Package ‘tost.suite’

February 6, 2026

Version 3.1.9

Date 2026-1-28

Title Two One-Sided Tests for Equivalence

Imports combinat, Hmisc, index0, lm.beta, mathjaxr, rlang, stringr, webuse

Depends R (>= 3.5.0)

RdMacros mathjaxr

BuildManual TRUE

Description Ports the 'Stata' ado package 'tost' which provides a suite of commands to perform two one-sided tests for equivalence following the approach by Schuirman (1987) <doi:10.1007/BF01068419>. Commands are provided for t tests on means, z tests on proportions, McNemar's test (1947) <doi:10.1007/BF02295996> on proportions and related tests, tests on the regression coefficients from OLS linear regression (not yet implementing all of the current regression options from the 'Stata' 'tostregress' command, e.g., survey regression options, estimation options, etc.), Wilcoxon's (1945) <doi:10.2307/3001968> signed rank tests, Wilcoxon-Mann-Whitney (1947) <doi:10.1214/aoms/1177730491> rank sum tests, supporting inference about equivalence for a number of paired and unpaired, parametric and nonparametric study designs and data types. Each command tests a null hypothesis that samples were drawn from populations different by at least plus or minus some researcher-defined level of tolerance, which can be defined in terms of units of the data or rank units (Delta), or in units of the test statistic's distribution (epsilon) except for tost.rrp() and tost.rrpi(). Enough evidence rejects this null hypothesis in favor of equivalence within the tolerance. Equivalence intervals for all tests may be defined symmetrically or asymmetrically.

License GPL-2

LazyData no

Encoding UTF-8

Contents

`canada` *Health Protection Branch of Canada equivalence trial for a generic drug*

Description

Example of doctor evaluation of two different drugs—one a test drug, and one a refrence drug—as either “effective” or “ineffective” as described on page 276 of Tu (1997).

Usage

```
data(canada)
```

Format

A data frame containing two binary variables, `drug`, where 0 means “Ineffective” and 1 means “Effective” and `group`, where 1 means “Test drug” and 2 means “reference drug” in 201 observations.

References

Tu, D. (1997) Two one-sided tests procedures in establishing therapeutic equivalence with binary clinical endpoints: Fixed sample performances and sample size determination. *Journal of Statistical Computing and Simmulation* **59**, 271–290.

hivfluid	*Outcomes of an HIV screening test*

Description

Example of two different tests—one from a blood plasma sample, and one from an alternate body fluid sample, neither being a 'gold standard' test—giving HIV positive and HIV negative status based on research by Lachenbruch and Lynch (1998).

Usage

```
data(hivfluid)
```

Format

A data frame containing two binary variables, `plasma` and `altenrate`, where 1 means "HIV Positive" and 1 means "HIV Negative" in 1157 observations.

References

Lachenbruch, P. A. and Lynch, C. J. (1998) Assessing screening tests: Extensions of McNemar's test. *Statistics In Medicine* **17**, 2207–2217.

tost.mcc	*Paired* z *test for equivalence of marginal probabilities in binary data*

Description

Performs two one-sided z tests for equivalence of marginal probabilities in binary data

Usage

```
tost.mcc(
  x           = NA,
  y           = NA,
  frequency   = NA,
  eqv.type    = equivalence.types,
  eqv.level   = 1,
  upper       = NA,
  ccontinuity = continuity.correction.methods,
  conf.level  = 0.95,
  relevance   = TRUE)

equivalence.types
#c("delta", "epsilon")

continuity.correction.methods
#c("none", "yates", "edwards")
```

Arguments

`x` — a (non-empty) vector of binary data values of equal length to `y`. The order of observations in `x` is assumed to correspond to the order of observations in `y` (i.e. `x` and `y` are paired.

`y` — a (non-empty) vector of binary data values of equal length to `x`. The order of observations in `y` is assumed to correspond to the order of observations in `x` (i.e. `x` and `y` are paired.

`frequency` — an optional (non-empty) vector of equal length to `x` and `y` containing non-negative integer frequencies indicating the number of duplicated observations corresponding to the paired `x` and `y` observations.

`eqv.type` — defines whether the equivalence interval will be defined in terms of Δ or ε (`"delta"`, or `"epsilon"`). These options change the way that `evq.level` is interpreted: when `"delta"` is specified, the `evq.level` is expressed in the units of marginal probabilities being tested, and when `"epsilon"` is specified, the `evq.level` is measured in units of the z distribution; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. The default is `"delta"`.

Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \le z_{\nu,\alpha}$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that it is not possible to reject any H_0^- if $\Delta \le \text{standard error} \times z_{\nu,\alpha}$. `tost.mcc` reports when either of these conditions obtain.

`eqv.level` — defines the equivalence threshold for the tests depending on whether `eqv.type` is `"delta"` or `"epsilon"` (see above). Researchers are responsible for choosing meaningful values of Δ or ε. The default value is 1, which is not a useful value for either `eqv.type="delta"` or `eqv.type="epsilon"`.

`upper` — defines the upper equivalence threshold for the test, is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper==abs(eqv.level)`, then `upper` will be ignored.

`conf.level` — confidence level of the interval, and complement of the test's nominal type I error rate α.

`ccontinuity` — calculates test statistics for both positivist and negativist tests using a continuity correction. The default is `"none"`, users may select a Yates continuity correction using the `"yates"` option, or the Hauck-Anderson continuity correction using the `"ha"` option. Note that the Hauck-Anderson continuity correction also adjusts the standard error of the proportion used to calculate test statistics.

`relevance` — reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

`tost.mcc` tests for equivalence of the marginal probabilities of exposure in matched case-control data. It calculates a Wald-type asymptotic z test (Liu, et al., 2002) in a two one-sided tests approach (Schuirmann, 1987). `tost.mcci` is the immediate form of `tost.mcc`. Typically the null hypotheses of the corresponding McNemar's χ^2 test (McNemar, 1947) for difference in marginal probabilities are framed from an assumption of equality of marginal probability of exposure between cases and controls (e.g., $\text{H}_0^+ : \frac{b}{n} - \frac{c}{n} = 0$, rejecting this assumption only with sufficient evidence. When performing tests for equivalence of marginal probabilities, the null hypothesis is framed as the difference in marginal probabilities is at least as much as the equivalence interval as defined by some chosen level of tolerance (as specified by `eqv.type` and `eqv.level`).

With respect to a z test, a negativist null hypothesis takes one of the following two forms depending on whether tolerance is defined in terms of Δ (equivalence expressed in the units of the marginal probability of counts of discordant pairs) or in terms of ε (equivalence expressed in the units of the z distribution):

$\text{H}_0^-: \left|\frac{b}{n} - \frac{c}{n}\right| \geq \Delta,$

where the equivalence interval ranges from $\left(\frac{b}{n} - \frac{c}{n}\right) - \Delta$ to $\left(\frac{b}{n} - \frac{c}{n}\right) + \Delta$, and where b is the count of pairs with cases exposed, but controls unexposed, and and c is the count of pairs with cases unexposed and controls exposed. This translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-: \frac{b}{n} - \frac{c}{n} \geq \Delta$, or
$\text{H}_{02}^-: \frac{b}{n} - \frac{c}{n} \leq -\Delta$.

–OR–

$\text{H}_0^-: |Z| \geq \varepsilon,$

where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-: Z \geq \varepsilon$; or
$\text{H}_{02}^-: Z \leq -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

$\text{H}_0^-: \frac{b}{n} - \frac{c}{n} \leq \Delta_{\text{lower}}$, or $\frac{b}{n} - \frac{c}{n} \geq \Delta_{\text{upper}}$

where the equivalence interval ranges from $\left(\frac{b}{n} - \frac{c}{n}\right) + \Delta_{\text{lower}}$ to $\left(\frac{b}{n} - \frac{c}{n}\right) + \Delta_{\text{upper}}$. This also translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-: \frac{b}{n} - \frac{c}{n} \geq \Delta_{\text{upper}}$; or
$\text{H}_{02}^-: \frac{b}{n} - \frac{c}{n} \leq \Delta_{\text{lower}}$.

–OR–

$\text{H}_0^-: Z \leq \varepsilon_{\text{lower}}$, or $Z \geq \varepsilon_{\text{upper}}$, with:

$\text{H}_{01}^-: Z \geq \varepsilon_{\text{upper}}$; or
$\text{H}_{02}^-: Z \leq \varepsilon_{\text{lower}}$.

NOTE: the appropriate level of $\alpha = (1-$`conf.level`$)$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H_{01}^- and H_{02}^- by comparing the resulting p-value to 0.01 (Tryon and Lewis, 2008; Wellek, 2010).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., $\text{H}_0^+: \frac{b}{n} - \frac{c}{n} = 0$ or H_0^+: Z = 0) and tests for equivalence (e.g., either $\text{H}_0^-: \left|\frac{b}{n} - \frac{c}{n}\right| \geq$

Δ, or H_0^-: $|Z| \geq \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H_0^+, but fail to reject H_0^-, and conclude that there is a **relevant difference** in marginal proportions at least as large as Δ or ε.
2. One may fail to reject H_0^+, but reject H_0^-, and conclude that there is **equivalence** in marjinal proportions within the equivalence range (i.e. defined by Δ or ε).
3. One may reject both H_0^+ and H_0^-, and conclude that there is a **trivial difference** in marjinal proportions which lies within the equivalence range (i.e. defined by Δ or ε).
4. One may fail to reject both H_0^+ and H_0^-, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.mcc` returns:

`statistics`	a vector containing the value of z_1 and z_2; if `relevance=TRUE`, these are followed by the value of the z statistic for the postivist test for difference.
`p.values`	a vector of p values for the z tests.
`estimate`	the estimated difference in proportion with exposure.
`threshold`	a scalar containing the equivalence threshold when `eqv.type="delta"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="delta"`. A scalar containing the equivalence threshold when `eqv.type="epsilon"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="epsilon"`.
`conclusion`	relevance test conclusion for a given α and Δ or ε.

Author(s)

Alexis Dinno (<`alexis.dinno@pdx.edu`>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

Suggested citation: Dinno, A. 2025. **tost.mcc**: Paired z test for equivalence of marginal probabilities in binary data. In: **tost.suite** R software package. URL: https://alexisdinno.com/Software/index.shtml#tost

References

Edwards, A. (1948) Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika* **13**, 185–187.

Liu, J., et al., (2002) Tests for equivalence or non-inferiority for paired binary data. *Statistics In Medicine* **21**, 231–245.

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Tryon, W. W., and C. Lewis. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277

Yates, F. (1934) Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*. **1**, 217–235.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, second edition. Chapman and Hall/CRC Press. p. 31

See Also

`mcnemar.test`, `tost.mcci`.

Examples

```
require("webuse")

# Setup
webuse("mccxmpl")

# Relevance test in paired binary data
tost.mcc(
  x=mccxmpl$case,
  y=mccxmpl$control,
  frequency=mccxmpl$pop,
  eqv.type="delta",
  eqv.level=.2,
  relevance=TRUE)
```

`tost.mcci` *Immediate paired z test for equivalence of marginal probabilities in binary data*

Description

Immediately performs two one-sided z tests for equivalence of marginal probabilities in binary data

Usage

```
tost.mcci(
    a = NA, b = NA, c = NA, d = NA,
    eqv.type    = equivalence.types,
    eqv.level   = 1,
    upper       = NA,
    ccontinuity = continuity.correction.methods,
    conf.level  = 0.95,
    relevance   = TRUE)

equivalence.types
#c("delta", "epsilon")

continuity.correction.methods
#c("none", "yates", "edwards")
```

Arguments

`a` a non-negative integer indicating the number of paired observations with both cases and controls exposed.

`b` a non-negative integer indicating the number of paired observations with cases exposed and controls unexposed.

`c` a non-negative integer indicating the number of paired observations with cases unexposed and controls exposed.

`d` a non-negative integer indicating the number of paired observations with both cases and controls unexposed.

`eqv.type` defines whether the equivalence interval will be defined in terms of Δ or ε (`"delta"`, or `"epsilon"`). These options change the way that `evq.level` is interpreted: when `"delta"` is specified, the `evq.level` is expressed in the units of marginal probabilities being tested, and when `"epsilon"` is specified, the `evq.level` is measured in units of the z distribution; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. The default is `"delta"`.

Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \le z_{\nu,\alpha}$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that it is not possible to reject any H_0^- if $\Delta \le \text{standard error} \times z_{\nu,\alpha}$. `tost.mcci` reports when either of these conditions obtain.

`eqv.level` defines the equivalence threshold for the tests depending on whether `eqv.type` is `"delta"` or `"epsilon"` (see above). Researchers are responsible for choosing meaningful values of Δ or ε. The default value is 1, which is not a useful value for either `eqv.type="delta"` or `eqv.type="epsilon"`.

`upper` defines the upper equivalence threshold for the test, is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper==abs(eqv.level)`, then `upper` will be ignored.

`conf.level` confidence level of the interval, and complement of the test's nominal type I error rate α.

`ccontinuity` calculates test statistics for both positivist and negativist tests using a continuity correction. The default is `"none"`, users may select a Yates continuity correction using the `"yates"` option, or an Edwards continuity correction using the `"edwards"` option. The Yates continuity correction (Yates, 1934) uses the term $\left[(b-c)-\frac{1}{2}\right]$ for z_1, and the term $\left[(b-c)+\frac{1}{2}\right]$ for z_2. The Edwards continuity correction (Edwards, 1947) uses the term $[(b-c)-1]$ for z_1, and the term $[(b-c)+1]$ for z_2.

`relevance` reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

Immediate commands perfom tests given summary statistics, rather than given data. `tost.mcci` tests for equivalence of the marginal probabilities of exposure in matched case-control data. It calculates a Wald-type asymptotic z test (Liu, et al., 2002) in a two one-sided tests approach (Schuirmann, 1987). `tost.mcc` is the non-immediate form of `tost.mcci`. Typically the null hypotheses of the corresponding McNemar's χ^2 test (McNemar, 1947) for difference in marginal probabilities are framed from an assumption of equality of marginal probability of exposure between cases and controls (e.g., $\text{H}_0^+: \frac{b}{n} - \frac{c}{n} = 0$, rejecting this assumption only with sufficient evidence. When performing tests for equivalence of marginal probabilities, the null hypothesis is framed as the difference in marginal probabilities is at least as much as the equivalence interval as defined by some chosen level of tolerance (as specified by `eqv.type` and `eqv.level`).

With respect to a z test, a negativist null hypothesis takes one of the following two forms depending on whether tolerance is defined in terms of Δ (equivalence expressed in the units of the marginal probability of counts of discordant pairs) or in terms of ε (equivalence expressed in the units of the z distribution):

$\text{H}_0^-\text{:}\ \left|\frac{b}{n} - \frac{c}{n}\right| \ge \Delta$,

where the equivalence interval ranges from $\left(\frac{b}{n} - \frac{c}{n}\right) - \Delta$ to $\left(\frac{b}{n} - \frac{c}{n}\right) + \Delta$, and where b is the count of pairs with cases exposed, but controls unexposed, and and c is the count of pairs with cases unexposed and controls exposed. This translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\text{:}\ \frac{b}{n} - \frac{c}{n} \ge \Delta$, or
$\text{H}_{02}^-\text{:}\ \frac{b}{n} - \frac{c}{n} \le -\Delta$.

–OR–

$\text{H}_0^-\text{:}\ |Z| \ge \varepsilon$,

where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\text{:}\ Z \ge \varepsilon$; or
$\text{H}_{02}^-\text{:}\ Z \le -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

$\text{H}_0^-\colon \frac{b}{n} - \frac{c}{n} \le \Delta_{\text{lower}}$, or $\frac{b}{n} - \frac{c}{n} \ge \Delta_{\text{upper}}$

where the equivalence interval ranges from $\left(\frac{b}{n} - \frac{c}{n}\right) + \Delta_{\text{lower}}$ to $\left(\frac{b}{n} - \frac{c}{n}\right) + \Delta_{\text{upper}}$. This also translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\colon \frac{b}{n} - \frac{c}{n} \ge \Delta_{\text{upper}}$; or
$\text{H}_{02}^-\colon \frac{b}{n} - \frac{c}{n} \le \Delta_{\text{lower}}$.

–OR–

$\text{H}_0^-\colon Z \le \varepsilon_{\text{lower}}$, or $Z \ge \varepsilon_{\text{upper}}$, with:

$\text{H}_{01}^-\colon Z \ge \varepsilon_{\text{upper}}$; or
$\text{H}_{02}^-\colon Z \le \varepsilon_{\text{lower}}$.

NOTE: the appropriate level of $\alpha = (1-\texttt{conf.level})$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H_{01}^- and H_{02}^- by comparing the resulting p-value to 0.01 (Tryon and Lewis, 2008; Wellek, 2010).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., $\text{H}_0^+\colon \frac{b}{n} - \frac{c}{n} = 0$ or H_0^+: Z = 0) and tests for equivalence (e.g., either $\text{H}_0^-\colon \left|\frac{b}{n} - \frac{c}{n}\right| \ge \Delta$, or $\text{H}_0^-\colon |Z| \ge \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H_0^+, but fail to reject H_0^-, and conclude that there is a **relevant difference** in marginal proportions at least as large as Δ or ε.
2. One may fail to reject H_0^+, but reject H_0^-, and conclude that there is **equivalence** in marjinal proportions within the equivalence range (i.e. defined by Δ or ε).
3. One may reject both H_0^+ and H_0^-, and conclude that there is a **trivial difference** in marjinal proportions which lies within the equivalence range (i.e. defined by Δ or ε).
4. One may fail to reject both H_0^+ and H_0^-, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.mcci` returns:

`statistics`	a vector containing the value of z_1 and z_2; if `relevance=TRUE`, these are followed by the value of the *z* statistic for the postivist test for difference.
`p.values`	a vector of *p* values for the *z* tests.
`estimate`	the estimated difference in proportion with exposure.
`threshold`	a scalar containing the equivalence threshold when `eqv.type="delta"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="delta"`. A scalar containing the equivalence threshold when `eqv.type="epsilon"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="epsilon"`.
`conclusion`	relevance test conclusion for a given α and Δ or ε.

Author(s)

Alexis Dinno (<alexis.dinno@pdx.edu>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

Suggested citation: Dinno, A. 2025. **tost.mcci**: Paired z test for equivalence of marginal probabilities in binary data. In: **tost.suite** R software package.

References

Edwards, A. (1948) Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika* **13**, 185–187.

Liu, J., et al., (2002) Tests for equivalence or non-inferiority for paired binary data. *Statistics In Medicine* **21**, 231–245.

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Tryon, W. W., and C. Lewis. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277

Yates, F. (1934) Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*. **1**, 217–235.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, second edition. Chapman and Hall/CRC Press. p. 31

See Also

`mcnemar.test`, `tost.mcc`.

Examples

```
# Immediate command for the relevance test in paired binary data in the help file
# for tost.mcc
tost.mcci(
    a=8, b=8, c=3, d=8,
    eqv.type="delta",
    eqv.level=.2,
    relevance=TRUE)
```

```
# Different example with an asymetric interval; the lower end of the equivalence
# interval = qnorm(.95)+.5 = 2.144854 meaning equivalence must lay no more
# than 0.5 sd beyond the critical value of Z for alpha = 0.05.  The upper end of
# the equivalence interval = qnorm(.95)+1 = 2.644854 meaning equivalence
# must lay no more than 1 sd beyond the critical value of Z for alpha = 0.05.
tost.mcci(
    a=4, b=9, c=8, d=5,
    eqv.type="epsilon",
    eqv.level=qnorm(.95)+.5,
    upper=qnorm(.95)+1,
    relevance=TRUE)
```

`tost.pr` *Mean-equivalence* z *tests*

Description

Performs two one-sided *z* tests for mean equivalence

Usage

```
tost.pr(
    x,
    y            = NULL,
    by           = NULL,
    by.names     = NULL,
    p0           = NA,
    eqv.type     = equivalence.types,
    eqv.level    = 1,
    upper        = NA,
    ccontinuity  = continuity.correction.methods,
    conf.level   = 0.95,
    x.name       = "",
    y.name       = "",
    relevance    = TRUE)

equivalence.types
#c("delta", "epsilon")

continuity.correction.methods
#c("none", "yates", "ha")
```

Arguments

`x`	a (non-empty) vector of binary data values.
`y`	an optional (non-empty) vector of binary data values.
`by`	an optional (non-empty) vector of group indicator values

`by.names`	an optional two-element character vector of group names. If none are supplied, the values of `by` will be used instead.
`p0`	a number indicating the true value of the proportion for a one-sample test. Implies `y=NULL` and `by=NULL`.
`eqv.type`	defines whether the equivalence interval will be defined in terms of Δ or ε (`"delta"`, or `"epsilon"`). These options change the way that `evq.level` is interpreted: when `"delta"` is specified, the `evq.level` is measured in the units of the variable(s) being tested, and when `"epsilon"` is specified, the `evq.level` is measured in units of the z distribution; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. The default is `"delta"`. Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \leq z_{\nu,\alpha}$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that it is not possible to reject any H_0^- if $\Delta \leq$ standard error $\times\, z_{\nu,\alpha}$. `tost.pr` reports when either of these conditions obtain.
`eqv.level`	defines the equivalence threshold for the tests depending on whether `eqv.type` is `"delta"` or `"epsilon"` (see above). Researchers are responsible for choosing meaningful values of Δ or ε. The default value is 1, which is not a useful value for either `eqv.type="delta"` or `eqv.type="epsilon"`.
`upper`	defines the upper equivalence threshold for the test, is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper==abs(eqv.level)`, then `upper` will be ignored.
`conf.level`	confidence level of the interval, and complement of the test's nominal type I error rate α.
`x.name`	specifies how the first variable will be labeled in the output. The default value of `x.name` is the variable name of `x`.
`y.name`	specifies how the second variable will be labeled in the output. The default value of `y.name` is the variable name of `y` when `y` is specified, or it has the same prefix as `x.name`, but with the higher/second of the two values of `by` or `by.names`.
`ccontinuity`	calculates test statistics for both positivist and negativist tests using a continuity correction. The default is `"none"`, users may select a Yates continuity correction using the `"yates"` option, or the Hauck-Anderson continuity correction using the `"ha"` option. Note that the Hauck-Anderson continuity correction also adjusts the standard error of the proportion used to calculate test statistics. The Yates option is included for convenience although the Hauck-Anderson correction is preferred (Tu, 1997).
`relevance`	reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

`tost.pr` tests for the equivalence of proportions within a symmetric equivalence interval defined by `eqvtype` and `eqvlevel` (or within an asymmetric interval when adding the `upper` argument) using a two one-sided z tests (TOST) approach (Schuirmann, 1987). Typically "positivist" null hypotheses are framed from an assumption of a lack of difference between two quantities, and reject this assumption only with sufficient evidence. When performing tests for equivalence, one frames a null hypothesis with the assumption that two quantities are different within an equivalence interval defined by some chosen level of tolerance.

With respect to an unpaired z test, an equivalence null hypothesis takes one of the following two forms depending on whether equivalence is defined in terms of Δ (equivalence expressed in the same units as proportions of the `x` and `y` variables) or in terms of ε (equivalence expressed in the units of the z distribution with the given degrees of freedom):

$\text{H}_0^-\colon |p_x - p_y| \geq \Delta$,

where the equivalence interval ranges from $(p_x - p_y) - \Delta$ to $(p_x - p_y) + \Delta$. This translates directly into two one-sided null hypotheses:

H^-_{01}: $p_x - p_y \geq \Delta$, or
H^-_{02}: $p_x - p_y \leq -\Delta$.

–OR–

$\text{H}_0^-\colon |Z| \geq \varepsilon$,

where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

H^-_{01}: $Z \geq \varepsilon$; or
H^-_{02}: $Z \leq -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

$\text{H}_0^-\colon p_x - p_y \leq \Delta_{\text{lower}}$, or $p_x - p_y \geq \Delta_{\text{upper}}$

where the equivalence interval ranges from $(p_x - p_y) + \Delta_{\text{lower}}$ to $(p_x - p_y) + \Delta_{\text{upper}}$. This also translates directly into two one-sided null hypotheses:

H^-_{01}: $p_x - p_y \geq \Delta_{\text{upper}}$; or
H^-_{02}: $p_x - p_y \leq \Delta_{\text{lower}}$.

–OR–

$\text{H}_0^-\colon Z \leq \varepsilon_{\text{lower}}$, or $Z \geq \varepsilon_{\text{upper}}$, with:

H^-_{01}: $Z \geq \varepsilon_{\text{upper}}$; or
H^-_{02}: $Z \leq \varepsilon_{\text{lower}}$.

NOTE: the appropriate level of $\alpha = (1 - \texttt{conf.level})$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H^-_{01} and H^-_{02} by comparing the resulting p-value to 0.01 (Tryon and Lewis, 2008; Wellek, 2010).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., $\text{H}_0^+\colon p_x - p_y = 0$ or) and tests for equivalence (e.g., either $\text{H}_0^-\colon |p_x - p_y| \geq \Delta$, or $\text{H}_0^-\colon |Z| \geq \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H_0^+, but fail to reject H_0^-, and conclude that there is a **relevant difference** in proportions at least as large as Δ or ε.
2. One may fail to reject H_0^+, but reject H_0^-, and conclude that there is **equivalence** in proportions within the equivalence range (i.e. defined by Δ or ε).
3. One may reject both H_0^+ and H_0^-, and conclude that there is a **trivial difference** in proportions which lies within the equivalence range (i.e. defined by Δ or ε).
4. One may fail to reject both H_0^+ and H_0^-, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.pr` returns:

`statistics`	a vector of the z statistics for the two one-sided tests; if `relevance=TRUE`, these are followed by the value of the z statistic for the postivist test for difference.
`p.values`	a vector of p values for the z tests.
`proportion`	a scalar estimate of the sample proportion in the one-sample test. A vector of the proportions in both groups, as well as the estimate of the proportion under the null hypothesis in the two-sample test.
`sample_size`	a scalar containing the sample size of the one-sample test. A vector of the sample size in both groups, as well as the combined sample size in the two-sample test.
`threshold`	a scalar containing the equivalence threshold when `eqv.type="delta"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="delta"`. A scalar containing the equivalence threshold when `eqv.type="epsilon"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="epsilon"`.
`conclusion`	a string containing the relevance test conclusion when `relevance=TRUE`.

Author(s)

Alexis Dinno (<alexis.dinno@pdx.edu>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

I am endebted to my winter 2013 and fall 2023 students for their inspiration. Much appreciation to Mick McVeety for troubleshooting the translation of my Stata **tost** package to R.

Suggested citation: Dinno, A. 2025. **tost.pr**: Mean-equivalence z tests. In: **tost.suite** R software package. URL: https://alexisdinno.com/Software/index.shtml#tost

References

Hauck, W. W., and Anderson, S. (1984) A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Pharmacodynamics*. **12**, 83–91.

Hauck, W. W., and Anderson, S. (1986) A comparison of large-sample confidence interval methods for the difference of two binomial probabilities. *The American Statistician*. **40**, 318–322.

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Tryon, W. W., and Lewis, C. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277

Tu, D. (1997) Two one-sided tests procedures in establishing therapeutic equivalence with binary clinical endpoints: Fixed sample performances and sample size determination. *Journal of Statistical Computing and Simmulation*. **59**, 271–290.

Yates, F. (1934) Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*. **1**, 217–235.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, second edition. Chapman and Hall/CRC Press. p. 31

See Also

prop.test, tost.pri.

Examples

```
require("webuse")

# Setup
webuse("auto")

# One-sample proportion equivalence test with asymmetric equivalence interval
tost.pr(
  auto$foreign,
  p0=0.4,
  eqv.type="delta",
  eqv.level=.15,
  upper=.2,
  relevance=FALSE)

# Setup
webuse("cure")

# Two-sample proportion relevance test; equivalence interval is +/- 1 sd
# beyond the critical value of Z for alpha = 0.05
tost.pr(
  x=cure$cure1,
  y=cure$cure2,
```

```
  eqv.type="epsilon",
  eqv.level=qnorm(.95)+1,
  conf.level=0.95,
  relevance=TRUE)

# Setup
data("canada")

# Two-group proportion equivalence test from Tu 1997, p 276, and incorporating
# a Hauck and Anderson continuity correction from that same example.

tost.pr(
  x=canada$drug,
  by=canada$group,
  eqv.type="delta",
  eqv.level=.2,
  ccontinuity="ha",
  conf.level=0.95,
  relevance=FALSE)
```

`tost.pri` *Immediate one- and two-sample* z *tests for proportion equivalence*

Description

Immediately performs two one-sided z tests for proportion equivalence

Usage

```
tost.pri(
    n1 = NA, obs1 = NA, n2 = NA, obs2 = NA, count = FALSE,
    eqv.type     = equivalence.types,
    eqv.level    = 1,
    upper        = NA,
    ccontinuity  = continuity.correction.methods,
    conf.level   = 0.95,
    x.name       = "x",
    y.name       = "y",
    relevance    = TRUE)

equivalence.types
#c("delta", "epsilon")

continuity.correction.methods
#c
```

Arguments

- `n1` required group 1 sample size.
- `obs1` required group 1 sample proportion if `count=FALSE`. If `count=TRUE`, then `obs1` is interpreted as the count of successes in the first sample (i.e. as the numerator of the group 1 sample proportion).
- `n2` an optional group 2 sample size. If `n2` is a positive integer, then `tost.pri` performs a two-sample test.
- `obs2` required true proportion (p_0) for the one-sample test when `n2=NA`. If `n2=NA` and `count=FALSE`, `obs2` is the group 2 sample proportion. If `n2=NA` and `count=TRUE`, `obs2` is still interpreted as the true population proportion (p_0) when `n2=NA`.
- `count` optionally indicates whether `n1` and `obs1` (but not `obs2`) are both to be treated as counts for a one-sample test, or, whether `n1`, `obs1`,`n2`, and `obs2` are to be treated as counts for a two-sample test.
- `eqv.type` defines whether the equivalence interval will be defined in terms of Δ or ε (`"delta"`, or `"epsilon"`). These options change the way that `evq.level` is interpreted: when `"delta"` is specified, the `evq.level` is expressed in the same units as proportion of the variable(s) being tested, and when `"epsilon"` is specified, the `evq.level` is expressed in units of the z distribution; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. The default is `"delta"`.

  Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \leq z_{\nu,\alpha}$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that it is not possible to reject any H_0^- if $\Delta \leq \text{standard error} \times z_{\nu,\alpha}$. `tost.pri` reports when either of these conditions obtain.
- `eqv.level` defines the equivalence threshold for the tests depending on whether `eqv.type` is `"delta"` or `"epsilon"` (see above). Researchers are responsible for choosing meaningful values of Δ or ε. The default value is 1, which is not a useful value for either `eqv.type="delta"` or `eqv.type="epsilon"`.
- `upper` defines the upper equivalence threshold for the test, is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper==abs(eqv.level)`, then `upper` will be ignored.
- `conf.level` confidence level of the interval, and complement of the test's nominal type I error rate α.
- `x.name` specifies how the first group will be labeled in the output. The default value of `x.name` is `"x"`.
- `y.name` specifies how the second group will be labeled in the output. The default value of `y.name` is `"y"`
- `ccontinuity` calculates test statistics for both positivist and negativist tests using a continuity correction. The default is `"none"`, users may select a Yates continuity correction using the `"yates"` option, or the Hauck-Anderson continuity correction using the `"ha"` option. Note that the Hauck-Anderson continuity correction (only available for the two-sample test) also adjusts the standard error of the proportion used to calculate test statistics.

`relevance` reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

Immediate commands perfom tests given summary statistics, rather than given data. `tost.pri` tests for the equivalence of proportions within a symmetric equivalence interval defined by `eqvtype` and `eqvlevel` (or within an asymmetric interval when adding the `upper` argument) using a two one-sided z tests (TOST) approach (Schuirmann, 1987). Typically "positivist" null hypotheses are framed from an assumption of a lack of difference between two quantities, and reject this assumption only with sufficient evidence. When performing tests for equivalence, one frames a null hypothesis with the assumption that two quantities are different within an equivalence interval defined by some chosen level of tolerance.

With respect to an unpaired z test, an equivalence null hypothesis takes one of the following two forms depending on whether equivalence is defined in terms of Δ (equivalence expressed in the same units as the proportions of the two variables) or in terms of ε (equivalence expressed in the units of the z distribution with the given degrees of freedom):

$\text{H}_0^-\colon |p_x - p_y| \geq \Delta,$

where the equivalence interval ranges from $(p_x - p_y) - \Delta$ to $(p_x - p_y) + \Delta$. This translates directly into two one-sided null hypotheses:

H_{01}^-: $p_x - p_y \geq \Delta$, or
H_{02}^-: $p_x - p_y \leq -\Delta$.

–OR–

$\text{H}_0^-\colon |Z| \geq \varepsilon,$

where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

H_{01}^-: $Z \geq \varepsilon$; or
H_{02}^-: $Z \leq -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

$\text{H}_0^-\colon p_x - p_y \leq \Delta_{\text{lower}}$, or $p_x - p_y \geq \Delta_{\text{upper}}$

where the equivalence interval ranges from $(p_x - p_y) + \Delta_{\text{lower}}$ to $(p_x - p_y) + \Delta_{\text{upper}}$. This also translates directly into two one-sided null hypotheses:

H_{01}^-: $p_x - p_y \geq \Delta_{\text{upper}}$; or
H_{02}^-: $p_x - p_y \leq \Delta_{\text{lower}}$.

–OR–

$\text{H}_0^-\colon Z \leq \varepsilon_{\text{lower}}$, or $Z \geq \varepsilon_{\text{upper}}$, with:

H_{01}^-: $Z \geq \varepsilon_{\text{upper}}$; or
H_{02}^-: $Z \leq \varepsilon_{\text{lower}}$.

NOTE: the appropriate level of $\alpha = (1-\texttt{conf.level})$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H^{-}_{01} and H^{-}_{02} by comparing the resulting p-value to 0.01 (Tryon and Lewis, 2008; Wellek, 2010).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., H^{+}_{0}: $p_x - p_y = 0$ or) and tests for equivalence (e.g., either H^{-}_{0}: $|p_x - p_y| \geq \Delta$, or H^{-}_{0}: $|Z| \geq \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H^{+}_{0}, but fail to reject H^{-}_{0}, and conclude that there is a **relevant difference** in proportions at least as large as Δ or ε.
2. One may fail to reject H^{+}_{0}, but reject H^{-}_{0}, and conclude that there is **equivalence** in proportions within the equivalence range (i.e. defined by Δ or ε).
3. One may reject both H^{+}_{0} and H^{-}_{0}, and conclude that there is a **trivial difference** in proportions which lies within the equivalence range (i.e. defined by Δ or ε).
4. One may fail to reject both H^{+}_{0} and H^{-}_{0}, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.pri` returns:

`statistics`	a vector of the *z* statistics for the two one-sided tests; if `relevance=TRUE`, these are followed by the value of the *z* statistic for the postivist test for difference.
`p.values`	a vector of *p* values for the *z* tests.
`proportion`	a scalar estimate of the sample proportion in the one-sample test. A vector of the proportions in both groups, as well as the estimate of the proportion under the null hypothesis in the two-sample test.
`sample_size`	a scalar containing the sample size of the one-sample test. A vector of the sample size in both groups, as well as the combined sample size in the two-sample test.
`threshold`	a scalar containing the equivalence threshold when `eqv.type="delta"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="delta"`. A scalar containing the equivalence threshold when `eqv.type="epsilon"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="epsilon"`.
`conclusion`	a string containing the relevance test conclusion when `relevance=TRUE`.

Author(s)

Alexis Dinno (<alexis.dinno@pdx.edu>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and

3. a copy of the exact output of the command.

I am endebted to my winter 2013 and fall 2023 students for their inspiration. Much appreciation to Mick McVeety for troubleshooting the translation of my Stata **tost** package to R.

Suggested citation: Dinno, A. 2025. **tost.pri**: Mean-equivalence z tests. In: **tost.suite** R software package.

References

Hauck, W. W. and S. Anderson. (1984) A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Pharmacodynamics*. **12**, 83–91.

Hauck, W. W. and Anderson, S. (1986) A comparison of large-sample confidence interval methods for the difference of two binomial probabilities. *The American Statistician*. **40**, 318–322.

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Tryon, W. W., and C. Lewis. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277

Yates, F. (1934) Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*. **1**, 217–235.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, second edition. Chapman and Hall/CRC Press. p. 31

See Also

prop.test, tost.pr.

Examples

```
# Immediate form of one-sample z test for proportion equivalence
# Note warning about value of Delta!
tost.pri(
    n1=50,
    obs1=.52,
    obs2=.70,
    eqv.type="delta",
    eqv.level=.1,
    relevance=FALSE)

# First two numbers are counts; equivalence interval is +/- 1 sd
# beyond the critical value of Z for alpha = 0.05
tost.pri(
    n1=30,
    obs1=4,
    obs2=.70,
    eqv.type="epsilon",
```

```
    eqv.level=qnorm(.95)+1,
    count=TRUE,
    conf.level=0.95,
    relevance=TRUE)

# Immediate form of two-sample z test for proportion equivalence using an
# example from Tu 1997, p 276, and incorporating the Hauck and Anderson
# continuity correction from that same example.
tost.pri(
    n1=101,
    obs1=.40594059,
    n2=100,
    obs2=.49,
    eqv.type="delta",
    eqv.level=.2,
    ccontinuity="ha",
    relevance=FALSE)

# The same example, but all numbers are counts
tost.pri(
    n1=101,
    obs1=41,
    n2=100,
    obs2=49,
    eqv.type="delta",
    eqv.level=.2,
    count=TRUE,
    ccontinuity="ha",
    relevance=FALSE)
```

`tost.rank.sum` *Two-sample rank sum test for stochastic equivalence*

Description

Performs two one-sided approximate z tests for stochastic equivalence between two independent samples.

Usage

```
tost.rank.sum(
    x, by,
    eqv.type     = equivalence.types,
    eqv.level    = 1,
    upper        = NA,
    conf.level   = 0.95,
    x.name       = "",
    by.name      = "",
    by.values    = NULL,
```

```
    ccontinuity  = FALSE,
    relevance    = TRUE)

equivalence.types
#c("delta", "epsilon")
```

Arguments

`x` a numeric vector of data values.

`by` a numeric or factor vector of exactly two values indicating group membership.

`eqv.type` defines whether the equivalence interval will be defined in terms of ε or Δ (`"epsilon"`, or `"delta"`). These options change the way that `evq.level` is interpreted: when `"epsilon"` is specified, the `evq.level` is measured in units of the z distribution, and when `"delta"` is specified, the `evq.level` is measured in the units of rank sums; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. Because units of rank sums is unlikely to be substantively meaningful, the default is `"epsilon"`.

Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \le z_\alpha$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that it is not possible to reject any H_0^- if $\Delta \le$ standard error $\times z_\alpha$. `tost.rank.sum` reports when either of these conditions obtain.

`eqv.level` defines the equivalence threshold for the tests depending on whether `eqv.type` is `"epsilon"` or `"delta"` (see above). Researchers are responsible for choosing meaningful values of ε or Δ. The default value is 1, which is not a useful value for either `eqv.type="delta"` or `eqv.type="epsilon"`.

`upper` defines the upper equivalence threshold for the test, is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper==abs(eqv.level)`, then `upper` will be ignored.

`conf.level` confidence level of the interval, and complement of the test's nominal type I error rate α.

`x.name` specifies how the outcome variable will be labeled in the output. The default value of `x.name` is the variable name of `x`.

`by.name` specifies how the grouping variable will be labeled in the output. The default value of `by.name` is the variable name of `by`.

`by.values` a string vector of exact two values specifying how group names will be labeled in the output. The default value of `by.names` are the factor labels or, if those are `NA` the factor levels of `by`.

`ccontinuity` calculates test statistics for both positivist and negativist tests using a continuity correction. For the positivist test the approximate statistic $z = \frac{\text{sgn}(W)\times(|W-\mu_W|-0.5)}{\sigma_W}$.

For the negativist test using ε the approximate test statsitics are $z_1 = \varepsilon_\text{u} - z$, and $z_2 = z - \varepsilon_\text{l}$ (where z is the continuity-corrected test statistic from the positivist test).

For the negativist test using Δ approximate statistics are $z_1 = \frac{\Delta_\text{u} - [\text{sgn}(W) \times (|W - \mu_W| - 0.5)]}{\sigma_W}$ and $z_2 = \frac{[\text{sgn}(W) \times (|W - \mu_W| - 0.5)] - \Delta_\text{l}}{\sigma_W}$.

`relevance` reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

`tost.rank.sum` tests the null hypothesis that the paired differences in measures are not symmetrically distributed and/or are not centered on the value of zero, and provides evidence for the distribution paired differences being equivalence to one that is symmetric and centered on zero. `tost.rank.sum` uses the z approximation to the rank sum test (Wilcoxon, 1945; Mann and Whitney, 1947) in a two one-sided tests approach (Schuirmann, 1987).

With respect to the rank sum test, a negativist null hypothesis takes one of the following two forms depending on whether tolerance is defined in terms of Δ (equivalence expressed in units of rank sums) or in terms of ε (equivalence expressed in the units of the z distribution):

H_0^-: $|W - \mu_W| \ge \Delta$,
where the equivalence interval ranges from $(W - \mu_W) - \Delta$ to $(W - \mu_W) + \Delta$ This translates directly into two one-sided null hypotheses:

H_{01}^-: $W - \mu_W \ge \Delta$, or
H_{02}^-: $W - \mu_W \le -\Delta$.

–OR–

H_0^-: $|Z| \ge \varepsilon$,
where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

H_{01}^-: $Z \ge \varepsilon$; or
H_{02}^-: $Z \le -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

H_0^-: $W - \mu_W \le \Delta_\text{l}$, or $W - \mu_W \ge \Delta_\text{u}$
where the equivalence interval ranges from $(W - \mu_W) + \Delta_\text{l}$ to $(W - \mu_W) + \Delta_\text{u}$. This also translates directly into two one-sided null hypotheses:

H_{01}^-: $W - \mu_W \ge \Delta_\text{u}$; or
H_{02}^-: $W - \mu_W \le \Delta_\text{l}$.

–OR–

H_0^-: $Z \le \varepsilon_\text{l}$, or $Z \ge \varepsilon_\text{u}$, with:

H_{01}^-: $Z \ge \varepsilon_\text{u}$; or
H_{02}^-: $Z \le \varepsilon_\text{l}$.

NOTE: the appropriate level of $\alpha = (1-$`conf.level`$)$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H_{01}^- and H_{02}^- by comparing the resulting p-value to 0.01 (Wellek, 2010).

Remarks: Following Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., H_0^+: $W - \mu_W = 0$ or) and tests for equivalence (e.g., either H_0^-: $|W - \mu_W| \geq \Delta$, or H_0^-: $|Z| \geq \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H_0^+, but fail to reject H_0^-, and conclude that there is **relevant 0th-order stochastic dominance** between the first and second groups which is at least as large as ε or Δ.
2. One may fail to reject H_0^+, but reject H_0^-, and conclude that there is **0th-order stochastic equivalence** between the first and second groups within the equivalence range (i.e. defined by ε or Δ).
3. One may reject both H_0^+ and H_0^-, and conclude that there is a **trivial 0th-order stochastic dominance** between the first and second groups which lies within the equivalence range (i.e. defined by ε or Δ).
4. One may fail to reject both H_0^+ and H_0^-, and draw an **indeterminate** conclusion, because the data are underpowered to detect either $0^{0\text{th}}$-order stochastic dominance or equivalence.

Value

`tost.rank.sum` returns:

`statistics`	a vector of the z statistics for the two one-sided tests; if `relevance=TRUE`, these are followed by the value of the z statistic for the postivist test for difference.
`p.values`	a vector of p values for the z tests.
`rank_sums`	a vector containing the rank sums in each group, and the rank sum expected under the positivist null hypothesis.
`sample_sizes`	a vector containing the sample sizes in both groups, as well as the combined sample size of both groups.
`var_adj`	a scalar containing the adjusted variance under the postivist null hypothesis.
`threshold`	a scalar containing the equivalence threshold when `eqv.type="delta"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="delta"`. A scalar containing the equivalence threshold when `eqv.type="epsilon"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="epsilon"`.
`conclusion`	a string containing the relevance test conclusion when `relevance=TRUE`.

Author(s)

Alexis Dinno (<`alexis.dinno@pdx.edu`>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

I am endebted to my winter 2013 and fall 2023 students for their inspiration. Much appreciation to Mick McVeety for troubleshooting the translation of my Stata **tost** package to R.

Suggested citation: Dinno, A. 2025. **tost.rank.sum**: Equivalence signed rank tests. In: **tost.suite** R software package.

References

Mann, H. B., and D. R. Whitney. (1947) On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**, 50–60.

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Snedecor, G. W., and W. G. Cochran. (1989) *Statistical Methods"*. 8th ed. Ames, IA: Iowa State University Press.

Tryon, W. W., and C. Lewis. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, Second edition. Chapman and Hall/CRC Press. p. 31.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin*. **1**, 80–83.

See Also

`tost.sign.rank`, `wilcox.test`, `Wilcoxon`.

Examples

```
require("webuse")

# Setup
webuse("fuel2")

# Perform two-sample rank-sum relevance test on mpg by using the two
# groups defined by treat; equivalence interval is +/- 1 sd beyond the
# critical value of Z for alpha = 0.1.
tost.rank.sum(
    x=fuel2$mpg,
    by=fuel2$treat,
    eqv.type="epsilon",
    eqv.level=qnorm(.9)+1,
    conf.level=.9,
    relevance=TRUE)

# Perform asymmetric rank-sum relevance test on mpg by using the two
# two groups defined by treat, and add a continuity correction.
# The lower end of the equivalence interval = qnorm(.9)+1=2.281552
# meaning equivalence must lay no more than 1 sd beyond the critical value
# of Z for alpha = 0.1.  The upper end of the equivalence interval
# = qnorm(.9)+1.5 = 1.781552 meaning equivalence must lay no more than
```

```
# 0.5 sd beyond the critical value of Z for alpha = 0.1.
tost.rank.sum(
    x=fuel2$mpg,
    by=fuel2$treat,
    eqv.type="epsilon",
    eqv.level=qnorm(.9)+1,
    upper=qnorm(.9)+.5,
    conf.level=.9,
    ccontinuity=TRUE,
    relevance=TRUE)
```

`tost.regress` *Linear regression tests for equivalence*

Description

Performs linear regression tests for equivalencee

Usage

```
tost.regress(
  formula,
  data        = NULL,
  eqv.type    = equivalence.types,
  eqv.level   = 1,
  upper       = NA,
  conf.level  = 0.95,
  relevance   = TRUE)

equivalence.types
#c("delta", "epsilon")
```

Arguments

`formula` — an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.

`data` — an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from `environment(formula)`, typically the environment from which `lm` is called.

`eqv.type` — either a single string (`"delta"`, or `"epsilon"`), or a vector of strings—one for each regression coefficient estimated; the first applying to the model constant term, and the remaining to each model variable in `formula` in order—which specifies whether the equivalence interval will be defined in terms of Δ or ε. If a single string, then each coefficient's equivalence region will use that definition. These options change the way that `evq.level` is interpreted: when `"delta"` is specified, the `evq.level` is measured in the units of the variable(s) being tested, and when `"epsilon"` is specified, the `evq.level` is measured in units of the *t* distribution; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. The default is `"delta"`.

Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \leq t_{\nu,\alpha}$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that it is not possible to reject any H_0^- if $\Delta \leq$ standard error $\times\, t_{\nu,\alpha}$. `tost.regress` reports when either of these conditions obtain.

`eqv.level` either a single numerical value, or a vector of numerical values—one for each regression coefficient estimated—defines the equivalence threshold for the tests depending on whether `eqv.type` is `"delta"` or `"epsilon"` (see `eqv.type` above). If a single value, then each coefficient's equivalence region will use that level. Researchers are responsible for choosing meaningful values of Δ or ε. The default value is 1.

`upper` either a single numerical value, or a vector of numerical values—one for each regression coefficient estimated—which defines the upper equivalence threshold for a coefficient's equivalence interval; is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper==abs(eqv.level)`, then `upper` will be ignored.

`conf.level` confidence level of the interval, and complement of the test's nominal type I error rate α.

`relevance` reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

`tost.regress` tests for the equivalence of each regression coefficient and zero within separate symmetric equivalence intervals defined by `eqv.type` and `eqv.level` for using a two one-sided *t* tests approach (Schuirmann, 1987). Typically ('positivist') null hypotheses are framed from an assumption of a lack of difference between two quantities, and reject this assumption only with sufficient evidence. When performing tests for equivalence, one frames a ('negativist') null hypothesis with the assumption that two quantities are different by at least as much as an equivalence interval defined by some chosen level of tolerance. **Note:** This version of `tost.regress` does not yet implement survey regression, bootstrap or jacknife estimation, or regression with robust or cluster standard errors, and currently implements only the simplest OLS functionality found in the Stata program `tostregress`.

An equivalence null hypothesis takes one of the following two forms depending on whether equivalence is defined in terms of Δ (equivalence expressed in the same units as the `x` and `y` varibales) or in terms of ϵ (equivalence expressed in the units of the *t* distribution with the given degrees of freedom):

$\text{H}_0^-\colon |\beta_x| \geq \Delta$,
where the equivalence interval ranges from $\beta_x - \Delta$ to $\beta_x + \Delta$ This translates directly into two one-sided null hypotheses:

H^-_{01}: $\beta_x \geq \Delta$, or
H^-_{02}: $\beta_x \leq -\Delta$.

–OR–

H^-_0: $|T| \geq \varepsilon$,
where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

H^-_{01}: $T \geq \varepsilon$; or
H^-_{02}: $T \leq -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

H^-_0: $\beta_x \leq \Delta_{\text{lower}}$, or $\beta_x \geq \Delta_{\text{upper}}$
where the equivalence interval ranges from $(\beta_x) + \Delta_{\text{lower}}$ to $(\beta_x) + \Delta_{\text{upper}}$. This also translates directly into two one-sided null hypotheses:

H^-_{01}: $\beta_x \geq \Delta_{\text{upper}}$; or
H^-_{02}: $\beta_x \leq \Delta_{\text{lower}}$.

–OR–

H^-_0: $T \leq \varepsilon_{\text{lower}}$, or $T \geq \varepsilon_{\text{upper}}$, with:

H^-_{01}: $T \geq \varepsilon_{\text{upper}}$; or
H^-_{02}: $T \leq \varepsilon_{\text{lower}}$.

NOTE: the appropriate level of $\alpha = (1-\texttt{conf.level})$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H^-_{01} and H^-_{02} by comparing the resulting p-value to 0.01 (Tryon and Lewis, 2008; Wellek, 2010).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., H^+_0: $\beta_x = 0$ or) and tests for equivalence (e.g., either H^-_0: $|\beta_x| \geq \Delta$, or H^-_0: $|T| \geq \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H^+_0, but fail to reject H^-_0, and conclude that there is a **relevant difference** in means at least as large as Δ or ε.
2. One may fail to reject H^+_0, but reject H^-_0, and conclude that there is **equivalence** in means within the equivalence range (i.e. defined by Δ or ε).
3. One may reject both H^+_0 and H^-_0, and conclude that there is a **trivial difference** in means which lies within the equivalence range (i.e. defined by Δ or ε).
4. One may fail to reject both H^+_0 and H^-_0, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.regress` returns:

`N`	the sample size.
`df_m`	the model degrees of freedom.
`df_r`	the residual degrees of freedom.

F	the F statistic.
r2	R^2.
rmse	root mean squared error.
mss	model sum of squares.
rss	residual sum of squares.
r2_a	adjusted R^2.
alpha	1 - conf.level.
T1	vector containing the value of the t_1 test statistics.
T2	vector containing the value of the t_2 test statistics.
T_pos	if relevance=TRUE a vector containing the value of the t test statistics for the positivist tests for the difference.
P1	vector of p values corresponding to the test statistics in T1.
P2	vector of p values corresponding to the test statistics in T2.
P_pos	if relevance=TRUE a vector of p values corresponding to the test statistics in T_pos.
SE	vector of estimated standard deviations of the regression coefficients corresponding to B, also corresponding to the square roots of the diagonal of V.
V	variance-covariance matrix corresponding to B.
Beta	vector of standardized regression coefficients corresponding to B, where the standardized coefficient for the effect of x on y is $\beta^*_x = \frac{s_x}{s_y}\beta_x$, and s_x is the sample standard deviation of x, s_y is the sample standard deviation of y, and β_x is the non-standardized coefficient of the effect of x on y.
thresholds_lower	vector containing the lower equivalence thresholds.
thresholds_upper	vector containing the upper equivalence thresholds.
conclusions	if relevance=TRUE a vector containing the relevance test conclusion string for a given α and the Δ or ε for the tests as specified for each coefficient.

Author(s)

Alexis Dinno (<alexis.dinno@pdx.edu>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

I am endebted to my winter 2013 and fall 2023 students for their inspiration. Much appreciation to Mick McVeety for troubleshooting the translation of my Stata **tost** package to R.

Suggested citation: Dinno, A. 2025. **tost.regress**: Linear regression tests for equivalence. In: **tost.suite** R software package.

References

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Tryon, W. W., and C. Lewis. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277

Wellek, S (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, second edition. Chapman and Hall/CRC Press. p. 31.

See Also

lm.

Examples

```
require("webuse")

# Setup
webuse("auto")

# Report equivalence tests for a linear regression; equivalence interval is
# +/- 1 sd beyond the critical value of T for alpha = 0.05 and df = 71, and
# where sd = sqrt(df/(df-2)).
tost.regress(
  auto$mpg ~ auto$weight + auto$foreign,
  eqv.type="epsilon",
  eqv.level=qt(.95, df=71)+1*sqrt(71/(71-2)),
  conf.level=0.95,
  relevance=FALSE)


# Report relevance tests for a linear regression; equivalence interval is
# +/- 1 sd beyond the critical value of T for alpha = 0.05 and df = 71.
tost.regress(
  auto$mpg ~ auto$weight + auto$foreign,
  eqv.type="epsilon",
  eqv.level=qt(.95, df=71)+1*sqrt(71/(71-2)),
  conf.level=0.95,
  relevance=TRUE)

# Setup
webuse("auto")
auto["gp100m"] <- 100/auto$mpg

# Fit a better linear regression, from a physics standpoint, but add
# asymmetric intervals, and report relevance test results.  The lower end of
# the equivalence interval = qt(.95, 71)+1.5*sqrt(71/(71-2)) = 3.188184 meaning
# eequivalence must lay no more than 1.5 sd beyond the critical value of T for
# alpha = 0.05 and df = 71.  The upper end of the equivalence interval =
# qt(.95, 71)+1*sqrt(71/(71-2)) = 2.680989 meaning equivalence must lay no more
```

```
# than 1 sd beyond the critical value of T for alpha = 0.05 and df = 71, and
# where sd = sqrt(df/(df-2)).gp100m <- 100/auto$mpg
tost.regress(
  auto$gp100m ~ auto$weight + auto$foreign,
  eqv.type="epsilon",
  eqv.level=qt(.95, df=71)+1.5*sqrt(71/(71-2)),
  upper=qt(.95, df=71)+1*sqrt(71/(71-2)),
  conf.level=0.95,
  relevance=TRUE)

# Obtain standardized regression coefficients from the above model
tost.regress(
  auto$gp100m ~ auto$weight + auto$foreign,
  eqv.type="epsilon",
  eqv.level=qt(.95, df=71)+1.5*sqrt(71/(71-2)),
  upper=qt(.95, df=71)+1*sqrt(71/(71-2)),
  conf.level=0.95,
  relevance=TRUE)$Beta

# Report equivalence tests when suppressing the intercept term
tost.regress(
  auto$weight ~ 0 + auto$length,
  eqv.type="delta",
  eqv.level=5,
  conf.level=0.95,
  relevance=FALSE)

# Report equivalence tests when the model already has constant; express
# equivalence interval in units of the variable only for length, and in units
# of the test statistic for each level of foreign. For the latter, the
# equivalence interval is +/- 1 sd beyond the critical value of T for
# alpha = 0.05.
tost.regress(
  auto$weight ~ 0 + auto$length + as.factor(auto$foreign),
  eqv.type=c("delta", "epsilon", "epsilon"),
  eqv.level=c(5, qt(.95, 71)+1*sqrt(71/(71-2)), qt(.95, 71)+1*sqrt(71/(71-2))),
  conf.level=0.95,
  relevance=FALSE)
```

---

tost.rrp *Test for equivalence of relative risk and unity in paired binary data*

Description

Performs two one-sided z tests for equivalence of marginal probabilities in binary data following Tang, Tang, and Chan, 2003

Usage

```
tost.rrp(
  x=NA, y=NA,
  delta0       = 1,
  deltaupper   = NA,
  exact.chisq  = FALSE,
  conf.level   = 0.95,
  treatment1   = "",
  treatment2   = "",
  outcome      = "",
  nooutcome    = "",
  relevance    = TRUE)
```

Arguments

`x` a (non-empty) vector of binary data values of equal length to `y`. The order of observations in `x` is assumed to correspond to the order of observations in `y` (i.e. `x` and `y` are paired.

`y` a (non-empty) vector of binary data values of equal length to `x`. The order of observations in `y` is assumed to correspond to the order of observations in `x` (i.e. `x` and `y` are paired.

`delta0` a required real value between 0 and 1 defining the lower threshold of an equivalence interval around RR=1. The upper boundary is `1/delta0`, unless `deltaupper` is used to define an assymetric upper interval. The default value is `delta0=1` which is not a useful value.

`deltaupper` an optional value greater than 1 which is other than `1/delta0` and which creates a geometrically asymmetric equivalence interval.

`exact.chisq` indicates that Fisher's exact p-value will be used for the positivist test (i.e. for McNemar's χ^2 test). This probability is calculated as $2\sum_{i=0}^{\min b,c}$ Binomial $(n = b + c, k = i, p = 0.5)$.

`conf.level` confidence level of the interval, and complement of the test's nominal type I error rate α.

`treatment1` an optional string to label the first treatment group in the output (e.g., "Treated"). If unspecified, `tost.rrp` will create a label from the `x` variable's label, names, or variable name (in that order).

`treatment2` an optional string to label the second treatment group in the output (e.g., "Untreated"). If unspecified, `tost.rrp` will create a label from the `y` variable's label, names, or variable name (in that order).

`outcome` an optional string to label those with the outcome (e.g., "Cases"). If unspecified `tost.rrp` will use the label "Positive".

`nooutcome` an optional string to label those without the outcome (e.g., "Not cases"). If unspecified `tost.rrp` will use the label "Negative".

`relevance` reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `delta0`, and, if used, `deltaupper`. See the Remarks section more details on inference from combined tests.

Details

`tost.rrp` tests for equivalence of the relative risk of a positive outcome and unity in paired (or matched) randomized control trial or paired (or matched) cohort design data. It calculates an asymptotic z test statistic based on a reparameterized multinomial model (Tang, et al., 2003) in a two one-sided tests approach (Schuirmann, 1987). The equivalence interval for the test is defined by a chosen level of tolerance, as specified by `delta0`.

The two one-sided null hypotheses take on the following form based on the relative risk (RR), and the threshold `delta0`:

H^-_{01}: RR $\le \delta_0$, or

H^-_{02}: RR $\ge \frac{1}{\delta_0}$.

where the equivalence interval ranges from δ_0 to $\frac{1}{\delta_0}$.

When a geometrically asymmetric equivalence interval is defined using the `deltaupper` option the two one-sided null hypotheses become:

H^-_{01}: RR $\le \delta_0$, or

H^-_{02}: RR $\ge \delta_{\text{upper}}$.

where the equivalence interval ranges from δ_0 to δ_{upper}.

The two z test statistics, z_1 and z_2, are both constructed with rejection probabilities in the upper tails. So $p_1 = P(Z \ge z_1)$, and $p_2 = P(Z \ge z_2)$.

NOTES: When $\delta_0 = 1$, the Tang-Tang-Chan test statistic reduces to McNemar's χ^2 test statistic (McNemar, 1947). When $a = b = c = 0$, there are no positve outcomes in either treatment group, and the RR and test statistics become undefined. If $a > 0$, and $b = c = 0$, then there is complete concordance, and $z_1 = z_2$, so $p_1 = p_2$. As is standard with two one-sided tests for equivalence, if one wishes to make a type I error %5 of the time, one simply conducts both of the one-sided tests of H^-_{01} and H^-_{02} by comparing the resulting p-values to 0.05 (Wellek, 2010).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (i.e. H^+_0: RR $= 1$) and tests for equivalence (i.e. H^-_{01}: RR $\le \delta_0$, or H^-_{02}: RR $\ge \frac{1}{\delta_0}$) are combined, there are four possible interpretations for a given α and δ_0:

1. One may reject H^+_0, but fail to reject both H^-_{01} and H^-_{02}, and conclude that there is a **relevant difference** between RR and 1 at least as large as the interval defined by δ_0.
2. One may fail to reject H^+_0, but reject both H^-_{01} and H^-_{02}, and conclude that there is **equivalence** between RR and 1 within the interval defined by δ_0.
3. One may reject H^+_0 and reject both H^-_{01} and H^-_{02}, and conclude that there is a **trivial difference** between RR and 1 which lies within the interval defined by δ_0.
4. One may fail to reject H^+_0 and fail to reject both H^-_{01} and H^-_{02}, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.rrp` returns:

`statistics` a vector containing the value of z_1 and z_2; if `relevance=TRUE`; these are followed by the value of the χ^2 statistic for the postivist test for difference.

`p.values` a vector of *p* values for the *z* tests, and, if `relevance=TRUE`, for the χ^2 test.

`estimate`	the estimated relative risk (aka incidence rate ratio) of positive outcome for treatment 2 vs. treatment 1.
`error`	the estimated standard deviation of relative risk based on the score statistic per (Tang, et al., 2003).
`threshold`	a scalar (δ_0) containing the equivalence threshold when `deltaupper=NA`. A vector (δ_l, δ_u) containing the asymmetric equivalence thresholds `delta0`, and `deltaupper`.
`conclusion`	relevance test conclusion for a given α and δ_0, or δ_l and δ_u.

Author(s)

Alexis Dinno (<alexis.dinno@pdx.edu>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

Suggested citation: Dinno, A. 2025. **tost.rrp**: Test for equivalence of relative risk and unity in paired binary data. In: **tost.suite** R software package.

References

Lachenbruch, P. A. and Lynch, C. J. (1998) Assessing screening tests: Extensions of McNemar's test. *Statistics In Medicine* **17**, 2207–2217.

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680.

Tang, N.-S., Tang, M.-L., and Chan, I. S. F. (2003) On tests of equivalence via non-unity relative risk for matched-pair design. *Statistics In Medicine* **22**, 1217–1233.

Tryon, W. W., and C. Lewis. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods* **13**, 272–277.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, second edition. Chapman and Hall/CRC Press. p. 31

See Also

`mcnemar.test`, `tost.rrpi`.

Examples

```
# Setup
data(hivfluid)

# Relevance test example from Tang, et al., 2003, Table II, based on data from
# Lachenbruch and Lynch, 1998 with equivalence interval .95 to 1.052632
#  (1/.95 = 1.052632)
tost.rrp(
  x=hivfluid$plasma,
  y=hivfluid$alternate,
  delta0=.95,
  outcome="HIV Positive",
  nooutcome="HIV Negative",
  relevance=TRUE)
```

`tost.rrpi` *Immediate test for equivalence of relative risk and unity in paired binary data*

Description

Immediately performs two one-sided z tests for equivalence of marginal probabilities in binary data following Tang, Tang, and Chan, 2003

Usage

```
tost.rrpi(
  a = NA, b = NA, c = NA, n = NA,
  delta0      = 1,
  deltaupper  = NA,
  exact.chisq = FALSE,
  conf.level  = 0.95,
  treatment1  = "",
  treatment2  = "",
  outcome     = "",
  nooutcome   = "",
  relevance   = TRUE)
```

Arguments

`a`	a non-negative integer indicating the number of paired observations with both first treatment and second treatment are positive for the outcome.
`b`	a non-negative integer indicating the number of paired observations with first treatment negative and second treatment positive for the outcome.
`c`	a non-negative integer indicating the number of paired observations with first treatment positive and second treatment negative for the outcome.

n — a non-negative integer indicating the total number of paired observations. $n = a + b + c + d$ (d, which is not directly provided, equals $n - a - b - c$).

delta0 — a required real value between 0 and 1 defining the lower threshold of an equivalence interval around RR=1. The upper boundary is `1/delta0`, unless `deltaupper` is used to define an assymetric upper interval. The default value is `delta0=1` which is not a useful value.

deltaupper — an optional value greater than 1 which is other than `1/delta0` and which creates a geometrically asymmetric equivalence interval.

exact.chisq — indicates that Fisher's exact p-value will be used for the positivist test (i.e. for McNemar's χ^2 test). This probability is calculated as $2\sum_{i=0}^{\min b,c}$ Binomial $(n = b + c, k = i, p = 0.5)$.

conf.level — confidence level of the interval, and complement of the test's nominal type I error rate α.

treatment1 — an optional string to label the first treatment group in the output (e.g., "Treated"). If unspecified, `tost.rrp` will create a label from the `x` variable's label, names, or variable name (in that order).

treatment2 — an optional string to label the second treatment group in the output (e.g., "Untreated"). If unspecified, `tost.rrp` will create a label from the `y` variable's label, names, or variable name (in that order).

outcome — an optional string to label those with the outcome (e.g., "Cases"). If unspecified `tost.rrp` will use the label "Positive".

nooutcome — an optional string to label those without the outcome (e.g., "Not cases"). If unspecified `tost.rrp` will use the label "Negative".

relevance — reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `delta0`, and, if used, `deltaupper`. See the Remarks section more details on inference from combined tests.

Details

Immediate commands perfom tests given summary statistics, rather than given data. `tost.rrpi` tests for equivalence of the relative risk of a positive outcome and unity in paired (or matched) randomized control trial or paired (or matched) cohort design data. It calculates an asymptotic z test statistic based on a reparameterized multinomial model (Tang, et al., 2003) in a two one-sided tests approach (Schuirmann, 1987). `tost.rrp` is the non-immediate form of `tost.rrpi`. The equivalence interval for the test is defined by a chosen level of tolerance, as specified by `delta0`.

The two one-sided null hypotheses take on the following form based on the relative risk (RR), and the threshold `delta0`:

H_{01}^{-}: RR $\le \delta_0$, or

H_{02}^{-}: RR $\ge \frac{1}{\delta_0}$.

where the equivalence interval ranges from δ_0 to $\frac{1}{\delta_0}$.

When a geometrically asymmetric equivalence interval is defined using the `deltaupper` option the two one-sided null hypotheses become:

H_{01}^{-}: RR $\le \delta_0$, or

H_{02}^{-}: RR $\ge \delta_{\text{upper}}$.

where the equivalence interval ranges from δ_0 to δ_{upper}.

The two z test statistics, z_1 and z_2, are both constructed with rejection probabilities in the upper tails. So $p_1 = P(Z \geq z_1)$, and $p_2 = P(Z \geq z_2)$.

NOTES: When $\delta_0 = 1$, the Tang-Tang-Chan test statistic reduces to McNemar's χ^2 test statistic (McNemar, 1947). When $a = b = c = 0$, there are no positve outcomes in either treatment group, and the RR and test statistics become undefined. If $a > 0$, and $b = c = 0$, then there is complete concordance, and $z_1 = z_2$, so $p_1 = p_2$. As is standard with two one-sided tests for equivalence, if one wishes to make a type I error %5 of the time, one simply conducts both of the one-sided tests of H^-_{01} and H^-_{02} by comparing the resulting p-values to 0.05 (Wellek, 2010).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (i.e. H^+_0: RR $= 1$) and tests for equivalence (i.e. H^-_{01}: RR $\leq \delta_0$, or H^-_{02}: RR $\geq \frac{1}{\delta_0}$) are combined, there are four possible interpretations for a given α and δ_0:

1. One may reject H^+_0, but fail to reject both H^-_{01} and H^-_{02}, and conclude that there is a **relevant difference** between RR and 1 at least as large as the interval defined by δ_0.
2. One may fail to reject H^+_0, but reject both H^-_{01} and H^-_{02}, and conclude that there is **equivalence** between RR and 1 within the interval defined by δ_0.
3. One may reject H^+_0 and reject both H^-_{01} and H^-_{02}, and conclude that there is a **trivial difference** between RR and 1 which lies within the interval defined by δ_0.
4. One may fail to reject H^+_0 and fail to reject both H^-_{01} and H^-_{02}, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.rrpi` returns:

`statistics`	a vector containing the value of z_1 and z_2; if `relevance=TRUE`; these are followed by the value of the χ^2 statistic for the postivist test for difference.
`p.values`	a vector of p values for the z tests, and, if `relevance=TRUE`, for the χ^2 test.
`estimate`	the estimated relative risk (aka incidence rate ratio) of positive outcome for treatment 2 vs. treatment 1.
`error`	the estimated standard deviation of relative risk based on the score statistic per (Tang, et al., 2003).
`threshold`	a scalar (δ_0) containing the equivalence threshold when `deltaupper=NA`. A vector (δ_l, δ_u) containing the asymmetric equivalence thresholds `delta0`, and `deltaupper`.
`conclusion`	relevance test conclusion for a given α and δ_0, or δ_l and δ_u.

Author(s)

Alexis Dinno (<`alexis.dinno@pdx.edu`>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),

2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

Suggested citation: Dinno, A. 2025. **tost.rrpi**: Test for equivalence of relative risk and unity in paired binary data. In: **tost.suite** R software package.

References

Lachenbruch, P. A. and Lynch, C. J. (1998) Assessing screening tests: Extensions of McNemar's test. *Statistics In Medicine* **17**, 2207–2217.

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680.

Tang, N.-S., Tang, M.-L., and Chan, I. S. F. (2003) On tests of equivalence via non-unity relative risk for matched-pair design. *Statistics In Medicine* **22**, 1217–1233.

Tryon, W. W., and C. Lewis. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods* **13**, 272–277.

Tango, T. (1998) Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics In Medicine* **17**, 891–908.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, second edition. Chapman and Hall/CRC Press. p. 31

See Also

`mcnemar.test`, `tost.rrp`.

Examples

```
# Same as the relevance test example from Tang, et al., 2003, Table II in
# tost.rpp, based on data from Lachenbruch and Lynch, 1998 with equivalence
# interval .95 to 1.052632, but using the immediate command.
tost.rrpi(a=446, b=5, c=16, n=1157,
  delta0=.95,
  treatment1="Plasma sample",
  treatment2="Alternate fluid",
  outcome="HIV Positive",
  nooutcome="HIV Negative",
  relevance=TRUE)

# Same as above, but using the exact p-value for the positivist test.
# Positivist test and relevance test conclusions change
tost.rrpi(a=446, b=5, c=16, n=1157,
  delta0=.95,
```

```
  treatment1="Plasma sample",
  treatment2="Alternate fluid",
  outcome="HIV Positive",
  nooutcome="HIV Negative",
  exact.chisq=TRUE,
  relevance=TRUE)

# Example from Tang, et al., 2003, Table V, based on data from Tango, 1998
# Using exact.chisq=TRUE because expected counts are tiny in some cells
tost.rrpi(a=43, b=0, c=1, n=44,
  delta0=.9,
  treatment1="Thermal",
  treatment2="Chemical",
  outcome="Effective",
  nooutcome="Ineffective",
  exact.chisq=TRUE,
  relevance=FALSE)
```

`tost.sign.rank` *Test for the distribution of paired or matched data being equivalent to one that is symmetrical & centered on zero*

Description

Performs two one-sided approximate z tests for equivalence between the distribution of paired differences and a distribution which is both symmetric and centered on zero.

Usage

```
tost.sign.rank(
  x, y,
  eqv.type     = equivalence.types,
  eqv.level    = 1,
  upper        = NA,
  ccontinuity  = FALSE,
  conf.level   = 0.95,
  x.name       = "",
  y.name       = "",
  relevance    = TRUE)

equivalence.types
#c("delta", "epsilon")
```

Arguments

x	a (non-empty) numeric vector of data values.
y	a (non-empty) numeric vector of data values.

`eqv.type` defines whether the equivalence interval will be defined in terms of ε or Δ (`"epsilon"`, or `"delta"`). These options change the way that `evq.level` is interpreted: when `"epsilon"` is specified, the `evq.level` is measured in units of the z distribution, and when `"delta"` is specified, the `evq.level` is measured in the units of the absolute value of sums of signed ranks of paired differences; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. Because units of absolute value of sums of signed ranks of paired differences is unlikely to be substantively meaningful, the default is `"epsilon"`.

Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \le z_\alpha$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that it is not possible to reject any H_0^- if $\Delta \le$ standard error $\times z_\alpha$. `tost.sign.rank` reports when either of these conditions obtain.

`eqv.level` defines the equivalence threshold for the tests depending on whether `eqv.type` is `"epsilon"` or `"delta"` (see above). Researchers are responsible for choosing meaningful values of ε or Δ. The default value is 1, which is not a useful value for either `eqv.type="delta"` or `eqv.type="epsilon"`.

`upper` defines the upper equivalence threshold for the test, is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper==abs(eqv.level)`, then `upper` will be ignored.

`ccontinuity` calculates test statistics for both positivist and negativist tests using a continuity correction. For the positivist test the approximate statistic $z = \frac{\text{sgn}(T)\times(|T-\mu_T|-0.5)}{\sigma_T}$.

For the negativist test using ε the approximate test statsitics are $z_1 = \varepsilon_{\text{u}} - z$, and $z_2 = z - \varepsilon_{\text{l}}$ (where z is the continuity-corrected test statistic from the positivist test).

For the negativist test using Δ approximate statistics are $z_1 = \frac{\Delta_{\text{u}} - [\text{sgn}(T)\times(|T-\mu_T|-0.5)]}{\sigma_T}$ and $z_2 = \frac{[\text{sgn}(T)\times(|T-\mu_T|-0.5)]-\Delta_{\text{l}}}{\sigma_T}$.

`conf.level` confidence level of the interval, and complement of the test's nominal type I error rate α.

`x.name` specifies how the first variable will be labeled in the output. The default value of `x.name` is the variable name of `x`.

`y.name` specifies how the second variable will be labeled in the output. The default value of `y.name` is the variable name of `y`.

`relevance` reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

`tost.sign.rank` tests the null hypothesis that the paired differences in measures are not symmetrically distributed and/or are not centered on the value of zero, and provides evidence for the distribution paired differences being equivalence to one that is symmetric and centered on zero. `tost.sign.rank` uses the z approximation to the Wilcoxon matched-pairs signed-ranks test (Wilcoxon 1945) in a two one-sided tests approach (Schuirmann, 1987).

With respect to the signed-rank test, a negativist null hypothesis takes one of the following two forms depending on whether tolerance is defined in terms of Δ (equivalence expressed in the same units as the absolute value of sums of signed ranks) or in terms of ε (equivalence expressed in the units of the z distribution):

$\text{H}_0^-\colon |T - \mu_T| \geq \Delta$,
where the equivalence interval ranges from $(T - \mu_T) - \Delta$ to $(T - \mu_T) + \Delta$ This translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\colon T - \mu_T \geq \Delta$, or
$\text{H}_{02}^-\colon T - \mu_T \leq -\Delta$.

–OR–

$\text{H}_0^-\colon |Z| \geq \varepsilon$,
where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\colon Z \geq \varepsilon$; or
$\text{H}_{02}^-\colon Z \leq -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

$\text{H}_0^-\colon T - \mu_T \leq \Delta_\text{l}$, or $T - \mu_T \geq \Delta_\text{u}$
where the equivalence interval ranges from $(T - \mu_T) + \Delta_\text{l}$ to $(T - \mu_T) + \Delta_\text{u}$. This also translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\colon T - \mu_T \geq \Delta_\text{u}$; or
$\text{H}_{02}^-\colon T - \mu_T \leq \Delta_\text{l}$.

–OR–

$\text{H}_0^-\colon Z \leq \varepsilon_\text{l}$, or $Z \geq \varepsilon_\text{u}$, with:

$\text{H}_{01}^-\colon Z \geq \varepsilon_\text{u}$; or
$\text{H}_{02}^-\colon Z \leq \varepsilon_\text{l}$.

NOTE: the appropriate level of $\alpha = (1-\texttt{conf.level})$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H_{01}^- and H_{02}^- by comparing the resulting p-value to 0.01 (Wellek, 2010).

Remarks: Following Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., $\text{H}_0^+\colon T - \mu_T = 0$ or) and tests for equivalence (e.g., either $\text{H}_0^-\colon |T - \mu_T| \geq \Delta$, or $\text{H}_0^-\colon |Z| \geq \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H_0^+, but fail to reject H_0^-, and conclude that there is a **relevant difference** between the distribution of paired differences and a distribution which is both symmetric and centered on zero which is at least as large as ε or Δ.
2. One may fail to reject H_0^+, but reject H_0^-, and conclude that there is **equivalence** between the distribution of paired differences and a distribution which is both symmetric and centered on zero within the equivalence range (i.e. defined by ε or Δ).

3. One may reject both H_0^+ and H_0^-, and conclude that there is a **trivial difference** between the distribution of paired differences and a distribution which is both symmetric and centered on zero which lies within the equivalence range (i.e. defined by ε or Δ).
4. One may fail to reject both H_0^+ and H_0^-, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.sign.rank` returns:

`statistics` a vector of the z statistics for the two one-sided tests; if `relevance=TRUE`, these are followed by the value of the z statistic for the postivist test for difference.

`p.values` a vector of p values for the z tests.

`signed_rank_sums` a vector containing the absolute value of positive and negative rank sums, and the signed rank sum expected under the positivist null hypothesis.

`sample_size` a scalar containing the sample size.

`counts` a vector containing the number of negative comparisons, number of positive comparisons, and number of tied comparisons.

`var_adj` a scalar containing the adjusted variance under the postivist null hypothesis.

`threshold` a scalar containing the equivalence threshold when `eqv.type="delta"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="delta"`. A scalar containing the equivalence threshold when `eqv.type="epsilon"` and `upper=NA`. A vector containing the asymmetric equivalence thresholds `upper`, and `eqv.level` when `eqv.type="epsilon"`.

`conclusion` a string containing the relevance test conclusion when `relevance=TRUE`.

Author(s)

Alexis Dinno (<alexis.dinno@pdx.edu>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
help
3. a copy of the exact output of the command.

Much appreciation to Mick McVeety for troubleshooting the translation of my Stata **tost** package to R.

Suggested citation: Dinno, A. 2025. **tost.sign.rank**: Test for the distribution of paired or matched data being equivalent to one that is symmetrical & centered on zero. In: **tost.suite** R software package.

References

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Snedecor, G. W., and W. G. Cochran. (1989) *Statistical Methods*. 8th ed. Ames, IA: Iowa State University Press.

Tryon, W. W., and Lewis, C. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority*, Second edition. Chapman and Hall/CRC Press. p. 31.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin*. **1**, 80–83.

See Also

`SignRank`, `tost.rank.sum`, `wilcox.test`.

Examples

```
require("webuse")

#Setup
webuse("fuel")

# Perform sign-rank relevance test between mpg1 and mpg2; equivalence
# interval is +/- 1.5 sd beyond the critical value of Z for alpha = 0.05.
tost.sign.rank(
  fuel$mpg1,
  fuel$mpg2,
  eqv.type="epsilon",
  eqv.level=qnorm(.95)+1.5,
  relevance=TRUE)

# Same example, but using an asymmetric equivalence interval and continuity
# correction.  The lower end of the equivalence interval = qnorm(.95)+1.5
# = 3.144854 meaning equivalence must lay no more than 1.5 sd beyond the
# critical value of Z for alpha = 0.05.  The upper end of the equivalence
# interval = qnorm(.95)+1 = 2.644854 meaning equivalence must lay
# no more than 1 sd beyond the critical value of Z for alpha = 0.05.
tost.sign.rank(
  fuel$mpg1,
  fuel$mpg2,
  eqv.type="epsilon",
  eqv.level=qnorm(.95)+1.5,
  upper=qnorm(.95)+1,
  ccontinuity=TRUE,
  relevance=TRUE)
```

tost.t *Mean-equivalence* t *tests*

Description

Performs two one-sided *t* tests for mean equivalence

Usage

```
tost.t(
  x,
  y           = NULL,
  mu          = NA,
  by          = NULL,
  eqv.type    = equivalence.types,
  eqv.level   = 1,
  upper       = NA,
  paired      = FALSE,
  var.equal   = FALSE,
  welch       = FALSE,
  conf.level  = 0.95,
  x.name      = "",
  y.name      = "",
  by.name     = "",
  by.values   = NULL,
  relevance   = TRUE)

equivalence.types
#c("delta", "epsilon")
```

Arguments

x	a (non-empty) numeric vector of data values.
y	an optional (non-empty) numeric vector of data values. Implies `by=NULL`.
mu	a number indicating the true value of the mean for a one-sample test. Implies `paired=FALSE`, and `y=NULL`.
by	an optional (non-empty) vector of group indicator values. Implies `y=NA`.
eqv.type	defines whether the equivalence interval will be defined in terms of Δ or ε (`"delta"`, or `"epsilon"`). These options change the way that `evq.level` is interpreted: when `"delta"` is specified, the `evq.level` is measured in the units of the variable(s) being tested, and when `"epsilon"` is specified, the `evq.level` is measured in units of the *t* distribution; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. The default is `"delta"`. Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \le t_{\nu,\alpha}$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that

it is not possible to reject any H_0^- if $\Delta \leq$ standard error $\times$ $t_{\nu,\alpha}$. `tost.t` reports when either of these conditions obtain.

`eqv.level` defines the equivalence threshold for the tests depending on whether `eqv.type` is `"delta"` or `"epsilon"` (see above). Researchers are responsible for choosing meaningful values of Δ or ε. The default value is 1, which should not automatically be assumed to be a meaningful value for any given research question.

`upper` defines the upper equivalence threshold for the test, is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper==abs(eqv.level)`, then `upper` will be ignored.

`paired` a logical variable indicating whether you want a paired *t* test. Requires `y` to be supplied.

`var.equal` a logical variable indicating whether to treat the two samples as being drawn from populations with equal variances. If `var.equal=TRUE` the pooled variance is used with degrees of freedom $\nu = n_x + n_y - 2$, otherwise Satterthwaite' approximation to the degrees of freedom is used (unless `welch=TRUE` is specified).

`welch` a logical variable indicating `tost.t` should use Welch's (1947) approximation for the degrees of freedom will be used in an unpaired *t* test assuming unequal variances. Specifying `welch=TRUE` requires that `var.equal==FALSE`.

`conf.level` confidence level of the interval, and complement of the test's nominal type I error rate α.

`x.name` specifies how the first variable will be labeled in the output. The default value of `x.name` is `names(x)`, but if that is not present will use the variable name of `x`.

`y.name` specifies how the second variable will be labeled in the output when `by=NULL`. The default value of `y.name` is `names(y)`, but if that is not present will use the variable name of `y`. If `by!=NULL`, then information in `names(x)`, `x`, `names(by)`, `by`, `x.name`, `y.name`, and `by.values` will be used to label the two groups depending on what information is present in these objects.

`by.name` an optional string to customize the grouping variable name in the output. If `by.name=""`, `names(by)` or the name of the `by` variable will be used instead.

`by.values` an optional two-element character vector of group names. If none are supplied, the names of the values of `names(by)` will be used if present, otherwise the raw values of the `by` variable will be used.

`relevance` reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

`tost.t` tests for the equivalence of means within a symmetric equivalence interval defined by `eqv.type` and `eqv.level` using a two one-sided *t* tests (TOST) approach (Schuirmann, 1987). Typically "positivist" null hypotheses are framed from an assumption of a lack of difference between two quantities, and reject this assumption only with sufficient evidence. When performing tests for equivalence, one frames a null hypothesis with the assumption that two quantities are different within an equivalence interval defined by some chosen level of tolerance.

With respect to an unpaired *t* test, an equivalence null hypothesis takes one of the following two forms depending on whether equivalence is defined in terms of Δ (equivalence expressed in the same units as the `x` and `y` variables) or in terms of ϵ (equivalence expressed in the units of the *t* distribution with the given degrees of freedom):

$\text{H}_0^-\colon |\mu_x - \mu_y| \ge \Delta$,
where the equivalence interval ranges from $(\mu_x - \mu_y) - \Delta$ to $(\mu_x - \mu_y) + \Delta$ This translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\colon \mu_x - \mu_y \ge \Delta$, or
$\text{H}_{02}^-\colon \mu_x - \mu_y \le -\Delta$.

–OR–

$\text{H}_0^-\colon |T| \ge \varepsilon$,
where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\colon T \ge \varepsilon$; or
$\text{H}_{02}^-\colon T \le -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

$\text{H}_0^-\colon \mu_x - \mu_y \le \Delta_{\text{lower}}$, or $\mu_x - \mu_y \ge \Delta_{\text{upper}}$
where the equivalence interval ranges from $(\mu_x - \mu_y) + \Delta_{\text{lower}}$ to $(\mu_x - \mu_y) + \Delta_{\text{upper}}$. This also translates directly into two one-sided null hypotheses:

$\text{H}_{01}^-\colon \mu_x - \mu_y \ge \Delta_{\text{upper}}$; or
$\text{H}_{02}^-\colon \mu_x - \mu_y \le \Delta_{\text{lower}}$.

–OR–

$\text{H}_0^-\colon T \le \varepsilon_{\text{lower}}$, or $T \ge \varepsilon_{\text{upper}}$, with:

$\text{H}_{01}^-\colon T \ge \varepsilon_{\text{upper}}$; or
$\text{H}_{02}^-\colon T \le \varepsilon_{\text{lower}}$.

NOTE: the appropriate level of $\alpha = (1-\texttt{conf.level})$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H_{01}^- and H_{02}^- by comparing the resulting p-value to 0.01 (Tryon and Lewis, 2008).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., $\text{H}_0^+\colon \mu_x - \mu_y = 0$ or) and tests for equivalence (e.g., either $\text{H}_0^-\colon |\mu_x - \mu_y| \ge \Delta$, or $\text{H}_0^-\colon |T| \ge \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H_0^+, but fail to reject H_0^-, and conclude that there is a **relevant difference** in means at least as large as Δ or ε.
2. One may fail to reject H_0^+, but reject H_0^-, and conclude that there is **equivalence** in means within the equivalence range (i.e. defined by Δ or ε).
3. One may reject both H_0^+ and H_0^-, and conclude that there is a **trivial difference** in means which lies within the equivalence range (i.e. defined by Δ or ε).
4. One may fail to reject both H_0^+ and H_0^-, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.t` returns:

- `statistics` — a vector of the t statistics for the two one-sided tests; if `relevance=TRUE`, these are followed by the value of the t statistic for the postivist test for difference.
- `p.values` — a vector of p values for the t tests.
- `estimate` — a scalar or vector of the estimated mean or means, mean difference, or difference in means depending on whether it was a one-sample test, paired test, or a two-sample test.
- `null.value` — the specified hypothesized value of the mean in a one-sample test, or 0 for a paired test or two-sample test.
- `sterr` — the standard error used in the denominator of the t statistic.
- `sd` — a vector containing the sample standard deviations of the two variables or two groups in paired and unpaired tests; not returned for one-sample tests.
- `sample_size` — a scalar (one-sample test) or vector (two-sample tests) containing the number of observations in the variable(s).
- `parameter` — the degrees of freedom for the t statistics.
- `threshold` — the value of the equivalence/relevance threshold: if `upper==NA` then returns the `eqv.level` argument. If `upper!=NA`, then returns a vector of (`eqv.level`,`upper`)
- `conclusion` — a string containing the relevance test conclusion when `relevance=TRUE`.

Author(s)

Alexis Dinno (<alexis.dinno@pdx.edu>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

I am endebted to my winter 2013 and fall 2023 students for their inspiration. Much appreciation to Mick McVeety for troubleshooting the translation of my Stata **tost** package to R.

Suggested citation: Dinno, A. 2025. **tost.t**: Mean-equivalence t tests. In: **tost.suite** R software package.

References

Satterthwaite, F. E. (1946) An approximate distribution of estimates of variance components. *Biometrics Bulletin*. **2**, 110–114.

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Tryon, W. W., and Lewis, C. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277

Welch, B. L. (1947) The generalization of "Student's" problem when several different population variances are involved. *Biometrika*. **34**, 28–35.

See Also

t.test, tost.ti.

Examples

```
require("webuse")

# Setup
webuse("auto")

# One-sample mean equivalence t test with asymmetric equivalence interval
tost.t(
  x=auto$mpg,
  mu=20,
  eqv.type="delta",
  eqv.level=2.5,
  upper=3,
  relevance=FALSE)

# Setup
webuse("fuel")

# Two-sample paired relevance t test of means; equivalence interval is
#   +/- 1.5 sd beyond the critical value of T with df = 11 for alpha = 0.05
tost.t(
  x=fuel$mpg1,
  y=fuel$mpg2,
  paired=TRUE,
  eqv.type="epsilon",
  eqv.level=qt(p=.95,df=11)+1.5*sqrt(11/9),
  conf.level=0.95,
  relevance=TRUE)

# Setup
webuse("fuel3")

# Two-group unpaired mean equivalence t test assuming equal variances
#   Notice warning about value of Delta!
tost.t(
  x=fuel3$mpg,
  by=fuel3$treated,
```

```
    eqv.type="delta",
    eqv.level=1.5,
    var.equal=TRUE,
    relevance=FALSE)

  # Same example but customizing output labels
  tost.t(
    x=fuel3$mpg,
    by=fuel3$treated,
    eqv.type="delta",
    eqv.level=1.5,
    var.equal=TRUE,
    by.name="Fuel",
    by.values=c("Treated", "Untreated"),
    relevance=FALSE)
```

---

tost.ti *Immediate mean-equivalence* t *tests*

Description

Immediately performs two one-sided *t* tests for mean equivalence

Usage

```
tost.ti(
  n1=NA, mean1=NA, sd1=NA, mu=NA,
  n2=NA, mean2=NA, sd2=NA,
  eqv.type    = equivalence.types,
  eqv.level   = 1,
  upper       = NA,
  var.equal   = FALSE,
  welch       = FALSE,
  conf.level  = 0.95,
  x.name      = "",
  y.name      = "",
  relevance   = TRUE)

equivalence.types
#c("delta", "epsilon")
```

Arguments

n1	a required positive integer value representing the sample size in group 1.
mean1	a required real value representing the sample mean in group 1.
sd1	a required non-negative real value representing the sample standard deviation (not standard error) in group 1.

`mu` — an optional real value representing the true value of the mean under the positivist null hypothesis for a one-sample test. Implies `n2=NA`, `mean2=NA` and `sd2=NA`.

`n2` — an optional positive integer value representing the sample size in group 2. Implies `mu=NA`, and also that `mean2` and `sd2` are provided.

`mean2` — an optional real value representing the sample mean in group 2. Implies `mu=NA`, and also that `n2` and `sd2` are provided.

`sd2` — an optional non-negative real value representing the sample standard deviation (not standard error) in group 2. Implies `mu=NA`, and also that `n2` and `mean2` are provided.

`eqv.type` — defines whether the equivalence interval will be defined in terms of Δ or ε (`"delta"`, or `"epsilon"`). These options change the way that `evq.level` is interpreted: when `"delta"` is specified, the `evq.level` is measured in the units of the variable(s) being tested, and when `"epsilon"` is specified, the `evq.level` is measured in units of the *t* distribution; put another way $\varepsilon = \frac{\Delta}{\text{standard error}}$. The default is `"delta"`.

Defining tolerance in terms of ε means that it is not possible to reject any test for mean equivalence's H_0^- if $\varepsilon \le t_{\nu,\alpha}$. Because $\varepsilon = \frac{\Delta}{\text{standard error}}$, we can see that it is not possible to reject any H_0^- if $\Delta \le \text{standard error} \times t_{\nu,\alpha}$. `tost.ti` reports when either of these conditions obtain.

`eqv.level` — defines the equivalence threshold for the tests depending on whether `eqv.type` is `"delta"` or `"epsilon"` (see above). Researchers are responsible for choosing meaningful values of Δ or ε. The default value is 1, which should not automatically be assumed to be a meaningful value for any given research question.

`upper` — defines the upper equivalence threshold for the test, is assumed to be positive, and transforms the meaning of `eqv.level` to mean the *lower* equivalence threshold for the test. Also, `eqv.level` is assumed to be a negative value. Taken together, these correspond to Schuirmann's (1987) asymmetric equivalence intervals. If `upper=abs(eqv.level)`, then `upper` will be ignored.

`var.equal` — a logical variable indicating whether to treat the two samples as being drawn from populations with equal variances. If `var.equal=TRUE` the pooled variance is used with degrees of freedom $\nu = n_x + n_y - 2$, otherwise Satterthwaite' approximation to the degrees of freedom is used (unless `welch=TRUE` is specified).

`welch` — a logical variable indicating `tost.ti` should use Welch's (1947) approximation for the degrees of freedom will be used in an unpaired *t* test assuming unequal variances. Specifying `welch=TRUE` requires that `var.equal=FALSE`.

`conf.level` — confidence level of the interval, and complement of the test's nominal type I error rate α.

`x.name` — specifies how the first variable will be labeled in the output. The default value of `x.name` is `names(x)`, but if that is not present `tost.ti` will use the variable name of `x`.

`y.name` — specifies how the second variable will be labeled in the output when `by=NULL`. The default value of `y.name` is `names(y)`, but if that is not present `tost.ti` will use the variable name of `y`. If `by!=NULL`, then information in `names(x)`, `x`, `names(by)`, `by`, `x.name`, `y.name`, and `by.values` will be used to label the two groups depending on what information is present in these objects.

`relevance` reports results and inference for combined tests for difference and for equivalence for a specific `conf.level`, `eqv.type`, `eqv.level`, and, if used, `upper`. See the Remarks section more details on inference from combined tests.

Details

Immediate commands perfom tests given summary statistics, rather than given data. `tost.ti` tests for the equivalence of means within a symmetric equivalence interval defined by `eqv.type` and `eqv.level` using a two one-sided *t* tests (TOST) approach (Schuirmann, 1987). Typically "positivist" null hypotheses are framed from an assumption of a lack of difference between two quantities, and reject this assumption only with sufficient evidence. When performing tests for equivalence, one frames a null hypothesis with the assumption that two quantities are different within an equivalence interval defined by some chosen level of tolerance.

With respect to an unpaired *t* test, an equivalence null hypothesis takes one of the following two forms depending on whether equivalence is defined in terms of Δ (equivalence expressed in the same units as `mean1` and `mean2`) or in terms of ϵ (equivalence expressed in the units of the *t* distribution with the given degrees of freedom):

$\text{H}_0^-: |\mu_x - \mu_y| \geq \Delta$,
where the equivalence interval ranges from $(\mu_x - \mu_y) - \Delta$ to $(\mu_x - \mu_y) + \Delta$ This translates directly into two one-sided null hypotheses:

H_{01}^-: $\mu_x - \mu_y \geq \Delta$, or
H_{02}^-: $\mu_x - \mu_y \leq -\Delta$.

–OR–

$\text{H}_0^-: |T| \geq \varepsilon$,
where the equivalence interval ranges from $-\varepsilon$ to ε. This also translates directly into two one-sided null hypotheses:

H_{01}^-: $T \geq \varepsilon$; or
H_{02}^-: $T \leq -\varepsilon$.

When an asymmetric equivalence interval is defined using the `upper` option the general negativist null hypothesis becomes:

$\text{H}_0^-: \mu_x - \mu_y \leq \Delta_{\text{lower}}$, or $\mu_x - \mu_y \geq \Delta_{\text{upper}}$
where the equivalence interval ranges from $(\mu_x - \mu_y) + \Delta_{\text{lower}}$ to $(\mu_x - \mu_y) + \Delta_{\text{upper}}$. This also translates directly into two one-sided null hypotheses:

H_{01}^-: $\mu_x - \mu_y \geq \Delta_{\text{upper}}$; or
H_{02}^-: $\mu_x - \mu_y \leq \Delta_{\text{lower}}$.

–OR–

$\text{H}_0^-: T \leq \varepsilon_{\text{lower}}$, or $T \geq \varepsilon_{\text{upper}}$, with:

H_{01}^-: $T \geq \varepsilon_{\text{upper}}$; or
H_{02}^-: $T \leq \varepsilon_{\text{lower}}$.

NOTE: the appropriate level of $\alpha = (1-$`conf.level`$)$ is precisely the same as in the corresponding two-sided test for mean difference, so that, for example, if one wishes to make a type I error %1 of the time, one simply conducts both of the one-sided tests of H_{01}^- and H_{02}^- by comparing the resulting p-value to 0.01 (Tryon and Lewis, 2008).

Remarks: As described by Tryon and Lewis (2008), when rejection decisions from both tests for difference (e.g., $\text{H}_0^+\colon \mu_x - \mu_y = 0$ or) and tests for equivalence (e.g., either $\text{H}_0^-\colon |\mu_x - \mu_y| \geq \Delta$, or $\text{H}_0^-\colon |T| \geq \varepsilon$) are combined, there are four possible interpretations for a given α and Δ or ε:

1. One may reject H_0^+, but fail to reject H_0^-, and conclude that there is a **relevant difference** in means at least as large as Δ or ε.
2. One may fail to reject H_0^+, but reject H_0^-, and conclude that there is **equivalence** in means within the equivalence range (i.e. defined by Δ or ε).
3. One may reject both H_0^+ and H_0^-, and conclude that there is a **trivial difference** in means which lies within the equivalence range (i.e. defined by Δ or ε).
4. One may fail to reject both H_0^+ and H_0^-, and draw an **indeterminate** conclusion, because the data are underpowered to detect either difference or equivalence.

Value

`tost.ti` returns:

`statistics`	a vector of the *t* statistics for the two one-sided tests; if `relevance=TRUE`, these are followed by the value of the *t* statistic for the postivist test for difference.
`p.values`	a vector of *p* values for the *t* tests.
`estimate`	the estimated mean or means, or difference in means depending on whether it was a one-sample test, or a two-sample test.
`null.value`	the specified hypothesized value of the mean in a one-sample test, or 0 for a paired test or two-sample test.
`sterr`	the standard error used in the denominator of the *t* statistic.
`sd`	a vector containing the sample standard deviations of the two variables or two groups in unpaired tests; not returned for one-sample tests.
`sample_size`	a scalar (one-sample test) or vector (two-sample tests) containing the number of observations in the variable(s).
`parameter`	the degrees of freedom for the *t* statistics.
`threshold`	the value of the equivalence/relevance threshold: if `upper=NA` then returns the `eqv.level` argument. If `upper!=NA`, then returns a vector of (`eqv.level`,`upper`)
`conclusion`	a string containing the relevance test conclusion when `relevance=TRUE`.

Author(s)

Alexis Dinno (<alexis.dinno@pdx.edu>)

Please contact me with any questions, bug reports or suggestions for improvement. Fixing bugs will be facilitated by sending along:

1. a copy of the data (de-labeled or anonymized is fine),
2. a copy of the command syntax used, and
3. a copy of the exact output of the command.

I am endebted to my winter 2013 and fall 2023 students for their inspiration. Much appreciation to Mick McVeety for troubleshooting the translation of my Stata **tost** package to R.

Suggested citation: Dinno, A. 2025. **tost.ti**: Mean-equivalence *t* tests. In: **tost.suite** R software package.

References

Satterthwaite, F. E. (1946) An approximate distribution of estimates of variance components. *Biometrics Bulletin*. **2**, 110–114.

Schuirmann, D. A. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. **15**, 657–680.

Tryon, W. W., and Lewis, C. (2008) An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*. **13**, 272–277

Welch, B. L. (1947) The generalization of "Student's" problem when several different population variances are involved. *Biometrika*. **34**, 28–35.

See Also

`t.test`, `tost.t`.

Examples

```
# Immediate one-sample mean equivalence test
tost.ti(
  n1=24,
  mean1=62.6,
  sd1=15.8,
  mu=75,
  eqv.type="delta",
  eqv.level=20,
  relevance=FALSE)

# Immediate two-sample relevance t test of means assuming unequal variances
# Note: n1=24 m1=62.6 sd1=15.8 n2=30 m2=76.6 sd2=16.6
# Satterthwaite's df = 50.3912, and equivalence interval is +/- 1.5 sd
# beyond the critical value of T with df = 50.3912
tost.ti(
  n1=24, mean1=62.6, sd1=15.8,
  n2=30, mean2=76.6, sd2=16.6,
  eqv.type="epsilon",
  eqv.level=qt(.95, df=50.3912)+1.5*sqrt(50.3912/(50.3912-2)),
  x.name="Intervention",
  y.name="Control",
  conf.level=0.95,
  relevance=TRUE)
```

Index