gsbounds — Boundaries for group sequential trials

Description

`gsbounds` computes stopping boundaries for group sequential designs (GSDs), a class of experimental design popular in clinical trials. GSDs incorporate planned interim analyses, or looks at the data, and provide criteria for stopping the trial early based on values of a test statistic. Stopping can be for efficacy, futility, or both. For a software-free introduction to GSDs, see [ADAPT] **GSD intro**; for an introduction to Stata's gs suite of commands, see [ADAPT] **gs**, and for associated sample-size calculations, see [ADAPT] **gsdesign**.

Quick start

Calculate boundaries using the default settings: a two-sided O'Brien–Fleming design with two evenly spaced analyses (one interim look, one final look), power of 0.8, and familywise significance level $\alpha = 0.05$

```
gsbounds
```

Same as above, but add a nonbinding O'Brien–Fleming futility boundary and conduct three evenly spaced analyses

```
gsbounds, efficacy(obfleming) futility(obfleming) nlooks(3)
```

Same as above, but plan the looks to occur with 50%, 75%, and 100% of the data, and visualize the bounds on a graph

```
gsbounds, efficacy(obfleming) futility(obfleming) ///
    information(0.5 0.75 1) graphbounds
```

Same as above, but use error-spending approximations of O'Brien–Fleming bounds

```
gsbounds, efficacy(errobfleming) futility(errobfleming) ///
    information(0.5 0.75 1) graphbounds
```

Nonbinding futility boundaries for an upper one-sided test using a five-look Wang–Tsiatis design with parameter $\Delta_f = 0.3$, power of 0.9, and significance level $\alpha = 0.01$

```
gsbounds, alpha(0.01) power(0.9) futility(wtsiatis(0.3)) nlooks(5) upper
```

Same as above, but use a binding futility bound

```
gsbounds, alpha(0.01) power(0.9) futility(wtsiatis(0.3), binding) ///
    nlooks(5) upper
```

Efficacy and nonbinding futility boundaries for a lower one-sided test using a seven-look error-spending Hwang–Shih–de Cani design with efficacy parameter $\gamma_e = -2$, futility parameter $\gamma_f = -4$, power of 0.9, and significance level $\alpha = 0.01$

```
gsbounds, alpha(0.01) power(0.9) efficacy(hsdecani(-2)) ///
    futility(hsdecani(-4)) nlooks(7) lower
```

Same as above, but use a binding Kim–DeMets futility boundary with parameter $\rho_f = 2.5$, and graph the boundaries but not the fixed-sample critical values

```
gsbounds, alpha(0.01) power(0.9) efficacy(hsdecani(-2)) ///
      futility(kdemets(2.5), binding) nlooks(7) lower ///
      graphbounds(nofixed)
```

Menu

Statistics > Power, precision, and sample size

Syntax

Calculate efficacy stopping boundaries

```
gsbounds [ , efficacy(boundary) options ]
```

Calculate futility stopping boundaries

```
gsbounds, futility(boundary[ , binding]) [ options ]
```

Calculate efficacy and futility stopping boundaries

```
gsbounds, efficacy(boundary) futility(boundary[ , binding]) [ options ]
```

boundary	Description
obfleming	classical O'Brien–Fleming bound
pocock	classical Pocock bound
wtsiatis(#)	classical Wang–Tsiatis bound with specified parameter value
errpocock	error-spending Pocock-style bound
errobfleming	error-spending O'Brien–Fleming-style bound
kdemets(#)	error-spending Kim–DeMets bound with specified parameter value
hsdecani(#)	error-spending Hwang–Shih–de Cani bound with specified parameter value

options	Description
Main	
`efficacy(`*boundary*`)`	boundary for efficacy stopping; if neither `efficacy()` nor `futility()` is specified, the default is `efficacy(obfleming)`
`futility(`*boundary*`[ , binding ])`	boundary for futility stopping; use `binding` to request binding futility bounds (default is nonbinding)
`nlooks(`#`)`	total number of analyses (`nlooks()` − 1 interim analyses and one final analysis)
`information(`*numlist*`)`	sequence of information levels for analyses; default is evenly spaced
`nopvalues`	suppress p-values
`alpha(`#`)`	overall significance level for all tests; default is `alpha(0.05)`
`power(`#`)`	overall power for all tests; default is `power(0.8)`
`beta(`#`)`	overall probability of type II error for all tests; default is `beta(0.2)`
`upper`	upper one-sided test; default is two-sided
`lower`	lower one-sided test; default is two-sided
`onesided`	synonym for `upper`
Graph	
`graphbounds[ (`*graphopts*`) ]`	graph boundaries
`matlistopts(`*general_options*`)`	control the display of boundaries; seldom used
optimopts	optimization options for boundary calculations; seldom used

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

`matlistopts()` and *optimopts* do not appear in the dialog box.

graphopts	Description
`xdiminformation`	label the x axis with the information fraction (default); use information levels if `information()` specified
`xdimlooks`	label the x axis with the number of each look
`noshade`	do not shade the rejection, acceptance, and continuation regions
`rejectopts(`*area_options*`)`	change the appearance of the rejection region
`acceptopts(`*area_options*`)`	change the appearance of the acceptance region
`continueopts(`*area_options*`)`	change the appearance of the continuation region
`efficacyopts(`*connected_options*`)`	change the appearance of the efficacy bound
`futilityopts(`*connected_options*`)`	change the appearance of the futility bound
`nolooklines`	do not draw vertical reference lines at each look
`looklinesopts(`*added_line_suboptions*`)`	change the appearance of the reference lines marking each look
`nofixed`	do not label critical values from a fixed study design
`fixedopts(`*marker_options*`)`	change the appearance of the fixed-study critical values
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***

optimopts	Description
intpointsscale(#)	scaling factor for number of quadrature points; default is intpointsscale(20)
initinfo(*initinfo_spec*)	initial value(s) for maximum information
initscale(#)	initial value for scaling factor C of classical bounds
infotolerance(#)	tolerance for bisection search for maximum information of error-spending bounds with futility stopping; default is infotol(1e-6)
marquardt	use the Marquardt stepping algorithm in nonconcave regions; default is to use a mixture of steepest descent and Newton
technique(*algorithm_spec*)	maximization technique
iterate(#)	perform maximum of # iterations; default is iterate(300)
[no]log	display an iteration log; default is nolog
trace	display current parameter vector in iteration log
gradient	display current gradient vector in iteration log
showstep	report steps within an iteration in iteration log
hessian	display current negative Hessian matrix in iteration log
showtolerance	report the calculated result that is compared with the effective convergence criterion
tolerance(#)	tolerance for the parameter being optimized; default is tolerance(1e-12)
ftolerance(#)	tolerance for the objective function; default is ftolerance(1e-10)
nrtolerance(#)	tolerance for the scaled gradient; default is nrtolerance(1e-16)
nonrtolerance	ignore the nrtolerance() option

Options

Main

efficacy(*boundary*) specifies the boundary for efficacy stopping. If neither efficacy() nor futility() is specified, the default is efficacy(obfleming).

futility(*boundary*[, binding]) specifies the boundary for futility stopping.

binding specifies binding futility bounds. With binding futility bounds, if the result of an interim analysis crosses the futility boundary and lies in the acceptance region, the trial must end or risk overrunning the specified type I error. With nonbinding futility bounds, the trial does not need to stop if the result of an interim analysis crosses the futility boundary; the familywise type I error rate is controlled even if the trial continues. By default, futility bounds are nonbinding.

nlooks(#) specifies the total number of analyses to be performed (nlooks() − 1 interim analyses and one final analysis). If neither nlooks() nor information() is specified, the default is nlooks(2).

information(*numlist*) specifies a sequence of information levels for interim and final analyses. This must be a sequence of increasing positive numbers, but the scale is unimportant because the information sequence will be automatically rescaled to ensure the maximum information is reached at the final look. By default, analyses are evenly spaced.

nopvalues suppresses the p-values from being reported in the table of boundaries for each look.

`alpha(#)` sets the overall significance level, which is the familywise type I error rate for all analyses (interim and final). The default is `alpha(0.05)`.

`power(#)` sets the overall power for all analyses. The default is `power(0.8)`. If `beta()` is specified, `power()` is set to be 1 − `beta()`. Only one of `power()` or `beta()` may be specified.

`beta(#)` sets the overall probability of a type II error. The default is `beta(0.2)`. If `power()` is specified, `beta()` is set to be 1 − `power()`. Only one of `beta()` or `power()` may be specified.

`upper` indicates an upper one-sided test, which means that the postulated value of the parameter is larger than the value under the null hypothesis. The default is two-sided.

`lower` indicates a lower one-sided test, which means that the postulated value of the parameter is smaller than the value under the null hypothesis. The default is two-sided.

`onesided` is a synonym for `upper`.

Graph

`graphbounds` and `graphbounds(`*graphopts*`)` produce graphical output showing the stopping boundaries.

graphopts are the following:

`xdiminformation` labels the x axis with the information fraction unless `information()` is specified, in which case information levels will be used. This is the default x-axis label.

`xdimlooks` labels the x axis with the number of each look.

`noshade` suppresses shading of the rejection, acceptance, and continuation regions of the graph.

`rejectopts(`*area_options*`)` affects the rendition of the rejection region. See [G-3] ***area_options***.

`acceptopts(`*area_options*`)` affects the rendition of the acceptance region. See [G-3] ***area_options***.

`continueopts(`*area_options*`)` affects the rendition of the continuation region. See [G-3] ***area_options***.

`efficacyopts(`*connected_options*`)` affects the rendition of the efficacy bound. See [G-3] ***cline_options*** and [G-3] ***marker_options***.

`futilityopts(`*connected_options*`)` affects the rendition of the futility bound. See [G-3] ***cline_options*** and [G-3] ***marker_options***.

`nolooklines` suppresses the vertical reference lines drawn at each look.

`looklinesopts(`*added_line_suboptions*`)` affects the rendition of reference lines marking each look. See *suboptions* in [G-3] ***added_line_options***.

`nofixed` suppresses the fixed-study critical values in the plot.

`fixedopts(`*marker_options*`)` affects the rendition of the fixed-study critical values. See [G-3] ***marker_options***.

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding `by()`. These include options for titling the graph (see [G-3] ***title_options***) and for saving the graph to disk (see [G-3] ***saving_option***).

The following options are available with gsbounds but are not shown in the dialog box:

matlistopts(*general_options*) affects the display of the matrix of boundaries. *general_options* are title(), tindent(), rowtitle(), showcoleq(), coleqonly, colorcoleq(), aligncolnames(), and linesize(); see *general_options* in [P] **matlist**. This option is seldom used.

optimopts control the iterative algorithm used to calculate stopping boundaries:

intpointsscale(#) specifies the scaling factor for the number of quadrature points used during the numerical evaluation of stopping probabilities at each look. The default is intpointsscale(20). See *Methods and formulas*.

initinfo(*initinfo_spec*) specifies either one or two initial values to be used in the iterative calculation of the maximum information.

The syntax initinfo(#) is applicable when using classical group sequential boundaries (Pocock bounds, O'Brien–Fleming bounds, and Wang–Tsiatis bounds), as well as with efficacy-only stopping when using error-spending boundaries (error-spending Pocock-style efficacy bounds, error-spending O'Brien–Fleming-style efficacy bounds, Kim–DeMets efficacy bounds, and Hwang–Shih–de Cani efficacy bounds). The default is to use the information from a fixed study design; see *Methods and formulas*.

The syntax initinfo(# #) is applicable when using error-spending group sequential boundaries with futility stopping (error-spending Pocock-style bounds, error-spending O'Brien–Fleming-style bounds, Kim–DeMets bounds, and Hwang–Shih–de Cani bounds). With this syntax, the first and second numbers specify the lower and upper starting values, respectively, for the bisection algorithm estimating the maximum information. The default is to use the information from a fixed study design for the lower initial value and the information corresponding to a Bonferroni correction for the upper initial value; see *Methods and formulas*. To specify just the lower starting value, use initinfo(# .), and to specify just the upper starting value, use initinfo(. #).

initscale(#) specifies the initial value to be used during the iterative calculation of scaling factor C for classical group sequential boundaries (Pocock bounds, O'Brien–Fleming bounds, and Wang–Tsiatis bounds). The default is to use the z-value corresponding to the specified value of alpha(). See *Methods and formulas*.

infotolerance(#) specifies the tolerance for the bisection algorithm used in the iterative calculation of the maximum information of error-spending group sequential boundaries with futility stopping (error-spending Pocock-style bounds, error-spending O'Brien–Fleming-style bounds, Kim–DeMets bounds, and Hwang–Shih–de Cani bounds). The default is infotolerance(1e-6). See *Methods and formulas*.

marquardt specifies that the optimizer should use the modified Marquardt algorithm when, at an iteration step, it finds that H is singular. The default is to use a mixture of steepest descent and Newton, which is equivalent to the difficult option in [R] **ml**.

technique(*algorithm_spec*) specifies how the objective function is to be maximized. The following algorithms are allowed. For details, see Pitblado, Poi, and Gould (2024).

technique(bfgs) specifies the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

technique(nr) specifies Stata's modified Newton–Raphson (NR) algorithm.

technique(dfp) specifies the Davidon–Fletcher–Powell (DFP) algorithm.

The default is technique(bfgs) when using classical group sequential boundaries (Pocock bounds, O'Brien–Fleming bounds, and Wang–Tsiatis bounds) and also for the second optimization step used to estimate the maximum information with efficacy-only stopping when using error-spending boundaries (error-spending Pocock-style efficacy bounds, error-spending O'Brien–Fleming-style efficacy bounds, Kim–DeMets efficacy bounds, and Hwang–Shih–de Cani efficacy bounds). The default is technique(nr) for the sequential optimization steps used to estimate critical values for error-spending boundaries. You can also switch between two algorithms by specifying the technique name followed by the number of iterations. For example, specifying technique(nr 10 bfgs 20) requests 10 iterations with the NR algorithm followed by 20 iterations with the BFGS algorithm, and then back to NR for 10 iterations, and so on. The process continues until convergence or until the maximum number of iterations is reached.

iterate(#) specifies the maximum number of iterations. If convergence is not declared by the time the number of iterations equals iterate(), an error message is issued. The default value of iterate(#) is the number set using set maxiter, which is 300 by default.

[no]log requests an iteration log showing the progress of the optimization. The default is nolog.

trace adds to the iteration log a display of the current parameter vector.

gradient adds to the iteration log a display of the current gradient vector.

showstep adds to the iteration log a report on the steps within an iteration. This option was added so that developers at StataCorp could view the stepping when they were improving the ml optimizer code. At this point, it mainly provides entertainment.

hessian adds to the iteration log a display of the current negative Hessian matrix.

showtolerance adds to the iteration log the calculated value that is compared with the effective convergence criterion at the end of each iteration. Until convergence is achieved, the smallest calculated value is reported. shownrtolerance is a synonym of showtolerance.

Below, we describe the three convergence tolerances. Convergence is declared when the nrtolerance() criterion is met and either the tolerance() or the ftolerance() criterion is also met.

tolerance(#) specifies the tolerance for the parameter vector. When the relative change in the parameter vector from one iteration to the next is less than or equal to tolerance(), the tolerance() convergence criterion is satisfied. The default is tolerance(1e-12).

ftolerance(#) specifies the tolerance for the objective function. When the relative change in the objective function from one iteration to the next is less than or equal to ftolerance(), the ftolerance() convergence is satisfied. The default is ftolerance(1e-10).

nrtolerance(#) specifies the tolerance for the scaled gradient. Convergence is declared when $\mathbf{gH}^{-1}\mathbf{g}' <$ nrtolerance(). The default is nrtolerance(1e-16).

nonrtolerance specifies that the default nrtolerance() criterion be turned off.

boundary

obfleming specifies a classical O'Brien–Fleming design for efficacy or futility bounds (O'Brien and Fleming 1979). O'Brien–Fleming efficacy bounds are characterized by being extremely conservative at early looks. The O'Brien–Fleming design is a member of the Wang–Tsiatis family and is equivalent to specifying a *boundary* of wtsiatis(0).

`pocock` specifies a classical Pocock design for efficacy or futility bounds (Pocock 1977). Pocock efficacy bounds are characterized by using the same critical value at all looks. The Pocock design is a member of the Wang–Tsiatis family and is equivalent to specifying a *boundary* of `wtsiatis(0.5)`.

`wtsiatis(#)` specifies a classical Wang–Tsiatis design for efficacy or futility bounds (Wang and Tsiatis 1987). The shape of Wang–Tsiatis bounds is determined by parameter $\Delta \in [-10, 0.7]$, where smaller values of Δ yield bounds that are more conservative at early looks.

`errpocock` specifies an error-spending Pocock-style design for efficacy or futility bounds (Lan and DeMets 1983). The critical values from error-spending Pocock-style bounds are very similar to those of classic Pocock bounds, but they are obtained using an error-spending function.

`errobfleming` specifies an error-spending O'Brien–Fleming-style design for efficacy or futility bounds (Lan and DeMets 1983). The critical values from error-spending O'Brien–Fleming-style bounds are very similar to those of classic O'Brien–Fleming bounds, but they are obtained using an error-spending function.

`kdemets(#)` specifies an error-spending Kim–DeMets design for efficacy or futility bounds (Kim and DeMets 1987). The shape of Kim–DeMets bounds is determined by power parameter $\rho \in (0, 10]$, where larger values of ρ yield bounds that are more conservative at early looks.

`hsdecani(#)` specifies an error-spending Hwang–Shih–de Cani design for efficacy or futility bounds (Hwang, Shih, and de Cani 1990). The shape of Hwang–Shih–de Cani bounds is determined by parameter $\gamma \in [-30, 3]$, where smaller values of γ yield bounds that are more conservative at early looks.

For a design with both efficacy and futility stopping boundaries, if you specify a classical boundary (that is, in the Wang–Tsiatis family) for one, then you must specify a classical boundary for the other. So, you could not specify a boundary in the Wang–Tsiatis family for one boundary and an error-spending boundary for the other. When specifying efficacy and futility boundaries from the same family, the efficacy parameter does not need to be the same as the futility parameter.

Boundaries that are conservative at early looks, such as the O'Brien–Fleming bound, offer little chance of early stopping unless the true effect size is quite large (in the case of efficacy bounds) or quite small (in the case of futility bounds). A trial employing a conservative bound is more likely to continue to the final look, yielding an expected sample size that is not dramatically smaller than the sample size required by an equivalent fixed-sample trial. However, the maximum sample size (that is, the sample size at the final look) of a trial with a conservative bound is generally not much greater than the sample size required by an equivalent fixed trial. Another direct result of specifying conservative bounds is that the critical value at the final look tends to be close to the critical value employed by an equivalent fixed design. In contrast, anticonservative boundaries such as the Pocock bound offer a much better shot at early stopping (often yielding a small expected sample size) at the cost of a larger maximum sample size and final critical values that are considerably larger than the critical value of an equivalent fixed design.

Remarks and examples

Remarks are presented under the following headings:

This entry describes the `gsbounds` command and the methodology for calculating stopping boundaries for GSDs. For a software-free introduction to GSDs, see [ADAPT] **GSD intro**; for an introduction to Stata's gs suite of commands, see [ADAPT] **gs**; and for associated sample-size calculations, see [ADAPT] **gsdesign**.

Introduction

Clinical trials, studies investigating the effects of a treatment on human participants, must address ethical concerns that are often not considered when designing other types of experiments. These ethical imperatives, such as not unnecessarily exposing participants to harmful or inferior treatments, must be met while also meeting scientific needs (such as type I error and power) and financial realities that can limit sample sizes.

In a classical fixed-sample design, an experiment of predetermined size is conducted and all data are collected before analysis. This approach is efficient if the data are all collected at once, but in the context of a large clinical trial, participants are typically enrolled over the course of months or years and data about the clinical endpoint are collected bit by bit. In this scenario, GSDs offer a tantalizing prospect: the ability to end a study early when preliminary data are overwhelmingly favorable or unfavorable. Early stopping, without sacrificing type I error, is beneficial because it saves resources and, more importantly, addresses the ethical need to avoid exposing participants to suboptimal treatments unnecessarily.

In a GSD, a number of interim analyses, or looks, are conducted at prespecified points during the collection of experimental data. At each look, the test statistic is calculated based on the data available at the time, and it is compared with critical values defined by the efficacy and futility boundaries. If the test statistic is more extreme than the critical values defined by the efficacy boundaries, then H_0 is rejected and the study is terminated early for efficacy. The complement to efficacy stopping is futility stopping, and if the test statistic crosses the futility boundaries, then H_0 is accepted and the study is terminated early for futility. The concept of accepting H_0, while taboo in many areas, is a long-established practice in GSDs (see *Origins of GSD* in [ADAPT] **GSD intro**) and is often thought of as "abandoning a lost cause" (Gould 1989). If H_0 is neither rejected nor accepted after the interim analysis, the trial continues until the next look.

Stata's `gsbounds` command allows the calculation of stopping boundaries for efficacy and futility, allows for both one-sided and two-sided tests, and implements the most popular boundary calculations. In the examples that follow, the `graphbounds` option is used to visualize the boundaries. The boundaries divide the range of possible test statistic values into regions: the rejection region, the acceptance region, and the continuation region. If the test statistic falls within the rejection region, then H_0 is rejected and the study is terminated due to treatment efficacy. If the test statistic lies within the acceptance region,

then H_0 is accepted and the study is terminated due to futility. If the test statistic is within the continuation region, the study proceeds as planned. Efficacy bounds separate the rejection region from the continuation region, and futility bounds separate the acceptance region from the continuation region. At the final look, there is no continuation region, and H_0 must be accepted or rejected.

Examples

Efficacy stopping

▷ Example 1: Two-sided Pocock efficacy bounds

Consider a two-sided test of the difference between two means with known standard deviations. The standardized test statistic z follows a normal distribution, and we wish to test for efficacy at five equally spaced looks using Pocock efficacy bounds. The familywise type I error allowed is 0.05, while the desired power (at a prespecified clinically significant effect size) is 80%.

We use `gsbounds` to calculate and graph the stopping boundaries and compare them with those of a fixed-sample trial. To calculate Pocock efficacy bounds, we specify the `efficacy(pocock)` option, while the `nlooks(5)` option specifies five equally spaced looks (four interim analyses and a final analysis). The `alpha()` and `power()` options are not specified, which leaves them at their default values of `alpha(0.05)` and `power(0.8)`.

```
. gsbounds, efficacy(pocock) nlooks(5)
Group sequential boundaries
Efficacy: Pocock
Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.8000
Info. ratio = 1.2286
Fixed-study crit. values = ±1.9600
Critical values and p-values for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Efficacy Upper	p-value
1	0.20	-2.4132	2.4132	0.0158
2	0.40	-2.4132	2.4132	0.0158
3	0.60	-2.4132	2.4132	0.0158
4	0.80	-2.4132	2.4132	0.0158
5	1.00	-2.4132	2.4132	0.0158

```
Note: Critical values are for z statistics;
      otherwise, use p-value boundaries.
```

`gsbounds` begins by displaying a summary of the α and power parameters used in the design, followed by a table of stopping boundaries. To facilitate comparing the GSD with a fixed study design, `gsbounds` also displays the fixed-study critical values and the information ratio, which is the ratio of the sample size at the final look of a GSD to the sample size from a fixed study design.

Pocock efficacy bounds are characterized by using the same critical value at all looks. To maintain a familywise type I error of 0.05, Pocock boundaries require that the z statistic reach or exceed ± 2.413 at any look (which corresponds to a p-value of 0.0158) to reject H_0. This is far larger than the critical value of ± 1.96 required by a fixed-sample test. Pocock bounds allow for the possibility of very early stopping if the effect size is large, but if the study continues to the final look, it will require approximately 22.9% more participants than an equivalently powered fixed design, as seen by the information ratio of 1.229.

To plot the bounds for visual inspection, we rerun the previous `gsbounds` command with the `graphbounds` option.

```
. gsbounds, efficacy(pocock) nlooks(5) graphbounds
```
(*output omitted*)

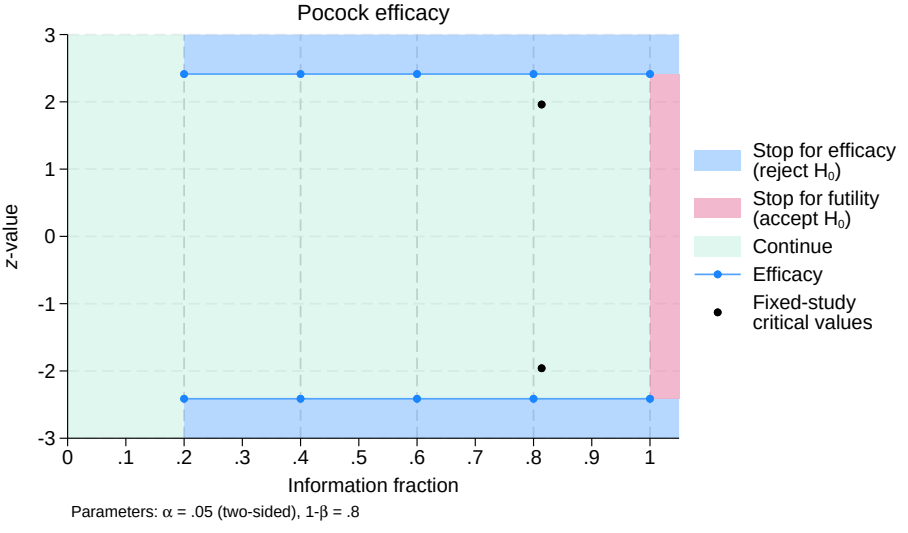

Figure 1. Pocock efficacy bounds

The graph displays the bounds visually, dividing the range of possible z-values into continuation, rejection, and acceptance regions. The vertical axis is the value of the z statistic, and the horizontal axis is the information fraction, or the fraction of the total information that has been collected at the time of the analysis. The information fraction is typically proportional to the sample size, except in time-to-event studies, in which case it is proportional to the number of events observed. The information fraction is reported in the `Info. frac.` column of the table above.

We progress from left to right in the graph as information is collected during the clinical trial. The efficacy bounds, which separate the rejection and continuation regions, are drawn in blue and marked with a dot at each look. Before the first look (that is, when the information fraction is < 0.2), it is impossible to reject H_0 because the data have not yet been analyzed, so all z-values fall within the continuation region. Beginning with the first look, the range of z-values is divided into rejection and continuation regions. Because we are conducting a two-sided test, the rejection region is made up of two areas: z-values ≥ 2.413 and z-values ≤ -2.413.

At the first look, a z test is performed using the command `ztest` or `ztesti`, and z statistic z_1 is calculated; see [R] **ztest**. z_1 is compared with the critical values of the efficacy bounds. If z_1 lies in the rejection region above the efficacy upper bound or below the efficacy lower bound, the null hypothesis

is rejected and the trial is terminated early for treatment efficacy. Mathematically, we would write that we reject H_0 if $z_1 \geq 2.413$ or $z_1 \leq -2.413$. If z_1 lies in the continuation region between the upper and lower efficacy bounds, written as $z_1 \in (-2.413, 2.413)$, then the trial continues.

Because Pocock efficacy bounds use the same critical values for each look, the procedure during the second, third, and fourth looks will be the same. At the final look, there is no continuation region. If $|z_5| < 2.413$, then H_0 is accepted, and if $|z_5| \geq 2.413$, then H_0 is rejected.

The graph also includes points marking the critical values that would be used in an equivalently powered fixed study design. These points appear at z-values of ± 1.96, which give a type I error of 0.05 in a fixed design with a single analysis. Compared with the GSD, the analysis in the fixed design occurs at an information fraction of 0.814. This is calculated as the inverse of the information ratio: $1/1.229 = 0.814$.

At the fifth look, the critical values of the Pocock design are more extreme than the critical values of the fixed design. If $|z_5| \in [1.96, 2.413)$, the researcher will be unable to reject H_0, because they used a Pocock design; they will likely regret not having chosen a fixed design, which would have allowed them to reject H_0 with the same z-value (and a smaller sample).

To avoid this uncomfortable situation, some researchers prefer to use O'Brien–Fleming boundaries, which are demonstrated in the following example.

◁

▷ Example 2: Two-sided O'Brien–Fleming efficacy bounds

O'Brien–Fleming efficacy boundaries are extremely conservative at early looks and far less so at later looks. The final critical values in an O'Brien–Fleming design are similar to those of a fixed study design. Here we calculate O'Brien–Fleming efficacy bounds for the scenario described in the previous example.

```
. gsbounds, efficacy(obfleming) nlooks(5) graphbounds
Group sequential boundaries
Efficacy: O'Brien-Fleming
Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.8000
Info. ratio = 1.0284
Fixed-study crit. values = ±1.9600
Critical values and p-values for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Efficacy Upper	p-value
1	0.20	-4.5617	4.5617	0.0000
2	0.40	-3.2256	3.2256	0.0013
3	0.60	-2.6337	2.6337	0.0084
4	0.80	-2.2809	2.2809	0.0226
5	1.00	-2.0401	2.0401	0.0413

```
Note: Critical values are for z statistics;
      otherwise, use p-value boundaries.
```

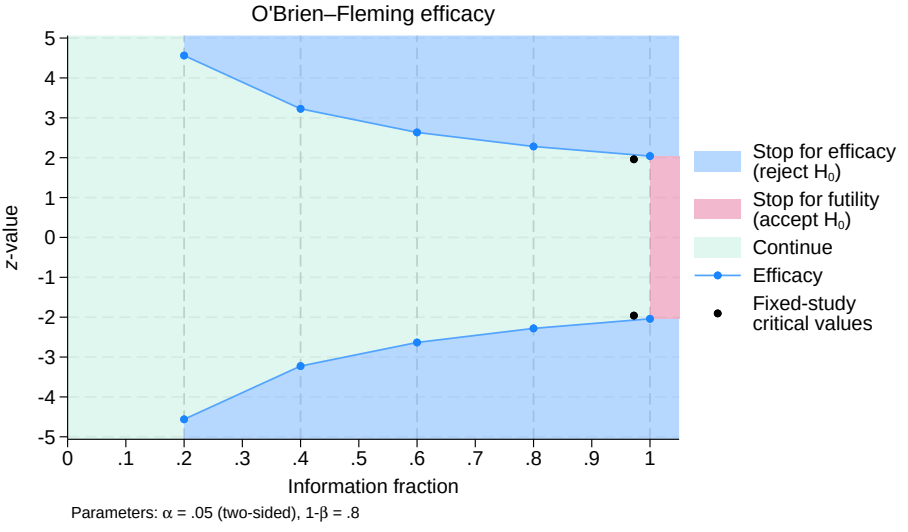

Figure 2. O'Brien–Fleming efficacy bounds

The O'Brien–Fleming design makes it difficult to reject H_0 at early looks but easier at later looks. At the first look, the critical values of ± 4.562 correspond to a p-value of 0.000005, while the critical values at the last look, ± 2.04, correspond to a p-value of 0.0413. The information ratio of 1.028 indicates that the maximum sample size is only 2.8% larger than that of a fixed design.

In the graph, the efficacy bounds take the shape of a funnel with the opening to the left; the continuation region shrinks as more information is collected. By the final look, the critical values of the efficacy bounds are nearly the same as the critical values from a fixed study design. The fixed design uses nearly the same amount of information as the final look of the O'Brien–Fleming design, with the data analysis in the fixed design occurring at information fraction $1/1.028 = 0.97$.

The procedure for interim analysis with O'Brien–Fleming bounds is equivalent to the procedure we used with Pocock bounds, except that the critical values change from one look to the next. At the first look, the continuation region is defined by $|z_1| < 4.562$ and the rejection region by $|z_1| \geq 4.562$. At the second look, the continuation region is defined by $|z_2| < 3.226$ and the rejection region by $|z_2| \geq 3.226$. The pattern continues until the fifth and final look, which has no continuation region. At the fifth look, the acceptance region is defined by $|z_5| < 2.04$ and the rejection region by $|z_5| \geq 2.04$.

◁

▷ Example 3: Two-sided Wang–Tsiatis efficacy bounds

Both Pocock and O'Brien–Fleming boundaries are special cases of a one-parameter family of boundaries described by Wang and Tsiatis (1987). This family of boundaries is indexed by power parameter Δ. Setting $\Delta = 0.5$ yields a Pocock boundary, whereas setting $\Delta = 0$ yields an O'Brien–Fleming boundary. Wang–Tsiatis boundaries with $\Delta \in (0, 0.5)$ offer a balance between the two designs.

We continue example 2, this time calculating Wang–Tsiatis efficacy bounds with power parameter $\Delta_e = 0.25$.

```
. gsbounds, efficacy(wtsiatis(0.25)) nlooks(5) graphbounds
Group sequential boundaries
Efficacy: Wang-Tsiatis, Delta = 0.2500
Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.8000
Info. ratio = 1.0718
Fixed-study crit. values = ±1.9600
Critical values and p-values for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Efficacy Upper	p-value
1	0.20	-3.1941	3.1941	0.0014
2	0.40	-2.6859	2.6859	0.0072
3	0.60	-2.4270	2.4270	0.0152
4	0.80	-2.2586	2.2586	0.0239
5	1.00	-2.1360	2.1360	0.0327

```
Note: Critical values are for z statistics;
      otherwise, use p-value boundaries.
```

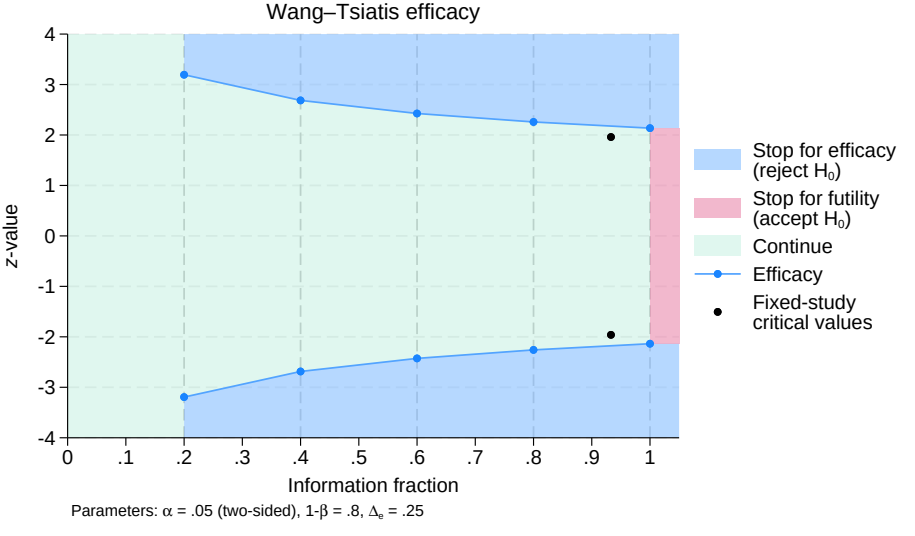

Figure 3. Wang–Tsiatis efficacy bounds, $\Delta = 0.25$

In addition to the values of α and power used to calculate the bounds, `gsbounds` now reports the efficacy parameter for the Wang–Tsiatis bounds. The boundaries themselves are a compromise between the two previous designs. The critical values at early looks are less conservative than those of the O'Brien–Fleming design, making it more likely that a study with a positive result will be stopped very early. At the first look, the critical values of ± 3.194 correspond to a p-value of 0.0014, while the second look critical values of ± 2.686 correspond to a p-value of 0.0072. If the study continues to its conclusion, the final critical values of ± 2.136 correspond to a p-value of 0.0327.

The maximum required sample size is 7.2% larger than that of a fixed study, which means that data analysis in a fixed study is conducted at information fraction $1/1.072 = 0.933$. Looking at the graph, we see that the funnel shape of the efficacy bounds is less pronounced than with the O'Brien–Fleming efficacy bounds, but the general form is similar.

◁

Efficacy and futility stopping

▷ Example 4: Two-sided Wang–Tsiatis efficacy and futility bounds

Efficacy boundaries allow early stopping to reject H_0, but in some cases, there is an ethical argument for early stopping to accept H_0, such as when the experimental treatment causes deleterious side effects. If we can demonstrate that the experimental treatment is not significantly better than a placebo, we can end the trial early and prevent participants from receiving a treatment that does more harm than good. Even in the absence of harmful side effects, ending a trial early by accepting H_0 means that participants who would have been recruited into a "dead-end" study can instead be recruited to test the next promising treatment.

We continue with the scenario of example 3, this time adding futility bounds to permit early stopping to accept H_0. We want to allow futility stopping, but we do not want to be hasty in abandoning a treatment just because the very first results are not promising. To accomplish this, we use an O'Brien–Fleming futility bound that creates a narrow acceptance region at early looks.

We specify a binding futility bound with `futility()` suboption `binding`. If the z statistic from an interim analysis crosses a binding futility bound, the trial must be stopped for futility or else it will risk exceeding the desired familywise type I error.

```
. gsbounds, efficacy(wtsiatis(0.25)) futility(obfleming, binding) nlooks(5)
> graphbounds
Group sequential boundaries
Efficacy: Wang-Tsiatis, Delta = 0.2500
Futility: O'Brien-Fleming, binding
Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.8000
Info. ratio = 1.1961
Fixed-study crit. values = ±1.9600
Critical values and p-values for a group sequential design
```

	Info.	Efficacy			Futility		
Look	frac.	Lower	Upper	p-value	Lower	Upper	p-value
1	0.20	-3.0960	3.0960	0.0020	.	.	.
2	0.40	-2.6034	2.6034	0.0092	-0.3669	0.3669	0.7137
3	0.60	-2.3525	2.3525	0.0186	-1.0907	1.0907	0.2754
4	0.80	-2.1892	2.1892	0.0286	-1.6297	1.6297	0.1032
5	1.00	-2.0704	2.0704	0.0384	-2.0704	2.0704	0.0384

```
Note: Critical values are for z statistics; otherwise, use p-value
      boundaries.
```

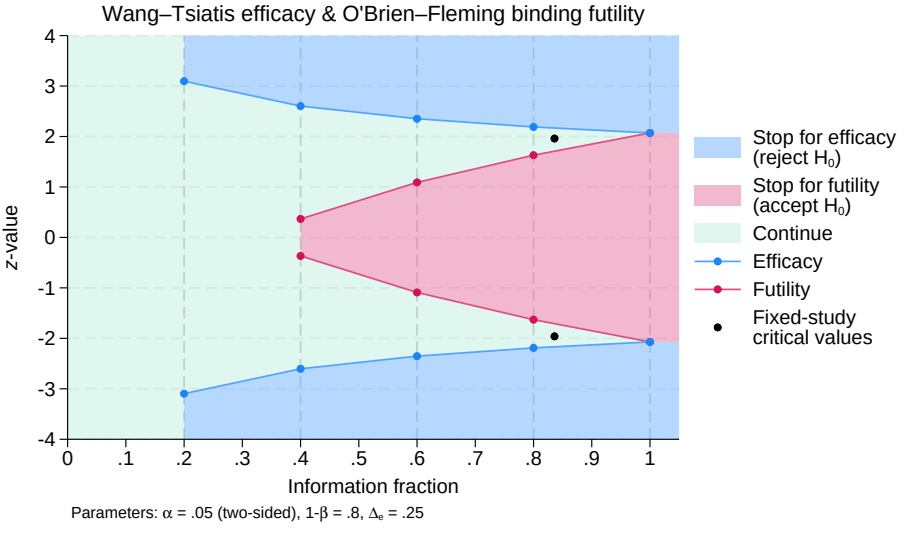

Figure 4. Wang–Tsiatis efficacy and futility bounds

The table of boundary values includes columns for futility lower and upper bounds, but the futility bounds for the first look are missing. This is because, to achieve the required significance level and power, the futility lower bound at the first look would have been above the futility upper bound. As such, the trial cannot be stopped for futility at the first look, and the futility bounds for this look are reported as missing. If z_1, the test statistic at the first look, lies within the continuation region of $(-3.096, 3.096)$, then the study will continue. If $|z_1| \geq 3.096$, then H_0 is rejected and the trial is stopped early for efficacy.

At the second look, there are three possibilities: If $|z_2| < 0.367$, then H_0 is accepted and the trial is terminated for futility. If $|z_2| \geq 2.603$, then H_0 is rejected and the trial is terminated due to treatment efficacy. If $|z_2| \in [0.367, 2.603)$, then the trial continues. A similar procedure is followed at the third and fourth looks, and by the fourth look, the continuation region has shrunk to $|z_4| \in [1.63, 2.189)$; if $|z_4| < 1.63$, the trial is terminated for futility, and if $|z_4| \geq 2.189$, the trial is terminated due to efficacy.

At the final look of a GSD with both efficacy and futility boundaries, the efficacy critical values are always the same as the futility critical values, and there is no continuation region. Here, if $|z_5| < 2.07$, H_0 is accepted; otherwise, H_0 is rejected. The sample size at the fifth look is 19.6% larger than that of a fixed study design, but the ability to stop the trial early due to futility has increased the chance that the trial will be terminated before the fifth look.

In the graph, we see the familiar funnel-shaped efficacy bounds, but now the futility bounds form a truncated "inner wedge" inside the efficacy bounds. The critical values from an equivalent fixed study design are similar to the critical values from the fifth look of the GSD, but the data analysis of the fixed study occurs at information fraction $1/1.196 = 0.836$.

Compared with the efficacy-only design of example 3 (which used the same significance level, power, efficacy bound type, and efficacy parameter as this example), we see that adding futility boundaries increases the maximum sample size from 107.2% to 119.6% of the fixed-study sample size. What's more, adding binding futility bounds has shrunk the efficacy critical values. Without futility bounds, the efficacy critical values at the first and fifth looks were ± 3.194 and ± 2.136, respectively (corresponding

to p-values of 0.0014 and 0.033). The addition of binding futility bounds has decreased those efficacy critical values to ± 3.096 and ± 2.07, respectively (equivalent to p-values of 0.002 and 0.038). Similar decreases in efficacy critical values are seen at the second, third, and fourth looks as well.

This decrease is best understood by considering the case of a true null hypothesis and examining the behavior of the two designs. In this case, the correct action would be to accept H_0; it is a type I error to reject H_0. When the null hypothesis is true, each interim look in the efficacy-only GSD presents the opportunity to continue the trial or to commit a type I error and mistakenly reject H_0. Only at the very last look do we have the option to correctly accept H_0. In the trial with both efficacy and futility bounds, we have more opportunities to correctly accept H_0, making it less likely that the trial will continue to later looks. If we were to use the same efficacy critical values as in the efficacy-only design, the actual probability of committing a type I error would be lower than the specified significance level, and the test would be conservative. By relaxing the efficacy critical values, the desired significance level is achieved.

◁

Nonbinding futility bounds

▷ Example 5: Two-sided Wang–Tsiatis efficacy and nonbinding futility bounds

The binding futility bounds we used in example 4 come with the restriction that the trial must be stopped if an interim analysis crosses the futility boundary. We can relax this requirement by removing `futility()` suboption `binding` to calculate nonbinding futility bounds. We omit the `graphbounds` option because the shape of this graph is nearly identical to that of the binding design.

```
. gsbounds, efficacy(wtsiatis(0.25)) futility(obfleming) nlooks(5)
Group sequential boundaries
Efficacy: Wang-Tsiatis, Delta = 0.2500
Futility: O'Brien-Fleming, nonbinding
Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.8000
Info. ratio = 1.2507
Fixed-study crit. values = ±1.9600
Critical values and p-values for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Efficacy Upper	Efficacy p-value	Futility Lower	Futility Upper	Futility p-value
1	0.20	-3.1941	3.1941	0.0014	.	.	.
2	0.40	-2.6859	2.6859	0.0072	-0.4050	0.4050	0.6855
3	0.60	-2.4270	2.4270	0.0152	-1.1396	1.1396	0.2544
4	0.80	-2.2586	2.2586	0.0239	-1.6875	1.6875	0.0915
5	1.00	-2.1360	2.1360	0.0327	-2.1360	2.1360	0.0327

```
Note: Critical values are for z statistics; otherwise, use p-value
      boundaries.
```

Examining the efficacy boundaries, we see that the critical values are identical to the efficacy critical values from the efficacy-only design of example 3. This is because nonbinding futility bounds do not affect the calculation of efficacy bounds.

At the end of example 4, we saw that binding futility bounds reduced the chance of erroneously rejecting a true null hypothesis because the trial is required to stop if the z statistic from an interim analysis crosses the futility bound. This is not the case with nonbinding futility bounds, where the experimenter can decide to continue the experiment even if the futility boundary is crossed.

Compared with the binding futility bounds of example 4, the nonbinding boundaries are slightly wider and the information ratio is larger (1.251 for the nonbinding design versus 1.196 for the binding design). The phenomenon of larger information ratios for designs with nonbinding futility bounds than for designs with binding futility bounds holds true, in general, and can be considered a cost associated with the increased flexibility offered by nonbinding designs.

◁

One-sided tests

▷ Example 6: One-sided O'Brien–Fleming efficacy bounds

The previous examples have all involved two-sided tests. When conducting a clinical trial of an experimental treatment, the researcher usually has a good idea of whether the effect will be positive or negative, but often two-sided tests are conducted to demonstrate impartiality. However, in some cases, it may be of interest to consider a one-sided alternative hypothesis. Here we plan to conduct a two-sample means test with a one-sided alternative hypothesis.

In example 2, we used a two-sided O'Brien–Fleming design with five equally spaced looks, a significance level of 0.05, and a power of 0.8. Here we use a similar design, but we restrict ourselves to a one-sided alternative hypothesis. This restricts the rejection region to positive values of a z statistic that are larger than the efficacy upper bound.

In the two-sided design with a significance level of 0.05, under the null hypothesis, there is a 2.5% probability that the observed z statistic is above the efficacy upper bound and a 2.5% probability that it is below the efficacy lower bound. To design a comparable study using a one-sided test, we adopt a significance level of 0.025 to match the efficacy upper bound of the two-sided design.

```
. gsbounds, alpha(0.025) efficacy(obfleming) nlooks(5) upper graphbounds
Group sequential boundaries
Efficacy: O'Brien-Fleming
Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.8000
Info. ratio = 1.0284
Fixed-study crit. value = 1.9600
Critical values and p-values
for a group sequential design
```

Look	Info. frac.	Efficacy Upper	p-value
1	0.20	4.5617	0.0000
2	0.40	3.2256	0.0006
3	0.60	2.6337	0.0042
4	0.80	2.2809	0.0113
5	1.00	2.0401	0.0207

```
Note: Critical values are for z statistics;
      otherwise, use p-value boundaries.
```

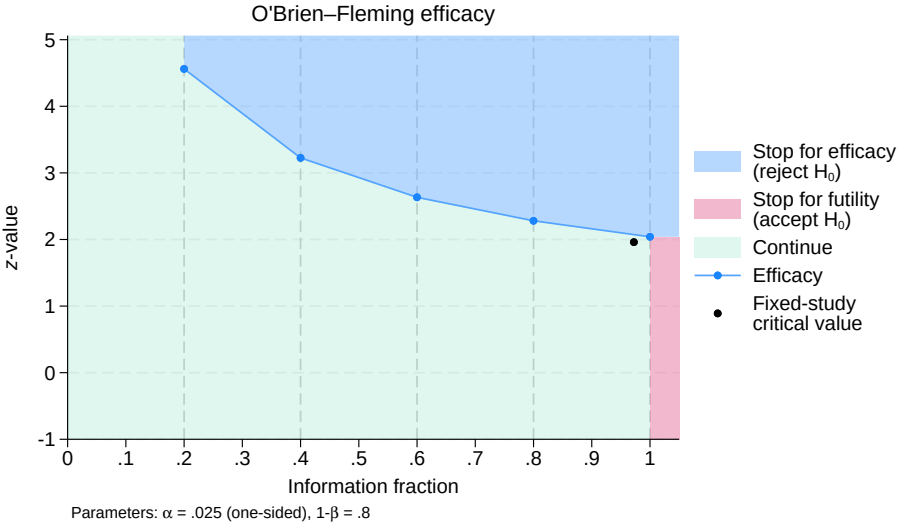

Figure 6. One-sided O'Brien–Fleming efficacy bounds

As expected, the efficacy upper bound for a one-sided design with significance level 0.025 is identical to the efficacy upper bound in the two-sided design with significance level 0.05. The graph of the one-sided bound is identical to the upper portion of the graph of the two-sided bound from example 2.

The procedure for comparing test statistics to the boundary critical values is somewhat simpler with a single bound: At the first through fourth looks, we reject H_0 if the z statistic exceeds the critical value; otherwise, we continue the experiment. At the final look, we reject H_0 if $z_5 \geq 2.04$; otherwise, we accept H_0.

◁

Error-spending bounds

▷ Example 7: One-sided error-spending O'Brien–Fleming-style efficacy bounds

In example 6, we used a one-sided O'Brien–Fleming design with five equally spaced looks, a significance level of 0.025, and a power of 0.8. O'Brien–Fleming efficacy bounds possess properties that appeal to clinical trialists: The conservative critical values at early looks ensure that a trial is not stopped very early unless the evidence against the null hypothesis is overwhelming, and the critical values at the final look are nearly the same as those from a fixed study design, reducing the risk of the group sequential trial being unable to reject H_0 despite a final z statistic that would have resulted in rejecting H_0 under a fixed study design.

The large critical values at early looks correspond to a very small probability of committing a type I error. Viewed from the perspective of the error-spending paradigm, we can say that the O'Brien–Fleming design spends very little error at early looks, instead saving the error for later looks. If we rerun the design from example 6, we can examine the cumulative type I error spent by displaying returned matrix `r(aspent)`.

```
. gsbounds, alpha(0.025) efficacy(obfleming) nlooks(5) upper
  (output omitted)
. matrix list r(aspent)
r(aspent)[5,1]
         alpha spent:
            per look
Look 1     2.537e-06
Look 2     .00062953
Look 3      .0044518
Look 4     .01279229
Look 5           .025
```

In the classical O'Brien–Fleming design, critical values are calculated directly, and the error spent at each look is a product of those critical values. Boundaries cannot be modified while the trial is underway because the critical value at each look depends on the critical values of all other looks. With error-spending boundaries, the error spent at each look is determined by the error-spending function, and the critical value is a product of the error spent. In this case, each critical value depends on the total information to be collected and the error spent at previous looks, but not on the critical values of future looks.

When Lan and DeMets (1983) developed the error-spending approach, they formulated an error-spending function that approximates the error spent at each look by O'Brien–Fleming bounds. By spending the type I error at nearly the same rate as the classic O'Brien–Fleming design, the error-spending approximation attains critical values that are nearly the same as those of the classic O'Brien–Fleming design.

Here we modify the design used in example 6 by specifying an efficacy boundary of `errobfleming` to calculate error-spending O'Brien–Fleming-style bounds.

```
. gsbounds, alpha(0.025) efficacy(errobfleming) nlooks(5) upper graphbounds
Group sequential boundaries
Efficacy: Error-spending O'Brien-Fleming style
Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.8000
Info. ratio = 1.0247
Fixed-study crit. value = 1.9600
Critical values and p-values
for a group sequential design
```

Look	Info. frac.	Efficacy Upper	p-value
1	0.20	4.8769	0.0000
2	0.40	3.3570	0.0004
3	0.60	2.6803	0.0037
4	0.80	2.2898	0.0110
5	1.00	2.0310	0.0211

```
Note: Critical values are for z statistics;
      otherwise, use p-value boundaries.
```

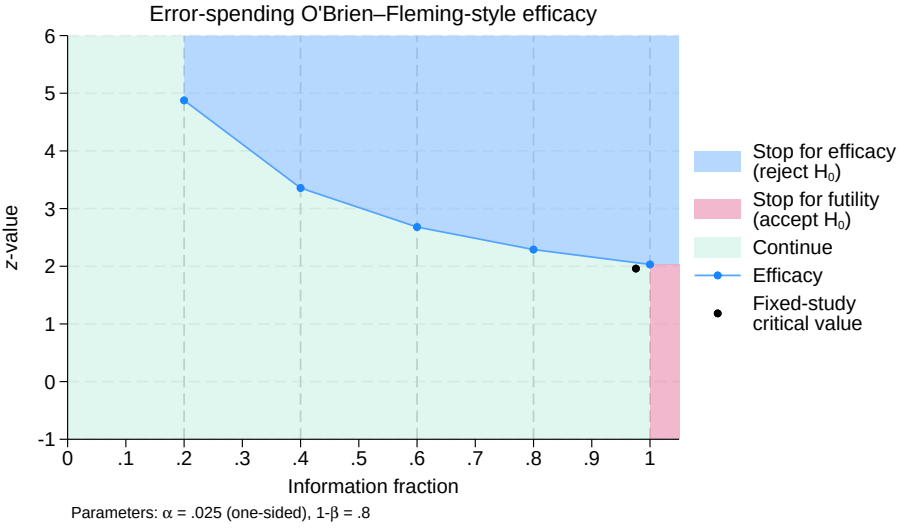

Figure 7. One-sided error-spending O'Brien–Fleming-style efficacy bounds

The critical values of the error-spending O'Brien–Fleming-style bounds are very similar to those of the classic O'Brien–Fleming design. Both start off conservatively at early looks and approach the fixed-study critical value by the final look. The information ratio of both designs is also very similar. At the final look, the classic O'Brien–Fleming design required 2.8% more information than an equivalent fixed design, while the error-spending approximation requires 2.5% more.

Examining the graph, it is difficult to distinguish the difference between the shape of the error-spending O'Brien–Fleming-style bounds and the classic O'Brien–Fleming bounds from example 6.

To see the cumulative type I error spent at each look, we examine `r(aspent)`.

```
. matrix list r(aspent)
r(aspent)[5,1]
         alpha spent:
            per look
Look 1      5.389e-07
Look 2      .00039415
Look 3      .00380806
Look 4      .01221179
Look 5           .025
```

Unsurprisingly, we see that the error-spending O'Brien–Fleming-style design spends the allotted α of 0.025 at nearly the same rate as the classic O'Brien–Fleming design.

◁

▷ Example 8: One-sided error-spending efficacy and futility bounds

Clinical trials using one-sided tests stand to benefit from futility stopping just as much as trials using two-sided tests. Consider a trial with the one-sided alternative hypothesis that the mean of the experimental group is less than the mean of the control group. We plan for three evenly spaced looks, and we use error-spending bounds.

We want an efficacy boundary that is conservative at early looks, so we choose Kim–DeMets efficacy bounds with parameter $\rho_e = 3$, which yields bounds that are similar in shape to O'Brien–Fleming bounds, if slightly less conservative at very early looks. To increase the chance that we can accept the null hypothesis at the first look if the evidence supports H_0, we want a futility boundary that is less conservative at early looks. Selecting Hwang–Shih–de Cani futility bounds with parameter $\gamma_f = 1$ accomplishes this by producing bounds that are similar in shape to Pocock bounds, and we make the futility bound nonbinding so that stopping is not required if it is crossed at an interim analysis. As in example 6, we use a significance level of 0.025, but here we specify the power to be 0.9.

```
. gsbounds, alpha(0.025) power(0.9) efficacy(kdemets(3)) futility(hsdecani(1))
> nlooks(3) lower graphbounds

Group sequential boundaries

Efficacy: Error-spending Kim-DeMets, rho = 3.0000
Futility: Error-spending Hwang-Shih-de Cani, nonbinding, gamma = 1.0000

Study parameters:
      alpha = 0.0250  (lower one-sided)
      power = 0.9000

Info. ratio = 1.2315

Fixed-study crit. value = -1.9600

Critical values and p-values for a group sequential design
```

	Info.	Efficacy		Futility	
Look	frac.	Lower	p-value	Upper	p-value
1	0.33	-3.1130	0.0009	-0.3798	0.3521
2	0.67	-2.4619	0.0069	-1.3016	0.0965
3	1.00	-2.0087	0.0223	-2.0087	0.0223

```
Note: Critical values are for z statistics; otherwise,
      use p-value boundaries.
```

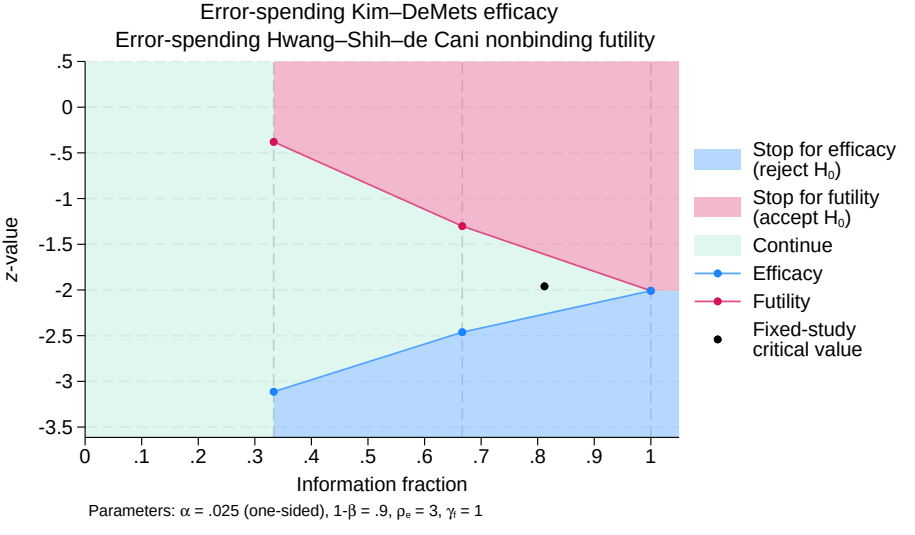

Figure 8. One-sided lower error-spending efficacy and futility bounds

At the first look, the continuation region is the interval between the efficacy lower bound of -3.113 and the futility upper bound of -0.38. If $z_1 > -0.38$, then H_0 may be accepted and the trial terminated for futility; if $z_1 \leq -3.113$, then H_0 is rejected and the trial is terminated due to treatment efficacy. At the second look, the continuation region has shrunk to $(-2.462, -1.302]$. At the third and final look, the critical values of the efficacy lower bound and the futility upper bound coincide, and there is no continuation region: If $z_3 \leq -2.009$, then H_0 is rejected; otherwise, it is accepted.

If the study continues to the last look, the final critical value is very close to the critical value for a fixed study design, but the GSD requires 23.1% more participants than a fixed design.

◁

Unevenly spaced looks

▷ Example 9: One-sided error-spending bounds with unevenly spaced looks

In example 8, we used a three-look GSD with evenly spaced information increments. Here we consider a similar scenario, but we add a new look halfway between the first and second looks. To specify four looks with uneven spacing, we use the `information()` option. Because `information()` is automatically rescaled, we need not specify the final information level as 1, so we can type `information(1 1.5 2 3)` to avoid repeating decimals.

```
. gsbounds, alpha(0.025) power(0.9) efficacy(kdemets(3)) futility(hsdecani(1))
> information(1 1.5 2 3) lower graphbounds

Group sequential boundaries

Efficacy: Error-spending Kim-DeMets, rho = 3.0000
Futility: Error-spending Hwang-Shih-de Cani, nonbinding, gamma = 1.0000

Study parameters:
      alpha = 0.0250  (lower one-sided)
      power = 0.9000

Info. ratio = 1.2456

Fixed-study crit. value = -1.9600

Critical values and p-values for a group sequential design
```

	Info.	Efficacy		Futility	
Look	frac.	Lower	p-value	Upper	p-value
1	0.33	-3.1130	0.0009	-0.3916	0.3477
2	0.50	-2.7889	0.0026	-0.7827	0.2169
3	0.67	-2.5133	0.0060	-1.2002	0.1150
4	1.00	-2.0120	0.0221	-2.0120	0.0221

```
Note: Critical values are for z statistics; otherwise,
      use p-value boundaries.
```

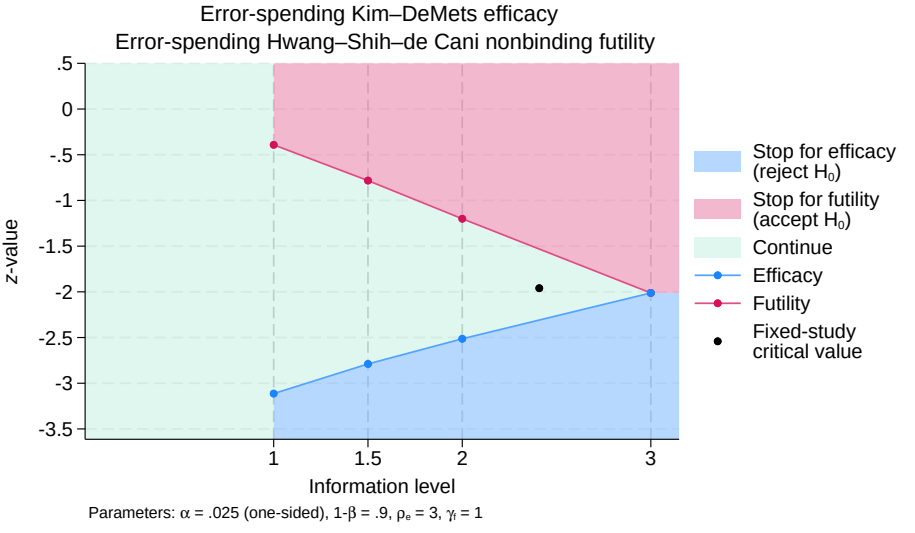

Figure 9. One-sided lower error-spending efficacy and futility bounds with unevenly spaced looks

The shape of the bounds is strikingly similar to the design in example 8, but the x axis of the graph has been labeled using the scale we specified in the `information()` option. The properties of the design, including the final critical value and the information ratio, are in line with the three-look design, but the additional look gives us one more opportunity to terminate the trial early.

◁

Futility-only stopping

▷ Example 10: One-sided error-spending Pocock-style futility bounds

The previous examples have all allowed early stopping due to efficacy, but occasionally only futility stopping is desired. This can occur, for example, if there is concern about uncommon but serious adverse events, which are harmful side effects of the treatment and negative medical outcomes not associated with an underlying disease. In this case, even if the interim results offer compelling evidence of treatment efficacy, the trial will continue in order to collect a sample large enough to evaluate the pattern of adverse events. If the interim results are not promising, the trial can be terminated early for futility.

Here critical values for the futility bounds are calculated for each look, but critical values for the efficacy bounds are only calculated for the final look because H_0 cannot be rejected until the end of the study. As in example 7, we will design a study with five equally spaced looks, an upper one-sided significance level of 0.025, and a power of 0.8. But we replace the error-spending O'Brien–Fleming-style efficacy bound with a nonbinding error-spending Pocock-style futility bound.

```
. gsbounds, alpha(0.025) futility(errpocock) nlooks(5) upper graphbounds
Group sequential boundaries
Futility: Error-spending Pocock style, nonbinding
Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.8000
Info. ratio = 1.3060
Fixed-study crit. value = 1.9600
Critical values and p-values for a group sequential design
```

Look	Info. frac.	Efficacy Upper	Efficacy p-value	Futility Lower	Futility p-value
1	0.20			-0.1307	0.5520
2	0.40			0.5751	0.2826
3	0.60			1.1163	0.1321
4	0.80			1.5672	0.0585
5	1.00	1.9600	0.0250	1.9600	0.0250

```
Note: Critical values are for z statistics; otherwise,
      use p-value boundaries.
```

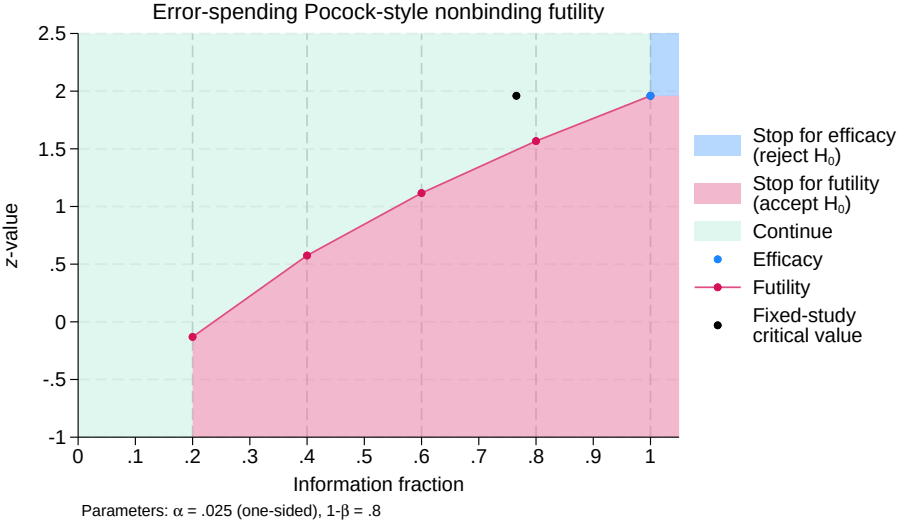

Figure 10. Error-spending Pocock-style nonbinding futility bound

At the first look, we are allowed, but not required, to accept H_0 if $z_1 < -0.131$; otherwise, the trial continues. No efficacy critical value is reported for the first look because we cannot stop the trial for efficacy at this point. This procedure is repeated at the second, third, and fourth looks, with progressively larger futility critical values. At the fifth look, which is the only look with an efficacy critical value, we reject H_0 if $z_5 \geq 1.96$; otherwise, we accept H_0.

The critical value at the fifth look is equal to the critical value from an equivalently powered fixed study design. This is because a GSD with futility-only stopping offers a single opportunity to reject H_0 at the end of the study, just as a fixed design does. If we had specified binding futility bounds, the critical value would have been even smaller than that of a fixed design. This is because, if the null hypothesis

is true, binding futility bounds reduce the probability of committing a type I error because the trial can be forced to stop for futility before reaching the opportunity to reject H_0 at the final look. To avoid underspending the desired type I error in the presence of binding futility bounds, efficacy critical values are reduced until the desired α level is reached.

On the graph, the efficacy bound is drawn as a single dot rather than a line because only the last look uses an efficacy bound. The dot for the efficacy bound covers the final dot marking the final futility bound because they share the same critical value.

◁

Stored results

`gsbounds` stores the following in `r()`:

Scalars

`r(alpha)`	overall significance level (familywise type I error)
`r(beta)`	overall probability of a type II error
`r(binding)`	`1` for binding futility bounds, `0` for nonbinding
`r(effparam)`	efficacy parameter (if `wtsiatis()`, `kdemets()`, or `hsdecani()` specified)
`r(futparam)`	futility parameter (if `wtsiatis()`, `kdemets()`, or `hsdecani()` specified)
`r(info_ratio)`	ratio of maximum information required to that of a fixed study design
`r(nlooks)`	number of analyses
`r(onesided)`	`1` for a one-sided test, `0` otherwise
`r(power)`	overall power
`r(stop)`	`0` for futility bounds, `1` for efficacy bounds, `2` for both
`r(z_fixed)`	critical value for an equivalent fixed study design

Macros

`r(cmd)`	`gsbounds`
`r(cmdline)`	command as typed
`r(direction)`	`upper`, `lower`, or `two-sided`
`r(effbnd)`	`pocock`, `obfleming`, `wtsiatis`, `errpocock`, `errobfleming`, `kdemets`, or `hsdecani`
`r(futbnd)`	`pocock`, `obfleming`, `wtsiatis`, `errpocock`, `errobfleming`, `kdemets`, or `hsdecani`

Matrices

`r(aspent)`	cumulative alpha spent per look (stored with efficacy-only stopping or when futility bounds are binding)
`r(aspent_fstop)`	cumulative alpha spent per look if futility stopping does occur (stored when futility bounds are nonbinding)
`r(aspent_nofstop)`	cumulative alpha spent per look if futility stopping does not occur (stored when futility bounds are nonbinding)
`r(bounds)`	stopping boundaries
`r(bspent)`	cumulative beta spent per look (when futility bounds are specified)
`r(info_frac)`	information fraction
`r(info_level)`	specified information level
`r(p_crit)`	p-values corresponding to boundary critical values

Methods and formulas

Methods and formulas are presented under the following headings:

Group sequential bounds

After each group of observations is collected, an analysis is performed and the test statistic Z is calculated. In the description that follows, we assume that Z follows a standard normal distribution under H_0. For test statistics that follow other distributions, the normal model is used to calculate boundaries that are then converted to the appropriate scale using the significance level approach.

In a GSD with K looks, let $(n_1, \ldots, n_K)$ be the cumulative sample sizes at looks 1 through K, with the maximum sample size of n_K attained at the final look. For any k in $(1, \ldots, K)$, let $\mathcal{I}_k$ denote the information fraction at look k. This is the fraction of the maximum sample size that has been observed, with $\mathcal{I}_k = n_k/n_K$ for k in $(1, \ldots, K)$. For studies with time-to-event outcomes, where information is proportional to the number of events observed, interpret n_k to be the cumulative number of events observed at stage k, and interpret n_K to be the maximum number of events.

Each test statistic Z_k is calculated using all observations collected through look k. This cumulative quality implies that $(Z_1, \ldots, Z_K)$ are not independent. Jennison and Turnbull (2000, 49) show that $(Z_1, \ldots, Z_K)$ is multivariate normal with

$$\text{Cov}(Z_j, Z_k) = \sqrt{\frac{\mathcal{I}_j}{\mathcal{I}_k}} \qquad \text{for } 1 \leq j \leq k \leq K \tag{1}$$

When $(Z_1, \ldots, Z_K)$ follow this distribution, the score statistics $(S_1, \ldots, S_K)$ that correspond to these z statistics are said to have the property of "independent increments". For any k in $(1, \ldots, K)$, S_k is equal to Z_k multiplied by the square root of the Fisher information for the parameter involved in the test. The independent increments property means that $S_1, (S_2 - S_1), \ldots, (S_K - S_{K-1})$ are independently distributed.

Without loss of generality, consider a GSD for an upper one-sided test with both efficacy and binding futility bounds. Denote critical values for efficacy stopping as $(e_1, \ldots, e_K)$ and critical values for futility stopping as $(f_1, \ldots, f_K)$. At interim look $k < K$, if test statistic $Z_k \geq e_k$, the trial is stopped for efficacy; if $Z_k < f_k$, the trial is stopped for futility; and if $f_k \leq Z_k < e_k$, the trial continues. At the final look, there is no continuation region because $f_K = e_K$.

Let α_k and β_k be the respective probabilities of type I and type II error at look k, and let $\alpha = \sum_{k=1}^{K} \alpha_k$ and $\beta = \sum_{k=1}^{K} \beta_k$ be the overall probabilities of type I and type II error (with power equal to $1 - \beta$). Using the result of Wassmer and Brannath (2016, 57), we write the probability of type I error during the first and subsequent looks as

$$\alpha_1 = \Pr_{H_0}(Z_1 \geq e_1) \quad \text{and} \quad \alpha_k = \Pr_{H_0}\left(Z_k \geq e_k \cap \bigcap_{j=1}^{k-1} f_j \leq Z_j < e_j\right) \text{ for } k \in (2, \ldots, K) \tag{2}$$

Similarly, the formula for the stagewise probability of type II error is

$$\beta_1 = \Pr_{H_a}(Z_1 < f_1) \quad \text{and} \quad \beta_k = \Pr_{H_a}\left(Z_k < f_k \cap \bigcap_{j=1}^{k-1} f_j \leq Z_j < e_j\right) \text{ for } k \in (2, \ldots, K) \tag{3}$$

where $\Pr_{H_0}(\cdot)$ indicates the probability under the null hypothesis and $\Pr_{H_a}(\cdot)$ indicates the probability under the alternative hypothesis.

For trials with efficacy stopping only, replace $(f_1, \ldots, f_{K-1})$ with $-\infty$ and let $f_K = e_K$ in the calculations above. For trials with nonbinding futility bounds, replace $(f_1, \ldots, f_{K-1})$ with $-\infty$ in (2) but not in (3). For trials with futility stopping only, replace $(e_1, \ldots, e_{K-1})$ with ∞ and let $e_K = f_K$ (in this case, stored result `r(bounds)` records interim efficacy critical values as `.z`). For two-sided trials, replace all instances of Z with $|Z|$ in (2), and replace Z_j with $|Z_j|$ in (3).

To calculate the probabilities in (2) and (3), cumulative multivariate normal distributions are evaluated with lower limit $(f_1, \ldots, f_K)$ and upper limit $(e_1, \ldots, e_K)$. Two-sided tests require additional integration from $(-e_1, \ldots, -e_K)$ to $(-f_1, \ldots, -f_K)$. The covariance matrix of the distribution, defined in (1), allows the multivariate normal integral to be decomposed into a series of univariate integrals using the recursive integration formula of Armitage, McPherson, and Rowe (1969).

The integrals are approximated using Simpson's rule, with quadrature points spaced closer together toward the center of the distribution than at the tails, as per Jennison and Turnbull (2000, 349). The number of quadrature points is $12r - 3$, with $r = 20$ by default. Jennison and Turnbull (2000) report that using $r = 16$ yields probabilities that are accurate to 10^{-6}. The value of r can be set with the `intpointsscale(#)` option. When integrating over narrow intervals, the number of quadrature points is increased adaptively to ensure sufficient precision.

Classical (Wang–Tsiatis) bounds

Wang and Tsiatis (1987) developed a class of group sequential boundaries with shape parameter Δ. The Wang–Tsiatis family includes the classical bounds of Pocock (1977) and O'Brien and Fleming (1979) as special cases. The Pocock boundary is equivalent to a Wang–Tsiatis design with $\Delta = 0.5$, and the O'Brien–Fleming boundary is a Wang–Tsiatis design with $\Delta = 0$. The implementation of classical boundaries `pocock`, `obfleming`, and `wtsiatis()` follows the work of Pampallona and Tsiatis (1994), who extended the Wang–Tsiatis family of bounds to include futility stopping.

To allow efficacy and futility bounds to use different parameters, we use the notation Δ_e and Δ_f. We define efficacy critical value $e_k = C * \mathcal{I}_k^{\Delta_e - 1/2}$, where Δ_e controls the shape of the efficacy bounds and C is a scaling factor. At the final look, $\mathcal{I}_K = 1$, so $e_K = C$. Futility critical value $f_k = C * \mathcal{I}_k^{\Delta_f - 1/2} + \mathcal{M}^{1/2}(\mathcal{I}_k^{1/2} - \mathcal{I}_k^{\Delta_f - 1/2})$, where $\mathcal{M}$ is the maximum information of the trial and Δ_f controls the shape of the futility bound. $\mathcal{M}$ can be thought of as a standardized version of the Fisher information, scaled to equal the expected information at the final look of a group sequential trial with an effect size of 1 under H_a. The expected information of an equivalent fixed-sample trial is denoted as $\mathcal{F}$. For a one-sided trial, $\mathcal{F} = \{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2$, where $\Phi^{-1}(\cdot)$ is the inverse standard normal cumulative distribution function. For a two-sided trial, α is replaced with $\alpha/2$.

Two-dimensional optimization is performed to find values of C and $\mathcal{M}$ that yield the desired probabilities of type I and type II errors. The starting value for C can be specified with the `initscale(#)` option. The default starting value for C is z_α for one-sided trials and $z_{\alpha/2}$ for two-sided trials, where $z_\alpha = \Phi^{-1}(1-\alpha)$. The starting value for $\mathcal{M}$ can be specified with the `initinfo(#)` option, and the default starting value for $\mathcal{M}$ is $\mathcal{F}$. Other aspects of the optimization process, such as the optimization technique and number of iterations, can be controlled by specifying additional optimization options (see *optimopts*).

Let R represent the information ratio, the ratio of the maximum sample size of a Wang–Tsiatis design to that of a fixed design with equivalent type I and type II error. We calculate $R = \mathcal{M}/\mathcal{F}$.

Error-spending bounds

Instead of calculating critical values e_k directly, the error-spending approach defines an α-spending function $\alpha^*(t)$. This function must be monotonically increasing over $t \in [0, 1]$, and it must satisfy $\alpha^*(0) = 0$ and $\alpha^*(t) = \alpha$ for $t \geq 1$. The α-spending function is used to partition α into $(\alpha_1, \ldots, \alpha_K)$ by setting $\alpha_1 = \alpha^*(\mathcal{I}_1)$ and $\alpha_k = \alpha^*(\mathcal{I}_k) - \alpha^*(\mathcal{I}_{k-1})$ for k in $(2, \ldots, K)$.

Lan and DeMets (1983) proposed error-spending functions that closely approximate classical Pocock and O'Brien–Fleming bounds. The α-spending function for Pocock-style bounds is $\alpha^*_{\text{P}}(t; \alpha) = \min[\alpha \log\{1+(e-1)t\},\ \alpha]$. The α-spending function for O'Brien–Fleming-style bounds is $\alpha^*_{\text{OBF}}(t; \alpha) = \min\{2-2\Phi(z_{\alpha/2}/\sqrt{t}),\ \alpha\}$ for one-sided bounds and $\alpha^*_{\text{OBF}}(t; \alpha) = \min\{4-4\Phi(z_{\alpha/4}/\sqrt{t}),\ \alpha\}$ for two-sided bounds (Wassmer and Brannath 2016), where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Kim and DeMets (1987) introduced a single parameter family of error-spending functions indexed by parameter $\rho > 0$, with α-spending function $\alpha^*_{\text{KD}}(t; \rho, \alpha) = \min(\alpha t^\rho,\ \alpha)$. Another popular error-spending function, proposed by Hwang, Shih, and de Cani (1990), uses parameter γ in α-spending function

$$\alpha^*_{\text{HSD}}(t; \gamma, \alpha) = \begin{cases} \alpha(1 - e^{-\gamma t})/(1 - e^{-\gamma}) & \text{for} \gamma \neq 0 \\ \alpha t & \text{for} \gamma = 0 \end{cases}$$

The error-spending approach can also be used to spend type II error, with the resulting β-spending function $\beta^*(\cdot)$ following rules analogous to those of the α-spending function. It is used to partition β into $\beta_1 = \beta^*(\mathcal{I}_1)$ and $\beta_k = \beta^*(\mathcal{I}_k) - \beta^*(\mathcal{I}_{k-1})$ for k in $(2, \ldots, K)$.

For trials with efficacy stopping only, $e_1 = \Phi^{-1}(1-\alpha_1)$ for a one-sided test and $e_1 = \Phi^{-1}(1-\alpha_1/2)$ for a two-sided test. The error spent at subsequent looks depends on the stopping boundaries of the previous stages, so boundary values are found sequentially through numerical optimization. A separate optimization step is then performed to determine the maximum information $\mathcal{M}$. The starting value for $\mathcal{M}$ can be specified with the `initinfo(#)` option. The default starting value for $\mathcal{M}$ is $\mathcal{F}$, the expected information from an equivalent fixed study design.

For trials allowing stopping for futility, calculation of the boundary critical values and maximum information cannot be decomposed into separate optimization steps. In this case, a numerical search for $\mathcal{M}$ is performed using the bisection method, and boundaries are recalculated at each step. The tolerance for the bisection search can be specified with the `infotol(#)` option, and the default value is `infotol(1e-6)`. The lower starting value in the search for $\mathcal{M}$ can be specified with the `initinfo(# .)` option, and the upper starting value can be specified as `initinfo(. #)`. To specify both lower and upper starting values, use syntax `initinfo(# #)`, specifying first the lower starting value and then the upper starting value. By default, the lower starting value for the bisection search is $\mathcal{F}$, and the upper starting value is the information required by a Bonferroni correction for repeated hypothesis tests.

Regardless of whether stopping is for efficacy, futility, or both, rarely modified aspects of the optimization process, such as the optimization technique and number of iterations, can be controlled by specifying additional optimization options (see *optimopts*).

As with classical Wang–Tsiatis designs, the information ratio for error-spending designs is calculated as $R = \mathcal{M}/\mathcal{F}$.

Significance level approach

The theory behind GSDs relies on the assumption that test statistics $(Z_1, \ldots, Z_K)$ follow a multivariate normal distribution with covariance specified in (1) and marginal standard normal distributions under H_0. The classic example is the difference of means between two normally distributed responses, scaled by a known standard deviation. However, many common test statistics are asymptotically normal, such as log odds-ratios and log-rank tests.

When the desired test does not produce an asymptotically normal test statistic, Pocock (1977) suggests the significance level approach to approximately control errors in GSDs. Jennison and Turnbull (2000, 80) and Wassmer and Brannath (2016, 103) advocate the use of this approximation, describing it as "remarkably accurate" and "stupendously accurate", respectively.

For test statistic T_k with cumulative distribution $F(\cdot)$ under H_0, we calculate standardized test statistic $T_k^* = \Phi^{-1}\{F(T_k)\}$ that has the same significance level as T_k. That is, $F(T_k) = \Phi(T_k^*)$. The standardized test statistic T_k^* can be compared directly with critical values e_k and f_k. Equivalently, we can calculate the p-value of test statistic T_k and compare it with the p-values corresponding to e_k and f_k. The p-value technique is straightforward to implement and is demonstrated in examples 2 and 3 of [ADAPT] **gsdesign onemean**, example 2 of [ADAPT] **gsdesign twomeans**, and examples 2 and 3 of [ADAPT] **gsdesign twoproportions**.

The significance level approach can be used as long as the assumption of independent increments is met. Many popular statistical tests satisfy this assumption; however, Jennison and Turnbull (2000) provide several examples of scenarios where this assumption does not hold, even asymptotically. One such example is the group sequential analysis of longitudinal data comparing the mean response of two groups, where the within-subject response has an autoregressive element. The significance level approach does not justify the use of group sequential testing when the assumption of independent increments is violated; it only applies when this assumption is satisfied but the test statistics are not normally distributed.

References

Armitage, P., C. K. McPherson, and B. C. Rowe. 1969. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society*, A ser., 132: 235–244. https://doi.org/10.2307/2343787.

Gould, A. L. 1989. "Abandoning lost causes (early termination of unproductive clinical trials)". In *Proceedings of the Biopharmaceutical Section*, 31–34. Washington, DC: American Statistical Association.

Hwang, I. K., W. J. Shih, and J. S. de Cani. 1990. Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* 9: 1439–1445. https://doi.org/10.1002/sim.4780091207.

Jennison, C., and B. W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman and Hall/CRC.

Kim, K., and D. L. DeMets. 1987. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 74: 149–154. https://doi.org/10.1093/biomet/74.1.149.

Lan, K. K. G., and D. L. DeMets. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70: 659–663. https://doi.org/10.1093/biomet/70.3.659.

O'Brien, P. C., and T. R. Fleming. 1979. A multiple testing procedure for clinical trials. *Biometrics* 35: 549–556. https://doi.org/10.2307/2530245.

Pampallona, S., and A. A. Tsiatis. 1994. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference* 42: 19–35. https://doi.org/10.1016/0378-3758(94)90187-2.

Pitblado, J. S., B. P. Poi, and W. W. Gould. 2024. *Maximum Likelihood Estimation with Stata*. 5th ed. College Station, TX: Stata Press.

Pocock, S. J. 1977. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64: 191–199. https://doi.org/10.1093/biomet/64.2.191.

Wang, S. K., and A. A. Tsiatis. 1987. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 43: 193–199. https://doi.org/10.2307/2531959.

Wassmer, G., and W. Brannath. 2016. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Cham, Switzerland: Springer.

Also see

[ADAPT] **GSD intro** — Introduction to group sequential designs

[ADAPT] **gs** — Introduction to commands for group sequential design

[ADAPT] **gsdesign** — Study design for group sequential trials

[ADAPT] **Glossary**

[PSS-2] **power** — Power and sample-size analysis for hypothesis tests