Best Practices in Equivalence Testing

John Castura

Compusense Inc.

COMPUSENSE®

10th Sensometrics, Rotterdam, July 2010

Equivalence Testing – Purposes

- Reformulation
  - e.g. Ingredient substitution
- Research and Development
  - e.g. Product matching
- Claims Substantiation
  - e.g. Detergent X cleans equivalently to the leading brand

ASTM E1958–07 Standard Guide for Sensory Claim Substantiation

Comparative
- Superiority
- Parity
  - Equality / Equivalence
  - Unsurpassed / Non-inferiority

Non-Comparative

ASTM

E1885-04 Standard Test Method for Sensory Analysis
- Triangle Test

E1958-08 Standard Guide for Sensory Claim Substantiation

E2139-05 Standard Test Method for Same-Different Test

E2164-08 Standard Test Method for Directional Difference Test

E2610-08 Standard Test Method for Sensory Analysis
- Duo-Trio Test

ISO

ISO 4120:2004 Sensory Analysis - Methodology - Triangle Test
ISO 5495:2005 Sensory Analysis - Methodology - Paired Comparison Test
ISO 10399:2004 Sensory Analysis - Methodology - Duo-Trio Test

Equivalence Testing - Background

In statistical hypothesis testing usually we have a distribution under H_0. The probability of observing a result in the tail regions is low if H_0 is true. This gives evidence to reject H_0 at the tails of the distribution.

How would a proper hypothesis test for equivalence be constructed?

H_0: Products not equivalent

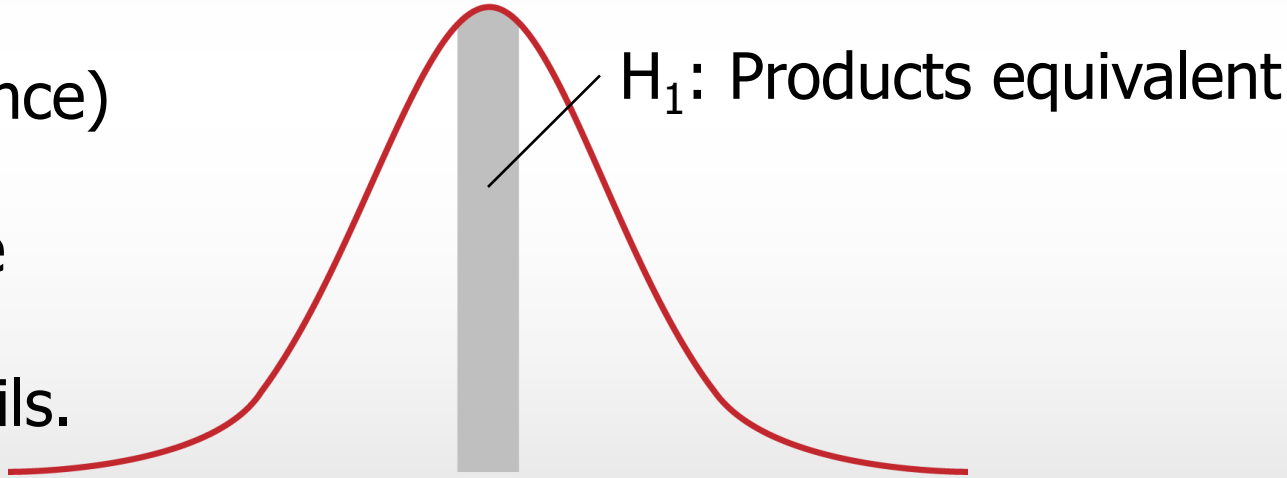
H_1: Products equivalent

What is the rejection region?

Equivalence Testing - Background

Consider the difference between two products evaluated for a sensory attribute by line scale.
Typically we reject H_0 in favour of H_1 at the tails of the distribution, which are improbable under H_0.

H_1 (equivalence) falls in the center of the distribution, not at the tails.



How do we reject H_0 in favour of H_1?

Power Approach

With the Power Approach, the difference hypothesis test is re-applied to address the equivalence scenario:

H_0: Products not different
H_1: Products different

Shift focus now to sensory difference testing methodologies...

Equivalence Testing – Power Approach

Decision \ Truth	Different	Not
Reject H_0	Correct	Type I Error α
Retain H_0	Type II Error β	Correct

Power Approach

With the Power Approach, power calculations are made to determine an appropriate sample size.

The idea is to ensure that Type II error is improbable.
β is set at some low value.
Power (1-β) is high.

Power Approach

In hypothesis testing the research hypothesis is the alternative hypothesis (H_1), not the null hypothesis (H_0).

Insufficient evidence to reject H_0 means that it is retained. It is not "proven" or "accepted".

Neither p=0.86, nor p=0.06, nor any other p-value "proves" H_0.

The hypothesis test logic has been contorted to meet the objectives.

Triangle Data Simulations - ASTM E1885-04

From Jian Bi's publication "Similarity testing in sensory and consumer research" (2005, *FQ&P*):

Select $\alpha=0.1$ and $\beta=0.05$
Assumed proportion of detectors: $p_d=0.3$
Proportion of correct responses:

$$p_c = p_d + (1/3)(1-p_d) = 0.533$$

Use "E-1885 04 Standard Test Method for Sensory Analysis – Triangle Test" to determine the number of assessors.

Triangle Data Simulations - ASTM E1885-04

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	54	64	97
0.10		30	43		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

Triangle Data Simulations - ASTM E1885-04

Assume the following is true: the products are more similar than we expected.

Proportion of detectors: $p_d=0.1$
Proportion correct responses:

$$p_c = p_d + (1/3)(1-p_d) = 0.1+0.3 = 0.4$$

If the power approach works we would expect to confirm similarity with high probability.

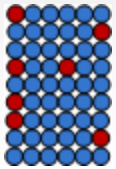
Triangle Critical Value - ASTM E1885-04

Retain H_0 when the number of correct responses is less than the number given in Table A1.2.

Standard indicates that values not in the table can be obtained from normal approximation

$x_{crit} = (n/3) + z_{\alpha} \sqrt{2n/9}$

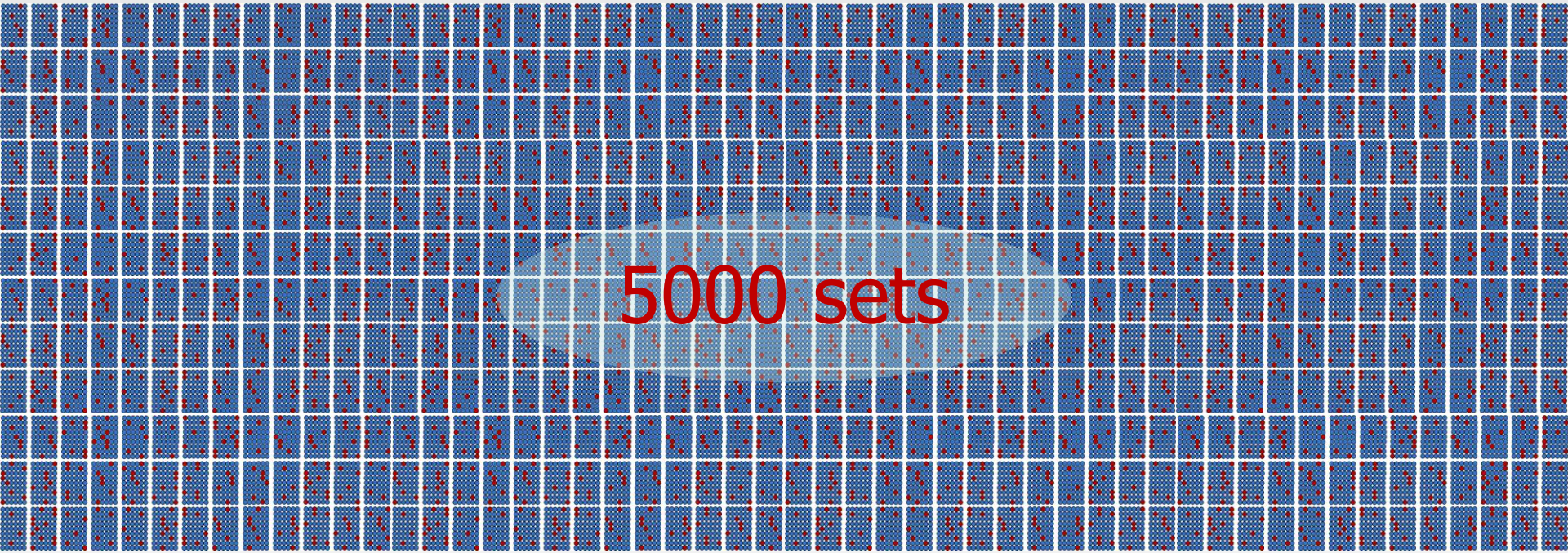
Triangle Data Simulations - ASTM E1885-04



n=54

Simulated data drawn from a population with a known proportion of detectors.

Triangle Data Simulations - ASTM E1885-04



H_0 is retained in some sets and rejected in others.
The power approach confirms similarity with probability 0.49.

Triangle Data Simulations - ASTM E1885-04

Table 1 in E1885-04 recommends a minimum of 457 assessors at $\alpha=0.1$, $\beta=0.05$, $p_d=0.1$.

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

H_0 is retained in some sets and rejected in some others.
The power approach confirms similarity with probability 0.02.

This is not good.

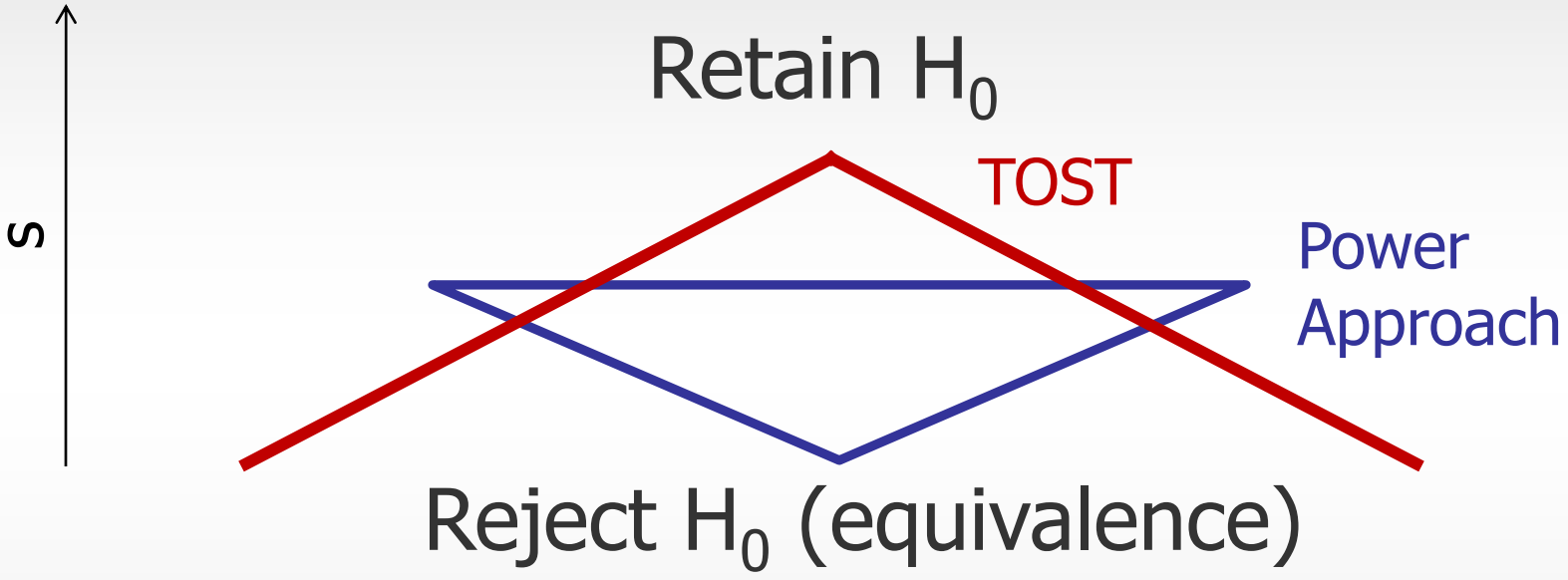
Triangle Critical Value - ASTM E1885-04

As n becomes large standard error gets small ($\sqrt{p(1-p)/n}$).
Probability of confirming similarity decreases.

Increased precision
= increased probability of conclusion of difference
= decreased probability of confirming similarity

Increasing n can be problematic.
In practice n is often increased to balance serving orders.

Equivalence Testing – Rejection Regions



Relationship between variance and rejection regions due to power approach (blue) and TOST (red) in an equivalence test with two treatments for a bioavailability variable (adapted from Schuirrmann, 1987). A similar issue exists with the power approach involving binomial data (where rejection region will follow a step function).

Triangle Critical Value - ISO 4120:2004

ISO standard 4120:2004 also provides guidance for the Triangle test. Selection of n follows the same procedure as ASTM E1885-04.

ISO 4120:2004 provides a table and formula for maximum correct responses for similarity testing significance:

$$x_{crit} = \{ x \mid p_d = (1.5(x/n)-0.5) + 1.5\, z_\beta \sqrt{(nx-x^2)/n^3} \}$$

Triangle Critical Values - ISO vs. ASTM

ISO tests whether $CI_{upper} < p_d$
p_d is defined by the researcher.

ASTM tests whether the CI includes zero.
p_d is defined by the researcher.
A CI within $(0, p_d)$ is not similar – zero must be included.

Using CI equations in E1885-04 Appendix X4 it is possible to make decisions following the ISO guidelines.

Triangle Simulation for 3 methods

Set $\alpha=\beta=0.05$ and $p_d=30\%$. Use n=66.

Let n={66, 660} and
p_d = {40%, 35%, 30%, 25%, 20%, 15%, 10%, 5%, 0%}.
2500 simulated datasets for each of the 18 scenarios.

Determine percentage of times that similarity is confirmed according to methods provided in…

(i) ASTM E1885-04
(ii) ISO 4120:2004
(iii) sensR::discrim()

Triangle Simulation for 3 methods

Percentages with which similarity confirmed when **n=66**

Method	p_d=40%	p_d=35%	p_d=30%	p_d=25%	p_d=20%
ASTM E1885-04	0.20	1.36	5.16	13.20	28.00
ISO 4120:2004	0.20	1.36	5.16	13.20	28.00
sensR::discrim()	0.20	1.36	5.16	13.20	28.00

Method	p_d=15%	p_d=10%	p_d=5%	p_d=0%
ASTM E1885-04	50.36	71.08	85.64	95.12
ISO 4120:2004	50.36	71.08	85.64	95.12
sensR::discrim()	50.36	71.08	85.64	95.12

Triangle Simulation for 3 methods

Percentages with which similarity confirmed when **n=660**

Method	p_d=40%	p_d=35%	p_d=30%	p_d=25%	p_d=20%
ASTM E1885-04	0.00	0.00	0.00	0.00	0.00
ISO 4120:2004	0.00	0.04	8.92	64.68	97.80
sensR::discrim()	0.00	0.00	5.00	51.12	95.64

Method	p_d=15%	p_d=10%	p_d=5%	p_d=0%
ASTM E1885-04	0.04	2.48	41.24	94.80
ISO 4120:2004	100.00	100.00	100.00	100.00
sensR::discrim()	100.00	100.00	100.00	100.00

Replicated Triangle Tests

Both ISO 4120:2004 and E1885-04 discourage the reader from using replicated triangle tests.

Vague wording in E1885-04 suggests that such an analysis is possible, but none is referenced.

Similarity Testing – Test Statistic

Detection shown experimentally to be stochastic, not deterministic (Ennis, 1993).

Binomial model still applies if assessors have identical detection abilities. But assessor variance means that test statistic follows different statistical distributions in H_0 and H_1.

Bi (2001) notes that it is incorrect for H_0 and H_1 to follow different distributions – power and sample size calculations based on the binomial are invalid when this principle is violated.

Duo-Trio Simulation for 3 methods

Set $\alpha=\beta=0.05$ and $p_d=30\%$. Use n=119.

Let n={119, 1190} and
p_d = {40%, 35%, 30%, 25%, 20%, 15%, 10%, 5%, 0%}.
2500 simulated datasets for each of the 18 scenarios.

Determine percentage of times that similarity is confirmed according to methods provided in…

(i) ASTM E2610-08
(ii) ISO 10399:2004
(iii) sensR::discrim()

Duo-Trio Simulation for 3 methods

Percentages with which similarity confirmed when **n=119**

Method	p_d=40%	p_d=35%	p_d=30%	p_d=25%	p_d=20%
ASTM E2610-08	0.20	1.16	4.48	13.64	29.60
ISO 10399:2004	0.20	1.16	4.48	13.64	29.60
sensR::discrim()	0.20	1.16	4.48	13.64	29.60

Method	p_d=15%	p_d=10%	p_d=5%	p_d=0%
ASTM E2610-08	50.8	72.32	86.6	94.76
ISO 10399:2004	50.8	72.32	86.6	94.76
sensR::discrim()	50.8	72.32	86.6	94.76

Duo-Trio Simulation for 3 methods

Percentages with which similarity confirmed when **n=1190**

Method	p_d=40%	p_d=35%	p_d=30%	p_d=25%	p_d=20%
ASTM E2610-08	0.00	0.00	0.00	0.00	0.00
ISO 10399:2004	0.00	0.04	5.24	56.28	97.44
sensR::discrim()	0.00	0.04	4.60	54.84	97.16

Method	p_d=15%	p_d=10%	p_d=5%	p_d=0%
ASTM E2610-08	0.00	3.68	46.80	95.68
ISO 10399:2004	100.00	100.00	100.00	100.00
sensR::discrim()	100.00	100.00	100.00	100.00

Equivalence Testing – Binomial Exact Solution

Equivalence Testing – null and alternative hypotheses

H_0: Products not equivalent

H_1: Products equivalent

$$p = \sum_{k=0}^{N-m} \binom{N}{k} (0.5-\theta)^k (0.5+\theta)^{N-k} - \sum_{k=0}^{m-1} \binom{N}{k} (0.5-\theta)^k (0.5+\theta)^{N-k}$$

Exact binomial solution from Ennis & Ennis (2008).

Equivalence Testing – Binomial Exact Solution

Equivalence and unsurpassed advertising claims using 2-AFC addressed in "Tables for Parity Testing" (Ennis, 2008).

H_0: $(p-0.5)^2 \geq 0.05^2$

H_1: $(p-0.5)^2 < 0.05^2$

Bounds defining equivalence are 45% and 55%, and true choice probability is p.

Table values based on normal approximation given in E1958-07 Standard Guide for Sensory Claim Substantiation.

Binomial Exact Test more limited in application than TOST.

Some key points…

- Confidence intervals are much preferable to hypothesis test decision
- Increasing n can have unintended consequences if following ASTM standards!
- Power approach contorts hypothesis test logic
- So far we are talking about equivalence of population average, not individual equivalence

Some key points…

- Assumption that all assessors have same detection probability is not believable
- Assumption that each assessor is either non-detector or detector is not believable
- Some interest in the beta-binomial
- Choose the best methods for the purpose
- Assessor selection and test procedure very important
- What do we really want to know?