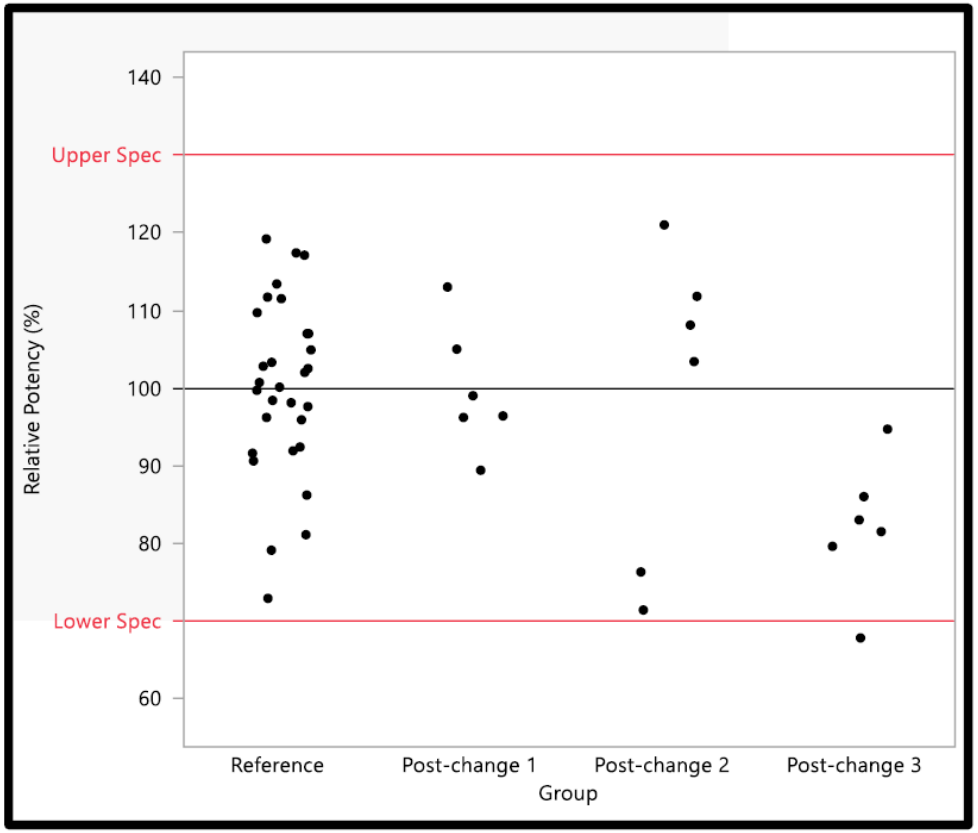




Reference and Post-Change

Which of these post-change processes are comparable with the reference?

- Scale and labels are important!
- We must always be considering the scientific relevance.



Comparability Tools

1. Side-by-side plots
2. Statistical tests
   - Equivalence testing of means
   - Non-inferiority of standard deviations
3. Quality ranges
   - Min-Max intervals
   - Tolerance intervals
   - 3-sigma intervals
   - Risk-based side-by-side intervals

Specifications

Meeting release specifications is not sufficient to demonstrate comparability

- Release specifications are as broad as possible without compromising efficacy/safety
- Comparability acceptance criteria must be narrow enough to detect meaningful change in product or process intermediate or process.

Considerations for Approach

Protect Navy/operators from consequences of concluding comparability when Reference (R) and New (N) are not comparable.

Protect vendors from consequences of concluding lack of comparability when R and N are comparable.

Incentivize vendors to acquire process knowledge concerning N (and perhaps R in similarity).

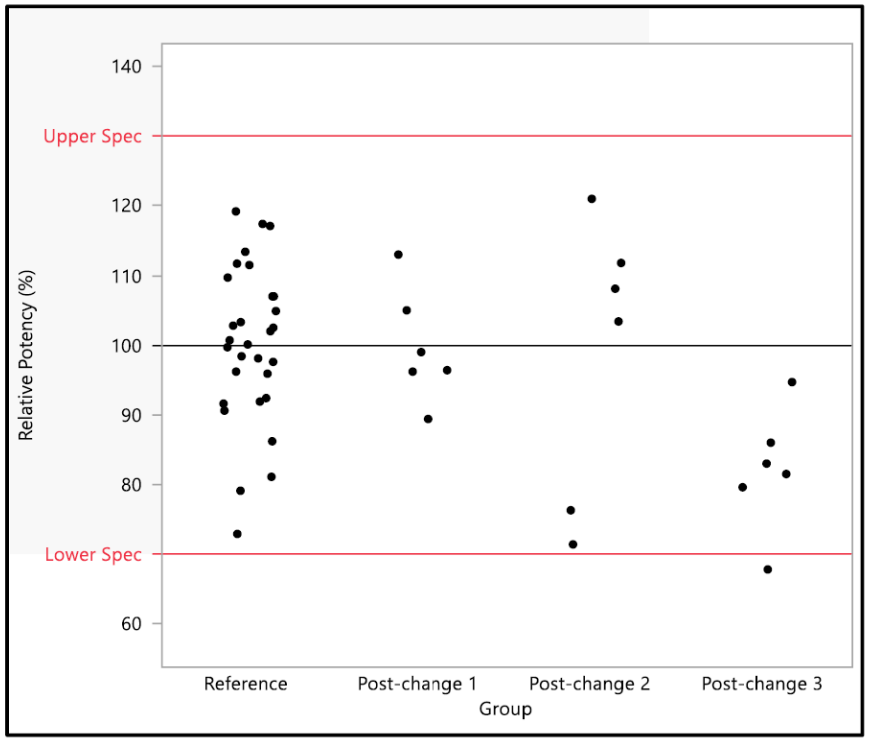
Enable decision making with practical sample sizes.

- The practical sample size for the qualification of a reference standard will be quite different than for the transfer of a manufacturing site.

Considerations for Approach

- Examine the entirety of the process distribution (mean and standard deviation).
- Statistical rigor should consider criticality and measurement scale of the attribute.
- Demonstrate robustness to violations of assumptions.
- Be transparent, easy to explain, and easy to compute by scientists with no formal statistical training.

Case Problem



Group	Mean	Standard deviation
Reference	100	10
Post-change 1	100	10
Post-change 2	100	20
Post-change 3	80	10

Defining Comparability

Key Concepts from EMA 2021 Reflection Paper

Similarity Condition

- What does it mean to be comparable?

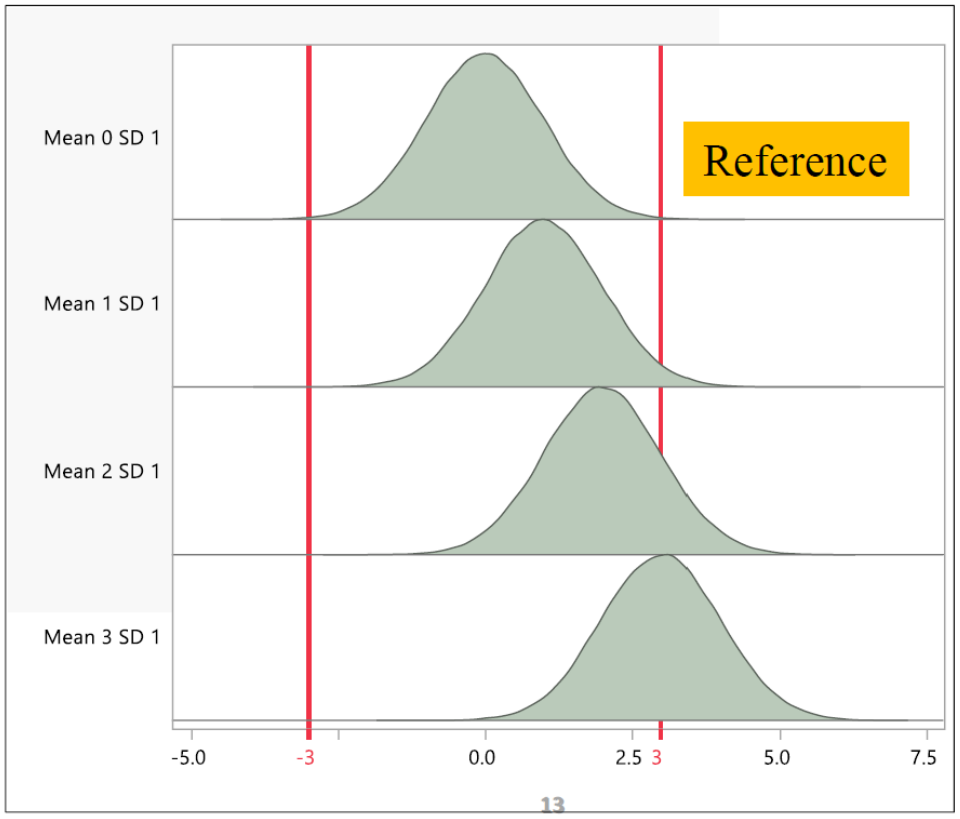
Similarity Criterion

- What evidence is needed to declare comparability?
- Based on expected operating characteristics (power curves)

Defining Comparability

Similarity Conditions

Shifts in process mean

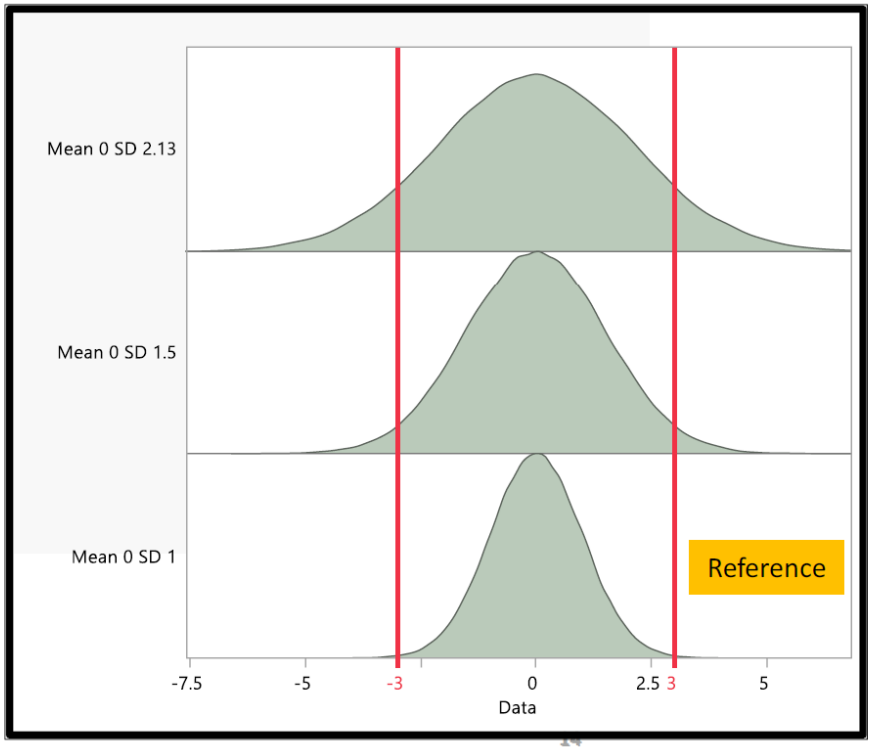


Distributions have different means but the same standard deviation.

Defining Comparability

Similarity Conditions

Shifts in process standard deviation



Distributions have different standard deviations but the same mean.

Defining Comparability

How Can We Quantify These Visualizations?

The following parameters will be used to describe differences between processes

$$K_1 = \frac{\left|\text{Difference in means}\right|}{\text{Standard deviation reference}}.$$ This is the effect size

$$K_2 = \frac{\text{Standard deviation new}}{\text{Standard deviation reference}}.$$ This is the ratio of standard deviations

$K_1 = 0$ and $K_2 = 1$ means R and N are identical

Defining Comparability

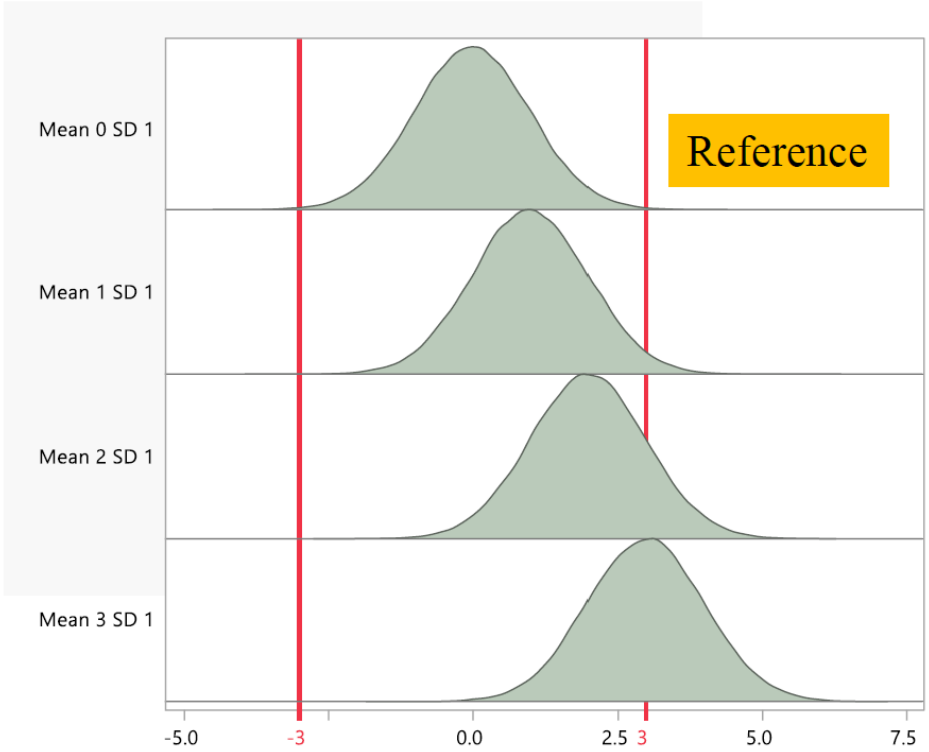
Similarity Conditions

Shifts in process mean with K_2=1

K_1=1

K_1=2

K_1=3



Dissimilarity increases from top to bottom.

Where do we draw the line between comparable and non-comparable?

Defining Comparability

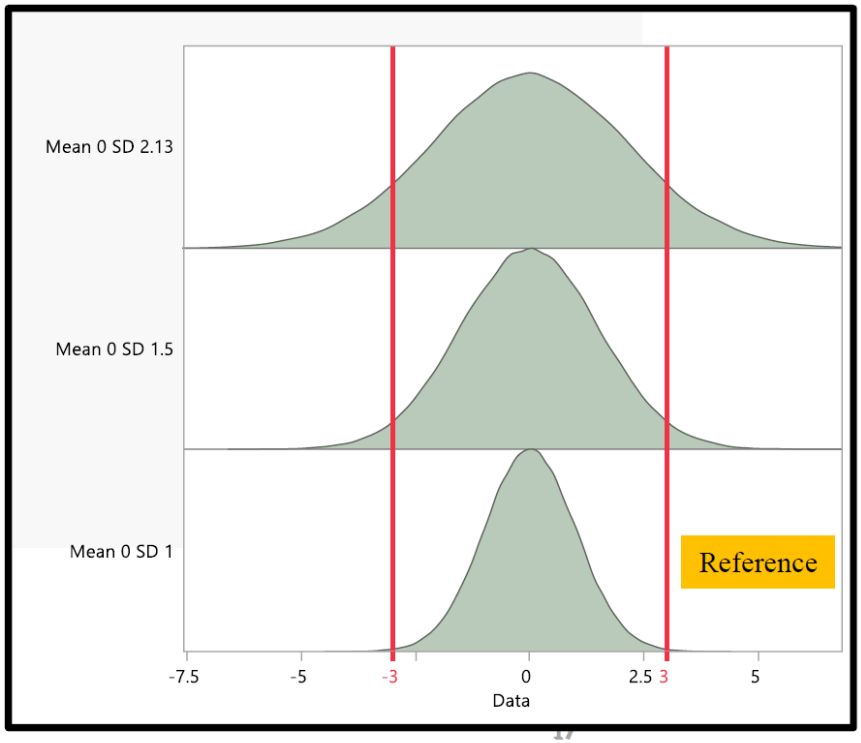
Similarity Conditions

Shifts in process standard deviation with $K_1=0$

$K_2=2.13$

$K_2=1.5$

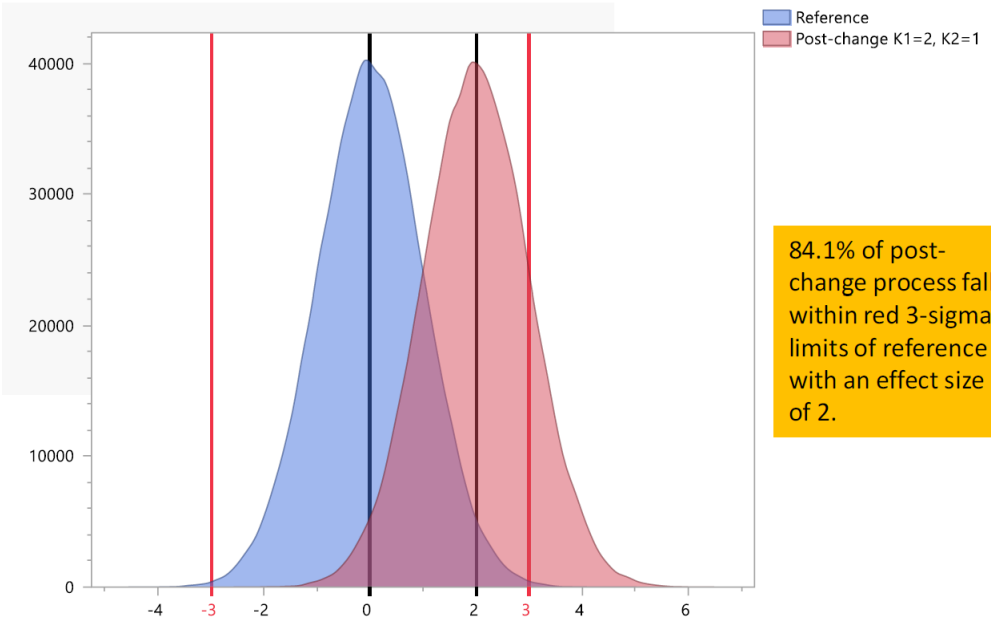
$K_2=1$



Dissimilarity increases from bottom to top.

Where do we draw the line between comparable and non-comparable?

Defining Comparability



84.1% of post-change process fall within red 3-sigma limits of reference with an effect size of 2.

A useful metric for defining an acceptable similarity condition is the "Capability" of the post-change distribution defined as the probability that a value from the post-change process falls within 3-sigma limits of the reference process.

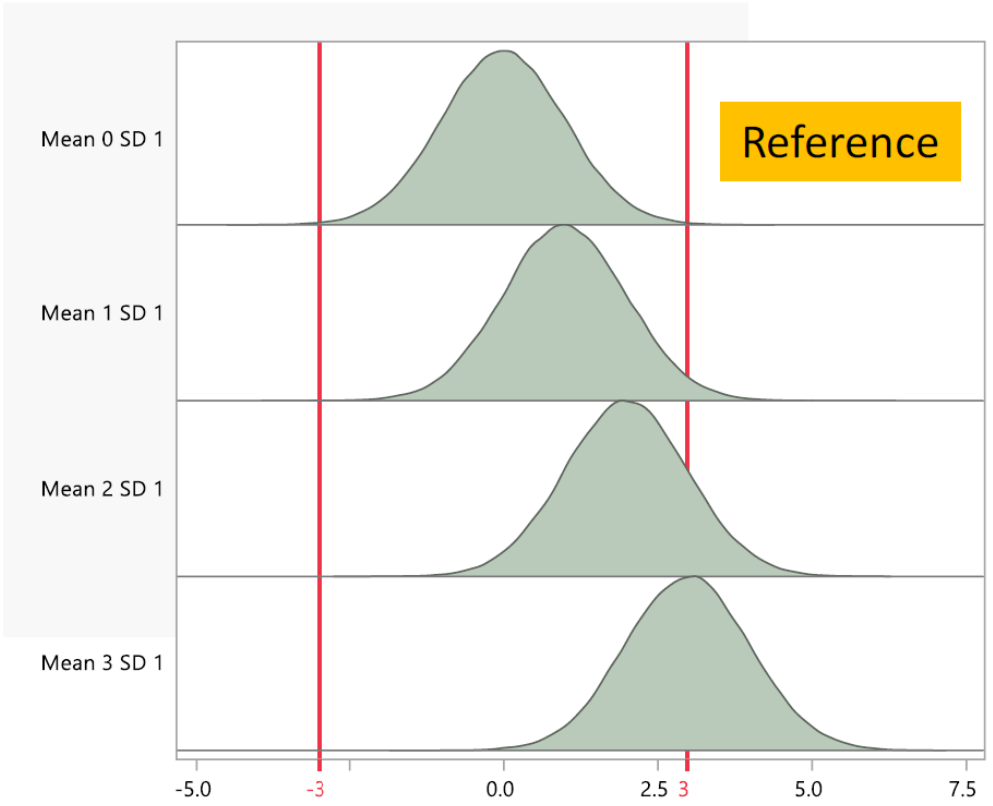
Defining Comparability

Capability with shift in means

K_1=1
Capability=97.7%

K_1=2
Capability=84.1%

K_1=3
Capability=50%



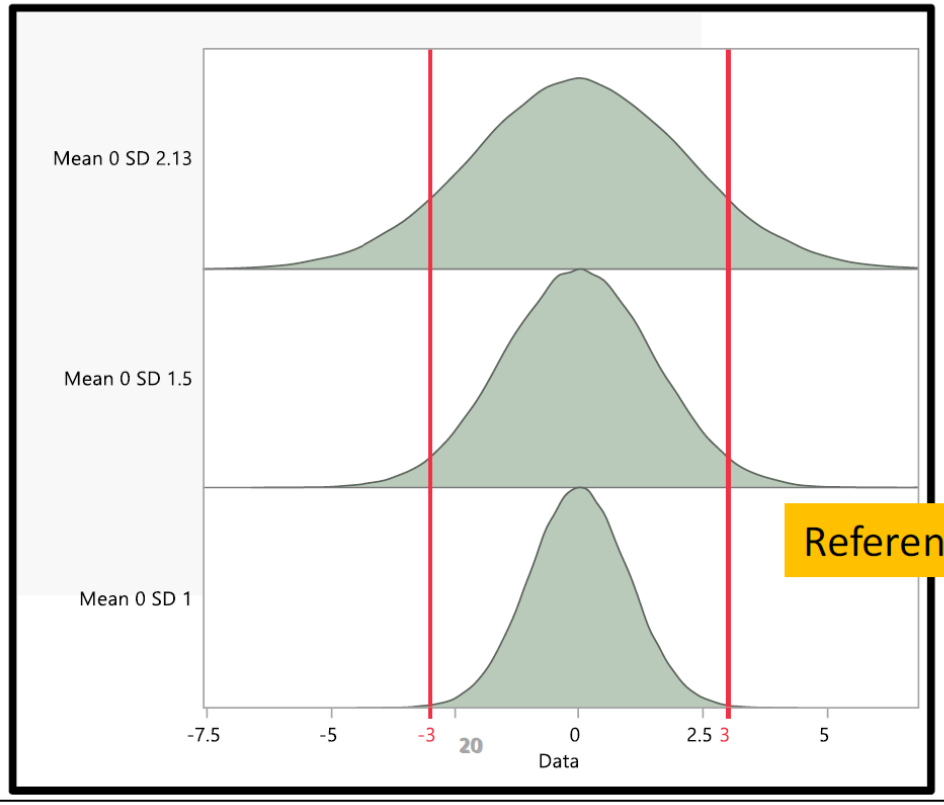
Defining Comparability

Capability with change in standard deviations

K_2=2.13
Capability=84.1%

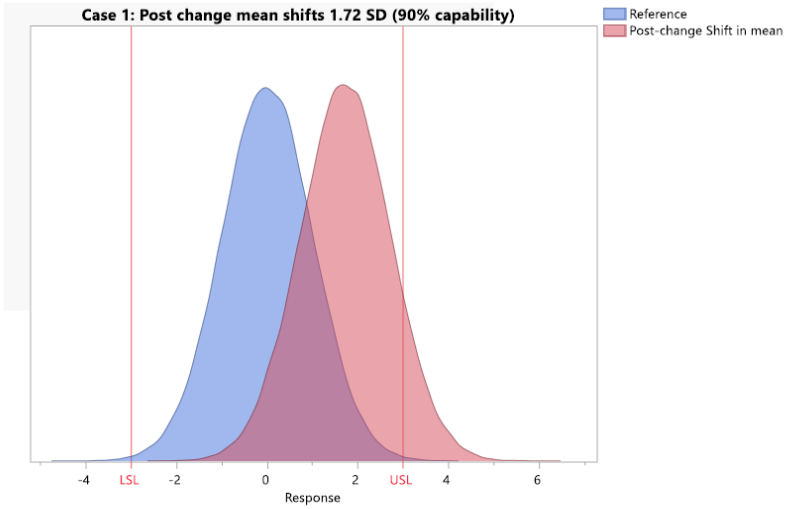
K_2=1.5
Capability=95.0%

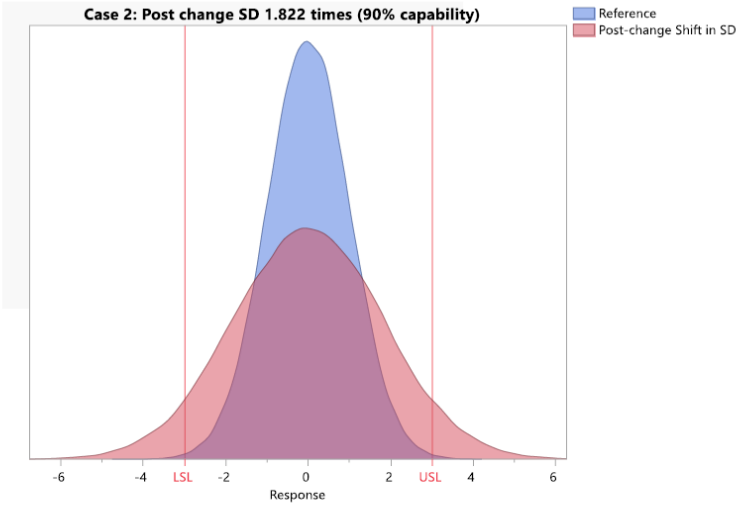
K_2=1
Capability=97.7%



Defining Comparability

Two Processes with 90% Capability Relative to Reference





These two process relationships should be treated comparably because they are equally capable.

Defining Comparability

How Can We Quantify These Visualizations?

1. Mathematically, the capability is

$$\text{Capability} = \Phi\left(\frac{3-K_1}{K_2}\right) - \Phi\left(\frac{-3-K_1}{K_2}\right)$$

$$K_1 = \frac{\mu_N - \mu_R}{\sigma_R}$$

$$K_2 = \frac{\sigma_N}{\sigma_R}$$

$\Phi(\bullet)$ is standard normal cdf

2. JMP script "Calculate Capability"
3. If actual LSL and USL values are available, this equation is modified to replace the values 3 and -3 with the number of standard deviations between the mean of the reference process and the closest specification.
4. Subject matter experts in combination with specification limits and risk profiles can be used to determine an acceptable similarity condition.

Defining Comparability

Process Capabilities as Functions K_1 and K_2

Effect size (K1) with K2=1	Capability	Sigma ratio (K2) with K1=0		Effect size (K1) with K2=1	Capability	Sigma ratio (K2) with K1=0
0.00	0.997	1		1.20	0.964	1.43
0.40	0.995	1.07		1.30	0.955	1.50
0.47	0.994	1.09		1.40	0.945	1.56
0.58	0.992	1.13		1.50	0.933	1.64
0.71	0.989	1.18		1.60	0.919	1.72
0.80	0.986	1.22		1.72	0.900	1.82
0.90	0.982	1.27		1.80	0.885	1.90
1.00	0.977	1.32		1.90	0.864	2.01
1.10	0.971	1.37		2.00	0.841	2.13

What do you define as “Comparable”?

Defining Comparability

What is an Appropriate Similarity Condition ?

A statistical test should assign a "small" probability of claiming comparability at the similarity condition.

Thus, the similarity condition should represent a "clearly unacceptable condition" as opposed to a "marginally acceptable condition".

- e.g., 95% capability might be marginally acceptable and 80% is clearly unacceptable.

Subject matter experts (SME) need to determine an appropriate similarity condition (not the statistician!).

Reasonable definitions are needed to enable decision making with practical sample sizes.

Equivalence Testing of Means

- Test the following set of hypotheses

$$H_0: |\mu_R - \mu_N| \geq \text{EAC}$$

$$H_1: |\mu_R - \mu_N| < \text{EAC}$$

where EAC is the Equivalence Acceptance Criterion.

- Type 1 error is claiming equivalence when such is not the case (Patient Risk).
- Type 2 error is failing to claim equivalence when equivalence exists (Sponsor Risk).
- Decision Rule: Reject H_o and claim equivalence of means if a 100(1-2α)% two-sided confidence interval on the difference in means falls between –EAC and +EAC. This will provide a type 1 error rate of α.
- This is referred to as the two one-sided test (TOST) approach.

Difference test versus equivalence test

Most statistical investigations that support a sponsor's position to a regulatory agency require a test of equivalence rather than a difference test.

This is because the equivalence test considers scientific importance of any difference that might exist.

Additionally, the equivalence test more properly puts the burden of proof (alternative hypothesis) on the sponsor.

Difference hypotheses

H_0 : No difference in means

H_A : Difference between means

Equivalence hypotheses

H_0 : Difference in means is of practical importance

H_A : Difference in means is of no practical importance

Selection of EAC

1. Subject matter experts can define an acceptable difference based on scientific understandings.
2. Conversion of Capability to EAC
   - Assume K_2=1 and iteratively solve the following equation for EAC given values of Capability and σ_R

$$\text{Capability} = \Phi\left(3 - \frac{\text{EAC}}{\sigma_R}\right) - \Phi\left(-3 - \frac{\text{EAC}}{\sigma_R}\right)$$

   - JMP script “Calculate EAC given comparability”
3. σ_R is based on historical information of the pre-change process.

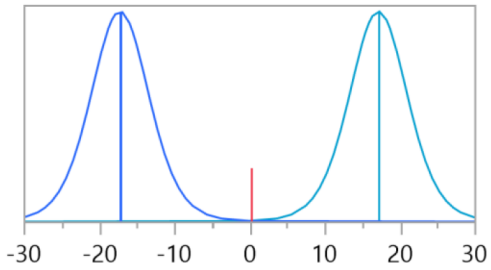
Statistical Tests Demonstatration

Equivalence Tests Pairwise Comparisons

Practical Equivalence between Reference and Post-change 1

Specified Practical Difference Threshold	17.2
Actual Difference in Means	0.233333
Std Error of Difference	3.925395

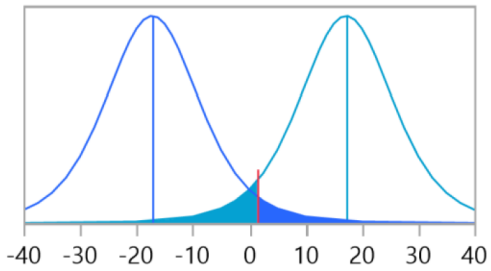
Null Hypothesis	DF	t Ratio	p-Value
Mean Difference ≥ 17.2	9.3099	-4.32228	0.0009*
Mean Difference ≤ -17.2	9.3099	4.441166	0.0007*
Max over both			0.0009*



Practical Equivalence between Reference and Post-change 2

Specified Practical Difference Threshold	17.2
Actual Difference in Means	1.4
Std Error of Difference	8.47361

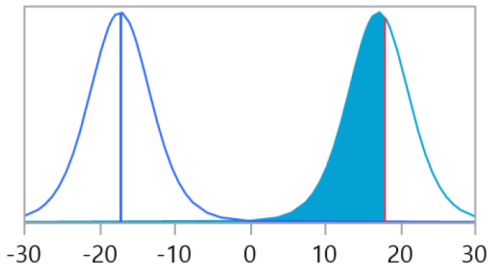
Null Hypothesis	DF	t Ratio	p-Value
Mean Difference ≥ 17.2	5.646	-1.86461	0.0573
Mean Difference ≤ -17.2	5.646	2.19505	0.0367*
Max over both			0.0573



Practical Equivalence between Reference and Post-change 3

Specified Practical Difference Threshold	17.2
Actual Difference in Means	17.96667
Std Error of Difference	4.137425

Null Hypothesis	DF	t Ratio	p-Value
Mean Difference ≥ 17.2	8.6918	0.1853	0.5714
Mean Difference ≤ -17.2	8.6918	8.499651	<.0001*
Max over both			0.5714



Non-inferiority of Standard Deviations

- Generally, we want the post-change process to have a standard deviation that is at least "not much worse" than the pre-change process standard deviation.
- To demonstrate this non-inferior condition, test the following set of hypotheses

$$H_0: \frac{\sigma_N}{\sigma_R} \geq \text{NIAC}$$

$$H_1: \frac{\sigma_N}{\sigma_R} < \text{NIAC}$$

where NIAC is the non-inferiority acceptance criterion.

- Decision Rule: Reject H_o and claim non-inferiority of standard deviations if a 100(1-α)% one-sided upper confidence bound on σ_N/σ_R is less than NIAC.
- This will provide a type 1 error rate of α.

Selection of NIAC

1. Subject matter experts can define an acceptable difference based on scientific understandings.
2. Conversion of Capability to EAC
   - Typically, assume K_1=0 and iteratively solve the following equation for NIAC for the given value of Capability.

$$\text{Capability} = \Phi\left(\frac{3}{\text{NAIC}}\right) - \Phi\left(\frac{-3}{\text{NAIC}}\right)$$

   - JMP script "Calculate NAIC given capability".

Statistical Tests Demonstration

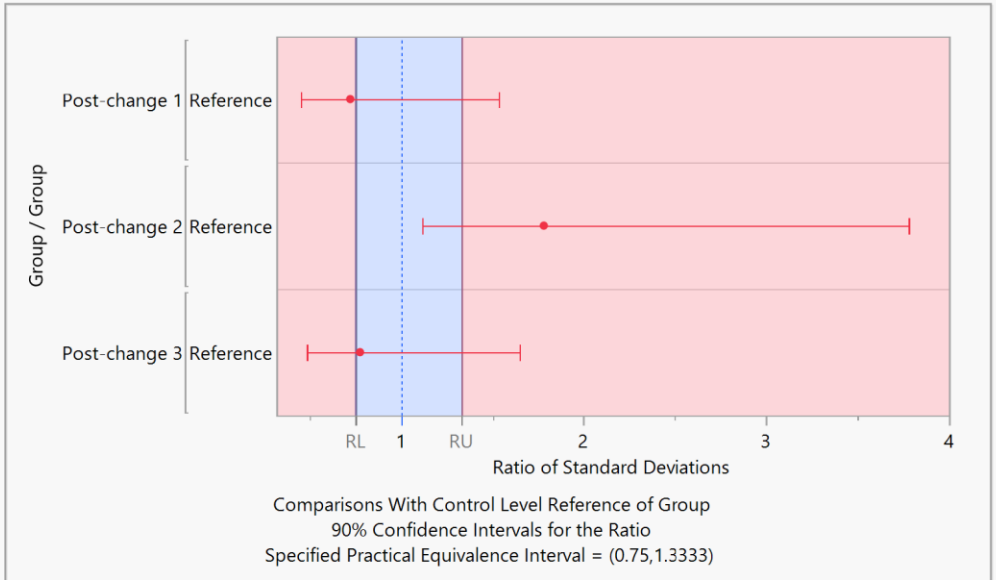
Equivalence Tests for the Ratio of Standard Deviations

Test	Alternative Hypothesis
Lower Bound	σT / σC > 0.75
Upper Bound	σT / σC < 1.33333333333333
Equivalence	0.75 < σT / σC < 1.33333333333333

TOST Tests

Group	Group	Ratio	Lower Bound F-Value	Upper Bound F-Value	Lower Bound p-Value	Upper Bound p-Value	Max p-Value	Two-Sided Lower 90%	Two-Sided Upper 90%	Assessment (α=0.05)
Post-change 1	Reference	0.723989	0.931839	0.294840	0.4749	0.0882	0.4749	0.453790	1.535825	Not Equivalent
Post-change 2	Reference	1.781702	5.643486	1.785634	0.0009*	0.8529	0.8529	1.116756	3.779592	Not Equivalent
Post-change 3	Reference	0.777498	1.074672	0.340033	0.3948	0.1156	0.3948	0.487329	1.649336	Not Equivalent

Forest Plot



Quality Ranges

- Decision Rule: Claim comparability if at least a large percentage (e.g., 90%) of sampled post-change items fall in a quality range (QR) based on the reference data where the QR is defined as

$$\bar{Y}_R \pm C \times S_R$$

$\bar{Y}_R =$ Sample mean of the reference sample

$S_R =$ Sample standard deviation of the reference sample

- Several rules have been suggested for determining C.

Methods of Selecting C in a QR

1. Tolerance Interval
   - C is determined from tolerance interval tables for given level of confidence and coverage.
   - Only sample size considered is the reference group.
2. 3-Sigma Interval
   - C=3 in all applications.
3. Risk Based Approach
   - Determination of C requires a boundary condition defined by Capability and a declared probability for such a condition passing the test.

Quality Ranges

Risk Based Approach

- The risk-based approach considers the risk of falsely declaring comparability as well as the sample sizes for both the reference and the post-change groups.
- This approach defines a boundary condition based on the defined similarity condition (Capability) that experts agree is non-comparable, and then assigns a low probability that such a condition will meet the comparability criterion.
- This process is analogous to the selection of EAC or NIAC.
- The value of C is determined as a function of the probability of the boundary condition passing, and the sample sizes of the pre- and post-change groups using computer simulation which we have provided in a JMP script.

Quality Ranges

Defining a Similarity Condition

Capabilitiy is defined by selecting values for K_1 and K_2:

$$K_1 = \frac{\left|\text{Difference in means}\right|}{\text{Standard deviation reference}}.$$ This is the effect size

$$K_2 = \frac{\text{Standard deviation new}}{\text{Standard deviation reference}}.$$ This is the ratio of standard deviations

$K_1 = 0$ and $K_2 = 1$ means R and N are identical

Tolerance and 3-sigma Intervals

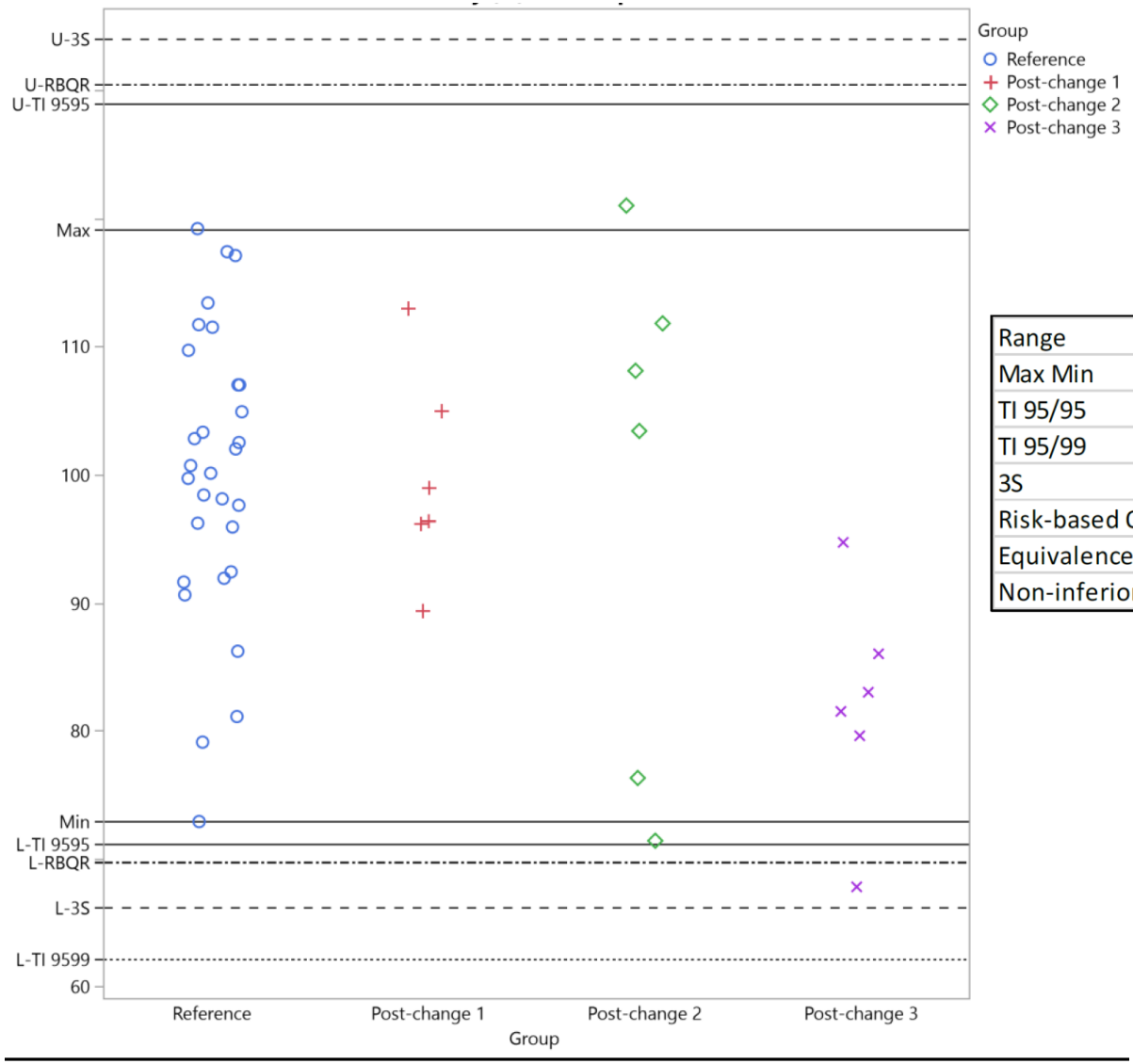
- Tolerance intervals are defined given a confidence and coverage level.
- There is no general recommendation for the confidence or coverage to select, but popular values are 95% confidence and either 95% or 99% coverage.
- Tolerance intervals will tighten as the reference sample size increases for a given confidence and coverage, but are not impacted by the size of the post-change process.
- FDA Draft Guidance (2019) "Based on our experience to date, methods such as tolerance intervals are not recommended for establishing the similarity acceptance criteria because a very large number of lots would be required to establish meaningful intervals." (Lines 812-814).
- 3-sigma intervals use C=3 regardless of the sample size of the reference group or the post-change process group.
- **Neither approach directly considers the probability of claiming comparability when the processes are different, or accounts for post-change sample sizes.**

A Note on Min-Max Ranges

- Min-max ranges should not be used for comparability because they are much too tight.

n_R	n_N	Probability of passing when R and N are identical
6	3	0.417
6	6	0.23
10	3	0.577
10	6	0.377
10	10	0.239
30	3	0.825

Quality Ranges Demonstration



Range	Post-change 1	Post-change 2	Post-change 3
Max Min	Pass	Fail	Fail
TI 95/95	Pass	Pass	Fail
TI 95/99	Pass	Pass	Pass
3S	Pass	Pass	Pass
Risk-based QR	Pass	Pass	Fail
Equivalence Test	Pass	Pass	Fail
Non-inferiority test	Pass	Fail	Pass

Method Comparison

Do These Methods Satisfy our Considerations?

Consideration	Equivalence test of means	Non-inferiority of SD	Risked base QR	Tolerance interval	3-Sigma
1. Protect patient by controlling type 1 error rate	Yes	Yes	Yes	No	No
2. Protect sponsor by controlling power	Yes	Yes	Yes	No	No
3. Incentivize sponsors to acquire process knowledge concerning post-change process	Yes as power increases	Yes as power increases	Yes as power increases	No power decreases as more added	No power decreases as more added
4. Enable decision making with practical sample sizes	Depends on level of Type 1 error and power				

Method Comparison

Do These Methods Satisfy our Considerations?

Consideration	Equivalence test of means	Non-inferiority of SD	Risked based QR	Tolerance interval	3-Sigma
5. Examine entirety of the process distribution of product.	Yes if combined with test of non-inferiority	Yes if combined with test of equivalence	Yes	Yes	Yes
6. Statistical rigor should consider criticality and measurement scale of the attribute.	All are based on quantitative measures.				
7. Demonstrate robustness to violations of assumptions.	All assume data are normally distributed. Equivalence test is most robust with respect to this assumption				
8. Be transparent, easy to explain, and easy to compute by scientists with no formal statistical training.	Calculations are reproducible and easy to compute				

Recommendations

1. If you are concerned with only the mean, use the test of statistical equivalence of means (e.g., qualification of a reference standard).
2. If you are concerned with both mean and SD, use either
   - Both tests of equivalence of means and non-inferiority of standard deviations, or a
   - Risk-based quality range.

Power and Sample Size

Statistical Power

- Power is the probability of claiming comparability for given values of K1, K2, type 1 error rate, and pre- and post-change samples sizes.
- Power=1-type 2 error rate.
- Practical limits on sample size will limit power.
- For a given sample size, power will increase as the type 1 error rate increases.

Determination of Type 1 and Type 2 Error Rates

1. Type 2 error rate when processes are identical should generally not be less than type 1 error rate unless type 1 error rate is sufficiently small (say 0.05).
2. Tests for non-inferiority and the risk-based quality range require a higher type 1 error rate than an equivalence test, or the type 2 error rate will be unacceptably large.
   - For a given sample size there is greater uncertainty in the estimate of a standard deviation than a mean.
3. Regulatory agencies and sponsors need to agree on what constitutes a “reasonable” and “practical” sample size from which power calculations can be performed.

Power and Sample Size

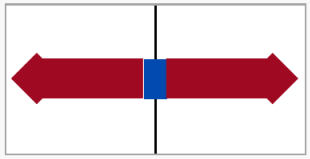
Power Explorer for One Sample Equivalence

Explorer Settings

Test Type

- (•) Equivalence
- ( ) Superiority
- ( ) Non-inferiority

H0
H1



Upper Margin 1.00
Lower Margin -1.00
[ ] Use symmetric bounds

Preliminary Information

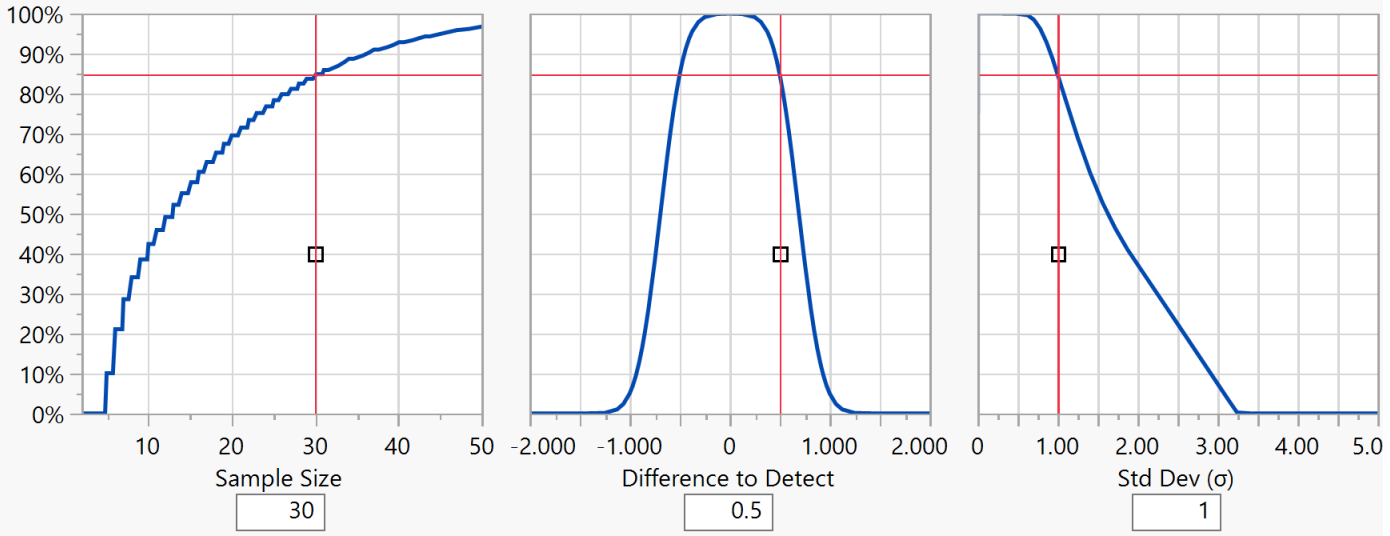
Alpha 0.05

Is the population standard deviation assumed to be known? ( ) Yes (•) No

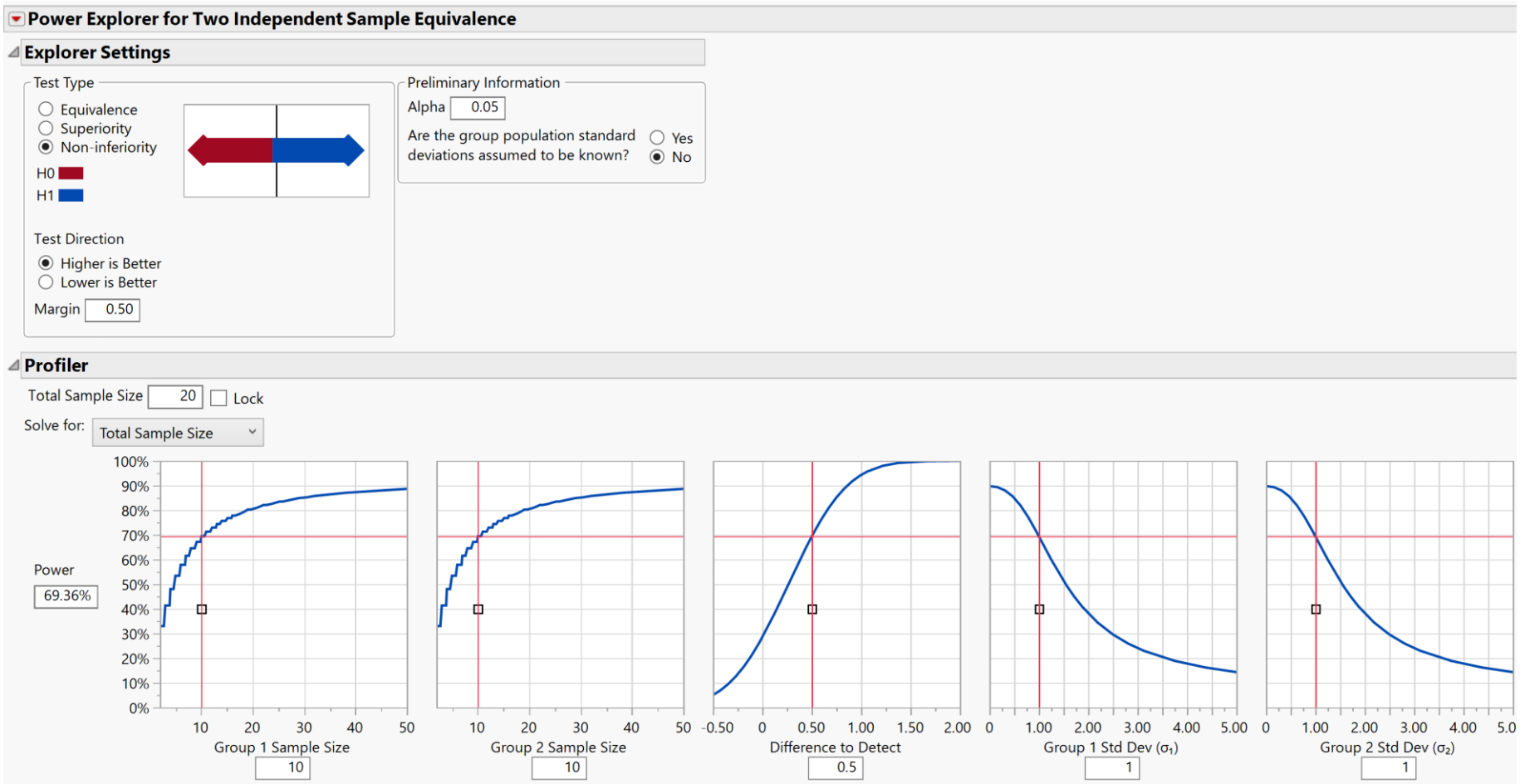
Profiler

Solve for: Sample Size

Power
84.83%

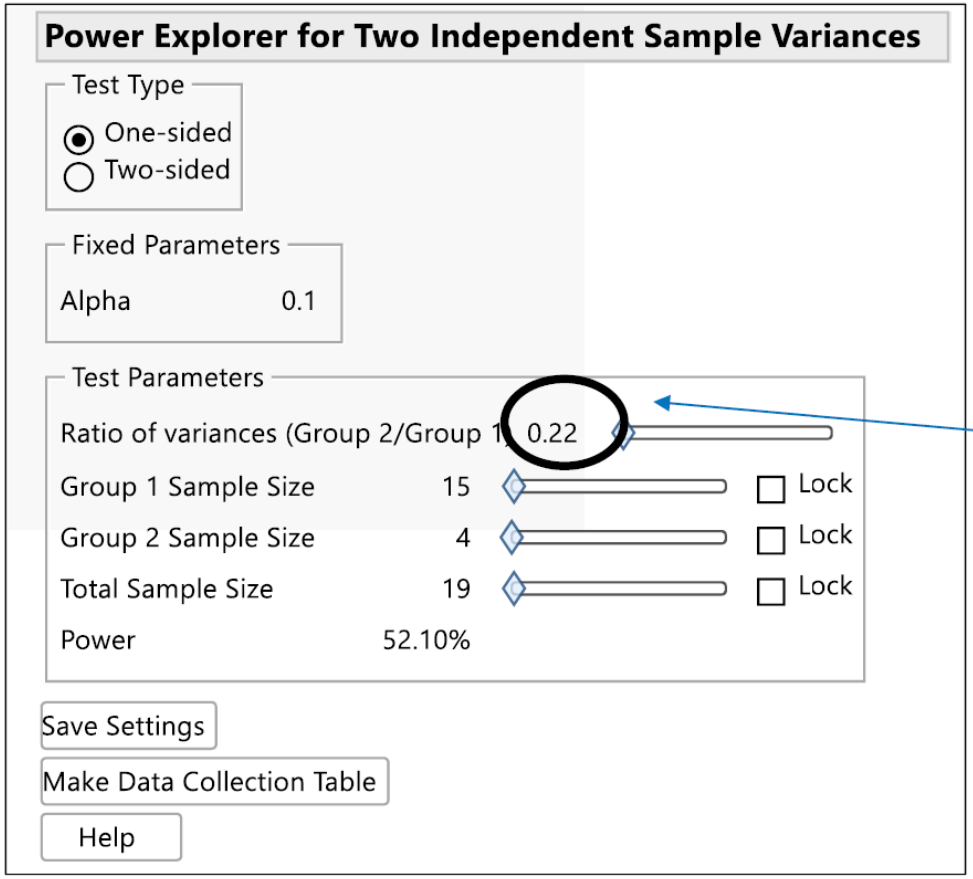


Power and Sample Size



Power and Sample Size

JMP Power Explorer Two Sample Variances



Calculator with 15 reference, 4 post-change, alpha=0.1, and NAIC=2.13.

This term is

$$\frac{\sigma_N^2}{\sigma_R^2 \times (NAIC)^2}$$

If variances are equal,

$$\frac{1}{(NAIC)^2} = \frac{1}{(2.13)^2} = 0.220$$

Final Recommendations

Determination of Acceptance Criteria

1. Criteria required
   - EAC in equivalence testing of means
   - NIAC in non-inferiority testing of SD
   - Boundary condition in Quality Range
2. In lieu of scientific information regarding the particular attribute, the Capability with 3-sigma intervals of the reference distribution is a useful metric.
3. Capability between 84% and 95% seem reasonable for most applications depending on the criticality of the attribute.

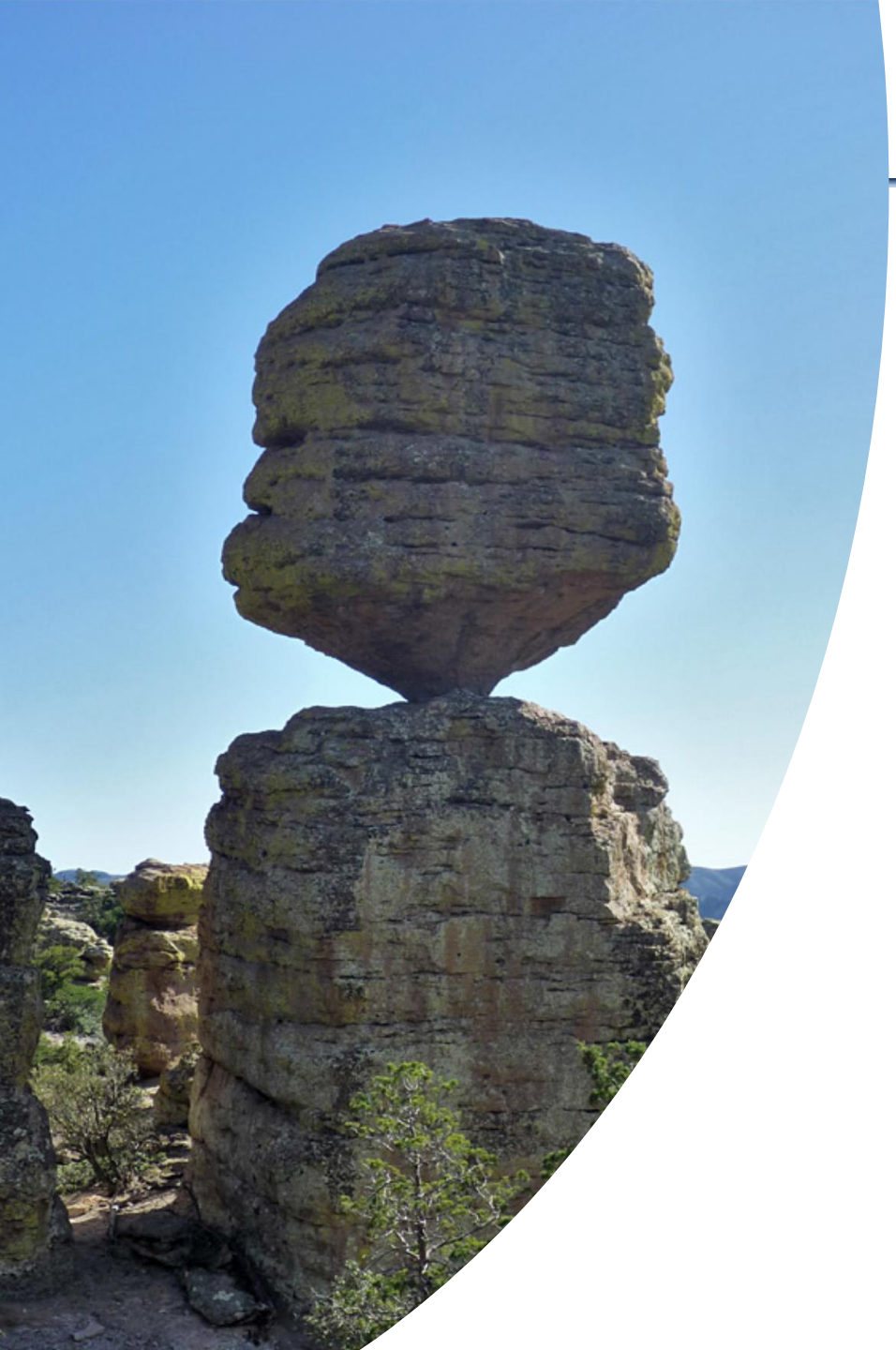
Final Recommendations

"Reasonable" Capability Values

Equivalence				Non-inferiority		
K1	K2	Capability		K1	K2	Capability
1.36	1	95.0%		0	1.53	95.0%
1.5	1	93.3%		0	1.82	90.1%
1.72	1	90.0%		0	2.13	84.1%
2	1	84.1%				

Take Home Messages

1. Sponsors should report patient risks and border conditions, and ensure sponsors take greater risk than patient.
2. Regulatory agencies should consider these values in light of reasonable sample sizes for the particular application.
3. Capability is a practical metric for defining comparability.
4. Do not use specifications, tolerance intervals, 3-sigma, or min-max ranges as quality ranges.
5. Quality ranges are better than equivalence tests of means because they are sensitive to differences in standard deviations.



Questions?