## Hypothesis testing

Hypothesis testing, introduced by Fisher, Neyman, and by Karl Pearson and his son Egon, is a method of statistical inference, used in making statistical decisions based on experimental data. It is basically an assumption made about the population parameter.

The usual process of hypothesis testing consists of the following steps:

1. formulate the so-called null hypothesis $H_0$ and the alternative hypothesis $H_1$;
2. identify a test statistic T, that can be used to assess the truth of the null hypothesis.
3. infer the distribution of the test statistic under the null hypothesis from the assumptions (e.g., the test statistic might follow a Student's t or a normal distribution).
4. select a significance level (α), a probability threshold below which the null hypothesis will be rejected. Common values are 5% and 1%, leading to respective critical values of T.
5. compute the observed value (realization) t of the statistic T from the observations and compare it with the critical value of T.

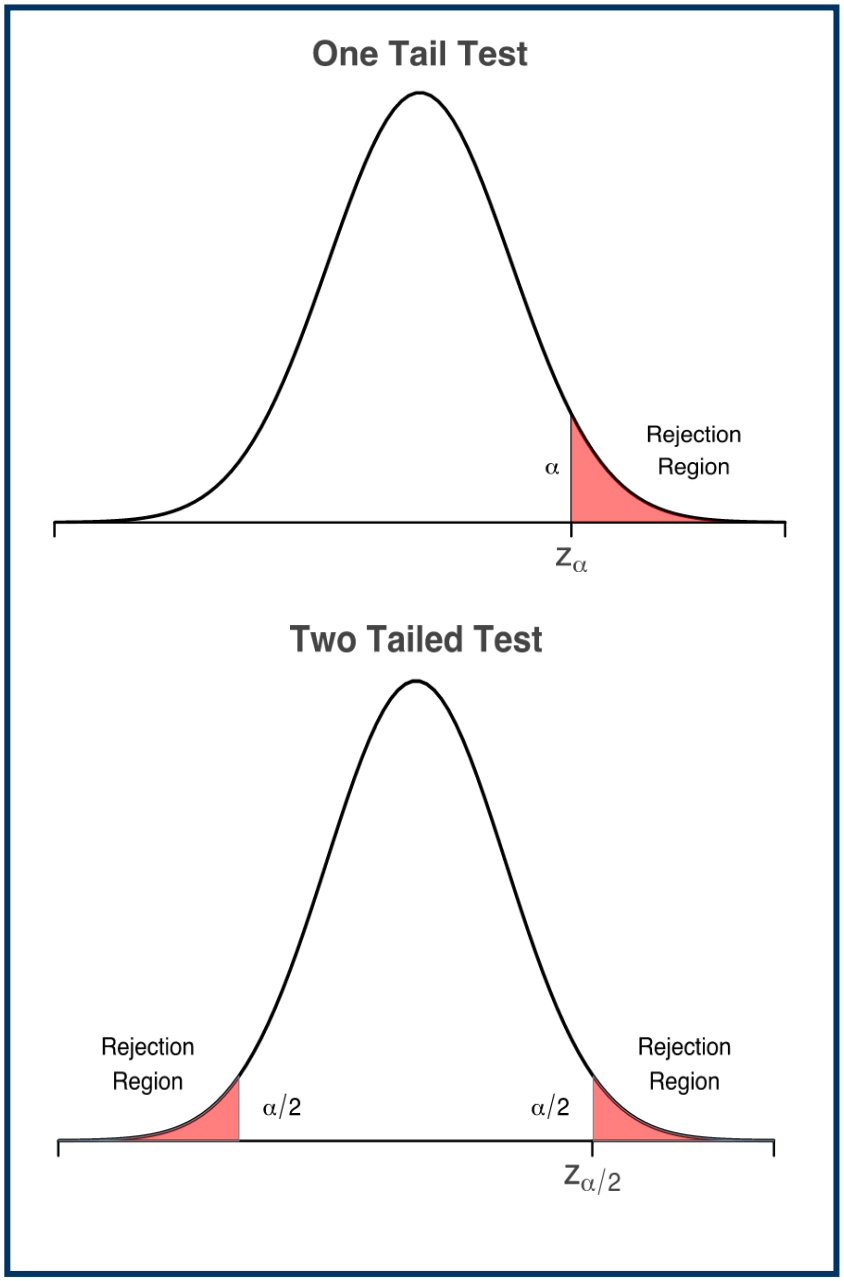
Hypothesis tests based on statistical significance are another way of expressing confidence intervals. In other words, every hypothesis test based on significance can be obtained via a confidence interval, and every confidence interval can be obtained via a hypothesis test based on significance.

The type of hypothesis testing depends on the formulation of the alternative hypothesis.

For tests related to mean (comparison between a mean and a known value or between two means) $H_0$ ad $H_1$ are formulated as follows:

| One-Tail Test (left tail) | Two-Tail Test | One-Tail Test (right tail) |
|---|---|---|
| $H_0 : \mu = \mu_0$ <br> $H_1 : \mu < \mu_0$ | $H_0 : \mu = \mu_0$ <br> $H_1 : \mu \neq \mu_0$ | $H_0 : \mu = \mu_0$ <br> $H_1 : \mu > \mu_0$ |
| $H_0: \mu_1 = \mu_2$ <br> $H_1: \mu_1 < \mu_2$ | $H_0: \mu_1 = \mu_2$ <br> $H_1: \mu_1 \neq \mu_2$ | $H_0: \mu_1 = \mu_2$ <br> $H_1: \mu_1 > \mu_2$ |

In the case of a two tailed test two critical values are defined.



**Hypothesis testing for normally distributed populations: comparison between a sampling mean and a known value**

| Case | hypotheses | statistic and type of test | rejection criteria for null hypothesis |
|---|---|---|---|
| 1: normal distribution $\sigma^2$ known | $H_0: \mu = \mu_0$ | $T = (\bar{X} - \mu_0) / (\sigma/\sqrt{n}) \sim N(0,1)$ | |
| | $H_1: \mu > \mu_0$ | one tail | $t \geq z_{(1-\alpha)}$ |
| | $H_1: \mu < \mu_0$ | one tail | $t \leq -z_{(1-\alpha)}$ |
| | $H_1: \mu \neq \mu_0$ | two tails | $\lvert t \rvert \geq z_{(1-\alpha/2)}$ |
| 2: normal distribution $\sigma^2$ unknown, n > 30 | $H_0: \mu = \mu_0$ | $T = (\bar{X} - \mu_0) / (s/\sqrt{n}) \sim N(0,1)$ | |
| | $H_1: \mu > \mu_0$ | one tail | $t \geq z_{(1-\alpha)}$ |
| | $H_1: \mu < \mu_0$ | one tail | $t \leq -z_{(1-\alpha)}$ |
| | $H_1: \mu \neq \mu_0$ | two tails | $\lvert t \rvert \geq z_{(1-\alpha/2)}$ |
| 3: normal distribution $\sigma^2$ unknown, n < 30 | $H_0: \mu = \mu_0$ | $T = (\bar{X} - \mu_0) / (s/\sqrt{n}) \sim t_{n-1}$ | |
| | $H_1: \mu > \mu_0$ | one tail | $t \geq t_{n-1,(1-\alpha)}$ |
| | $H_1: \mu < \mu_0$ | one tail | $t \leq -t_{n-1,(1-\alpha)}$ |
| | $H_1: \mu \neq \mu_0$ | two tails | $\lvert t \rvert \geq t_{n-1,(1-\alpha/2)}$ |

# Hypothesis testing for normally distributed populations: comparison between two means

| Case | hypotheses | statistic | rejection criteria |
|---|---|---|---|
| 1: normal distributions<br>$\sigma_1^2$ and $\sigma_2^2$ known | $H_0: \mu_1 - \mu_2 = 0$<br>$H_1: \mu_1 - \mu_2 > 0$<br>$H_1: \mu_1 - \mu_2 < 0$<br>$H_1: \mu_1 - \mu_2 \neq 0$ | $T = \dfrac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$ | $t \geq z_{(1-\alpha)}$<br>$t \leq -z_{(1-\alpha)}$<br>$\lvert t \rvert \geq z_{(1-\alpha/2)}$ |
| 2: normal distributions<br>$\sigma_1^2$ and $\sigma_2^2$ unknown<br>$n_1$ and $n_2 > 30$ | $H_0: \mu_1 - \mu_2 = 0$<br>$H_1: \mu_1 - \mu_2 > 0$<br>$H_1: \mu_1 - \mu_2 < 0$<br>$H_1: \mu_1 - \mu_2 \neq 0$ | $T = \dfrac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim N(0,1)$ | $t \geq z_{(1-\alpha)}$<br>$t \leq -z_{(1-\alpha)}$<br>$\lvert t \rvert \geq z_{(1-\alpha/2)}$ |
| 3: normal distributions<br>$\sigma_1^2$ and $\sigma_2^2$ unknown<br>but equal<br>$n_1$ and/or $n_2 > 30$ | $H_0: \mu_1 - \mu_2 = 0$<br>$H_1: \mu_1 - \mu_2 > 0$<br>$H_1: \mu_1 - \mu_2 < 0$<br>$H_1: \mu_1 - \mu_2 \neq 0$ | $T = \dfrac{(\bar{X} - \bar{Y})}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n1+n2-2}$ | $t \geq t_{n1+n2-2,\,(1-\alpha)}$<br>$t \leq -t_{n1+n2-2,\,(1-\alpha)}$<br>$\lvert t \rvert \geq t_{n1+n2-2,\,(1-\alpha/2)}$ |
| 4: normal distributions<br>$\sigma_1^2$ and $\sigma_2^2$ unknown<br>but different<br>$n_1$ and/or $n_2 > 30$ | $H_0: \mu_1 - \mu_2 = 0$<br>$H_1: \mu_1 - \mu_2 > 0$<br>$H_1: \mu_1 - \mu_2 < 0$<br>$H_1: \mu_1 - \mu_2 \neq 0$ | $T = \dfrac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim t_\nu$ | $t \geq t_{\nu,\,(1-\alpha)}$<br>$t \leq -t_{\nu,\,(1-\alpha)}$<br>$\lvert t \rvert \geq t_{\nu,\,(1-\alpha/2)}$ |

$$s = \sqrt{\frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2}}$$

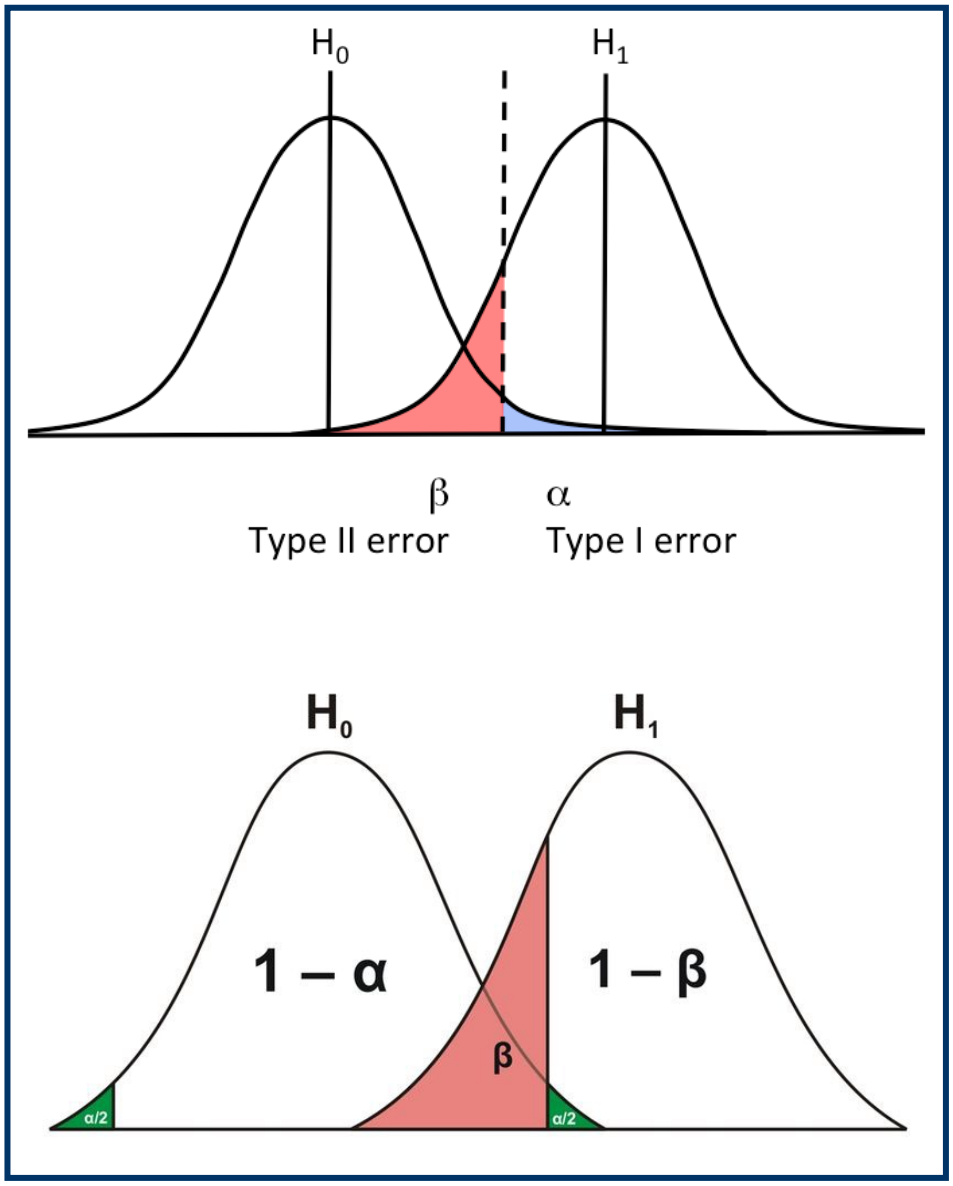
$$\frac{1}{\nu} = \frac{1}{n_1 - 1}\left(\frac{s_1^2 / n_1}{s_1^2 / n_1 + s_2^2 / n_2}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2 / n_2}{s_1^2 / n_1 + s_2^2 / n_2}\right)^2$$

# Hypothesis testing: type I and type II errors, power of the test

In statistical hypothesis testing a **type I error** is the rejection of a true null hypothesis (also known as a "false positive" finding or conclusion), while a **type II error** is the acceptance of a false null hypothesis (also known as a "false negative" finding or conclusion).

The probability of type I error corresponds to the significance level $\alpha$; the probability of the type II error is denoted by the Greek letter $\beta$ and is related to the power of a test, i.e., to the probability of accepting a true $H_1$ hypothesis, which equals $1-\beta$.

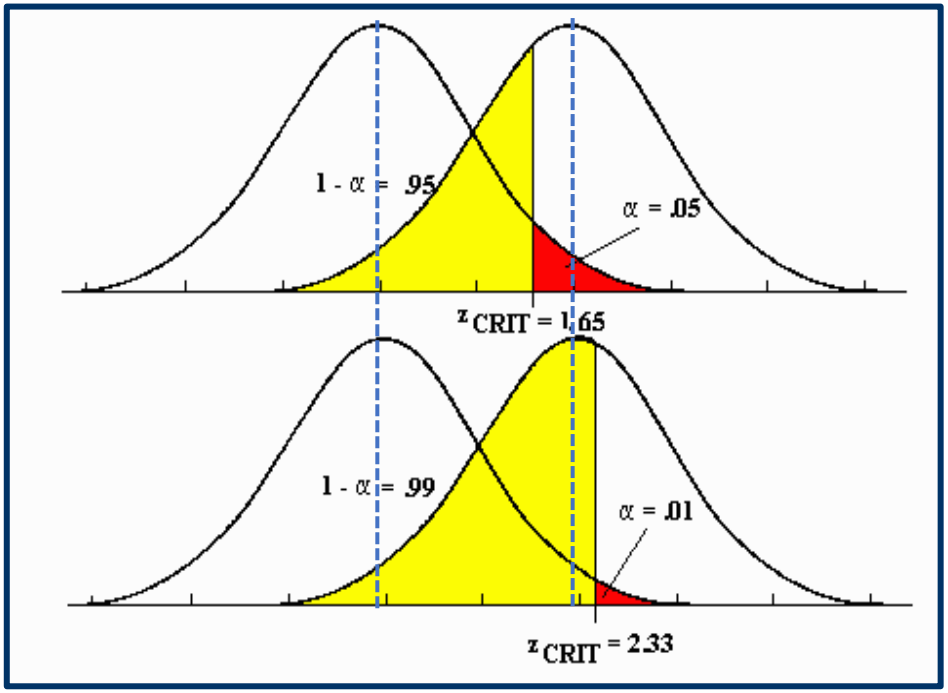
The representations related to a one tail and to a two tails test are shown in the picture on the right.



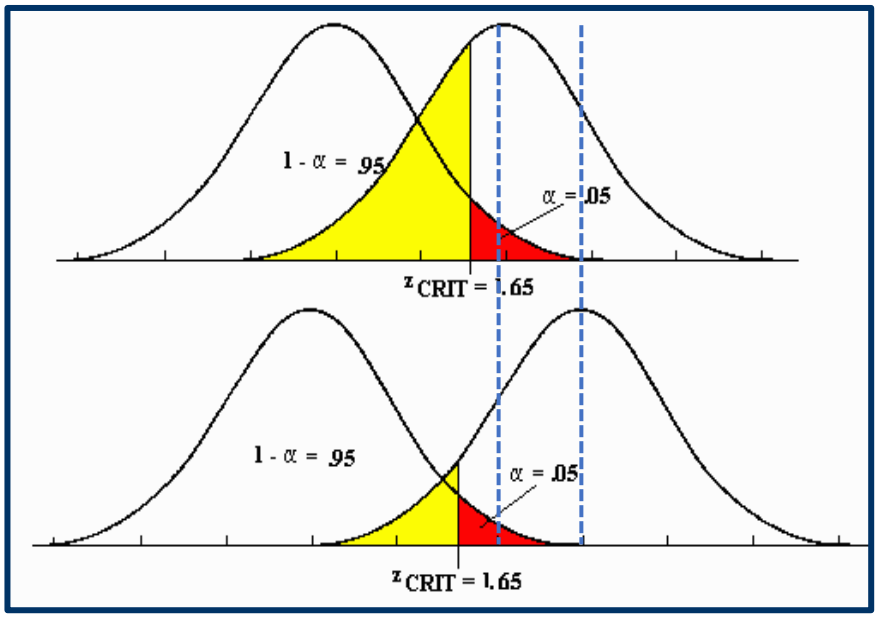
As shown in the table, in the context of analytical chemistry the power of the test represents its ability to recognize correctly a positivity, i.e., to state the presence of an analyte, when the analyte is actually present in a sample.
This ability is very important, since a positivity could lead, for example, to discard an entire lot of product.

| | | Reality | |
|---|---|---|---|
| | | Positive | Negative |
| Study Finding | Positive | **True Positive** (Power) (1-β) | False Positive **Type I Error** (α) |
| | Negative | False Negative **Type II Error** (β) | **True Negative** |

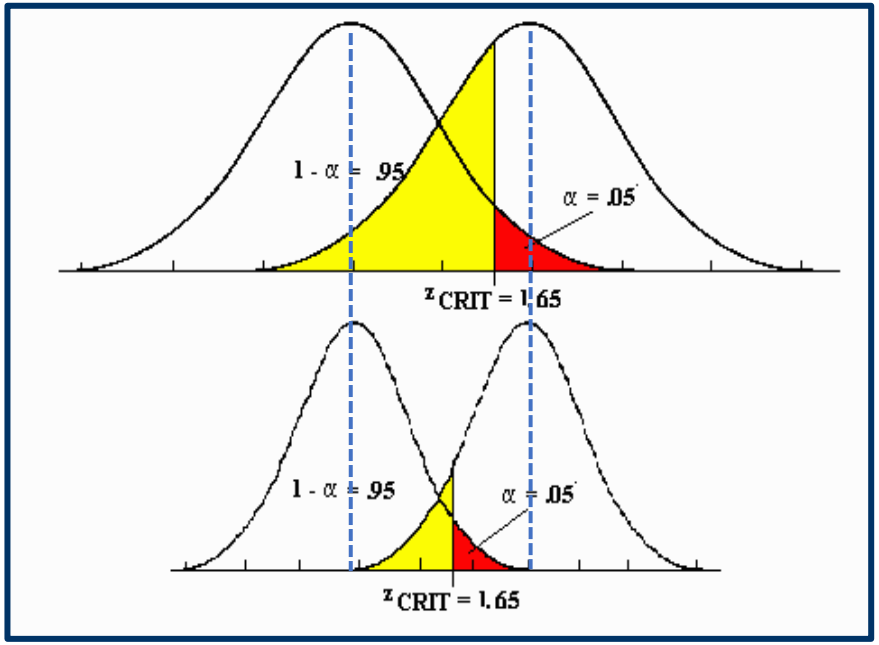
The relationship between α and β can be visualized under different conditions. If the critical value, e.g., based on a standardized normal distribution ($Z_{CRIT}$), is changed, in accordance with the adopted Type I error probability (α), the Type II error probability (β) increases at the decrease of α, for a specific value referred to the $H_1$ hypothesis:



As shown in the figure on the right, once $\alpha$ is fixed (for example, 0.05) $\beta$ is decreased if the value referred to the $H_1$ hypothesis is increased.




If values related to $H_0$ and $H_1$ hypotheses and $\alpha$ are fixed, $\beta$ will be decreased if the variance related to the statistic under test is decreased:



## Hypothesis testing for normally distributed populations: comparison between a variance and a known value

| Hypotheses | statistic type of test | rejection criteria for null hypothesis |
|---|---|---|
| $H_0 : \sigma^2 = \sigma_0^2$ | $T = (n-1)s^2/\sigma_0^2 \sim \chi^2_{n-1}$ | |
| $H_1 : \sigma^2 > \sigma_0^2$ | one tail | $t \geq \chi^2_{n-1\ (1-\alpha)}$ |
| $H_1 : \sigma^2 < \sigma_0^2$ | one tail | $t \leq \chi^2_{n-1(\alpha)}$ |
| $H_1 : \sigma^2 \neq \sigma_0^2$ | two tails | $t \leq \chi^2_{n-1\ (\alpha/2)}$ or $t \geq \chi^2_{n-1\ (1-\alpha/2)}$ |

## Hypothesis testing for normally distributed populations: comparison between two variances

| Hypotheses | statistic type of test | rejection criteria for null hypothesis |
|---|---|---|
| $H_0 : \sigma_1^2 = \sigma_2^2$ | $T = s^2_{num}/s^2_{den} \sim F_{\nu_{num},\nu_{den}}$ | |
| $H_1 : \sigma_1^2 > \sigma_2^2$ | one tail | $t \geq F_{\nu_{num},\nu_{den}\ (1-\alpha)}$ |
| $H_1 : \sigma_1^2 < \sigma_2^2$ | one tail | $t \geq F_{\nu_{num},\nu_{den}\ (1-\alpha)}$ |
| $H_1 : \sigma_1^2 \neq \sigma_2^2$ | two tails | $t \geq F_{\nu_{num},\nu_{den}\ (1-\alpha/2)}$ |

Note that $s^2_{num}$ and $s^2_{den}$ are chosen so that $s^2_{num} > s^2_{den}$

## Statistic equivalence testing

The challenge of assessing the comparability of different groups (or treatments or methods) is an issue facing many researchers and evaluators. Actually, the question of interest is often not whether two (or more) groups (or treatments or methods) are different from one another, but, rather, whether the groups (or treatments or methods) can be considered "practically the same" (i.e., equivalent).

Equivalence testing should assess whether mean differences between two groups are small enough that the groups can be considered equivalent (the differences found are considered practically unimportant).

The problem with using hypothesis testing methods is that a statistically non-significant value (failure to find a group difference – i.e., acceptance of $H_0$) is used to imply that the groups are comparable. However, a statistically non-significant finding only indicates that there is not enough evidence to support that two (or more) groups are statistically different.

It is possible that the two groups are effectively comparable, but it is also possible that:

- ✓ the study did not have enough power to detect a statistical difference
- ✓ there was high variability in the sample
- ✓ the study was poorly designed.

## Approaches to equivalence testing

**Equivalence testing** is recommended as a possible alternative to demonstrating comparability through the examination of whether mean differences between two groups are small enough that these differences can be considered practically unimportant and, thus, the groups can be treated as equivalent.

There are three general categories of equivalence tests:

1) the two one-sided t-tests (TOST) - procedure
2) the non-equivalence null hypothesis approach
3) Bayesian methods

The TOST procedure is the most popular, thanks to the ease of use and interpretation, and will be thus described in detail in the following slides.

## Two one-sided t-test (TOST) for equivalence testing

Designed specifically for bioequivalence testing of pharmaceutical products, i.e., to assess the expected *in vivo* biological equivalence of two proprietary preparations of a drug, **TOST** has recently been expanded into broader applications in pharmaceutical science, process engineering, psychology, medicine, chemistry, and environmental science.

The most important step in conducting equivalence testing is to give an operative definition of equivalence prior to statistical testing, e.g., by defining an acceptance criterion.

Briefly stated, this approach calculates a confidence interval around the mean difference between two groups. If this confidence interval is located within a specified range (the equivalence interval) then the groups are considered equivalent.

TOST is based on a null hypothesis stating that the two mean values are not equivalent, then tries to demonstrate that they are equivalent within a practical, preset limit.

This is conceptually opposite to the two-sample t-test procedure, included in hypothesis testing:

**Two-sample t-test**

$$\begin{cases} H_0 : \bar{y}_1 - \bar{y}_2 = 0 \\ H_1 : \bar{y}_1 - \bar{y}_2 \neq 0 \end{cases}$$

$$\begin{cases} H_0 : \bar{y}_1 = \bar{y}_2 \\ H_1 : \bar{y}_1 \neq \bar{y}_2 \end{cases}$$

**Equivalence test**

$$\begin{cases} H_0 : \bar{y}_1 - \bar{y}_2 \neq 0 \\ H_1 : \bar{y}_1 - \bar{y}_2 = 0 \end{cases}$$

$$\begin{cases} H_0 : \bar{y}_1 \neq \bar{y}_2 \\ H_1 : \bar{y}_1 = \bar{y}_2 \end{cases}$$

The acceptance criterion $\theta$ is the limit beyond which the difference in mean values should be considered practically and statistically significant, thus $\theta$ is used to construct the equivalence interval $[-\theta, \theta]$.

The two hypotheses can thus be written as:

$$\underbrace{H_0 : |\bar{y}_1 - \bar{y}_2| \geq \theta}_{\text{not } \textit{equivalent}} \quad \textit{versus} \quad \underbrace{H_1 : |\bar{y}_1 - \bar{y}_2| < \theta}_{\textit{equivalent}}$$

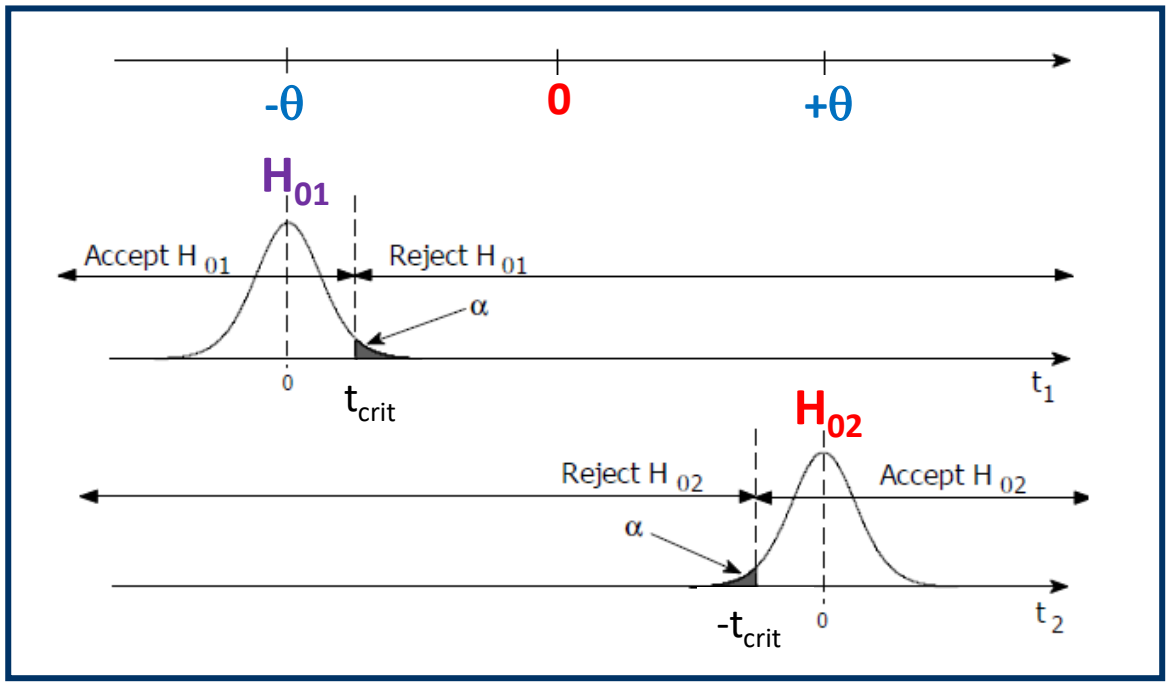
The absolute values in the original equivalence test hypotheses are broken up into two separate tests:

$$H_{01} : \bar{y}_1 - \bar{y}_2 \leq -\theta \quad \text{or} \quad H_{02} : \bar{y}_1 - \bar{y}_2 \geq \theta$$

**tested against the alternative hypothesis** $H_{11} : \bar{y}_1 - \bar{y}_2 > -\theta$

**tested against the alternative hypothesis** $H_{12} : \bar{y}_1 - \bar{y}_2 < \theta$

The two groups means are declared equivalent if, and only if, both $H_{01}$ and $H_{02}$ are rejected, in favor of hypotheses $H_{11}$ and $H_{12}$.




$$\begin{cases} H_{01}: \bar{y}_1 - \bar{y}_2 \le -\theta \\ H_{11}: \bar{y}_1 - \bar{y}_2 > -\theta \end{cases}$$

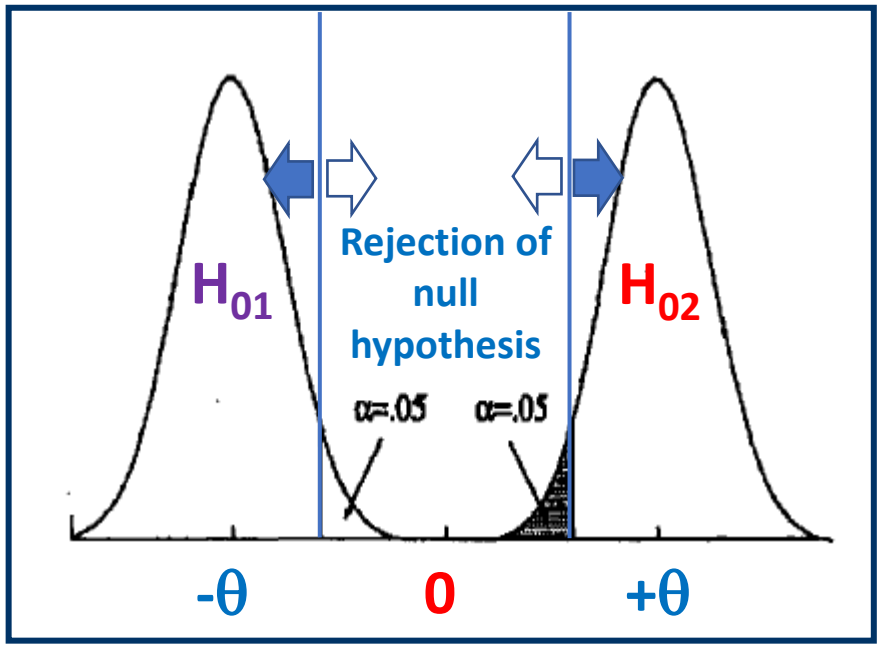
$$\begin{cases} H_{02}: \bar{y}_1 - \bar{y}_2 \ge \theta \\ H_{12}: \bar{y}_1 - \bar{y}_2 < \theta \end{cases}$$

When samples to which compared means are referred are extracted from gaussian populations having equal variances and their sizes are lower than 30 (like in case 3 of hypothesis tests for mean differences), the test statistics and the criteria for $H_{11}$ and $H_{12}$ acceptance are:

$$T_1 = \frac{(\bar{y}_1 - \bar{y}_2) + \theta}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{1-\alpha, n_1+n_2-2} \qquad T_2 = \frac{(\bar{y}_1 - \bar{y}_2) - \theta}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{1-\alpha, n_1+n_2-2}$$

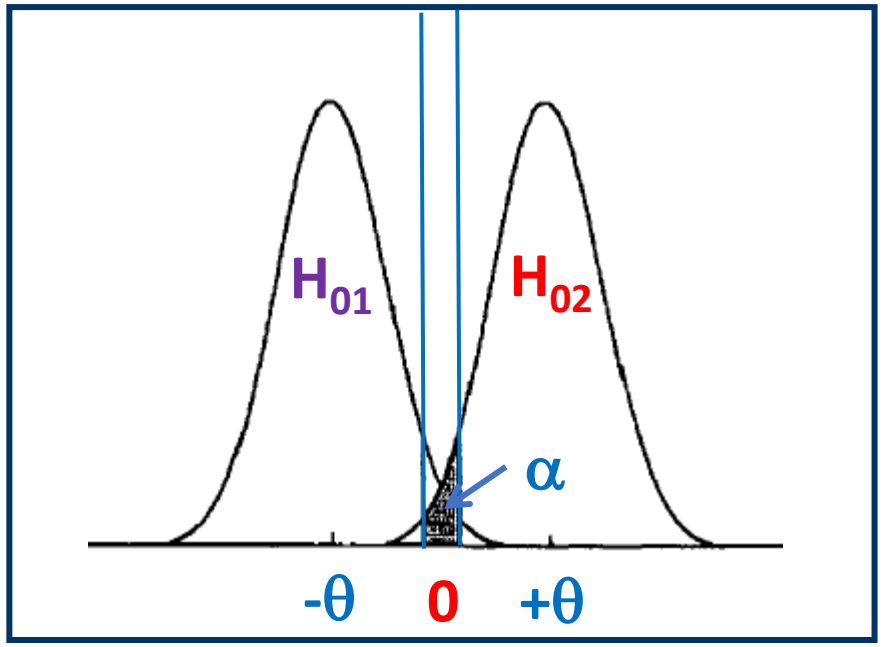
where $n_1$ and $n_2$ are the sizes of the two samples, $s_p$ is the weighed standard deviation and $t_{crit}$ is the 100 × (1-α)-th percentile of the t distribution with ($n_1$+$n_2$-2) degrees of freedom.

It is worth noting that $H_{01}$ and $H_{02}$ hypotheses cannot be both true, therefore Type I error rate will be determined by the critical region of only one of the curves representing $H_{01}$ and $H_{02}$:




If the standard deviation of the test statistic is too large and/or the width of the equivalence interval is too small, the two critical regions will overlap to a significant degree, thereby producing a conservative test.

A statistical test is conservative if, when constructed for a given nominal significance level, the true probability of incorrectly rejecting the null hypothesis is never greater than the nominal level.



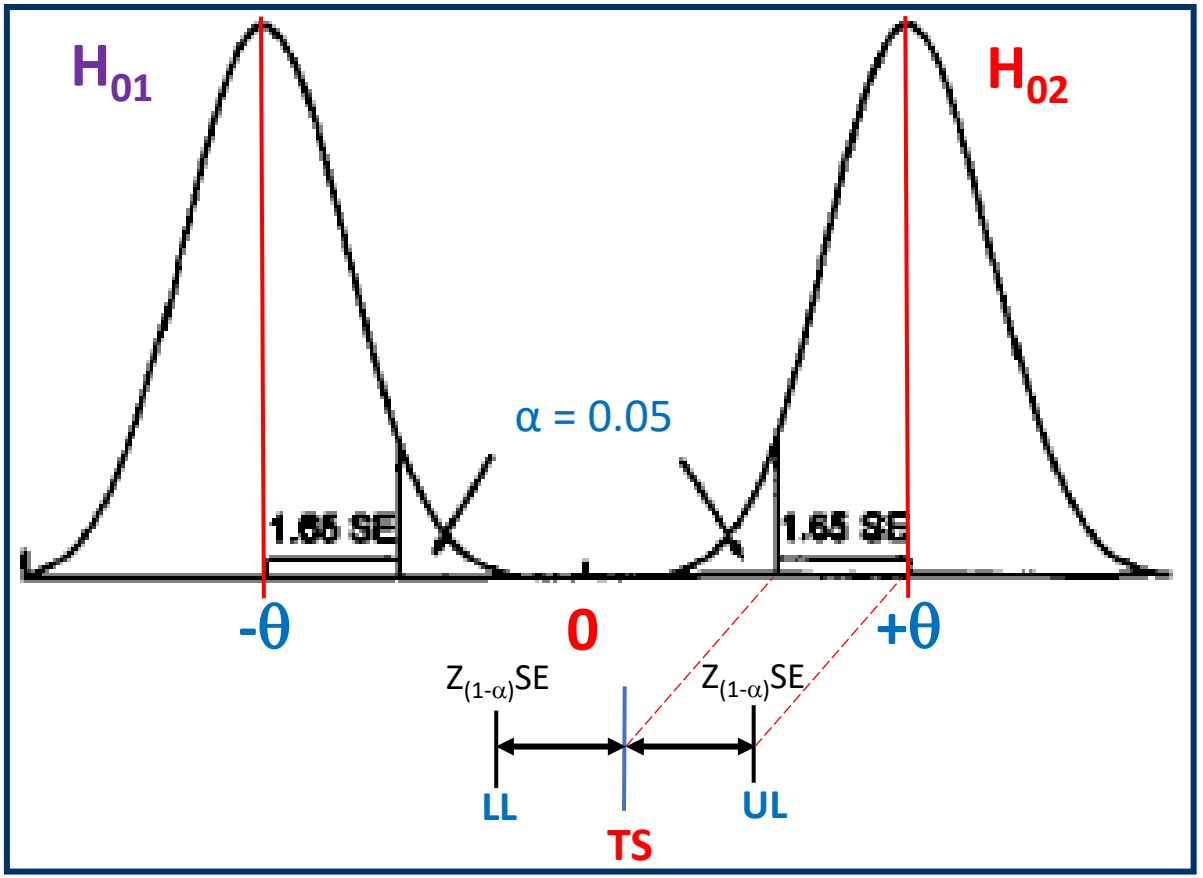
## Relationship between a confidence interval and a TOST

The graphical representation of a TOST with a [100 × (1-$\alpha$)]% confidence level can be compared with that of a [100 × (1-2$\alpha$)]% confidence interval centered on the value assumed by the test statistic TS, i.e., by the difference between the two sampling means under comparison.

For the sake of simplicity, let us consider the case in which a gaussian distribution with standard deviation SE can be used. Moreover, the value assumed by the TS will be reasonably different from 0.

For $2\alpha = 0.10$, $1-\alpha = 0.95$ and then $z_{(1-\alpha)} = 1.65$, thus 1.65 × SE is the half width of the 90% confidence interval, i.e., the distance between TS and the lower (LL) or the upper (UL) limits of the interval.

That value is also the distance between the hypothesized null value of the mean difference, $\theta$ or -$\theta$, and the beginning of the critical region:



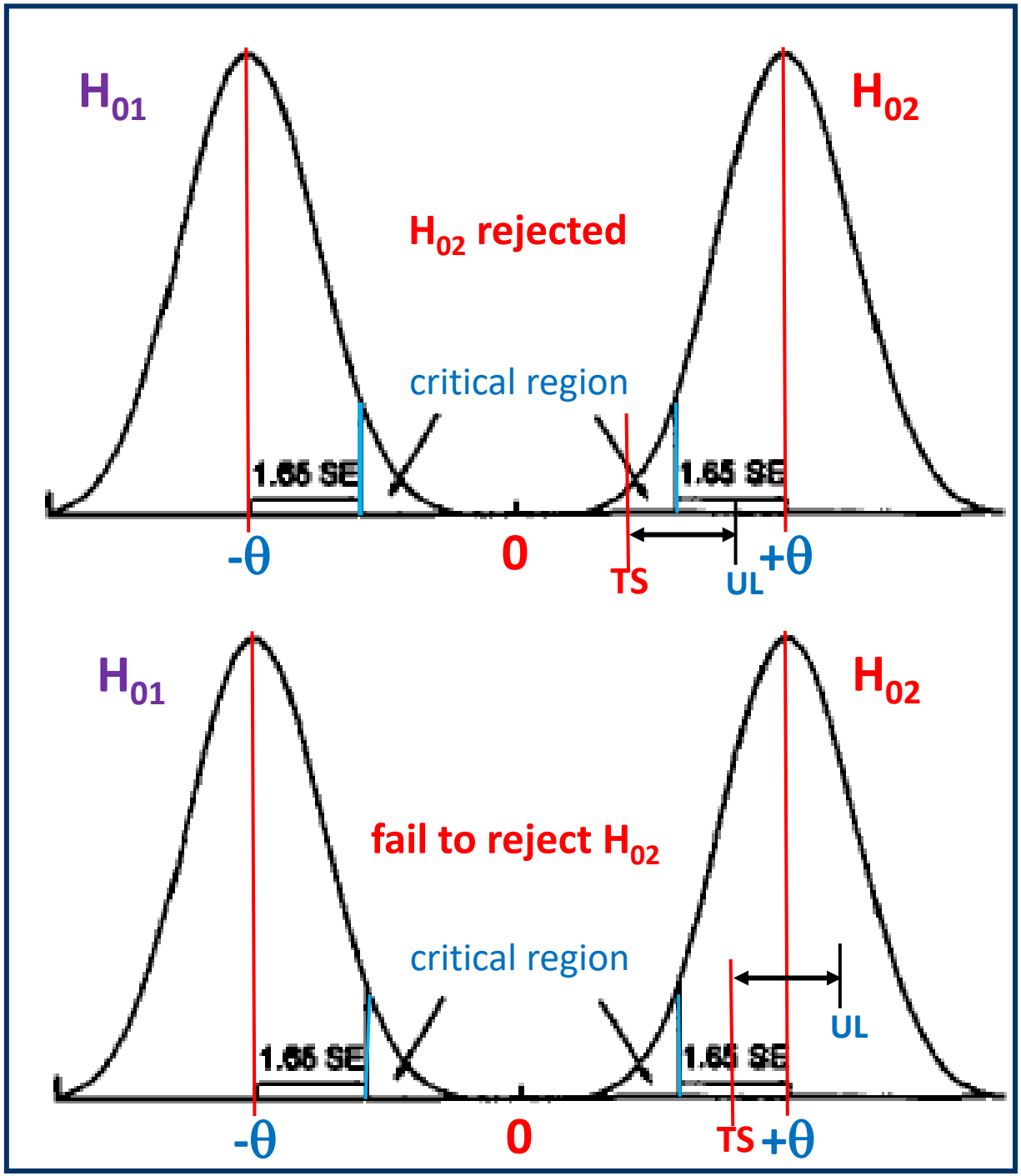
Different situations can be found:

1) TS is in the critical region for $H_{02}$, thus $H_{02}$ is rejected:

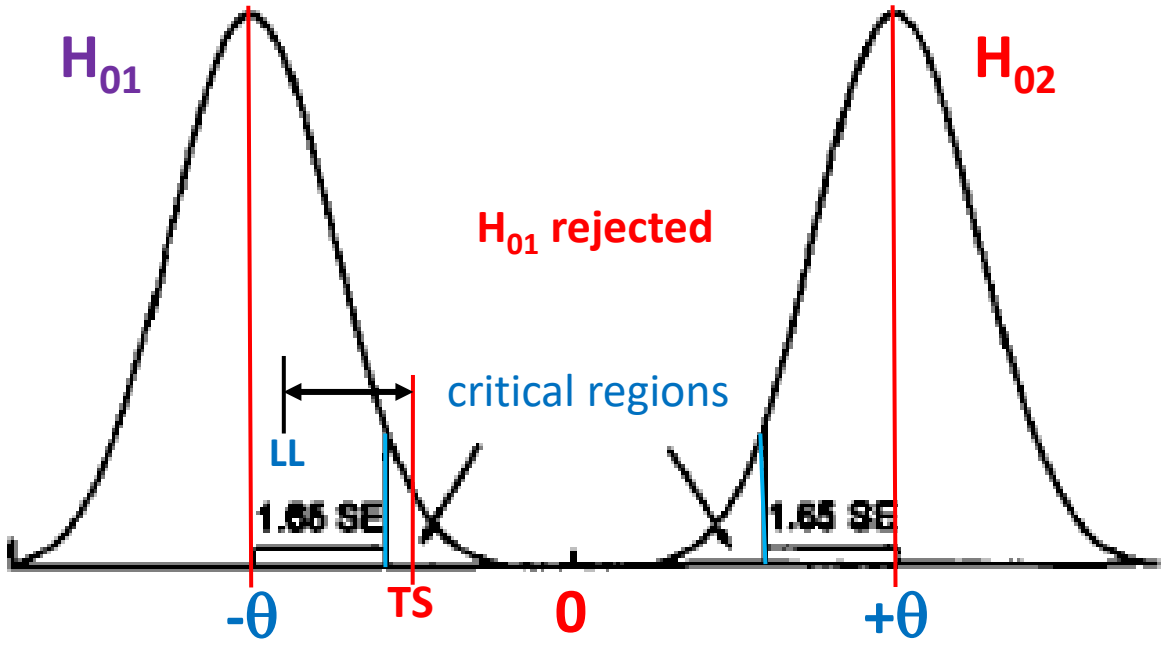
   In this case the UL of the confidence interval is lower than $\theta$.

2) TS is not in the critical region for $H_{02}$, thus $H_{02}$ is not rejected:

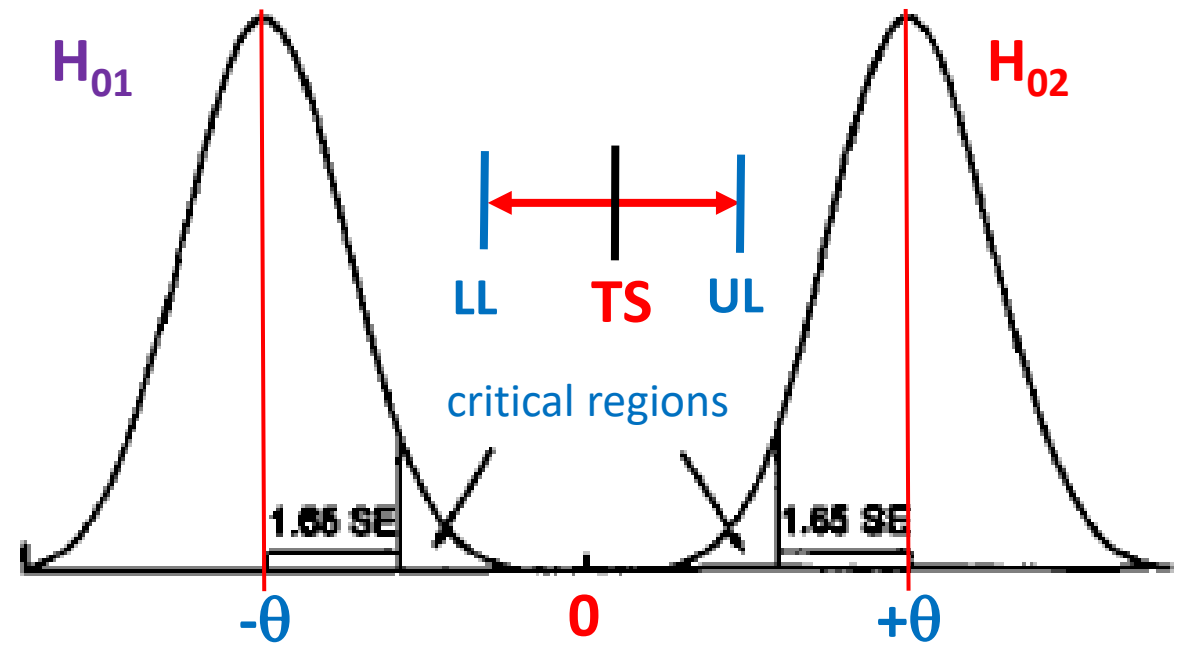
   In this case the UL of the confidence interval is higher than $\theta$.



The same consideration applies for the lover limit (LL) of the confidence interval when $H_{01}$ is tested:




A more direct assessment can thus be carried out by an inspection based on the entire 90% confidence interval on the difference between means (TS):



If concepts are summarized, it can be said that:

two one-sided tests (TOST) to establish significance at an $\alpha$ level can be conveniently performed also using a $100 \times (1-2\alpha)\%$ confidence interval centered on the test statistic (TS).

Indeed, if such an interval is completely contained within the $[-\theta, \theta]$ interval, the mean values of the two data sets are declared equivalent.

Note that the interval has to be calculated using the Student's t distribution, if cases 3 or 4 apply:

Case 3) $$(\bar{y}_1 - \bar{y}_2) \pm t_{1-\alpha, n_1+n_2-2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

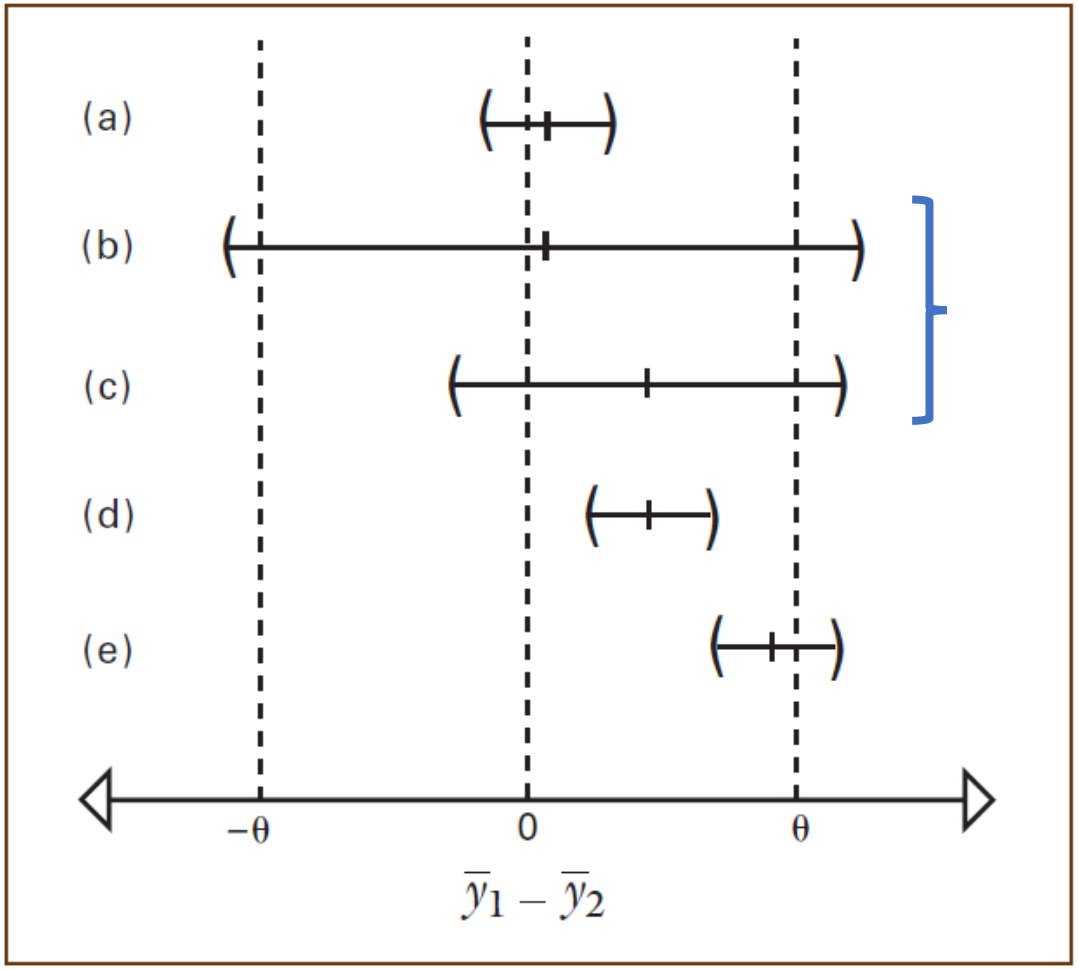
where $s_p^2$ is the variance obtained by a weighted average of the two sampling variances.

Case 4) $$(\bar{y}_1 - \bar{y}_2) \pm t_{1-\alpha, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\nu$ is obtained from the equation defined for the Fisher-Behrens problem.

## Significance vs equivalence test

The conclusions for each scenario with a ***t*-test** and **TOST**, respectively, would be:




**equal and equivalent**

**equal but not equivalent,**

**not equal but equivalent,**

**not equal and not equivalent.**

Consequently:

- ✓ acceptance of null hypothesis in a t-test does not necessarily imply equivalence
- ✓ rejection of null hypothesis in a t-test does not necessarily imply non equivalence.

## Two-sample *t-test* (or significance test) point of view

The width of the interval, which depends on the measurement precision, represents the range of plausible true differences in mean values of the two data sets.

If these intervals were created with the traditional two-sample t-test, in cases a, b and c the analyst would conclude that there is no difference between the mean values, because the confidence interval includes a difference of 0.
The confidence intervals in cases d and e do not include 0; therefore, the mean values would be declared different.

## TOST - Two one-sided *t-test* (or equivalence test) point of view

If confidence intervals were created with TOST, the mean values of the two data sets would be declared equivalent only in cases a and d, because those confidence intervals would be completely contained in the $[-\theta, \theta]$ interval.

The mean values in case d would be declared equivalent even though the confidence interval does not include 0, because the bias represented by the difference in means is small and within the interval $[-\theta, \theta]$.
The confidence intervals in cases b and c are too wide for the mean values of the data sets to be declared equivalent.

## Choice of an appropriate θ value for TOST

Choosing an appropriate value for θ in a TOST can be a challenge.
The following step-by-step process can be followed:

1) Choosing a value for the absolute value of the true difference, δ

The first parameter that must be specified before an analyst performs equivalence testing is δ, the absolute value of the true difference between the groups' mean values; δ is a hypothetical value, such that if the absolute value of the observed difference is not greater than δ there is a strong probability of concluding that the two data sets represent equivalent results.
The most conservative approach implies setting δ = 0.

2) Determining the n value needed for the test

The required value of n for TOST is related to θ and to other parameters, like Type I and Type II errors, i.e., α and β, respectively, the true difference δ and the upper confidence limit of method precision, s*.

Tables relating n and θ values for various combinations of the other parameters are available.

**Table 1. θ for various *n* and upper limit of method precision *s**.**

($\alpha = \beta = 0.05, \delta = 0$)

| *s** | θ for *n* = 5 | θ for *n* = 10 | θ for *n* = 12 | θ for *n* = 30 |
|---|---|---|---|---|
| 0.5 | 1.3 | 0.9 | 0.8 | 0.5 |
| 1.0 | 2.6 | 1.7 | 1.5 | 0.9 |
| 1.5 | 4.0 | 2.6 | 2.3 | 1.4 |
| 2.0 | 5.3 | 3.4 | 3.1 | 1.9 |
| 2.5 | 6.6 | 4.3 | 3.9 | 2.4 |
| 3.0 | 7.9 | 5.1 | 4.6 | 2.8 |

As expected, once s* is fixed, the decrease of n leads to an increase in θ, thus making it easier to prove equivalence, or, in other terms, making the proof of non-equivalence more difficult. This is typical of statistical testing.

On the other hand, once n is fixed, the increase in s* determines an increase in θ. This effect is due to the fact that if the precision is low, small differences between the two means under comparison may be not significant.

## 3) Finding an estimate for s*

Once the sample size n is chosen, an estimate of the precision of a measurement, expressed as standard deviation s, is made through replicated analyses, usually performed by a single analyst or a single laboratory.

Although this approach can lead to underestimate precision, it is a pragmatic compromise between appropriate statistical application and real-world constraints on resources.

To ensure a better representation of the true measurement precision, it is recommended that an upper confidence limit (e.g., the upper limit from a one-sided 80% confidence interval), indicated as s*, be used as an estimate of measurement precision.

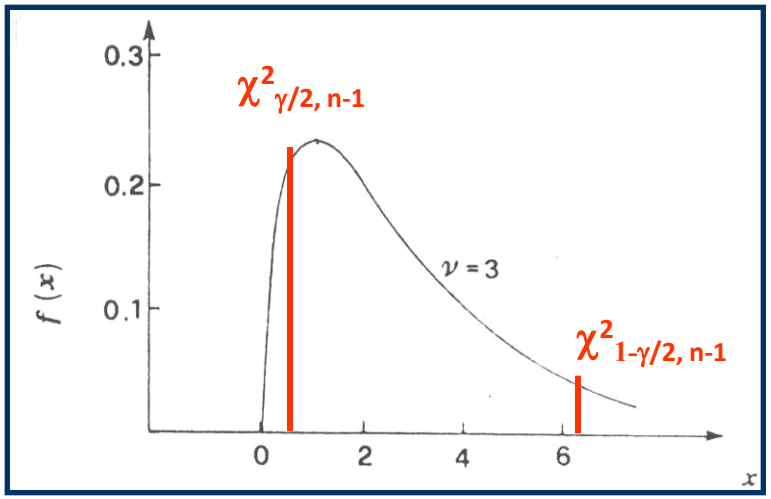
It can be demonstrated that the upper 100(1-$\gamma$)% confidence limit s* for s can be calculated as:

$$s^* = s\sqrt{\frac{n-1}{\chi^2_{(\gamma,\,n-1)}}}$$

where $\chi^2_{(\gamma,\,n-1)}$ represents the $(100\,\gamma)^{th}$ percentile of a $\chi^2$ distribution with n-1 degrees of freedom.

Indeed, let us consider that, for a random variable distributed according to a Gaussian distribution with variance $\sigma^2$, the following relationship is true:

$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$$




It thus follows that, for a two-sided confidence interval with a significance level $\gamma$:

$$\chi^2_{\gamma/2,\,n-1} \leq (n-1)s^2/\sigma^2 \leq \chi^2_{1-\gamma/2,\,n-1}$$

consequently:

$$\sigma^2\,\chi^2_{\gamma/2,\,n-1} \leq (n-1)s^2 \leq \sigma^2\,\chi^2_{1-\gamma/2,\,n-1}$$

and thus:

$$s\,(n-1)^{1/2} / [\chi^2_{1-\gamma/2,\,n-1}]^{1/2} \leq \sigma \leq s\,(n-1)^{1/2} / [\chi^2_{\gamma/2,\,n-1}]^{1/2}$$

The upper limit for a one-sided confidence interval with a significance level $\gamma$, i.e., a confidence level $1-\gamma$, is thus given by:

$$s\,(n-1)^{1/2} / [\chi^2_{\gamma,\,n-1}]^{1/2}$$

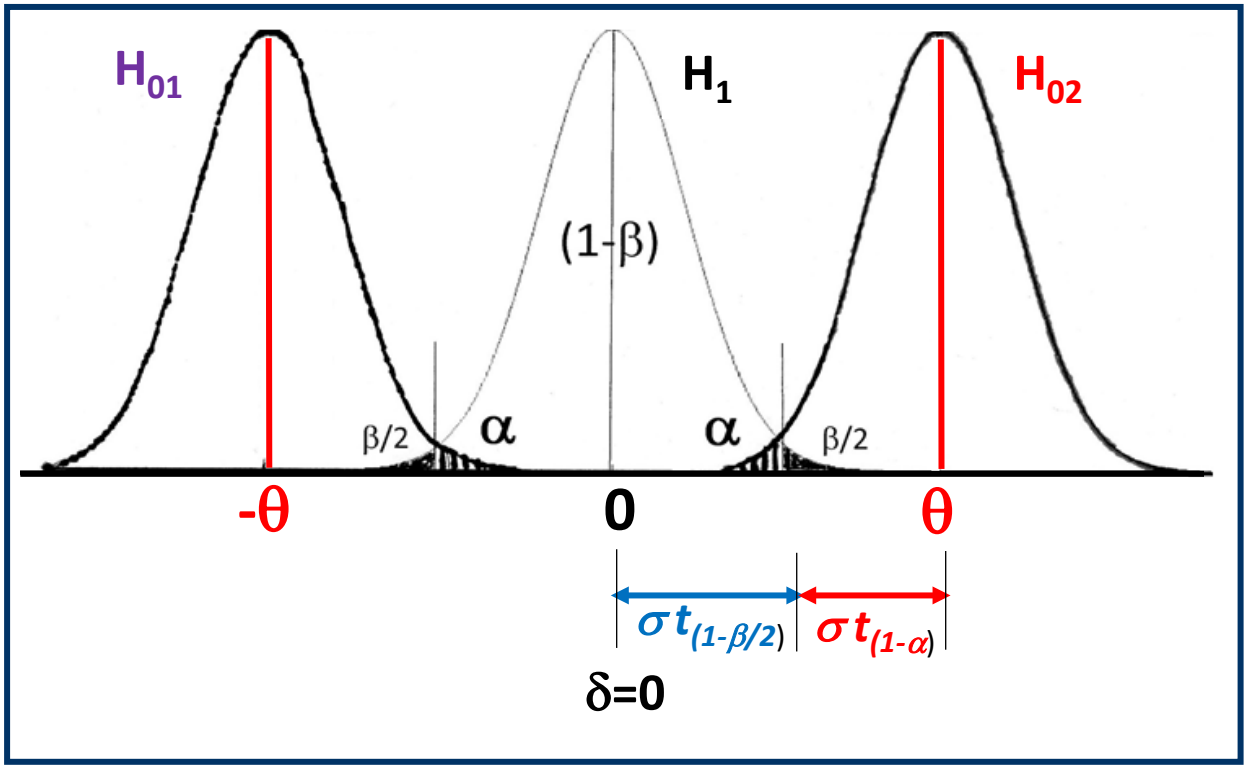
## 4) Refining the θ value by considering also Type II error (β)

In previous graphical representations of the equivalence test the distribution related to the $H_1$ hypothesis

$$H_1 : |\bar{y}_1 - \bar{y}_2| < \theta$$

was not reported for simplicity and only type I error (α) was evidenced.

The following representation is required to show both types of error, including β:



Since equal variances are related to the three distributions shown in the figure, the population standard deviation, σ, can be estimated using the expression (case 3):

$$s_p \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

The following general expression can be thus obtained for θ:

$$\theta = \delta + \left[ t_{(1-\beta/2),\, n_1+n_2-2} + t_{(1-\alpha),\, n_1+n_2-2} \right] s_p \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

This equation can be slightly modified when $n_1 = n_2 = n > 30$, since the Student's t distribution values can be replaced by those obtained from the Gaussian distribution; moreover, s* can be used instead of s, thus the expression for θ becomes:

$$\theta = \delta + s^* \left[ z_{(1-\beta/2)} + z_{(1-\alpha)} \right] \sqrt{\frac{2}{n}}$$

Interestingly, an equation providing a value for n, once s* is known, can be obtained starting from the last equation:

$$\theta = \delta + s^* \left[ z_{(1-\beta/2)} + z_{(1-\alpha)} \right] \sqrt{\frac{2}{n}}$$

$$(\theta - \delta)^2 = s^{*2} \left[ z_{(1-\beta/2)} + z_{(1-\alpha)} \right]^2 \frac{2}{n}$$

$$n = \frac{2 s^{*2} \left[ z_{(1-\beta/2)} + z_{(1-\alpha)} \right]^2}{(\theta - \delta)^2}$$

## Example of sample size calculation

Two inhalers used for the relief of asthma attacks have to be assessed for equivalence. They will be considered equivalent if the 95% two-sided confidence interval for the treatment difference, based on morning peak expiratory rate (L/min), falls entirely within the interval $\pm 15$ L/min.

Under these conditions, $\theta - \delta = 15$ (considering $\delta = 0$, as usual) and $\alpha = (1-0.95)/2 = 0.025$.

From a previous trial $s^{*2}$ was estimated to be equal to 1600 $(L/min)^2$.

Supposing that a power of 0.8 has to be obtained, thus $(1-\beta) = 0.8$ and then $\beta/2 = 0.1$, the following calculations can be made:

$z_{(1-\alpha)} = z_{(1-0.025)} = z_{0.975} = 1.96$

$z_{(1-\beta/2)} = z_{0.90} = 1.28$

The size of each group of patients for the evaluation of the two inhalers should be:

$$n = \frac{2s^{*2}\left[z_{(1-\beta/2)} + z_{(1-\alpha)}\right]^2}{(\theta-\delta)^2} = \frac{2 \times 1600}{15^2}[1.96 + 1.28]^2 = 149.3 \cong 149$$

# Comparison between t-test and equivalence test: an example

In Table 2 data obtained by a method development laboratory and by a manufacturing quality control (QC) laboratory for a dissolution test performed on 12 drug tablets, expressed as percentage of content indicated on the drug label, are reported:

Starting from a 0.05 value for $\alpha$ and $\beta$, a 0 value for $\delta$, n = 12, and an s value of 1.9%, which is then transformed into s*, $\theta$ is calculated to be equal to 3.7%.

Since the difference between the two means is: 89.3 - 87.7 = 1.6%, and the 90% confidence interval goes from 0.5 to 2.7%, the null hypothesis, stating that they are not equivalent, is rejected, thus the two laboratory methods are declared equivalent.

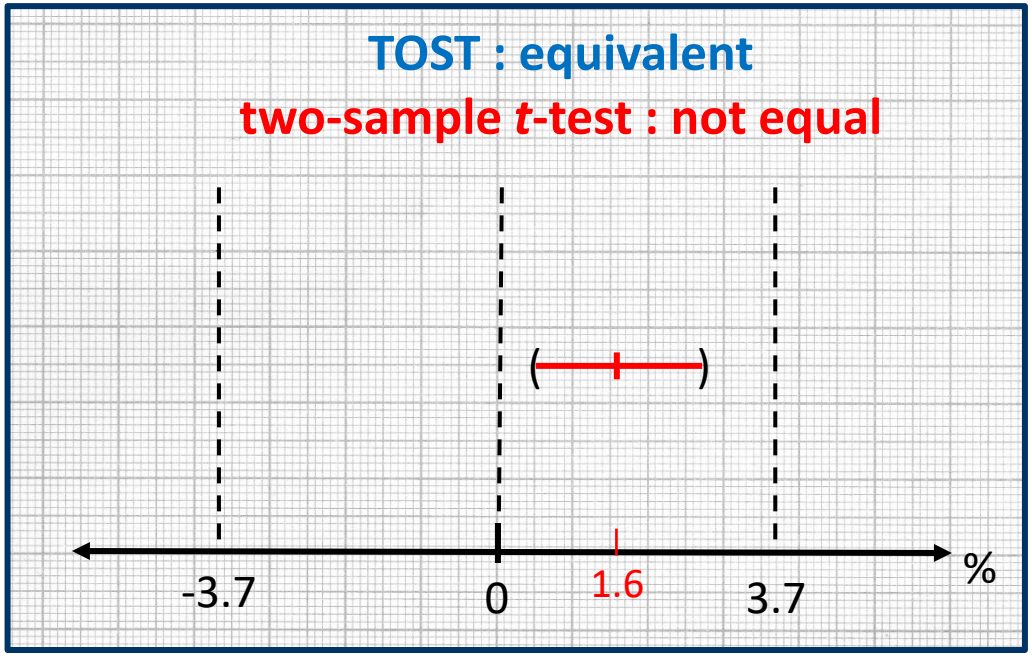
**Table 2. Statistical significance vs practical relevance in the transfer of a dissolution method.**

($\theta$ calculated on the basis of the development laboratory *s* of 1.9%, with $\alpha = \beta = 0.05$, $\delta = 0$, $n = 12$)

| | % Label strength dissolved | |
|---|---|---|
| Tablet number | Development lab | Manufacturing QC lab |
| 1 | 90.8 | 86.2 |
| 2 | 88.0 | 87.4 |
| 3 | 90.5 | 88.2 |
| 4 | 90.0 | 89.7 |
| 5 | 91.0 | 87.3 |
| 6 | 86.0 | 87.6 |
| 7 | 88.3 | 88.0 |
| 8 | 89.3 | 86.5 |
| 9 | 88.9 | 89.6 |
| 10 | 91.1 | 89.1 |
| 11 | 86.2 | 86.1 |
| 12 | 91.3 | 86.2 |
| $\bar{y}$ | 89.3 | 87.7 |
| *s* | 1.9 | 1.3 |
| % RSD | 2.1 | 1.5 |
| TOST: 90% confidence interval for difference ($\theta = 3.7\%$) | (0.5, 2.7) | |
| Two-sample *t*-test p-value | 0.02 | |

As shown in the graph on the right, the 90% confidence interval clearly does not include 0, thus the classical two-sample t-test indicates that a significant difference exist between the two methods.

Notably, this outcome is indicated by the p value, 0.02, reported in the previous table. In fact, the p-value is the probability of observing a T value more extreme than the one that would be observed if the means were not statistically significant. Consequently, if p is lower than the significance level, a significant difference between the means is inferred.




The example highlights a key advantage of TOST over a two- sample *t*-test for showing equivalence: TOST allows small, scientifically irrelevant differences to exist without leading to the conclusion that the laboratory means are not equivalent.

# Consequences of poor precision

Table 3 is an example of a tablet dissolution method that was transferred from a development laboratory to a contract laboratory during the early stages of product development.
In this study, n = 6 for each laboratory because of limited sample availability.

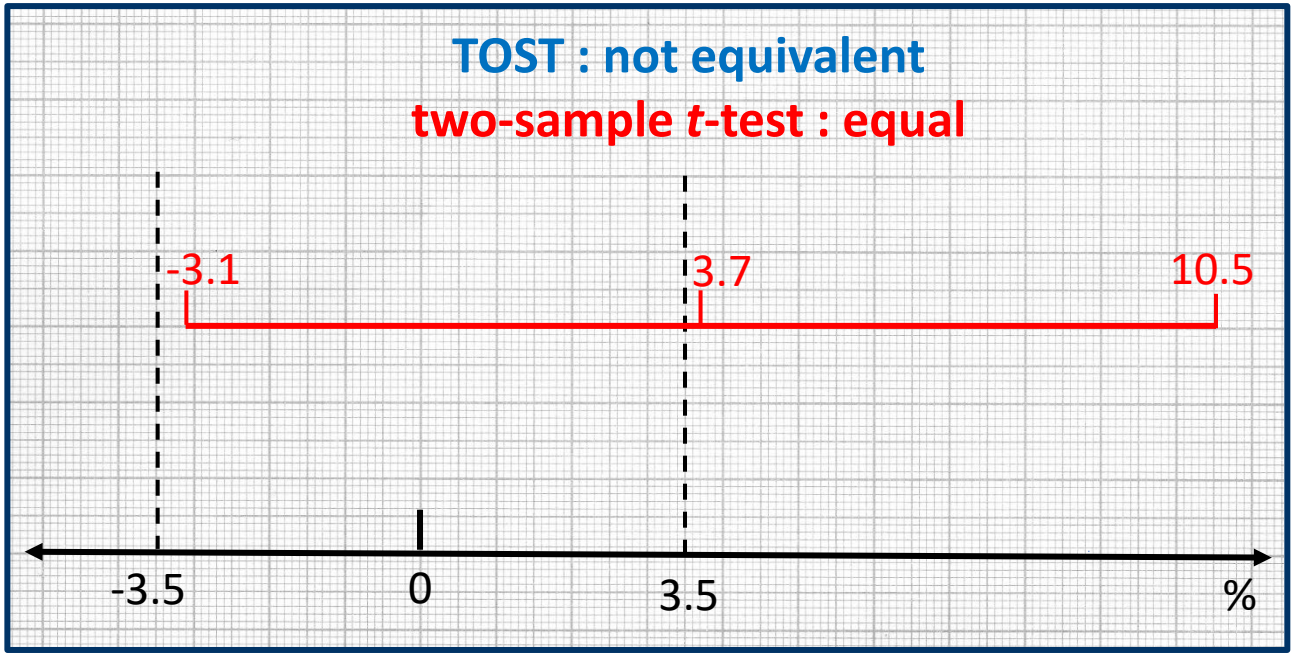
With an initial *s** estimate of 1.5%, arising from previous analyses, θ was calculated as 3.5%, which would generally be considered acceptable for a method of this type.

However, the actual *s* from each laboratory was much larger than the initial estimate of 1.5%. This was determined to be the result of poor sample homogeneity caused by degradation during storage.

**Table 3. Comparison of the effect of poor precision on the outcome of TOST vs a two-sample *t*-test.**
(θ calculated on the basis of the development laboratory *s** from previous experiments of 1.5%, with $\alpha = \beta = 0.05$, $\delta = 0$, $n = 6$)

| | % Label strength dissolved | |
|---|---|---|
| Tablet number | Development lab | Contract lab |
| 1 | 82 | 74 |
| 2 | 92 | 70 |
| 3 | 78 | 84 |
| 4 | 85 | 76 |
| 5 | 77 | 90 |
| 6 | 79 | 77 |
| $\bar{y}$ | 82 (82.2) | 79 (78.5) |
| *s* | 5.6 | 7.3 |
| % RSD | 6.9 | 9.2 |
| Difference between $\bar{y}$ values | 3.7 | |
| TOST: 90% confidence interval for difference (θ = 3.5%) | (–3.1, 10.5) | |
| Two-sample *t*-test p-value | 0.35 | |

When data are compared by an equivalence test with $\theta$ = 3.5%, there is not enough evidence to declare the laboratories' methods equivalent.




Conversely, if the laboratory mean values are compared with a two-sample t-test, the resulting confidence interval includes 0, thus data do not provide enough evidence to conclude that the laboratories' methods are different (the p-value of 0.35 is another way to express the same result).

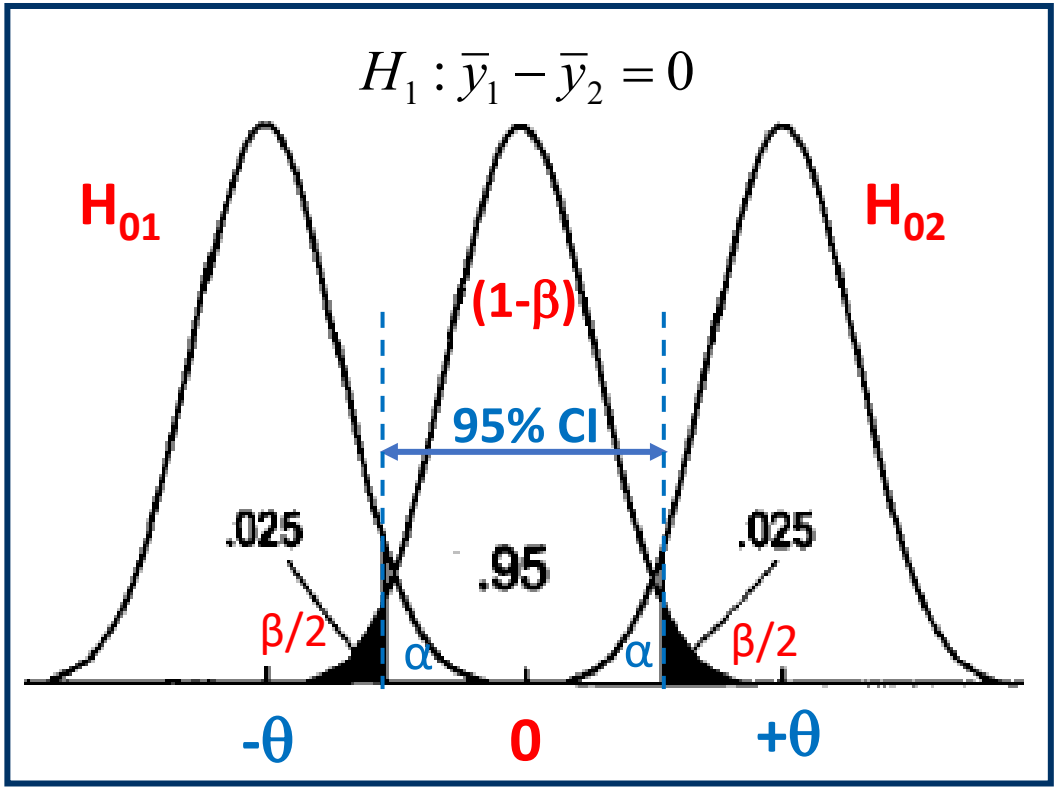
In this case, the traditional two-sample t-test does not reject the hypothesis that the data sets are equal, because s values for the two datasets are too large with respect to the difference between the mean values of y.

This example highlights another key advantage of TOST over a two-sample t-test: TOST appropriately penalizes the analyst if the observed variance is too large.

## Considerations on test power

For the present situation, power is defined as the probability of attaining significance when $\bar{y}_1 - \bar{y}_2$ is contained in the interval [-θ, θ].

Equivalently, power may be defined as the probability that a properly constructed confidence interval will be completely contained in the interval [-θ, θ] when $\bar{y}_1 - \bar{y}_2$ is in the interval [-θ, θ].




Power calculations are usually carried out under the assumption that $\bar{y}_1 - \bar{y}_2 = 0$ although other values may be chosen.

The unshaded area of the central distribution in the figure depicts a power of 0.95 for a two-sided equivalence test. Notice that the probability of failing to obtain a significant result in this situation is the shaded portion in the tails of the central distribution. Because power is 0.95, the total shaded area is 0.05 with 0.025 allocated to each tail.

# Example of calculation of test power for a TOST as a function of sample size

Two parallel groups of patients are enrolled to compare the effect of two drugs on diastolic (minimum) blood pressure.
The diastolic blood pressure is known to be close to 96 mmHg with the reference drug and is thought to be 92 mmHg with the experimental drug. Based on similar studies, the within-group standard deviation is set to 18 mmHg.

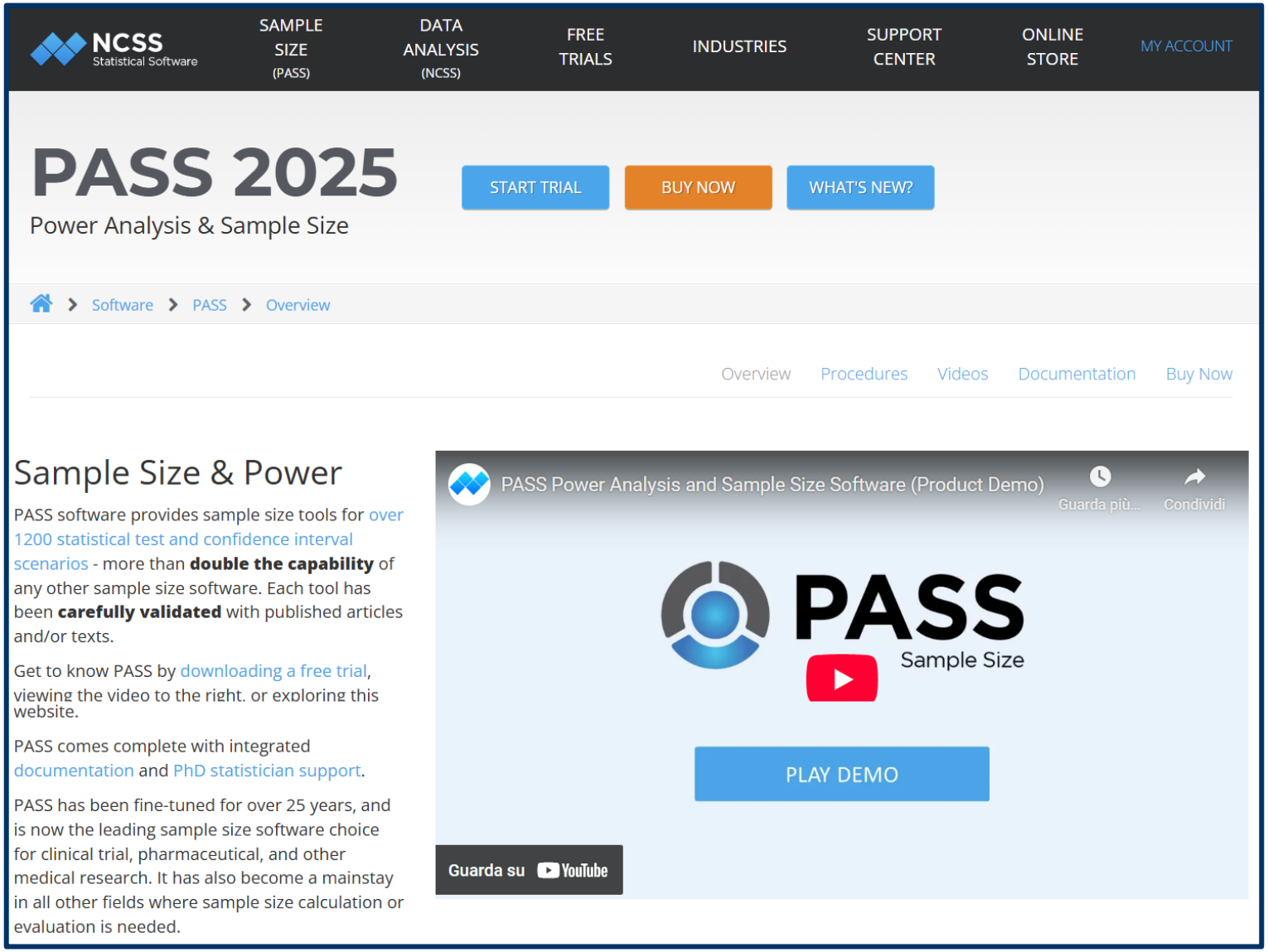
Following the United States Food and Drug Administration (FDA) guidelines, the researchers want to show that the diastolic blood pressure with the experimental drug is within 20% of the diastolic blood pressure with the reference drug. Note that 20% of 96 is 19.2.
They decide to calculate the test power for a range of sample sizes between 3 and 60. The significance level is $\alpha = 0.05$.

The following conditions are thus posed:

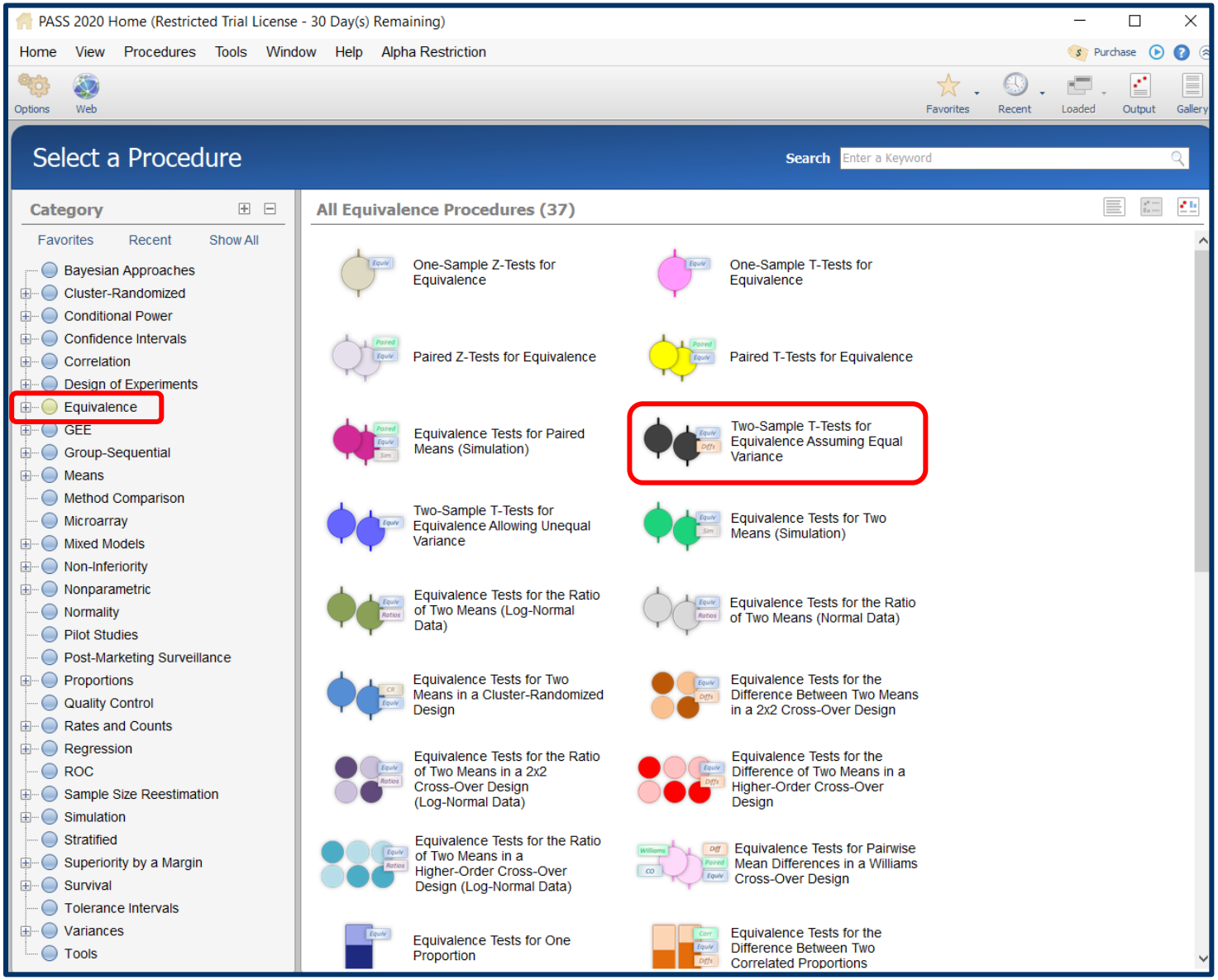
| | |
|---|---|
| $\alpha$..... | 0.05 |
| Group Allocation ..... | Equal ($n_1 = n_2$) |
| Sample Size Per Group..... | 3, 5, 8, 10, 15, 20, 30, 40, 50, 60 |
| Lower, Upper Equivalence Limit $(-\theta, \theta)$.... | -19.2, 19.2 |
| $\delta$ (True Difference)..... | (92-96) = -4 mmHg |
| s (Standard Deviation) ..... | 18 mmHg |

Power calculation can be performed using the PASS 2025 software, available on the Internet (for a 30-days free trial), at the following address:

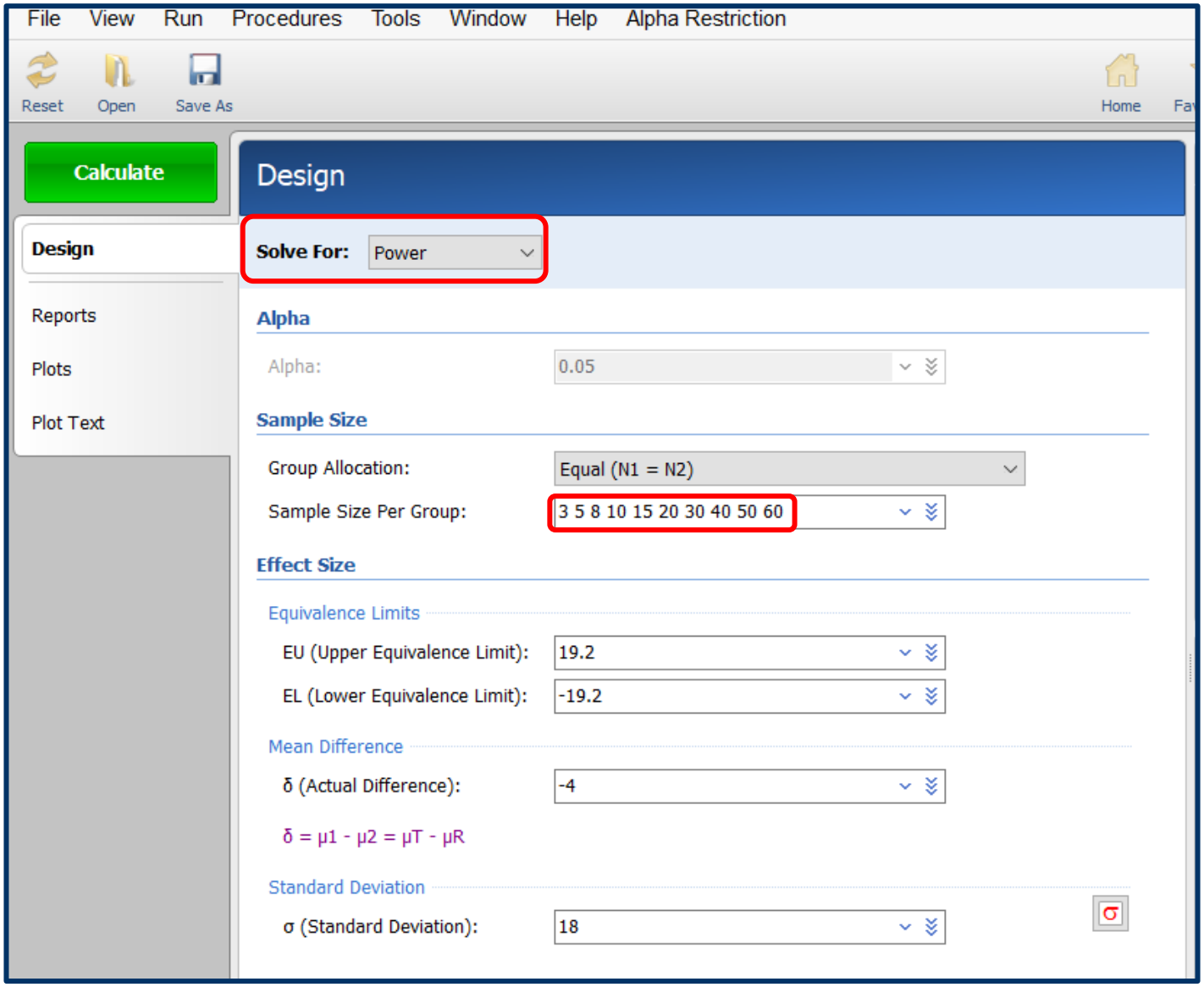
https://www.ncss.com/software/pass/



As far as equivalence is concerned, the PASS software provides 37 different procedures, including Two-Sample T-Tests for Equivalence Assuming Equal Variance:



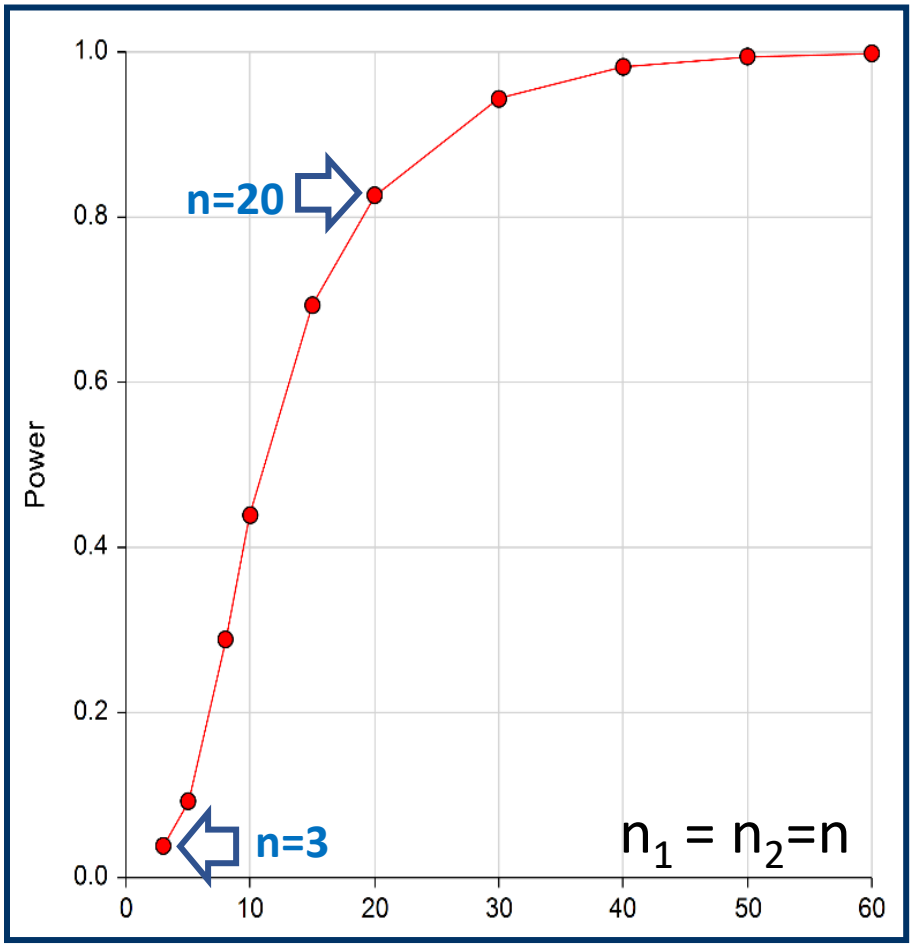
When accessing the corresponding page, all parameters required for the equivalence test can be set. In the following case, once values for $\alpha$, $\theta$, $\delta$ and $\sigma$ are specified, along with the condition N1 = N2, the test power can be calculated for a range of sample sizes:



A graph reporting Power vs sample size is also generated by the program:

As expected, the test power is increased at the increase of the sample size, with a value of 0.8 achieved already for n = 20.

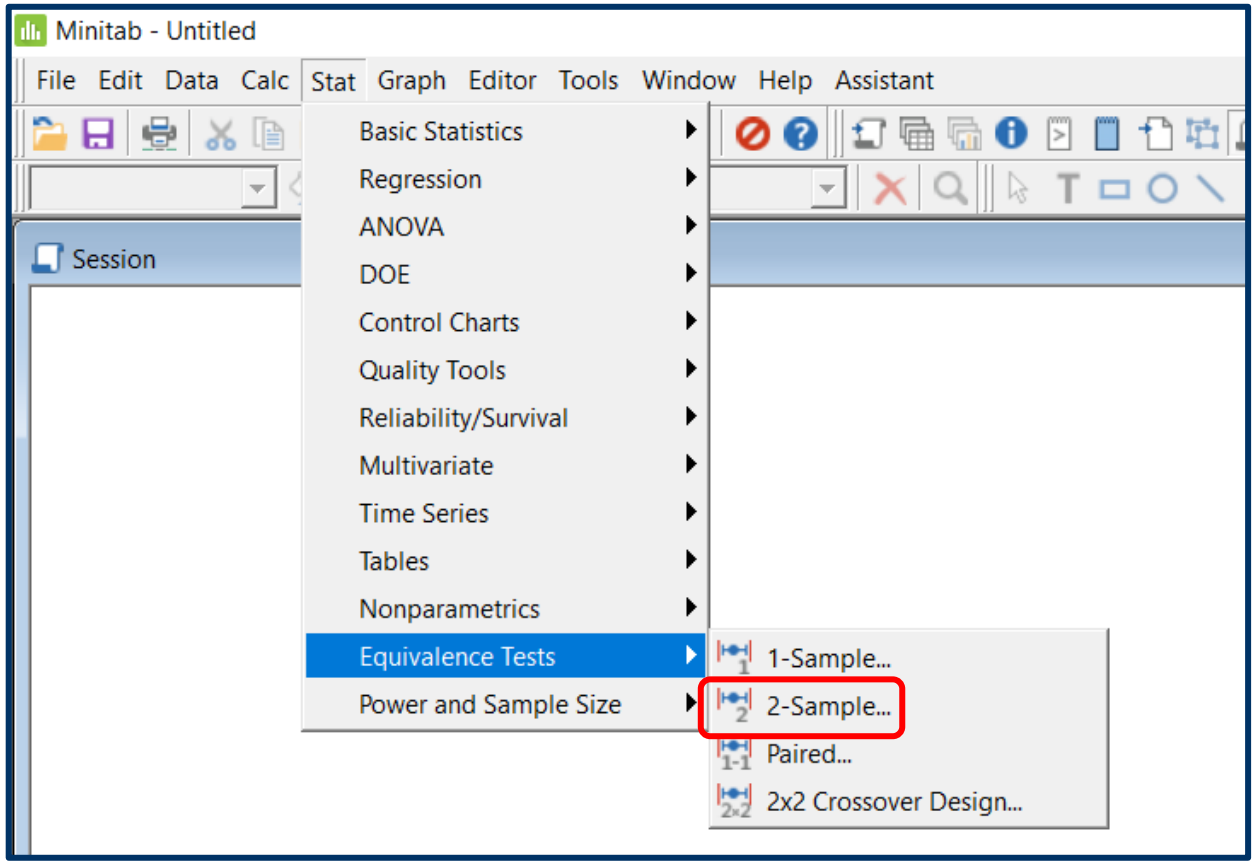
As apparent, further increases of the sample size lead to smaller increases in the test power, that asymptotically tends towards unity, thus an enlargement of the sample set is not very useful.



# Equivalence test performed using the Minitab® 18 software

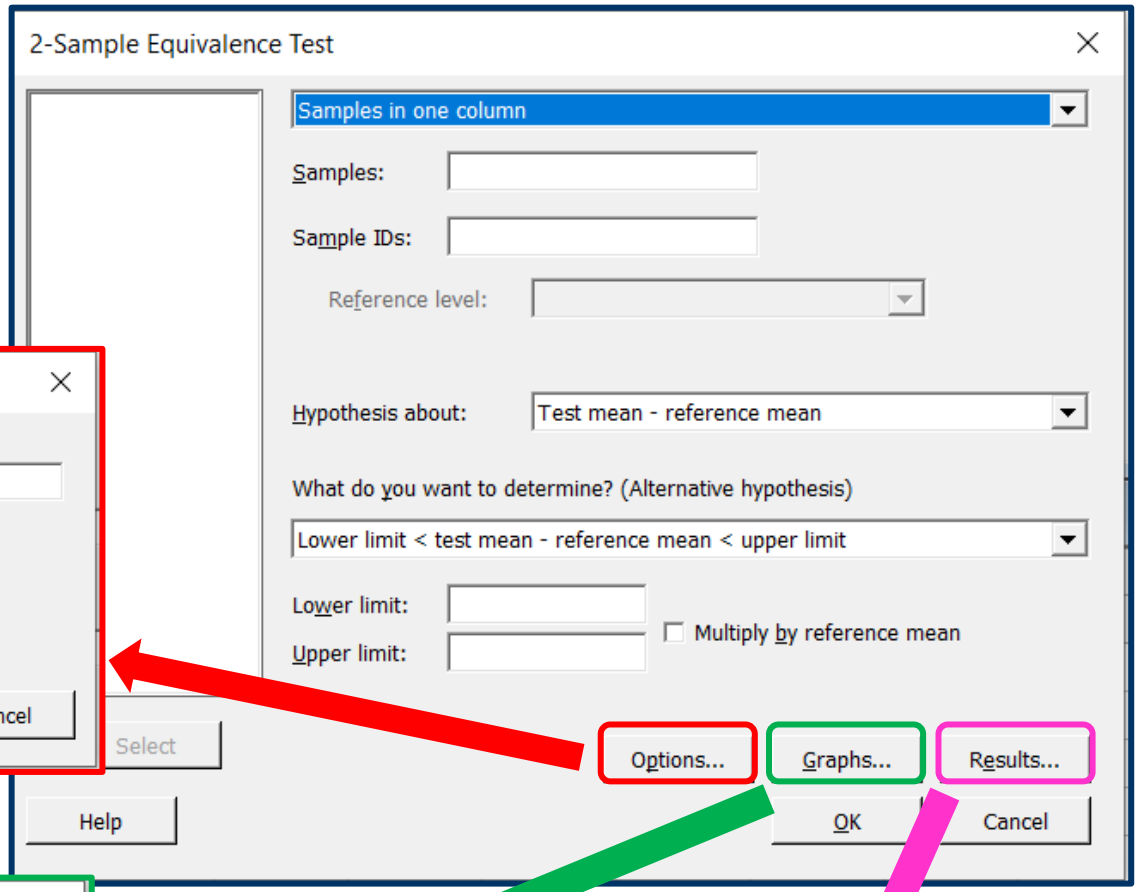
First step

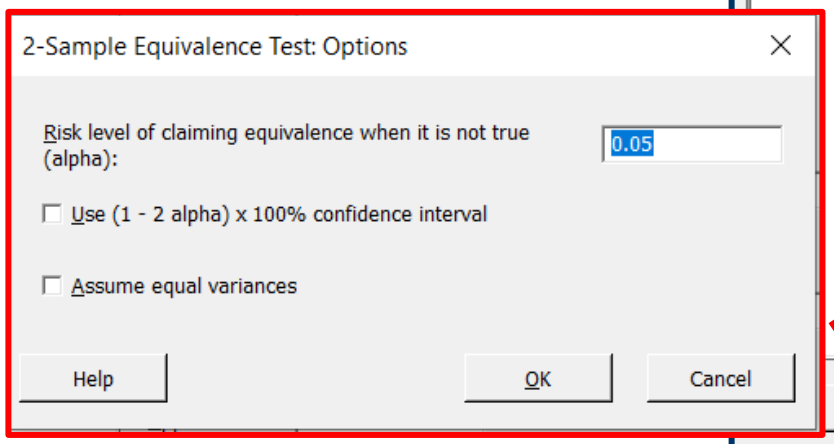
select **Equivalence Tests** in the **Stat menu** and then choose the **2-Sample…** option

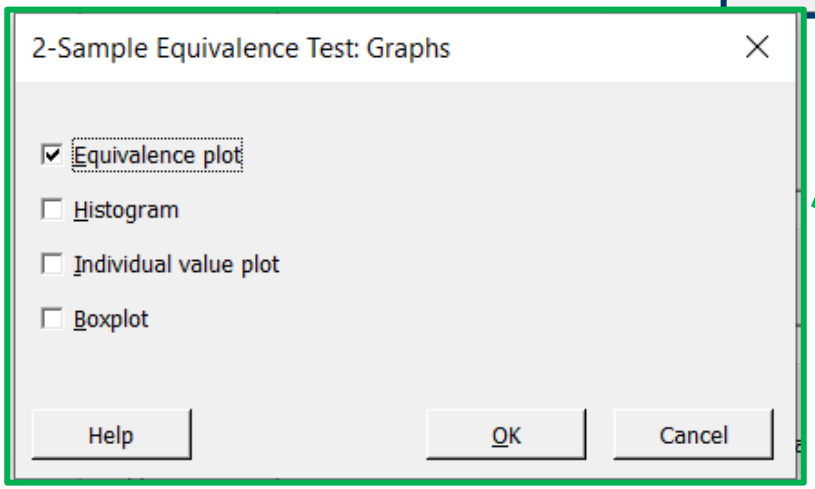


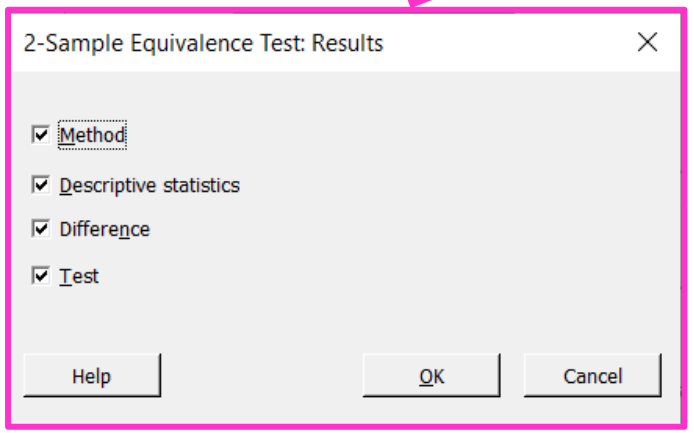
## Second step

Evaluation of parameters and choices in the **2-Sample Equivalence Test** window:



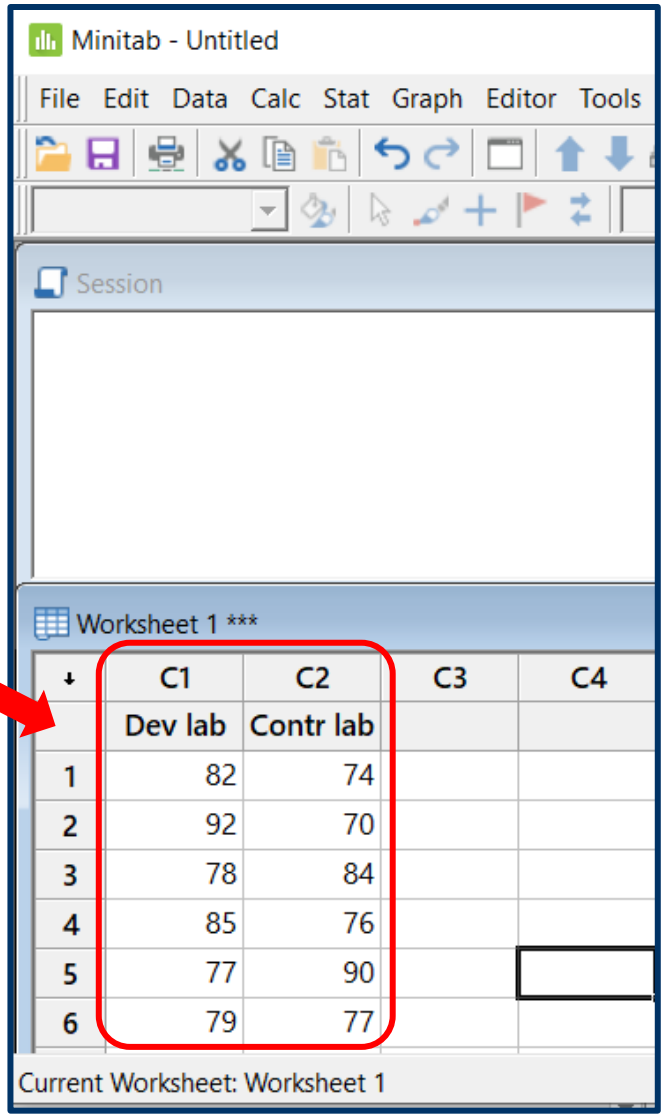
## Third step

Data (in this example taken from Table 2 shown before) are introduced into the Minitab 18 Worksheet, with each set corresponding to a specific column:

**Table 3. Comparison of the effect of poor precision on the outcome of TOST vs a two-sample *t*-test.**

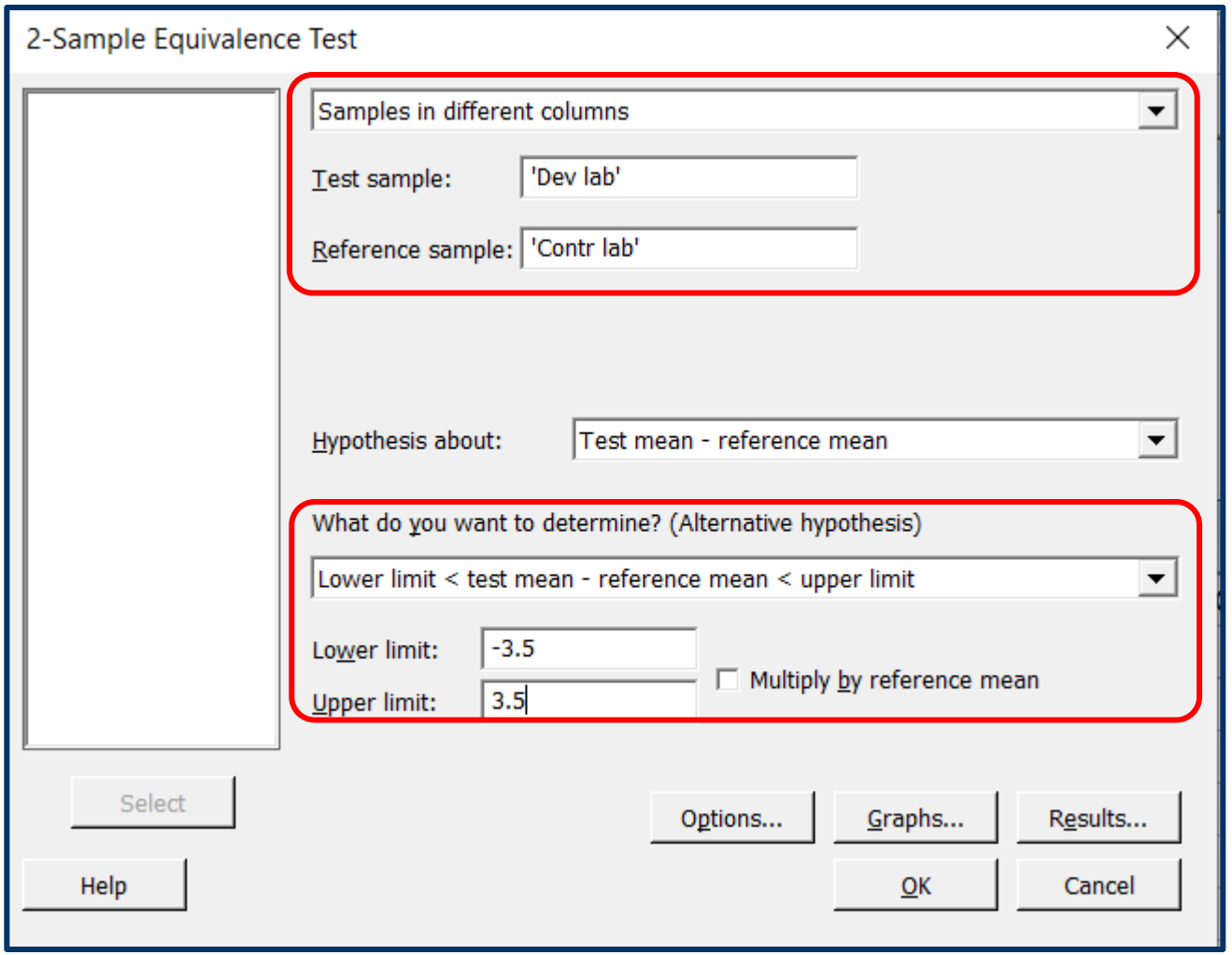
(θ calculated on the basis of the development laboratory *s** from previous experiments of 1.5%, with $\alpha = \beta = 0.05$, $\delta = 0$, $n = 6$)

| Tablet number | % Label strength dissolved Development lab | Contract lab |
|---|---|---|
| 1 | 82 | 74 |
| 2 | 92 | 70 |
| 3 | 78 | 84 |
| 4 | 85 | 76 |
| 5 | 77 | 90 |
| 6 | 79 | 77 |
| $\bar{y}$ | 82.2 | 78.5 |
| *s* | 5.6 | 7.3 |
| % RSD | 6.9 | 9.2 |
| Difference between $\bar{y}$ values | 3.7 | |
| TOST: 90% confidence interval for difference (θ = 3.5%) | (–3.1, 10.5) | |
| Two-sample *t*-test p-value | 0.35 | |



## Fourth step

Selection of samples columns in the Worksheet and setting of lower and upper limits, i.e. of –θ and + θ values, which were calculated to be equal to -3.5 and 3.5 in the specific case:

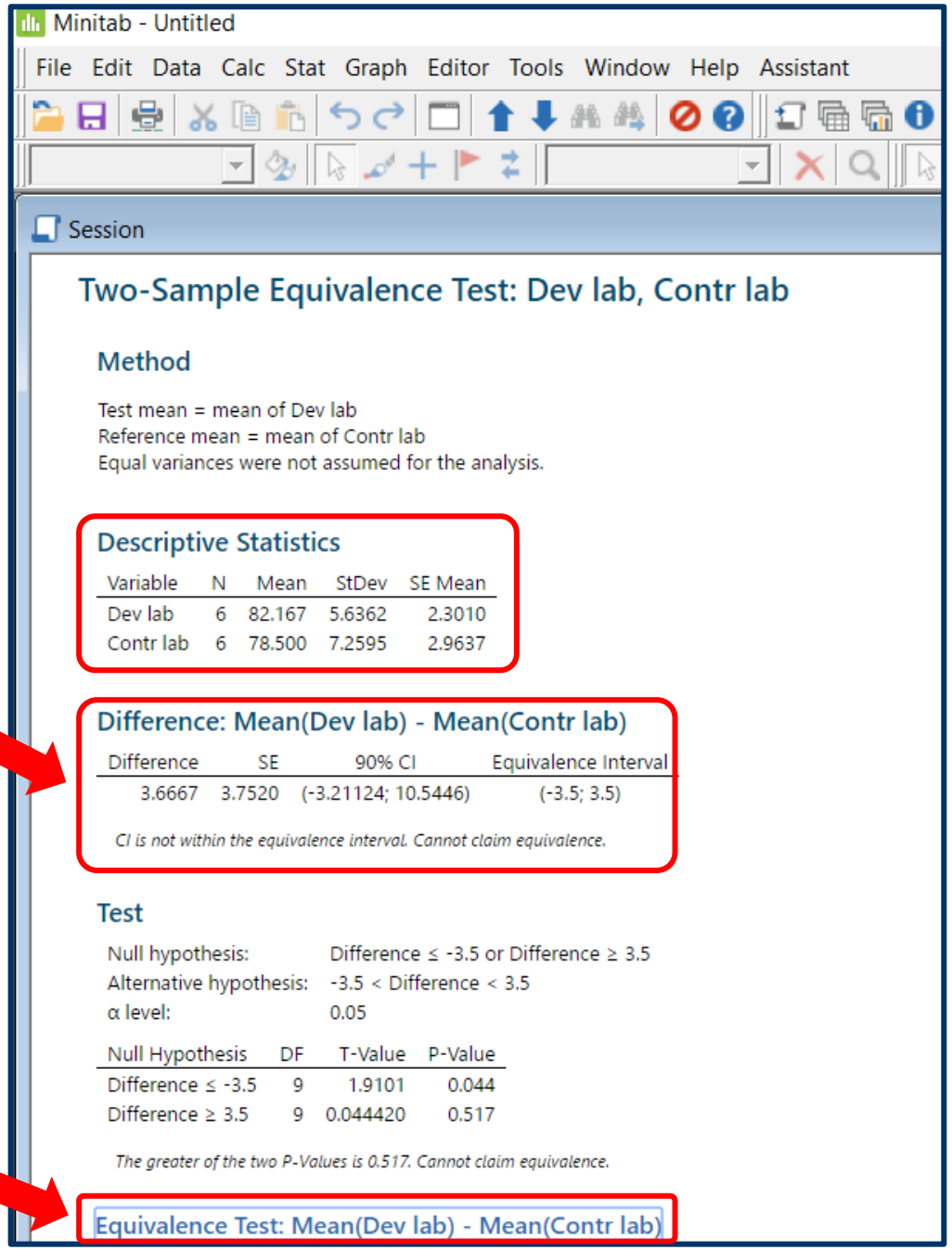


## Results

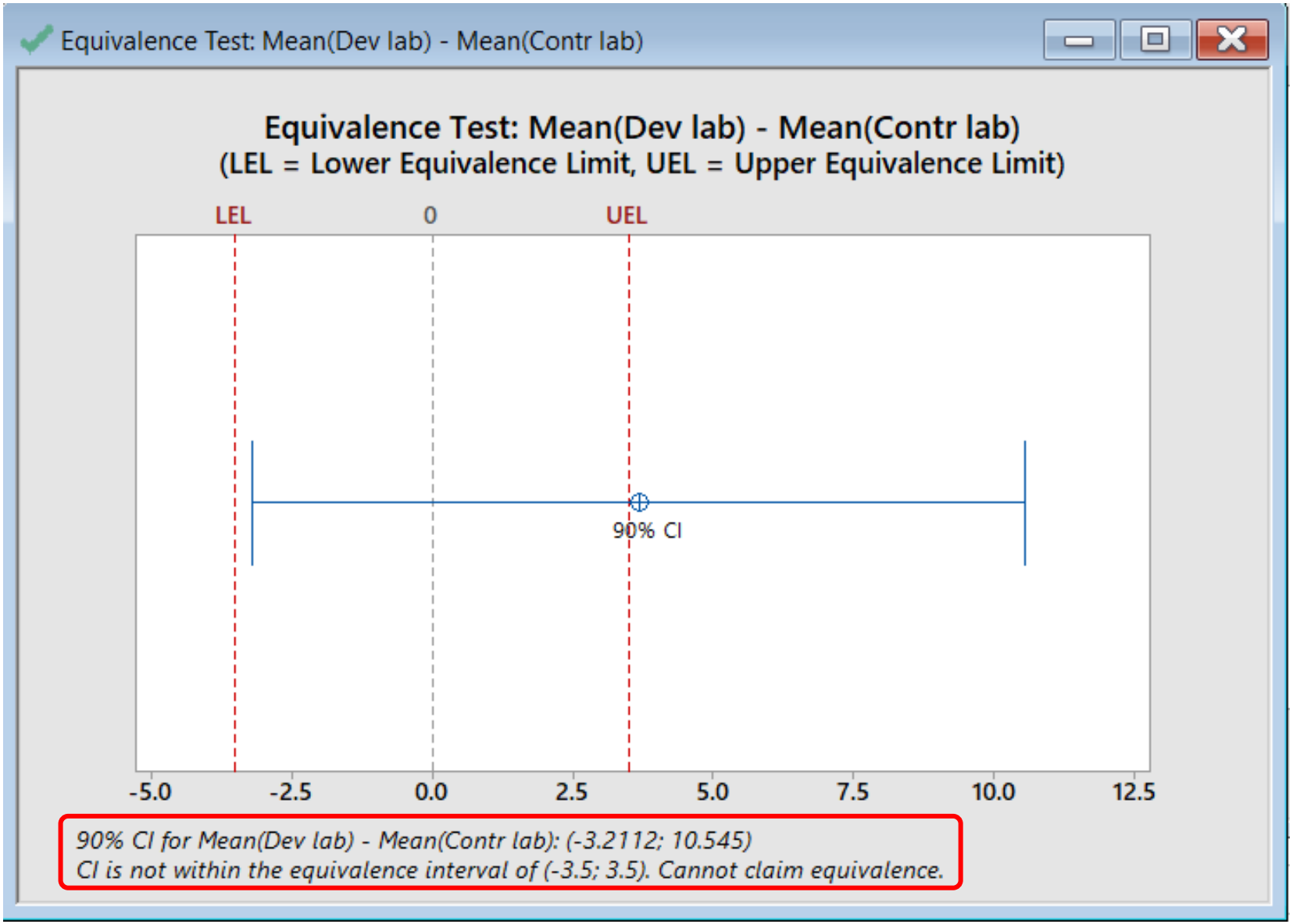
Once the calculation is completed, an output summary is reported inside the Session window:

As apparent, the conclusion is that the confidence interval (90% CI, in this case) is NOT within the equivalence interval, thus equivalence cannot be claimed.

By clicking on the bottom link, a graphical representation of the test can be obtained.

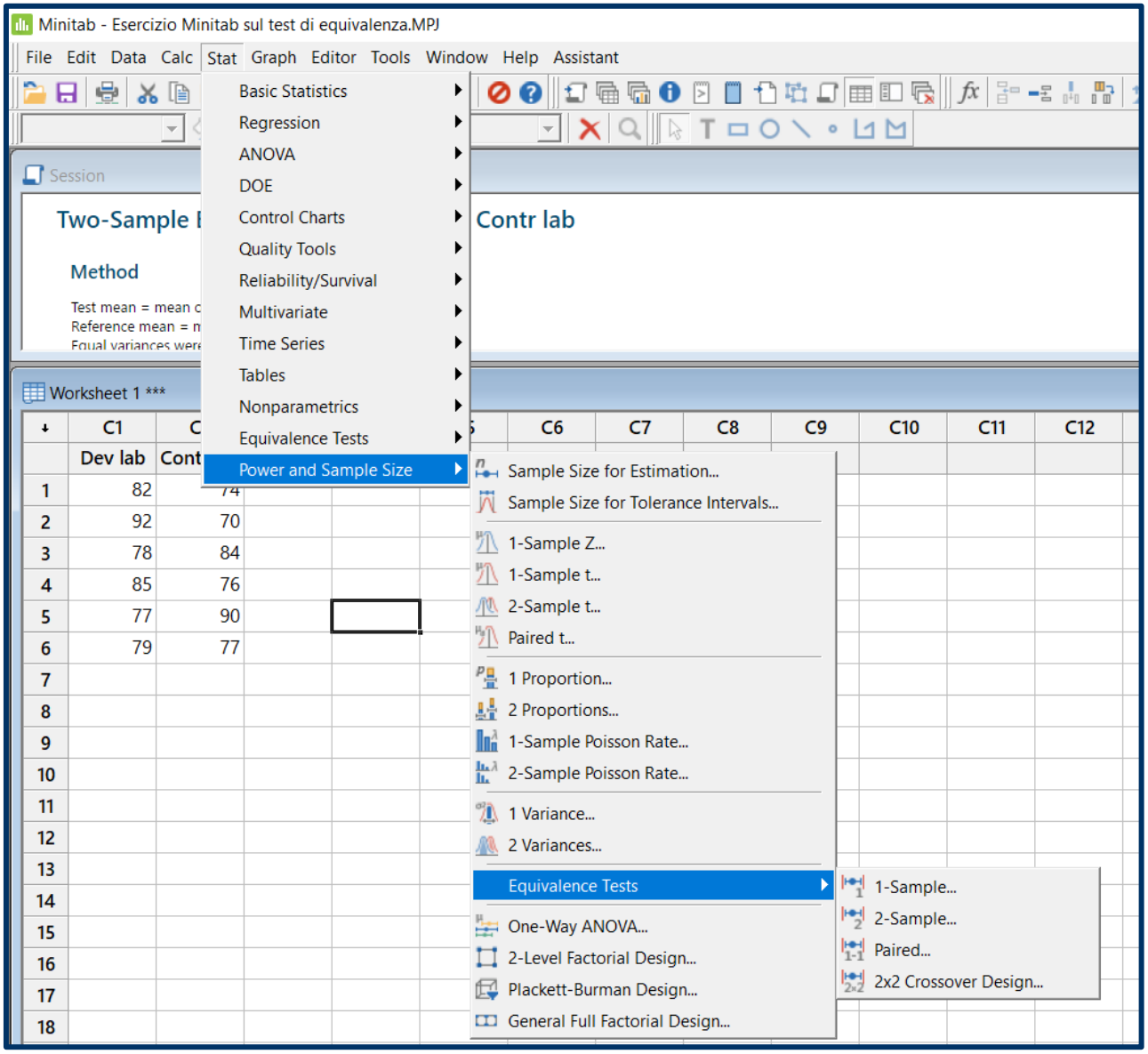


# Graphical output of the equivalence test



# Power and Sample Size calculations with Minitab 18

A further option of the Stat menu of Minitab 18 enables calculations of power and sample size also for equivalence tests:



The window opened after choosing the option includes the specification of $-\theta$ and $\theta$ values (Lower and Upper limit) and of experimental standard deviation.

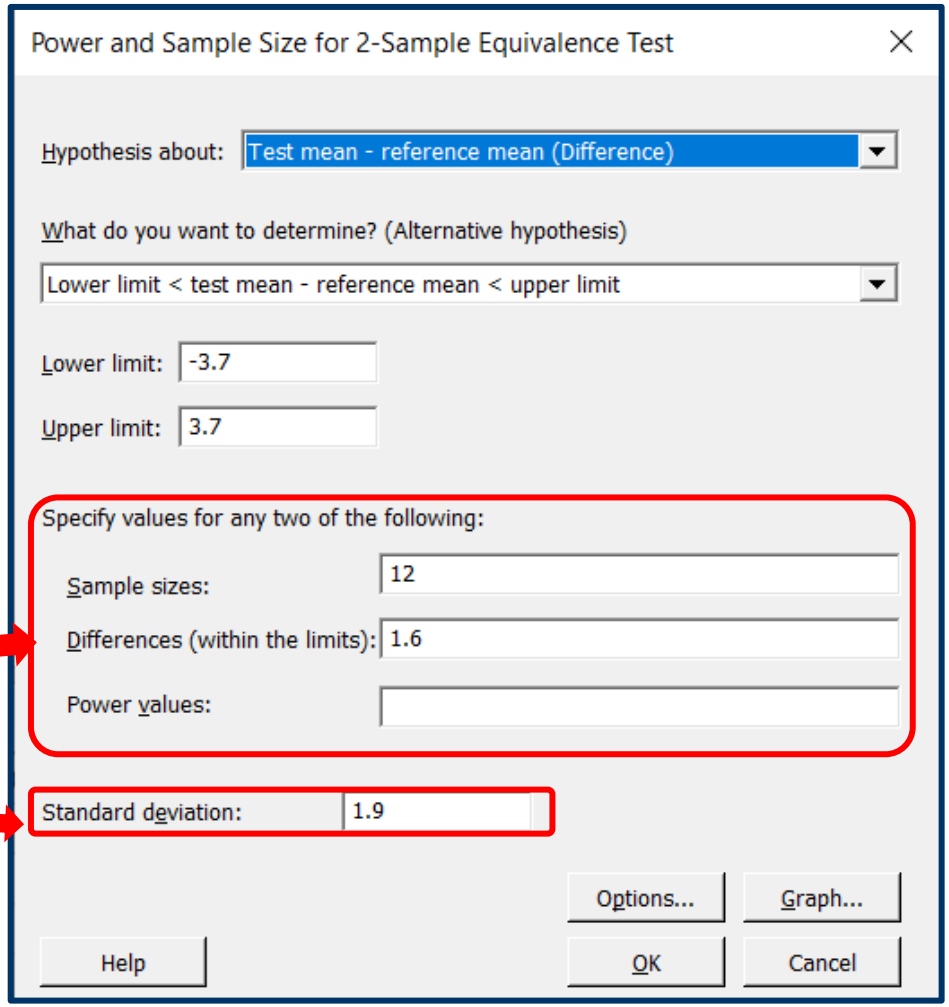
Afterwards, if a specific sample size and a specific difference (like the one between experimental means) are specified, the program can calculate power values as a function of differences and sample sizes.

Considering data of Table 2 shown before ($\theta$ = **3.7**):

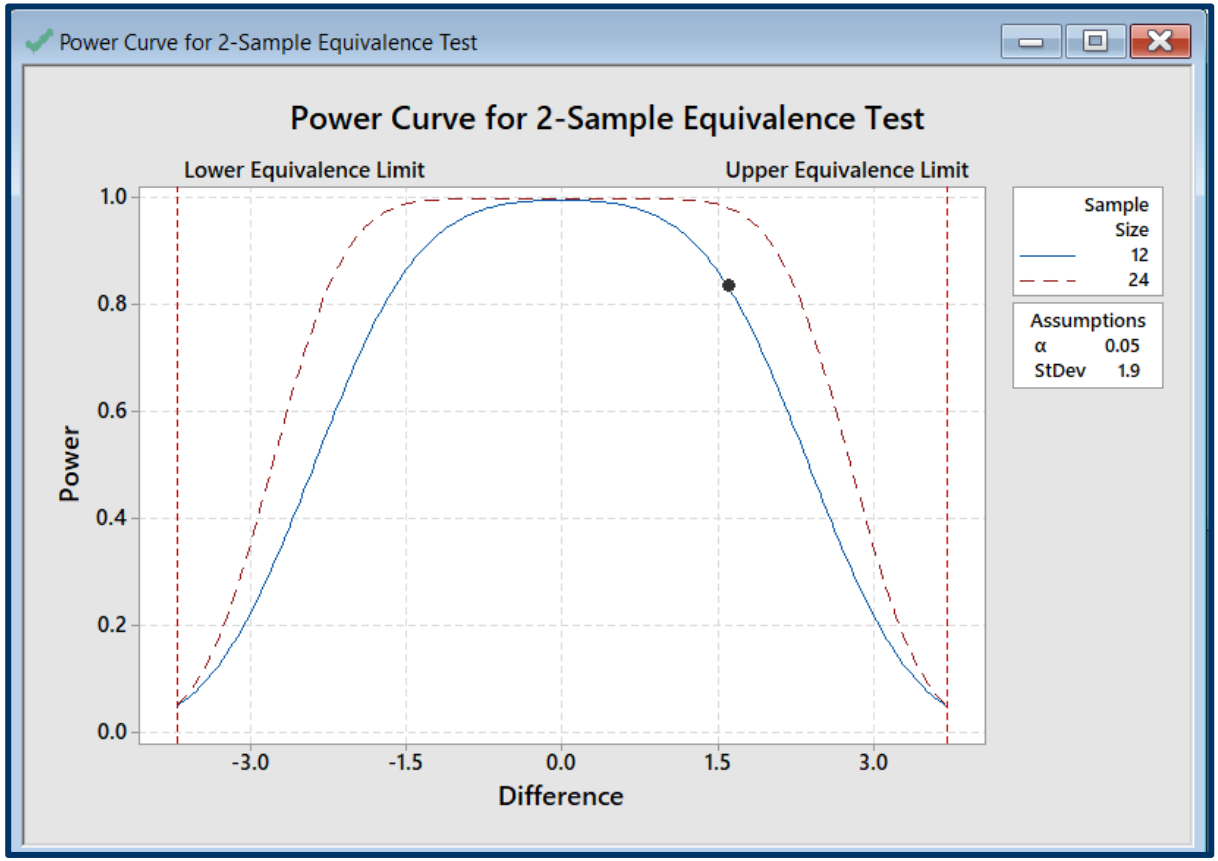
**Table 2. Statistical significance vs practical relevance in the transfer of a dissolution method.**

($\theta$ calculated on the basis of the development laboratory *s* of 1.9%, with $\alpha = \beta = 0.05$, $\delta = 0$, $n = 12$)

| Tablet number | % Label strength dissolved Development lab | Manufacturing QC lab |
|---|---|---|
| 1 | 90.8 | 86.2 |
| 2 | 88.0 | 87.4 |
| 3 | 90.5 | 88.2 |
| 4 | 90.0 | 89.7 |
| 5 | 91.0 | 87.3 |
| 6 | 86.0 | 87.6 |
| 7 | 88.3 | 88.0 |
| 8 | 89.3 | 86.5 |
| 9 | 88.9 | 89.6 |
| 10 | 91.1 | 89.1 |
| 11 | 86.2 | |
| 12 | 91.3 | 86.2 |
| $\bar{y}$ | 89.3 | 87.7 |
| *s* | 1.9 | 1.3 |
| % RSD | 2.1 | 1.5 |
| TOST: 90% confidence interval for difference ($\theta$ = 3.7%) | (0.5, 2.7) | |
| Two-sample *t*-test p-value | 0.02 | |



The following «Test power vs difference» curves are obtained:




Notably, the user can make the program generate a further power curve, for a different sample size (n = 24 was chosen in the present case), using the «Graph…» option in the window shown before.

As expected, the test power :

- ✓ is decreased, for a given sample size, at the increase of the difference absolute value;
- ✓ is increased, for a given difference, at the increase of sample size.

The power curve generation output depends, obviously, also on the selected standard deviation.

In the case of Table 2 data, the upper limit of a one-sided 80% confidence interval for standard deviation can be calculated as follows:

$$s^* = s\sqrt{\frac{n-1}{\chi^2_{(\gamma,\, n-1)}}}$$

For n = 12, the $\chi^2_{(0.2,\ 11)}$ value has to be calculated and it is equal to 6.989.

Given the maximum experimental standard deviation reported in Table 2, s = 1.9, it follows that s* = 2.38.

As shown by the figures, given a specific sample size and difference, the test power is decreased at the increase of standard deviation.