Little Stats Seminar

Equivalence Tests and Non-inferiority Tests for Sensory Applications

J. Castura

Compusense Inc.

Motivating Example

Should a proposed ingredient substitution advance to the next stage?

this is an equivalence question

Motivating Example

Should a proposed ingredient substitution advance to the next stage?

this is an equivalence question

Before answering this question, let’s take a moment to consider the hypothesis testing framework.

Hypothesis Tests

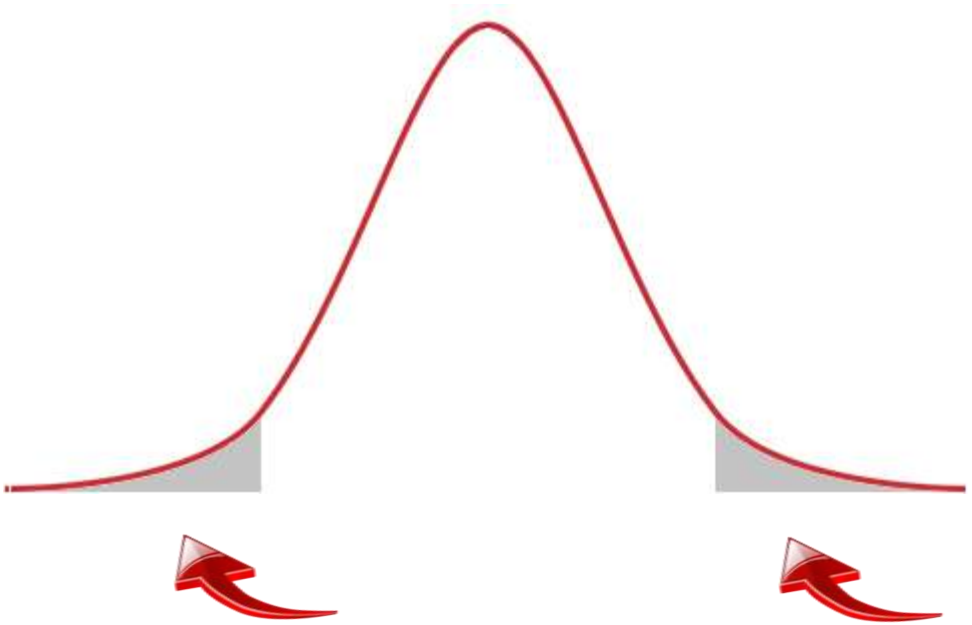
We assume a distribution under the null hypothesis (**H_0**).
The probability of observing a result in the tail regions is low.

We calculate a test statistic from the observed data.
An extreme test statistic gives evidence to reject **H_0**.
We reject at the tails of the distribution.

Now think about what this test statistic looks like in a typical statistical test (for difference).

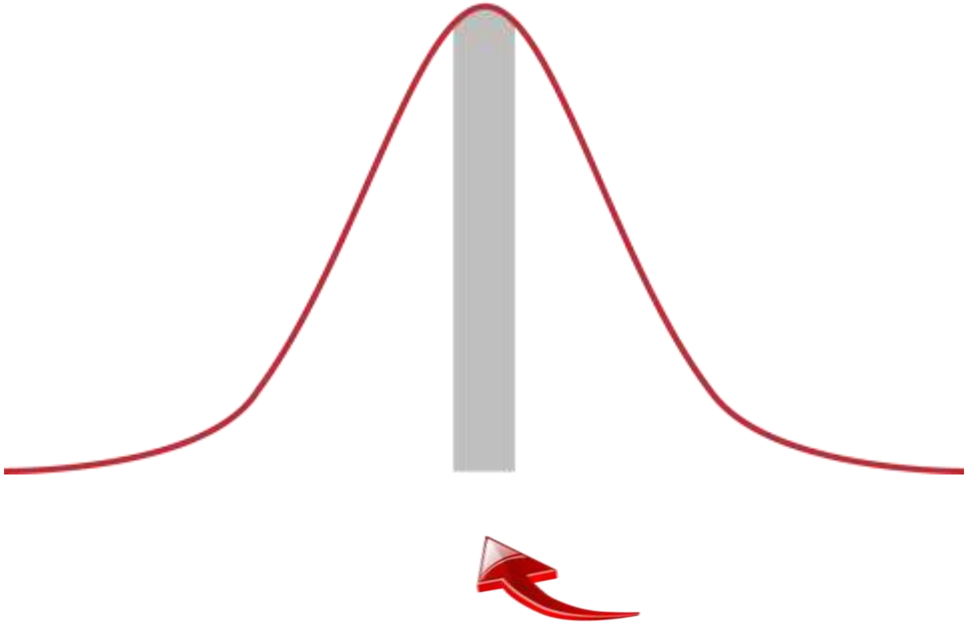
A Test for Difference

Typically we reject $\mathbf{H_0}$ in favor of the alternative ($\mathbf{H_1}$) at the tails of the distribution.
Observing a test statistic that falls in the tails are improbable under $\mathbf{H_0}$.



A Test for Similarity

If the test statistic comes from a difference testing paradigm, it falls in the center of the distribution. How do we reject $\mathbf{H_0}$ in favor of $\mathbf{H_1}$?



Merchandise: Genuine or Knock-off?

Suppose I only want to buy genuine merchandise...

		Truth	
		Knock-off	Genuine
Decision	Don't buy	Correct	**Missed a good deal**
	Buy	**Bought junk**	Correct

Type I and Type II errors: α-risk and β-risk

		Truth	
		Different	Not
Decision	Reject H_0	Correct **$1-\beta$**	**Type I Error** **α**
	Retain H_0	**Type II Error** **β**	Correct **$1-\alpha$**

Power Approach

Power calculations are made to determine an appropriate sample size.

A low value is selected for **β**.

We want to avoid concluding that products are not significantly different, when the products are, in fact, different.

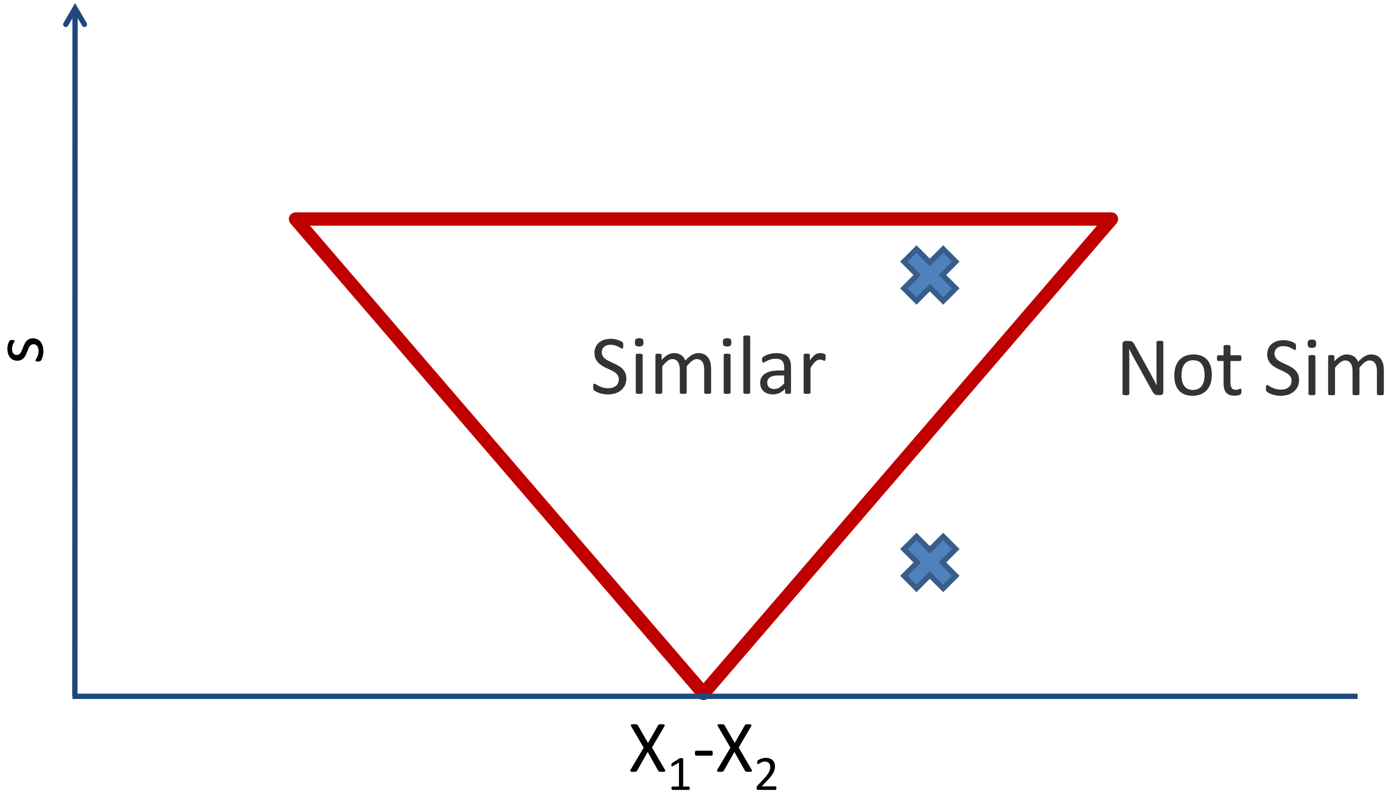
How does the Power Approach work?

If the null hypothesis ($\mathbf{H_0}$) is retained, it is reasoned that power was high enough that it would have been rejected in favor of the alternative ($\mathbf{H_1}$) if the products were dissimilar. Thus the products must be the same.

But really no **p**-value “proves” or leads us to “accept” $\mathbf{H_0}$.

Hypothesis test logic being is contorted to meet the objectives of trying to determine similarity.

Rejection Region for Power Approach



How does the Power Approach work?

p-values can be affected by…

Effect size: magnitude of difference between products.

Sample size: underpowering the test misses meaningful differences, and overpowering the test enables detection of trivial differences.

Relying only on **p**-values for decision-making is not a good practice.

Rejection Region Illustrates Problems

It could be that the confidence interval for the difference falls completely within the equivalence bounds, yet there is no conclusion of similarity.

More precise measurement leads to decreased power for detecting similarity.

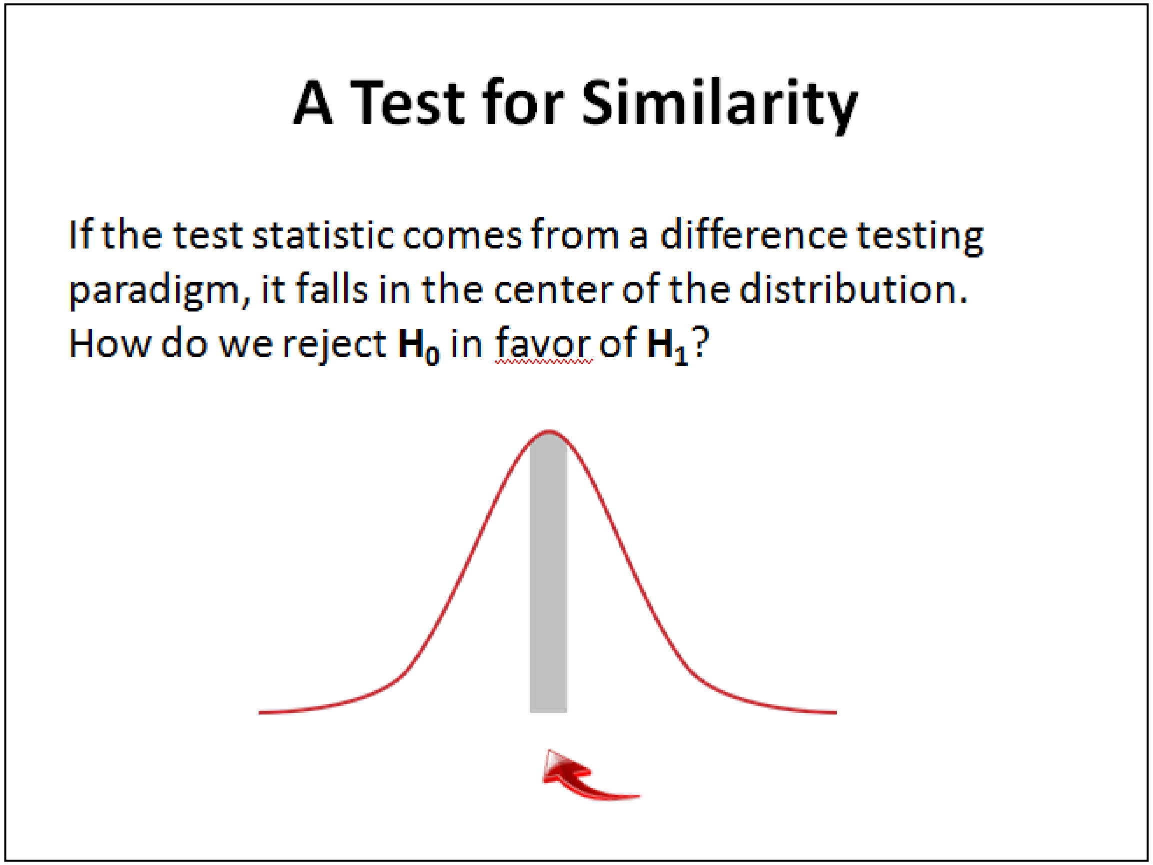
Hypothesis Testing for Equivalence

A proper hypothesis test for equivalence:

H_0: Products not equivalent
H_1: Products equivalent

Let’s try a different approach...

If we are looking for a hypothesis test, it is one that rejects the $\mathbf{H_0}$ hypothesis of non-equivalence in the center of the distribution.



A Test for Similarity

If the test statistic comes from a difference testing paradigm, it falls in the center of the distribution. How do we reject $\mathbf{H_0}$ in favor of $\mathbf{H_1}$?

Two One-Sided Tests (TOST) procedure

Two hypotheses are tested:

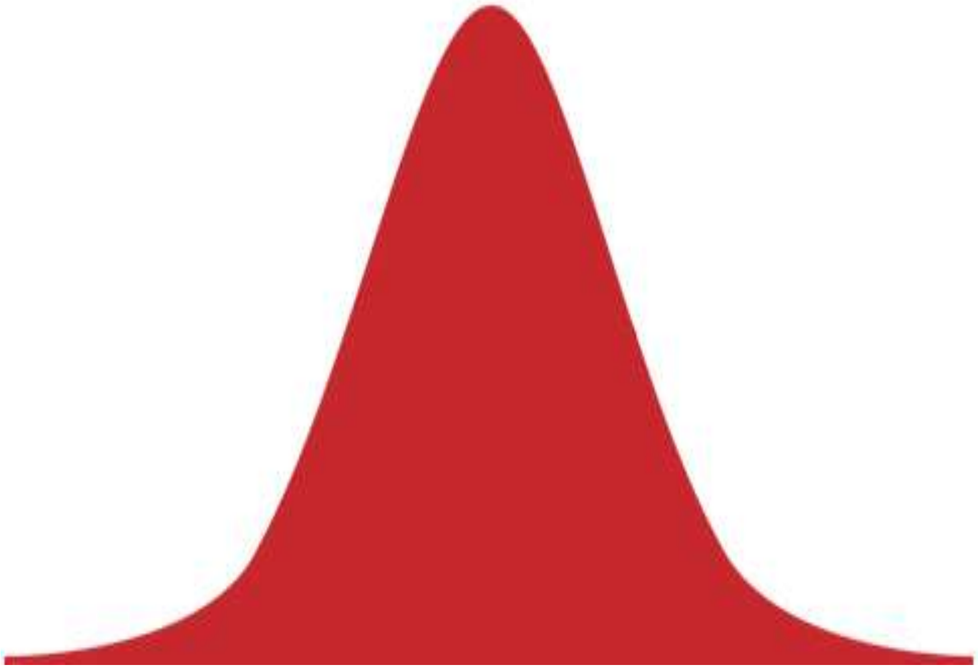
1) $\mathbf{H}_{01}$: $\theta < \theta_0 - \delta_1$ *vs.* $\mathbf{H}_{11}$: $\theta \geq \theta_0 - \delta_1$

2) $\mathbf{H}_{02}$: $\theta > \theta_0 + \delta_2$ *vs.* $\mathbf{H}_{12}$: $\theta \leq \theta_0 + \delta_2$

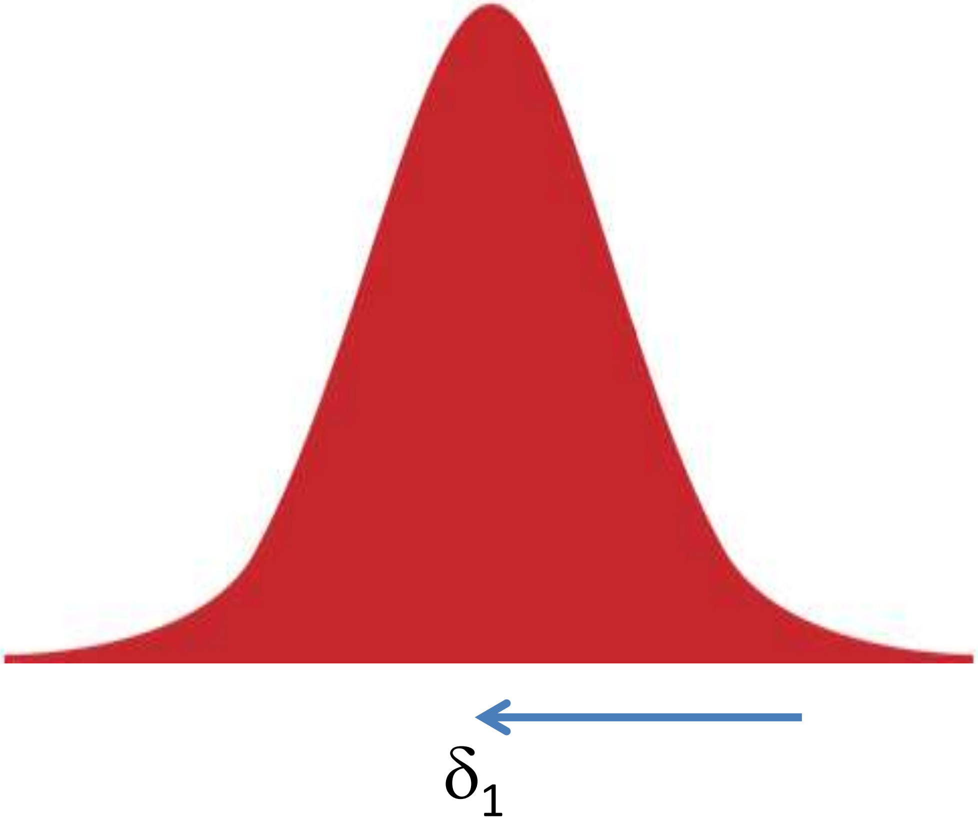
If both **p**-values are significant (at level α) then we can reject the complete $\mathbf{H}_0$ in favor of the $\mathbf{H}_1$ and declare **Equivalence**.

The procedure gives a valid test of the complete alternative hypothesis $\mathbf{H}_1$: $\theta_0 - \delta_1 \leq \theta \leq \theta_0 + \delta_2$.

TOST in Action



TOST in Action

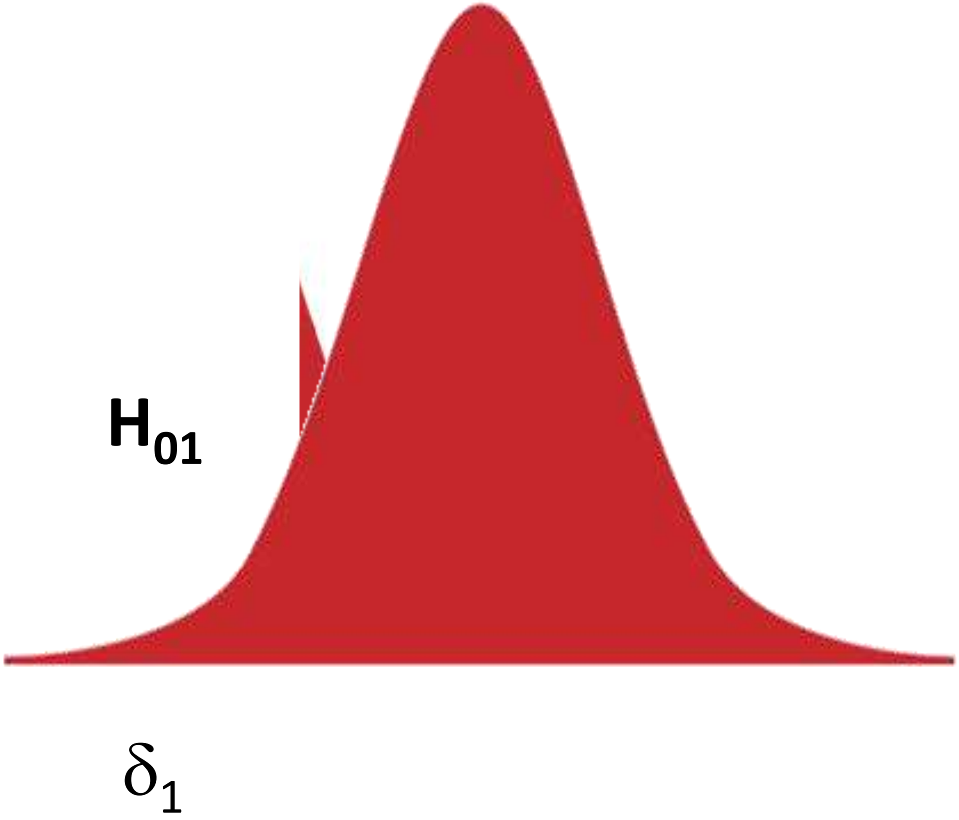


TOST in Action

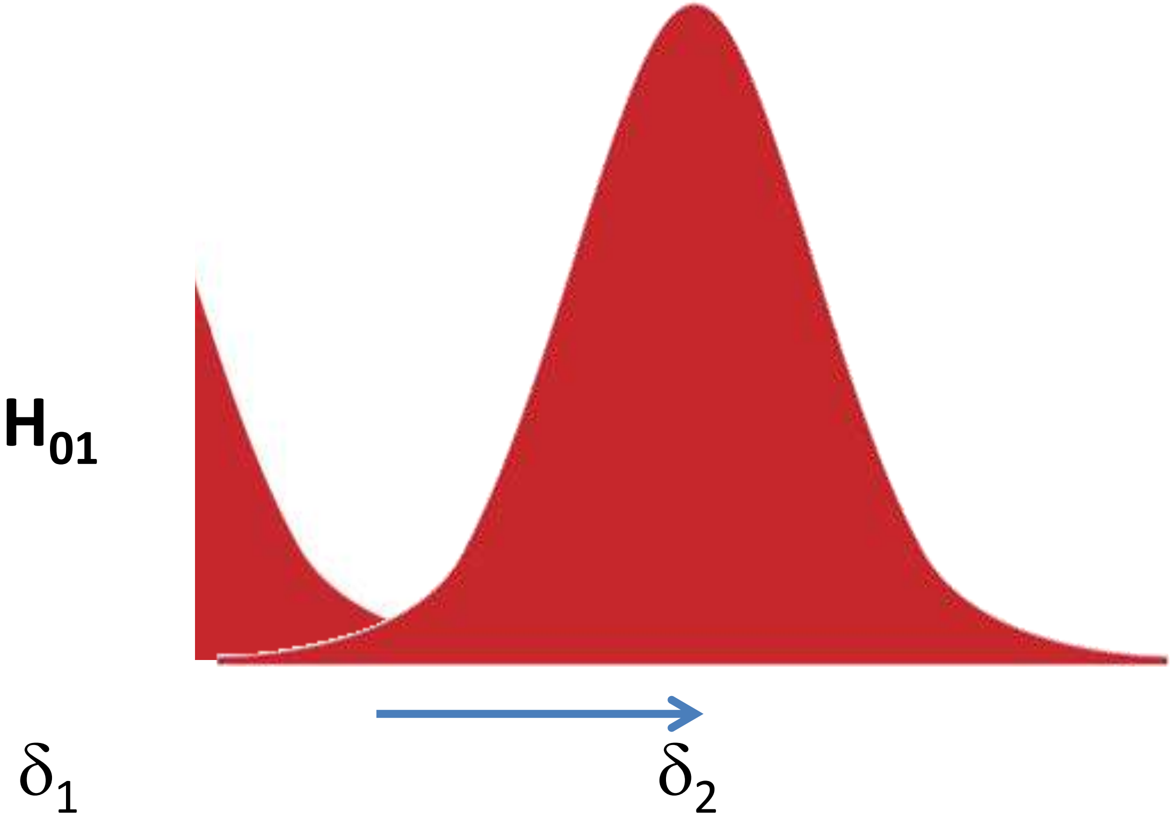
H_{01}

δ_1

TOST in Action

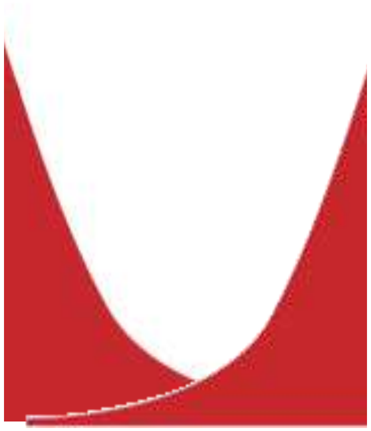


TOST in Action



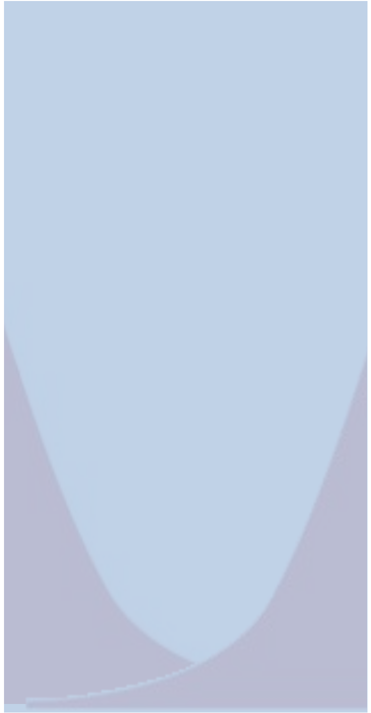
TOST in Action

H_{01} **H_{02}**



δ_1 δ_2

TOST in Action



H_{01} H_{02}

δ_1 δ_2

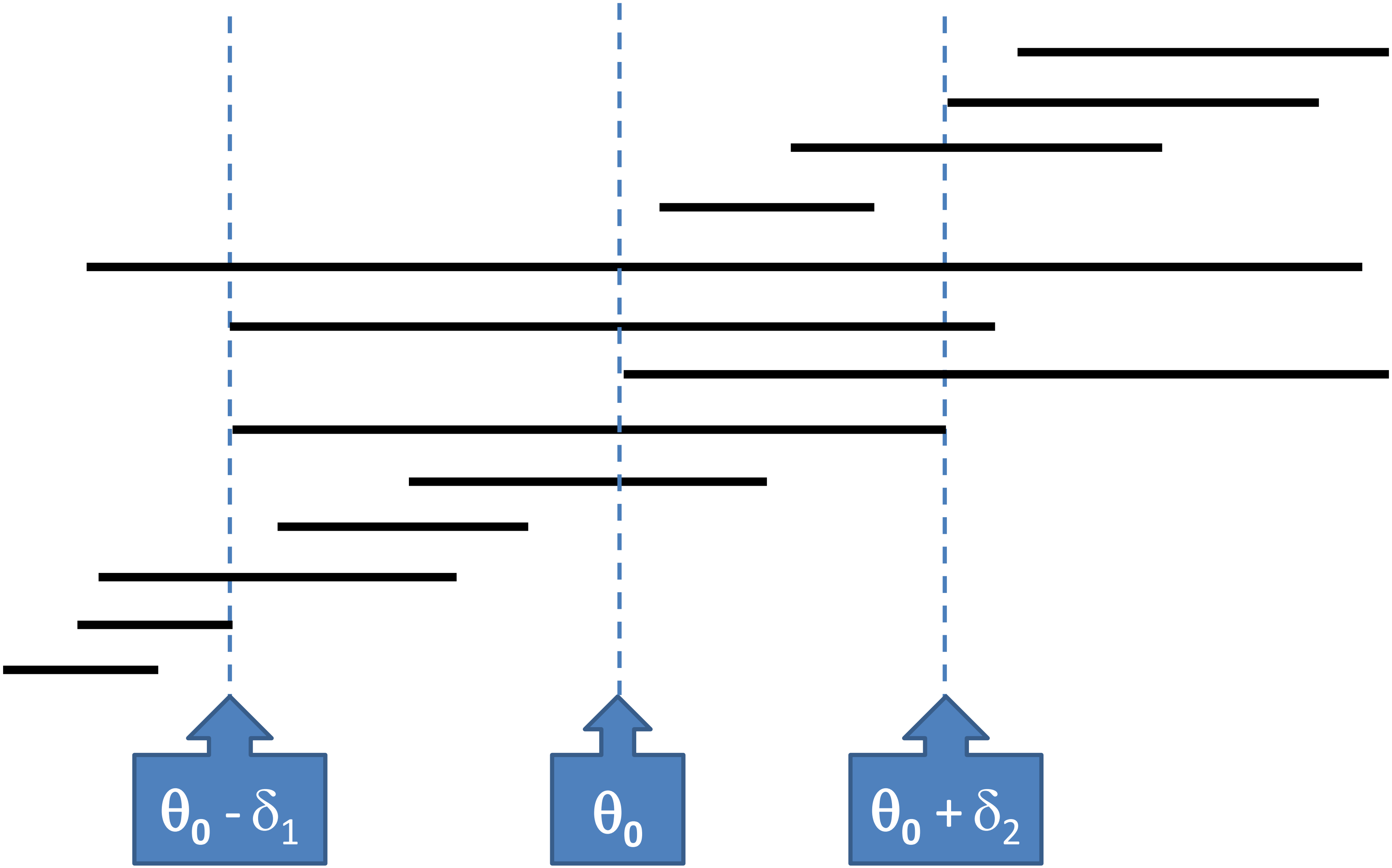
TOST & Confidence Interval Inclusion

When the same α is used in both tests. It possible to construct (1-2α)100% confidence intervals. Products are **Equivalent** if the confidence interval is contained within the equivalence bounds.

The confidence intervals for a hypothesis test for difference are (1-α)100% confidence intervals (which are not as wide).

Note: products might be different, yet equivalent!

Confidence Interval Inclusion



When can the TOST be applied?

- The TOST procedure is flexible
  - Parametric or non-parametric tests
  - Discrete or continuous data
  - We want to ensure that some parameter falls between a lower equivalence bound and an upper equivalence bound

Examples for the TOST procedure

2-AFC consumer test on saltiness perception.

Suppose we set the equivalence margin to δ=0.08.

Equivalence: [0.42, 0.58]

Non-equivalence:

[0.00, 0.42), (0.58, 1.00]

Not salty enough *Too salty*

Examples for the TOST procedure

Suppose bitterness is characteristic of the product but undesirable at high intensity.

We can tolerate more of a decrease in bitterness, and less of an increase (e.g. δ_1=6, δ_2=3).

A trained descriptive sensory panel evaluates *bitterness* intensity.

H_0: $\mu_d < -6$ or $\mu_d > 3$

H_1: $-6 \leq \mu_d \leq 3$

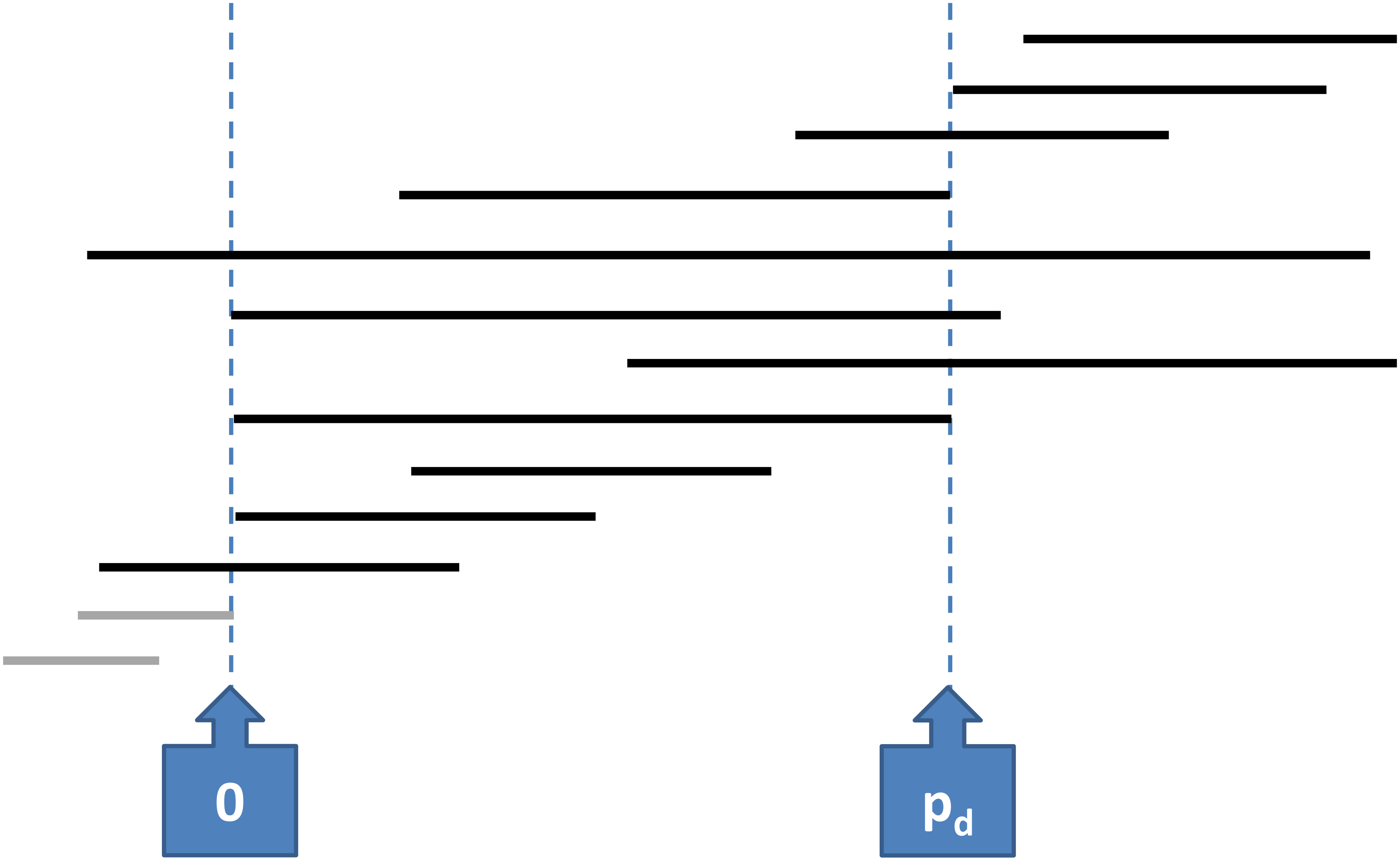
Examples for the TOST procedure

Does the TOST procedure make sense for analyzing sensory difference tests, such as triangle, duo-trio, etc.?

These are 1-sided tests – we don’t usually test whether incorrect responses are being made systematically!

If $\mathbf{p}_c$ falls below guessing probability the $\mathbf{p}_d$=0 is used (Bi, 2005; Christensen & Brockhoff, 2011).

Confidence Interval Inclusion



Similarity Testing in E18 Standards

E 1885 *Standard Test Method for Sensory Analysis - Triangle Test*

E 1958 *Standard Guide for Sensory Claim Substantiation*

E 2139 *Standard Test Method for Same-Different Test*

E 2164 *Standard Test Method for Directional Difference Test*

E 2610 *Standard Test Method for Sensory Analysis - Duo-Trio Test*

From E 1885-04

“8.1 Choose the number of assessors to yield the level of sensitivity called for by the test objectives. The sensitivity of the test is a function of three values: the **α**-risk, and the **β**-risk, and the maximum allowable proportion of distinguishers, $\mathbf{p_d}$.”

“...$\mathbf{p_d}$ is the proportion of the entire population of assessors who can distinguish between the two products. It is a strictly statistical “guessing model” of the assessor’s behavior.”

Number of Assessor (E1885)

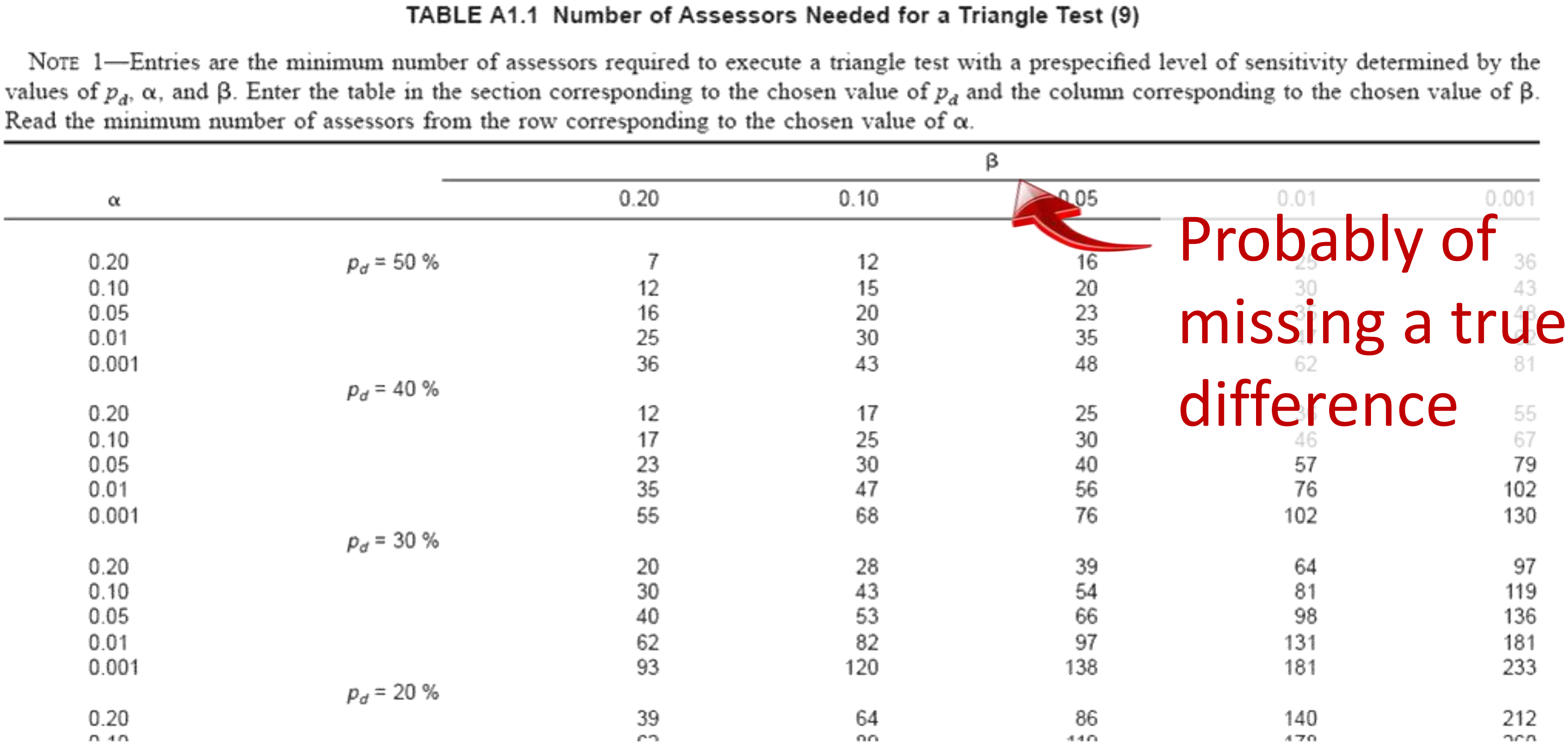


TABLE A1.1 Number of Assessors Needed for a Triangle Test (9)

NOTE 1—Entries are the minimum number of assessors required to execute a triangle test with a prespecified level of sensitivity determined by the values of p_d, α, and β. Enter the table in the section corresponding to the chosen value of p_d and the column corresponding to the chosen value of β. Read the minimum number of assessors from the row corresponding to the chosen value of α.

		β				
α		0.20	0.10	0.05	0.01	0.001
0.20	p_d = 50 %	7	12	16	25	36
0.10		12	15	20	30	43
0.05		16	20	23	35	48
0.01		25	30	35	47	62
0.001		36	43	48	62	81
	p_d = 40 %					
0.20		12	17	25	35	55
0.10		17	25	30	46	67
0.05		23	30	40	57	79
0.01		35	47	56	76	102
0.001		55	68	76	102	130
	p_d = 30 %					
0.20		20	28	39	64	97
0.10		30	43	54	81	119
0.05		40	53	66	98	136
0.01		62	82	97	131	181
0.001		93	120	138	181	233
	p_d = 20 %					
0.20		39	64	86	140	212

Probably of missing a true difference

Proportion of distinguishers

Probably of falsely declaring a difference

Example 1*

Select $\boldsymbol{\alpha}$=0.1 and $\boldsymbol{\beta}$=0.05

Assumed proportion of detectors: $\mathbf{p_d}$=0.3

Assumed proportion of correct responses:

$$\mathbf{p_c} = \mathbf{p_d} + (1/3)(1-\mathbf{p_d}) = 0.533$$

Use E 1885 to determine number of assessors.

* Example 1 from Bi (2005)

Example 1

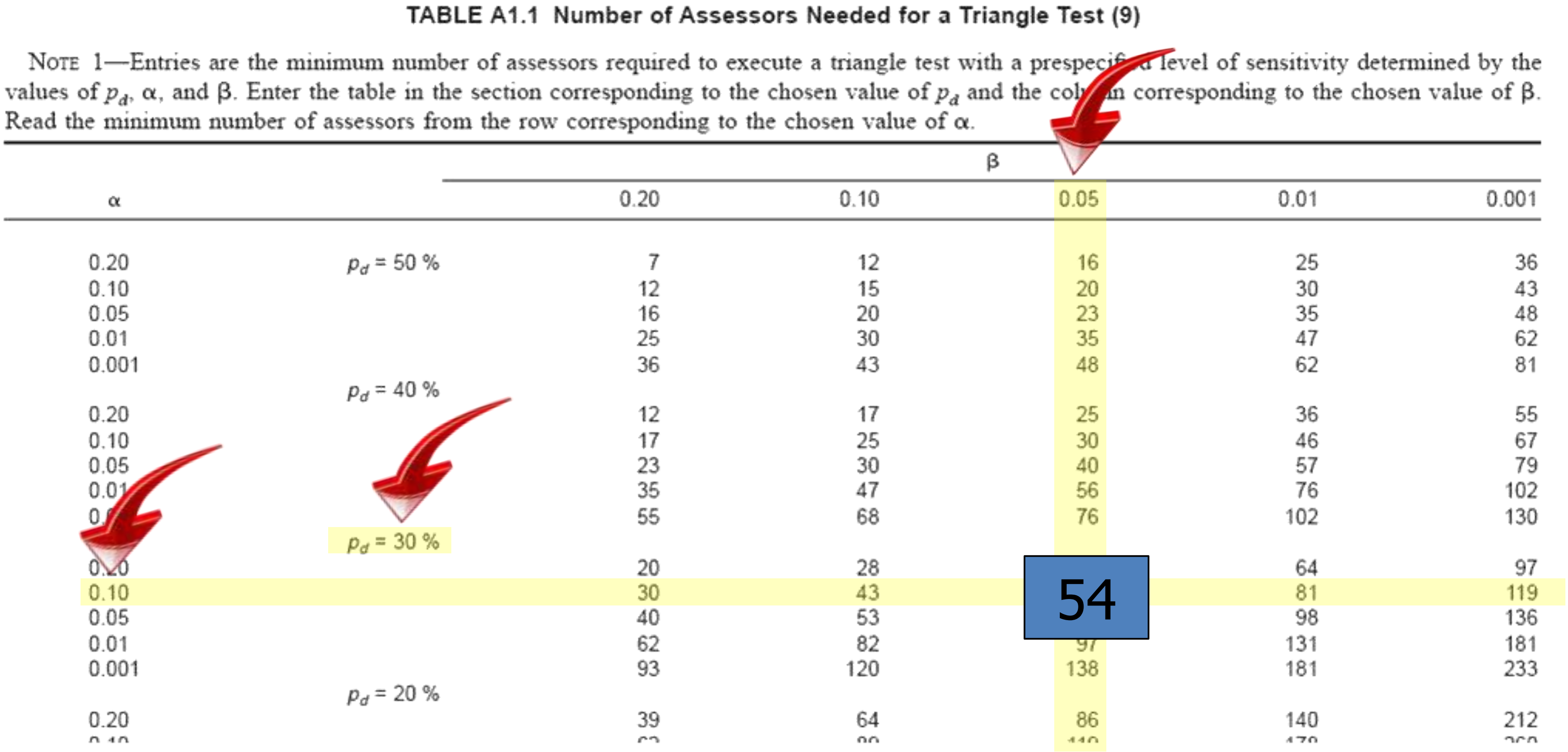


TABLE A1.1 Number of Assessors Needed for a Triangle Test (9)

NOTE 1—Entries are the minimum number of assessors required to execute a triangle test with a prespecified level of sensitivity determined by the values of p_d, α, and β. Enter the table in the section corresponding to the chosen value of p_d and the column corresponding to the chosen value of β. Read the minimum number of assessors from the row corresponding to the chosen value of α.

		β				
α		0.20	0.10	0.05	0.01	0.001
0.20	p_d = 50 %	7	12	16	25	36
0.10		12	15	20	30	43
0.05		16	20	23	35	48
0.01		25	30	35	47	62
0.001		36	43	48	62	81
	p_d = 40 %					
0.20		12	17	25	36	55
0.10		17	25	30	46	67
0.05		23	30	40	57	79
0.01		35	47	56	76	102
0.001		55	68	76	102	130
	p_d = 30 %					
0.20		20	28		64	97
0.10		30	43	54	81	119
0.05		40	53		98	136
0.01		62	82	97	131	181
0.001		93	120	138	181	233
	p_d = 20 %					
0.20		39	64	86	140	212

Example 1

Assume the following is true:

Products are more similar than we expected.

True proportion of distinguishers is $\mathbf{p}_{\mathbf{d0}}$=0.1

True proportion correct responses:

$$\mathbf{p}_{\mathbf{c}} = \mathbf{p}_{\mathbf{d0}} + (1/3)(1-\mathbf{p}_{\mathbf{d0}}) = 0.1+0.3 = 0.4$$

We expect to confirm similarity with high probability.

Simulation studies allow us to investigate what happens in this scenario.

Table of Critical Values

Table A1.2 shows how many “correct” responses indicate a statistically significant result.

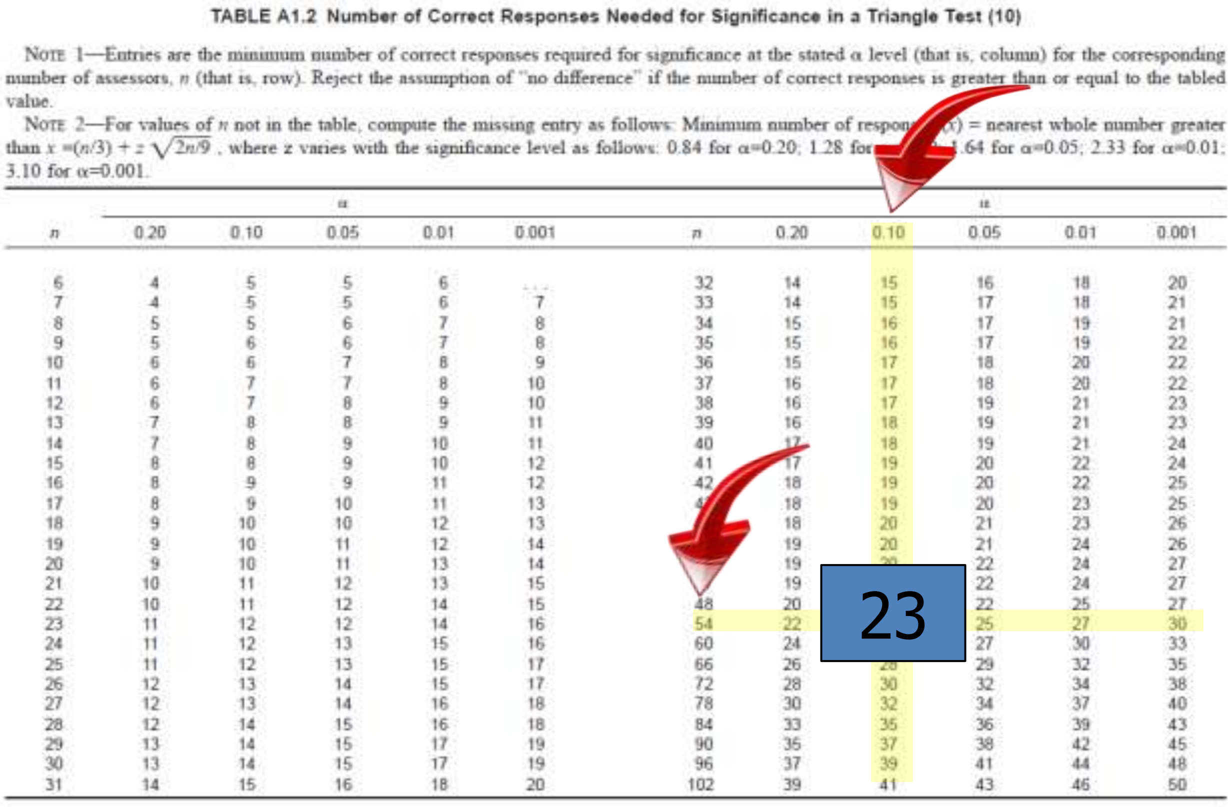


TABLE A1.2 Number of Correct Responses Needed for Significance in a Triangle Test (10)

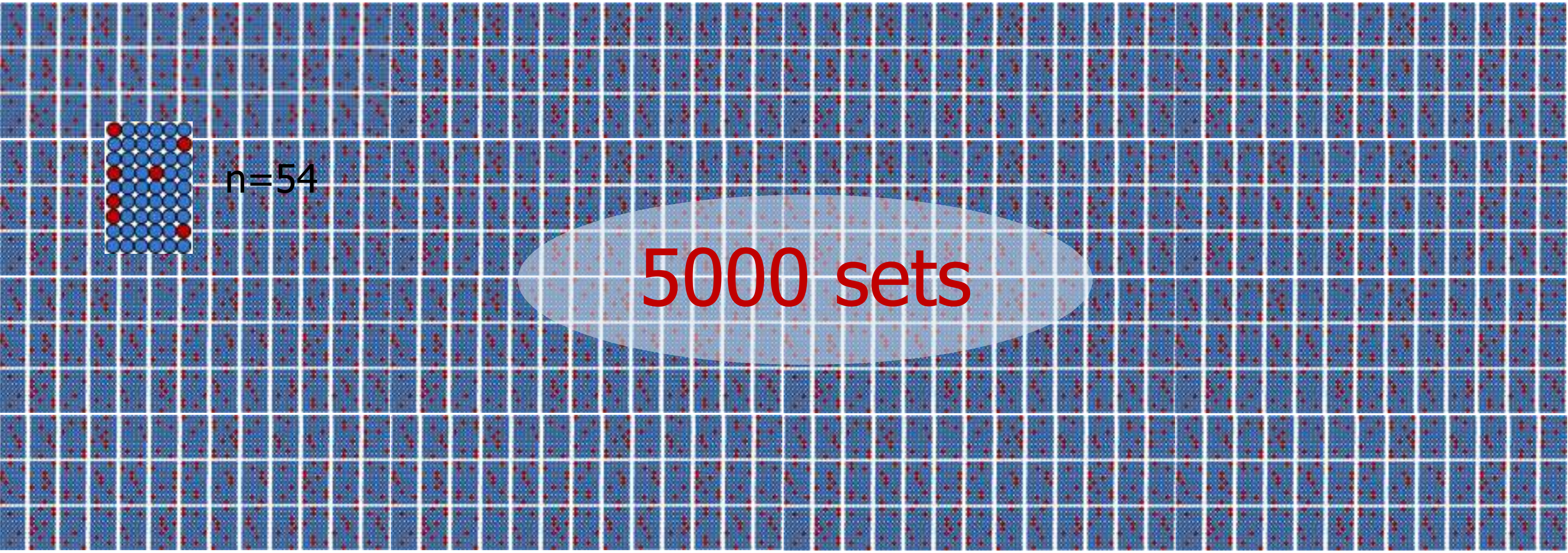
NOTE 1—Entries are the minimum number of correct responses required for significance at the stated α level (that is, column) for the corresponding number of assessors, *n* (that is, row). Reject the assumption of "no difference" if the number of correct responses is greater than or equal to the tabled value.

NOTE 2—For values of *n* not in the table, compute the missing entry as follows: Minimum number of respon[illegible](*x*) = nearest whole number greater than $x = (n/3) + z\sqrt{2n/9}$, where z varies with the significance level as follows: 0.84 for α=0.20; 1.28 for [illegible] 1.64 for α=0.05; 2.33 for α=0.01; 3.10 for α=0.001.

	α						α				
n	0.20	0.10	0.05	0.01	0.001	*n*	0.20	0.10	0.05	0.01	0.001
6	4	5	5	6	. . .	32	14	15	16	18	20
7	4	5	5	6	7	33	14	15	17	18	21
8	5	5	6	7	8	34	15	16	17	19	21
9	5	6	6	7	8	35	15	16	17	19	22
10	6	6	7	8	9	36	15	17	18	20	22
11	6	7	7	8	10	37	16	17	18	20	22
12	6	7	8	9	10	38	16	17	19	21	23
13	7	8	8	9	11	39	16	18	19	21	23
14	7	8	9	10	11	40	17	18	19	21	24
15	8	8	9	10	12	41	17	19	20	22	24
16	8	9	9	11	12	42	18	19	20	22	25
17	8	9	10	11	13	[illegible]	18	19	20	23	25
18	9	10	10	12	13	[illegible]	18	20	21	23	26
19	9	10	11	12	14	[illegible]	19	20	21	24	26
20	9	10	11	13	14	[illegible]	19	[illegible]	22	24	27
21	10	11	12	13	15	[illegible]	19	[illegible]	22	24	27
22	10	11	12	14	15	48	20	[illegible]	22	25	27
23	11	12	12	14	16	54	22	[illegible]	25	27	30
24	11	12	13	15	16	60	24	[illegible]	27	30	33
25	11	12	13	15	17	66	26	[illegible]	29	32	35
26	12	13	14	15	17	72	28	30	32	34	38
27	12	13	14	16	18	78	30	32	34	37	40
28	12	14	15	16	18	84	33	35	36	39	43
29	13	14	15	17	19	90	35	37	38	42	45
30	13	14	15	17	19	96	37	39	41	44	48
31	14	15	16	18	20	102	39	41	43	46	50

23

Example 1



H_0 is retained in some sets and rejected in others.
Similarity is confirmed with probability **0.49**.

Example 2*

Table A1.1 in E1885-04 recommends a *minimum* of 457 assessors at $\boldsymbol{\alpha}$=0.1, $\boldsymbol{\beta}$=0.05, $\mathbf{p_d}$=0.1.

Bi lets **n**=540 and re-runs the simulation to obtain 5000 sets.

$\mathbf{H_0}$ is retained in some sets and rejected in some others. The power approach confirms similarity with probability 0.02.

* Example 2 from Bi (2005)

Similarity based on Triangle

Following E 1885, set $\boldsymbol{\alpha}$=$\boldsymbol{\beta}$=0.05 and $\mathbf{p_d}$=0.3. Use **n**=66.

Let **n**={66, 660} and
$\mathbf{p_d}$ = {0.4, 0.35, 0.30, 0.25, 0.20, 0.15, 0.1, 0.05, 0}.

2500 simulated datasets for each of the 18 scenarios.

Similarity based on Triangle

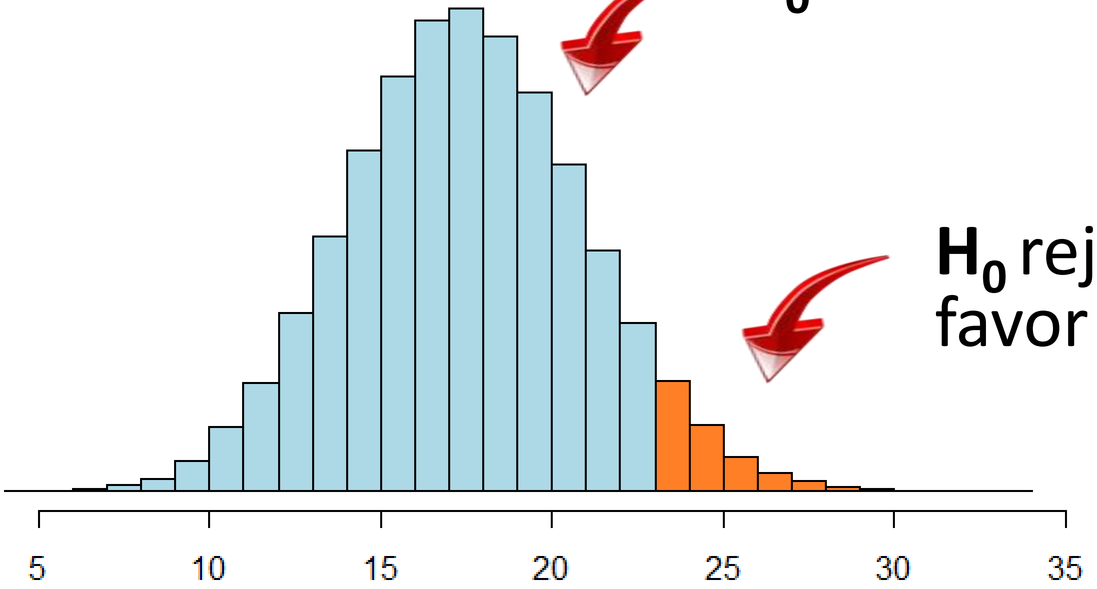
Proportions in which similarity is confirmed **n**={66, 660}

True p_d	n=66	n=660
0.40	0.0020	0.0000
0.35	0.0136	0.0000
0.30	0.0516	0.0000
0.25	0.1320	0.0000
0.20	0.2800	0.0000
0.15	0.5036	0.0004
0.10	0.7108	0.0248
0.05	0.8564	0.4124
0.00	0.9512	0.9480

So what is the ASTM approach?

The person running the test sets α, β, and $\mathbf{p_d}$ to get the number of assessors (**n**).

The power of the test is such that if $\mathbf{H_0}$ is retained then the products are deemed **Similar**.



$\mathbf{H_0}$ retained

$\mathbf{H_0}$ rejected in favor of $\mathbf{H_1}$

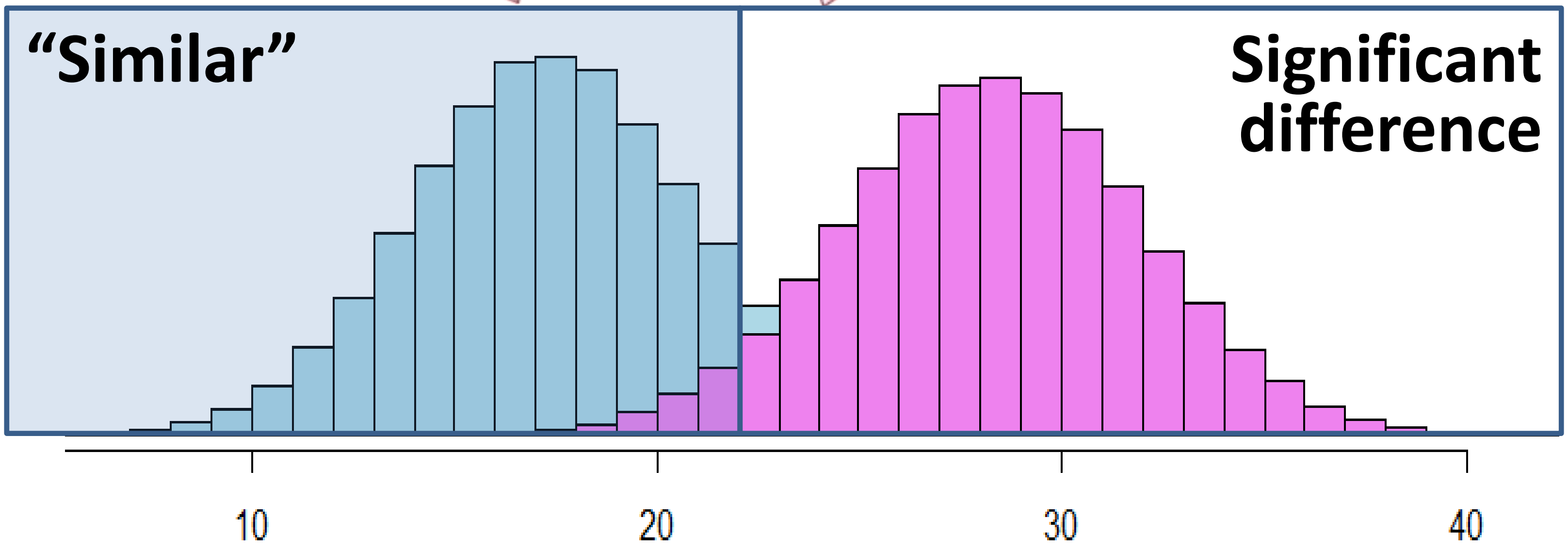
Distributions under H_0 and H_1

The distribution under $\mathbf{H_0}$ is fixed.

Suppose that there really are distinguishers in the population. The distribution under $\mathbf{H_1}$ is right-shifted.

$\mathbf{H_0}$ retained

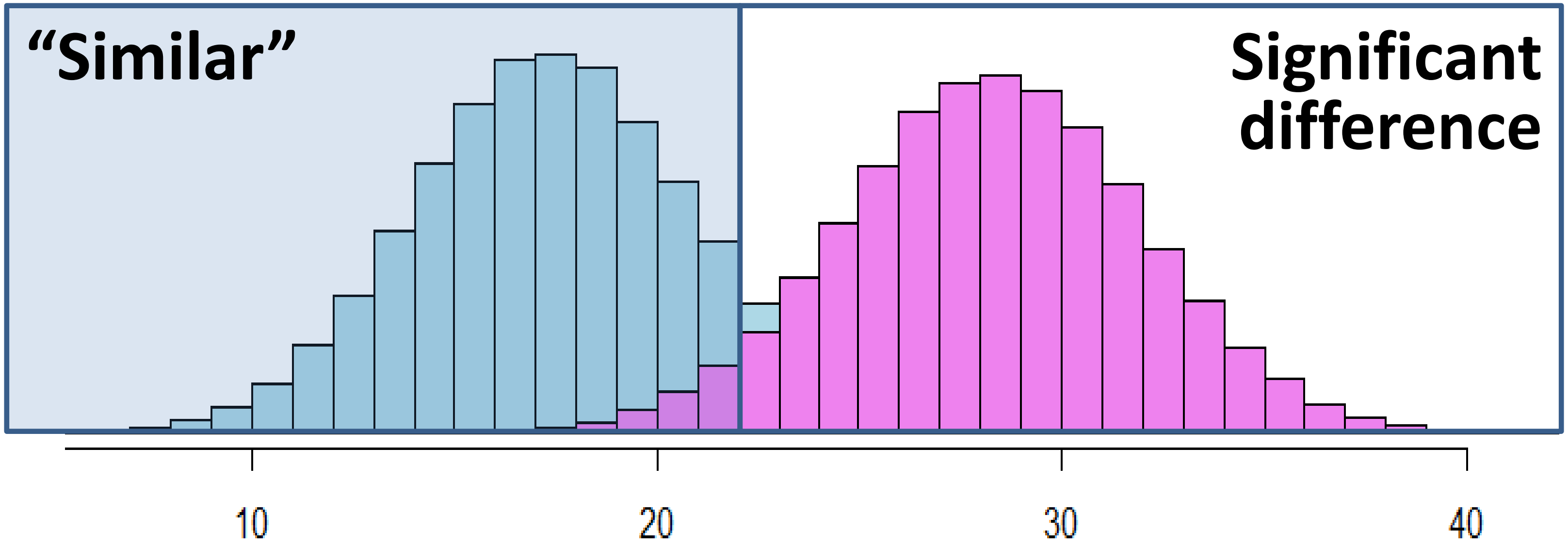
$\mathbf{H_0}$ rejected in favor of $\mathbf{H_1}$



Similarity is Affected by Shifts Large

The more right-shifted the $\mathbf{H_1}$ distribution is, the less often we concluded that products are deemed **Similar**.

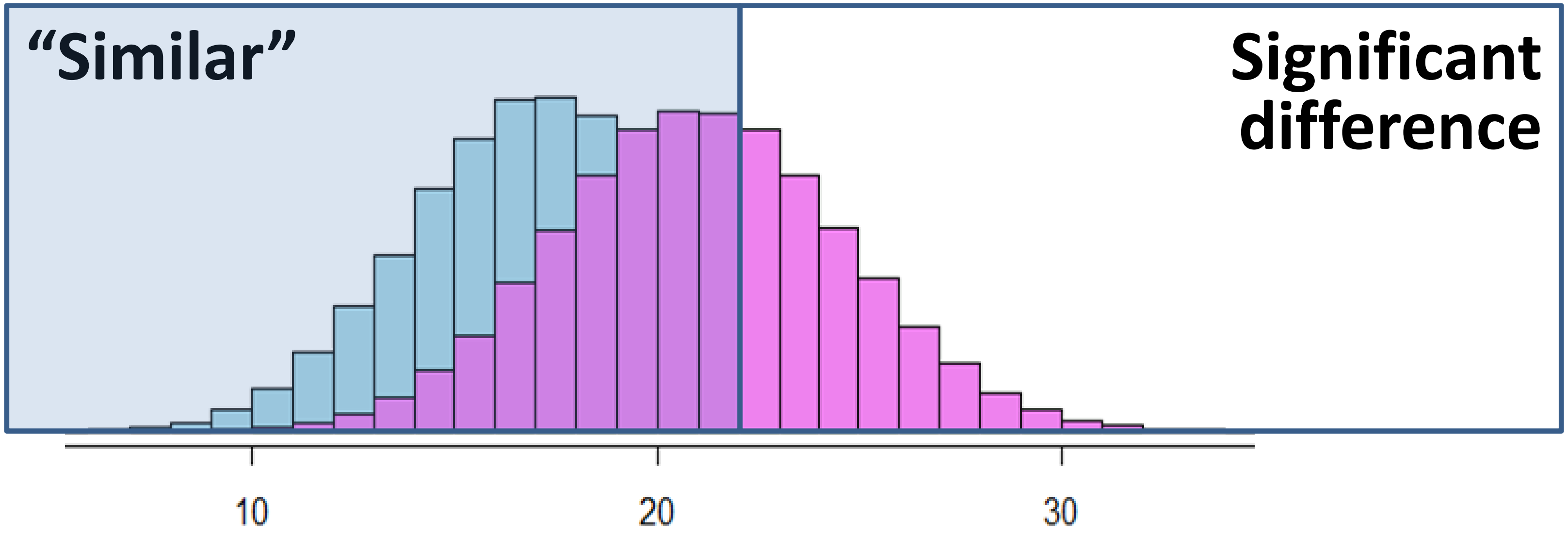
Any shift to the right illustrates this trend. This is because the distribution under $\mathbf{H_0}$ is fixed.



Similarity is Affected by Shifts Small

The more right-shifted the $\mathbf{H_1}$ distribution is, the less often we concluded that products are deemed **Similar**.

But the decrease in the proportion of **Similar** conclusions occurs for ***any*** shift of the $\mathbf{H_1}$ distribution to the right.



Changing n Affects the p-value

In practice **n** might be increased to balance serving orders or because additional assessors were invited in anticipation of no-shows. If there are no-shows for the test **n** might be reduced. Changing **n** can be problematic.

As **n** becomes larger, `sqrt(p(1-p)/n)` becomes smaller. The probability of confirming similarity decreases.
Increased precision
= increased probability of conclusion of difference
= decreased probability of confirming similarity

Returning to Example 1

That similarity was only confirmed in **49%** of the time in simulation is disappointing. (We might as well have flipped a coin.)

But it's clear why this happened: the products were not identical – *10% of people can distinguish a difference*! There was a small shift in the distribution under $\mathbf{H_1}$ against the distribution under $\mathbf{H_0}$.

A different way of thinking about this is that we always compare with $\mathbf{p_{d0}}$=0, not with the $\mathbf{p_d}$ selected by the researcher.

Another Perspective

But let’s look at this example from yet another perspective… the confidence interval of $\mathbf{p_{d(obs)}}$.

The test works such that products are determined to be **Similar** if $\mathbf{p_{d(lower)}} < 0 < \mathbf{p_{d(upper)}}$.

The confidence interval must include zero.

Confidence Interval

From the data observed in the Triangle test, Annex X4 of E 1885-04 gives instructions for obtaining a confidence interval for the proportion of distinguishers in the population:

$(\mathbf{p}_{\mathbf{d(lower)}}, \mathbf{p}_{\mathbf{d(upper)}}) = \mathbf{p}_{\mathbf{d(obs)}} \pm \mathbf{z}_{\alpha}\mathbf{s}_{\mathbf{d(obs)}}$

…where

$\mathbf{p}_{\mathbf{d(obs)}}=1.5\mathbf{p}_{\mathbf{d(obs)}}-0.5$

$\mathbf{z}_{\alpha}$ is the critical value at α from the normal distribution

$\mathbf{s}_{\mathbf{d(obs)}}=1.5\mathrm{sqrt}(\mathbf{p}_{\mathbf{c(obs)}}(1-\mathbf{p}_{\mathbf{c(obs)}})/\mathbf{n})$

Example 3

Select $\boldsymbol{\alpha}$=$\boldsymbol{\beta}$=0.05, $\mathbf{p_d}$=0.2. Thus **n**=147.

Let the true $\mathbf{p_{d(pop)}}$=0.12 and we get a very representative sample in which we observe…

$\mathbf{p_c}$ = (1-$\mathbf{p_d}$)/3 + $\mathbf{p_d}$ = 0.88/3 + 0.12 = 0.41

…thus 60 correct responses, which is greater than the critical value

$\mathbf{x_{crit}}$ = 147/3 + 1.64*sqrt(2*147/9) = 58

So we reject $\mathbf{H_0}$, and declare that the samples are **not similar**.

Example 3

We **reject** $\mathbf{H_0}$ because the test statistic indicates that the products are different.

Now, with the same data, let's get the 95% confidence interval for $\mathbf{p_{d(obs)}}$ using the method given in E 1885 X4. It is (0.054, 0.186).

You should notice two things about this interval.

Two Things to Notice…

1. The 95% confidence interval (0.054, 0.186) does not include zero ($\mathbf{p}_{\mathbf{d(lower)}} > 0$).

   There is a real shift between the $\mathbf{H_0}$ distribution and the $\mathbf{H_1}$ distribution, so this makes perfect sense.

2. The researcher set $\mathbf{p_d}$=0.2, and the 95% confidence interval is completely within the researcher's specification ($\mathbf{p}_{\mathbf{d(upper)}} < \mathbf{p_d}$).

   Yet the result is "not similar" because 0 is not in the confidence interval.

Example 2 Revisited

There was a shift under the $\mathbf{H_1}$ distribution ($\mathbf{p_d}$=0.1).

Increasing **n** enabled detection of this small but real difference. $\mathbf{H_0}$ was often rejected.

$\mathbf{H_0}$ was retained infrequently. Products were deemed **Similar** in only 2% of the tests.

Example 4

Another triangle test for similarity, with α=β=0.1, p_d=0.4. Table A1.1 in E1885-04 recommended a minimum of 25 assessors. We wouldn't normally run a similarity test with so few respondents, but this is for illustrative purposes only.

Unbeknownst to me, the true proportion of distinguishers in the population is 0.2.

Now imagine 3 alternate realities…

Example 4

Universe 1: There were 3 no-shows. The 22 assessors gave 10 correct responses, but $\mathbf{x}_{\text{crit}}$=11. Declare Similarity!

Universe 2: All 25 assessors attend. There are 12 correct responses, and $\mathbf{x}_{\text{crit}}$=12. The products are different.

Universe 3: We had 28 assessors show up and collected all results. There were 13 correct responses, and $\mathbf{x}_{\text{crit}}$=12. The products are different.

What just happened?

We lost power for detecting differences when **n** decreased. As our measurement error increased (larger variances, wider confidence intervals), so we failed to reject $\mathbf{H_0}$ more often.

Result: products were declared **Similar** more often.

Power for detecting differences increases with **n**.

Measurement error decreases (smaller variances, tighter confidence intervals), so we rejected $\mathbf{H_0}$ more often.

Result: Products are declared **Similar** less often.

An Undesirable Property

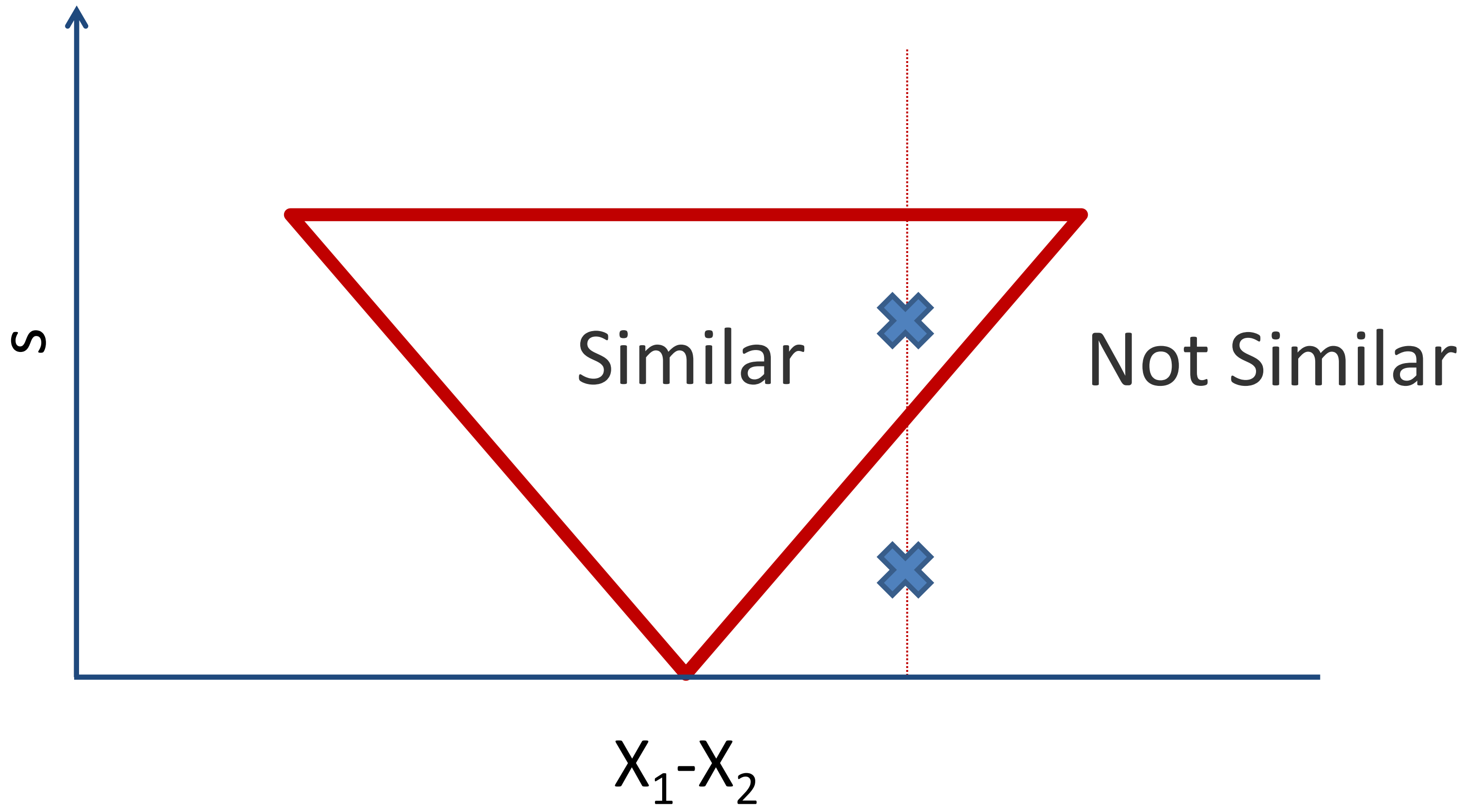
The Power Approach has the undesirable property that different products are deemed “similar” if they are observed with high variance but not with low variance.

Similarity based on Triangle

Proportions in which similarity is confirmed **n**={66, 660}

True **p_d**	n=66	n=660
0.40	0.0020	0.0000
0.35	0.0136	0.0000
0.30	0.0516	0.0000
0.25	0.1320	0.0000
0.20	0.2800	0.0000
0.15	0.5036	0.0004
0.10	0.7108	0.0248
0.05	0.8564	0.4124
0.00	0.9512	0.9480

Recall this Rejection Region…



Similarity tests will not have exactly this shape, but will share the same properties.

Back to the Context

After an ingredient substitution, current and new products don’t need to be **identical**...

...but should be similar enough.

Who decides how much is enough?

This is not a statistical question (although historical data might help to answer this question).

The researcher sets equivalence bounds, based on what is of practical relevance.

Questioning proportion of distinguishers

Proportion of distinguishers (**p_d**) is a controversial framework, because this proportion varies depending on the test method (Ennis, 1993).

Bi (2011) presents equivalence testing based on force-choice methods, in which the parameter of interest is the Thurstonian discriminal distance (***d’***; ASTM E 2262), rather than **p_d**.

Tetrads

At the Tetrad work group yesterday, Tom Carr proposed incorporating estimates of **d’** and the confidence interval of **d’** to evaluate differences and similarities.

There will be a few issues that must be worked out, but this looks like a promising direction for the Tetrads document, as well as documents that cover other sensory difference tests.

Motivating Examples

Is my product at least as good as the competitor’s product?

this is a non-inferiority question

Win. Lose. Or Draw.

You run a head-to-head test against a big competitor in a major market.

Outcomes are win, lose, or draw.

An equivalence test is inappropriate.
You would be satisfied with a draw, but an equivalence test is not appropriate here – a win is not a failure!

This is a meet-or-beat test.

Non-inferiority Test

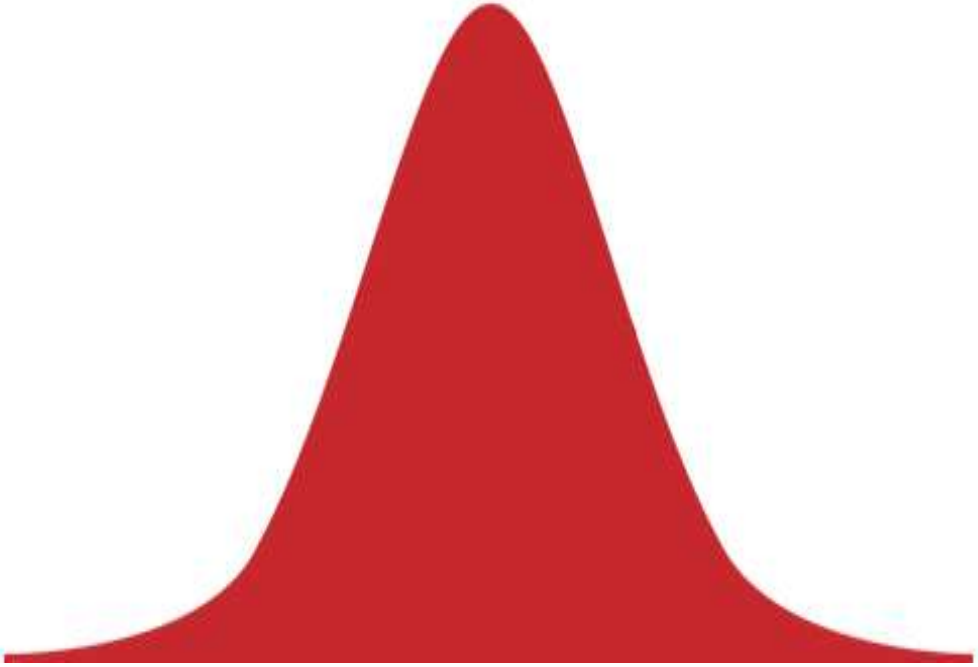
Non-inferiority can be tested using a procedure that is linked to the TOST.

One hypothesis is tested:

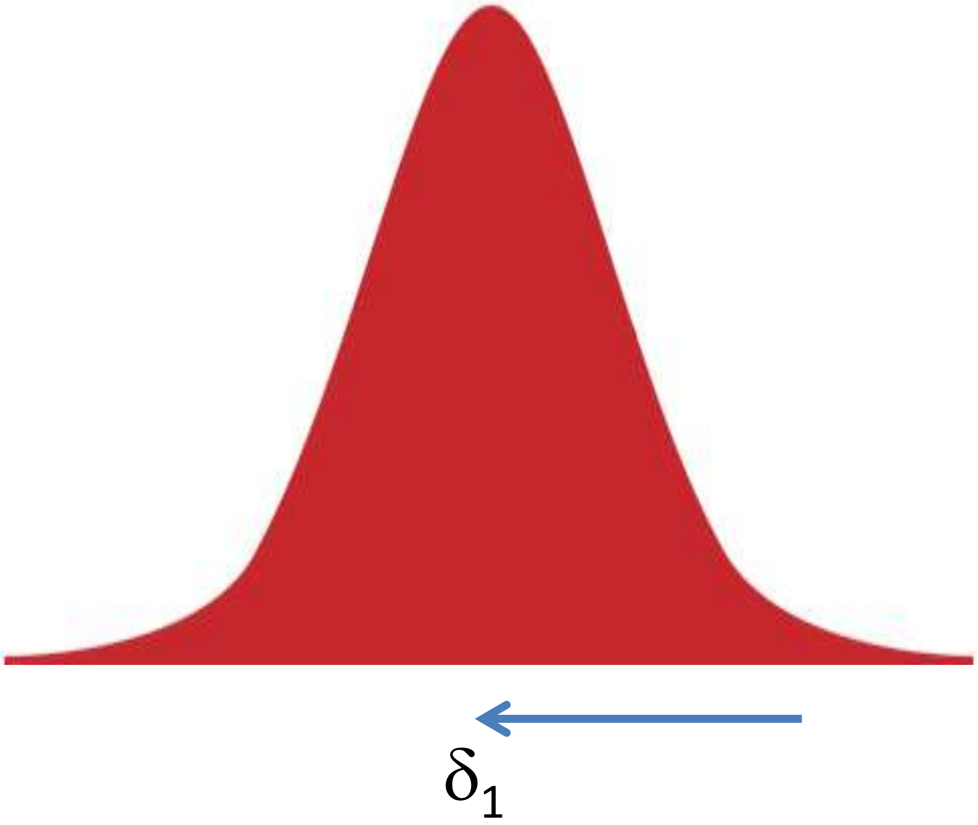
H_{01}: $\theta < \theta_{\mathbf{0}}\text{-}\delta_{\mathbf{1}}$ *vs.* H_{11}: $\theta \geq \theta_{\mathbf{0}}\text{-}\delta_{\mathbf{1}}$

If the **p**-value is significant (at level α) then we can reject $\mathbf{H_0}$ in favor of the $\mathbf{H_1}$ and declare **Non-inferiority**.

Non-inferiority Test in Action



Non-inferiority Test in Action

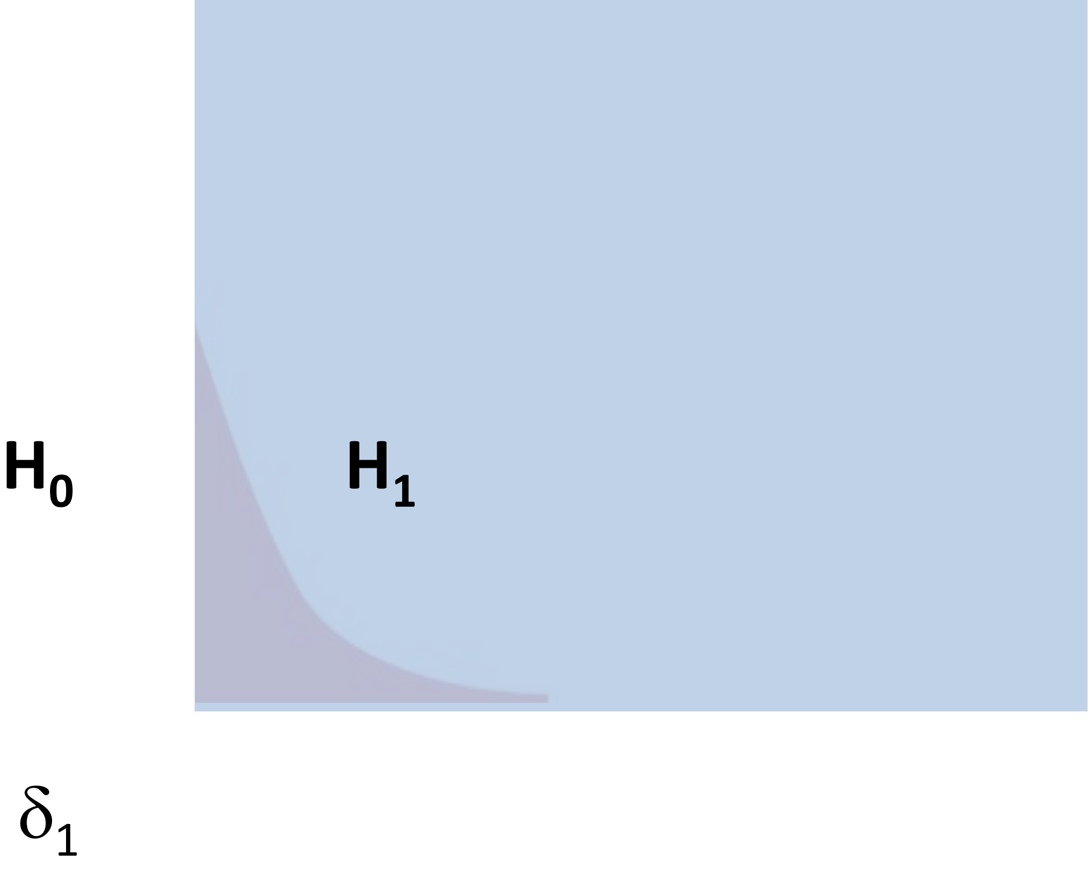


Non-inferiority Test in Action

H_0

δ_1

Non-inferiority Test in Action

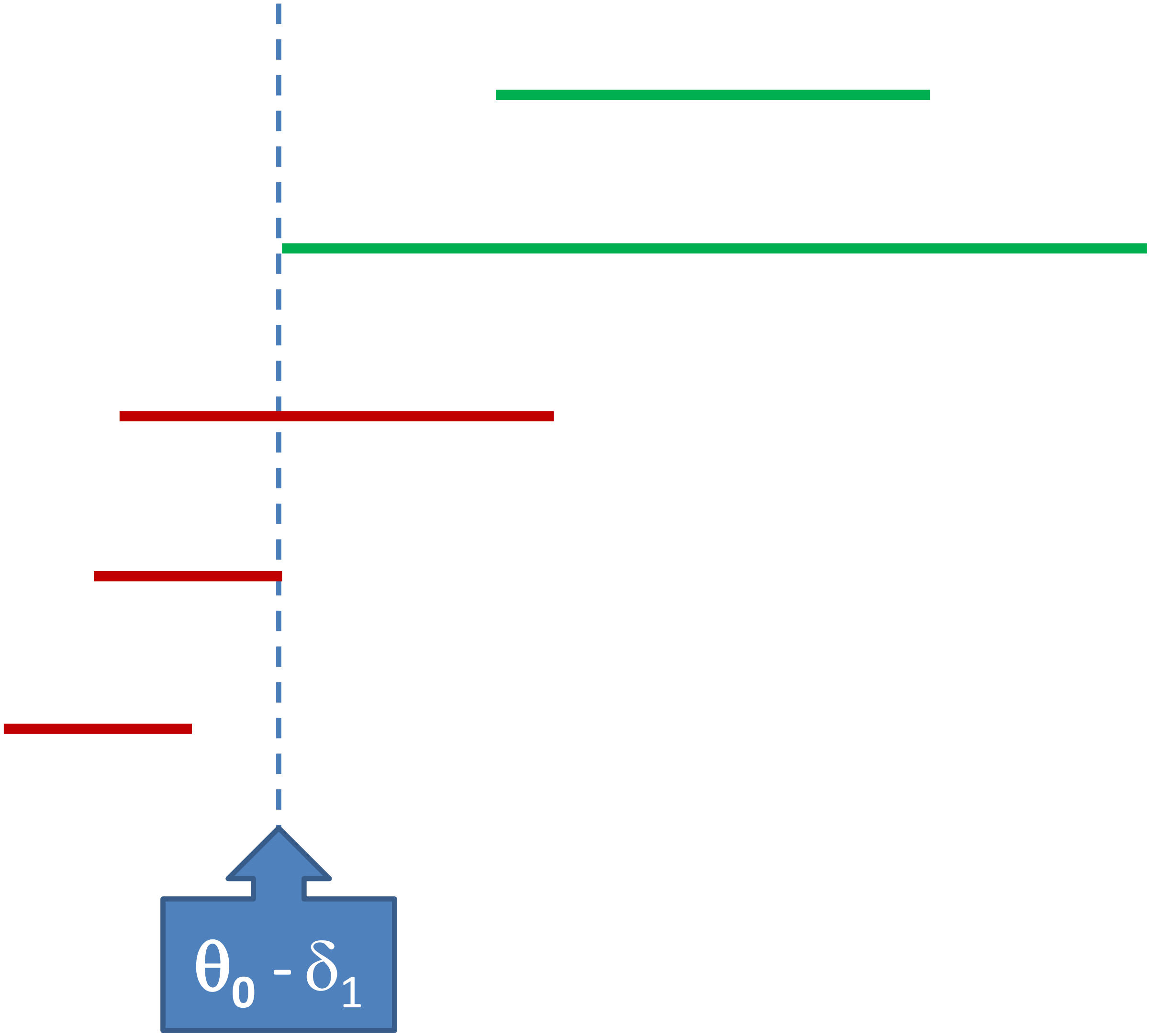


Non-inferiority & Confidence Intervals

It possible to construct (1-2α)100% confidence intervals* just as with equivalence tests, and then to determine whether the upper confidence limit is above the lower non-inferiority bound.

**See:* “Guidance for Industry E9 Statistical Principles for Clinical Trials” (Section 5.5.E)

Confidence Interval Inclusion



Thank you for your attention!