UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Design and Analysis of Experiments

08 - Testing Equivalence and Non-Inferiority

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
April 2015

“*Distinctions drawn by the mind are not necessarily equivalent to distinctions in reality.*”

Thomas Aquinas
1225 - 1274
Italian philosopher and theologian.

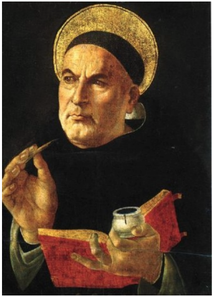


Image: `http://www.bfriars.ox.ac.uk/hall/aquinas-ins/`

Testing equivalence
Introduction

The tests introduced in the preceding chapters deal with situations in which one is interested in detecting *differences* between a population parameter θ – e.g., a population mean μ or a difference between population means $(\mu_1 - \mu_2)$ – and its nominal value θ_0 under a null hypothesis;

Another useful class of experiments in engineering and science is one in which the experimenter is interesting in investigating *equivalence* (within a given margin of error), for instance:

- Conformity/compliance testing (industrial certification);
- Equivalence of effects (pharmaceutical industry);

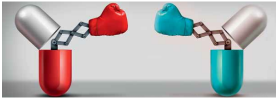


Image adapted from: `http://goo.gl/iXeCiY`

Testing equivalence

Introduction

In principle, one could express this as a shift in focus from trying to establish whether a population parameter is different from a given reference to trying to determine whether it is equal to that reference.

In usual (two-sided) comparative studies, the alternative hypothesis (i.e., the one that presents novelty in relation to the current state of knowledge) is the one of difference between the parameters of interest - that is, unless there is strong evidence of differences, one cannot rule out the null hypothesis of equality;

Testing equivalence

Introduction

In equivalence testing, the situation is reversed: the (approximate) equality of two parameters is the novelty one hopes to establish. Consequently, the burden of proof shifts to providing evidence that there is no difference.

The term *equivalent* is not used strictly, but to mean the absence of practical differences - that is, any differences that might exist fall within an *equivalence margin* or *limit of practical significance* δ^*.

Using this approach, the equivalence of two parameters can be established if a sample provides enough evidence that the true difference is smaller than δ^* units.

Testing Non-inferiority

Definition

A similar concept to equivalence testing is the definition of non-inferiority of a given treatment/ process/ method in relation to another (e.g., a standard solution).

In non-inferiority tests, one can declare that a given process is not worse than a standard one only if enough evidence is provided to conclude that the performance of the proposed process is no more than δ^* units worse than that of the standard.

In the case of non-inferiority tests, one can in principle use a regular test of differences with a one-sided alternative (which would be equivalent to setting $\delta^* = 0$), or define the null hypothesis in a way that includes δ^* in its formulation.

Comparison of studies

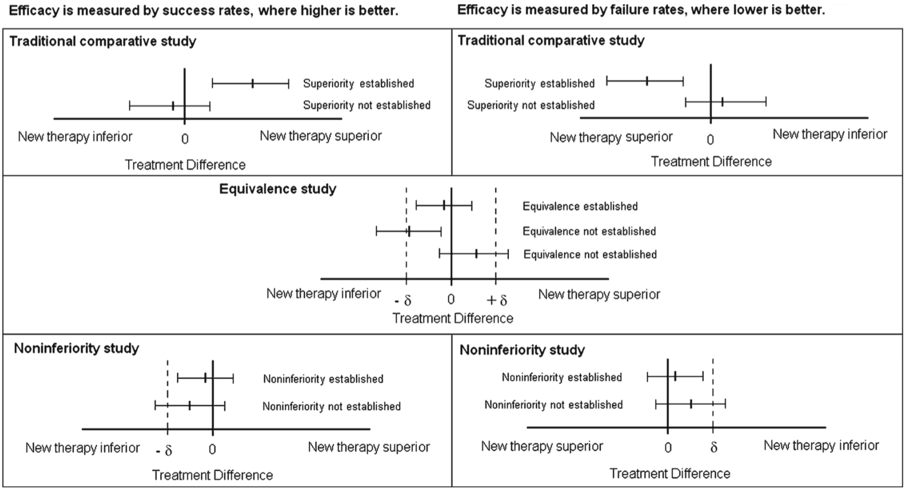


Image: Walker and Nowacki (2011), J. General Internal Medicine 26(2):192-196.

Testing Equivalence

Quick-and-dirty approach

A simple way of thinking about testing equivalence of two means is to observe confidence intervals instead of p-values:

> "*Equivalence can be established at the α significance level if a $(1-2\alpha)$-confidence interval for the difference between the two means is contained within a interval $\pm\delta^*$.*"

The difference between testing for differences and for equivalence can be easily illustrated using this approach:

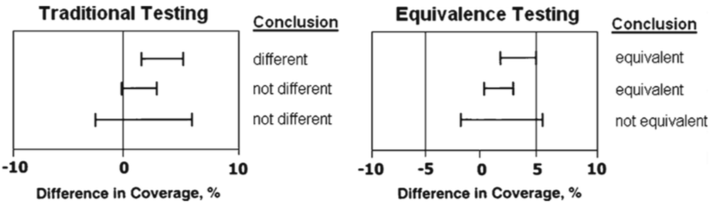


Image: Walker and Nowacki (2011), J. General Internal Medicine 26(2):192-196.

Equivalence test for a single mean

Hypotheses

An equivalence test for a single population mean can be expressed by the hypotheses:

$$\begin{cases} H_0 : |\mu - \mu_0| = & \Delta\mu \geq \delta^* \\ H_1 : & \Delta\mu < \delta^* \end{cases}$$

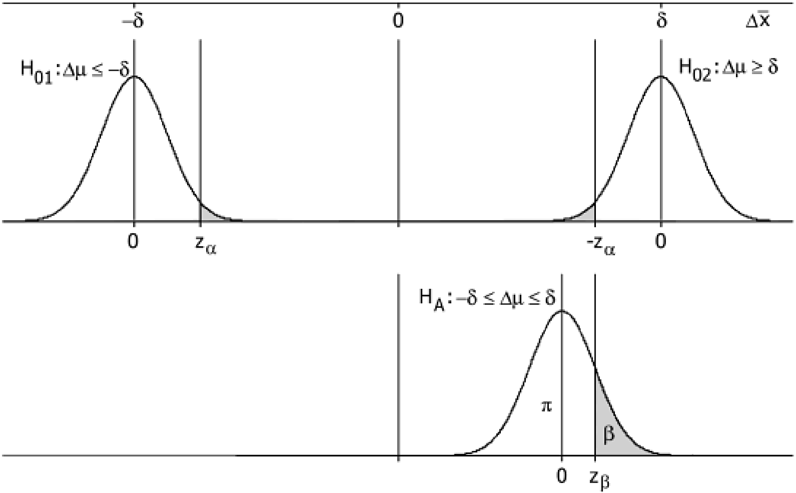
The most usual way of testing these hypotheses is the TOST (*two one-sided tests*) method. As the name suggests, two one-sided significance tests are constructed so that the desired statistical properties can be achieved. Using our standard notation:

$$\begin{cases} H_0^1 : & \Delta\mu = -\delta^* \\ H_1^1 : & \Delta\mu > -\delta^* \end{cases} \qquad \begin{cases} H_0^2 : & \Delta\mu = \delta^* \\ H_1^2 : & \Delta\mu < \delta^* \end{cases}$$

If both tests reject their respective H_0, then equivalence (within the equivalence margin δ^*) can be declared with significance level α.

Equivalence test for a single mean

Graphical interpretation



$-\delta$
0
δ
$\Delta\bar{x}$
$H_{01}: \Delta\mu \leq -\delta$
$H_{02}: \Delta\mu \geq \delta$
0
z_α
$-z_\alpha$
0
$H_A: -\delta \leq \Delta\mu \leq \delta$
π
β
0
z_β

Image: Matthews (2010), Sample Size Calculations, MMB. pg. 46

Equivalence of a single mean

Sample size

Sample sizes for testing equivalence of a single mean can be derived using essentially the same considerations used for the usual tests. In the case of a single sample:

$$n \geq \left(\frac{(t_\alpha + t_\beta)\,\hat{\sigma}}{\delta^* - \Delta\mu} \right)^2$$

As in the previous cases, iteration is needed to solve for *n* (since the quantiles of the t distribution depend on *n*). Use $t_x = z_x$ for the first iteration.

Equivalence of two means

Hypotheses

Analogously to the single sample test of equivalence, the hypotheses for testing the equivalence of two population means can be described as:

$$\begin{cases} H_0: & \mu_1 - \mu_2 \geq \delta^* \\ H_1: & \mu_1 - \mu_2 < \delta^* \end{cases}$$

$$\begin{cases} H_0^1: & \mu_1 - \mu_2 = -\delta^* \\ H_1^1: & \mu_1 - \mu_2 > -\delta^* \end{cases} \qquad \begin{cases} H_0^2: & \mu_1 - \mu_2 = \delta^* \\ H_1^2: & \mu_1 - \mu_2 < \delta^* \end{cases}$$

Just as in the previous case, both hypotheses are tested at the desired α value, and the rejection of both H_0 indicates evidence of equivalence.

Equivalence of two means

Sample size

Sample size for the $n_1 = n_2 = n$ case can be approximated based on the Zhang formula[a]:

$$n \geq \left(t_{\alpha;\nu} + t_{(1-c)\beta;\nu}\right)^2 \left(\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{\delta^* - \Delta\mu^*}\right)^2$$

with $\Delta\mu^* < \delta^*$ as the maximum real difference between the two means for which a power of $(1 - \beta)$ is desired, and:

$$c = \frac{1}{2} \exp\left(-7.06 \frac{\Delta\mu^*}{\delta^*}\right)$$

The degrees of freedom ν of the t-quantiles are given by the Welch t-test formula (see Chapter 6).

[a] Zhang (2003), J. Biopharm. Stat. 13(3):529-538.

Example

Laboratory certification

A ballistics laboratory is in the process of being certified for the evaluation of shielding technology, and needs to provide evidence of equivalence of a given callibration procedure with the reference equipment;

The certification authority demands that the mean hole area generated by this procedure in the lab be the same as the one from the reference equipment, and tolerates deviations no greater than $4mm^2$;

From previous measurements, the standard deviations can be roughly estimated as $\hat{\sigma}_{Lab} = 5mm^2$ and $\hat{\sigma}_{ref} = 10mm^2$.

The desired error levels for the comparison are $\alpha = 0.01$ and $\beta = 0.1$.

Image: http://www.everydaynodaysoff.com/2013/08/05/ballistic-shield-for-operators-only/

Example

Laboratory certification

To calculate the required sample size, assume that $\Delta\mu^* = 0.5$. Then:

```
> # load functions to calculate sample size for TOST
> source("calcN_tost.R")
>
> # Calculate sample size
> calcN_tost2(alpha = 0.01,
+             beta = 0.1,
+             diff_mu = 0.5,
+             tolmargin = 4,
+             s1 = 5,
+             s2 = 10)
[1] 144.1999
```

We'll need 145 observations from each group to test for equivalence with the desired experimental properties.

Example

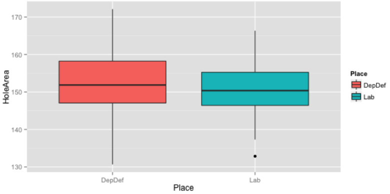
Laboratory certification

After collecting the observations, we proceed to the analysis:

```
> data<-read.table("../data files/labdata-example.csv",
+                  header = T, sep = ",")

> # Two one-sided t-tests
> t.test(HoleArea~Place,  data = data,  alternative = "less", mu = 4,
+       conf.level = 0.99)$p.value
[1] 0.00304124
> t.test(HoleArea~Place, data = data, alternative = "greater", mu = -4,
+       conf.level = 0.99)$p.value
[1] 6.586193e-10

> # Get (1-2*alpha) CI
> t.test(HoleArea~Place, data = data,  conf.level = 0.98)$conf.int
[1] -0.5117627  3.6244386
```

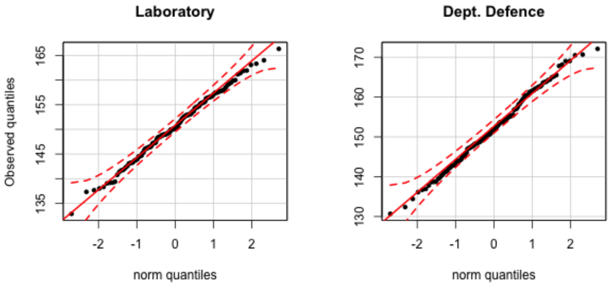


Example

Laboratory certification

Verification of test assumptions:

```
> par(mfrow=c(1,2))
> qqPlot(subset(data, Place=="Lab")[,2],
+        pch=20,
+        main = "Laboratory",
+        ylab = "Observed quantiles")
> qqPlot(subset(data, Place=="DepDef")[,2],
+        pch=20,
+        main = "Dept. Defence",
+        ylab = " ")
```



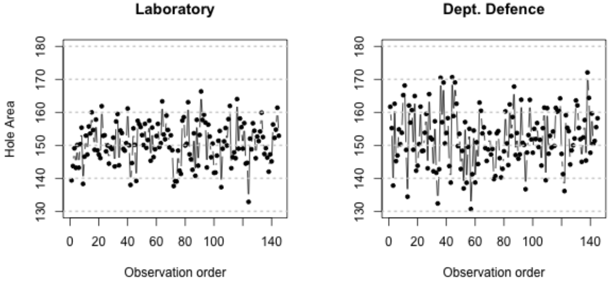
Example

Laboratory certification

Verification of test assumptions:

```
> dwtest(HoleArea~Place, data=data)
DW = 1.8116, p-value = 0.04757

> par(mfrow=c(1,2))
> plot(seq_along(subset(data, Place=="Lab")[,2]),
+       subset(data, Place=="Lab")[,2], ...)
> plot(seq_along(subset(data, Place=="DepDef")[,2]),
+       subset(data, Place=="DepDef")[,2], ...)
```



Bibliography

Required reading

1. E. Walker, A.S. Nowacki, *Understanding Equivalence and Noninferiority Testing*, Journal of General Internal Medicine 26(2):192-196, 2011.

Recommended reading

1. P. Mathews, *Sample Size Calculations: Practical Methods for Engineers and Scientists*, Ch. 2.4, 1st ed., MMB, 2010.
2. P. Zhang, *A Simple Formula for Sample Size Calculation in Equivalence Studies*, Journal of Biopharmaceutical Statistics 13(3):529-538, 2003.

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments*.
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.11, Chapter 8; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 8; Creative Commons BY-NC-SA 4.0.},
}
```