gs — Introduction to commands for group sequential design

Description

The gs suite of commands is useful for planning group sequential trials. These commands compute stopping boundaries and sample sizes for each look of a group sequential design (GSD). The gs commands can be used to calculate critical values for efficacy boundaries, futility boundaries, or both. Boundary-calculation procedures include those of Pocock (1977), O'Brien and Fleming (1979), Wang and Tsiatis (1987), Kim and DeMets (1987), and Hwang, Shih, and de Cani (1990).

The `gsbounds` command calculates stopping boundaries that can be applied to any group sequential clinical trial. The `gsdesign` *method* set of commands calculates both stopping boundaries and sample sizes for interim analyses with five different hypothesis tests: one- and two-sample means tests, one- and two-sample proportions tests, and the log-rank test. Interim analyses using other hypothesis tests are supported through the ability to incorporate user-defined sample-size calculations. Study designs can be displayed in a table and a graph.

Menu

Statistics > Power, precision, and sample size

Syntax

Compute stopping boundaries

```
gsbounds, gsboundopts
```

where *gsboundopts* are options described in [ADAPT] **gsbounds**.

Compute sample size and stopping boundaries

```
gsdesign method ... [ , designopts boundopts ]
```

where *designopts* are options controlling the sample-size calculation and *boundopts* are options controlling the calculation of the stopping boundaries.

method	Description
One sample	
`onemean`	One-sample mean test
`oneproportion`	One-sample proportion test
Two independent samples	
`twomeans`	Two-sample means test
`twoproportions`	Two-sample proportions test
Survival analysis	
`logrank`	Log-rank test
User-defined methods	
usermethod	Add your own method to `gsdesign`

Remarks and examples

Remarks are presented under the following headings:

This section describes how to compute boundaries and sample sizes for GSDs using the gs suite of commands. For a software-free introduction to GSDs, see [ADAPT] **GSD intro**.

Introduction

Clinical trials are studies investigating the effects of a treatment on human participants, and unlike some other types of studies, clinical trials rarely collect data all at once. It is common for large clinical trials to recruit participants over the course of months or years. Depending on the outcome of interest, known as the clinical endpoint, the study could follow up with participants over the course of several years.

Sponsors of clinical trials have both ethical and economic motivations for making trials as efficient as possible. One way of accomplishing this is to analyze trial data while the study is still underway. A positive result at an interim analysis can lead to early termination of the study due to treatment efficacy, sparing future participants from being assigned to the control group and receiving an inferior treatment. If the interim analysis demonstrates that the new treatment is ineffective, the trial can stop early and resources can be allocated to testing more promising treatments.

It is widely known that conducting multiple hypothesis tests at a nominal significance level will inflate type I error, but applying a simplistic technique like the Bonferroni correction to the results of interim analyses is overly conservative and will cause excessive type II error. GSDs provide a framework for conducting multiple interim analyses of clinical trial data while maintaining control of familywise type I and type II errors.

The gs suite of commands comprises the `gsbounds` command and the `gsdesign` *method* commands. This suite can be used to design group sequential clinical trials by calculating stopping boundaries and sample sizes for interim analyses, or looks. The `gsbounds` command calculates stopping boundaries that can be applied to any clinical trial following a GSD. The `gsdesign` *method* commands calculate both stopping boundaries and sample sizes for each look. The `gsbounds` and `gsdesign` *method* commands provide the same features and syntax for computing stopping boundaries; `gsdesign` extends the capabilities of `gsbounds` and additionally computes sample sizes. In the examples below, we first introduce `gsbounds` and focus on features for stopping boundaries. Then we move to examples that include sample-size calculations with `gsdesign`, which will be more commonly used in practice.

`gsbounds` and `gsdesign` provide four options—`efficacy()`, `futility()`, `nlooks()`, and `information()`—that allow us to specify the boundary-calculation procedure and the number and spacing of looks. Below, we introduce the syntax with `gsbounds`, but the options are specified in the same way with `gsdesign`.

By default, O'Brien–Fleming efficacy bounds are computed. The `efficacy()` option allows you to select from among seven available boundary-calculation procedures, such as the Pocock boundary:

```
gsbounds, efficacy(pocock) ...
```

To request futility bounds instead of efficacy bounds, replace the `efficacy()` option with `futility()`. All boundary-calculation procedures available for efficacy bounds are also available for futility bounds.

```
gsbounds, futility(pocock) ...
```

To compute both efficacy and futility bounds, specify both options:

```
gsbounds, efficacy(pocock) futility(pocock) ...
```

To request more than 2 equally spaced looks (the default), specify the `nlooks()` option:

```
gsbounds, nlooks(5) ...
```

To request that looks be performed at specific information levels rather than being equally spaced, use the `information()` option:

```
gsbounds, information(50 60 70 80 90) ...
```

In addition to the options demonstrated above for specifying boundaries, the `gsdesign` *method* commands allow both common and *method*-specific arguments and options for specifying your desired power and sample-size settings. See [PSS-2] **power** for discussion of the *method*-specific specifications such as effect size. Here we demonstrate the common options `alpha()`, `power()`, `beta()`, `onesided`, and `nfractional`.

To specify a significance level other than the default of 0.05, use the `alpha()` option:

```
gsdesign method ..., alpha(0.01) ...
```

Option `power()` specifies the desired power; alternatively, `beta()` can be used to specify type II error. For 90% power, specify

```
gsdesign method ..., power(0.9) ...
```

or, equivalently, specify

```
gsdesign method ..., beta(0.1) ...
```

For a one-sided test instead of a two-sided test, specify option `onesided`:

```
gsdesign method ..., onesided ...
```

To see fractional sample sizes instead of sample-sizes rounded up to a whole number, use option `nfractional`:

```
gsdesign method ..., nfractional ...
```

As the examples below demonstrate, these options as well as the *method*-specific syntax can be combined to obtain your desired boundary and sample-size computations for a GSD.

Efficacy stopping

The boundary-calculation procedure developed by Pocock (1977) was the first widely accepted stopping rule that allowed clinical trials to be terminated early due to treatment efficacy while maintaining desired levels of type I and type II errors. The theory underlying Pocock's boundary was formulated in the context of a z test for the difference in means between two normal responses with known variance, and it was extended to many other cases.

Pocock's stopping rule, and other efficacy bounds that have come since, defines critical values for a test statistic that is normally distributed under the null hypothesis with 0 mean and unit variance. At each interim look, the test is conducted and the test statistic is compared with the efficacy critical value. If the test statistic is equal to or exceeds the critical value, the null hypothesis is rejected early and the trial is terminated; if the test statistic is less extreme than the critical value, the trial continues to the following look.

▷ 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. Suppose that we wish to test for efficacy at three equally spaced looks using Pocock efficacy bounds. The familywise type I error allowed is 5%, while the desired power is 90%.

We use `gsbounds` to calculate and graph the stopping boundaries and compare them with those of a fixed-sample trial. If we wanted to additionally calculate sample sizes at each look, we would use command `gsdesign twomeans`; see example 9 for a demonstration. To calculate Pocock efficacy bounds, we use the `efficacy(pocock)` option. The `nlooks(3)` option specifies three equally spaced looks (two interim analyses and a final analysis). The `alpha(0.05)` and `power(0.9)` options specify the familywise significance level and power of the test, respectively.

```
. gsbounds, alpha(0.05) power(0.9) efficacy(pocock) nlooks(3)
Group sequential boundaries
Efficacy: Pocock
Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
Info. ratio = 1.1506
Fixed-study crit. values = ±1.9600
Critical values and p-values for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Upper	p-value
1	0.33	-2.2895	2.2895	0.0221
2	0.67	-2.2895	2.2895	0.0221
3	1.00	-2.2895	2.2895	0.0221

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

gsbounds displays a summary of the alpha 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 the z statistic to reach or exceed ± 2.29 at any look (which corresponds to a p-value of 0.022) 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 15% more participants than an equivalently powered fixed design, as seen by the information ratio of 1.151.

◁

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

O'Brien–Fleming boundaries have critical values that are conservative for early looks and less conservative as more data are collected. The final critical values in an O'Brien–Fleming design are similar to those of a fixed study design. Here we use the efficacy(obfleming) option to calculate O'Brien–Fleming efficacy bounds for the scenario described in the previous example.

```
. gsbounds, alpha(0.05) power(0.9) efficacy(obfleming) nlooks(3)
Group sequential boundaries
Efficacy: O'Brien-Fleming
Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
Info. ratio = 1.0161
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	0.33	-3.4711	3.4711	0.0005
2	0.67	-2.4544	2.4544	0.0141
3	1.00	-2.0040	2.0040	0.0451

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

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 ± 3.471 correspond to a p-value of 0.0005, while the critical values at the last look, ± 2.004, correspond to a p-value of 0.045. The information ratio of 1.016 indicates that the maximum sample size is only 1.6% larger than that of a fixed design.

The procedure for interim analysis with O'Brien–Fleming bounds is equivalent to the procedure we used with Pocock bounds with the exception that the critical values change from one look to the next. At the first look, we compare the test statistic z_1 against critical values ± 3.471. If $|z_1| \geq 3.471$, we reject H_0 and terminate the trial due to treatment efficacy.

If $|z_1| < 3.471$, the trial continues to the second look, where a second hypothesis test is conducted, yielding test statistic z_2. If $|z_2| \geq 2.454$, we reject H_0 and stop the trial at the second look. But if $|z_2| < 2.454$, we continue to the third and final look, where we calculate test statistic z_3.

At the final look, test statistic z_3 is compared with critical values ± 2.004. If $|z_3| \geq 2.004$, then we reject H_0, and if $|z_3| < 2.004$, then we fail to reject H_0. In the context of GSDs, it is not uncommon to discuss accepting H_0, a concept that is unheard-of in many other areas of practice. As we will see in the next section, the concept of accepting the null hypothesis holds particular appeal when applied to GSDs because it allows trials to be stopped early for futility, a practice that can be thought of as "abandoning a lost cause" (Gould 1989).

◁

Futility stopping

When the alternative hypothesis is true, the efficacy stopping rules described above can stop a trial early to reject H_0 and provide dramatic savings in sample size. But when H_0 is true, it is a type I error to reject H_0; by design, we limit the type I error probability to a small number, α. To achieve similar savings in sample size when H_0 is true, futility bounds allow us to stop a trial early to accept the null hypothesis.

There are two types of futility bounds, binding and nonbinding. If the test statistic at an interim analysis crosses a binding futility bound, H_0 must be accepted and the trial must be stopped early for futility. A trial that continues after crossing a binding futility bound is no longer subject to the familywise type I error control specified in the design. For this reason, many researchers prefer to use nonbinding futility bounds, which may be crossed without the obligation to stop the trial.

▷ Example 3: Two-sided O'Brien–Fleming efficacy and nonbinding Pocock futility bounds

Here we include the `futility(pocock)` option to add Pocock futility bounds to the design from example 2. By default, futility bounds are nonbinding. As before, we plan for three evenly spaced looks and allow an overall significance level of 5% and power of 90%.

```
. gsbounds, alpha(0.05) power(0.9) efficacy(obfleming) futility(pocock)
> nlooks(3)

Group sequential boundaries

Efficacy: O'Brien-Fleming
Futility: Pocock, nonbinding

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000

Info. ratio = 1.2601

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	0.33	-3.4711	3.4711	0.0005	-0.4661	0.4661	0.6411
2	0.67	-2.4544	2.4544	0.0141	-1.3363	1.3363	0.1814
3	1.00	-2.0040	2.0040	0.0451	-2.0040	2.0040	0.0451

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

Adding nonbinding futility bounds does not affect the calculation of the efficacy bounds, which take the same values as they did in example 2. At any analysis, if the test statistic is above the efficacy upper bound or below the efficacy lower bound, the trial will be stopped for efficacy. However, if the test statistic at an interim analysis lies between the futility lower bound and the futility upper bound, we have the option to accept H_0 and stop the trial for futility, saving resources. In practice, the decision to terminate a clinical trial is often made by an independent Data Monitoring Committee.

At the first look, we compare test statistic z_1 against the efficacy and futility critical values. If $|z_1| \geq 3.471$, we reject H_0 and stop the trial for efficacy. If $|z_1| < 0.466$, we have the option to accept H_0 and stop the trial for futility. If $|z_1| \in [0.466, 3.471)$, the trial must continue to the second look.

The procedure at the second look is the same, except the critical values are different and the continuation region, the interval between the efficacy and futility critical values, has shrunk. Test statistic z_2 is compared with the efficacy critical values, and if $|z_2| \geq 2.454$, we reject H_0 and terminate the trial. If $|z_2| < 1.336$, we have the option of stopping for futility, and if $|z_2| \in [1.336, 2.454)$, we must continue to the third and final look.

At the final look of a GSD, the efficacy bounds and the futility bounds take the same critical value because there is no continuation region at the final analysis: H_0 must be rejected or accepted. Test statistic z_3 is compared with critical values ± 2.004. If $|z_3| \geq 2.004$, then H_0 is rejected; otherwise, it is accepted.

◁

▷ Example 4: One-sided error-spending efficacy and binding futility bounds

It is common for GSDs that allow futility stopping to specify a one-sided alternative hypothesis. Here we consider the two-sided trial from example 3, but we specify a one-sided test with an overall significance level of 2.5%, half of what was used in the two-sided case. Instead of the classic Pocock and O'Brien–Fleming bounds from previous examples, here we choose error-spending Kim–DeMets bounds with parameter $\rho = 3$ for both efficacy and futility, and we make the futility bound binding.

```
. gsbounds, alpha(0.025) power(0.9) efficacy(kdemets(3))
> futility(kdemets(3), binding) nlooks(3) onesided

Group sequential boundaries

Efficacy: Error-spending Kim-DeMets, rho = 3.0000
Futility: Error-spending Kim-DeMets, binding, rho = 3.0000

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

Info. ratio = 1.0308

Fixed-study crit. value = 1.9600

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

	Info.	Efficacy		Futility	
Look	frac.	Upper	p-value	Lower	p-value
1	0.33	3.1130	0.0009	-0.7779	0.7817
2	0.67	2.4619	0.0069	0.7788	0.2180
3	1.00	1.9920	0.0232	1.9920	0.0232

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

With an efficacy upper bound and a futility lower bound, we have three possible outcomes at interim looks: efficacy stopping, futility stopping, and continuation of the trial. At the first look, we calculate test statistic z_1. If $z_1 < -0.778$, we must accept H_0 and stop the trial for futility; if $z_1 \geq 3.113$, we must reject H_0 and stop the trial for efficacy; and if $-0.778 \leq z_1 < 3.113$, we must continue to the second look.

At the second look, the efficacy and futility bounds are closer together. The testing procedure is similar to the first look, but now the test statistic z_2 is compared with a futility lower bound of 0.779 and an efficacy upper bound of 2.462. At the third and final look, the efficacy and futility bounds are equal. If $z_3 < 1.992$, we accept H_0, and if $z_3 \geq 1.992$, we reject H_0.

◁

Graphing stopping boundaries

`gsbounds` and `gsdesign` support the `graphbounds` option to display a visual representation of the stopping boundaries. This can be very helpful when designing a clinical trial and considering different configurations of stopping rules and interim analyses.

▷ Example 5: Graphing one-sided efficacy and binding futility bounds

Here we graph the stopping boundaries from the design in example 4.

```
. gsbounds, alpha(0.025) power(0.9) efficacy(kdemets(3))
> futility(kdemets(3), binding) nlooks(3) onesided graphbounds
```

(*output omitted*)

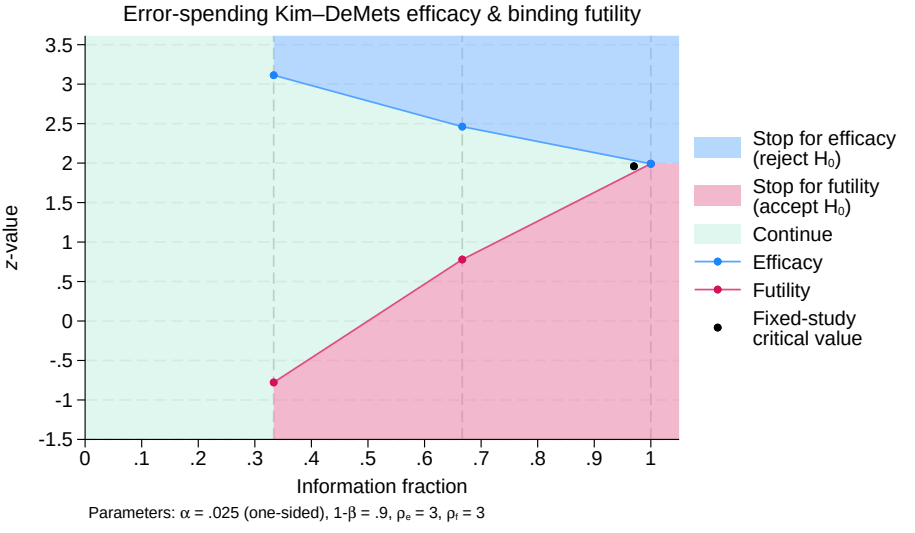

Figure 1. One-sided efficacy and futility bounds

The graph displays the bounds visually, dividing the range of possible z-values into rejection, acceptance, and continuation regions. The vertical axis is the value of the z statistic and the horizontal axis is the information fraction, 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.

We progress from left to right in the graph as information is collected during the clinical trial. The efficacy bounds, which separate the rejection region from the continuation region, are drawn in blue and marked with a dot at each look. Futility bounds separate the acceptance region from the continuation region and are drawn in red.

Before the first look (that is, when the information fraction is < 0.33), it is impossible to reject or accept H_0 because the data have not yet been analyzed, so all z-values fall within the continuation region. Beginning at the first look, the range of z-values is divided into rejection, acceptance, and continuation regions.

The continuation region at the first look is wide, encompassing z-values in the range $[-0.778, 3.113)$. By the second look, occurring with an information fraction of 0.67, the continuation region has shrunk to $[0.779, 2.462)$. At the final look, there is no continuation region because the efficacy and futility bounds meet.

The graph also includes a point marking the critical value that would be used in an equivalently powered fixed study design. This point appears at a z-value of 1.96, which gives a one-sided type I error of 0.025 in a fixed design with a single analysis. Compared with the GSD, the analysis in the fixed design occurs at an information fraction of 0.97. This is calculated as the inverse of the information ratio: $1/1.03 = 0.97$.

◁

▷ Example 6: Graphing two-sided efficacy and nonbinding futility bounds

Graphing the stopping boundaries is a particularly useful technique with complicated stopping rules and many interim analyses. Here we consider a two-sided design with efficacy and futility bounds, and interim analyses conducted at seven unevenly spaced looks.

We choose an O'Brien–Fleming efficacy bound and a nonbinding Wang–Tsiatis futility bound. Wang and Tsiatis (1987) introduced a single-parameter family of stopping bounds that includes both Pocock and O'Brien–Fleming bounds as special cases. The shape of Wang–Tsiatis bounds is determined by parameter Δ, with a Pocock bound equivalent to a Wang–Tsiatis bound with $\Delta = 0.5$, and an O'Brien–Fleming bound equivalent to a Wang–Tsiatis bound with $\Delta = 0$. Here we let $\Delta = 0.25$ to yield a futility bound that has characteristics halfway between a Pocock futility bound and an O'Brien–Fleming futility bound.

Instead of using the `nlooks()` option to specify evenly spaced looks, we use the `information()` option to provide a *numlist* of the information levels at each of the seven looks. We graph the boundaries and specify `graphbounds()` suboption `xdimlooks` to label the horizontal axis with the number of looks rather than the information fraction.

```
. gsbounds, alpha(0.05) power(0.9) efficacy(obfleming) futility(wtsiatis(0.25))
> information(0.25 0.5 0.65 0.75 0.84 0.92 1) graphbounds(xdimlooks)

Group sequential boundaries

Efficacy: O'Brien-Fleming
Futility: Wang-Tsiatis, nonbinding, Delta = 0.2500

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000

Info. ratio = 1.2409

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	0.25	-4.1845	4.1845	0.0000	.	.	.
2	0.50	-2.9589	2.9589	0.0031	-0.7473	0.7473	0.4549
3	0.65	-2.5951	2.5951	0.0095	-1.2198	1.2198	0.2225
4	0.75	-2.4159	2.4159	0.0157	-1.4952	1.4952	0.1349
5	0.84	-2.2828	2.2828	0.0224	-1.7231	1.7231	0.0849
6	0.92	-2.1813	2.1813	0.0292	-1.9128	1.9128	0.0558
7	1.00	-2.0923	2.0923	0.0364	-2.0923	2.0923	0.0364

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

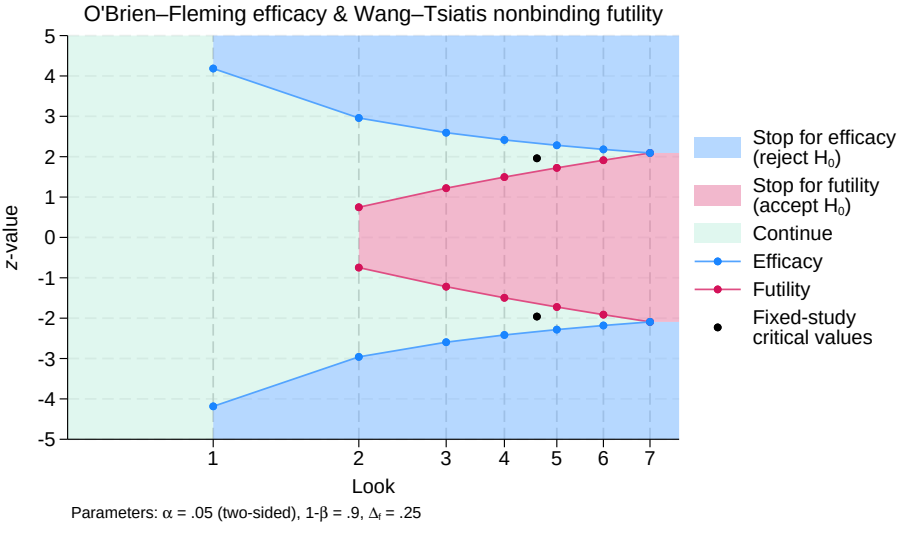

Figure 2. Two-sided efficacy and futility bounds

On the graph, we see the acceptance region displayed as a truncated inner wedge, and on the table of stopping boundaries, we see that the futility critical values for the first look are missing. This is because, to attain the specified significance level and power, the futility lower bound would have been greater than the futility upper bound, implying that futility stopping is impossible at the first look.

Boundary and sample-size calculations using gsdesign

The previous examples have used `gsbounds` to calculate stopping bounds, but when designing a group sequential clinical trial, you will want to know the sample size at each look as well as the boundary critical values. This is done using the `gsdesign` *method* set of commands, where *method* is `onemean`, `oneproportion`, `twomeans`, `twoproportions`, `logrank`, or even a user-defined method.

One-sample tests

The gold standard for clinical trials is the randomized controlled trial, where participants are randomly assigned to one of two groups: one group receives the experimental treatment while the other group is kept as a control. The groups are often called arms, and the experimental arm will receive the experimental treatment. The control arm will receive either a placebo (an inactive substance such as a sugar pill, or a "sham" procedure for nonpharmacological trials) or an active control (typically the standard of care, a treatment that has been previously studied and is known to be effective).

However, there are some scenarios where randomizing subjects to a control group would be impractical or unethical, such as a clinical trial of a treatment for a serious condition where there is a moral argument against giving participants a placebo but there is no existing standard of care. In these cases, a single-arm clinical trial is desired, and a one-sample test is conducted.

Example 7: Boundary and sample-size calculations for a one-sample mean test

We consider a clinical trial of the chemotherapy medicine sunitinib as a treatment for advanced non–small cell lung cancer. Suppose that we are interested in developing a treatment for patients whose cancers have not responded to the standard treatment options. There is no possibility of forming an active control group with this population because the standard of care has already proven ineffective for them. The clinical outcomes for patients with untreated advanced non–small cell lung cancer are known to be very poor, so we have ethical reasons to avoid creating a placebo control group. We decide to conduct a single-arm clinical trial and perform a one-sample test.

The clinical endpoint of this study is the tumor shrinkage rate (TSR), a measure of how quickly a participant's largest tumor is shrinking (or growing, in the case of negative TSR values). We want to test whether the mean TSR is greater than 0 with a one-sided test and a familywise significance level of 2.5%. We anticipate the standard deviation of the TSR to be 2, and we require 90% power to detect a mean TSR of 0.5. We plan on conducting two evenly spaced looks at the data, and we will use an O'Brien–Fleming efficacy bound.

```
. gsdesign onemean 0 0.5, sd(2) alpha(0.025) power(0.9) efficacy(obfleming)
> nlooks(2) onesided

Group sequential design for a one-sample mean test
t test
H0: m = m0 versus Ha: m > m0

Efficacy: O'Brien-Fleming

Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.9000
      delta = 0.2500
         m0 = 0.0000
         ma = 0.5000
         sd = 2.0000

Expected sample size:
         H0 = 171.78
         Ha = 145.20

Info. ratio = 1.0071
    N fixed =    171
      N max =    172

Fixed-study crit. value = 1.9600

Critical values, p-values, and sample sizes
for a group sequential design
```

Look	Info. frac.	Efficacy Upper	Efficacy p-value	Sample size N
1	0.50	2.7965	0.0026	86
2	1.00	1.9774	0.0240	172

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

`gsdesign onemean` displays the specified study parameters, including `m0`, the mean under the null hypothesis; `ma`, the mean under the alternative hypothesis; and `delta`, the difference in means divided by the standard deviation.

The next section of output displays the expected sample size, which is the average sample size if the group sequential trial were to be repeated many times. The average sample size under H_0 is 171.78, nearly the same as the maximum of 172 participants at the final look. This is expected because our

design does not allow for early stopping to accept H_0. If H_a is true, we expect an average of only 145.2 participants because of the probability of early stopping to reject H_0, a savings over the 171 participants required by the fixed design.

Next we see the information ratio, the sample size for a fixed study with an equivalent significance level and power (`N fixed`), and the maximum sample size of the GSD (`N max`). The information ratio is the ratio of the maximum sample size of the GSD to the fixed-study sample size. We then see the critical value for a fixed study with an equivalent significance level.

At the end of the display is a table of stopping boundaries, p-values, and sample sizes for the two looks. The efficacy critical values in the table can be compared directly with the z statistic from a one-sided z test of whether the mean TSR is equal to 0. We do not presume to know the population standard deviation a priori (which is why we did not specify the `knownsds` option), so we must estimate the standard deviation when conducting the one-sample mean test. This would indicate that the proper one-sample mean test for this study is a t test, which yields a t statistic, not a z statistic.

With these rather large sample sizes (especially at the second look), it would be common to conduct a large-sample z test in this scenario. The use of this test relies on the fact that the estimate of the population standard deviation improves with increasing sample size. The distribution of the test statistic asymptotically approaches a normal distribution, enabling the use of a z test with large samples, even with unknown standard deviation. However, if we prefer to conduct a t test, we can instead use the significance level approach and compare the p-value from the t test against the p-values corresponding to the boundary critical values, which are also reported in this table.

For more examples of `gsdesign onemean`, see [ADAPT] **gsdesign onemean**.

◁

▷ Example 8: Boundary and sample-size calculations for a one-sample proportion test

We consider an alternate endpoint for the clinical trial of sunitinib as a treatment for advanced non–small cell lung cancer described in example 7. Instead of measuring the TSR, suppose we are interested in the objective response rate (ORR), defined as the proportion of participants that exhibit at least a partial response to therapy. It is important to emphasize that the outcome of each participant is binary (either they exhibit a response to therapy or they do not), and we calculate the proportion as the number of participants who exhibit a response divided by the total number of participants.

We can use `gsdesign oneproportion` to determine the required sample sizes if we wish to determine whether the ORR of participants receiving sunitinib is greater than 5%, and we plan to conduct a one-sided proportion test at the 2.5% familywise significance level. We require 90% power to detect an ORR of 10%. We will conduct two evenly spaced looks using an O'Brien–Fleming efficacy bound and a nonbinding Pocock futility bound, which we graph.

```
. gsdesign oneproportion 0.05 0.1, alpha(0.025) power(0.9) efficacy(obfleming)
> futility(pocock) nlooks(2) onesided graphbounds

Group sequential design for a one-sample proportion test
Score z test
H0: p = p0 versus Ha: p > p0

Efficacy: O'Brien-Fleming
Futility: Pocock, nonbinding

Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.9000
      delta = 0.0500
         p0 = 0.0500
         pa = 0.1000

Expected sample size:
         H0 = 181.12
         Ha = 251.76

Info. ratio = 1.1662
    N fixed =    264
      N max =    308

Fixed-study crit. value = 1.9600

Critical values, p-values, and sample sizes for a group sequential design
```

Look	Info. frac.	Efficacy Upper	Efficacy p-value	Futility Lower	Futility p-value	Sample size N
1	0.50	2.7965	0.0026	0.9521	0.1705	154
2	1.00	1.9774	0.0240	1.9774	0.0240	308

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

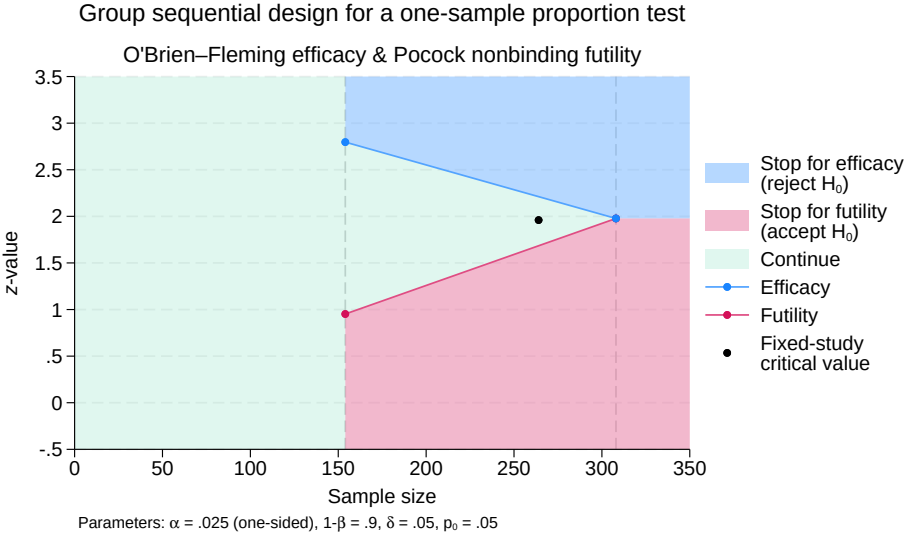

Figure 3. One-sided test of one proportion with efficacy and futility bounds

Once we have collected data from 154 participants, we could conduct a large-sample test of one proportion with command `prtest`, which yields a z statistic, z_1; see [R] **prtest**. If $z_1 \geq 2.797$, we reject H_0 and declare the treatment to be effective, and if $z_1 < 0.952$, we can choose to accept H_0 and terminate the trial due to futility or we can continue the trial. If $z_1 \in [0.952, 2.797)$, we must continue the trial because z_1 lies in the continuation region. At the second and final look, there is no continuation region; if $z_2 \geq 1.977$, we reject H_0, and if $z_2 < 1.977$, we accept H_0.

Compared with a fixed study design with equivalent significance level and power, this GSD has a larger maximum sample size (308 participants versus 264 for the fixed trial). But the group sequential trial has a smaller expected sample size than the fixed trial under both the null and the alternative hypotheses. If this trial were to be repeated many times, on average it would require only 181.12 participants if H_0 was true and only 251.76 participants if H_a was true, which is fewer than the 264 required for the fixed trial.

For more examples of `gsdesign oneproportion`, see [ADAPT] **gsdesign oneproportion**.

◁

Two-sample tests

In a classic randomized controlled trial, participants are randomly assigned to one of two groups: the experimental group (which receives the treatment being tested) and the control group (which receives either a placebo or the existing standard of care, if one exists). The two groups are often called arms, making this a two-arm trial. Examples of treatments include new drugs, medical devices, and medical procedures. To determine the efficacy of the treatment, the responses of participants in the experimental arm are compared with the responses of participants in the control arm.

When the responses are continuous, a two-sample test of means can be performed to determine whether the mean of the experimental arm is the same as that of the control arm. When the response from each participant is binary, a two-sample test of proportions can be performed to determine whether the proportion of "successes" in the control arm is the same as the proportion in the experimental arm.

▷ Example 9: Boundary and sample-size calculations for a two-sample means test

Subarachnoid hemorrhage (SAH) is a type of stroke that is typically caused by head trauma or a brain aneurysm, and a large proportion of patients who survive SAH are affected by cerebral vasospasm during their recovery. Fatal vasospasm occurs in approximately 5 to 10% of patients who are hospitalized for SAH (Macdonald, Pluta, and Zhang 2007). One way to detect vasospasm is by measuring peak systolic velocity (PSV) of blood in the middle cerebral artery. In a preliminary study of high-dose intraarterial nicardipine as a treatment for cerebral vasospasm, Badjatia et al. (2004) defined mild vasospasm as time-averaged PSV of 200–249 cm/s, moderate vasospasm as PSV of 250–299 cm/s, and severe vasospasm as PSV in excess of 300 cm/s. Suppose that we want to design a clinical trial that compares nicardipine to papaverine, the standard intraarterial treatment for vasospasm following SAH. We assign participants to the experimental and control arms in a 1:1 ratio, and we measure the ΔPSV (percent reduction in PSV) of each participant.

The analysis will compare the average ΔPSV in the control arm, μ_1, against the average ΔPSV in the experimental arm, μ_2. We will test the null hypothesis $H_0 : \mu_1 = \mu_2$ versus the one-sided alternative $H_a : \mu_2 > \mu_1$ with a familywise significance level of 2.5%. We use `gsdesign twomeans` to calculate sample sizes for a GSD that requires 90% power to detect the difference between a 15% reduction in mean ΔPSV in the control arm and a 20% mean reduction in the experimental arm, with a common standard deviation of 20.

We specify `efficacy(wtsiatis(0.25))` to use a Wang–Tsiatis efficacy bound with parameter $\Delta_e = 0.25$, and we specify `futility(obfleming)` to use a nonbinding O'Brien–Fleming futility bound. The nonbinding futility bound allows us to accept H_0 and terminate the trial for futility if it is crossed, but if we choose to continue the trial despite crossing the nonbinding futility bound, the familywise type I error is still controlled at the 2.5% significance level. We specify four analyses with 30%, 60%, 80%, and 100% of the data.

```
. gsdesign twomeans 15 20, sd(20) alpha(0.025) power(0.9)
> efficacy(wtsiatis(0.25)) futility(obfleming)
> information(30 60 80 100) onesided graphbounds

Group sequential design for a two-sample means test
t test assuming sd1 = sd2 = sd
H0: m2 = m1 versus Ha: m2 > m1

Efficacy: Wang-Tsiatis, Delta = 0.2500
Futility: O'Brien-Fleming, nonbinding

Study parameters:
      alpha =  0.0250  (upper one-sided)
      power =  0.9000
      delta =  5.0000
         m1 = 15.0000
         m2 = 20.0000
         sd = 20.0000

Expected sample size:
         H0 =  438.96
         Ha =  518.27

Info. ratio =  1.1631
    N fixed =     676
      N max =     786
     N1 max =     393
     N2 max =     393

Fixed-study crit. value = 1.9600

Critical values, p-values, and sample sizes for a group sequential design
```

Look	Info. frac.	Efficacy Upper	Efficacy p-value	Futility Lower	Futility p-value
1	0.30	2.8703	0.0021	-0.5895	0.7222
2	0.60	2.4136	0.0079	0.9371	0.1743
3	0.80	2.2461	0.0123	1.5933	0.0555
4	1.00	2.1243	0.0168	2.1243	0.0168

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

Look	Sample size N1	N2	N
1	118	118	236
2	236	236	472
3	314	314	628
4	393	393	786

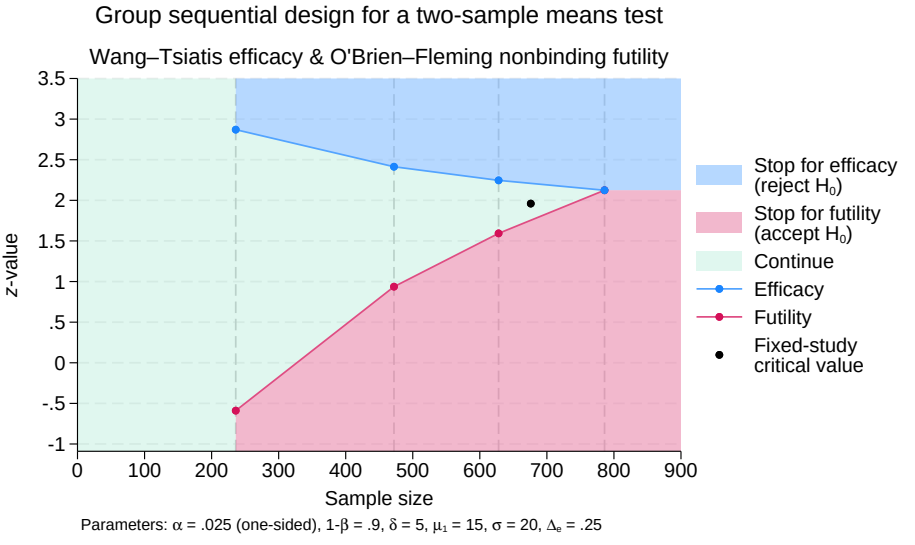

Figure 4. One-sided test of the equality of two means with efficacy and futility bounds

`gsdesign twomeans` begins by displaying a description of the test being performed, a list of the requested boundaries, and a summary of the parameters used in the design.

The next section of output displays the expected sample size, which is the average sample size if the group sequential trial were to be repeated many times. On average, we expect this trial to require 438.96 participants if H_0 is true and 518.27 participants if H_a is true.

Next we see the information ratio, the sample size for a fixed study with an equivalent significance level and power (`N fixed`), the maximum sample size of the GSD (`N max`), and the maximum sample sizes for each group (`N1 max` and `N2 max`). The information ratio is the ratio of the maximum sample size of the GSD to the fixed-study sample size. We then see the critical value for a fixed study with an equivalent significance level.

Finally, `gsdesign twomeans` displays tables with the critical values and p-values for the stopping boundaries as well as the sample sizes at each look. The first look occurs once ΔPSV has been recorded from 118 participants in each arm. With such a large sample, we conduct a z test instead of a t test because the two tests are asymptotically equivalent as the sample size increases. The z statistic from this large-sample z test, z_1, is compared with the boundary critical values. If $z_1 \geq 2.87$, we will reject H_0 and terminate the trial early due to treatment efficacy. If $z_1 < -0.59$, we have the option to stop the trial for futility, but the familywise type I error will still be controlled at the 2.5% level should the trial proceed. If $z_1 \in [-0.59, 2.87)$, the trial must continue.

When we have ΔPSV for 236 participants in each arm, we will perform another large-sample z test and compare the test statistic, z_2, with the boundary critical values for the second look. If $z_2 \geq 2.414$, we reject H_0 and end the trial for efficacy, while if $z_2 < 0.937$, we have the option of stopping the trial for futility and accepting H_0. If $z_2 \in [0.937, 2.414)$, we must continue the trial. At the third look, the testing procedure is similar, but the continuation region has shrunk to $z_3 \in [1.593, 2.246)$. If the trial continues to the fourth and final look, with a total of 786 participants, there is no continuation region, because the futility critical value is the same as the efficacy critical value. If $z_4 \geq 2.124$, we reject H_0; otherwise, we accept H_0.

For more examples of `gsdesign twomeans`, see [ADAPT] **gsdesign twomeans**.

◁

▷ Example 10: Boundary and sample-size calculations for a two-sample proportions test

We consider a variation of the study of nicardipine as a treatment for vasospasm, as described in example 9. Suppose we are interested in an alternate endpoint: the proportion of participants whose vasospasm is resolved because of the treatment. We will record a participant's response as 1 if their time-averaged PSV in the middle cerebral artery is below 200 cm/s after treatment, and we will record their response as 0 if their PSV is 200 cm/s or above.

Participants will be randomly assigned to the experimental arm, whose members receive intraarterial nicardipine, or to the control group, whose members receive the standard of care, which is intraarterial papaverine, in a 1:1 ratio. Based on previous research from Badjatia et al. (2004) and others, we anticipate that a single treatment will resolve vasospasm in 50% of control-group participants and 60% of experimental-group participants. We will test whether the two proportions are the same by using a one-sided Pearson's χ^2 test with familywise significance level of 2.5% and power of 90% to detect the difference between $p_1 = 0.5$ and $p_2 = 0.6$.

To stop the trial early for evidence of treatment efficacy, we will use an error-spending approximation of the O'Brien–Fleming bound, and for futility stopping, we will use a nonbinding error-spending Hwang–Shih–de Cani bound with parameter $\gamma_f = -2$. If the test statistic from an interim analysis crosses a nonbinding futility bound, we have the option to accept H_0 and terminate the trial, saving resources and "abandoning a lost cause," but if we continue the trial, the familywise type I error is still controlled. We plan three evenly spaced looks, two interim analyses, and one final analysis.

```
. gsdesign twoproportions .5 .6, alpha(0.025) power(0.9) efficacy(errobfleming)
> futility(hsdecani(-2)) nlooks(3) onesided graphbounds

Group sequential design for a two-sample proportions test
Pearson's chi-squared test
H0: p2 = p1 versus Ha: p2 > p1

Efficacy: Error-spending O'Brien-Fleming style
Futility: Error-spending Hwang-Shih-de Cani, nonbinding, gamma = -2.0000

Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.9000
      delta = 0.1000  (difference)
         p1 = 0.5000
         p2 = 0.6000

Expected sample size:
         H0 = 650.03
         Ha = 869.55

Info. ratio = 1.0665
    N fixed =  1,038
      N max =  1,106
     N1 max =    553
     N2 max =    553

Fixed-study crit. value = 1.9600

Critical values, p-values, and sample sizes for a group sequential design
```

	Info.	Efficacy		Futility	
Look	frac.	Upper	p-value	Lower	p-value
1	0.33	3.7103	0.0001	-0.2418	0.5955
2	0.67	2.5114	0.0060	0.9367	0.1745
3	1.00	1.9930	0.0231	1.9930	0.0231

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

	Sample size		
Look	N1	N2	N
1	185	185	370
2	369	369	738
3	553	553	1,106

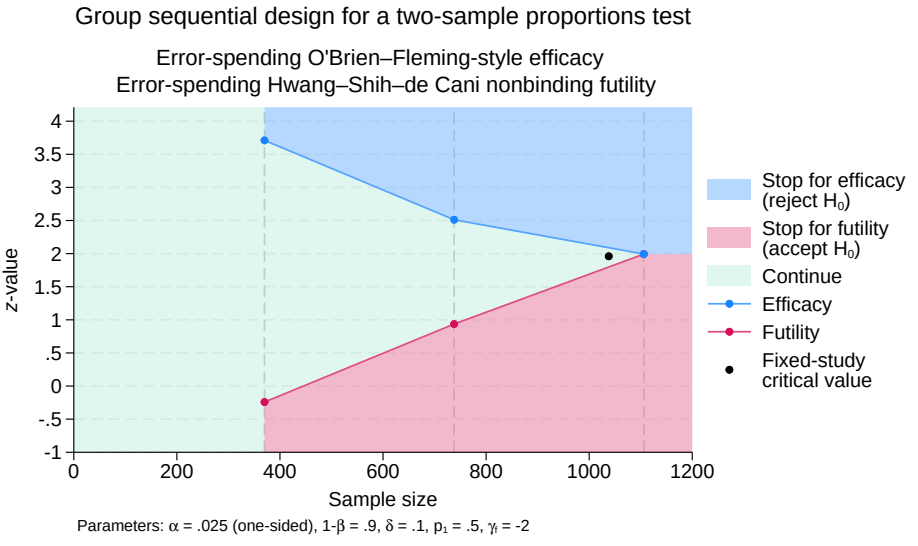

Figure 5. One-sided test of the equality of two proportions with efficacy and futility bounds

`gsdesign twoproportions` shows the specified study parameters, including the control-group proportion `p1`, the experimental-group proportion `p2`, and the difference in proportions `delta`.

The next section of output displays the expected sample size under the null and alternative hypotheses. The expected sample size is the average sample size (taking into account early stopping) that would be observed if this trial were to be repeated many times. If H_0 is true, our trial will require an average of 650.03 participants, and if H_a is true, we will require an average of 869.55 participants.

Next we see the information ratio, the sample size for a fixed study with an equivalent significance level and power (`N fixed`), the maximum sample size of the GSD (`N max`), and the maximum sample sizes for each group (`N1 max` and `N2 max`). The information ratio is the ratio of the maximum sample size of the GSD to the fixed-study sample size. We then see the critical value for a fixed study with an equivalent significance level.

Finally, `gsdesign twoproportions` displays tables with the critical values and p-values for the stopping boundaries as well as the sample sizes at each look. At the first look, we will conduct Pearson's χ^2 test with command `prtest`, which reports a z statistic, z_1, that can be compared directly with the boundary critical values. Just like the classical O'Brien–Fleming boundary, the error-spending O'Brien–Fleming-style efficacy bound is very conservative at early looks, with a critical value at the first look of 3.71, which corresponds to a p-value of 0.0001.

On the graph, we see that if $z_1 \geq 3.71$, it lies in the blue rejection region, so we will reject H_0 and stop the trial early for efficacy. If $z_1 < -0.242$, it lies in the red acceptance region, and we have the option of accepting H_0 and stopping the trial for futility or continuing the trial without overrunning the 2.5% familywise type I error. If $z_1 \in [-0.242, 3.71)$, then z_1 lies in the green continuation region and the trial must continue.

At the second look, the testing procedure is similar, but the efficacy and futility critical values are closer together, shrinking the continuation region to $z_2 \in [0.937, 2.511)$. At the third and final look, the efficacy critical values equal the futility critical values, so there is no continuation region. If $z_3 \geq 1.993$, we reject H_0; otherwise, we accept H_0.

For more examples of `gsdesign twoproportions`, see [ADAPT] **gsdesign twoproportions**.

◁

Survival analysis

When analyzing time-to-event data, we often want to compare the survivor functions of two groups. If we denote the time of failure as T, we can define the survivor function as the probability of surviving beyond time t, expressed mathematically as $S(t) = \Pr(T > t)$. A related term is the hazard function, the instantaneous rate of failure at time t, conditional on survival until time t, written as $h(t)$.

Consider a survival study comparing survivor functions in two groups by using the log-rank test, and let $S_1(t)$ and $S_2(t)$ denote the survivor functions of the control and the experimental groups, respectively. The log-rank test is most appropriate when the hazard functions are thought to be proportional across the groups, in which case it is the most powerful nonparametric test of $S_1(\cdot) = S_2(\cdot)$. The proportional-hazards assumption can be written as $h_2(t) = \Delta h_1(t)$ for all t or, equivalently, $S_2(t) = \{S_1(t)\}^{\Delta}$, where Δ is the hazard ratio. If $\Delta < 1$, then survival in the experimental group is higher than survival in the control group, which means that the experimental treatment is superior to the control treatment. If $\Delta > 1$, then the control treatment is superior to the experimental treatment.

Sample-size calculations for the log-rank test compute the number of events observed in the study. The required sample size is equal to the required number of events if a failure event is observed for every participant in the trial. Often, the time of failure is not known for some participants, a phenomenon known as censoring. Administrative censoring occurs when a trial ends before all participants have experienced a failure event. Nonadministrative censoring occurs when participants withdraw from the study or are lost to follow-up. If censoring occurs in the study, the required number of participants will be greater than the required number of events.

▷ Example 11: Boundary and sample-size calculations for a log-rank test

The Beta-Blocker Heart Attack Trial (BHAT) was one of the first large-scale clinical trials to adopt a group sequential monitoring plan (DeMets et al. 1984). This was a double-blind study in which participants who had experienced a heart attack were randomized to one of two groups: the control group (which received a placebo) and the intervention group (which received the beta-blocker propranolol). The endpoint, or outcome of interest, was time until death by any cause, and survival analysis was conducted using a log-rank test.

The BHAT's independent Policy and Data Monitoring Board adopted the then-recently published O'Brien–Fleming method for calculating efficacy bounds, but here we consider how the trial could have been designed using methods that were not available at the time. The original BHAT was powered to detect the difference between nonadherence-adjusted three-year survival probabilities of 82.54% for the control group and 86.25% for the intervention group. Seven biannual analyses were scheduled for 11, 16, 21, 28, 34, 40, and 48 months into the study. The log-rank test statistic crossed the O'Brien–Fleming boundary at the sixth of seven looks, and the BHAT was terminated for treatment efficacy eight months before the trial was scheduled to end.

Here we use `gsdesign logrank` to calculate sample sizes for a design that is inspired by the BHAT but that allows for both efficacy and futility stopping. We will conduct a one-sided test of hazard ratio Δ, with $H_0 : \Delta = 1$ versus $H_a : \Delta < 1$. We will allow a one-sided familywise type I error rate of 2.5%, and we require 90% power to detect the difference in survival probability described above. We will use the error-spending approximation of O'Brien–Fleming bounds for efficacy stopping and nonbinding Kim–DeMets futility bounds with parameter $\rho_f = 3$. Instead of spacing the seven looks evenly, we use the `information()` option and follow Method 2 from Lan and DeMets (1989, 1195) to specify the timing of interim looks based on calendar time, which we use as the horizontal axis of our graph.

```
. gsdesign logrank 0.8254 0.8625, alpha(0.025) power(0.9)
> efficacy(errobfleming) futility(kdemets(3))
> information(11 16 21 28 34 40 48) onesided
> graphbounds(xdiminformation xtitle("Months"))

Group sequential design for two-sample comparison of survivor functions
Log-rank test, Freedman method
H0: HR = 1 versus Ha: HR < 1

Efficacy: Error-spending O'Brien-Fleming style
Futility: Error-spending Kim-DeMets, nonbinding, rho = 3.0000

Study parameters:
      alpha = 0.0250  (lower one-sided)
      power = 0.9000
      delta = 0.7709  (hazard ratio)
     hratio = 0.7709

Censoring:
         s1 = 0.8254
         s2 = 0.8625
       Pr_E = 0.1560

Expected number of events:
         H0 = 378.92
         Ha = 469.55

Info. ratio = 1.0727
    E fixed =    628
    N fixed =  4,024
      N max =  4,316
     N1 max =  2,158
     N2 max =  2,158

Fixed-study crit. value = -1.9600

Critical values, p-values, and sample sizes for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Efficacy p-value	Futility Upper	Futility p-value	Events E
1	0.23	-4.5380	0.0000	1.4276	0.9233	155
2	0.33	-3.7128	0.0001	0.7980	0.7876	225
3	0.44	-3.2081	0.0007	0.2509	0.5991	295
4	0.58	-2.7361	0.0031	-0.4339	0.3322	393
5	0.71	-2.4739	0.0067	-0.9312	0.1759	477
6	0.83	-2.2717	0.0116	-1.3987	0.0810	562
7	1.00	-2.0473	0.0203	-2.0473	0.0203	674

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

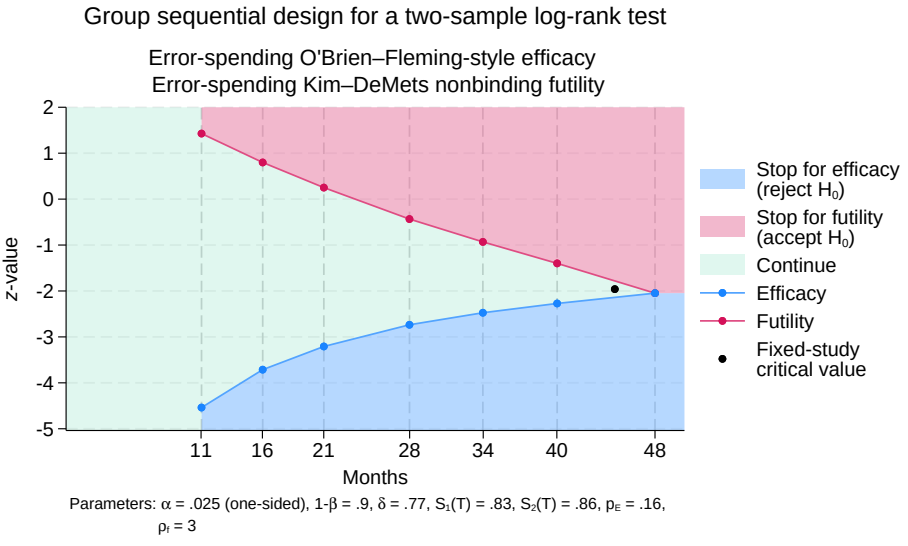

Figure 6. One-sided log-rank test with efficacy and futility bounds

At the top of the output, `gsdesign logrank` displays a description of the trial with null and alternative hypotheses as well as study parameters. We see that the survival probabilities 0.8254 and 0.8625 correspond to a hazard ratio of 0.7709, which is the effect size used when calculating the number of events necessary to achieve 90% power.

If the null hypothesis is correct (the hazard ratio is 1) and this trial were to be repeated many times, we would expect to observe an average of 378.92 events per trial. If the hazard ratio is truly 0.7709 (the value under the alternative hypothesis) and the trial were to be repeated many times, we would expect an average of 469.55 events per trial.

A fixed study would require 628 events (deaths) to detect a hazard ratio of 0.7709 with 90% power, which, with the specified survival probabilities, corresponds to a sample size of 4,024. The GSD requires a maximum of 674 events (corresponding to a sample of size 4,316) if it continues to the final look.

The table at the end of the output displays the critical values and p-values for stopping boundaries and the sample sizes at each look, where sample size is reported as the number of events observed. Boundary critical values are reported on the z scale and are designed to be compared against the z statistic from a log-rank test. Command `sts test` (see [ST] **sts test**) conducts the log-rank test and reports a χ^2 test statistic, which is not directly comparable with the z scale critical values. However, the square root of the χ^2 test statistic is a z statistic, which can be directly compared with the boundary critical values.

The first look occurs 11 months into the study, at which point 155 events are expected to have occurred, and a log-rank test is performed. We denote the square root of the χ^2 test statistic from the first look as z_1, and we note that the sign of z_1 depends on whether the observed hazard ratio was greater than 1 (in which case z_1 is positive) or less than 1 (in which case z_1 is negative). If $z_1 \leq -4.538$, we say that z_1 lies in the rejection region (shaded blue on the graph) and we reject H_0, terminating the trial early due to treatment efficacy. If $z_1 > 1.428$, it lies in the acceptance region and we may terminate the trial for futility; however, if the trial proceeds, the familywise type I error is still controlled at the 2.5% level. If $z_1 \in (-4.538, 1.428]$, then z_1 lies in the green continuation region and the trial must continue.

The testing procedure is similar at each of the following interim looks, with the efficacy bound increasing and the futility bound decreasing at each look, shrinking the continuation region. At the seventh and final look, the efficacy critical value is equal to the futility critical value and there is no continuation region. If $z_7 \leq -2.047$, we reject H_0; otherwise, we accept H_0.

For more examples, see [ADAPT] **gsdesign logrank**.

◁

Add your own methods

The `gsdesign` command provides several built-in methods, and additional `power` methods can be used with the `methodok` option. However, if you want to design a clinical trial using a method that is not included, you can write your own sample-size calculation and use it with `gsdesign`.

All you need to do is write a program that computes the sample size for a fixed study; `gsdesign` will calculate the stopping boundaries, information ratio, and sample sizes at each look. The procedure for adding a method to `gsdesign` is identical to the procedure for adding a sample-size calculation to the `power` command. Detailed instructions can be found in [ADAPT] ***gsdesign usermethod***, but a quick guide is as follows:

1. Create a program that computes a fixed-study sample size and follows `power`'s naming convention: `power_cmd_`*mymethod*, where *mymethod* is the name of your method.
2. Ensure your program accepts the `nfractional` option. This is necessary because `gsdesign` uses the fractional sample size when calculating the sample required at each look.
3. Store the resulting sample size following `power`'s simple naming conventions. Store the total sample size in `r(N)`. For two-sample methods, additionally store control-group and experimental-group sample sizes in `r(N1)` and `r(N2)`, respectively. For time-to-event methods, additionally store the number of events in `r(E)` and store macro `r(endpoint)` as "`survival`".
4. Place your program `power_cmd_`*mymethod* where Stata can find it.

▷ Example 12: Group sequential design with user-defined methods

To show how easy this is, let's write a program to compute sample size for a fixed-study one-sample z test given standardized difference, significance level, and power. For simplicity, we assume a two-sided test.

We will call our new method `myztest`.

```
program power_cmd_myztest, rclass
        version 19.5      // (or version 19 if you do not have StataNow)

        syntax, STDDiff(real)      /// standardized difference (effect size)
                [ Alpha(real 0.05) /// significance level
                  Power(real 0.8)  /// power
                  NFRACtional      /// report fractional sample size
                ]

        tempname N
        scalar `N' = ((invnormal(`power') + invnormal(1 - `alpha' / 2)) / `stddiff')^2
        if ("`nfractional'" == "") {
                scalar `N' = ceil(`N')
        }

        return scalar power   = `power'
        return scalar N       = `N'
        return scalar alpha   = `alpha'
        return scalar stddiff = `stddiff'
end
```

The computation in this program is trivial, but yours could be as complicated as you like. It could even involve simulation to compute the sample size.

With our program in hand, we can design a clinical trial using the default values of 5% familywise significance level, 80% power, and an O'Brien–Fleming efficacy boundary with two evenly spaced looks. We need only specify the effect size by using `stddiff()`.

```
. gsdesign myztest, stddiff(0.7)

Group sequential design for myztest
Two-sided test

Efficacy: O'Brien-Fleming

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.8000

Expected sample size:
         H0 =  16.96
         Ha =  15.06

Info. ratio = 1.0078
    N fixed =     17
      N max =     17

Fixed-study crit. values = ±1.9600

Critical values, p-values, and sample sizes for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Efficacy Upper	p-value	Sample size N
1	0.50	-2.7965	2.7965	0.0052	9
2	1.00	-1.9774	1.9774	0.0480	17

```
Notes: Critical values are for z statistics; otherwise,
       use p-value boundaries.
       Requested information fraction not attained.
```

gsdesign called our program power_cmd_myztest for the sample-size calculation for a fixed design and used the stored result r(N) to calculate the sample sizes at both looks. In this case, the use of a user-defined program to calculate sample size was purely for didactic purposes; the same calculation could have been conducted with built-in command gsdesign onemean, diff(0.7) sd(1) knownsd.

This example was simple, but all the standard gsdesign options apply to user-defined methods. For example, suppose we wanted to design a trial using a one-sample z test at the familywise 10% level with 90% power to detect a standardized difference of 0.3. We use Wang–Tsiatis efficacy bounds with parameter $\Delta_e = 0.25$ and binding Wang–Tsiatis futility bounds with parameter $\Delta_f = 0.3$. We require six looks, spaced at 30%, 50%, 70%, 80%, 90%, and 100% of the data, and we graph the bounds with a custom subtitle.

```
. gsdesign myztest, stddiff(0.3) alpha(0.1) power(0.9) efficacy(wtsiatis(0.25))
> futility(wtsiatis(0.3), binding) information(30 50 70 80 90 100)
> graphbounds(subtitle("One-sample z test"))

Group sequential design for myztest
Two-sided test

Efficacy: Wang-Tsiatis, Delta = 0.2500
Futility: Wang-Tsiatis, binding, Delta = 0.3000

Study parameters:
      alpha = 0.1000  (two-sided)
      power = 0.9000

Expected sample size:
         H0 =  74.34
         Ha =  64.92

Info. ratio = 1.2596
    N fixed =     96
      N max =    120

Fixed-study crit. values = ±1.6449

Critical values, p-values, and sample sizes for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Efficacy Upper	Efficacy p-value	Futility Lower	Futility Upper	Futility p-value
1	0.30	-2.4353	2.4353	0.0149	.	.	.
2	0.50	-2.1434	2.1434	0.0321	-0.6200	0.6200	0.5352
3	0.70	-1.9704	1.9704	0.0488	-1.1563	1.1563	0.2476
4	0.80	-1.9057	1.9057	0.0567	-1.3880	1.3880	0.1651
5	0.90	-1.8504	1.8504	0.0642	-1.6022	1.6022	0.1091
6	1.00	-1.8023	1.8023	0.0715	-1.8023	1.8023	0.0715

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

Look	Sample size N
1	36
2	60
3	84
4	96
5	108
6	120

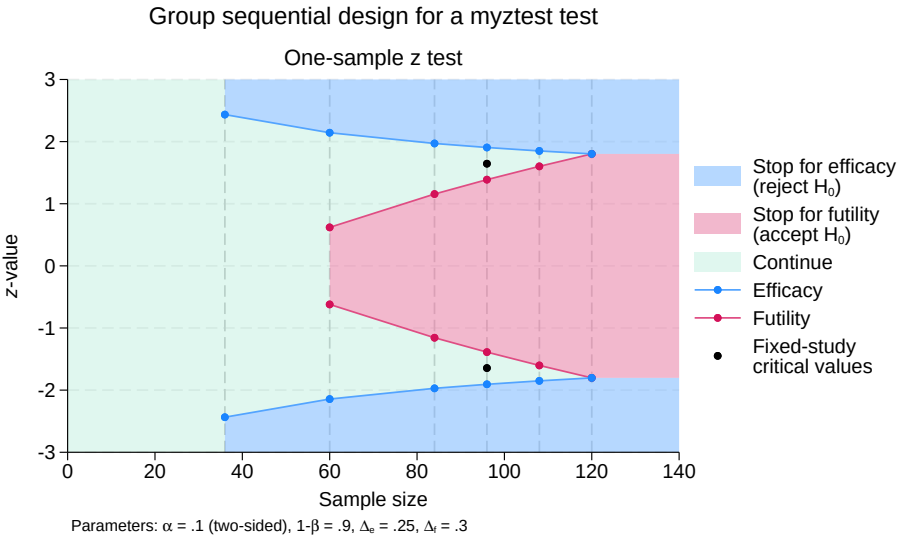

Figure 7. User-written one-sample z test with efficacy and futility bounds

Our program `power_cmd_myztest` need only handle the sample-size calculation in the case of a fixed study design; `gsdesign` handles the rest, including the graph.

For more examples, see [ADAPT] ***gsdesign usermethod***.

Stored results

See *Stored results* in [ADAPT] **gsbounds**.

See *Stored results* in [ADAPT] **gsdesign**.

Also see *Stored results* in the `gsdesign` *method*-specific entries.

Acknowledgments

Stata has an active research community adding features to the area of GSD. We would like to acknowledge their previous and ongoing contributions: `doubletriangular`, `haybittlepto`, `innerwedge`, `powerfamily`, `triangular`, and `wangtsiatis` by Michael J. Grayling, James M. S. Wason, and Adrian P. Mander; `desma` by Michael J. Grayling; `nstage` by Alexandra Blenkinsop and Babak Choodari-Oskooei; `stopbound` by Bryan Fellman; and more. Type `search group sequential design` to see Stata's official and community-contributed features for GSD.

References

Badjatia, N., M. A. Topcuoglu, J. C. Pryor, J. D. Rabinov, C. S. Ogilvy, B. S. Carter, and G. A. Rordorf. 2004. Preliminary experience with intra-arterial nicardipine as a treatment for cerebral vasospasm. *American Journal of Neuroradiology* 25: 819–826.

DeMets, D. L., R. J. Hardy, L. W. Friedman, and K. K. G. Lan. 1984. Statistical aspects of early termination in the beta-blocker heart attack trial. *Controlled Clinical Trials* 5: 362–372. https://doi.org/10.1016/S0197-2456(84)80015-X.

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.

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. 1989. Group sequential procedures: Calendar versus information time. *Statistics in Medicine* 8: 1191–1198. https://doi.org/10.1002/sim.4780081003.

Macdonald, R. L., R. M. Pluta, and J. H. Zhang. 2007. Cerebral vasospasm after subarachnoid hemorrhage: The emerging revolution. *Nature Clinical Practice Neurology* 3: 256–263. https://doi.org/10.1038/ncpneuro0490.

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.

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.

Also see

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

[ADAPT] **Glossary**