Two One-Sided t-Test

In this leaflet we present the Two One-Sided Test (TOST) procedure for testing the equivalence of two values

Introduction

Researchers often face the task of equivalence testing for two (mean) values, or for a mean value and a reference value. Unlike the conventional t-test, which assesses whether two values are different, the TOST procedure evaluates whether the difference is acceptable.

What is the essence of the TOST procedure?

In the TOST procedure, the difference between two values (*D*) is compared with a predefined value (the equivalence limit, *E*), reflecting the practical significance of the difference. Deviations in both directions are tested separately using a one-sided t-test at the chosen significance level.

What must be set in advance?

(1) The maximum allowable difference between two values (*E*), usually defined by expert consensus. (2) The risk α, i.e. the risk of falsely accepting equivalence, controlled by a low probability level (usually α = 0.05, corresponding to a 95% confidence level). (3) The risk *β*, i.e. the risk of falsely rejecting equivalence, controlled by requirements for experimental data precision and sample size.

How are statistical hypotheses constructed for the TOST procedure?

Null hypothesis H_0: the two values differ by more than *E* (non-equivalence);

Alternative hypothesis H_a: the two values differ by no more than *E* (equivalence).

If the null hypothesis is rejected, the alternative hypothesis is accepted. In the TOST procedure, the claim of equivalence is therefore placed in the alternative hypothesis, so that the risk α is directly controlled by the chosen significance level. To conclude which hypothesis is true, the upper and lower confidence limits (*UCL* and *LCL*) are constructed:

$$UCL = D + t \times s_D, \text{ and } LCL = D - t \times s_D,$$

where *D* is the difference between two values, *t* is the upper 100(1 - α) % percentile of the Student's *t* distribution, and s_D is the standard error of *D*, calculated from experimental data as the standard deviation divided by the square root of the number of test results.

That is, the TOST procedure evaluates whether confidence limits are within the equivalence limits.

In the TOST procedure, H_a is treated as consisting of two hypotheses that are tested in turn:

$$H_{a1}: UCL \leq E_+, \text{ and } H_{a2}: LCL \geq E_-,$$

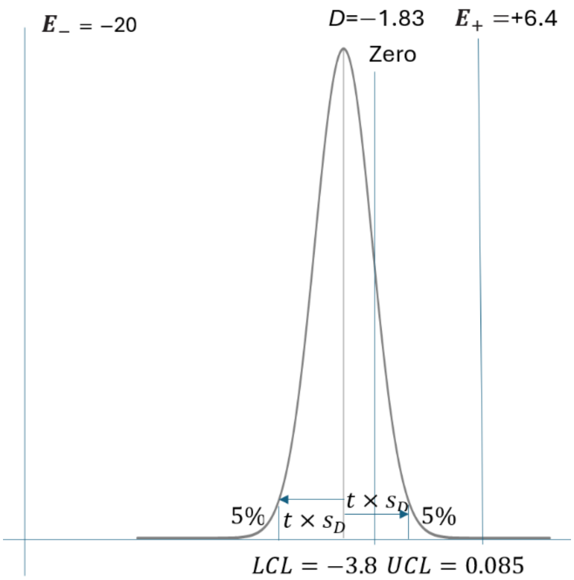
where E_+ and E_- are the maximum allowable differences in the positive and negative directions.

The TOST procedure distinguishes two directions in which the measured difference is compared with the maximum allowable difference *E* (see Figure on page 2).

The 100×(1−α)% one-sided confidence interval is used to define a decision-relevant bound for *D*, i.e. a value that will not be exceeded by 100×(1−α)% of results. Thus, for α = 0.05 (95%), the *t* value cuts off 5% of the area in the upper tail of the distribution (application of t-test in one direction) and, correspondingly, in the lower tail (application of *t*-test in the opposite direction).

An example of the practical application of the TOST procedure

Validation results of an analytical method for monitoring cleaning of process equipment after production of a medicinal product are presented. Washes are analysed after applying model mixtures to equipment surfaces. The method is considered suitable if the population mean of



Recovery (μ_R) is ≥80%. Recovery values (R) exceeding 100% indicate problems with the method; so, the upper limit for μ_R was set at 106.4%, based on conventional practice in the sector. Individual values of R were determined six times (R_i, n = 6), and the average value ($\bar{R}$) is is used as the best estimate of μ_R. The precision of analytical results is acceptable if the standard deviation (s) of R_i is ≤3.2%.

μ_R is compared to the reference value of 100 % with zero uncertainty, this case is the *bias equivalence*, a special case of *means equivalence*.

For Recovery, deviations in negative and positive directions are not equivalent. Therefore, the limits are set separately as E_- and E_+. For the deviation of μ_R in the negative direction $E_- = 80 - 100 = -20$, and in the positive direction $E_+ = 106.4 - 100 = 6.4$.

FIGURE. The distribution of possible values of R, $\bar{R}$*, D,* confidence interval (C.I.)*,* E_- *and* E_+*, LCL/UCL, and the 5% cut-off areas for the given example of the TOST procedure is shown in the D coordinate.*

The statistical hypotheses are constructed as:

Null hypothesis (non-equivalence)	H_{02}: $\mu_R - 100\ \% < E_-$	H_{01}: $\mu_R - 100\ \% > E_+$
Alternative hypothesis (equivalence)	H_{a2}: $\mu_R - 100\ \% \geq E_-$	H_{a1}: $\mu_R - 100\ \% \leq E_+$

For correct decisions, equivalence limits belong to the equivalence region, not the null hypothesis.

To ensure an acceptable risk of mistakenly accepting an unsuitable method (α, the risk in routine use), a 95% confidence level is used. To limit the risk of mistakenly rejecting a suitable method (β, the risk during validation), s should not exceed 3.2% (n = 6), corresponding to a one-sided 95% confidence deviation of about 6.4% (t = 2.015), i.e. the upper equivalence limit E_+.

Experimental results:

R_1	R_2	R_3	R_4	R_5	R_6
95.1	101.3	100.4	98.3	97	96.9

Calculation results:

$$\bar{R} = \frac{\sum_{i=1}^{n} x_i}{n} = 98.17;\ D = 98.17 - 100 = -1.83;\ \nu = 5;\ s = \sqrt{\sum_{i=1}^{n} \frac{(R_i - \bar{R})^2}{n-1}} = 2.33;\ s_D = \frac{s}{\sqrt{n}} = \frac{2.33}{\sqrt{6}} = 0.95;$$

where n is the number of analyses; ν is the number of degrees of freedom for $\bar{R}$; $\nu = n - 1$.

Because $s \leq 3.2$ %, results are suitable for equivalence estimation. The 95th percentile of Student's t distribution ($\nu = 5$) is 2.015:

$$UCL = D + s_D \times t = -1.83 + 0.95 \times 2.015 = 0.085;\ LCL = D - s_D \times t = -1.83 - 0.95 \times 2.015 = -3.8.$$

That is, –3.8 > –20 and 0.085 < 6.4. The equivalence is accepted for a 95 % confidence level. The risk of making an incorrect decision is acceptably low. The method is acceptable for the determination of product residues on the surface of process equipment.

More information / further reading

ASTM E2935-21. *Standard Practice for Evaluating Equivalence of Two Testing Processes.*

USP–NF. General Chapter <1010> *Analytical Data—Interpretation and Treatment.*

Produced by the Eurachem Measurement Uncertainty and Traceability Working Group
First English edition March 2026
www.eurachem.org