

PREPRINT

Exploring Equivalence Testing with the TOSTER R package

Aaron R. Caldwell^a

^aNatick, MA, <https://orcid.org/0000-0002-4541-6283>

ARTICLE HISTORY

Compiled February 13, 2022

ABSTRACT

This is an article detailing the “avocado TOST” update to the TOSTER R package.

KEYWORDS

statistics, bootstrapping, minimal effects test, NHST, TOST

1. Introduction

Researchers often erroneously declare that no statistical effect exists based on a single “non-significant” p-value (Altman and Bland 1995). In many of these cases, the data may corroborate the researchers claim but the interpretation of a null hypothesis significance test (NHST), wherein the null hypothesis is “no effect”, is nonetheless incorrect. In order to statistically test for whether there is “no effect” or “no difference” researchers could explore using equivalence testing (CITE Welleck and Senn books). A very simple equivalence testing approach is the “two one-sided tests” (TOST) procedure (Schuirmann 1987). In the TOST procedure, an upper (Δ_U) and lower (Δ_L) equivalence bound is specified based on the smallest effect size of interest (SESOI). If the TOST procedure indicates the effect is close enough to zero to be practically equivalent then (CITE Lakens).

Both the complaints about erroneous conclusions regarding equivalence (Altman and Bland 1995) and proposed statistical solutions (Schuirmann 1987) have existed for decades now. Yet the problem appears to persist in many applied disciplines. I estimate that the cause of this continued dissonance is due to a lack of education on equivalence testing and struggle for many applied researchers to implement equivalence testing. In my experience, most researchers have received some degree of statistical training in their doctoral or master’s studies¹, but it is rare that any have idea of equivalence testing. It may also be difficult to implement equivalence testing for many researchers. This may be caused by most statistical software defaulting to a null hypothesis of zero, or even completely lacking an ability to change the null hypothesis.

The TOSTER R package was originally developed in 2018 (CITE Lakens) to introduce experimental psychologists to the concept of equivalence testing. Since then I have made a significant update to the package in order to improve the user interface and expand the tools available within the package. TOSTER (Avocado TOST edition

CONTACT Aaron R. Caldwell. Email: arcaldwell149@gmail.com

¹I do not mean to blame educators for this problem. I merely want to point out that

>v0.4.0) is a good tool to introduce the idea of equivalence testing to the average researcher who is already familiar with equivalence testing.

One possible source of this misinterpretation may be due to the fact that many researchers are

An experienced R programmer may have no problem performing equivalence testing within R but beginners may struggle with both writing the code and interpreting the output.

2. Basics of Equivalence Testing

2.1. *The TOSTER R Package*

In an effort to make TOSTER more informative and easier to use, a new function `t.TOST` has been created. This function operates very similarly to base R's `t.test` function, but performs 3 t-tests (one two-tailed and two one-tailed tests). In addition, this function has a generic method where two vectors can be supplied or a formula can be given (e.g., `y ~ group`). This function also makes it easier to switch between types of t-tests. All three types (two sample, one sample, and paired samples) can be performed/calculated from the same function. Moreover, the summary information and visualizations have been upgraded. This should make the decisions derived from the function more informative and user-friendly.

Also, `t.TOST` is not limited to equivalence tests. Minimal effects testing (MET) is possible. MET is useful for situations where the hypothesis is about a minimal effect and the null hypothesis *is* equivalence.

2.2. *Vignettes with TOSTER*

In the general introduction to this package we detailed how to look at *old* results and how to apply TOST to interpreting those results. However, in many cases, users may have new data that needs to be analyzed. Therefore, `t.TOST` can be applied to new data. This vignette will use the `bugs` data from the `jmv` R package and the `sleep` data.

```
data('sleep')
library(jmv)
data('bugs')
```

2.2.1. *Independent Groups*

For this example, we will use the `sleep` data. In this data there is a `group` variable and an outcome `extra`.

```
head(sleep)
```

```
##   extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
```

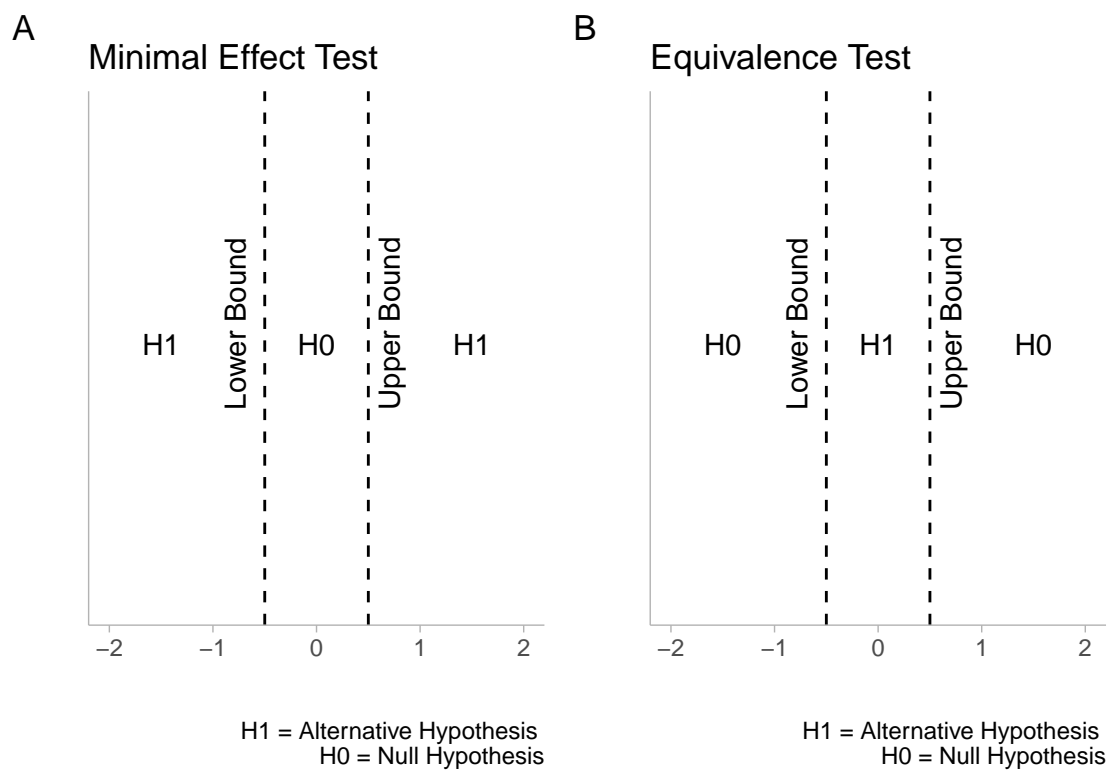


Figure 1. Type of Hypothesis

```
## 5  -0.1      1  5
## 6   3.4      1  6
```

We will assume the data are independent, and that we have equivalence bounds of ± 0.5 . All we need to do is provide the `formula`, `data`, and `eqbound` arguments for the function to run appropriately. In addition, we can set the `var.equal` argument (to assume equal variance), and the `paired` argument (sets if the data is paired or not). Both are logical indicators that can be set to TRUE or FALSE. The `alpha` is automatically set to 0.05 but this can also be adjusted by the user. The Hedges correction is also automatically calculated, but this can be overridden with the `bias_correction` argument. The `hypothesis` is automatically set to “EQU” for equivalence but if a minimal effect is of interest then “MET” can be supplied

```
# Formula Interface
res1 = t_TOST(formula = extra ~ group,
              data = sleep,
              low_eqbound = -.5,
              high_eqbound = .5)

# x & y interface
res1a = t_TOST(x = subset(sleep,group==1)$extra,
              y = subset(sleep,group==2)$extra,
              low_eqbound = -.5,
              high_eqbound = .5)
```

Once the function has run, we can print the results with the `print` command. This provides a verbose summary of the results.

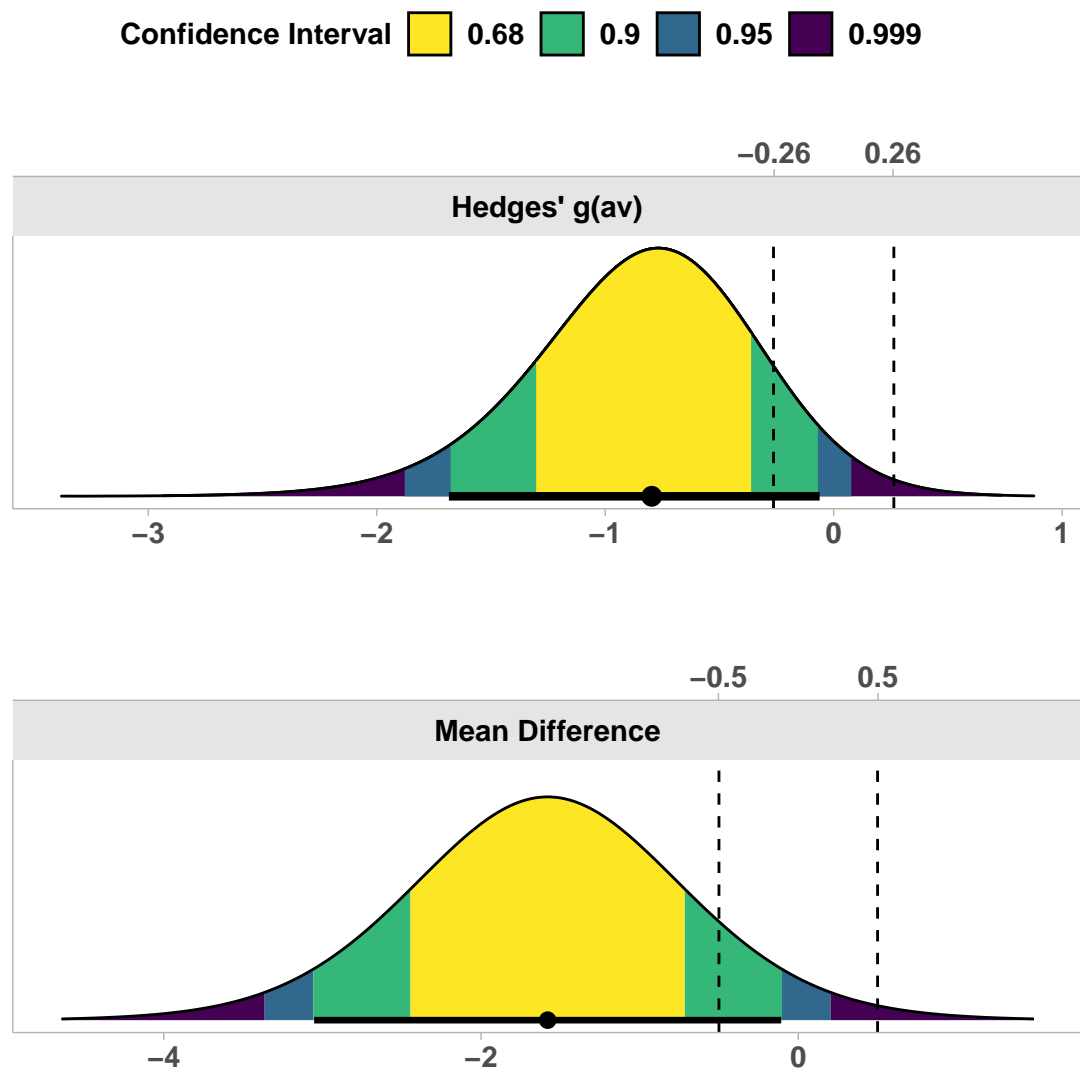
```
print(res1)

##
## Welch Two Sample t-test
## Hypothesis Tested: Equivalence
## Equivalence Bounds (raw):-0.500 & 0.500
## Alpha Level:0.05
## The equivalence test was non-significant, t(17.78) = -1.272, p = 8.9e-01
## The null hypothesis test was non-significant, t(17.78) = -1.861, p = 7.94e-02
## NHST: don't reject null significance hypothesis that the effect is equal to zero
## TOST: don't reject null equivalence hypothesis
##
## TOST Results
##           t           SE           df      p.value
## t-test    -1.860813  0.849091  17.77647  0.07939414
## TOST Lower -1.271948  0.849091  17.77647  0.89010996
## TOST Upper -2.449678  0.849091  17.77647  0.01245133
##
## Effect Sizes
##           estimate           SE  lower.ci  upper.ci  conf.level
## Raw          -1.5800000  0.849091 -3.053381 -0.10661850         0.9
## Hedges' g(av) -0.7964846  0.497633 -1.684326 -0.06154947         0.9
##
```

```
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

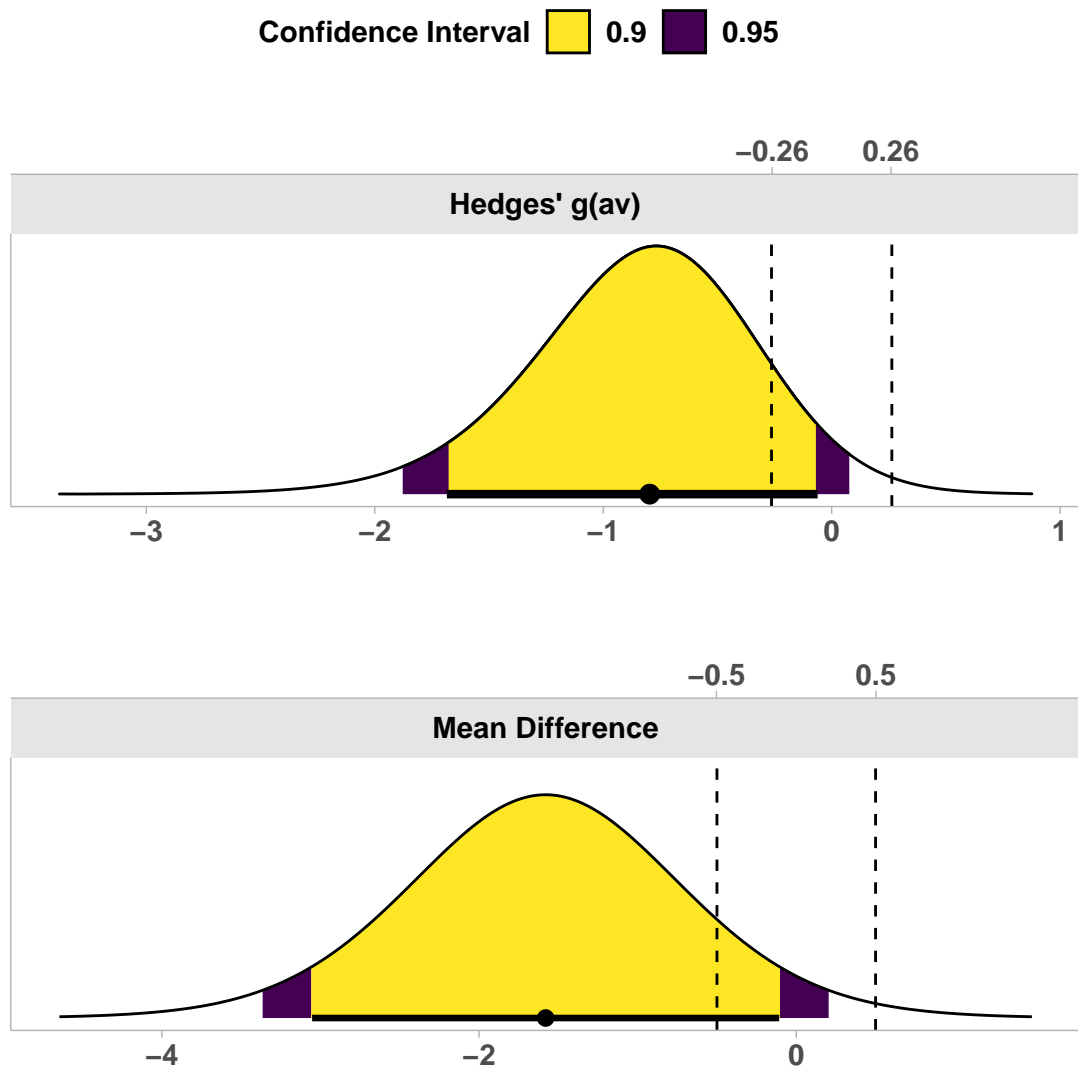
Another nice feature is the generic `plot` method that can provide a visual summary of the results. All of the plots in this package were inspired by the `concurve` R package. There are two types of plots that can be produced. The first, and default, is the consonance density plot (`type = "cd"`).

```
plot(res1, type = "cd")
```



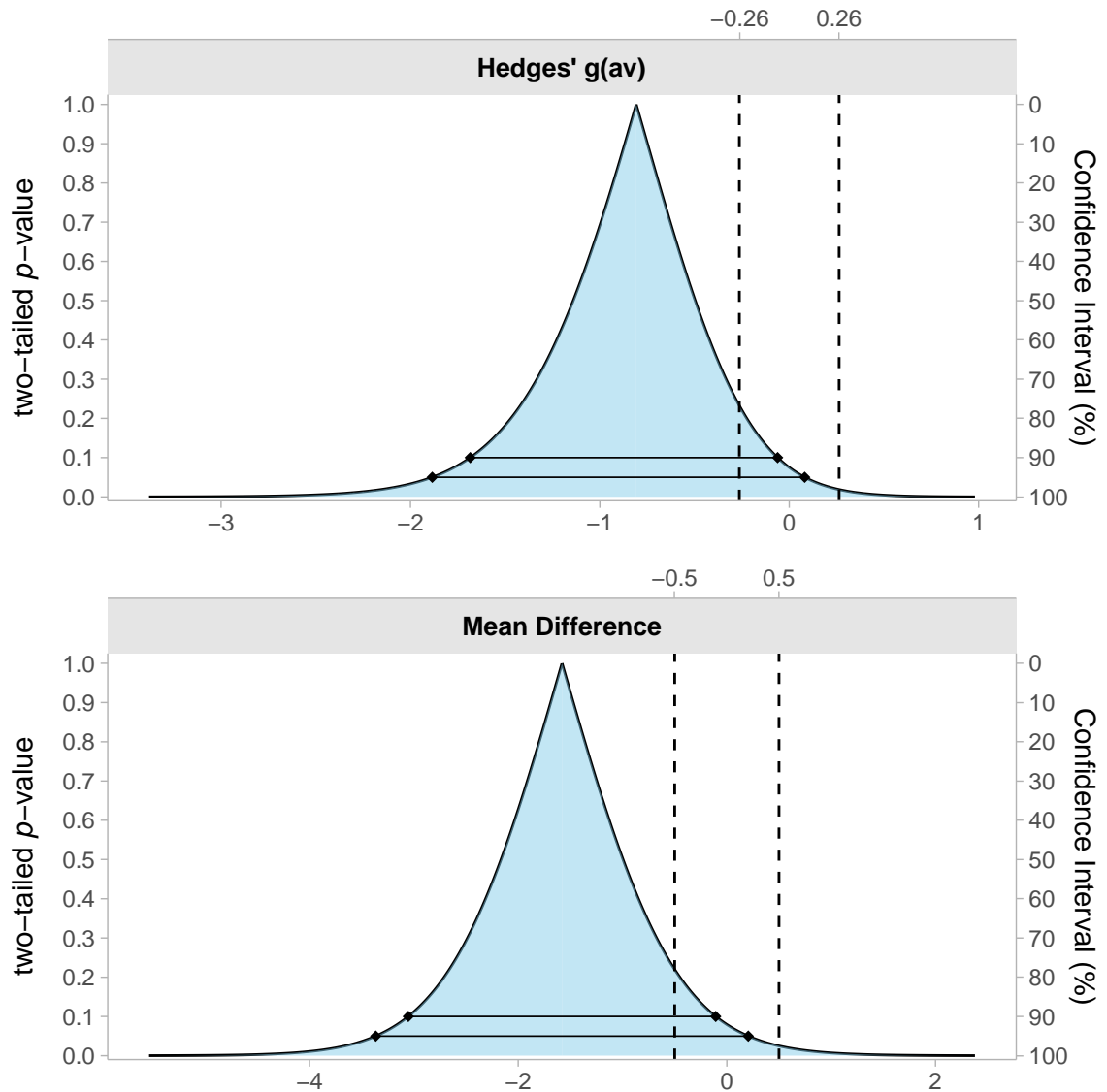
The shading pattern can be modified with the `ci_shades`.

```
plot(res1, type = "cd",  
      ci_shades = c(.9, .95))
```



Consonance plots, where all confidence intervals can be simultaneous plotted, can also be produced. The advantage here is multiple confidence interval lines can be plotted at once.

```
plot(res1, type = "c",
     ci_lines = c(.9,.95))
```



2.2.2. Paired Samples

To perform a paired samples TOST, the process does not change much. We could process the test the same way by providing a formula. All we would need to then is change `paired` to `TRUE`.

```
res2 = t.TOST(formula = extra ~ group,
              data = sleep,
              paired = TRUE,
              low_eqbound = -.5,
              high_eqbound = .5)
res2
```

```
##
## Paired t-test
## Hypothesis Tested: Equivalence
```

```
## Equivalence Bounds (raw):-0.500 & 0.500
## Alpha Level:0.05
## The equivalence test was non-significant, t(9) = -2.777, p = 9.89e-01
## The null hypothesis test was significant, t(9) = -4.062, p = 2.83e-03
## NHST: reject null significance hypothesis that the effect is equal to zero
## TOST: don't reject null equivalence hypothesis
##
## TOST Results
##           t           SE df         p.value
## t-test    -4.062128 0.3889587  9 0.0028328902
## TOST Lower -2.776644 0.3889587  9 0.9892407566
## TOST Upper -5.347611 0.3889587  9 0.0002319027
##
## Effect Sizes
##           estimate           SE lower.ci upper.ci conf.level
## Raw          -1.580000 0.3889587 -2.293005 -0.8669947         0.9
## Hedges' g(z) -1.230152 0.2008070 -1.848296 -0.8362302         0.9
##
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

However, we may have two vectors of data that are paired. So instead we may want to just provide those separately rather than using a data set and setting the formula. This can be demonstrated with the “bugs” data.

```
res3 = t_TOST(x = bugs$LDHF,
              y = bugs$LDLF,
              paired = TRUE,
              low_eqbound = -1,
              high_eqbound = 1)
res3
```

```
##
## Paired t-test
## Hypothesis Tested: Equivalence
## Equivalence Bounds (raw):-1.000 & 1.000
## Alpha Level:0.05
## The equivalence test was non-significant, t(90) = 2.655, p = 9.95e-01
## The null hypothesis test was significant, t(90) = 6.649, p = 2.22e-09
## NHST: reject null significance hypothesis that the effect is equal to zero
## TOST: don't reject null equivalence hypothesis
##
## TOST Results
##           t           SE df         p.value
## t-test     6.648618 0.2504032 90 2.223690e-09
## TOST Lower 10.642177 0.2504032 90 6.681389e-18
## TOST Upper  2.655058 0.2504032 90 9.953114e-01
##
## Effect Sizes
##           estimate           SE lower.ci upper.ci conf.level
## Raw          1.6648352 0.25040321 1.2486748 2.0809956         0.9
```



```
## Hedges' g(z) 0.6940558 0.09548981 0.5296194 0.8707979      0.9
##
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

We may want to perform a Minimal Effect Test with the `hypothesis` argument set to “MET”.

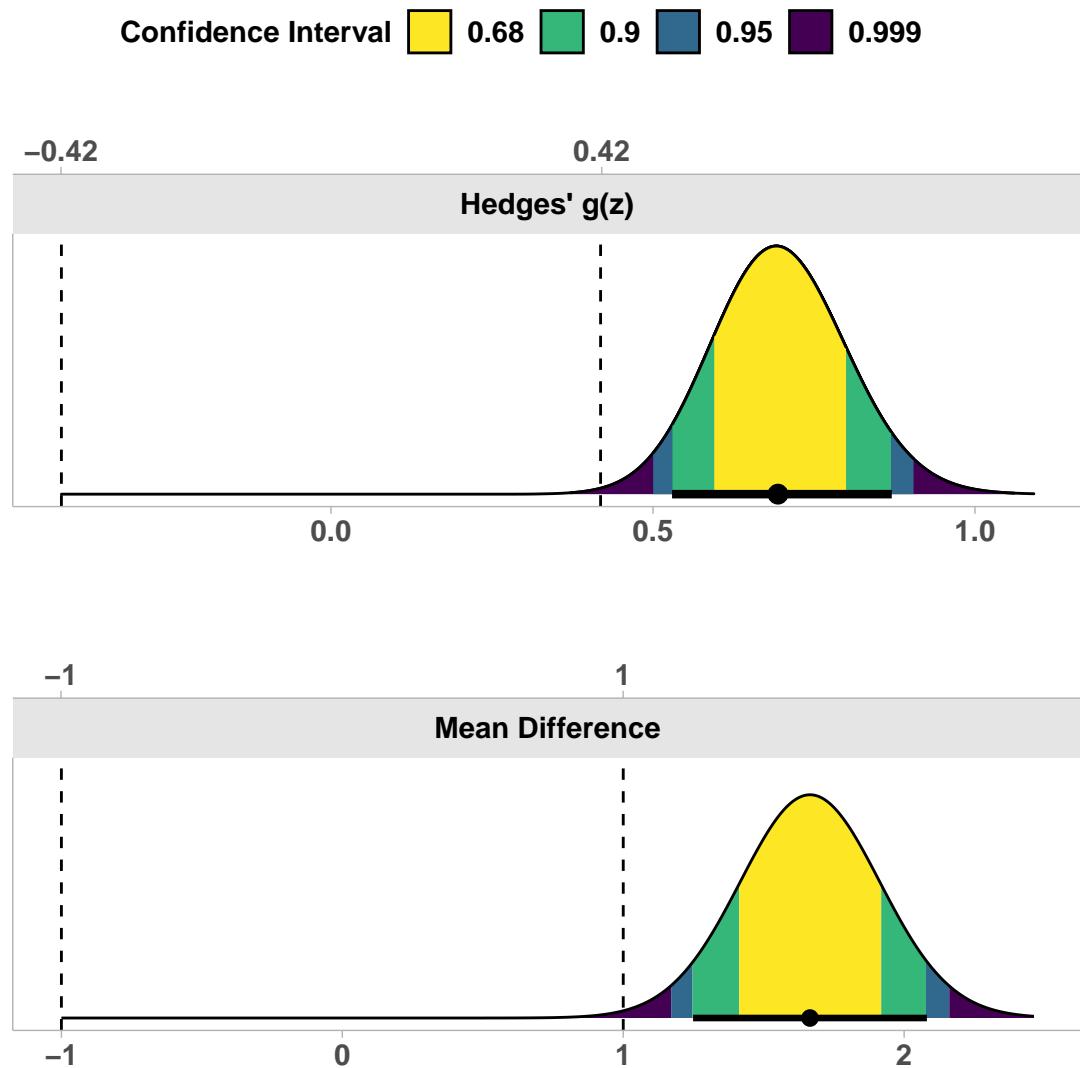
```
res3a = t_TOST(x = bugs$LDHF,
               y = bugs$LDLF,
               paired = TRUE,
               hypothesis = "MET",
               low_eqbound = -1,
               high_eqbound = 1)
```

```
res3a
```

```
##
## Paired t-test
## Hypothesis Tested: Minimal Effect
## Equivalence Bounds (raw):-1.000 & 1.000
## Alpha Level:0.05
## The minimal effect test was significant, t(90) = 10.642, p = 4.69e-03
## The null hypothesis test was significant, t(90) = 6.649, p = 2.22e-09
## NHST: reject null significance hypothesis that the effect is equal to zero
## TOST: reject null MET hypothesis
##
## TOST Results
##           t           SE df      p.value
## t-test      6.648618 0.2504032 90 2.223690e-09
## TOST Lower 10.642177 0.2504032 90 1.000000e+00
## TOST Upper  2.655058 0.2504032 90 4.688649e-03
##
## Effect Sizes
##           estimate           SE lower.ci upper.ci conf.level
## Raw           1.6648352 0.25040321 1.2486748 2.0809956      0.9
## Hedges' g(z) 0.6940558 0.09548981 0.5296194 0.8707979      0.9
##
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

Again, we can plot the effects from the `t_TOST` result.

```
plot(res3a)
```



2.2.3. One Sample t-test

In other cases we may just have a one sample test. If that is the case all we have to do is supply the `x` argument for the data. For this test we may hypothesis that the mean of LDHF is not more than 1.5 points greater or less than 7.

```
res4 = t.TOST(x = bugs$LDHF,
              hypothesis = "EQU",
              low_eqbound = 5.5,
              high_eqbound = 8.5)
```

```
res4
```

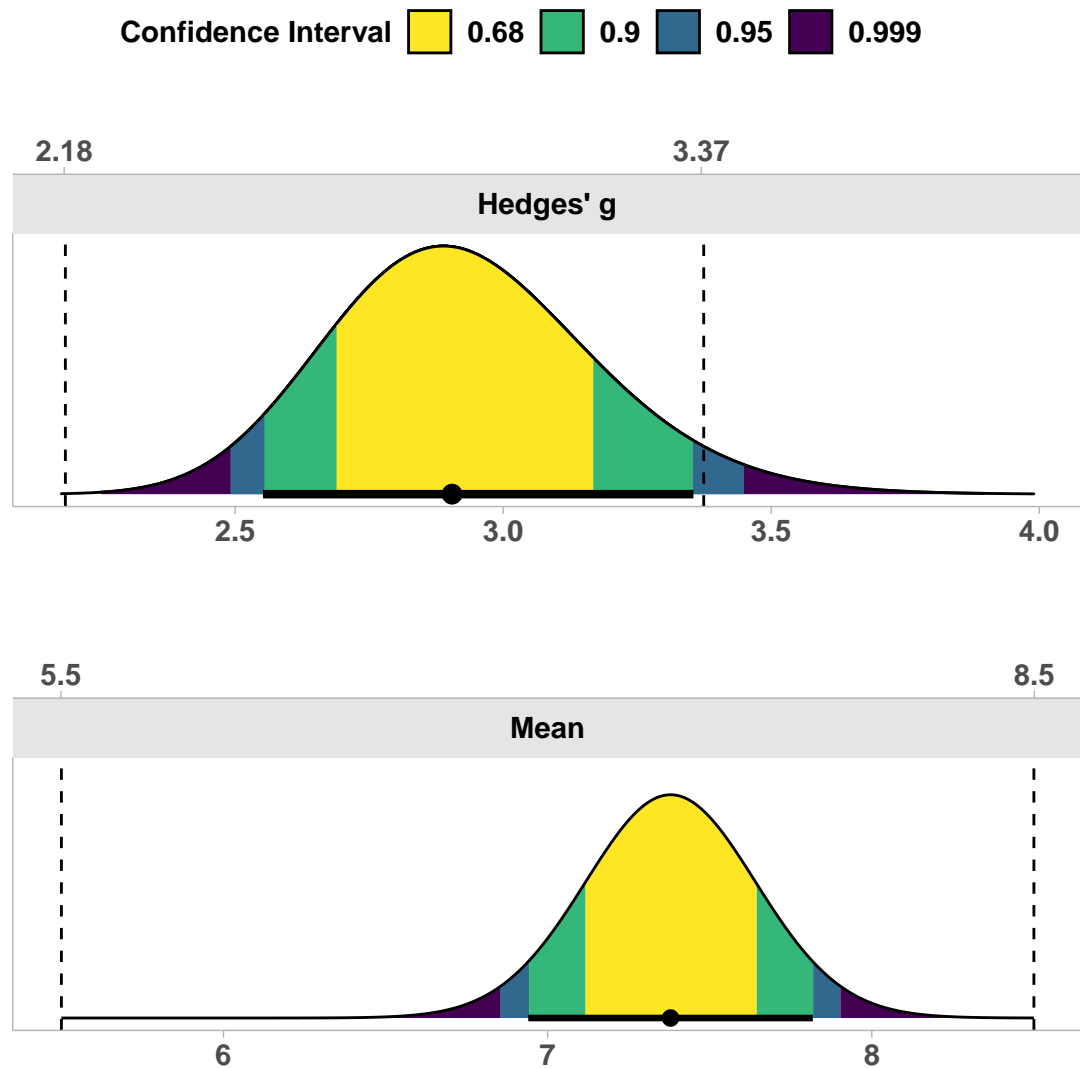
```
##
## One Sample t-test
## Hypothesis Tested: Equivalence
## Equivalence Bounds (raw):5.500 & 8.500
```

```

## Alpha Level:0.05
## The equivalence test was significant, t(90) = -4.244, p = 2.66e-05
## The null hypothesis test was significant, t(90) = 27.942, p = 3.91e-46
## NHST: reject null significance hypothesis that the effect is equal to zero
## TOST: reject null equivalence hypothesis
##
## TOST Results
##           t           SE df         p.value
## t-test    27.942403 0.2640833 90 3.906969e-46
## TOST Lower  7.115638 0.2640833 90 1.298575e-10
## TOST Upper -4.244416 0.2640833 90 2.658068e-05
##
## Effect Sizes
##           estimate           SE lower.ci upper.ci conf.level
## Raw          7.379121 0.2640833 6.940225 7.818017          0.9
## Hedges' g    2.904671 0.2452067 2.551811 3.354884          0.9
##
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").

plot(res4)

```



2.2.4. Using Summary Statistics

In some cases you may only have access to the summary statistics. Therefore, we created a function, `tsum_TOST`, to perform the same tests just based on the summary statistics. This involves providing the function with a number of different arguments.

- `n1` & `n2` the sample sizes (only `n1` needs to be provided for one sample case)
- `m1` & `m2` the sample means
- `sd1` & `sd2` the sample standard deviation
- `r12` the correlation between each if paired is set to `TRUE`

The results from above can be replicated with the `tsum_TOST`

```
res.tsum = tsum_TOST(
  m1 = mean(bugs$LDHF, na.rm=TRUE),
  sd1 = sd(bugs$LDHF, na.rm=TRUE),
```

```

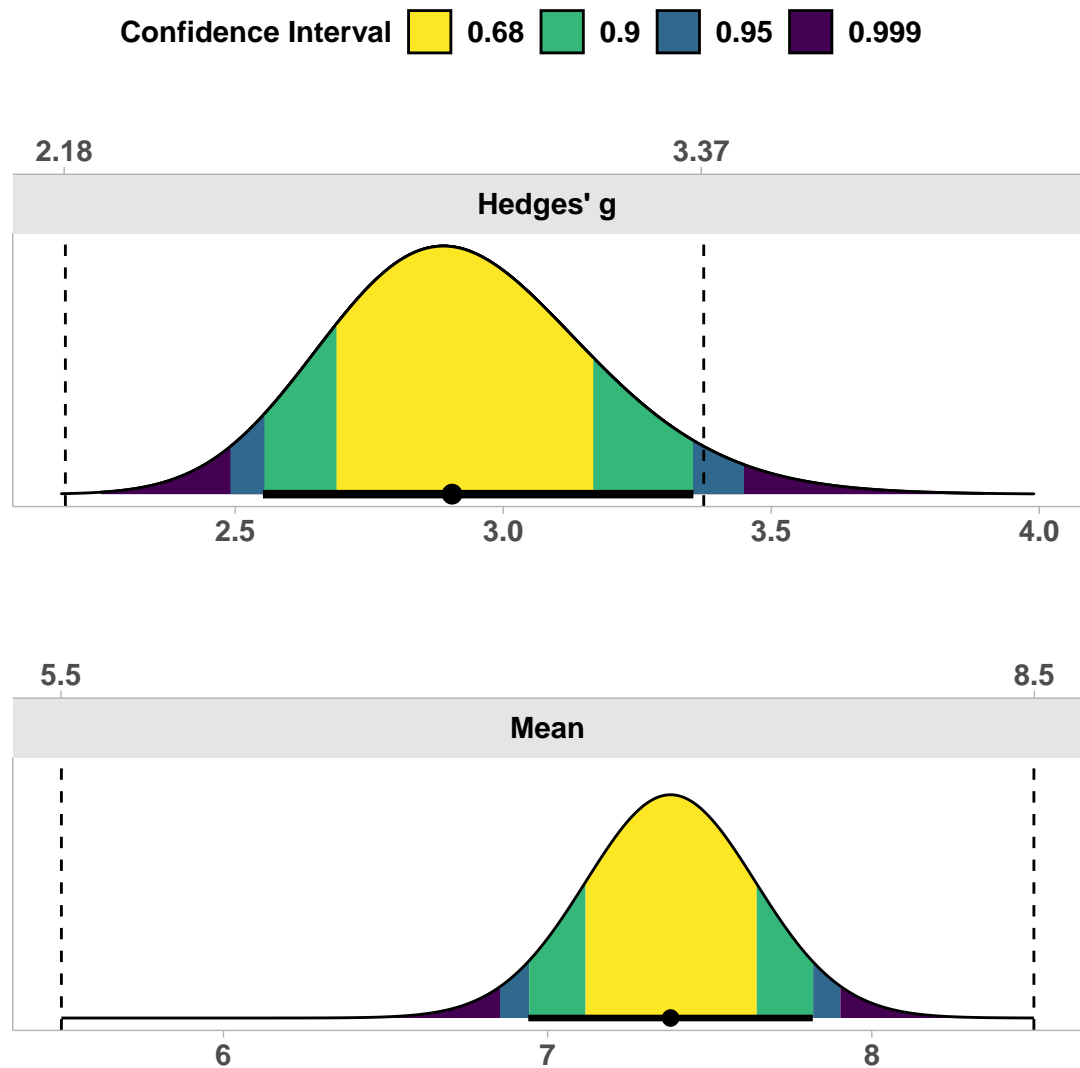
n1 = length(na.omit(bugs$LDHF)),
hypothesis = "EQU",
low_eqbound = 5.5,
high_eqbound = 8.5
)

res_tsum

##
## One-sample t-Test
## Hypothesis Tested: Equivalence
## Equivalence Bounds (raw):5.500 & 8.500
## Alpha Level:0.05
## The equivalence test was significant, t(90) = -4.244, p = 2.66e-05
## The null hypothesis test was significant, t(90) = 27.942, p = 3.91e-46
## NHST: reject null significance hypothesis that the effect is equal to zero
## TOST: reject null equivalence hypothesis
##
## TOST Results
##           t           SE df       p.value
## t-test      27.942403 0.2640833 90 3.906969e-46
## TOST Lower   7.115638 0.2640833 90 1.298575e-10
## TOST Upper  -4.244416 0.2640833 90 2.658068e-05
##
## Effect Sizes
##           estimate           SE lower.ci upper.ci conf.level
## Raw           7.379121 0.2640833 6.940225 7.818017         0.9
## Hedges' g     2.904671 0.2452067 2.551811 3.354884         0.9
##
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").

plot(res_tsum)

```



3. Robust Methods for Equivalence Testing

Why this may be useful Describe these aren't tests of medians but symmetry tests (myth busting) Paired samples may want to be rank transformed

In this package there are currently 2 functions that provide robust alternatives to the `t.TOST` function.

3.1. *Tests of Symmetry*

The Wilcoxon group of tests (includes Mann-Whitney U-test) provide a non-parametric test of differences between groups, or within samples, based on ranks. This provides a test of location shift, which is a fancy way of saying differences in the center of the distribution (i.e., in parametric tests the location is mean). With TOST, there are two separate tests of directional location shift to determine if the location

shift is within (equivalence) or outside (minimal effect). The exact calculations can be explored via the documentation of the `wilcox.test` function.

In the TOSTER package, we accomplish this with the `wilcox.TOST` function. Overall, this function operates extremely similar to the `t.TOST` function. However, the standardized mean difference (SMD) is *not* calculated. Instead the rank-biserial correlation is calculated for *all* types of comparisons (e.g., two sample, one sample, and paired samples). Also, there is no plotting capability at this time for the output of this function.

As an example, we can use the sleep data to make a non-parametric comparison of equivalence.

```
data('sleep')
library(TOSTER)

test1 = wilcox.TOST(formula = extra ~ group,
                    data = sleep,
                    paired = FALSE,
                    low.eqbound = -.5,
                    high.eqbound = .5)

print(test1)
```

```
##
## Wilcoxon rank sum test with continuity correction
## Hypothesis Tested: Equivalence
## Equivalence Bounds (raw):-0.500 & 0.500
## Alpha Level:0.05
## The equivalence test was non-significant W = 20.000, p = 8.94e-01
## The null hypothesis test was non-significant W = 25.500, p = 6.93e-02
## NHST: don't reject null significance hypothesis that the effect is equal to zero
## TOST: don't reject null equivalence hypothesis
##
## TOST Results
##           statistic    p.value
## NHST           25.5 0.06932758
## TOST Lower      34.0 0.89385308
## TOST Upper      20.0 0.01287404
##
## Effect Sizes
##           estimate  lower.ci  upper.ci  conf.level
## Median of Differences    -1.346388 -3.3999651 -0.09995341      0.9
## rank-biserial correlation -0.490000 -0.7492521 -0.10053222      0.9
```

3.2. Bootstrap TOST

The bootstrap refers to resampling with replacement and can be used statistical estimation and inference. Bootstrapping techniques are very useful because they are considered somewhat robust to the violations of assumptions for a simple t-test. Therefore we added a bootstrap option, `boot.t.TOST` to the package to provide another robust

alternative to the `t_TOST` function.

In this function we provide a percentile bootstrap solution outlined by Efron and Tibshirani (1993) (see chapter 16, page 220). The bootstrapped p-values are derived from the “studentized” version of a test of mean differences (Efron and Tibshirani 1993). Overall, the results should be similar to the results of `t_TOST`. **However**, for paired samples, the Cohen’s $d(rm)$ effect size *cannot* be calculated at this time.

3.2.1. Two Sample Algorithm

1. Form B bootstrap data sets from x^* and y^* wherein x^* is sampled with replacement from $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$ and y^* is sampled with replacement from $\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_n$
2. t is then evaluated on each sample, but the mean of each sample (y or x) and the overall average (z) are subtracted from each

$$t(z^{*b}) = \frac{(\bar{x}^* - \bar{x} - \bar{z}) - (\bar{y}^* - \bar{y} - \bar{z})}{\sqrt{sd_y^*/n_y + sd_x^*/n_x}}$$

3. An approximate p-value can then be calculated as the number of bootstrapped results greater than the observed t-statistic from the sample.

$$p_{boot} = \frac{\#t(z^{*b}) \geq t_{sample}}{B}$$

The same process is completed for the one sample case but with the one sample solution for the equation outlined by $t(z^{*b})$. The paired sample case in this bootstrap procedure is equivalent to the one sample solution because the test is based on the difference scores.

3.2.2. Example of Bootstrapping

Again, we can use the sleep data to see the bootstrapped results. Notice that the plots show how the resampling via bootstrapping indicates the instability of Hedges’ $d(z)$.

```
data('sleep')

test1 = boot_t_TOST(formula = extra ~ group,
                    data = sleep,
                    paired = TRUE,
                    low_eqbound = -.5,
                    high_eqbound = .5,
                    R = 999)

print(test1)

##
## Bootstrapped Paired t-test
```



```

## Hypothesis Tested: Equivalence
## Equivalence Bounds (raw):-0.500 & 0.500
## Alpha Level:0.05
## The equivalence test was non-significant, t(9) = -2.777, p = 1e+00
## The null hypothesis test was significant, t(9) = -4.062, p = 0e+00
## NHST: reject null significance hypothesis that the effect is equal to zero
## TOST: don't reject null equivalence hypothesis
##
## TOST Results
##           t           SE df p.value
## t-test    -4.062128 0.3889587  9      0
## TOST Lower -2.776644 0.3889587  9      1
## TOST Upper -5.347611 0.3889587  9      0
##
## Effect Sizes
##           estimate           SE lower.ci upper.ci conf.level
## Raw          -1.580000 0.3498821 -2.170000 -1.0490000        0.9
## Hedges' g(z) -1.230152 0.7258598 -2.881678 -0.9220578        0.9
##
## Note: percentile bootstrap method utilized.

```

4. Equivalence Testing with ANOVAs

For an open access tutorial paper explaining how to set equivalence bounds, and how to perform and report equivalence for ANOVA models see Campbell and Lakens (2021). These functions are meant to be omnibus tests, and additional testing may be necessary. For example, comparison of the estimated marginal means, in addition to or as an alternative of with may be prudent.

4.1. *F-test Calculations*

Statistical equivalence testing (or “omnibus non-inferiority testing” as Campbell and Lakens (2021)) for F -tests are special use case of the cumulative distribution function of the non-central F distribution. As Campbell and Lakens (2021) state, these type of questions answer the question: “Can we reject the hypothesis that the total proportion of variance in outcome Y attributable to X is greater than or equal to the equivalence bound Δ ?”

$$H_0 : 1 > \eta_p^2 \geq \Delta H_1 : 0 \geq \eta_p^2 < \Delta$$

In TOSTER we go a tad farther and calculate a more generalizable non-centrality parameter that allows the equivalence test for F -tests to be applied to variety of designs.

Campbell and Lakens (2021) calculate the p -value as:

$$p = p_f(F; J - 1, N - J, \frac{N \cdot \Delta}{1 - \Delta})$$

However, this approach could not be applied to factorial ANOVA and the paper only outlines how to apply this approach to a one-way ANOVA and an extension to Welch's one-way ANOVA.

However, the non-centrality parameter ($\text{ncp} = \lambda$) can be calculated with the equivalence bound and the degrees of freedom:

$$\lambda_{eq} = \frac{\Delta}{1 - \Delta} \cdot (df_1 + df_2 + 1)$$

The p -value for the equivalence test (p_{eq}) could then be calculated from traditional ANOVA results and the distribution function:

$$p_{eq} = p_f(F; df_1, df_2, \lambda_{eq})$$

4.2. Example of Equivalence ANOVA Test

Using the `InsectSprays` data set in R and the base R `aov` function we can demonstrate how this omnibus equivalence testing can be applied with `TOSTER`.

From the initial analysis we can see a clear “significant” effect (the p -value listed is zero but it is just very small) of the factor `spray`. However, we *may* be interested in testing if the effect is practically equivalent. I will arbitrarily set the equivalence bound to a partial eta-squared of 0.35 ($H_0 : \eta_p^2 > 0.35$)

```
library(TOSTER)
# Get Data
data("InsectSprays")
# Build ANOVA
aovtest = aov(count ~ spray,
              data = InsectSprays)

# Display overall results
knitr::kable(broom::tidy(aovtest),
             caption = "Traditional ANOVA Test")
```

Table 1.: Traditional ANOVA Test

term	df	sumsq	meansq	statistic	p.value
spray	5	2668.833	533.76667	34.70228	0
Residuals	66	1015.167	15.38131	NA	NA

We can then use the information in the table above to perform an equivalence test using the `equ_ftest` function. This function returns an object of the S3 class `htest` and the output will look very familiar to the t -test. The main difference is the estimates, and confidence interval, are for partial η_p^2 .

```
equ_ftest(Fstat = 34.70228,
          df1 = 5,
          df2 = 66,
          eqbound = 0.35)
```

```
##
##  Equivalence Test from F-test
##
## data:  Summary Statistics
## F = 34.702, df1 = 5, df2 = 66, p-value = 1
## 95 percent confidence interval:
##  0.5806263 0.7804439
## sample estimates:
## [1] 0.724439
```

Based on the results above we would conclude there is a significant effect of “spray” and the differences due to spray are *not* statistically equivalent. In essence, we reject the traditional null hypothesis of “no effect” but accept the null hypothesis of the equivalence test.

The `equ_ftest` is very useful because all you need is very basic summary statistics. However, if you are doing all your analyses in R then you can use the `equ_anova` function. This function accepts objects produced from `stats::aov`, `car::Anova` and `afex::aov_car` (or any ANOVA from derived from `afex`).

```
equ_anova(aovtest,
          eqbound = 0.35)
```

```
##          effect df1 df2  F.value      p.null    pes eqbound    p.equ
## 1 spray          5  66 34.70228 3.182584e-17 0.724439    0.35 0.9999965
```

5. Visualizing Equivalence

Add plots from the package

Acknowledgement(s)

An unnumbered section, e.g. `\section*{Acknowledgements}`, may be used for thanks, etc. if required and included *in the non-anonymous version* before any Notes or References.

Disclosure statement

An unnumbered section, e.g. `\section*{Disclosure statement}`, may be used to declare any potential conflict of interest and included *in the non-anonymous version* before any Notes or References, after any Acknowledgements and before any Funding information.

Funding

An unnumbered section, e.g. `\section*{Funding}`, may be used for grant details, etc. if required and included *in the non-anonymous version* before any Notes or References.

Notes on contributor(s)

An unnumbered section, e.g. `\section*{Notes on contributors}`, may be included *in the non-anonymous version* if required. A photograph may be added if requested.

Nomenclature/Notation

An unnumbered section, e.g. `\section*{Nomenclature}` (or `\section*{Notation}`), may be included if required, before any Notes or References.

Notes

An unnumbered **Notes** section may be included before the References (if using the `endnotes` package, use the command `\theendnotes` where the notes are to appear, instead of creating a `\section*`).

6. References

6.1. *References cited in the text*

6.2. *The list of references*

References should be listed at the end of the main text in alphabetical order by authors' surnames, then chronologically (earliest first). If references have the same author(s), editor(s), etc., arrange by year of publication, with undated works at the end. A single-author entry precedes a multi-author entry that begins with the same name. If the reference list contains two or more items by the same author(s) in the same year, add a, b, etc. and list them alphabetically by title. Successive entries by two or more authors when only the first author is the same are alphabetized by co-authors' surnames. If a reference has more than ten named authors, list only the first seven, followed by 'et al.'. If a reference has no author or editor, order by title; if a date of publication is impossible to find, use 'n.d.' in its place.

The following list shows some sample references prepared in the Taylor & Francis Chicago author-date style.

(??)

References

Altman, Douglas G, and J Martin Bland. 1995. "Statistics notes: Absence of evidence is not evidence of absence." *BMJ* 311 (7003): 485.

- Campbell, Harlan, and Daniël Lakens. 2021. “Can we disregard the whole model? Omnibus non-inferiority testing for R^2 in multi-variable linear regression and in ANOVA.” *British Journal of Mathematical and Statistical Psychology* 74 (1): e12201. <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bmsp.12201>.
- Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Boca Raton, Florida, USA: Chapman & Hall/CRC.
- Schirmann, Donald J. 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 (6): 657–680.