pkequiv — Perform bioequivalence tests

Description

`pkequiv` tests for average bioequivalence between two drugs. By default, `pkequiv` computes a confidence interval (CI) for the ratio of treatment geometric means, calculated using log-scaled data. `pkequiv` can also calculate CIs using the original data, such as the classic shortest CI and a CI based on Fieller's theorem. Additionally, `pkequiv` performs interval hypothesis tests for bioequivalence, such as Schuirmann's two one-sided tests.

`pkequiv` is one of the pk commands. Please read [R] **pk** before reading this entry.

Quick start

Log-scale CI for the geometric mean ratio of pharmacokinetic (PK) outcome `auc`—the area under the concentration-time curve (AUC)—between drugs identified by `treat`, given during period `period` in sequence `sequence`, with subjects identified by `id`

```
pkequiv auc treat period sequence id
```

Same as above, but use an equivalence limit of 0.1, do not perform Schuirmann's two one-sided tests, and do not perform a bootstrap analysis of the long-run probability of bioequivalence

```
pkequiv auc treat period sequence id, limit(0.1) notost noboot
```

Analyze data using the original scale, and calculate a CI for the ratio using Fieller's theorem

```
pkequiv auc treat period sequence id, fieller
```

Same as above, but specify that treatment 2 is the reference drug and treatment 1 is the test drug

```
pkequiv auc treat period sequence id, fieller compare(2 1)
```

Menu

Statistics > Epidemiology and related > Other > Bioequivalence tests

Syntax

pkequiv *outcome treatment period sequence id* [*if*] [*in*] [, *options*]

options	Description
Options	
compare(# #)	compare the two specified values of the treatment variable
limit(#)	set equivalence limit (between 0.01 and 0.99); default is limit(0.2)
level(#)	set confidence level; default is level(90)
log	calculate a CI using the log scale; the default
classic	calculate the classic shortest CI
fieller	calculate a CI for the ratio using Fieller's theorem
symmetric	calculate a CI for the difference in means that is symmetric around 0
[no]tost	perform or suppress two one-sided hypothesis tests for bioequivalence
anderson	perform Anderson and Hauck hypothesis test for bioequivalence
noboot	do not compute the bootstrap probability that the CI lies within the equivalence limits
reps(#)	perform # bootstrap replications; default is reps(1000)

collect is allowed; see **[U] 11.1.10 Prefix commands**.

Options

Options

compare(# #) specifies the two treatments being compared, where the first value corresponds to the reference drug and the second value corresponds to the test drug. If there are more than two treatments, compare() is required; bioequivalence can be determined between any two treatments.

limit(#) specifies the equivalence limit. The default is limit(0.2). Using the log scale, this yields equivalence limits of 80% to 125% for the ratio of geometric means. Using the original scale, this yields limits of 80% to 120% for the ratio of arithmetic means and limits of ±20% for the difference between arithmetic means. See *Equivalence limits* in *Methods and formulas*.

level(#) specifies the confidence level, as a percentage, for CIs. The default is level(90). This setting is not controlled by the set level command.

log specifies a CI for the ratio of treatment geometric means, calculated using the log-transformed outcome. This is the default CI, and its use is recommended by the US Food and Drug Administration (FDA) (2001, 2022) and the European Medicines Agency (2010). log may not be combined with classic, fieller, symmetric, or anderson.

classic specifies the classic shortest CI for the difference between treatment arithmetic means, calculated using the original scale of the data. This CI is also expressed as a CI for the ratio of arithmetic means. classic may not be combined with log, fieller, or symmetric.

fieller specifies a CI based on Fieller's theorem for the ratio of treatment arithmetic means, calculated using the original scale of the data. fieller may not be combined with log, classic, symmetric, tost, notost, anderson, noboot, or reps().

`symmetric` specifies a CI for the difference between treatment arithmetic means that is symmetric around zero, calculated using the original scale of the data. Use of this CI for bioequivalence testing has been criticized by Mantel (1977), Kirkwood (1981), and others, so it is included mainly for historical purposes. `symmetric` may not be combined with `log`, `classic`, `fieller`, `tost`, `notost`, `anderson`, `noboot`, or `reps()`.

`tost` and `notost` perform or suppress the calculation of Schuirmann's (1987) two one-sided tests (TOST) for bioequivalence. `tost` is the default whenever log-scale or classic CIs are computed. With a log-scale CI, the TOST will be performed on log-transformed outcome data; with a classic CI, the TOST will be performed on untransformed data. The TOST procedure may not be combined with CIs based on Fieller's theorem or with symmetric CIs, so `tost` and `notost` may not be combined with `fieller` or `symmetric`.

`anderson` specifies that the Anderson and Hauck (1983) hypothesis test for bioequivalence be computed. Use of this test has been criticized by the US FDA (1988), Schuirmann (1987), and others, so it is included mainly for historical purposes. This test is supported only with `classic`.

`noboot` prevents the computation of the probability that the CI lies within the equivalence limits. By default, this probability is estimated via bootstrap resampling with 1,000 replications. This probability is not computed with CIs based on Fieller's theorem or with symmetric CIs, so `noboot` may not be combined with `fieller` or `symmetric`, nor may it be combined with `reps()`.

`reps(#)` specifies the number of bootstrap replications to be performed when estimating the probability that the CI lies within the equivalence limits. The default is `reps(1000)`. `reps()` may not be combined with `fieller`, `symmetric`, or `noboot`.

Remarks and examples

Remarks are presented under the following headings:

Introduction

`pkequiv` is used to determine average bioequivalence based on data from a nonreplicated crossover study. A crossover study is a longitudinal study where each subject undergoes treatment with both reference (R) and test (T) drugs. In a nonreplicated crossover study, subjects receive each treatment only once, so the sequence of treatments is either R followed by T or T followed by R.

Two drugs are bioequivalent if their bioavailability profiles are sufficiently similar. The most common PK metrics used as measures of bioavailability are the AUC and the maximum concentration ($C_{\max}$).

"Average bioequivalence" between two drugs is declared if the average PK metric of the test drug is close enough to the average PK metric of the reference drug. This is traditionally done by constructing a CI for a measure of disparity between the average PK metrics. If the CI is entirely contained within prespecified equivalence limits, the two drugs are said to demonstrate average bioequivalence.

Two measures of disparity can be used: the difference between treatment means (T − R) and the ratio of treatment means (T/R). Equivalence limits for the difference and for the ratio are controlled with the `limit()` option; see *Equivalence limits* in *Methods and formulas* for details.

By default, pkequiv analyzes PK data using the log scale and calculates a CI for the geometric mean ratio, which is the ratio of exponentiated means of the log-transformed outcome. This approach is recommended by the US FDA (2001, 2022) and the European Medicines Agency (2010) for the analysis of most PK metrics. One motivation for taking the log of PK metrics AUC and $C_{\max}$ is that the biological model postulated for each of those metrics contains a multiplicative term that is a function of the subject; this term becomes additive on the log scale (US FDA 2001, 2022).

The classic option analyzes PK data using the original scale and calculates the classic shortest CI for the difference between treatment arithmetic means. This CI is transformed into an approximate CI for the ratio of treatment arithmetic means; see *Classic interval* in *Methods and formulas* for details. The CIs for the difference and for the ratio are compared with their respective equivalence limits, and if the CIs are within the equivalence limits, the two drugs are bioequivalent. The comparison of the difference is mathematically identical to the comparison of the ratio, so the two comparisons will always agree.

The fieller option calculates an exact CI for the ratio of treatment arithmetic means, computed using the untransformed data. This approach is most frequently used when the outcome variable is a clinical outcome instead of a PK metric. These trials, called comparative clinical endpoint studies, can have binary outcomes (such as success or failure) or continuous outcomes (such as percent improvement versus baseline), and are generally not analyzed using the log scale (US FDA 2022).

The symmetric option uses the original scale of the data to calculate a CI for the difference between treatment arithmetic means that is symmetric around zero. The symmetric CI has received considerable criticism since its proposal by Westlake (1976), so it is included chiefly for historical purposes. See, for example, the criticisms of Mantel (1977) and Kirkwood (1981). Although this CI is symmetric around zero, it is asymmetric with respect to the estimated difference between means. This shift away from the observed difference between means causes unequal tail probabilities. As the difference between means increases, the symmetric CI is shifted ever farther from the estimated difference, and the tail probability becomes dominated by one side.

In addition to calculating CIs, pkequiv can conduct interval hypothesis tests of bioequivalence. When a log-scale or classic CI is requested, pkequiv performs Schuirmann's (1987) TOST unless the notost option is specified. The TOST procedure consists of two hypothesis tests: one to determine whether the average PK metric of the test drug exceeds the upper equivalence limit and one to determine whether it is less than the lower equivalence limit. In both cases, the null hypothesis is that the two drugs are not bioequivalent, so rejecting both null hypotheses leads to a conclusion of bioequivalence. The formula for the TOST is an inversion of the formula for the CI; see *Schuirmann's two one-sided tests* in *Methods and formulas* for details. This ensures that the conclusion reached by the TOST with two α-level tests will agree with the conclusion reached by comparing the $(1-2\alpha)100\%$ CI against the equivalence limits.

The anderson option is used to perform Anderson and Hauck's (1983) test, where the null hypothesis is that the two drugs are not bioequivalent. This procedure offers the appeal of a single test of bioequivalence that is more powerful than Schuirmann's TOST for any nominal significance level α. However, two serious criticisms of this test have prevented its widespread adoption (US FDA 1988), so it is included mainly for historical reasons. The actual size of Anderson and Hauck's test generally exceeds the nominal size of the test, meaning that the probability of committing a type I error and erroneously concluding bioequivalence is greater than desired (Frick 1987). Additionally, the rejection region of Anderson and Hauck's test is nonconvex and unbounded (Schuirmann 1987), meaning that H_0 will be rejected for any arbitrarily large difference between means as long as the corresponding standard error is sufficiently large.

Examples

pkequiv requires that the user specify the *outcome*, *treatment*, *period*, *sequence*, and *id* variables. The dataset must be in the same format as that produced by pkshape; see [R] **pkshape**. The examples below assume the use of a $2 \times 2 \times 2$ design: a crossover study with 2 treatments, 2 periods, and 2 sequences. If your dataset includes results for more than two treatments, you will need to use the compare() option to specify which two treatments are being compared.

▷ Example 1

We use pkdata.dta from example 2 of [R] **pk**. We use pkcollapse and pkshape on the data as discussed in that example. After we sort the data with the sort command, our data appear as follows:

```
. use https://www.stata-press.com/data/r19/pkdata
(Fictional drug concentration data)
. pkcollapse time conc1 conc2, id(id) keep(seq) stat(auc)
.............................
. pkshape id seq auc*, order(RT TR)
. sort period id
. list, sep(4)
```

	id	sequence	outcome	treat	carry	period
1.	1	RT	150.9643	R	0	1
2.	2	RT	146.7606	R	0	1
3.	3	RT	160.6548	R	0	1
4.	4	RT	157.8622	R	0	1
5.	5	RT	133.6957	R	0	1
6.	6	RT	160.639	R	0	1
7.	7	RT	131.2604	R	0	1
8.	8	RT	168.5186	R	0	1
9.	9	TR	137.0627	T	0	1
10.	10	TR	153.4038	T	0	1
11.	11	TR	163.4593	T	0	1
12.	12	TR	146.0462	T	0	1
			(*output omitted*)			
25.	9	TR	139.7382	R	T	2
26.	10	TR	202.3942	R	T	2
27.	11	TR	136.7848	R	T	2
28.	12	TR	104.5191	R	T	2
29.	13	TR	165.8654	R	T	2
30.	14	TR	139.235	R	T	2
31.	15	TR	166.2391	R	T	2
32.	16	TR	158.5146	R	T	2

The outcome variable is the AUC, and we use pkequiv to conduct a bioequivalence test between reference drug "R" and test drug "T" using the log scale.

```
. set seed 123
. pkequiv outcome treat period sequence id
Log-transformed confidence interval for bioequivalence
Outcome:    ln(outcome)
Treatment:  treat = R (reference)
            treat = T (test)
Period:     period
Sequence:   sequence
Subject ID: id
```

	[Equivalence limits]		[90% conf. interval]		Estimate
Geometric mean ratio (%)	80	125	91.087	116.210	102.885

```
Bootstrap prob. that confidence interval is within equivalence limits = 0.901
Schuirmann's two one-sided tests
    t statistics: Upper = -2.816                    p-value = 0.0069
                  Lower =  3.638                    p-value = 0.0013
```

The equivalence table for the log-transformed CI contains columns for the equivalence limits (which can be controlled by using the limit() option), the confidence interval (the width of which can be controlled by using the level() option), and the point estimate. The default value of limit(0.2) yields equivalence limits of [80%, 125%] for the ratio of geometric means. The default value of level(90) specifies a 90% CI.

The CI of [91.087%, 116.210%] is symmetric around the estimated geometric mean ratio of 102.885%. The interval is entirely contained within the equivalence limits, so we declare average bioequivalence between the test and reference drugs.

To determine the probability that the CI would lie within the equivalence limits if this experiment were to be repeated, pkequiv conducts a bootstrap with 1,000 replications. In this example, there is a 90.1% assurance that the observed CI would be within the equivalence limits in the long run. Because of the nature of the bootstrap, we expect the bootstrap probability calculation to vary slightly from run to run, so we specified the command set seed 123 before executing pkequiv for you to match the bootstrap probability exactly when you run this example.

The output also displays Schuirmann's two one-sided tests. The upper one-sided test is a formal statistical test of whether the geometric mean ratio exceeds the upper equivalence limit, and the lower one-sided test evaluates whether the geometric mean ratio is below the lower equivalence limit. The null hypothesis for both of Schuirmann's two one-sided tests is that the two drugs are not bioequivalent, so rejecting both of these null hypotheses supports a conclusion of average bioequivalence.

Schuirmann's two one-sided tests are constructed by inverting the formula for the CI (see *Schuirmann's two one-sided tests* in *Methods and formulas*), so conducting the TOST at the α level will always agree with the conclusion of bioequivalence or bioinequivalence reached by comparing the $(1-2\alpha)100\%$ CI against the equivalence limits. The default of level(90) yields $\alpha = 0.05$, and because we have already demonstrated that the 90% CI is within its equivalence limits, it comes as no surprise that we can reject the null hypotheses of both one-sided tests at the 0.05 level.

We now repeat this same analysis using the classic shortest CI instead of the log-transformed CI. For speed, we specify the noboot option to omit the bootstrap.

```
. pkequiv outcome treat period sequence id, classic noboot
Classic confidence interval for bioequivalence
Outcome:    outcome
Treatment:  treat = R (reference)
            treat = T (test)
Period:     period
Sequence:   sequence
Subject ID: id
```

	[Equivalence limits]		[90% conf. interval]		Estimate
Arithmetic mean difference	-30.296	30.296	-11.332	26.416	7.542
ratio (%)	80	120	92.519	117.439	104.979

```
Schuirmann's two one-sided tests
    t statistics: Upper = -2.123                    p-value = 0.0260
                  Lower =  3.531                    p-value = 0.0017
```

The equivalence table for the classic CI has two rows: one for the difference between the arithmetic means of the test and reference groups and one for the ratio of arithmetic means. The default value of limit(0.2) indicates that bioequivalence is declared if the CI for the difference is within the range $[-30.296, 30.296]$, which corresponds to a CI for the ratio within [80%, 120%]. If the CI for the difference is within its equivalence limits, the CI for the ratio must be too, and vice versa.

The 90% CI for the ratio of arithmetic means, [92.519%, 117.439%], is not exactly the same as the 90% CI for the ratio of geometric means, but it is entirely contained within its equivalence limits, so it yields the same conclusion of bioequivalence. However, now the upper equivalence limit for the ratio is only 120% (as opposed to 125% when conducting log-scale analysis), so here the upper endpoint of the CI comes closer to its equivalence limit. As expected, the 90% CI for the difference between means, $[-11.332, 26.416]$, is also contained within its equivalence limits, supporting a conclusion of bioequivalence.

When combined with the classic option, Schuirmann's two one-sided tests are performed on the untransformed outcomes, which is why the test statistics and p-values above differ slightly from those reported with the log-transformed CI in the first part of this example. But this ensures that the TOST with level α tests will agree with the conclusion of bioequivalence or bioinequivalence reached by comparing the classic $(1 - 2\alpha)100\%$ CI against the equivalence limits.

Schuirmann's upper one-sided test presents evidence that the difference between the average AUCs of the test and reference drugs does not exceed the upper equivalence limit ($p = 0.026$), while Schuirmann's lower one-sided test presents evidence that the difference between the average AUCs of the test and reference drugs is not less than the lower equivalence limit ($p = 0.0017$). Considering these two tests together, we reject the null hypothesis of bioinequivalence at the $\alpha = 0.05$ level, which is the same conclusion that we reached when we compared the CIs with their equivalence limits.

◁

▷ Example 2

We now use the data published in Chow and Liu (2009, 71), which we describe in example 1 of [R] **pkshape**. As before, the *outcome* variable is AUC, and we use `pkshape` to reshape the data. We use the same `pkequiv` specification as we did in example 1 above to conduct a bioequivalence test.

```
. use https://www.stata-press.com/data/r19/chowliu, clear
. pkshape id seq period1 period2, order(RT TR)
. set seed 123
. pkequiv outcome treat period sequence id
Log-transformed confidence interval for bioequivalence
Outcome:    ln(outcome)
Treatment:  treat = R (reference)
            treat = T (test)
Period:     period
Sequence:   sequence
Subject ID: id
```

	[Equivalence limits]		[90% conf. interval]		Estimate
Geometric mean ratio (%)	80	125	88.313	106.928	97.175

```
Bootstrap prob. that confidence interval is within equivalence limits = 0.985
Schuirmann's two one-sided tests
    t statistics: Upper = -4.521                    p-value = 0.0001
                  Lower =  3.492                    p-value = 0.0010
```

For these data, the 90% CI for the ratio of treatment geometric means is well within the equivalence limits, supporting a conclusion of average bioequivalence. The bootstrap probability that the CI is within the equivalence limits is 98.5%, indicating that if this experiment were to be repeated many times, the vast majority of replications would support a conclusion of bioequivalence. Examining the TOST, we reject both null hypotheses at the 5% level, which agrees with the conclusion we reached by comparing the 90% CI against the equivalence limits.

If we want to conduct a test of average bioequivalence using the original scale of the data, we can employ Fieller's theorem to calculate an exact CI for the ratio of arithmetic means. Here we repeat the previous analysis, but we add the `fieller` option to calculate a CI based on Fieller's theorem.

```
. pkequiv outcome treat period sequence id, fieller
Confidence interval for bioequivalence based on Fieller's theorem
Outcome:    outcome
Treatment:  treat = R (reference)
            treat = T (test)
Period:     period
Sequence:   sequence
Subject ID: id
```

	[Equivalence limits]		[90% conf. interval]		Estimate
Arithmetic mean ratio (%)	80	120	89.787	105.193	97.229

The exact 90% CI for the ratio of treatment arithmetic means is quite similar to the CI for the ratio of geometric means. Because we performed this analysis using the original scale of the data, the upper equivalence limit for the ratio is 120%, not 125% (the value used for log-scale analysis). But the narrow CI for the ratio of arithmetic means is well contained within these narrower equivalence limits, supporting a conclusion of average bioequivalence.

◁

Stored results

pkequiv stores the following in r():

Scalars

r(ratio)	estimated ratio of treatment means
r(ratio_l)	lower CI limit for ratio of treatment means
r(ratio_u)	upper CI limit for ratio of treatment means
r(ratio_limit_l)	lower equivalence limit for ratio of treatment means
r(ratio_limit_u)	upper equivalence limit for ratio of treatment means
r(diff)	estimated difference between treatment means, if classic or symmetric specified
r(diff_l)	lower CI limit for difference between means , if classic or symmetric specified
r(diff_u)	upper CI limit for difference between means, if classic or symmetric specified
r(diff_limit_l)	lower equivalence limit for difference between means, if classic or symmetric specified
r(diff_limit_u)	upper equivalence limit for difference between means, if classic or symmetric specified
r(bootprob)	bootstrap probability that CI is within equivalence limits, if bootstrap performed
r(reps)	number of bootstrap replications, if bootstrap performed
r(tost_l)	lower test statistic for Schuirmann's two one-sided tests
r(tost_u)	upper test statistic for Schuirmann's two one-sided tests
r(p_tost_l)	lower p-value for Schuirmann's two one-sided tests
r(p_tost_u)	upper p-value for Schuirmann's two one-sided tests
r(anderson)	test statistic for Anderson and Hauck's test, if anderson specified
r(ncp_anderson)	noncentrality parameter for Anderson and Hauck's test, if anderson specified
r(p_anderson)	p-value for Anderson and Hauck's test, if anderson specified
r(sd)	pooled-sample standard deviation of period differences from both sequences
r(delta)	delta value used in calculating a symmetric CI, if symmetric specified
r(limit)	equivalence limit
r(level)	confidence level used for calculating CIs
r(refval)	value of reference treatment
r(testval)	value of test treatment

Macros

r(outcome)	name of outcome variable
r(treatment)	name of treatment variable
r(period)	name of period variable
r(sequence)	name of sequence variable
r(id)	name of id variable
r(citype)	log, classic, fieller, or symmetric
r(reflab)	value label of reference treatment
r(testlab)	value label of test treatment

Matrices

r(limit_table)	table of results

Methods and formulas

Methods and formulas are presented under the following headings:

Data and model

Let y_{ij} be the observed PK outcome for subject $i = 1, \ldots, n$ in period $j = 1, 2$. Let $\mathcal{RT}$ be the set of subject IDs for subjects who received the reference drug before the test drug, which we will designate as sequence 1. Let $\mathcal{TR}$ be the set of IDs for subjects who received the test drug before the reference drug, which we will call sequence 2. Let n_1 and n_2 represent the number of subjects in sequences 1 and 2, respectively, with $n = n_1 + n_2$.

Consider a model for the log-transformed outcome first:

$$\ln(y_{ij}) = \mu^*_{ij} + s^*_i + p^*_j + \epsilon^*_{ij} \tag{1}$$

s^*_i is the random effect of subject i, where $s^*_i \sim$ i.i.d. $N(0, \sigma^{*2}_s)$ and σ^{*2}_s controls between-subjects variability. p^*_j is the fixed effect of period j, subject to the constraint $p^*_1 + p^*_2 = 0$. Random error ϵ^*_{ij} is independent of s^*_i, and $\epsilon^*_{ij} \sim$ i.i.d. $N(0, \sigma^{*2}_e)$. Fixed effect μ^*_{ij} is defined as follows,

$$\mu^*_{ij} = \begin{cases} \mu^*_R & \text{if } (i \in \mathcal{RT} \text{ and } j = 1) \text{ or } (i \in \mathcal{TR} \text{ and } j = 2) \\ \mu^*_T & \text{if } (i \in \mathcal{RT} \text{ and } j = 2) \text{ or } (i \in \mathcal{TR} \text{ and } j = 1) \end{cases}$$

where μ^*_R and μ^*_T are the population means of the log-transformed PK outcomes for the reference and test drugs, respectively. The model above assumes there are no unequal carryover effects.

By default, `pkequiv` produces a CI for the geometric mean ratio of the PK outcome, defined as $R_g = \exp(\mu^*_T)/\exp(\mu^*_R)$. `pkequiv` also produces CIs for the difference $D_a = \mu_T - \mu_R$ and ratio $R_a = \mu_T/\mu_R$ of arithmetic means of the PK outcome. The arithmetic means μ_T and μ_R are estimated from (1) applied to the untransformed PK outcome,

$$y_{ij} = \mu_{ij} + s_i + p_j + \epsilon_{ij} \tag{2}$$

$$\mu_{ij} = \begin{cases} \mu_R & \text{if } (i \in \mathcal{RT} \text{ and } j = 1) \text{ or } (i \in \mathcal{TR} \text{ and } j = 2) \\ \mu_T & \text{if } (i \in \mathcal{RT} \text{ and } j = 2) \text{ or } (i \in \mathcal{TR} \text{ and } j = 1) \end{cases}$$

where μ_R and μ_T are the population means of the untransformed PK outcomes for the reference and test drugs, respectively. Subject-specific random effect $s_i \sim$ i.i.d. $N(0, \sigma^2_s)$ is independent of random error $\epsilon_{ij} \sim$ i.i.d. $N(0, \sigma^2_e)$, where σ^2_s controls between-subjects variability and σ^2_e controls within-subjects variability. Period effect p_j is defined similarly to (1).

Equivalence limits

Equivalence limits are controlled by using the `limit()` option. Depending on the type of CI requested, equivalence limits will be calculated for the ratio of treatment means, the difference between treatment means, or both.

The lower equivalence limit for the ratio of treatment means is $l_R = \{1 - \texttt{limit()}\}100\%$. The upper equivalence limit for the ratio is $u_R = [1/\{1 - \texttt{limit()}\}]100\%$ when performing analysis on the log scale, and it is $u_R = \{1 + \texttt{limit()}\}100\%$ when performing calculations on the original scale of the data. The default value of `limit()` is 0.2, which yields equivalence limits of 80% to 125% when using the log scale and equivalence limits of 80% to 120% when using the original scale.

The lower and upper equivalence limits for the difference between treatment arithmetic means are

$$l_D = -\texttt{limit()} \times \hat{\mu}_R \qquad \text{and} \qquad u_D = \texttt{limit()} \times \hat{\mu}_R \tag{3}$$

respectively. The estimated marginal mean $\hat{\mu}_R$, also known as the least-squares mean, is used for calculations because the true value of μ_R in the population is unknown.

When we conduct inference using the ratio of treatment means, bioequivalence is declared when $(l_R < L_R)$ and $(U_R < u_R)$, where L_R and U_R are the lower and upper endpoints, respectively, of the CI for the ratio. When we consider the difference between treatment means, average bioequivalence is declared when $(l_D < L_D)$ and $(U_D < u_D)$, where L_D is the lower endpoint of the CI for the difference and U_D is the upper endpoint.

Confidence intervals

`pkequiv` supports four types of CIs: a CI calculated using the log scale, the classic shortest CI, a CI symmetric around zero, and a CI based on Fieller's theorem. The width of the CI is controlled by the `level()` option, with a default of `level(90)` yielding a 90% CI. This option also sets the significance level α of the bioequivalence test, with the default of $\alpha = 0.05$ given by the relationship $\alpha = \{100 - \texttt{level()}\}/200$. The formula for the log-scale CI is based on the formula for the classic CI, which is presented first.

Classic interval

The lower limit of the classic shortest CI for the difference between two treatment means is

$$L_{D,\text{classic}} = (\hat{\mu}_T - \hat{\mu}_R) - t_{(\alpha,n_1+n_2-2)}\hat{\sigma}_d\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{4}$$

and the upper limit is

$$U_{D,\text{classic}} = (\hat{\mu}_T - \hat{\mu}_R) + t_{(\alpha,n_1+n_2-2)}\hat{\sigma}_d\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5}$$

Here $\hat{\mu}_R$ and $\hat{\mu}_T$ are the estimated marginal means of the reference and test drugs, respectively, estimated using (2). The term $t_{(\alpha,n_1+n_2-2)}$ represents the upper α critical value of a central t distribution with $n_1 + n_2 - 2$ degrees of freedom. $\hat{\sigma}_d$ is the pooled sample standard deviation of the period differences from both sequences, defined as

$$\hat{\sigma}_d = \sqrt{\frac{1}{n_1 + n_2 - 2}\Big\{ \sum_{i\in\mathcal{RT}} \left(d_i - \overline{d}_{RT}\right)^2 + \sum_{i\in\mathcal{TR}} \left(d_i - \overline{d}_{TR}\right)^2 \Big\}} \tag{6}$$

The period difference for subject i is $d_i = (y_{i2} - y_{i1})/2$, the average period difference in sequence 1 is $\overline{d}_{RT} = (1/n_1)\sum_{i\in\mathcal{RT}} d_i$, and the average period difference in sequence 2 is $\overline{d}_{TR} = (1/n_2)\sum_{i\in\mathcal{TR}} d_i$.

The limits of the approximate CI for the ratio are

$$L_{R,\text{classic}} = \left(\frac{L_{D,\text{classic}}}{\hat{\mu}_R} + 1\right) 100\%$$

and

$$U_{R,\text{classic}} = \left(\frac{U_{D,\text{classic}}}{\hat{\mu}_R} + 1\right) 100\%$$

This CI for the ratio is approximate because the denominator $\hat{\mu}_R$ is an estimate of the true population average μ_R. The conclusion of average bioequivalence will be the same regardless of whether you compute the classic CI for the difference or for the ratio. If one CI is within its equivalence limits, the other one necessarily is too.

The finite sample performance of the classic CI is assessed via bootstrap simulation. Bootstrap samples are drawn with replacement using the patient IDs as clusters, and the default of `reps(1000)` calls for 1,000 bootstrap samples. For each sample, a classic CI is constructed and compared with the equivalence limits. The bootstrap probability is computed as the proportion of CIs that are within the equivalence limits.

Intervals using log-transformed data

Calculation of the log-scale CI for the ratio of treatment geometric means, $[L_{R,\log}, U_{R,\log}]$, begins with taking the natural logarithm of the PK outcome and using (1) to calculate estimated marginal means $\hat{\mu}^*_R$ and $\hat{\mu}^*_T$.

Next, the estimated standard deviation of log-scale period differences, $\hat{\sigma}^*_d$, is calculated by substituting $\ln(y_{ij})$ for y_{ij} in (6). A classic CI for the difference between means of the log-PK metric is calculated, substituting $\hat{\mu}^*_R$ for $\hat{\mu}_R$ and $\hat{\mu}^*_T$ for $\hat{\mu}_T$ in (4) and (5).

The endpoints of this interval, $L_{D,\log}$ and $U_{D,\log}$, are exponentiated to yield $L_{R,\log} = \exp(L_{D,\log})$ and $U_{R,\log} = \exp(U_{D,\log})$. These are the endpoints of an exact CI for the geometric mean ratio: $R_g = \exp(\mu^*_T)/\exp(\mu^*_R)$.

Use of this approach is recommended by the US FDA (2001, 2022) and the European Medicines Agency (2010). PK metrics such as AUC and C_{max} are often right-skewed, but the US FDA's rationale for requesting log-scale bioequivalence analysis is not based on the distribution of the data but instead on properties of the model of log-PK data. The US FDA has, since 1991, recommended that analysis of bioequivalence be conducted using the ratio of treatment geometric means (US FDA 2001). Additionally, the models postulated for PK parameters in bioequivalence studies contain a multiplicative subject effect when analyzed on the original scale, but this effect is additive on the logarithmic scale (US FDA 2022).

The finite sample performance of the log-transformed CI is assessed via bootstrap simulation. This is equivalent to the bootstrap procedure used to determine the long-run performance of the classic CI, where patient IDs are used as clusters during the resampling process.

Intervals symmetric around zero

The lower and upper limits for the symmetric CI are, respectively, $L_{D,\text{sym}} = \hat{\mu}_R - \Delta$ and $U_{D,\text{sym}} = \hat{\mu}_R + \Delta$. To calculate Δ, we must solve a system of two equations,

$$\Delta = b_1 \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} - (\hat{\mu}_R - \hat{\mu}_T)$$

and (simultaneously)

$$\Delta = -b_2 \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} + (\hat{\mu}_R - \hat{\mu}_T)$$

Values b_1 and b_2 are computed iteratively to satisfy the above equalities and the condition

$$\int_{b_1}^{b_2} f(t)dt = 1 - 2\alpha$$

where $f(t)$ is the probability density function of a central t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Intervals using Fieller's theorem

To calculate an exact CI for the ratio of treatment means without using the log scale, Locke (1984) employed Fieller's theorem (1954) to obtain $[L_{R,\text{Fieller}}, U_{R,\text{Fieller}}]$. See Locke (1984) or Chow and Liu (2009, 88–92) for details about calculating the exact CI based on Fieller's theorem.

Interval hypothesis tests

`pkequiv` offers two types of interval hypothesis tests of bioequivalence: Schuirmann's two one-sided tests (1987) and Anderson and Hauck's test (1983). The null hypothesis for both types of tests is that the two treatments are not bioequivalent, so bioequivalence is declared when the null hypotheses are rejected.

Schuirmann's two one-sided tests

The hypotheses for Schuirmann's lower one-sided test are H_{0L}: $\mu_T - \mu_R \leq l_D$ and H_{aL}: $\mu_T - \mu_R > l_D$. For the upper one-sided test, the hypotheses are H_{0U}: $\mu_T - \mu_R \geq u_D$ and H_{aU}: $\mu_T - \mu_R < u_D$. The test statistics for the TOST are

$$T_L = \frac{(\hat{\mu}_T - \hat{\mu}_R) - l_D}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{and} \quad T_U = \frac{(\hat{\mu}_T - \hat{\mu}_R) - u_D}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{7}$$

The p-values for the lower and upper one-sided tests are $p_L = F_t(T_L)$ and $p_U = 1 - F_t(T_U)$, respectively, where F_t is the cumulative distribution function of the central t distribution with n_1+n_2-2 degrees of freedom.

Schuirmann's two one-sided tests with a classic CI

Schuirmann's two one-sided tests are inversions of the formulas for the endpoints of the classic CI for the difference in treatment means. Thus, rejecting H_{0L} at the α level is the same as concluding $l_D < L_{D,\text{classic}}$ for a $(1-2\alpha)100\%$ CI, and rejecting H_{0U} is the same as concluding $U_{D,\text{classic}} < u_D$. If both null hypotheses are rejected, the entire CI lies within the equivalence limits.

Schuirmann's two one-sided tests with a log-scale CI

The log-scale CI is calculated using the same formula as the classic CI, but the outcome is the natural log of the PK metric. Thus, the procedure for conducting the TOST on the log scale is the same as on the original scale, but the log-transformed PK outcome is used.

In (7), terms $\hat{\mu}_R$, $\hat{\mu}_T$, and $\hat{\sigma}_d$ are replaced with the values used in the calculation of the log-scale CI: $\hat{\mu}^*_R$, $\hat{\mu}^*_T$, and $\hat{\sigma}^*_d$. The terms l_D and u_D in (7) are likewise replaced with l^*_D and u^*_D, the equivalence limits for the difference between log-transformed treatment means, which are calculated by replacing $\hat{\mu}_R$ with $\hat{\mu}^*_R$ in (3).

Rejecting both null hypotheses of the TOST on the log scale is equivalent to concluding $l^*_D < L_{D,\log}$ and $U_{D,\log} < u^*_D$, which, in turn, is equivalent to concluding $l_R < L_{R,\log}$ and $U_{R,\log} < u_R$.

Anderson and Hauck's test

Anderson and Hauck's test may be conducted only when calculating the classic shortest CI on the original scale of the data. The null hypothesis for Anderson and Hauck's test is $H_0 : (\mu_T - \mu_R \leq l_D) \cup (\mu_T - \mu_R \geq u_D)$. The alternative hypothesis for Anderson and Hauck's test is $H_a : l_D < (\mu_T - \mu_R) < u_D$. The test statistic is

$$T_{\text{AH}} = \frac{(\hat{\mu}_T - \hat{\mu}_R) - (l_D + u_D)/2}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and the noncentrality parameter is estimated by

$$\hat{\delta} = \frac{u_D - l_D}{2\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The p-value is calculated as

$$p = F_t\left(|T_{\text{AH}}| - \hat{\delta}\right) - F_t\left(-|T_{\text{AH}}| - \hat{\delta}\right)$$

where F_t is the cumulative distribution function of the central t distribution with $n_1 + n_2 - 2$ degrees of freedom.

References

Anderson, S., and W. W. Hauck. 1983. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics—Theory and Methods* 12: 2663–2692. https://doi.org/10.1080/03610928308828634.

Chow, S.-C., and J.-P. Liu. 2009. *Design and Analysis of Bioavailability and Bioequivalence Studies*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781420011678.

European Medicines Agency. 2010. Guideline on the Investigation of Bioequivalence. Doc. Ref.: CPMP/EWP/QWP/1401/98 Rev. 1/Corr. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf.

Fieller, E. C. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society*, B ser., 16: 175–185. https://doi.org/10.1111/j.2517-6161.1954.tb00159.x.

Frick, H. 1987. On level and power of Anderson and Hauck's procedure for testing equivalence in comparative bioavailability. *Communications in Statistics—Theory and Methods* 16: 2771–2778. https://doi.org/10.1080/03610928708829538.

Kirkwood, T. B. L. 1981. Bioequivalence testing—a need to rethink. *Biometrics* 37: 589–594. https://doi.org/10.2307/2530573.

Locke, C. S. 1984. An exact confidence interval from untransformed data for the ratio of two formulation means. *Journal of Pharmacokinetics and Biopharmaceutics* 12: 649–655. https://doi.org/10.1007/BF01059558.

Mantel, N. 1977. Do we want confidence intervals symmetrical about the null value? *Biometrics* 33: 759–760.

Schuirmann, D. 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: 657–680. https://doi.org/10.1007/BF01068419.

US Food and Drug Administration. 1988. Report by the Bioequivalence Task Force on Recommendations from the Bioequivalence Hearing Conducted by the Food and Drug Administration, 9/29–10/1/1986.

———. 2001. Statistical Approaches to Establishing Bioequivalence: Guidance for Industry. Docket No.01D-0027. https://www.fda.gov/media/70958/download.

———. 2022. Statistical Approaches to Establishing Bioequivalence: Guidance for Industry (Draft Guidance). Docket No. FDA-2001-D-0197. https://www.fda.gov/media/163638/download.

Westlake, W. J. 1976. Symmetrical confidence intervals for bioequivalence trials. *Biometrics* 32: 741–744. https://doi.org/10.2307/2529259.

Also see

[R] **pk** — Pharmacokinetic (biopharmaceutical) data