UNSW SYDNEY | Australia's Global University

STATS CENTRAL

Mark Wainwright Analytical Central

No difference does not imply equivalence:

misuse of P value in equivalence/ non-inferiority testing

Zhixin Liu

August 8, 2019

E: stats.central@unsw.edu.au
W: https://www.analytical.unsw.edu.au/facilities/stats-central

Outline

- *Type of study in terms of hypothesis testing*
- *Why* no difference does not imply equivalence
- *How* to conduct equivalence/non-inferiority test
  - Margin specification
  - Confidence interval & P-value
- Sample size estimate
- More on **P-value**

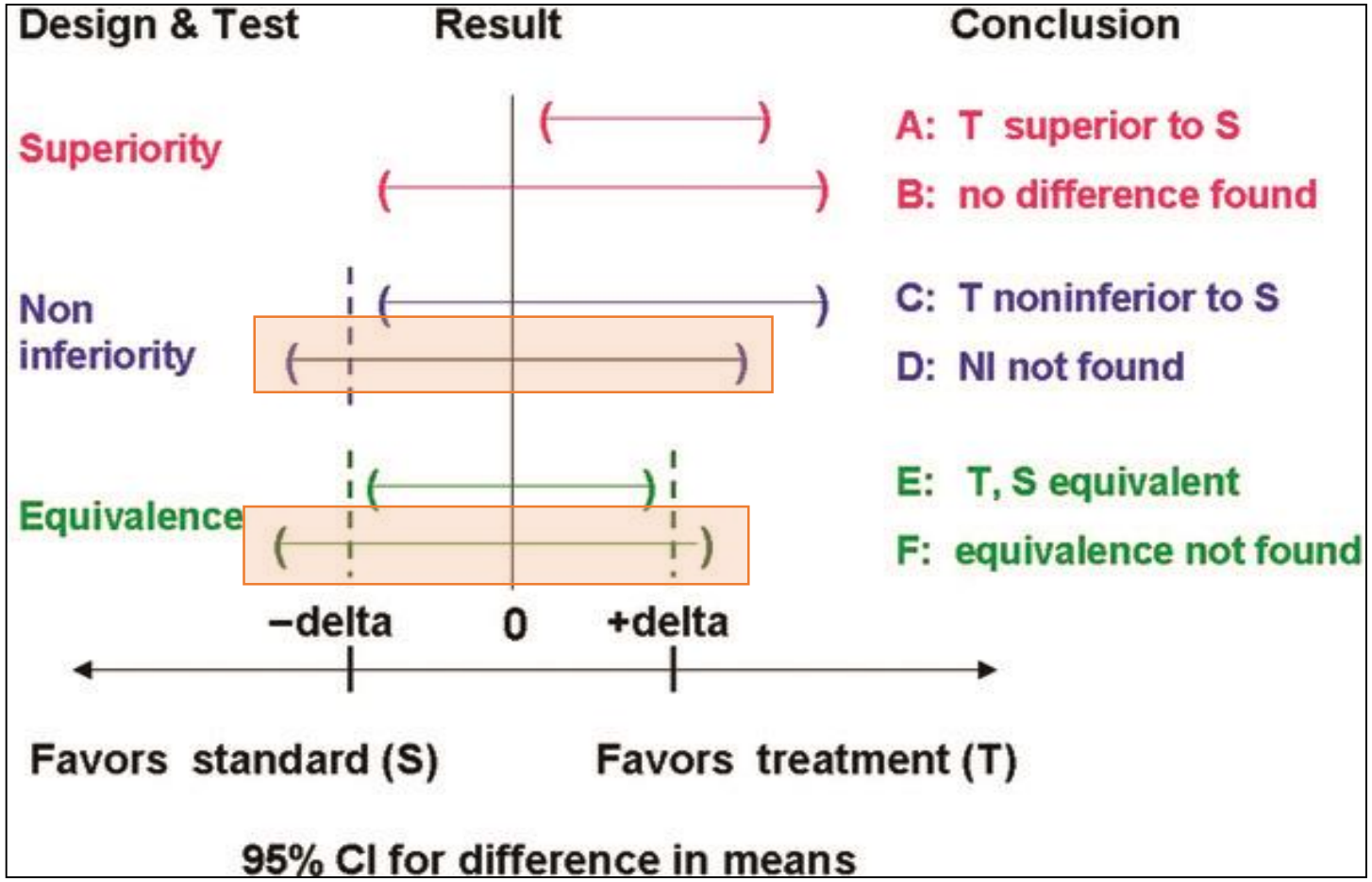
Study with different hypothesis testing

- Superiority testing
- Non- inferiority testing
  - a preferred treatment is “at least as good as” or “not worse than” a competitor or standard treatment, and it is safer and less expensive.
- Equivalence testing
  - Therapeutic equivalence, bioequivalence

Hypothesis

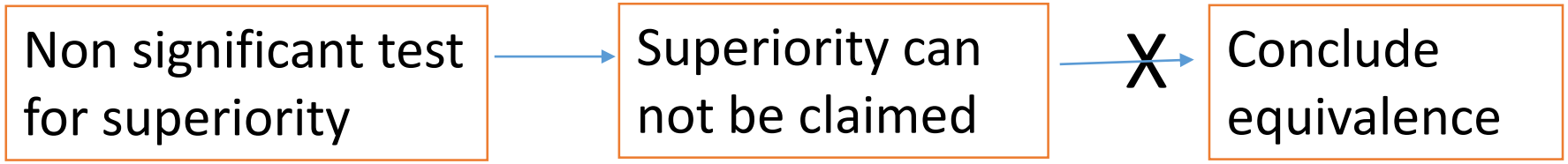
Study type	Null Hypothesis	Research hypothesis	Hypothesis test*
Superiority	There is no difference between the comparative groups	There is a difference between the comparative groups	H_0: $\mu_E - \mu_S = 0$ H_1: $\mu_E - \mu_S \neq 0$
Equivalence	The groups are **not** equivalent	The experimental group is equivalent to standard group	H_0: $\mu_E - \mu_S \leq -\delta$ OR $\mu_E - \mu_S \geq +\delta$ H_1: $\mu_E - \mu_S > -\delta$ AND $\mu_E - \mu_S < +\delta$
Noninferiority	The experimental group is inferior to the current therapy	The experimental group is not inferior to standard group	H_0: $\mu_E - \mu_S \leq -\delta$ H_0: $\mu_E - \mu_S > -\delta$

Equivalence/non-inferiority verse superiority



Equivalence and Noninferiority Testing in Regression Models and Repeated-Measures Designs. Mascha, Edward; Sessler, Daniel Anesthesia & Analgesia. 112(3):678-687, March 2011.

No difference does not imply equivalence



- Non-superiority ≠ equivalence
- It is **not valid** to assess non-inferiority or equivalence in a study designed for superiority

Construct equivalence/noninferiority testing -- Margin specification

- Smallest effect size of interest
- Minimal clinically important difference
  - *should be considerably smaller than the "clinically important difference" that would be used in a power analysis for assessing superiority of treatment versus placebo.*
  - *An equivalence/noninferiority study should be designed to minimize the possibility that a new therapy that is found to be equivalent/noninferior to the current therapy can be nonsuperior to a placebo.*
- **The value of the equivalence margin should be determined before the data is recorded.**

Construct equivalence/noninferiority testing -- inference

Study type	Hypothesis test	Test	Confidence interval*
Superiority	$H_0: \mu_E - \mu_S = 0$ $H_1: \mu_E - \mu_S \neq 0$	Two sided test (or one sided test)	100(1-α)% two sided CI
Equivalence	$H_0: \mu_E - \mu_S \leq -\delta$ OR $\mu_E - \mu_S \geq +\delta$ $H_1: \mu_E - \mu_S > -\delta$ **AND** $\mu_E - \mu_S < +\delta$	Two one sided test (TOST)	100(1-α)% one sided CI for each test 100(1-2α)% two sided CI
Noninferiority	$H_0: \mu_E - \mu_S \leq -\delta$ $H_0: \mu_E - \mu_S > -\delta$	One sided test	100(1-α)% one sided CI

*significance level α

P-value for equivalence /non-inferiority

Superiority

$$T_{\text{sup}} = \frac{\hat{\mu}_{\text{E}} - \hat{\mu}_{\text{S}}}{\sqrt{s_{\text{p}}^2 \left(1/n_{\text{E}} + 1/n_{\text{S}}\right)}},$$

Non-inferiority

$$T_{\text{L}} = \frac{\hat{\mu}_{\text{E}} - \hat{\mu}_{\text{S}} + \delta}{\sqrt{S_{\text{p}}^2 \left(1/n_{\text{E}} + 1/n_{\text{S}}\right)}}.$$

Two sided P value: P > |t|

One sided P-value: P <t or P>t

Depending on the direction of the one-tailed hypothesis, its p-value is either 0.5*(two-tailed p-value) or 1-0.5*(two-tailed p-value) if the test statistic symmetrically distributed about zero.

Practical Example

- Research question:

  Patients core temperature during surgery are kept similar under two intraoperative warming technics: Circulating water sleeve or Forced air.

- Hypothesis:

  H_0: $\mu_C - \mu_F \leq$ **-0.5** OR $\mu_C - \mu_F \geq$ **+0.5**

  H_1: $\mu_c - \mu_F >$ **-0.5** **AND** $\mu_C - \mu_F <$ **+0.5**

- Equivalence test:
  - TOST procedure: two one sided test
  - $\alpha = 0.05$

Practical Example --TOST Independent Samples T-Test

- Data summary:

	N	MEAN (SD)	Pool SD	Margin (raw)	Margin (standardized)
Circulating	37	35.96 (0.43)	0.45	-0.5, +0.5	-1.11, +1.11 (0.5/0.45)
Forced air	34	35.87 (0.47)			

- Statistical analysis:
  - install.packages("TOSTER")
  - library(TOSTER)
  - TOSTtwo(m1 = 35.96, m2 = 35.87, sd1 = 0.43, sd2 = 0.47, n1 = 37, n2 = 34, low_eqbound_d = -1.1111, high_eqbound_d = 1.1111, alpha = 0.05, var.equal = FALSE)
  - dataTOSTtwo(data, deps, group, var_equal = FALSE, low_eqbound = -0.5, high_eqbound = 0.5, eqbound_type = "raw", alpha = 0.05, desc = FALSE, plots = FALSE)

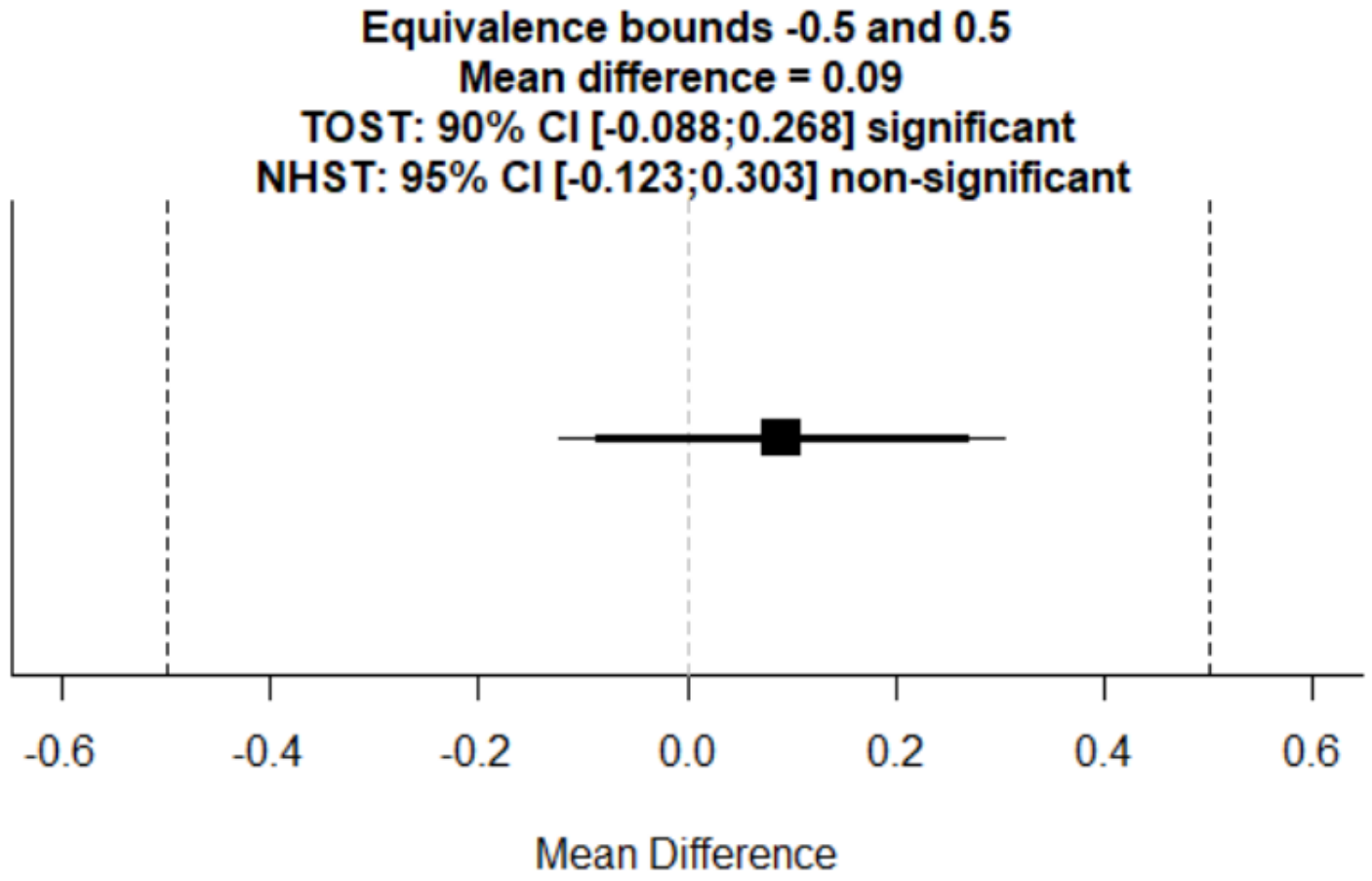
Practical Example -results

```
TOST results:
t-value lower bound: 5.52        p-value lower bound: 0.0000003
t-value upper bound: -3.83       p-value upper bound: 0.0001
degrees of freedom : 69

Null Hypothesis Test Result:
The null hypothesis test was non-significant, t(69) = 0.843, p = 0
.402, given an alpha of 0.05.

Based on the equivalence test and the null-hypothesis test combine
d, we can conclude that the observed effect is statistically not d
ifferent from zero and statistically equivalent to zero.
```

Practical Example -results



R packages

- TOSTer

https://cran.r-project.org/web/packages/TOSTER/TOSTER.pdf

- EQUIVNONINF:

https://cran.r-project.org/web/packages/EQUIVNONINF/EQUIVNONINF.pdf

Sample size

		In Reality	
		H_0 is True	H_0 is False
Decision	Fail to Reject H_0	Correct	Type II Error
	Reject H_0	Type I Error	Correct

- Equivalence margin: should be considerably smaller than the "clinically important difference" that would be used in a power analysis for assessing superiority of treatment versus placebo.
- Type I error: α (one sided),
- Type II error (power): β
  - Equivalence (Two one sided tests): power=1- β/2
  - Non-inferiority(One one sided test), power=1- β

Sample size

- Online calculator:
  - https://www.sealedenvelope.com/power/
- R package:
  - SampleSize4ClinicalTrials
  - TrialSize
  - powerTOST (bioequivalence)

Reporting

- Justification for testing an equivalence/noninferiority hypothesis as opposed to a superiority criterion.
- Clear statement and justification of the equivalence margin.
- Detailed method (including software) used to calculate the sample size needed to achieve the desired power. The method should take into account the equivalence margin. All the elements necessary to reproduce the calculation, including the proportion of dropouts anticipated, should be reported.
- The analysis section should report clearly the sets of patients analyzed and report the results of **both, the ITT and PP analyses.**
- The statistical methods should state whether the confidence interval is one- or two-sided and match the significance level used in the sample size calculation to that of confidence interval. Recall that the correct procedure to test equivalence at significance level α is to use a $(1–2\alpha) \times 100\%$ confidence interval.

Moving to a world beyond “p<0.05”

- The **ASA** Statement on p-Values
- https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.XUtrl-gzaUk
- https://www.tandfonline.com/doi/pdf/10.1080/00031305.2019.1583913?needAccess=true (recommending that declarations of “statistical significance” be abandoned)

- The **NEJM** changes to its statistical guidelines to authors

https://www.nejm.org/author-center/new-manuscripts

https://www.nejm.org/doi/full/10.1056/NEJMe1906559 (rationale for changes))

The ASA Statement on p-Values

- 1. P-values can indicate how incompatible the data are with a specified statistical model.
- 2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- 3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- 4. Proper inference requires full reporting and transparency.
- 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- 6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Key reference

- Equivalence and Noninferiority Testing in Regression Models and Repeated-Measures Designs. Mascha, Edward; Sessler, Daniel Anesthesia & Analgesia. 112(3):678-687, 2011.
- Understanding Equivalence and Noninferiority Testing. Esteban Walker and Amy S. Nowacki. Journal of General Internal Medicine. 26 (2): 192-196, 2010.
- Ng T (2015) Noninferiority Testing in Clinical Trials Boca Raton FL:Chapman & Hall/CRC.
- Wellek S (2010). Testing Statistical Hypotheses of Equivalence and Noninferiority, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

Thank you!