STATA ADAPTIVE DESIGNS: GROUP SEQUENTIAL TRIALS

REFERENCE MANUAL

RELEASE 19

A Stata Press Publication
StataCorp LLC
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845

ISBN-10: 1-59718-417-9
ISBN-13: 978-1-59718-417-5

The suggested citation for this software is

StataCorp. 2025. *Stata 19*. Statistical software. StataCorp LLC.

The suggested citation for this manual is

StataCorp. 2025. *Stata 19 Adaptive Designs: Group Sequential Trials Reference Manual*. College Station, TX: Stata Press.

www.stata.com

Contents

Intro . . . . Introduction to adaptive designs for clinical trials 1

GSD intro . . . . Introduction to group sequential designs 4

gs . . . . Introduction to commands for group sequential design 14

gsbounds . . . . Boundaries for group sequential trials 42

gsdesign . . . . Study design for group sequential trials 73

gsdesign onemean . . . . Group sequential design for a one-sample mean test 100

gsdesign twomeans . . . . Group sequential design for a two-sample means test 121

gsdesign oneproportion . . . . Group sequential design for a one-sample proportion test 143

gsdesign twoproportions . . . . Group sequential design for a two-sample proportions test 165

gsdesign logrank . . . . Group sequential design for a log-rank test 187

gsdesign usermethod . . . . Add your own methods to the gsdesign command 210

Glossary 234

Subject and author index 248

Cross-referencing the documentation

When reading this manual, you will find references to other Stata manuals, for example, **[U] 27 Overview of Stata estimation commands**; **[R] regress**; and **[D] reshape**. The first example is a reference to chapter 27, *Overview of Stata estimation commands*, in the *User's Guide*; the second is a reference to the `regress` entry in the *Base Reference Manual*; and the third is a reference to the `reshape` entry in the *Data Management Reference Manual*.

All the manuals in the Stata Documentation have a shorthand notation:

[GSM]	*Getting Started with Stata for Mac*
[GSU]	*Getting Started with Stata for Unix*
[GSW]	*Getting Started with Stata for Windows*
[U]	*Stata User's Guide*
[R]	*Stata Base Reference Manual*
[ADAPT]	*Stata Adaptive Designs: Group Sequential Trials Reference Manual*
[BAYES]	*Stata Bayesian Analysis Reference Manual*
[BMA]	*Stata Bayesian Model Averaging Reference Manual*
[CAUSAL]	*Stata Causal Inference and Treatment-Effects Estimation Reference Manual*
[CM]	*Stata Choice Models Reference Manual*
[D]	*Stata Data Management Reference Manual*
[DSGE]	*Stata Dynamic Stochastic General Equilibrium Models Reference Manual*
[ERM]	*Stata Extended Regression Models Reference Manual*
[FMM]	*Stata Finite Mixture Models Reference Manual*
[FN]	*Stata Functions Reference Manual*
[G]	*Stata Graphics Reference Manual*
[H2OML]	*Machine Learning in Stata Using H2O: Ensemble Decision Trees Reference Manual*
[IRT]	*Stata Item Response Theory Reference Manual*
[LASSO]	*Stata Lasso Reference Manual*
[XT]	*Stata Longitudinal-Data/Panel-Data Reference Manual*
[META]	*Stata Meta-Analysis Reference Manual*
[ME]	*Stata Multilevel Mixed-Effects Reference Manual*
[MI]	*Stata Multiple-Imputation Reference Manual*
[MV]	*Stata Multivariate Statistics Reference Manual*
[PSS]	*Stata Power, Precision, and Sample-Size Reference Manual*
[P]	*Stata Programming Reference Manual*
[RPT]	*Stata Reporting Reference Manual*
[SP]	*Stata Spatial Autoregressive Models Reference Manual*
[SEM]	*Stata Structural Equation Modeling Reference Manual*
[SVY]	*Stata Survey Data Reference Manual*
[ST]	*Stata Survival Analysis Reference Manual*
[TABLES]	*Stata Customizable Tables and Collected Results Reference Manual*
[TS]	*Stata Time-Series Reference Manual*
[I]	*Stata Index*
[M]	*Mata Reference Manual*

Intro — Introduction to adaptive designs for clinical trials

Description Remarks and examples References Also see

Description

This entry provides a brief introduction to adaptive designs for clinical trials. For a general introduction to group sequential designs and their implementation in Stata, see [ADAPT] **GSD intro** and [ADAPT] **gs**, respectively.

Remarks and examples

Armitage (1993) observes that “classical theory of experimental design deals predominantly with experiments of predetermined size, presumably because the pioneers of the subject, particularly R. A. Fisher, worked in agricultural research, where the outcome of a field trial is not available until long after the experiment has been designed and started.” This type of study, where the target sample size is fixed during the design stage, is known as a fixed-sample design (FSD). In other applications, it is common for data to trickle in, providing researchers the opportunity to conduct interim analyses of a partial dataset. This is especially common in clinical trials—studies examining the effects of treatments on humans—where participants are usually accrued over time.

An alternative to an FSD is an adaptive design, a type of experimental design increasingly popular for clinical trials. The US Food and Drug Administration (2019) describes an adaptive design as “a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial.” By providing a framework to modify aspects of the study design, an adaptive design allows the trial to be adjusted to account for information that was unavailable during the design stage.

Adaptive designs for clinical trials offer several potential advantages over FSDs. Many adaptive designs offer the prospect of increased statistical efficiency, often in the form of a smaller expected sample size than that of an equivalently powered FSD. This can save resources for the sponsor of the trial. Resources can also be saved by employing adaptations that modify the recruitment practices or the desired sample size of an ongoing trial to maximize the probability of identifying a clinically meaningful treatment effect. There is also an ethical argument for some adaptations, particularly those that reduce the number of participants assigned to ineffective treatments. Some adaptations even allow the trial to test additional hypotheses that were not considered during the design stage, such as whether a treatment is particularly effective in some subgroups of the population.

Adaptive designs are not a panacea for all challenges encountered during a clinical trial, and this has led some authors to caution against viewing adaptive designs as a distinct class of clinical trials. Piantadosi (2017, 416), for example, advocates using the term “adaptive design features” to emphasize that adaptations are tools for a trialist, not an alternative to addressing underlying issues in a clinical trial design.

Adaptive designs are not without their drawbacks, often in the form of increased complexity. Sample-size calculations and statistical analysis of adaptive designs are typically more complicated than the equivalent methods for FSDs. The implementation of an adaptive design adds logistical challenges: interim analyses require timely and accurate data to be reported multiple times over the course of the trial, and adaptations to the way participants are assigned to treatment groups add complexity to the recruitment process. Also, even if an adaptive design has a smaller expected sample size than a similarly powered

FSD, the adaptive design may have a larger maximum sample size. This is because the expected sample size is the average sample size if the trial were to be repeated many times, while the maximum sample size is the largest possible sample under the adaptive design.

The most popular forms of adaptive designs for clinical trials fall into several broad categories.

- **Group sequential designs** provide the ability to stop a trial early if an interim analysis of the data provides compelling evidence that a treatment is effective or ineffective. This is one of the most widely used adaptive designs and will be the focus of this manual.
- **Adaptive methods for sample-size modification** allow the desired sample size to be adjusted while the trial is underway. Blinded sample-size reestimation adjusts the sample size based on estimates of nuisance parameters (such as the variance of a normal mean) that have been pooled over all treatment groups, while unblinded sample-size reestimation can use estimates of nuisance parameters from individual treatment groups or even interim estimates of the treatment effect.
- **Adaptive randomization designs** modify the way participants are randomized (allocated) to treatment groups. Covariate-adaptive randomization seeks to reduce differences in the distribution of covariates in the treatment groups by modifying the probability that a participant will be assigned to a treatment group based on covariate data collected from the participant before randomization. Response-adaptive randomization modifies allocation probabilities based on interim estimates of treatment effects and can be used to reduce the number of participants assigned to less effective treatments.
- **Adaptive designs for treatment-arm modification** allow the addition or removal of treatment groups, or arms, during the course of the study. Examples include early-phase dose-finding trials that add or remove arms at different dosage levels, and late-phase multiarm trials that "drop the loser", terminating one arm at a time. Large-scale ongoing adaptive platform trials follow a prespecified master protocol to compare multiple experimental arms against a single treatment arm; new experimental arms are added as new treatments become available, and experimental arms may be terminated based on the results of interim analyses.
- **Adaptive enrichment designs** typically begin by enrolling participants from a diverse population and use interim data about treatment efficacy to restrict subsequent recruitment to targeted population subgroups. This approach is particularly appealing when participant characteristics, such as genetic markers, are believed to play a role in treatment efficacy.

Adaptive design of clinical trials is a topic of active research, and the list above is by no means exhaustive. In what follows, we focus on group sequential designs. For more information about adaptive designs, see Pong and Chow (2010), Chow and Chang (2012), Bhatt and Mehta (2016), Pallmann et al. (2018), and US Food and Drug Administration (2019).

References

Armitage, P. 1993. "Interim analyses in clinical trials". In *Multiple Comparisons, Selection, and Applications in Biometry*, edited by F. M. Hoppe, 391–402. Boca Raton, FL: CRC Press.

Bhatt, D. L., and C. R. Mehta. 2016. Adaptive designs for clinical trials. *New England Journal of Medicine* 375: 65–74. https://doi.org/10.1056/NEJMra1510061.

Choodari-Oskooei, B., D. J. Bratton, and M. K. B. Parmar. 2023. Facilities for optimizing and designing multiarm multistage (MAMS) randomized controlled trials with binary outcomes. *Stata Journal* 23: 774–798.

Chow, S.-C., and M. Chang. 2012. *Adaptive Design Methods in Clinical Trials*. 2nd ed. Boca Raton, FL: CRC Press.

Pallmann, P., A. W. Bedding, B. Choodari-Oskooei, M. Dimairo, L. Flight, L. V. Hampson, J. Holmes, A. P. Mander, L. Odondi, M. R. Sydes, S. S. Villar, J. M. S. Wason, C. J. Weir, G. M. Wheeler, C. Yap, and T. Jaki. 2018. Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medicine* 16: art. 29. https://doi.org/10.1186/s12916-018-1017-7.

Piantadosi, S. 2017. *Clinical Trials: A Methodologic Perspective*. 3rd ed. Hoboken, NJ: Wiley.

Pong, A., and S.-C. Chow, eds. 2010. *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/b10279.

US Food and Drug Administration. 2019. Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. https://www.fda.gov/media/78495/download.

Also see

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

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

[ADAPT] **Glossary**

GSD intro — Introduction to group sequential designs

Description Remarks and examples References Also see

Description

This entry provides a general introduction to group sequential designs (GSDs) and describes relevant statistical terminology. For an introduction to Stata's commands for GSDs, see [ADAPT] **gs**.

Remarks and examples

Remarks are presented under the following headings:

Introduction
FSDs
GSDs
Components of GSD
Origins of GSD
Brief overview of GSD
Graphing group sequential boundaries

Introduction

For a brief introduction to adaptive designs, see [ADAPT] **Intro**. In this section, we describe GSDs for clinical trials in more detail.

Clinical trials are experimental studies in which the investigator assigns treatments to the participants. Each clinical trial begins with a design that determines the number of participants to recruit and how to allocate the participants to the treatments. GSDs are a subset of clinical trial designs that incorporate preplanned analyses of interim data.

In a GSD, the data-collection step is split into multiple predefined stages, and an interim analysis is performed at each stage as the data accumulate [Pocock (1977), O'Brien and Fleming (1979), Lan and DeMets (1983), Jennison and Turnbull (2000), Wassmer and Brannath (2016)]. Each analysis of the data is known as a look. Stopping boundaries are calculated for each look such that analyses from multiple looks are guaranteed not to exceed a predefined overall false-positive error rate, ensuring control of familywise type I error. Unlike fixed-sample designs (FSDs), GSDs can be stopped early in the presence of compelling evidence against or in favor of the null hypothesis.

FSDs

To understand the process of creating and implementing GSDs, it is helpful to begin by considering FSDs. To plan an FSD, the investigator will begin by calculating the required sample size based on several factors, such as the size of a clinically meaningful effect, the desired power and significance level, results of previous studies, and practical considerations like cost and ability to recruit participants.

The next step is to recruit participants to the study. Depending on the scale of the study, recruitment could take place at a single site or at many sites. Recruitment often continues for months or even years. When a participant is recruited to the study, they are assigned, or randomized, to a treatment group. The gold standard of clinical trial design is a randomized controlled trial, where participants are randomly assigned to control or experimental groups, or arms. Trials without a control group are common in early-phase clinical trials designed to explore the appropriate dosage of a therapeutic agent and to investigate how the treatment affects participants. Uncontrolled trials are less common in late-phase clinical trials designed to demonstrate treatment efficacy, though there are circumstances warranting their use (see *Remarks and examples* in [ADAPT] **gsdesign onemean** and in [ADAPT] **gsdesign oneproportion** for examples).

In a classical two-arm randomized controlled trial, one group receives the experimental treatment, while the other group receives a control treatment. If there are no existing treatments that are comparable with the experimental treatment, the control group will typically receive a placebo. When a standard of care exists, there is often an ethical argument against using a placebo. In this case, an active control is used, wherein participants receive the existing standard of care.

After being assigned to a treatment arm, participants are monitored to collect data on the outcome of interest, which is typically referred to as the endpoint. In studies with multiple endpoints, it is common to designate one primary endpoint or to combine multiple endpoints into a single composite endpoint. Depending on the endpoint, the follow-up period may last for years. This is especially common in trials with survival outcomes (also known as time-to-event endpoints). Some participants might leave the study before their primary endpoint data are collected, a phenomenon known as loss to follow-up.

In a large clinical trial, it is not uncommon for several years to elapse before all the endpoint data are collected. If the trial follows an FSD, no analysis of treatment efficacy is conducted until all endpoint data have been obtained. At the end of an FSD, the data are analyzed and the null hypothesis is either rejected or not. In contrast with some other disciplines, in the context of clinical trials, it is common to describe the failure to reject H_0 as “accepting the null hypothesis”. The flowchart in figure 1 details the course of an FSD.

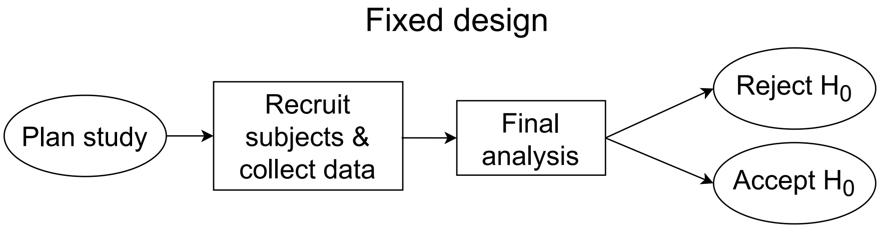


Figure 1. FSD flowchart

GSDs

In the context of a long clinical trial, there is the potential for substantial benefit to both participants and sponsors if a treatment can be declared effective or ineffective before the trial is scheduled to end. GSDs accomplish this by allowing for multiple preplanned analyses of interim trial data while controlling the familywise error rate.

At each interim look, a statistical test is performed, and the test statistic is compared with sets of critical values called stopping boundaries to determine whether H_0 can be rejected (known as efficacy stopping) or accepted (known as futility stopping). If the interim test is inconclusive, the study continues to the next look. At the final look, H_0 must be rejected or accepted.

Planning a clinical trial using a GSD is similar to planning a trial using an FSD, but some additional considerations are required. On the logistical side, preparations must be made to ensure that interim data are of high quality and are quickly available to the data analysis team, often an independent group called a Data Monitoring Committee. On the statistical side, investigators must determine the type of stopping rule to apply (efficacy stopping, futility stopping, or both), the number and spacing of interim analyses, and the boundary-calculation procedure to be used.

Stopping boundaries are calculated, and sample-size calculations that account for the planned interim analyses are performed. The recruitment, randomization, treatment, and follow-up of a GSD are akin to those of an FSD. But instead of waiting to collect all endpoint data before analysis, interim analyses are performed, and the study can be terminated early for efficacy (if H_0 is rejected) or for futility (if H_0 is accepted).

The flowchart in figure 2 details the course of a GSD. Each interim analysis offers the opportunity to terminate the trial to reject H_0 (if efficacy bounds are used) and the opportunity to terminate the trial to accept H_0 (if futility bounds are used). If the interim test is inconclusive, the trial continues to collect more data until the next look. The process continues until an interim analysis determines that the trial should stop or until all possible data are collected and the final analysis is performed.

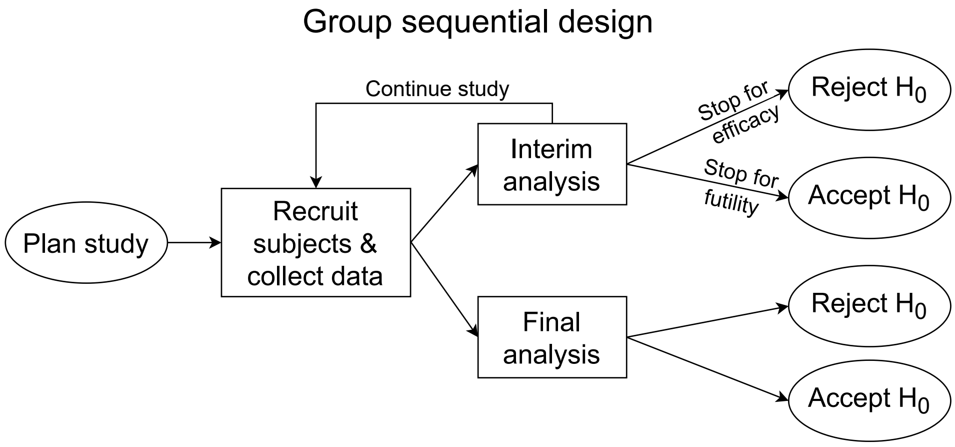


Figure 2. GSD flowchart

Components of GSD

The components of power and sample-size analysis for FSDs are also relevant to GSDs. Please read *Components of PSS analysis* in [PSS-2] **Intro (power)** before reading this section. The key components of GSDs include variations of components of FSDs as well as components that are specific to GSDs. We describe them below.

- **Statistical analysis method:** A clinical trial following a GSD must identify the intended statistical analysis method during the design stage. The type of statistical test to be used dictates the methodology used in calculating the sample sizes at interim analyses (but not the critical values for the stopping boundaries). A GSD is able to provide strong control of familywise type I error when, under the null hypothesis, the sequence of test statistics from interim analyses follows a multivariate normal distribution with a covariance matrix that depends only on the amount of the data analyzed at each interim look. To use a GSD with a test that does not produce a normally distributed test statistic under the null hypothesis, the significance level approach, which uses stopping boundaries based on p-values instead of critical values, may be used (see *Methods and formulas* in [ADAPT] **gsbounds** for details).

- **Significance level, α:** GSDs must account for multiple hypothesis tests being conducted. It is not sufficient to conduct each test at the desired overall significance level α because conducting multiple tests means that the chance of committing a type I error at one or more interim analyses will be greater than the desired α. Instead, the familywise error rate is controlled, ensuring that $\Pr(\text{reject } H_0 \text{ at any look} \mid H_0 \text{ is true}) = \alpha$.

- **Power, $1-\beta$:** The power of a group sequential test is the probability of rejecting a false null hypothesis at any look, or $\Pr(\text{reject } H_0 \text{ at any look} \mid H_0 \text{ is false}) = 1-\beta$. Power is calculated relative to a prespecified effect size, and the smaller the effect size, the larger the sample required to achieve a given power.

- **Accrual and endpoints:** In a GSD, as in most clinical trial designs, participants are generally recruited, or accrued, over time. The outcome of interest is known as the endpoint. GSDs offer the most benefit when the collection of endpoint data is rapid compared with accrual. If the time between randomization and endpoint-data collection is excessively long, there will be less benefit in terminating a trial early because resources already will have been expended to recruit many participants who are still in follow-up but whose endpoints have not yet been collected.

- **Interim looks:** Interim looks, or interim analyses of the data available to date, are the defining feature of GSDs. To conduct a GSD properly, it is necessary to ensure that endpoint data are collected in a timely and reliable manner and provided to the statistical analysis group or Data Monitoring Committee without unblinding individuals who should remain blinded. In the context of a clinical trial, blinding refers to knowledge of which treatment group a participant was assigned to.

- **Stopping rule:** GSDs can allow for efficacy stopping (early rejection of H_0) as well as futility stopping (early acceptance of H_0). During the design stage, a set of critical values known as stopping boundaries is calculated. At each interim analysis, the test statistic is compared with the critical values for that look. If the statistic is more extreme than the efficacy critical value, we say that it has crossed the efficacy boundary and the trial is stopped for treatment efficacy. If the statistic is less extreme than the futility critical value, we say that it has crossed the futility boundary and the trial is stopped for futility. Futility bounds can be either binding or nonbinding. If a study with binding futility bounds is not stopped after crossing the futility bound, it risks overrunning the desired type I error. Nonbinding futility bounds are similar to binding futility bounds, but if a nonbinding futility bound is crossed, investigators have the option of stopping for futility or continuing the trial in the hope that more evidence will accumulate in favor of the experimental treatment, and there is no risk of excessive type I error. The cost for nonbinding futility bounds is a slightly larger sample size than required by binding futility bounds with equivalent type I error and power.

- **Expected sample size:** If a group sequential trial, a clinical trial using a GSD, stops early, it can use a substantially smaller sample size than an equivalently powered FSD. But if all the interim tests are inconclusive, the study will continue to the final look and use the maximum possible sample size, which is always larger than that of an equivalent FSD. The expected sample size of a GSD is the average sample size that would be used if the trial were to be repeated many times. Expected sample size is calculated relative to a given effect size; the expected sample size of a GSD with efficacy stopping will decrease when the effect size is large, reflecting the increased probability of early stopping for efficacy. GSDs with futility stopping will often have a smaller expected sample size than an equivalent FSD when the effect size is 0 because of the ability to accept the null hypothesis and stop the trial early for futility.

- **Boundary-calculation procedure:** Frequently, an investigator will decide which stopping rule to employ (efficacy stopping, futility stopping, or both) and how many interim looks to perform before picking the procedure they will use to calculate the bounds. There are several different formulas available to calculate stopping boundaries, but the most popular ones fall into two broad categories: classical bounds and error-spending bounds. Classical boundary-calculation procedures compute the boundary critical values directly, while error-spending procedures define an error-spending function that partitions the type I error (for efficacy bounds) or type II error (for futility bounds) between the planned looks. For designs with both efficacy and futility stopping, it is not necessary to use the same boundary-calculation method for both efficacy and futility bounds, but classical bounds cannot be combined with error-spending bounds.

  Some boundary-calculation procedures are conservative, which means they 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 FSD. In contrast, anticonservative boundaries 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 FSD. Other boundary-calculation procedures use a parameter to control their shape; depending on the value of the parameter, these bounds can be conservative, anticonservative, or somewhere in the middle.

- **Information:** The amount of information a dataset contains about an unknown parameter is known as the Fisher information. Generally, information is proportional to the sample size, but not always. For example, with time-to-event data, the amount of information is proportional to the number of events observed. When designing a group sequential trial, the timing of the interim looks is specified in terms of the information fraction, the fraction of the maximum possible information to be collected by the study. For example, in a GSD with four equally spaced looks, analyses will occur when 25%, 50%, 75%, and 100% of the data are collected. If a group sequential trial continues to its final look, the maximum amount of information will be collected, which is always greater than the information of an equivalently powered FSD. The ratio of the maximum information required by a GSD to the information of an equivalent FSD is known as the information ratio, and the maximum sample size (or number of events) of a GSD is the product of the information ratio and the sample size of an equivalent FSD.

Origins of GSD

Clinical trials are studies investigating the effects of a treatment on human participants. Large clinical trials, such as those designed to determine the efficacy of an experimental treatment, typically enroll participants over months or years, randomizing some participants to the experimental treatment group and others to a control group. In an FSD, no analysis is conducted until all data are collected.

In the context of a clinical trial, there is an ethical imperative not to expose participants to inferior treatments. GSDs address this ethical consideration by providing a protocol for the interim analyses of clinical trial data. If an interim analysis demonstrates that the new treatment is effective, the trial can stop early, hastening regulatory approval and sparing future participants from being assigned to the control group. If an interim analysis demonstrates that the new treatment is ineffective, the trial can stop early and resources can be allocated to testing more promising treatments.

When done naïvely, conducting multiple analyses at a nominal significance level will inflate type I error. Wassmer and Brannath (2016) note that traditional methods of controlling the familywise error rate, such as the Bonferroni correction, are overly conservative because they do not exploit the covariance structure of test statistics from a sequential analysis. Wallis (1980) recounts the origin of modern sequential analysis theory, which arose not in the context of clinical trials, but as a more efficient way to test weaponry during the Second World War. Abraham Wald, a Hungarian Jewish mathematician who immigrated to the United States and participated in the war effort as a member of the Statistical Research Group at Columbia University, developed the sequential probability ratio test (SPRT) in 1943. (Wald is known for several contributions to statistics, including the eponymous Wald test; see the vignette in the [TS] **varwle** entry for more information about his life.)

The SPRT was so useful to the military that access to Wald's report was restricted to prevent it from falling into enemy hands. Two years later, the restriction was lifted and Wald (1945) published the first public account of the SPRT, which uses fixed stopping rules but does not fix the maximum sample size. Applying the Neyman–Pearson theory of hypothesis testing, practitioners of the SPRT begin by formulating two hypotheses, H_0 and H_a, which are compared using a sequence of likelihood-ratio tests. A continuation interval (a, b) is defined, with critical values a and b chosen so that the probabilities of type I and type II errors are equal to prespecified levels. After each sample is collected, the investigator calculates the likelihood ratio of the two hypotheses; if it falls within the continuation interval, the experiment continues and another sample is taken. If the likelihood ratio lies outside (a, b), the experiment ends and either H_0 or H_a is rejected (depending on whether the likelihood ratio is above or below the continuation interval). In contrast with the prevailing modern interpretation of null-hypothesis significance testing, the SPRT provides a mechanism to reject H_a and accept H_0.

Brief overview of GSD

More recent developments in sequential experimental design have introduced classes of sequential tests with different properties, including a fixed maximum sample size. But the appeal of a controlled framework for accepting H_0 has endured in sequential experimental designs, in no small part because accepting H_0 provides grounds for terminating the experiment due to futility. The term "accept H_0" is widely used in literature about sequential clinical trials, and we will use it (without quotes) in the remainder of this manual to refer to the demonstration of futility in a sequential design. The complement of futility stopping is efficacy stopping, where the experiment is terminated because H_0 can be rejected, even if the maximum sample size has not been reached.

In most clinical trials, it is not feasible to perform statistical analysis after each sample is collected, so fully sequential designs are rare in practice. GSDs address this logistical challenge by scheduling interim analyses after groups of samples have been collected. A major advance in GSDs came when Pocock (1977) established clear guidelines for calculating efficacy stopping boundaries that attain desired levels of type I and type II errors.

Pocock published critical values for a test statistic that follows a standard normal distribution under H_0. For test statistics following other distributions, Pocock recommends using a critical value that has an equivalent significance level to the published z score. For a demonstration, see example 2 in [ADAPT] **gsdesign onemean**.

While Pocock's boundaries use the same critical value at each interim look, O'Brien and Fleming (1979) introduced group sequential boundaries with critical values that are conservative for early looks and less so as more data are collected. O'Brien–Fleming boundaries have proven popular among researchers who are wary of stopping a trial very early for anything less than the strongest evidence, but who appreciate the smaller maximum sample size and final-look critical values compared with those of Pocock's boundaries.

Wang and Tsiatis (1987) developed a one-parameter family of boundaries that includes the Pocock and O'Brien–Fleming boundaries as special cases. Wang–Tsiatis bounds are popular with researchers who want a boundary that is less conservative than O'Brien–Fleming bounds but more conservative than Pocock bounds. However, it is possible to select values of the Wang–Tsiatis parameter that create bounds that are more conservative than the O'Brien–Fleming bound or more anticonservative than the Pocock bound.

Lan and DeMets (1983) introduced the error-spending approach to constructing stopping boundaries. This approach controls the overall probability of type I error by "spending" error probability at interim looks. This allows the number and timing of interim looks to be updated while the trial is in progress.

Lan and DeMets (1983) presented error-spending functions that correspond to boundaries that approximate both Pocock and O'Brien–Fleming bounds. Kim and DeMets (1987) created a useful family of error-spending functions indexed by a power parameter, and Hwang, Shih, and de Cani (1990) introduced another one-parameter family of error-spending functions. While the parameters for Kim–DeMets and Hwang–Shih–de Cani bounds use different scales, both boundary-calculation procedures are quite flexible and can produce bounds that are as conservative or anticonservative as desired.

The process of conducting interim analyses with a GSD is the same regardless of the procedure used to calculate the stopping boundaries. The boundaries comprise a series of critical values, one for each look. At each interim look, the data are analyzed and a test statistic is calculated. If the design includes efficacy bounds, the test statistic is compared with the efficacy critical value, and H_0 is rejected if the statistic is more extreme than the efficacy critical value. If the design includes futility bounds, the test statistic is compared with the futility critical value, and H_0 is accepted if the statistic is less extreme than the futility critical value.

Graphing group sequential boundaries

When comparing different GSDs, it is often helpful to visualize the boundaries of different methods. We begin by presenting a simple GSD using O'Brien–Fleming efficacy boundaries for a two-sided test of means in figure 3 below. Here we plan on conducting up to five analyses: four interim looks and one final analysis. At each interim look, if the test statistic calculated from the available data is within the green continuation region, then the study continues accruing more participants, but if the test statistic is outside the efficacy bounds and in the blue rejection region, then H_0 is rejected and the experiment is stopped early for efficacy. At the fifth and final look, there is no continuation region; if the final test statistic is not outside the efficacy bounds, it will lie in the red acceptance region and H_0 is accepted.

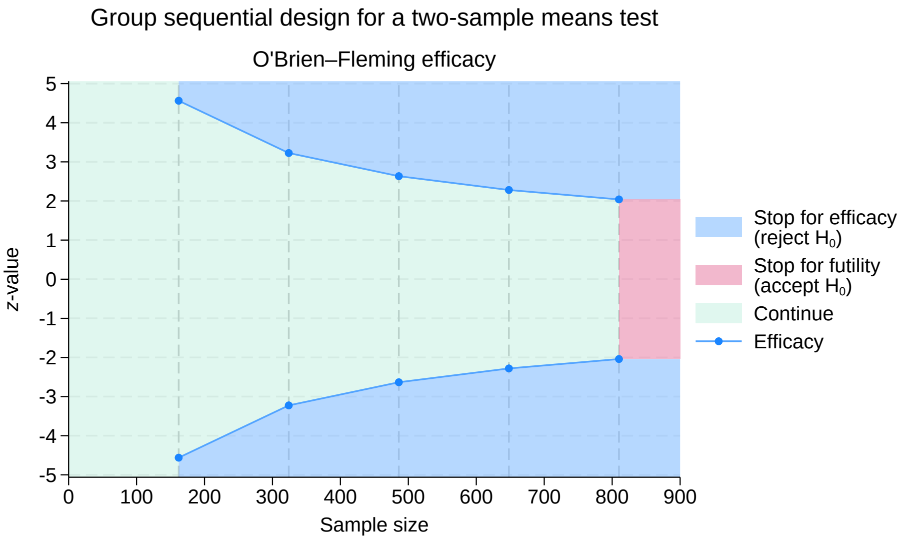


Figure 3. Two-sided O'Brien–Fleming efficacy bounds for a test of the equality of two means

Next we consider a similar scenario that includes futility bounds as well as efficacy bounds. Efficacy bounds separate the rejection region from the continuation region, and futility bounds separate the acceptance region from the continuation region. If the test statistic from an interim analysis falls within the continuation region, then the study proceeds as planned. If it 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. As in the previous example, at the final look, there is no continuation region and H_0 must be accepted or rejected.

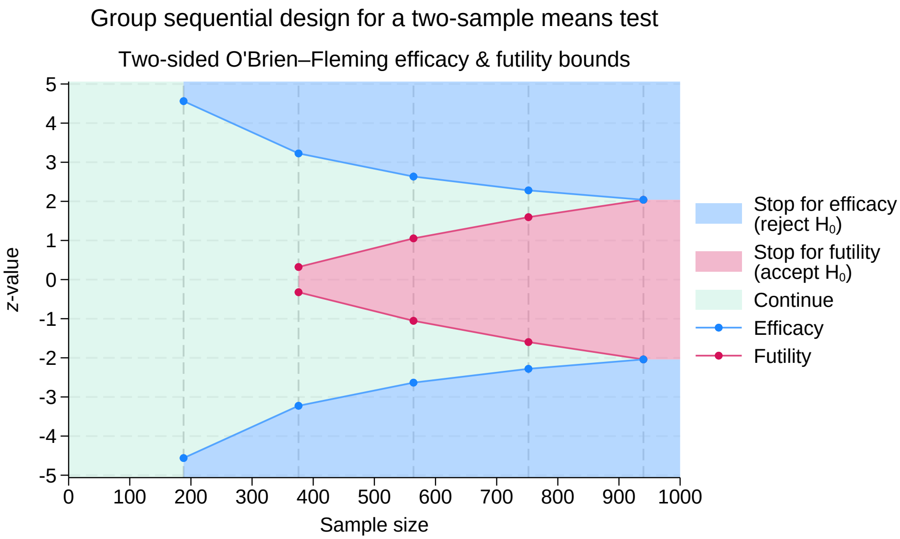


Figure 4. Two-sided O'Brien–Fleming efficacy and futility bounds

In both graphs, the vertical axis is labeled "z-value" because the theory underlying a GSD's ability to control familywise type I error is based on a sequence of test statistics whose marginal distribution under the null hypothesis is normal with a mean of 0 and a variance of 1. For details about how to calculate group sequential boundaries in Stata, including how to incorporate test statistics that are not normally distributed, see [ADAPT] **gsbounds**. To additionally calculate sample sizes for interim analyses, see [ADAPT] **gsdesign**.

References

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.

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.

Wald, A. 1945. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* 16: 117–186. https://doi.org/10.1214/aoms/1177731118.

Wallis, W. A. 1980. The statistical research group, 1942–1945. *Journal of the American Statistical Association* 75: 320–330. https://doi.org/10.2307/2287451.

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] **Intro** — Introduction to adaptive designs for clinical trials

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

[ADAPT] **Glossary**

gs — Introduction to commands for group sequential design

Description Menu Syntax Remarks and examples
Stored results Acknowledgments References Also see

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:

Introduction
Efficacy stopping
Futility stopping
Graphing stopping boundaries
Boundary and sample-size calculations using gsdesign
  One-sample tests
  Two-sample tests
  Survival analysis
  Add your own methods

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

──────────────────────────────────────────────
      │ Info.             Efficacy
 Look │ frac.      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

─────────────┬─────────────────────────────────────
             │  Info.            Efficacy
 Look        │  frac.      Lower     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)

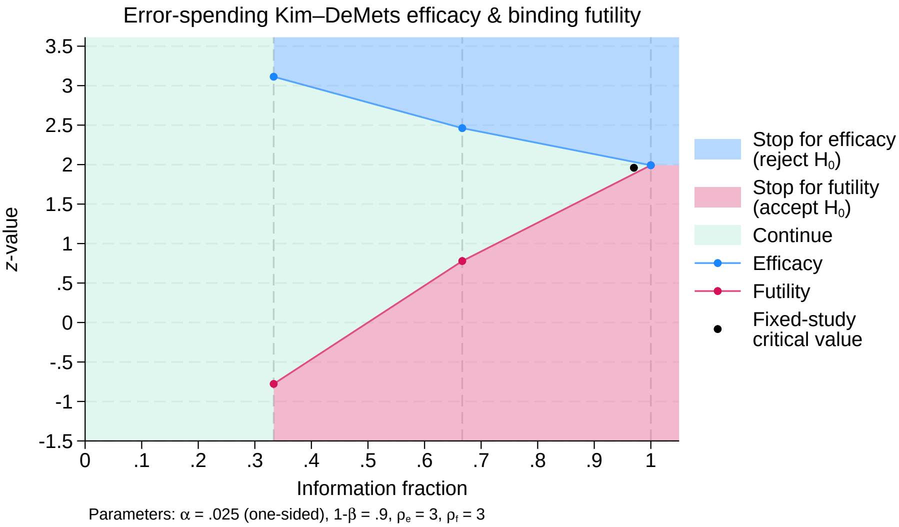


Figure 1. One-sided efficacy and futility bounds

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.
```

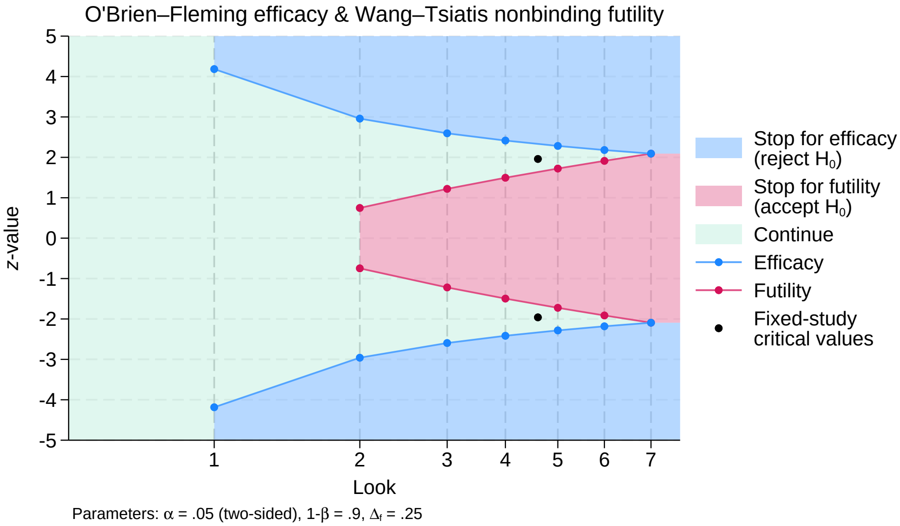


Figure 2. Two-sided efficacy and futility bounds

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

─────────────────────────────────────────────────
      │ Info.        Efficacy       │ Sample size
 Look │ frac.      Upper   p-value  │           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
------+-------------------------------------------------+------------
      | Info.        Efficacy             Futility      | Sample size
 Look | frac.      Upper   p-value      Lower   p-value |           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.
```

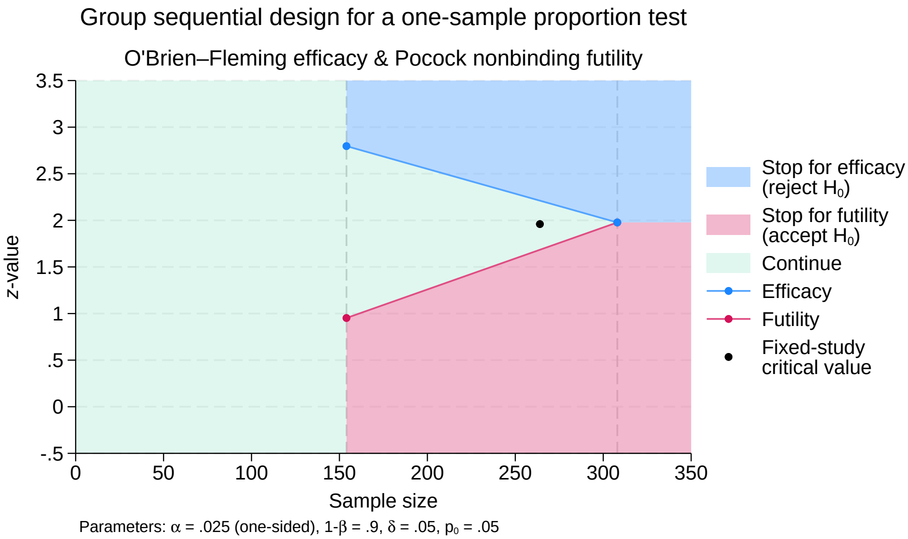


Figure 3. One-sided test of one proportion with efficacy and futility bounds

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

  ─────────────────────────────────────────────────────────
       │ Info.        Efficacy            Futility
  Look │ frac.     Upper   p-value     Lower   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.

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

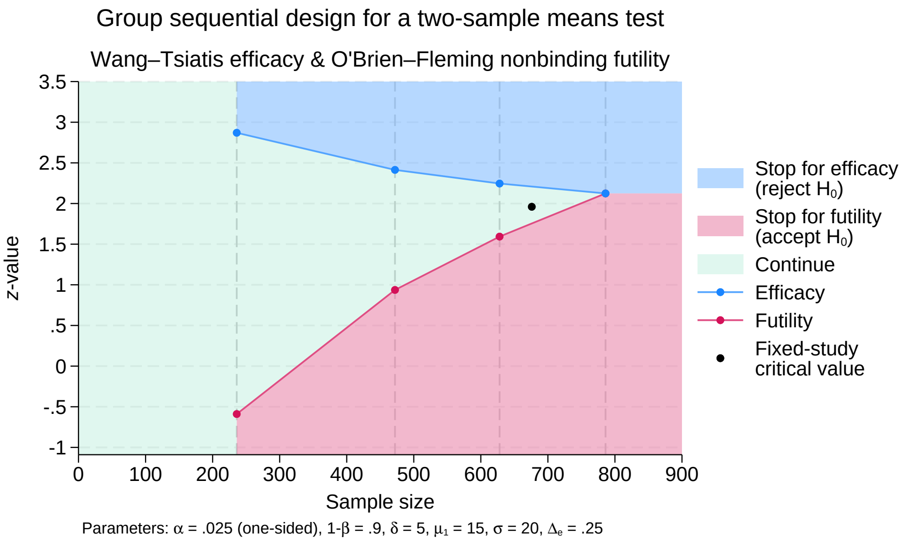


Figure 4. One-sided test of the equality of two means with efficacy and futility bounds

`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**.

◁

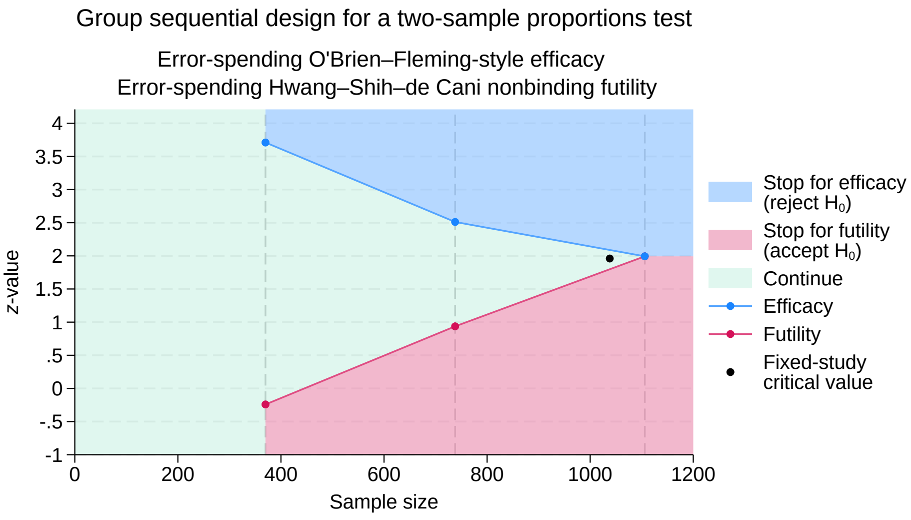
▷ Example 10: Boundary and sample-size calculations for a two-sample proportions test

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
------------------------------------
```



Parameters: α = .025 (one-sided), 1-β = .9, δ = .1, p_1 = .5, γ_f = -2

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
──────┬─────────────────────────────────────────────────┬────────────
      │ Info.        Efficacy             Futility      │    Events
 Look │ frac.      Lower   p-value     Upper   p-value │           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.
```

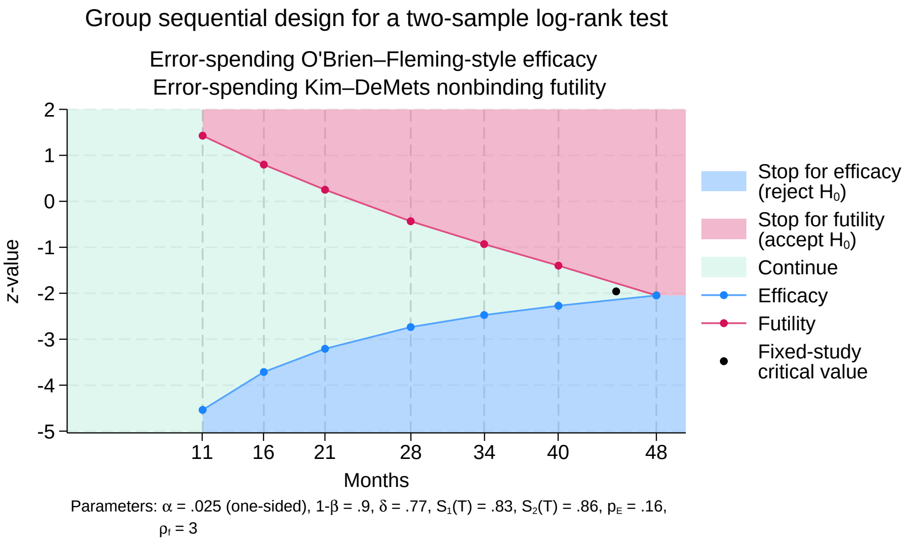


Figure 6. One-sided log-rank test with efficacy and futility bounds

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

─────────────────────────────────────────────────────────
      │ Info.            Efficacy               │ Sample size
 Look │ frac.       Lower      Upper    p-value │           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

     |  Info.           Efficacy                      Futility
Look |  frac.     Lower     Upper   p-value     Lower     Upper   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.

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

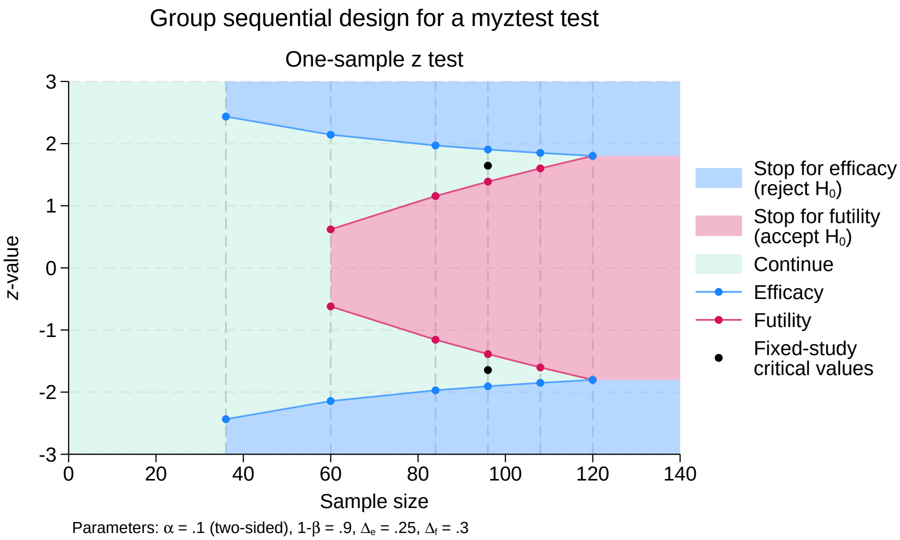


Figure 7. User-written one-sample z test with efficacy and futility bounds

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**

gsbounds — Boundaries for group sequential trials

Description Quick start Menu Syntax Options
Remarks and examples Stored results Methods and formulas References Also see

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 $\mathbf{g}\mathbf{H}^{-1}\mathbf{g}' <$ 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:

Introduction
Examples
  Efficacy stopping
  Efficacy and futility stopping
  Nonbinding futility bounds
  One-sided tests
  Error-spending bounds
  Unevenly spaced looks
  Futility-only stopping

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

            Info.              Efficacy
  Look      frac.       Lower     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)
```

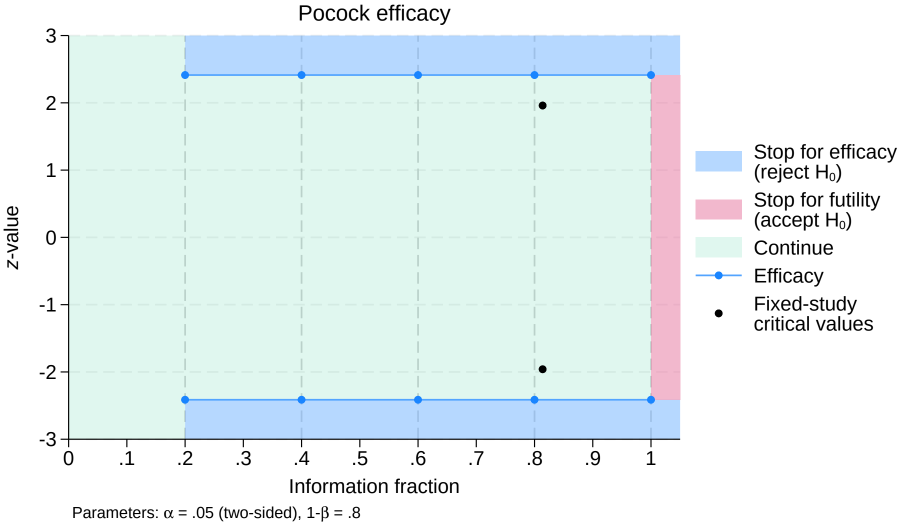


Figure 1. Pocock efficacy bounds

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

─────────────────────────────────────────────────────────
      │  Info.              Efficacy
 Look │  frac.       Lower      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.
```

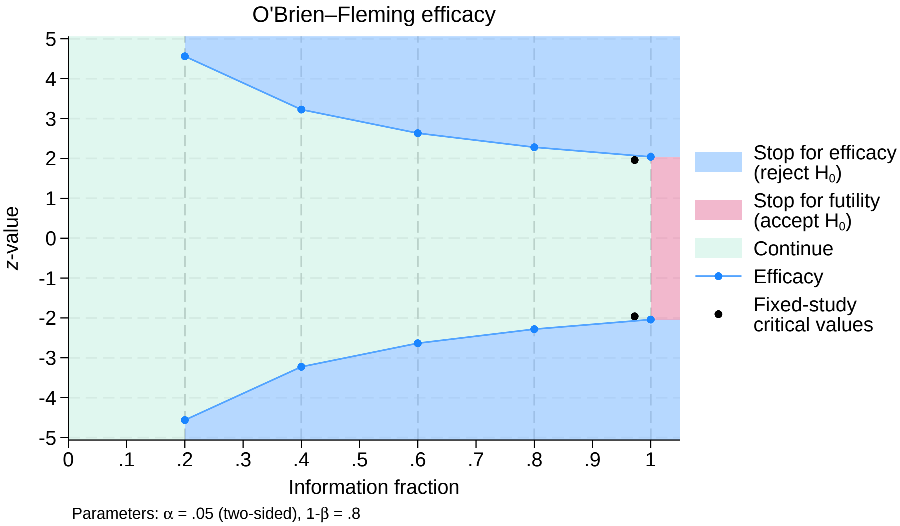


Figure 2. O'Brien–Fleming efficacy bounds

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 $|z_5| \geq 2.04$.

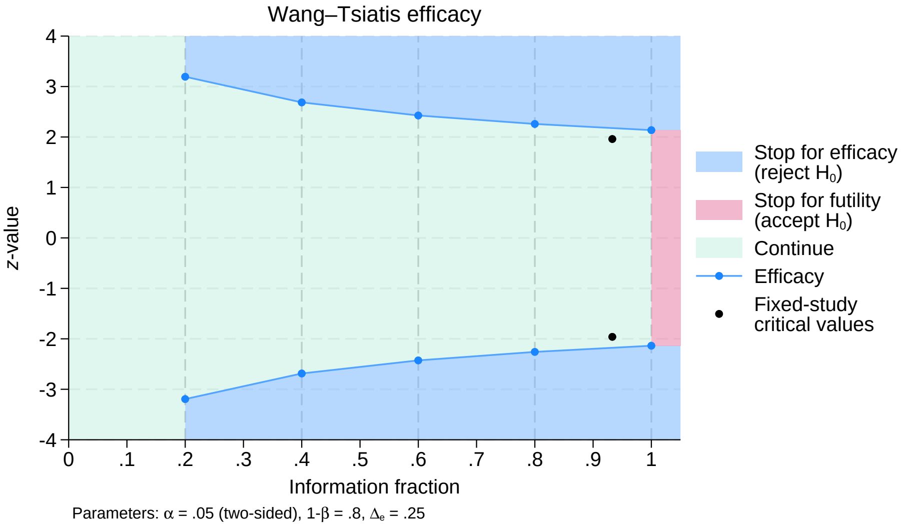
◁

▷ 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

─────────────┬────────────────────────────────────────
             │  Info.                Efficacy
  Look       │  frac.        Lower      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.
```



Parameters: α = .05 (two-sided), 1-β = .8, Δ_e = .25

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.
```

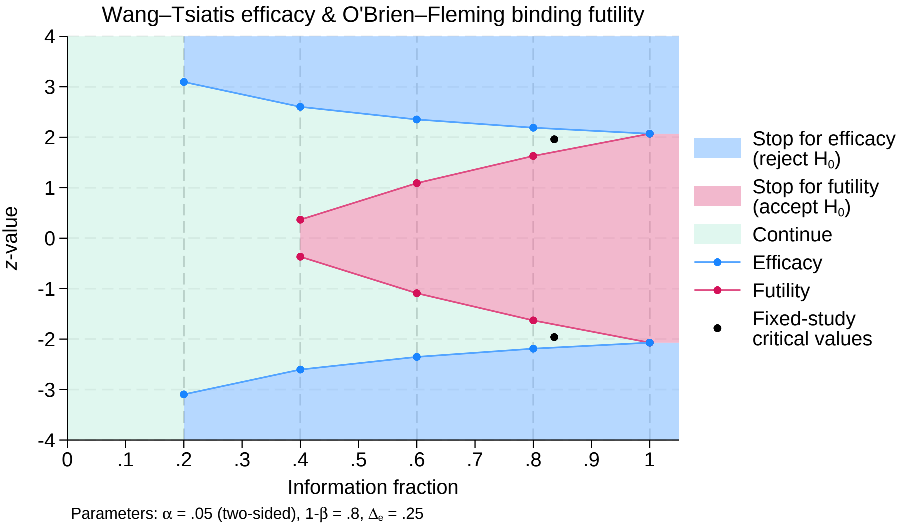


Figure 4. Wang–Tsiatis efficacy and futility bounds

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

--------------------------------------------------------------------------
       |  Info.            Efficacy                      Futility
  Look |  frac.     Lower     Upper   p-value     Lower     Upper   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

-------------------------------------
        |  Info.          Efficacy
   Look |  frac.       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.
```

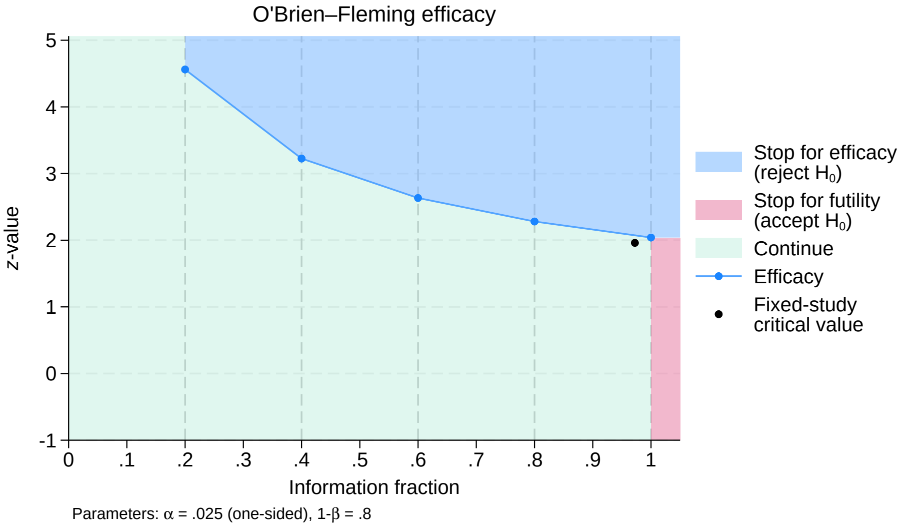


Figure 6. One-sided O'Brien–Fleming efficacy bounds

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
----------------------------------
      |  Info.       Efficacy
 Look |  frac.      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.
```

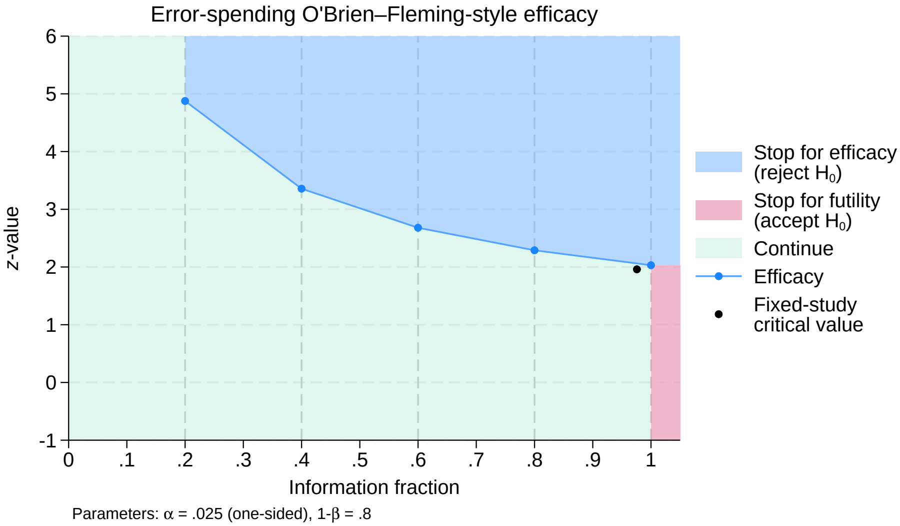


Figure 7. One-sided error-spending O'Brien–Fleming-style efficacy bounds

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.
```

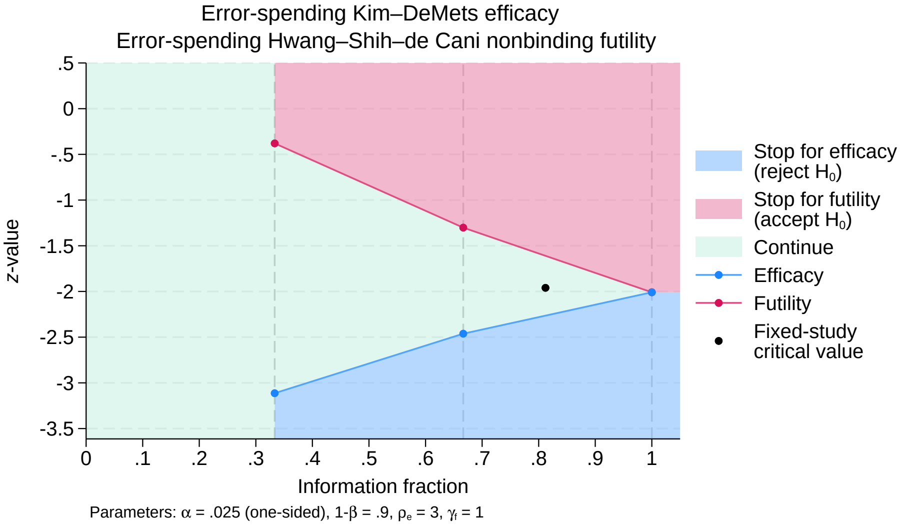


Figure 8. One-sided lower error-spending efficacy and futility bounds

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.
```

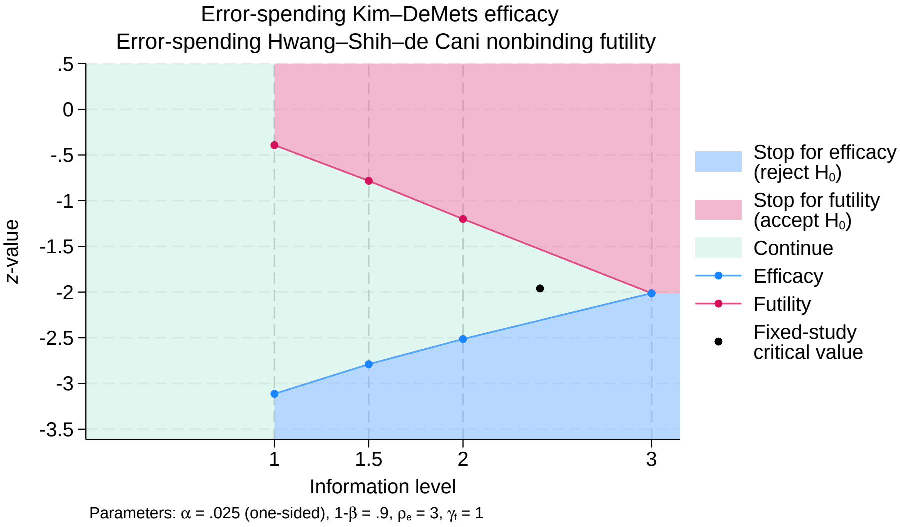


Figure 9. One-sided lower error-spending efficacy and futility bounds with unevenly spaced looks

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
─────────────────────────────────────────────────────────────
      │ Info.        Efficacy            Futility
 Look │ frac.     Upper   p-value     Lower    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.
```

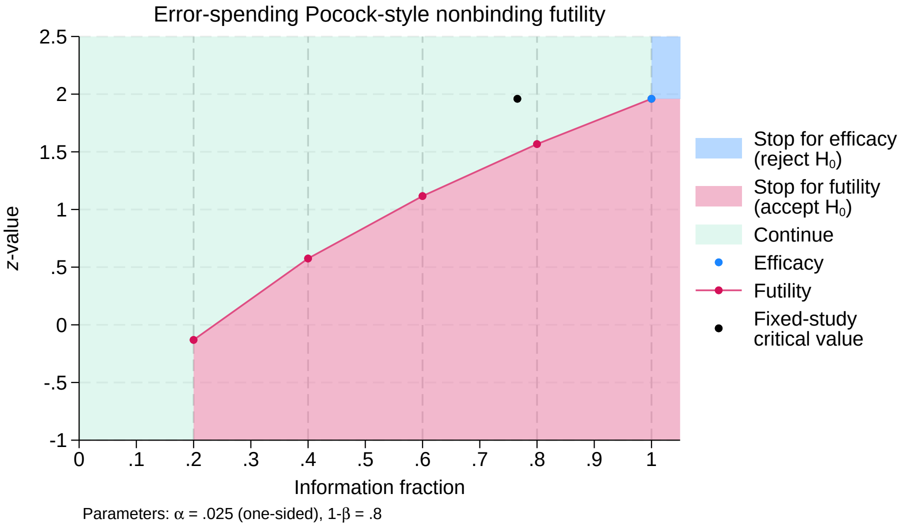


Figure 10. Error-spending Pocock-style nonbinding futility bound

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
Classical (Wang–Tsiatis) bounds
Error-spending bounds
Significance level approach

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 $(n_1, \dots, n_K)$ 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 $(1, \dots, K)$, 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 $(1, \dots, K)$. 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 $(Z_1, \dots, Z_K)$ are not independent. Jennison and Turnbull (2000, 49) show that $(Z_1, \dots, Z_K)$ is multivariate normal with

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

When $(Z_1, \dots, Z_K)$ follow this distribution, the score statistics $(S_1, \dots, S_K)$ that correspond to these z statistics are said to have the property of "independent increments". For any k in $(1, \dots, K)$, 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 $S_1, (S_2 - S_1), \dots, (S_K - S_{K-1})$ 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 $(e_1, \dots, e_K)$ and critical values for futility stopping as $(f_1, \dots, f_K)$. At interim look $k < K$, if test statistic $Z_k \ge e_k$, the trial is stopped for efficacy; if $Z_k < f_k$, the trial is stopped for futility; and if $f_k \le Z_k < e_k$, 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

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

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

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

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 $(f_1,\dots,f_{K-1})$ with $-\infty$ and let $f_K = e_K$ in the calculations above. For trials with nonbinding futility bounds, replace $(f_1,\dots,f_{K-1})$ with $-\infty$ in (2) but not in (3). For trials with futility stopping only, replace $(e_1,\dots,e_{K-1})$ 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 $(f_1,\dots,f_K)$ and upper limit $(e_1,\dots,e_K)$. Two-sided tests require additional integration from $(-e_1,\dots,-e_K)$ to $(-f_1,\dots,-f_K)$. 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 $(\alpha_1, \dots, \alpha_K)$ by setting $\alpha_1 = \alpha^*(\mathcal{I}_1)$ and $\alpha_k = \alpha^*(\mathcal{I}_k) - \alpha^*(\mathcal{I}_{k-1})$ for k in $(2, \dots, K)$.

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 $(2, \dots, K)$.

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 $(Z_1, \dots, Z_K)$ 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

gsdesign — Study design for group sequential trials

Description | Quick start | Menu | Syntax | Options
Remarks and examples | Stored results | Methods and formulas | References | Also see

Description

gsdesign computes stopping boundaries and sample sizes for interim analyses of clinical trials using group sequential designs (GSDs). Stopping can be for efficacy, futility, or both. gsdesign can be used with sample-size calculations from a variety of [PSS-2] **power** methods, including user-defined methods. For stopping boundary calculations without sample sizes, see [ADAPT] **gsbounds**. For a software-free introduction to GSDs, see [ADAPT] **GSD intro**; for an introduction to Stata's gs suite of commands, see [ADAPT] **gs**.

Quick start

Sample sizes and stopping boundaries for a two-sided test of two sample means, with $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$ and a shared standard deviation of 9, with default power of 0.8 to detect the difference between control-group mean $m_1 = 8$ and experimental-group mean $m_2 = 12$ at default overall significance level $\alpha = 0.05$, using default group sequential specifications of O'Brien–Fleming efficacy boundaries with two analyses (one interim, one final)

```
gsdesign twomeans 8 12, sd(9)
```

Same as above, but with an overall significance level of 0.01 and using an O'Brien–Fleming design with three looks to calculate both efficacy and nonbinding futility boundaries

```
gsdesign twomeans 8 12, sd(9) alpha(0.01) efficacy(obfleming)   ///
      futility(obfleming) nlooks(3)
```

Same as above, but use Kim–DeMets boundaries with parameters $\rho_e = 4$ and $\rho_f = 2.5$, and assign twice as many participants to the experimental arm as to the control arm

```
gsdesign twomeans 8 12, sd(9) nratio(2) alpha(0.01)        ///
      efficacy(kdemets(4)) futility(kdemets(2.5)) nlooks(3)
```

Sample size and stopping boundaries for one-sample proportion test of $H_0: \pi = 0.2$ versus $H_a: \pi \neq 0.2$ with power of 0.9 to detect the difference between null proportion $p_0 = 0.2$ and alternative proportion $p_a = 0.3$ at overall significance level $\alpha = 0.1$, using Wang–Tsiatis efficacy boundaries with eight analyses and efficacy parameter $\Delta_e = 0.25$

```
gsdesign oneproportion 0.2 0.3, alpha(0.1) power(0.9)        ///
      efficacy(wtsiatis(0.25)) nlooks(8)
```

Same as above, but report fractional sample sizes and graph the boundaries without shading

```
gsdesign oneproportion 0.2 0.3, alpha(0.1) nfractional power(0.9) ///
      efficacy(wtsiatis(0.25)) nlooks(8) graphbounds(noshade)
```

Sample size and number of events for the log-rank test of $H_0: HR = 1$ versus $H_a: HR < 1$ with default significance level $\alpha = 0.05$ and power of 0.8 to detect the difference between a control-group survival probability of 0.3 and an experimental-group survival probability of 0.5, using error-spending O'Brien–Fleming-style efficacy boundaries with five analyses

```
gsdesign logrank 0.3 0.5, onesided efficacy(errobfleming) nlooks(5)
```

Same as above, but time the looks to occur with 40%, 60%, 80%, 90%, and 100% of the data, adjust the sample size for 5% withdrawal, and graph the boundaries

```
    gsdesign logrank 0.3 0.5, wdprob(0.05) onesided          ///
        efficacy(errobfleming) information(0.4 0.6 0.8 0.9 1)  ///
        graphbounds
```

Menu

Statistics > Power, precision, and sample size

Syntax

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

where *method* ... refers to a `power` *method* that is used for sample-size calculation, *designopts* are options controlling the sample-size calculation, and *boundopts* are options controlling the calculation of the stopping boundaries.

method	Description
`onemean`	GSD for one-sample mean test
`twomeans`	GSD for two-sample means test
`oneproportion`	GSD for one-sample proportion test
`twoproportions`	GSD for two-sample proportions test
`logrank`	GSD for a log-rank test
usermethod	user-defined sample-size calculation

`gsdesign` supports the above methods when they are used to calculate sample size with simple random sampling. To use an unsupported method, specify option `methodok`; see *designopts* table below.

designopts	Description
Main	
methodopts	method-specific options
`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)`
`onesided`	request a one-sided test; default is two-sided
`nfractional`	report fractional sample size
`force`	allow calculation with unsupported *methodopts*
`methodok`	allow calculation with unsupported *method*
`poweriteration(`*powiteropts*`)`	iteration options for the calculation of fixed-study sample size; not available with *method* `logrank`; seldom used

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

`force`, `methodok`, and `poweriteration()` do not appear in the dialog box.

methodopts	[ADAPT] entry
onemeanopts	[ADAPT] **gsdesign onemean**
twomeansopts	[ADAPT] **gsdesign twomeans**
onepropopts	[ADAPT] **gsdesign oneproportion**
twopropopts	[ADAPT] **gsdesign twoproportions**
logrankopts	[ADAPT] **gsdesign logrank**
usermethodopts	[ADAPT] ***gsdesign usermethod***

powiteropts	Description
`init(#)`	initial value for fixed-study sample size
`iterate(#)`	maximum number of iterations; default is `iterate(500)`
`tolerance(#)`	parameter tolerance; default is `tolerance(1e-12)`
`ftolerance(#)`	function tolerance; default is `ftolerance(1e-12)`

boundopts	Description
Bounds	
`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(#`[`, equal`]`)`	total number of analyses (`nlooks()` − 1 interim analyses and one final analysis); use `equal` to enforce equal information increments; if neither `nlooks()` nor `information()` is specified, the default is `nlooks(2)`
`information(`*numlist*`)`	sequence of information levels for analyses; default is evenly spaced
`nopvalues`	suppress *p*-values
Graph	
`graphbounds`[`(`*graphopts*`)`]	graph boundaries
`matlistopts(`*general_options*`)`	control the display of boundaries and sample size; seldom used
optimopts	optimization options for boundary calculations; seldom used

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

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

graphopts	Description
xdimsampsize	label the x axis with the sample size collected (default)
xdiminformation	label the x axis with the information fraction; 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

`alpha(#)` sets the overall significance level, which is the familywise type I error rate for all analyses (interim and final). `alpha()` must be in $(0, 0.5)$. The default is `alpha(0.05)`.

`power(#)` sets the overall power for all analyses. `power()` must be in $(0.5, 1)$. 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. `beta()` must be in $(0, 0.5)$. 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.

`onesided` requests a study design for a one-sided test. The direction of the test is inferred from the effect size.

`nfractional` specifies that fractional sample sizes be reported.

Bounds

`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(#` [`, equal`]`)` 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)`.

`equal` indicates that equal information increments be enforced, which is to say that the same number of new observations will be collected at each look. The default behavior is to start by dividing information evenly among looks, then proceed by rounding up to a whole number of observations at each look. This can cause slight differences in the information collected at each look.

`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.

Graph

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

graphopts are the following:

xdimsampsize labels the x axis with the sample size collected (the default).

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

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 gsdesign but are not shown in the dialog box:

force indicates that gsdesign should allow unsupported method options, such as options specifying a finite population correction or a cluster randomized design. Even with option force, the method options specified must be compatible with sample-size determination, not effect size or power calculation. In addition, *numlist*s are not supported in method options or in arguments as they are with power, even when force is specified.

methodok indicates that gsdesign should allow unsupported methods. Option methodok is not required to run gsdesign with user-defined methods, but it is required to use power methods other than those described in *method*. Option methodok implies option force.

`poweriteration(`*powiteropts*`)` controls the iterative algorithm used to calculate the fixed-study sample size. This is seldom used.

powiteropts are the following:

`init(#)` specifies an initial value for the sample size when iteration is used to compute the fixed-study sample size. The default is to use a closed-form normal approximation to compute an initial sample size.

`iterate(#)` specifies the maximum number of iterations for the Newton method during calculation of the fixed-study sample size. The default is `iterate(500)`.

`tolerance(#)` specifies the tolerance used to determine whether successive parameter estimates have converged when calculating the fixed-study sample size. The default is `tolerance(1e-12)`. See *Convergence criteria* in [M-5] **solvenl()** for details.

`ftolerance(#)` specifies the tolerance used when calculating the fixed-study sample size to determine whether the proposed solution of a nonlinear equation is sufficiently close to 0 based on the squared Euclidean distance. The default is `ftolerance(1e-12)`. See *Convergence criteria* in [M-5] **solvenl()** for details.

`matlistopts(`*general_options*`)` affects the display of the matrix of boundaries and sample sizes. *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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**. 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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

`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 $\mathbf{g}\mathbf{H}^{-1}\mathbf{g}' <$ `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:

Introduction
Examples
    Design for GSD with tests of two means
    Background on the BHAT study
    Design for GSD with survival analysis

This entry describes the `gsdesign` command and the methodology for calculating stopping boundaries and sample sizes for group sequential designs, or 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**; to calculate stopping boundaries without sample sizes, see [ADAPT] **gsbounds**; and to calculate sample sizes for fixed study designs, see [PSS-2] **power**.

Introduction

Clinical trials are studies investigating the effects of a treatment on human participants, and 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.

When done naïvely, conducting multiple analyses at a nominal significance level will inflate type I error. Group sequential experimental designs provide a protocol for the interim analysis of clinical trial data and a framework in which the trial can be stopped early for efficacy or futility while maintaining control of familywise type I and type II errors.

A GSD lays out a sequence of looks, or analyses of the clinical trial data. Interim analyses, which take place before the trial is scheduled to end, provide the ability to terminate the trial early if the result at the interim look is sufficiently unambiguous. Efficacy stopping occurs when the null hypothesis, H_0, is rejected at an interim look and the clinical trial is terminated early due to treatment efficacy. The complement to efficacy stopping is futility stopping, in which the trial is terminated because H_0 has been accepted during an interim look. The concept of accepting the null hypothesis runs counter to the prevailing modern interpretation of null hypothesis significance testing, but accepting H_0 has a long history in the context of sequential trials and is commonly performed in the literature about sequential clinical trials. See *Origins of GSD* in [ADAPT] **GSD intro** for a history of GSDs.

The decision to terminate a clinical trial is frequently made by an independent monitoring group, often called a Data Monitoring Committee. The committee may decide to terminate the trial early because of demonstrated treatment efficacy or futility at an interim analysis. The Data Monitoring Committee can also stop a clinical trial for reasons such as safety and the prevalence of adverse events, which are harmful side effects of the treatment and negative medical outcomes not associated with an underlying disease. When determining whether to terminate a trial because of efficacy or futility, the committee can compare the test statistic from the interim analysis against the critical values of the efficacy or futility bounds. Test statistics with asymptotically standard normal distributions under H_0 can be compared directly with the boundary critical values, and statistics that follow other distributions under H_0 may be evaluated using the significance level approach.

The critical values of the group sequential efficacy and futility bounds depend on several factors: the overall power $(1-\beta)$ and significance level (α) of the design, the type of boundary (`gsdesign` supports seven types of boundaries), whether the test has a one- or two-sided alternative hypothesis, and the information fraction at which the analyses occur. Technically, the information fraction is the proportion of the maximum possible Fisher information that has been collected about the parameter being estimated as part of the test, but this definition is too abstract to be useful. In most cases, the information fraction is the proportion of the maximum sample size that has been collected. For survival data, the information fraction is the proportion of the total number of events (failures) that have been observed, not the total number of participants. To calculate the maximum sample size of a GSD, `gsdesign` scales up the sample size of an equivalently powered fixed-sample design by a factor known as the information ratio.

Examples

Design for GSD with tests of two means

▷ Example 1: Pocock efficacy bounds for a test of two sample means

Jennison and Turnbull (2000, 27) demonstrate the use of Pocock efficacy bounds by considering a test of two means: μ_1 and μ_2. The null hypothesis is $H_0: \mu_1 = \mu_2$, and the two-sided alternative hypothesis is $H_a: \mu_1 \neq \mu_2$. They assume a known standard deviation of 2 for both groups and desire a test with 90% power to detect a difference in means of one unit, while maintaining an overall significance level of $\alpha = 0.05$ over five evenly spaced looks.

Given these specifications, we use `gsdesign twomeans` with a control group mean, m_1, of 0 and a difference in means of 1, specified with the `diff(1)` option. The `efficacy(pocock)` and `nlooks(5)` options request the efficacy boundaries and sample size for a Pocock design with five evenly spaced looks. `alpha()` is omitted because it is left at its default value of 0.05, and `beta()` is omitted because `power()`, defined as $(1-\beta)$, is specified instead. The `graphbounds` option instructs Stata to draw a graph of the boundaries and sample size at each look. The `sd()` option specifies the shared standard deviation of both groups, and the `knownsd` option indicates that the population standard deviation is known for both control and treatment groups.

```
. gsdesign twomeans 0, diff(1) sd(2) knownsds power(0.9) efficacy(pocock)
> nlooks(5) graphbounds

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

Efficacy: Pocock

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
      delta = 1.0000
         m1 = 0.0000
         m2 = 1.0000
       diff = 1.0000
         sd = 2.0000

Expected sample size:
         H0 = 199.00
         Ha = 115.43

Info. ratio = 1.2066
    N fixed =    170
      N max =    204
     N1 max =    102
     N2 max =    102

Fixed-study crit. values = ±1.9600

Critical values, p-values, and sample sizes for a group sequential design
------+--------------------------------------+------------------------------
      | Info.             Efficacy           |         Sample size
 Look | frac.      Lower     Upper   p-value |       N1        N2         N
------+--------------------------------------+------------------------------
    1 |  0.20   -2.4132    2.4132    0.0158 |       21        21        42
    2 |  0.40   -2.4132    2.4132    0.0158 |       41        41        82
    3 |  0.60   -2.4132    2.4132    0.0158 |       61        61       122
    4 |  0.80   -2.4132    2.4132    0.0158 |       82        82       164
    5 |  1.00   -2.4132    2.4132    0.0158 |      102       102       204
------+--------------------------------------+------------------------------

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

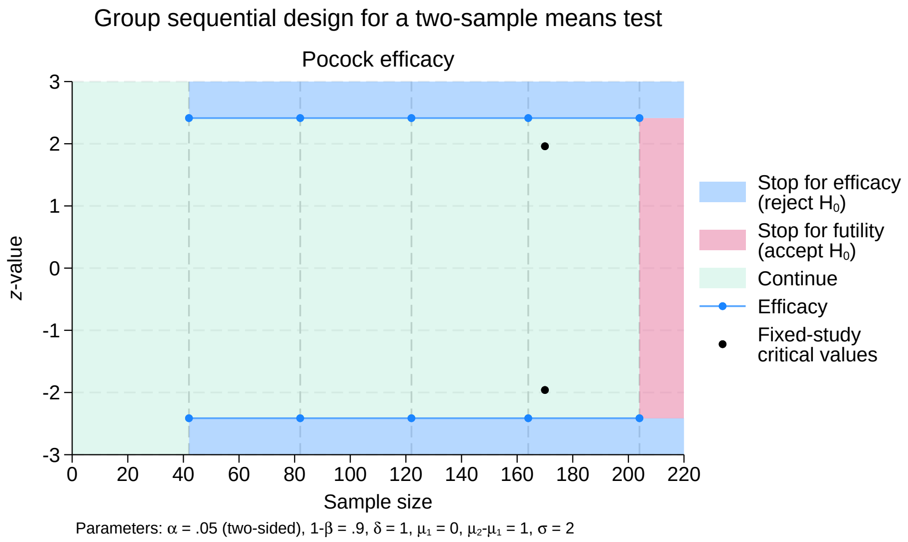


Figure 1. Pocock efficacy bounds for a test of the equality of two means

According to this design, the first look will occur when results have been collected from 21 participants in the control group and 21 participants in the experimental group. A z test of the two means will be conducted, and if the z statistic from that test, z_1, lies in the rejection region ($z_1 \geq 2.413$ or $z_1 \leq -2.413$), then H_0 will be rejected and the trial will end due to treatment efficacy. The efficacy boundary separates the rejection region from the continuation region; if $|z_1| < 2.413$, the test statistic lies within the continuation region and the trial will continue to the second look.

At each successive look, the same procedure is repeated. A defining characteristic of Pocock efficacy bounds is that the same critical value is used at all looks, so at each look the test statistic is compared with ± 2.413. At the fifth and final look, there is no continuation region: if $|z_5| \geq 2.413$, then the null hypothesis is rejected, and if $|z_5| < 2.413$, then the null hypothesis is accepted.

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 sample size. We progress from left to right in the graph as samples are collected during the course of the trial. The efficacy bounds, which separate the continuation and rejection regions, are drawn in blue and marked with a dot at each look. Before the first look (that is, when fewer than 42 samples have been collected), 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 continuation and rejection 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 final look, there is no continuation region; it has been replaced by the acceptance region because the trial cannot be continued beyond the fifth look.

To facilitate comparison with a fixed-sample study design, `gsdesign` displays the estimated sample size and critical values for a fixed study along with the information ratio, the ratio of the maximum sample size from a GSD to the sample size of a fixed design. The Pocock design allows the trial to end after collecting data from as few as 42 participants, but if the trial continues to completion, it will require 20% more participants to attain the same power and significance level as a fixed-sample trial.

When comparing the efficiency of a GSD versus a fixed-sample design, it is useful to examine the expected sample size of the GSD. The expected sample size, which is calculated relative to a given effect size, is the average sample size that a group sequential trial would need if the experiment were to be repeated many times. In the output above, we see that the expected sample size under H_0 is 199. This means that if the true difference between group means is 0 and the trial is repeated many times, the average sample size will be 199. The expected sample size under H_a of 115.43 means that if the true difference between group means is 1, the average sample size over repeated experiments will be 115.43, a substantial savings over the 170 subjects required by the fixed-sample design.

When designing this study, Jennison and Turnbull (2000) reported the maximum sample size as 210 participants, slightly more than the 204 calculated by `gsdesign`. The difference is due to the fact that Jennison and Turnbull forced the spacing of the looks to be exactly equal by requiring each arm of the study to collect data from 21 new participants between each look. By default, `gsdesign` begins by dividing information evenly among looks, and then `gsdesign` rounds the sample sizes up to whole numbers (which can cause slight differences in the spacing between looks). To match the calculation of Jennison and Turnbull (2000), we add the `equal` suboption in the `nlooks()` option.

```
. gsdesign twomeans 0, diff(1) sd(2) knownsds power(0.9) efficacy(pocock)
> nlooks(5, equal)

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

Efficacy: Pocock

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
      delta = 1.0000
         m1 = 0.0000
         m2 = 1.0000
       diff = 1.0000
         sd = 2.0000

Expected sample size:
         H0 = 204.80
         Ha = 116.94

Info. ratio = 1.2066
    N fixed =    170
      N max =    210
     N1 max =    105
     N2 max =    105

Fixed-study crit. values = ±1.9600

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

------+--------------------------------------+-----------------------------
      | Info.             Efficacy           |         Sample size
 Look | frac.      Lower     Upper   p-value |       N1        N2         N
------+--------------------------------------+-----------------------------
    1 |  0.20    -2.4132    2.4132    0.0158 |       21        21        42
    2 |  0.40    -2.4132    2.4132    0.0158 |       42        42        84
    3 |  0.60    -2.4132    2.4132    0.0158 |       63        63       126
    4 |  0.80    -2.4132    2.4132    0.0158 |       84        84       168
    5 |  1.00    -2.4132    2.4132    0.0158 |      105       105       210
------+--------------------------------------+-----------------------------

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

If we enforce equal information increments, we arrive at a maximum sample size of 210. The increased sample size causes a slight increase in attained power, stored as `r(power_a)`.

```
. display "Power attained at final analysis: " r(power_a) * 100
Power attained at final analysis: 91.020745
```

We see that the additional observations yield an attained power of 91%. To understand why the information increments were not exactly equal in the original design, it is informative to view the fractional sample-size calculations by specifying the nfractional option.

```
. gsdesign twomeans 0, diff(1) sd(2) knownsds nfractional power(0.9)
> efficacy(pocock) nlooks(5)

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

Efficacy: Pocock

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
      delta = 1.0000
         m1 = 0.0000
         m2 = 1.0000
       diff = 1.0000
         sd = 2.0000

Expected sample size:
         H0 = 197.83
         Ha = 115.15

Info. ratio = 1.2066
    N fixed = 168.12
      N max = 202.85
     N1 max = 101.43
     N2 max = 101.43

Fixed-study crit. values = ±1.9600

Critical values, p-values, and sample sizes for a group sequential design
---------------------------------------------------------------------------
      |  Info.             Efficacy          |          Sample size
 Look |  frac.     Lower     Upper   p-value |      N1        N2          N
------+--------------------------------------+------------------------------
    1 |   0.20   -2.4132    2.4132    0.0158 |  20.285    20.285     40.571
    2 |   0.40   -2.4132    2.4132    0.0158 |  40.571    40.571     81.141
    3 |   0.60   -2.4132    2.4132    0.0158 |  60.856    60.856     121.71
    4 |   0.80   -2.4132    2.4132    0.0158 |  81.141    81.141     162.28
    5 |   1.00   -2.4132    2.4132    0.0158 |  101.43    101.43     202.85
---------------------------------------------------------------------------

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

. display "Power attained at final analysis: " r(power_a) * 100
Power attained at final analysis: 90.003222
```

Option nfractional instructs gsdesign not to round sample sizes up to the nearest whole number. We can see that the first look occurs with 20.285 observations per arm, and the second occurs with 40.571 observations per arm. Rounding up to whole numbers of participants, this gives us 21 observations per arm for the first look, and an additional 20 observations (for a total of 41) at the second look. If this trial were to continue to the fifth look, it would require 202.85 participants to attain 90% power to detect a difference in means of one unit. As the sample size increases, the relative impact of rounding up to a whole number of observations diminishes.

◁

▷ Example 2: Pocock bounds with efficacy and futility stopping

In example 1, we saw that the GSD resulted in a substantially smaller expected sample size than an equivalent fixed study design if the alternative hypothesis was true but not if the null hypothesis was true. To increase the potential to stop the trial early if the treatment is ineffective, we now add futility bounds to the experimental design. Futility bounds separate the continuation region from the acceptance region and allow early acceptance of H_0 when there is evidence that the treatment is not meaningfully different from the control.

We use the same design as in example 1, this time adding the `futility(pocock)` option to add nonbinding Pocock futility bounds.

```
. gsdesign twomeans 0, diff(1) sd(2) knownsds power(0.9) efficacy(pocock)
> futility(pocock) nlooks(5) graphbounds

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

Efficacy: Pocock
Futility: Pocock, nonbinding

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
      delta = 1.0000
         m1 = 0.0000
         m2 = 1.0000
       diff = 1.0000
         sd = 2.0000

Expected sample size:
         H0 = 124.55
         Ha = 132.66

Info. ratio = 1.5966
    N fixed =    170
      N max =    270
     N1 max =    135
     N2 max =    135

Fixed-study crit. values = ±1.9600

Critical values, p-values, and sample sizes for a group sequential design
──────────────────────────────────────────────────────────────────────────
      │ Info.              Efficacy                         Futility
 Look │ frac.      Lower     Upper   p-value      Lower      Upper   p-value
──────┼───────────────────────────────────────────────────────────────────
    1 │  0.20    -2.4132    2.4132    0.0158    -0.1490     0.1490    0.8815
    2 │  0.40    -2.4132    2.4132    0.0158    -0.9078     0.9078    0.3640
    3 │  0.60    -2.4132    2.4132    0.0158    -1.4900     1.4900    0.1362
    4 │  0.80    -2.4132    2.4132    0.0158    -1.9808     1.9808    0.0476
    5 │  1.00    -2.4132    2.4132    0.0158    -2.4132     2.4132    0.0158
──────────────────────────────────────────────────────────────────────────

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

```
──────────────────────────────────────────────
            │         Sample size
  Look      │     N1         N2           N
──────────────────────────────────────────────
     1      │     27         27          54
     2      │     54         54         108
     3      │     81         81         162
     4      │    108        108         216
     5      │    135        135         270
──────────────────────────────────────────────
```

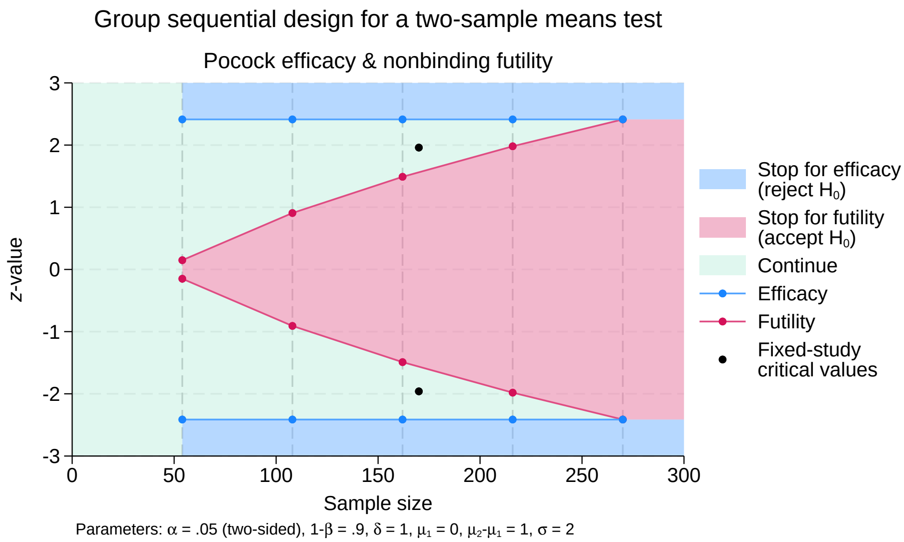


Figure 2. Pocock efficacy and futility bounds for a test of the equality of two means

The maximum sample size required by this design is even larger than that of the efficacy-only design, but the ability to end the trial early for futility can result in a considerably smaller sample size if H_0 is true. The efficacy bounds for this design are the same as they were in example 1; this is because adding nonbinding futility bounds to a group sequential trial does not affect the calculation of efficacy bound critical values.

As before, if $|z_1| \geq 2.413$, we reject H_0 and end the trial early for efficacy. With the addition of the futility bounds, we have the option of ending the trial early for futility if $|z_1| < 0.149$. If $|z_1| \in [0.149, 2.413)$, the trial must continue. While the Pocock efficacy bounds use the same critical values for all looks, the futility bounds do not; they grow from ± 0.149 at the first look to ± 1.981 by the fourth look, coinciding with the efficacy bounds at the fifth look.

As we move from left to right on the graph by collecting additional samples, we see the futility region grow and the continuation region shrink. The narrowing continuation region means that the trial is increasingly likely to stop due to futility or efficacy as more samples are collected. But if the test statistics do not cross the boundaries and the trial continues to the fifth look, the group sequential trial will require about 60% more participants than an equivalently powered fixed study.

One way to reduce the maximum sample size would be to use a boundary that is more conservative at early looks, such as an O'Brien–Fleming boundary. Another option is to use binding futility bounds instead of nonbinding bounds. While nonbinding futility bounds offer the option to stop the trial for efficacy if they are crossed, binding futility bounds require the termination of the trial if they are crossed. Continuing a trial that has crossed a binding futility bound can inflate the type I error, and any conclusions reached by the trial will be viewed with suspicion.

We rerun the previous example with `futility()` suboption `binding` to specify binding futility bounds, omitting the `graphbounds` option.

```
. gsdesign twomeans 0, diff(1) sd(2) knownsds power(0.9) efficacy(pocock)
> futility(pocock, binding) nlooks(5)

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

Efficacy: Pocock
Futility: Pocock, binding

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
      delta = 1.0000
         m1 = 0.0000
         m2 = 1.0000
       diff = 1.0000
         sd = 2.0000

Expected sample size:
         H0 = 120.18
         Ha = 113.00

Info. ratio = 1.5453
    N fixed =    170
      N max =    260
     N1 max =    130
     N2 max =    130

Fixed-study crit. values = ±1.9600

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

────────────────────────────────────────────────────────────────────────────
      │ Info.              Efficacy                          Futility
 Look │ frac.      Lower      Upper    p-value      Lower      Upper    p-value
──────┼─────────────────────────────────────────────────────────────────────
    1 │  0.20    -2.3564     2.3564     0.0185    -0.1290     0.1290     0.8974
    2 │  0.40    -2.3564     2.3564     0.0185    -0.8754     0.8754     0.3813
    3 │  0.60    -2.3564     2.3564     0.0185    -1.4482     1.4482     0.1476
    4 │  0.80    -2.3564     2.3564     0.0185    -1.9310     1.9310     0.0535
    5 │  1.00    -2.3564     2.3564     0.0185    -2.3564     2.3564     0.0185
────────────────────────────────────────────────────────────────────────────

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

────────────────────────────────────
      │          Sample size
 Look │       N1         N2          N
──────┼─────────────────────────────
    1 │       26         26         52
    2 │       52         52        104
    3 │       78         78        156
    4 │      104        104        208
    5 │      130        130        260
────────────────────────────────────
```

The binding futility bounds give a modest reduction in maximum sample size, down from 270 to 260. Compared with the nonbinding design, the binding design uses slightly smaller futility critical values. Also, while the efficacy-only design and the design with nonbinding futility bounds used efficacy critical values of ±2.413, here the efficacy critical values have shrunk to ±2.356.

To understand why, consider what happens when the null hypothesis is true. In this case, the correct action is to accept H_0, and it is a type I error to reject H_0. In the efficacy-only design of example 1, each interim look presents the opportunity to continue the trial or to commit a type I error and mistakenly reject H_0; only at the final look do we have the option to correctly accept H_0. With binding futility bounds, every look offers the possibility of crossing the futility boundary and correctly accepting 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. We do not relax the efficacy critical values when nonbinding futility boundaries are used because there is no guarantee that the trial will be stopped after crossing a futility boundary.

See [ADAPT] **gsdesign twomeans** for more examples of GSDs for tests of two sample means.

◁

Background on the BHAT study

The Beta-Blocker Heart Attack Trial (BHAT) was one of the first large-scale clinical trials to adopt a group sequential monitoring plan (Cook and DeMets 2008). 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 total mortality, and survival analysis was conducted using a log-rank test with a two-sided alternative hypothesis.

Recruitment ran from June 1978 to October 1980, with follow-up scheduled to continue until June 1982. Oversight was provided by an independent Policy and Data Monitoring Board (PDMB), which contained physicians, biostatisticians, and an ethicist. While the BHAT's study protocol did not set strict rules for early termination, the PDMB adopted the then-recently published O'Brien–Fleming method early on (DeMets et al. 1984).

Based on a combination of factors, including a log-rank test statistic that crossed the O'Brien–Fleming boundary at the sixth of seven looks, the PDMB stopped the BHAT for treatment efficacy in October of 1981, eight months before follow-up was scheduled to end in June 1982. Lan and DeMets (1989) report the values of the log-rank test statistic at each of the interim looks:

	May 1979	October 1979	March 1980	October 1980	April 1981	October 1981
test statistic	1.68	2.24	2.37	2.30	2.34	2.82

DeMets, Furberg, and Friedman (2006, Case 2) report that the BHAT was designed with a two-tailed alpha level of 0.05 and 90% power 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. A total of seven biannual analyses were planned, and O'Brien–Fleming efficacy bounds were calculated assuming seven evenly spaced looks.

Design for GSD with survival analysis

▷ Example 3: BHAT study

To re-create the design of the BHAT, we run `gsdesign logrank` with survival probabilities 0.8254 and 0.8625 for the control and intervention arms, respectively. We specify a power of 90% and O'Brien–Fleming efficacy bounds with seven evenly spaced looks.

```
. gsdesign logrank 0.8254 0.8625, power(0.9) efficacy(obfleming) nlooks(7)
> graphbounds

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

Efficacy: O'Brien-Fleming

Study parameters:
      alpha = 0.0500  (two-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 = 642.71
         Ha = 459.40

Info. ratio = 1.0323
    E fixed =    628
    N fixed =  4,024
      N max =  4,152
     N1 max =  2,076
     N2 max =  2,076

Fixed-study crit. values = ±1.9600

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

----------------------------------------------------------
      |  Info.            Efficacy          |      Events
 Look |  frac.      Lower     Upper  p-value|           E
------+-------------------------------------+-------------
    1 |   0.14    -5.4590    5.4590   0.0000|          93
    2 |   0.29    -3.8601    3.8601   0.0001|         186
    3 |   0.43    -3.1518    3.1518   0.0016|         278
    4 |   0.57    -2.7295    2.7295   0.0063|         371
    5 |   0.71    -2.4413    2.4413   0.0146|         463
    6 |   0.86    -2.2286    2.2286   0.0258|         556
    7 |   1.00    -2.0633    2.0633   0.0391|         648
----------------------------------------------------------

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

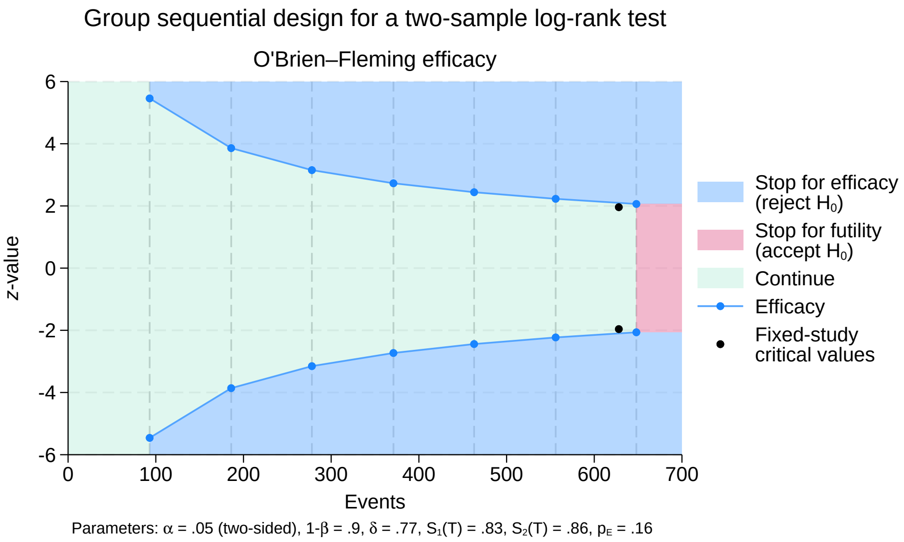


Figure 3. BHAT trial with O'Brien–Fleming efficacy bounds

At the top of the output, gsdesign 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. A fixed study would require 628 events (deaths) to detect a hazard ratio of 0.7709 with 90% power, and with the specified survival probabilities, this corresponds to a sample size of 4,024.

The GSD requires a maximum of 648 events (corresponding to a sample of size 4,152) if it continues to the final look. If the null hypothesis is correct (the hazard ratio is 1) and the BHAT were to be repeated many times using this design, we would expect to observe an average of 642.71 events per trial. This is near the maximum because if the null hypothesis is true, in most replications the trial will continue to the final look; only rarely will the trial be stopped early for efficacy (which would be a type I error). 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 459.4 events per trial. The substantial sample-size savings (try saying that five times fast) is due to the fact that many replications of the trial will correctly be stopped early for efficacy.

The log-rank statistic is asymptotically normally distributed with independent information increments, and can be compared directly against group sequential critical values (Tsiatis 1982). The critical values we calculate match those used by the PDMB Cook and DeMets (2008, 306).

At the first look, the test statistic $z_1 = 1.68 < 5.459$, so the trial continued. The test statistics at the following four looks are also in the continuation region ($z_2 = 2.24 < 3.86$, $z_3 = 2.37 < 3.152$, $z_4 = 2.30 < 2.73$, and $z_5 = 2.34 < 2.441$), bringing the trial to the sixth of seven planned looks. At the sixth look, the test statistic crosses the efficacy bound, $z_6 = 2.82 > 2.229$, which supports the PDMB's decision to stop the trial for treatment efficacy.

Two aspects of the O'Brien–Fleming bound that the PDMB found appealing were the conservative critical values early in the trial and the final critical value that is only marginally larger than the fixed-study critical value (DeMets et al. 1984). An additional advantage is that even if the trial were to continue to the final look, the O'Brien–Fleming design requires only 3% more information (deaths, in this case) than a fixed study.

While the BHAT was a success story for the use of group sequential clinical trials, it was not without its challenges (DeMets, Furberg, and Friedman 2006). The number of participants recruited was nearly equal to the desired sample size, so the power would have been almost 90% to detect the difference between the anticipated survival probabilities of 82.54% and 86.25%, but survival was higher than anticipated for both the control and intervention groups. At the sixth look, only 318 of the anticipated 556 events had been observed, and a smaller-than-anticipated number of events can reduce the power of the test. Fortunately, adherence was also better than anticipated, and the effect size was larger than anticipated. The reduced number of events observed impacted the power of the test but did not influence the probability of committing a type I error.

A potentially more vexing issue is that the efficacy critical values were calculated under the assumption of equal information increments, but the interim analyses were scheduled based on calendar time, making it impossible to enforce an evenly spaced information sequence. Severe violations of this assumption can cause excessive type I error, but the number of deaths between looks was approximately equal, and type I error control is robust to minor violations of this assumption (DeMets et al. 1984).

◁

▷ Example 4: Error-spending bounds

One of the members of the PDMB from the BHAT, David DeMets, was inspired by the experience to find a more flexible method of calculating group sequential boundaries. Lan and DeMets (1983) developed error-spending methods, which depend on the total information to be collected and the interim analyses already conducted but not on the critical values of future looks. This flexibility allows error-spending bounds to adjust to scenarios such as the BHAT, where the precise information fraction at each look is not known a priori. This framework was further extended by Lan and DeMets (1989), who introduced methods for calculating stopping boundaries based on calendar time.

Here we reimagine the BHAT trial using an error-spending approximation to the classical O'Brien–Fleming boundary (Lan and DeMets 1983). Instead of specifying evenly spaced looks, we use Method 2 from Lan and DeMets (1989, 1195) to specify the timing of interim looks based on calendar time. To do this, we use the `information()` option instead of the `nlooks()` option, and we specify the timing of each look as the number of months since June 1979, when the study began accruing participants. We graph the bounds and label the x axis with the number of months since June 1979.

```
. gsdesign logrank 0.8254 0.8625, power(0.9) efficacy(errobfleming)
> information(11 16 21 28 34 40 48)
> 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

Study parameters:
      alpha = 0.0500  (two-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 = 641.04
          Ha = 461.13

Info. ratio = 1.0280
    E fixed =    628
    N fixed =  4,024
      N max =  4,136
     N1 max =  2,068
     N2 max =  2,068

Fixed-study crit. values = ±1.9600

Critical values, p-values, and sample sizes for a group sequential design
─────────────┬───────────────────────────────────────────┬──────────────
             │ Info.            Efficacy                 │       Events
  Look       │ frac.      Lower       Upper     p-value  │            E
─────────────┼───────────────────────────────────────────┼──────────────
         1   │  0.23    -4.5380     4.5380     0.0000   │          148
         2   │  0.33    -3.7128     3.7128     0.0002   │          216
         3   │  0.44    -3.2081     3.2081     0.0013   │          283
         4   │  0.58    -2.7361     2.7361     0.0062   │          377
         5   │  0.71    -2.4739     2.4739     0.0134   │          458
         6   │  0.83    -2.2717     2.2717     0.0231   │          538
         7   │  1.00    -2.0473     2.0473     0.0406   │          646
─────────────┴───────────────────────────────────────────┴──────────────
Note: Critical values are for z statistics; otherwise, use
      p-value boundaries.
```

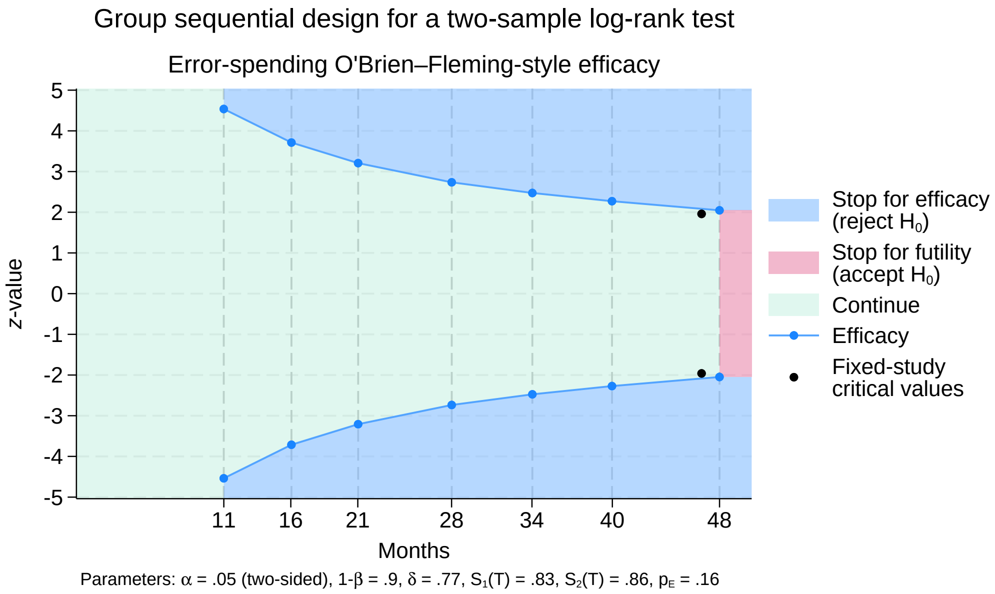


Figure 4. BHAT trial with error-spending bounds

The new design maintains the same familywise significance level, power, and effect size as the original BHAT design, so the fixed-study equivalent of the new design requires the same 628 events as the fixed equivalent of the original BHAT. Comparing the stopping boundaries of the new error-spending design against those of the original design, we see that the new critical values are quite close to those calculated using classical O'Brien–Fleming bounds with evenly spaced looks. The maximum number of events remains nearly the same, with the new design calling for 646 events at the final analysis versus 648 for the classical O'Brien–Fleming design.

More importantly, when the new error-spending boundaries are used to determine stopping for the BHAT trial, they support the same conclusion as the classical O'Brien–Fleming boundaries: to terminate the trial for efficacy at the sixth look. The first five tests statistics lie in the continuation region of the new design, but at the sixth look, $z_6 = 2.82 > 2.272$.

◁

Stored results

To calculate the fixed-study sample size, `gsdesign` *method* runs `power` *method* and returns all the method-specific stored results as well as the following common results 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(E_fixed)`	total number of events (failures) in a fixed study design (survival analysis only)
`r(E_max)`	maximum observed events if the study continues to completion (survival analysis only)
`r(effparam)`	efficacy parameter (if `wtsiatis()`, `kdemets()`, or `hsdecani()` specified)
`r(Efrac_fixed)`	fractional total number of events (failures) in a fixed study design (survival analysis only)
`r(ESS0)`	expected sample size under null hypothesis
`r(ESS1)`	expected sample size under alternative hypothesis
`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(N_fixed)`	sample size of a fixed study design
`r(N_fixedfrac)`	fractional sample size of a fixed study design
`r(N_max)`	maximum sample size if the study continues to completion
`r(N1_fixed)`	sample size of the control group in a fixed study design (multiarm trials only)
`r(N1_fixedfrac)`	fractional sample size of the control group in a fixed study design (multiarm trials only)
`r(N1_max)`	maximum sample size of the control group if the study continues to completion (multiarm trials only)
`r(N2_fixed)`	sample size of the experimental group in a fixed study design (multiarm trials only)
`r(N2_fixedfrac)`	fractional sample size of the experimental group in a fixed study design (multiarm trials only)
`r(N2_max)`	maximum sample size of the experimental group if the study continues to completion (multiarm trials only)
`r(nfractional)`	`1` if `nfractional` is specified, `0` otherwise
`r(nlooks)`	number of analyses
`r(onesided)`	`1` for a one-sided test, `0` otherwise
`r(power)`	specified overall power
`r(power_a)`	attained 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)`	`gsdesign`
`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`
`r(method)`	method name

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(bspent_a)`	attained cumulative beta spent per look (when futility bounds are specified)
`r(design)`	sample size and stopping boundaries at interim looks
`r(info_frac)`	specified information fraction
`r(info_frac_a)`	fraction of attained information
`r(info_level)`	specified information level
`r(p_crit)`	p-values corresponding to boundary critical values
`r(sampsize)`	sample size at interim looks

Methods and formulas

See *Methods and formulas* in [ADAPT] **gsbounds** for the formulas used to calculate the stopping boundaries, information fraction, and information ratio. See *Methods and formulas* in [PSS-2] **power** for the formulas used to calculate sample size of a fixed study design.

Methods and formulas are presented under the following headings:

Sample sizes at interim analyses
Expected sample size

Sample sizes at interim analyses

When planning a study using a GSD with K looks, we must specify the information fraction at each look, denoted as $(\mathcal{I}_1, \dots, \mathcal{I}_K)$. For any k in $(1, \dots, K)$, let $\mathcal{I}_k$ represent the proportion of trial data that has been collected by look k. In most cases, the information fraction is the proportion of the maximum sample size that has been collected, but for time-to-event data, the information fraction is the proportion of the total number of failure events that have been observed, not the total number of participants.

With `gsdesign`, the `information(`*numlist*`)` option can be used to specify the information fraction as a strictly increasing sequence, which is then scaled so that $\mathcal{I}_K = 1$. Alternatively, the `nlooks()` option can be used to specify the number of evenly spaced looks, and the information fraction is calculated automatically.

To determine the sample size required at each look of a GSD, we begin by calculating n_{fix}, the sample size of a fixed study design with equivalent type I and type II error. Next we calculate the information ratio, R, which is the ratio of the maximum sample size of the GSD to n_{fix}. Regardless of the properties of the study, R is always greater than 1 (see *Methods and formulas* in [ADAPT] **gsbounds** for more information).

Let $(n_1, \dots, n_K)$ 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 look k in $(1, \dots, K)$, the sample size $n_k = \mathcal{I}_k \times n_{\text{fix}} \times R$. In practice, sample sizes must be rounded up to whole numbers of participants, so `gsdesign` rounds up sample sizes unless the `nfractional` option is specified.

Expected sample size

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 boundary critical values, and then p-values for the test statistics are compared with p-values corresponding to the boundary critical values. The p-value comparison is known as the significance level approach and is described in [ADAPT] **gsbounds**.

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 $(e_1, \dots, e_K)$ and critical values for futility stopping as $(f_1, \dots, f_K)$. 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 $f_K = e_K$.

The probability of stopping the trial at look k is a function of the effect size δ and is denoted as $\omega_k(\delta)$, where $\omega_1(\delta) = \text{Pr}_\delta(Z_1 < f_1) + \text{Pr}_\delta(Z_1 \geq e_1)$ and

$$\omega_k(\delta) = \text{Pr}_\delta \left\{ (Z_k < f_k \cup Z_k \geq e_k) \cap \bigcap_{j=1}^{k-1} f_j \leq Z_j < e_j \right\} \quad \text{for } k \in (2, \dots, K)$$

For trials with efficacy stopping only, replace $(f_1, \dots, f_{K-1})$ with $-\infty$ and let $f_K = e_K$. For trials with nonbinding futility bounds, replace $(f_1, \dots, f_{K-1})$ with $-\infty$ when $\delta = 0$ but not when $\delta \neq 0$. For trials with futility stopping only, replace $(e_1, \dots, e_{K-1})$ with ∞ and let $e_K = f_K$. For two-sided trials, replace Z_k with $|Z_k|$.

The expected sample size is a function of effect size δ and is calculated as

$$\text{ESS}(\delta) = \sum_{k=1}^{K} n_k * \omega_k(\delta)$$

References

Cook, T. D., and D. L. DeMets. 2008. *Introduction to Statistical Methods for Clinical Trials.* Boca Raton, FL: Chapman and Hall/CRC.

DeMets, D. L., C. D. Furberg, and L. M. Friedman, eds. 2006. *Data Monitoring in Clinical Trials: A Case Studies Approach.* New York: Springer. https://doi.org/10.1007/0-387-30107-0.

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.

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.

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

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.

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.

Tsiatis, A. A. 1982. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* 77: 855–861. https://doi.org/10.1080/01621459.1982.10477898.

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

[ADAPT] **gsbounds** — Boundaries for group sequential trials

[ADAPT] **Glossary**

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

gsdesign onemean — Group sequential design for a one-sample mean test

Description Quick start Menu Syntax Options
Remarks and examples Stored results Methods and formulas References Also see

Description

gsdesign onemean computes stopping boundaries and sample sizes for interim analyses of clinical trials using a one-sample mean test with a group sequential design (GSD). Stopping can be for efficacy, futility, or both. For stopping boundary calculations without sample sizes, see [ADAPT] **gsbounds**. For sample-size calculations for a fixed-sample test of one mean, see [PSS-2] **power onemean**.

Quick start

Sample size and stopping boundaries for a two-sided test of $H_0: \mu = 10$ versus $H_a: \mu \neq 10$, with default power of 0.8 to detect the difference between the mean under the null hypothesis, $m_0 = 10$, and an observed mean of $m_a = 15$, with standard deviation of 12 and at default familywise significance level $\alpha = 0.05$, using default group sequential specifications of O'Brien–Fleming efficacy boundaries with two analyses (one interim, one final)

```
gsdesign onemean 10 15, sd(12)
```

Same as above, but specified as m_0 and difference $m_a - m_0 = 5$

```
gsdesign onemean 10, diff(5) sd(12)
```

Same as above, but for a one-sided test with familywise significance level $\alpha = 0.025$, power of 0.9, and Pocock bounds with three analyses

```
gsdesign onemean 10, diff(5) sd(12) alpha(0.025) power(0.9) onesided ///
      efficacy(pocock) nlooks(3)
```

Same as above, but use an error-spending O'Brien–Fleming-style efficacy bound and a nonbinding error-spending Pocock-style futility bound

```
gsdesign onemean 10, diff(5) sd(12) alpha(0.025) power(0.9) onesided ///
      efficacy(errobfleming) futility(errpocock) nlooks(3)
```

Same as above, but treat the standard deviation as known, use a nonbinding Hwang–Shih–de Cani futility bound with parameter $\gamma_f = -1$, time the looks to occur with 50%, 75%, and 100% of the data, and graph the bounds

```
gsdesign onemean 10, diff(5) sd(12) knownsd alpha(0.025) power(0.9) ///
      onesided efficacy(errobfleming) futility(hsdecani(-1))    ///
      information(50 75 100) graphbounds
```

Same as above, but use a Kim–DeMets efficacy bound with parameter $\rho_e = 2.5$, label the x axis of the graph with information instead of sample size, and do not plot the critical values from a fixed study design

```
gsdesign onemean 10, diff(5) sd(12) knownsd alpha(0.025) power(0.9) ///
      onesided efficacy(kdemets(2.5)) futility(hsdecani(-1))    ///
      information(50 75 100) graphbounds(xdiminformation nofixed)
```

Menu

Statistics > Power, precision, and sample size

Syntax

`gsdesign onemean` m_0 m_a [, *onemeanopts boundopts*]

where m_0 is the value of the mean under the null hypothesis and m_a is the value of the mean under the alternative hypothesis.

onemeanopts	Description
Main	
`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)`
`onesided`	request a one-sided test; default is two-sided
`nfractional`	report fractional sample size
`diff(#)`	difference between the alternative mean and the null mean, $m_a - m_0$; specify instead of the alternative mean m_a
`sd(#)`	standard deviation; default is `sd(1)`
`knownsd`	request computation assuming known standard deviation; default is to assume an unknown standard deviation
`force`	allow calculation with unsupported `power onemean` options
`poweriteration(`*powiteropts*`)`	iteration options for the calculation of fixed-study sample size; seldom used

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

`force` and `poweriteration()` do not appear in the dialog box.

powiteropts	Description
`init(#)`	initial value for fixed-study sample size
`iterate(#)`	maximum number of iterations; default is `iterate(500)`
`tolerance(#)`	parameter tolerance; default is `tolerance(1e-12)`
`ftolerance(#)`	function tolerance; default is `ftolerance(1e-12)`

boundopts	Description
Bounds	
`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(#`[`, equal`]`)`	total number of analyses (`nlooks()` − 1 interim analyses and one final analysis); use `equal` to enforce equal information increments; if neither `nlooks()` nor `information()` is specified, the default is `nlooks(2)`
`information(`*numlist*`)`	sequence of information levels for analyses; default is evenly spaced
`nopvalues`	suppress p-values
Graph	
`graphbounds`[`(`*graphopts*`)`]	graph boundaries
`matlistopts(`*general_options*`)`	control the display of boundaries and sample size; seldom used
optimopts	optimization options for boundary calculations; seldom used

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

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

graphopts	Description
`xdimsampsize`	label the x axis with the sample size collected (default)
`xdiminformation`	label the x axis with the information fraction; 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

`alpha(#)` sets the overall significance level, which is the familywise type I error rate for all analyses (interim and final). `alpha()` must be in $(0, 0.5)$. The default is `alpha(0.05)`.

`power(#)` sets the overall power for all analyses. `power()` must be in $(0.5, 1)$. 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. `beta()` must be in $(0, 0.5)$. 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.

`onesided` requests a study design for a one-sided test. The direction of the test is inferred from the effect size.

`nfractional` specifies that fractional sample sizes be reported.

`diff(#)` specifies the difference between the alternative mean and the null mean, $m_a - m_0$. You can either specify the alternative mean m_a as a command argument or specify the difference between the two means in `diff()`. If you specify `diff(#)`, the alternative mean is computed as $m_a = m_0 + \#$.

`sd(#)` specifies the sample standard deviation or the population standard deviation. The default is `sd(1)`. By default, `sd()` specifies the sample standard deviation. If `knownsd` is specified, `sd()` specifies the population standard deviation.

`knownsd` requests that the standard deviation be treated as known in the computation. By default, the standard deviation is treated as unknown and the computation is based on a t test, which uses a Student's t distribution as a sampling distribution of the test statistic. If `knownsd` is specified, the computation is based on a z test, which uses a normal distribution as the sampling distribution of the test statistic. In either case, critical values for efficacy and futility boundaries calculated by `gsdesign onemean` are reported on the standardized z scale. When a t test is performed, you can use the significance level approach and compare the p-value from the t test to the p-value boundaries reported by `gsdesign onemean`, as demonstrated in example 2.

Bounds

`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(#` [`, equal`]`)` 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)`.

`equal` indicates that equal information increments be enforced, which is to say that the same number of new observations will be collected at each look. The default behavior is to start by dividing information evenly among looks, then proceed by rounding up to a whole number of observations at each look. This can cause slight differences in the information collected at each look.

`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.

Graph

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

graphopts are the following:

`xdimsampsize` labels the x axis with the sample size collected (the default).

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

`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 gsdesign onemean but are not shown in the dialog box:

force indicates that gsdesign onemean should allow unsupported power onemean options, such as options specifying a finite population correction or a cluster randomized design. Even with option force, the power onemean options specified must be compatible with sample-size determination, not effect size or power calculation. In addition, *numlist*s are not supported in options or in arguments as they are with power, even when force is specified.

poweriteration(*powiteropts*) controls the iterative algorithm used to calculate the fixed-study sample size. This is seldom used.

powiteropts are the following:

init(#) specifies an initial value for the sample size when iteration is used to compute the fixed-study sample size. The default is to use a closed-form normal approximation to compute an initial sample size.

iterate(#) specifies the maximum number of iterations for the Newton method during calculation of the fixed-study sample size. The default is iterate(500).

tolerance(#) specifies the tolerance used to determine whether successive parameter estimates have converged when calculating the fixed-study sample size. The default is tolerance(1e-12). See *Convergence criteria* in [M-5] **solvenl( )** for details.

ftolerance(#) specifies the tolerance used when calculating the fixed-study sample size to determine whether the proposed solution of a nonlinear equation is sufficiently close to 0 based on the squared Euclidean distance. The default is ftolerance(1e-12). See *Convergence criteria* in [M-5] **solvenl( )** for details.

matlistopts(*general_options*) affects the display of the matrix of boundaries and sample sizes. *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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**. 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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

`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 $\mathbf{g}\mathbf{H}^{-1}\mathbf{g}' <$ 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:

Introduction
Using gsdesign onemean
Background for examples
Computing sample size and stopping boundaries
Unknown standard deviation and hypothesis tests on mean
Stopping for both efficacy and futility

This entry describes the use of the `gsdesign onemean` command for designing a group sequential analysis for a one-sample mean test. See [ADAPT] **GSD intro** for a general introduction to GSDs for clinical trials; see [ADAPT] **gsbounds** for information about group sequential bounds; and see [ADAPT] **gsdesign** for information about designing group sequential clinical trials with the `gsdesign` command. Also see [PSS-2] **Intro (power)** for a general introduction to power and sample-size analysis, and see [PSS-2] **power onemean** for details about study design for a one-sample mean test.

Introduction

The gold standard for clinical trials is the randomized controlled trial, a clinical 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. If no existing treatments are comparable with the experimental treatment, the control group will typically receive a placebo. When a standard of care exists, there is often an ethical argument against using a placebo. In this case, an active control is used, in which control-group participants receive the existing standard of care.

Sometimes, however, it is desirable to omit the control group entirely and perform a single-arm trial. This could occur with trials not intended for regulatory submission (such as early-phase trials), trials where a placebo would be unethical and there is no existing comparable treatment, or clinical trials where the population of interest is small and recruiting sufficient participants to create both treatment and control groups would be difficult.

If the outcome of interest, or endpoint, of a randomized controlled trial is continuous, researchers will commonly want to test whether the mean of the treatment group is equal to the mean of the control group. If there is no control group, the observed mean of the treatment group, μ, is compared with reference value μ_0, which is the value under the null hypothesis.

Here we consider the null hypothesis $H_0: \mu = \mu_0$ versus the two-sided alternative hypothesis $H_a: \mu \neq \mu_0$, the upper one-sided alternative $H_a: \mu > \mu_0$, or the lower one-sided alternative $H_a: \mu < \mu_0$. The actual test conducted will depend on whether the standard deviation of the sampling process is known a priori. In the case of a known standard deviation, the test statistic follows a standard normal distribution under the null hypothesis, and the corresponding test is known as a z test. In the case of an unknown standard deviation, an estimate of the population standard deviation is used to form a test statistic that follows Student's t distribution under the null hypothesis, and the corresponding test is a t test. The estimate of the population standard deviation improves with increasing sample size, and the distribution of the test statistic approaches a normal distribution, enabling the use of a z test with large samples, even with unknown standard deviation.

Historically, clinical trials were not analyzed until after the trial was completed and all data were collected. This enabled the use of traditional statistical methods such as those designed for agricultural field experiments, where all data are typically collected over a short period during the harvest. However, data from clinical trials tend to be collected over time, providing incentive to analyze the incomplete trial data while the study is still underway. Several methods have been developed to enable interim analysis of clinical trial data while controlling the type I error rate, and group sequential methods are among the most popular designs.

A GSD plans for a sequence of looks, or analyses of the clinical trial data. If the result at an interim look is sufficiently unambiguous, the trial can be stopped early. GSDs with efficacy bounds allow the trial to be terminated early for treatment efficacy if the interim test statistic crosses the efficacy bound, and designs with futility bounds allow early termination for futility if the interim test statistic crosses the futility bound.

The required sample size estimated by `gsdesign onemean` will depend on whether the standard deviation is known, but the stopping boundaries will not; they are reported on a standardized z scale. The critical values from the boundaries may be compared directly with the z statistic from a z test but must be transformed before being compared with the t statistic from a t test. This is demonstrated in example 2.

Using gsdesign onemean

`gsdesign onemean` calculates sample size and stopping boundaries for a group sequential trial comparing the population mean of one group against a prespecified reference value. `gsdesign onemean` can be thought of as a combination of `power onemean` for sample-size calculations and `gsbounds` for stopping boundary calculations.

To compute sample size, you must specify the effect size in one of two ways: by specifying m_0 and m_a, the means under the null and alternative hypotheses, respectively, or by specifying the difference $m_a - m_0$ in the `diff()` option. There is no default value for `diff()`, so either m_0 and m_a or `diff()` must be included as part of the command specification. m_0 may be specified with `diff()`; if m_0 is omitted, it is assumed to be 0. Another aspect of the effect size is the standard deviation of the response. This is specified with the `sd()` option, and the default standard deviation is 1. By default, the true population standard deviation is assumed to be unknown, and sample-size calculations are based on a t test; if the population standard deviation is known, the `knownsd` option requests sample-size computations based on a z test.

Options alpha(), power(), beta(), and onesided are used for both sample-size and stopping-boundary calculations. The default significance level, known as the familywise type I error rate, is 0.05 and can be changed by specifying the alpha() option. The default power is 0.8, which corresponds to a type II error rate of 0.2. This can be modified either by specifying the power in the power() option or by specifying the type II error in the beta() option. The default test is two-sided, and the onesided option requests a one-sided test, the direction of which is indicated by the sign of the effect size.

The group sequential stopping rule is determined by the efficacy() and futility() options. Stopping can be for efficacy, futility, or both, and if no stopping rule is specified, the default is to use an O'Brien–Fleming efficacy bound. If futility bounds are requested, the default behavior is to treat them as nonbinding. A trial that crosses a nonbinding futility bound can be stopped for futility, but the familywise type I error is controlled even if the trial continues. Binding futility bounds can be requested with futility() suboption binding. A trial that crosses a binding futility bound must be stopped for futility. If it continues, the familywise type I error will not be controlled at the specified significance level.

The number of looks, or analyses of the trial data, is specified with nlooks(). Alternatively, the information() option can be used to specify the spacing of the looks as a *numlist* of increasing information levels. In this case, values of the numlist are automatically rescaled so that the final look has the maximum information required by the design. If neither nlooks() nor information() is specified, the default is two looks.

By default, the sample size is rounded up to a whole number at each look, but option nfractional can be used to report fractional sample sizes. If nlooks() is specified, the default behavior is to divide information evenly among looks before rounding. Rounding can cause slight differences in the amount of information collected at each look, and nlooks() suboption equal can be specified to enforce equal information increments by requiring the same number of new observations at each look.

Background for examples

Oncology is an area where single-arm trials have become increasingly popular, and single-arm oncology trials have even led to regulatory approval (Tenhunen et al. 2020). There are multiple reasons for this trend, but Simon et al. (2015) identify several common factors of single-arm oncology trials that have been approved by regulators, including "unprecedented effect on tumor response is observed in a setting of high unmet medical need" as well as the existence of a well-defined target population.

Salvage therapy, also known as rescue therapy, is not a specific cancer treatment but rather a term to describe treatments that are considered for a patient when all standard treatment protocols have failed. We consider the design of a clinical trial of sunitinib malate as a salvage therapy for advanced unresectable non–small cell lung cancer, where "unresectable" describes tumors that cannot be removed surgically. Sunitinib is a tyrosine kinase inhibitor that has been used in the treatment of renal cancer and gastrointestinal stromal tumors, and it has been considered more recently as a treatment for other cancers. Novello et al. (2009) report the results of a phase 2 clinical trial of sunitinib for participants with advanced non–small cell lung cancer that did not respond to platinum-based chemotherapy, the standard of care.

The trial's endpoint, or outcome of interest, is the tumor shrinkage rate (TSR) of the participant's largest tumor, defined as TSR $= (D_b - D_a)/(D_b \times t) \times 100\%$, where D_b is the longest diameter of the tumor before treatment, D_a is the longest diameter after treatment, and t is the time elapsed in days. This endpoint was chosen based on the results of Yu et al. (2019), who identified TSR as a useful predictor of long-term outcomes in participants with advanced unresectable non–small cell lung cancer who received treatment with tyrosine kinase inhibitors.

By focusing on participants with unresectable tumors that have not responded to existing treatments, we have clearly defined a target population with high unmet medical need. There can be no active control because all standard treatment protocols have failed, and the only ethical justification for a placebo control would be if it were necessary to draw valid conclusions about the effect of sunitinib on TSR. As it so happens, non–small cell lung cancer is a well-studied disease, and it is known that spontaneous tumor shrinkage in patients with advanced-stage disease is exceedingly rare (Shatola et al. 2020).

Computing sample size and stopping boundaries

Suppose that we wish to conduct a single-arm trial, where participants will undergo a CT scan to measure tumor diameter both before and after a four-week course of sunitinib. We will test the null hypothesis $H_0 : \mu = \mu_0$ versus the upper one-sided alternative $H_a : \mu > \mu_0$, where μ is the mean TSR observed in study participants and $\mu_0 = 0$, indicating no tumor shrinkage under the null hypothesis. We choose a one-sided test because tumor shrinkage is of interest, not tumor growth. If we were interested in testing whether sunitinib slowed the rate of tumor growth, we would need a control group to compare against.

▷ Example 1: Efficacy bounds for a one-sample mean test

Yu et al. (2019) observed that participants with TSR $\geq$ 0.49% had a nearly 50% increase in overall survival time and in progression-free survival time when compared with participants with TSR $<$ 0.49%, so we desire 80% power to detect $\mu_a = 0.49$. We will conduct a one-sided trial with a familywise significance level of 2.5%, and we will assume that TSR has a known standard deviation of 1.1%. Assuming a known standard deviation is often unrealistic, and this assumption is relaxed in example 2. The stopping rule we employ is a Pocock efficacy bound, with four looks at the data: three interim analyses and one final analysis. We graph the bounds for visual inspection.

```
. gsdesign onemean 0 0.49, sd(1.1) knownsd alpha(0.025) efficacy(pocock)
> nlooks(4) onesided graphbounds

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

Efficacy: Pocock

Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.8000
      delta = 0.4455
         m0 = 0.0000
         ma = 0.4900
         sd = 1.1000

Expected sample size:
         H0 =  47.45
         Ha =  32.02

Info. ratio = 1.2025
    N fixed =     40
      N max =     48

Fixed-study crit. value = 1.9600
```

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

             Info.        Efficacy        Sample size
  Look       frac.     Upper    p-value             N

     1        0.25    2.3613     0.0091            12
     2        0.50    2.3613     0.0091            24
     3        0.75    2.3613     0.0091            36
     4        1.00    2.3613     0.0091            48

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

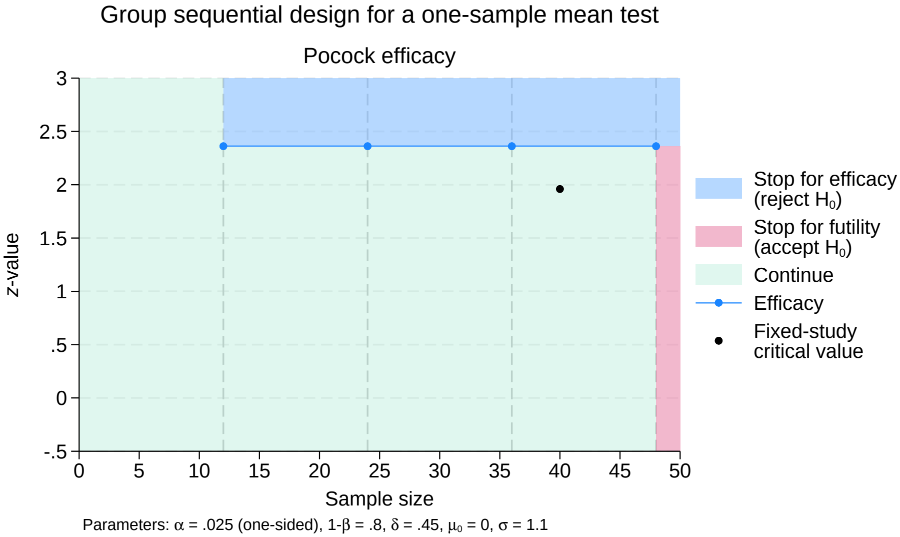


Figure 1. Sample size for a test of one mean with one-sided Pocock efficacy bounds

`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 47.45, close to the maximum of 48 participants at the fourth look. This is expected because our design does not allow for early stopping to accept the null hypothesis. If H_a is true, we expect an average of only 32.02 participants due to the probability of early stopping to reject H_0, a savings over the 40 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 four looks. Because we indicated that the population standard deviation was known, the appropriate analysis is a z test, which yields a z statistic. The first analysis, look 1, will be conducted once data have been collected from 12 participants. The second and third looks take place with 24 and 36 observations, respectively, and the final analysis occurs once data from all 48 participants have been collected. A defining feature of Pocock efficacy bounds is that all looks use the same critical value, which in this case is 2.361. If the test statistic at the first look, which we will label z_1, is greater than or equal to 2.361, then it lies in the rejection region, shaded blue on the graph. In this case, H_0 will be rejected and the study will terminate early due to treatment efficacy. If z_1 is less than 2.361, then it is said to lie in the continuation region and the study will proceed to the next look. The same procedure is repeated at the second and third looks. At the final look, there is no continuation region, only rejection and acceptance regions. H_0 is rejected if $z_4 \geq 2.361$, and H_0 is accepted if $z_4 < 2.361$. In addition to displaying the bounds visually, the graph also marks the critical value from a fixed study with equivalent power and significance level.

◁

Unknown standard deviation and hypothesis tests on mean

▷ Example 2: Unknown standard deviation, specifying difference between means

Instead of specifying the alternative mean of 0.49 as in example 1, we can specify the difference between the mean TSRs under the null and alternative hypotheses in the `diff()` option. Additionally, we will omit option `knownsd` because we do not know the population standard deviation a priori. This yields sample sizes for a t test, which will be demonstrated. The overall shape of the bounds is unchanged from figure 1, so we omit the graph.

```
. gsdesign onemean, diff(0.49) sd(1.1) alpha(0.025) efficacy(pocock)
> nlooks(4) onesided

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

Efficacy: Pocock

Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.8000
      delta = 0.4455
         m0 = 0.0000
         ma = 0.4900
       diff = 0.4900
         sd = 1.1000

Expected sample size:
         H0 =  49.44
         Ha =  33.48

Info. ratio = 1.2025
    N fixed =     42
      N max =     50

Fixed-study crit. value = 1.9600
```

```
Critical values, p-values, and sample sizes
for a group sequential design
------------------------------------------------------
       |  Info.        Efficacy      |  Sample size
  Look |  frac.     Upper    p-value |            N
-------+-----------------------------+----------------
     1 |   0.25    2.3613     0.0091 |           13
     2 |   0.50    2.3613     0.0091 |           25
     3 |   0.75    2.3613     0.0091 |           38
     4 |   1.00    2.3613     0.0091 |           50
------------------------------------------------------
Notes: Critical values are for z statistics;
       otherwise, use p-value boundaries.
       Requested information fraction not
       attained.
```

This alternate method of specifying the difference in means has not changed the fixed-study parameters reported by `gsdesign`, nor has it changed the critical values of the efficacy boundaries. What has changed is the required sample size to achieve 80% power with a fixed study, which has increased from 40 participants in example 1 to 42 participants now that we do not assume a known standard deviation. The information ratio is the same, but the larger fixed-study sample size yields larger required samples in the GSD as well, with a maximum sample of 50 participants and correspondingly larger expected sample sizes under the null and alternative hypotheses.

The biggest change, however, is the process of testing our hypothesis at the interim analyses. Previously, a z test was performed and the resulting z statistic was compared directly with the critical values of the efficacy bounds. Now that the standard deviation is estimated, we use a t test, which yields t statistics. These t statistics should not be compared directly with the standardized z critical values, as the note under the table indicates. Instead, the p-value from the test can be compared with the boundaries by using the significance level approach described in [ADAPT] **gsbounds**.

Imagine that the first look is conducted; the mean TSR of the first 13 participants is 0.9 and the sample standard deviation is 1.3. We conduct a t test using `ttesti`, the immediate form of the [R] **ttest** command. We specify `ttesti 13 0.9 1.3 0`, with the sample size as the first number, the observed mean as the second, the sample standard deviation as the third, and the mean under H_0 as the fourth.

```
. ttesti 13 0.9 1.3 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       x |      13          .9    .3605551         1.3    .1144179    1.685582
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   2.4962
H0: mean = 0                                     Degrees of freedom =       12

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9859         Pr(|T| > |t|) = 0.0281          Pr(T > t) = 0.0141
```

The output displays a t statistic of 2.496 with 12 degrees of freedom. `ttesti` also displays the p-value of the upper one-sided test: `Pr(T > t)` $= 0.0141$. This value, p_1, can be compared directly with the p-value of the efficacy bound. Because $p_1 > 0.0091$, we cannot reject H_0 at this look.

As the sample size increases and the degrees of freedom of the t test increase, the t distribution approaches a normal distribution. In practice, it is common to see a z test used with large samples, even when the population standard deviation is unknown and is estimated from the sample. But with small samples, the significance level approach is best.

◁

Stopping for both efficacy and futility

▷ Example 3: Stopping for both efficacy and futility, graphing bounds

The previous two examples have used stopping rules allowing for efficacy stopping, but further efficiency gains can be realized if the trial is also allowed to stop early due to treatment futility. This is done by creating a framework for accepting H_0, a practice that is unheard-of in some areas of statistical analysis but commonly discussed in the literature about GSDs.

Continuing with the scenario of example 2, we add a Wang–Tsiatis futility bound with parameter $\Delta_f = 0.3$. The boundary is nonbinding, the default for futility bounds, meaning that if the trial crosses the futility bound, stopping is optional. If the trial continues despite an interim result below the futility bound, the familywise type I error is still controlled at the desired significance level of 2.5%.

```
. gsdesign onemean, diff(0.49) sd(1.1) alpha(0.025) efficacy(pocock)
> futility(wt(0.3)) nlooks(4) onesided graphbounds

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

Efficacy: Pocock
Futility: Wang-Tsiatis, nonbinding, Delta = 0.3000

Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.8000
      delta = 0.4455
         m0 = 0.0000
         ma = 0.4900
       diff = 0.4900
         sd = 1.1000

Expected sample size:
         H0 =  23.96
         Ha =  37.78

Info. ratio = 1.5281
    N fixed =     42
      N max =     64

Fixed-study crit. value = 1.9600

Critical values, p-values, and sample sizes for a group sequential design
------------------------------------------------------------------------
     | Info.        Efficacy             Futility        | Sample size
Look | frac.     Upper   p-value     Lower    p-value    |           N
-----+---------------------------------------------------+-------------
   1 |  0.25    2.3613    0.0091    0.2776     0.3906    |          16
   2 |  0.50    2.3613    0.0091    1.1831     0.1184    |          32
   3 |  0.75    2.3613    0.0091    1.8321     0.0335    |          48
   4 |  1.00    2.3613    0.0091    2.3613     0.0091    |          64
------------------------------------------------------------------------
Note: Critical values are for z statistics; otherwise, use p-value
      boundaries.
```

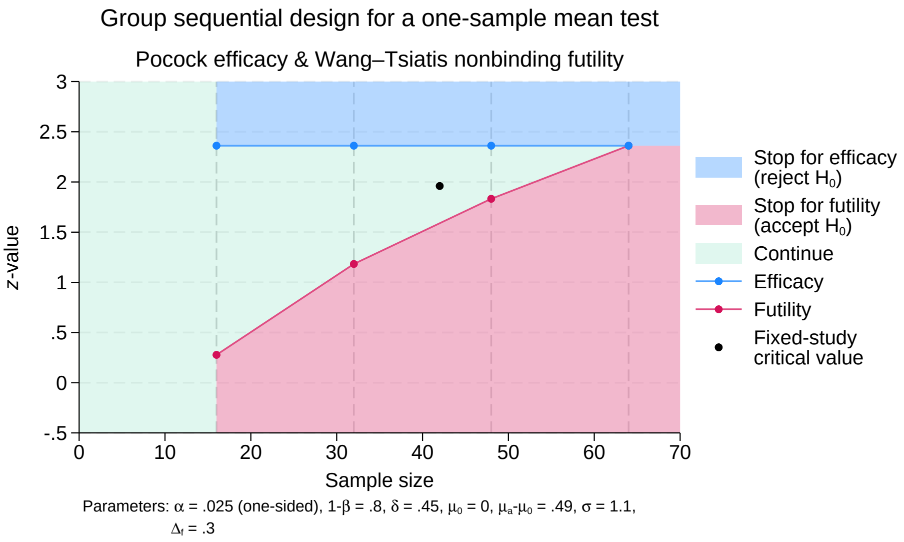


Figure 2. Sample size for a test of one mean with one-sided efficacy and futility bounds

The fixed-study properties are unchanged from example 2, but the information ratio is increased, reflecting the larger maximum sample size of the new GSD. However, the expected sample size under H_0 has decreased dramatically due to the ability to terminate the trial early for futility.

As before, the Pocock efficacy bounds use the same critical z-value of 2.361 for all looks, which corresponds to a p-value of 0.0091. Efficacy critical values are unaffected by the addition of nonbinding futility bounds, but as is demonstrated in example 2 of [ADAPT] **gsdesign**, using binding futility bounds would impact the critical values of efficacy bounds.

As in the previous example, we will conduct t tests at each look, and we will compare the p-values from each t test with the corresponding efficacy and futility critical p-values. The graph, however, uses the standardized z scale on the vertical axis to facilitate comparison with graphs of other stopping boundaries.

The first test is conducted after 16 observations have been collected, and we will let p_1 denote the p-value from the first look. If $p_1 > 0.391$, then the test statistic lies in the acceptance region, and researchers have the option to stop the trial early for futility. If $0.391 \geq p_1 > 0.0091$, then the test statistic falls within the continuation region, and the trial must proceed to the next look. If $p_1 \leq 0.0091$, then it is inside the rejection region and H_0 is rejected, terminating the trial early due to treatment efficacy. If the trial does not stop at the first look, a similar procedure is performed at the second look once 32 observations have been recorded. This time, the futility boundary p-value has decreased to 0.118, shrinking the continuation region and expanding the acceptance region. If the trial does not stop at the second look, the procedure is repeated at the third look with 48 observations, where the futility p-value has decreased to 0.034. If the trial proceeds to the final analysis with 64 observations, there is no continuation region, because the futility bound critical value is identical to the efficacy bound p-value of 0.0091. If $p_4 \leq 0.0091$, then H_0 is rejected; otherwise, H_0 is accepted.

◁

Stored results

gsdesign onemean 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(delta)	effect size
r(diff)	difference between the alternative and null means
r(effparam)	efficacy parameter (if wtsiatis(), kdemets(), or hsdecani() specified)
r(ESS0)	expected sample size under null hypothesis
r(ESS1)	expected sample size under alternative hypothesis
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(knownsd)	1 if option knownsd is specified, 0 otherwise
r(m0)	mean under the null hypothesis
r(ma)	mean under the alternative hypothesis
r(N_fixed)	sample size of a fixed study design
r(N_fixedfrac)	fractional sample size of a fixed study design
r(N_max)	maximum sample size if the study continues to completion
r(nfractional)	1 if nfractional is specified, 0 otherwise
r(nlooks)	number of analyses
r(onesided)	1 for a one-sided test, 0 otherwise
r(pow_converged)	1 if power calculation iteration algorithm converged, 0 otherwise
r(pow_deltax)	final parameter tolerance achieved for power calculation
r(pow_ftolerance)	requested distance of power calculation objective function from 0
r(pow_function)	final distance of power calculation objective function from 0
r(pow_init)	initial value for power calculation sample size
r(pow_iter)	number of iterations performed for power calculation
r(pow_maxiter)	maximum number of iterations for power calculation
r(pow_tolerance)	requested parameter tolerance for power calculation
r(power)	specified overall power
r(power_a)	attained overall power
r(sd)	standard deviation
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)	gsdesign
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
r(method)	onemean

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(bspent_a)	attained cumulative beta spent per look (when futility bounds are specified)
r(design)	sample size and stopping boundaries at interim looks
r(info_frac)	specified information fraction
r(info_frac_a)	fraction of attained information

`r(info_level)`	specified information level
`r(p_crit)`	p-values corresponding to boundary critical values
`r(sampsize)`	sample size at interim looks

Methods and formulas

Sample sizes at interim analyses are calculated as the product of the information fraction, the information ratio, and the sample size of a fixed-sample study.

See *Methods and formulas* in [ADAPT] **gsbounds** for the formulas used to calculate the stopping boundaries, information fraction, and information ratio. See *Methods and formulas* in [PSS-2] **power onemean** for the formulas used to calculate the sample size for a fixed study. See *Methods and formulas* in [ADAPT] **gsdesign** for the formulas used to calculate the expected sample size.

References

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. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70: 659–663. https://doi.org/10.1093/biomet/70.3.659.

Novello, S., G. V. Scagliotti, R. Rosell, M. A. Socinski, J. R. Brahmer, J. N. Atkins, C. Pallares, R. Burgess, L. Tye, P. Selaru, E. Wang, R. C. Chao, and R. Govindan. 2009. Phase II study of continuous daily sunitinib dosing in patients with previously treated advanced non-small cell lung cancer. *British Journal of Cancer* 101: 1543–1548. https://doi.org/10.1038/sj.bjc.6605346.

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.

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.

Shatola, A., K. N. Nguyen, E. Kamangar, and M. E. Daly. 2020. Spontaneous regression of non-small cell lung cancer: A case report and literature review. *Cureus* 12: e6639. https://doi.org/10.7759/cureus.6639.

Simon, R., G. M. Blumenthal, M. L. Rothenberg, J. Sommer, S. A. Roberts, D. K. Armstrong, L. M. LaVange, and R. Pazdur. 2015. The role of nonrandomized trials in the evaluation of oncology drugs. *Clinical Pharmacology and Therapeutics* 97: 502–507. https://doi.org/10.1002/cpt.86.

Tenhunen, O., F. Lasch, A. Schiel, and M. Turpeinen. 2020. Single-arm clinical trials as pivotal evidence for cancer drug approval: A retrospective cohort study of centralized European marketing authorizations between 2010 and 2019. *Clinical Pharmacology and Therapeutics* 108: 653–660. https://doi.org/10.1002/cpt.1965.

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.

Yu, S., X. Wang, X. Wang, X. Wu, R. Xu, X. Wang, X. Zhang, C. Zhang, K. Chen, D. Cheng, and L. Wenfeng. 2019. Tumor shrinkage rate as a potential marker for the prediction of long-term outcome in advanced non-small cell lung cancer treated with first-line tyrosine kinase inhibitors. *Journal of Cancer Research and Therapeutics* 15: 1574–1580. https://doi.org/10.4103/jcrt.jcrt_481_19.

Also see

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

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

[ADAPT] **gsbounds** — Boundaries for group sequential trials

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

[ADAPT] **gsdesign twomeans** — Group sequential design for a two-sample means test

[ADAPT] **Glossary**

[PSS-2] **power onemean** — Power analysis for a one-sample mean test

[R] **ttest** — t tests (mean-comparison tests)

[R] **ztest** — z tests (mean-comparison tests, known variance)

gsdesign twomeans — Group sequential design for a two-sample means test

Description	Quick start	Menu	Syntax	Options
Remarks and examples	Stored results	Methods and formulas	References	Also see

Description

gsdesign twomeans computes stopping boundaries and sample sizes for interim analyses of clinical trials using a two-sample mean test with a group sequential design (GSD). Stopping can be for efficacy, futility, or both. For stopping boundary calculations without sample sizes, see [ADAPT] **gsbounds**. For sample-size calculations for a fixed-sample test of two means, see [PSS-2] **power twomeans**.

Quick start

Sample size and stopping boundaries for a two-sided test of $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$, with default significance level $\alpha = 0.05$ and power of 0.8 to detect the difference between a control-group mean of $m_1 = 3$ and an experimental-group mean of $m_2 = 7$, with shared standard deviation of 9, using default group sequential specifications of O'Brien–Fleming efficacy boundaries with two analyses (one interim, one final)

```
gsdesign twomeans 3 7, sd(9)
```

Same as above, but for an upper one-sided test with $\alpha = 0.025$, and replace the O'Brien–Fleming efficacy bound with an error-spending Pocock-style bound with three looks

```
gsdesign twomeans 3 7, sd(9) alpha(0.025) onesided      ///
    efficacy(errpocock) nlooks(3)
```

Same as above, but add a nonbinding error-spending O'Brien–Fleming-style futility bound, and specify the difference between means instead of the experimental-group mean

```
gsdesign twomeans 3, diff(4) sd(9) alpha(0.025) onesided    ///
    efficacy(errpocock) futility(errobfleming) nlooks(3)
```

Same as above, but specify a control-group standard deviation of 6 and an experimental-group standard deviation of 12, and allocate twice as many subjects to the experimental group as the control group

```
gsdesign twomeans 3, diff(4) sd1(6) sd2(12) nratio(2)     ///
    alpha(0.025) onesided efficacy(errpocock)         ///
    futility(errobfleming) nlooks(3)
```

Same as above, but time the looks to occur with 50%, 75%, and 100% of the data, and plot the boundaries

```
gsdesign twomeans 3, diff(4) sd1(6) sd2(12) nratio(2)     ///
    alpha(0.025) onesided efficacy(errpocock)         ///
    futility(errobfleming) information(50 75 100)
```

Menu

Statistics > Power, precision, and sample size

Syntax

`gsdesign twomeans` m_1 m_2 [, *twomeansopts boundopts*]

where m_1 is the mean of the control (reference) group and m_2 is the mean of the experimental (treatment) group.

twomeansopts	Description
Main	
`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)`
`onesided`	request a one-sided test; default is two-sided
`nfractional`	report fractional sample size
`nratio(#)`	ratio of sample sizes of experimental to control groups; default is `nratio(1)`, meaning equal group sizes
`diff(#)`	difference between the experimental-group mean and the control-group mean, $m_2 - m_1$; specify instead of the experimental-group mean m_2
`sd(#)`	common standard deviation of the control and the experimental groups assuming equal standard deviations in both groups; default is `sd(1)`
`sd1(#)`	standard deviation of the control group; requires `sd2()`
`sd2(#)`	standard deviation of the experimental group; requires `sd1()`
`knownsds`	request computation assuming known standard deviations for both groups; default is to assume unknown standard deviations
`force`	allow calculation with unsupported `power twomeans` options
`poweriteration(`*powiteropts*`)`	iteration options for the calculation of fixed-study sample size; seldom used

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

`force` and `poweriteration()` do not appear in the dialog box.

powiteropts	Description
`init(#)`	initial value for fixed-study sample size
`iterate(#)`	maximum number of iterations; default is `iterate(500)`
`tolerance(#)`	parameter tolerance; default is `tolerance(1e-12)`
`ftolerance(#)`	function tolerance; default is `ftolerance(1e-12)`

boundopts	Description
Bounds	
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(#[, equal])	total number of analyses (nlooks() − 1 interim analyses and one final analysis); use equal to enforce equal information increments; if neither nlooks() nor information() is specified, the default is nlooks(2)
information(*numlist*)	sequence of information levels for analyses; default is evenly spaced
nopvalues	suppress *p*-values
Graph	
graphbounds[(*graphopts*)]	graph boundaries
matlistopts(*general_options*)	control the display of boundaries and sample size; seldom used
optimopts	optimization options for boundary calculations; seldom used

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

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

graphopts	Description
xdimsampsize	label the x axis with the sample size collected (default)
xdiminformation	label the x axis with the information fraction; 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

alpha(#) sets the overall significance level, which is the familywise type I error rate for all analyses (interim and final). alpha() must be in $(0, 0.5)$. The default is alpha(0.05).

power(#) sets the overall power for all analyses. power() must be in $(0.5, 1)$. 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. beta() must be in $(0, 0.5)$. 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.

onesided requests a study design for a one-sided test. The direction of the test is inferred from the effect size.

nfractional specifies that fractional sample sizes be reported.

nratio(#) specifies the sample-size ratio of the experimental group relative to the control group, $N2/N1$. The default is nratio(1), meaning equal allocation between the two groups.

diff(#) specifies the difference between the experimental-group mean and the control-group mean, $m_2 - m_1$. You can either specify the experimental-group mean m_2 as a command argument or specify the difference between the two means in diff(). If you specify diff(#), the experimental-group mean is computed as $m_2 = m_1 + \#$.

sd(#) specifies the common standard deviation of the control and the experimental groups assuming equal standard deviations in both groups. The default is sd(1).

sd1(#) specifies the standard deviation of the control group. If you specify sd1(), you must also specify sd2().

sd2(#) specifies the standard deviation of the experimental group. If you specify sd2(), you must also specify sd1().

knownsds requests that standard deviations of each group be treated as known in the computations. By default, standard deviations are treated as unknown and the computations are based on a two-sample t test, which uses Student's t distribution as a sampling distribution of the test statistic. If knownsds is specified, the computation is based on a two-sample z test, which uses a normal distribution as the sampling distribution of the test statistic. In either case, critical values for efficacy and futility boundaries calculated by gsdesign twomeans are reported on the standardized z scale. When a t test is performed, you can use the significance level approach and compare the p-value from the t test to the p-value boundaries reported by gsdesign twomeans, as demonstrated in example 2.

Bounds

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(`#[`, equal`]`)` 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)`.

`equal` indicates that equal information increments be enforced, which is to say that the same number of new observations will be collected at each look. The default behavior is to start by dividing information evenly among looks, then proceed by rounding up to a whole number of observations at each look. This can cause slight differences in the information collected at each look.

`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.

Graph

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

graphopts are the following:

`xdimsampsize` labels the x axis with the sample size collected (the default).

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

`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 gsdesign twomeans but are not shown in the dialog box:

force indicates that gsdesign twomeans should allow unsupported power twomeans options, such as options specifying a cluster randomized design. Even with option force, the power twomeans options specified must be compatible with sample-size determination, not effect size or power calculation. In addition, *numlist*s are not supported in options or in arguments as they are with power, even when force is specified.

poweriteration(*powiteropts*) controls the iterative algorithm used to calculate the fixed-study sample size. This is seldom used.

powiteropts are the following:

init(#) specifies an initial value for the sample size when iteration is used to compute the fixed-study sample size. The default is to use a closed-form normal approximation to compute an initial sample size.

iterate(#) specifies the maximum number of iterations for the Newton method during calculation of the fixed-study sample size. The default is iterate(500).

tolerance(#) specifies the tolerance used to determine whether successive parameter estimates have converged when calculating the fixed-study sample size. The default is tolerance(1e-12). See *Convergence criteria* in [M-5] **solvenl( )** for details.

ftolerance(#) specifies the tolerance used when calculating the fixed-study sample size to determine whether the proposed solution of a nonlinear equation is sufficiently close to 0 based on the squared Euclidean distance. The default is ftolerance(1e-12). See *Convergence criteria* in [M-5] **solvenl( )** for details.

matlistopts(*general_options*) affects the display of the matrix of boundaries and sample sizes. *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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**. 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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

`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:

Introduction
Using gsdesign twomeans
Background for examples 1 and 2
Computing sample size and stopping boundaries with known standard deviation
Unknown standard deviation and hypothesis tests on means
Background for example 3
Efficacy and futility stopping

This entry describes the use of the gsdesign twomeans command for designing a group sequential analysis for a two-sample means test. See [ADAPT] **GSD intro** for a general introduction to GSDs for clinical trials; see [ADAPT] **gsbounds** for information about group sequential bounds; and see [ADAPT] **gsdesign** for information about designing group sequential clinical trials with the gsdesign command. Also see [PSS-2] **Intro (power)** for a general introduction to power and sample-size analysis, and see [PSS-2] **power twomeans** for details about study design for a two-sample means test.

Introduction

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.

gsdesign twomeans calculates sample size and stopping boundaries for a group sequential trial comparing the population mean of the experimental group against that of the control group. Specifically, we consider the null hypothesis $H_0 : \mu_1 = \mu_2$ versus the two-sided alternative hypothesis $H_a : \mu_1 \neq \mu_2$, the upper one-sided alternative $H_a : \mu_1 > \mu_2$, or the lower one-sided alternative $H_a : \mu_1 < \mu_2$.

The actual test conducted will depend on whether the population standard deviation of both groups is known. In the case of a known standard deviation, the test statistic follows a standard normal distribution under the null hypothesis, and the corresponding test is known as a two-sample z test. A z test is also commonly used when sample sizes are large, even when the population standard deviations are unknown. This is because the distribution of the test statistic approaches a normal distribution as the sample size increases.

If the sample is not of sufficient size to use a large-sample z test and the standard deviations are unknown but assumed to be equal, then the test statistic has an exact Student's t distribution under the null hypothesis and the corresponding test is referred to as a two-sample t test. If the two unknown standard deviations are not equal, then the distribution of the test statistic under the null hypothesis can be approximated by a t distribution with degrees of freedom estimated using Satterthwaite's method, and the resulting test is known as Satterthwaite's t test.

The required sample size estimated by gsdesign twomeans will depend on whether the standard deviation is known, but the stopping boundaries will not; they are reported on a standardized z scale. The critical values from the boundaries may be compared directly with the z statistic from a z test. If the analysis is performed using a t test, the p-value from the t test can be compared with the p-values corresponding to the critical values for the boundaries. This is demonstrated in example 2.

Using gsdesign twomeans

gsdesign twomeans calculates sample size and stopping boundaries for a group sequential trial comparing the means of two populations. gsdesign twomeans can be thought of as a combination of power twomeans for sample-size calculations and gsbounds for stopping boundary calculations.

To compute sample size, you must specify the effect size. There are two ways to do this: by specifying the means of the control and experimental groups, m_1 and m_2, or by specifying m_1 and the difference $m_2 - m_1$ in the diff() option. There is no default value for diff(), so either m_1 and m_2 or m_1 and diff() must be included as part of the command specification. Another aspect of the effect size is the standard deviation of the responses. This is specified with the sd() option if both groups share a common standard deviation and specified with the sd1() and sd2() options otherwise. The default behavior is to assume a common standard deviation of 1 and to assume that the standard deviation must be estimated from the sample. If the true population standard deviation is known a priori, the knownsds option requests that sample-size calculations be performed for a z test, not a t test.

By default, gsdesign twomeans assumes that the control and experimental arms will be the same size. If participants are not allocated equally between the two arms, the nratio() option is used to specify the ratio of participants in the experimental arm to the control arm.

The alpha(), power(), beta(), and onesided options are used for both sample-size and stopping-boundary calculations. The default significance level, known as the familywise type I error rate, is 0.05 and can be changed by specifying the alpha() option. The default power is 0.8, which corresponds to a type II error rate of 0.2. This can be modified either by specifying the power in the power() option or by specifying the type II error in the beta() option. The default test is two-sided, and the onesided option requests a one-sided test, the direction of which is indicated by the sign of the effect size.

The group sequential stopping rule is determined by the efficacy() and futility() options. Stopping can be for efficacy, futility, or both, and if no stopping rule is specified, the default is to use an O'Brien–Fleming efficacy bound. If futility bounds are requested, the default behavior is to treat them as nonbinding. A trial that crosses a nonbinding futility bound can be stopped for futility, but the familywise type I error is controlled even if the trial continues. Binding futility bounds can be requested with futility() suboption binding. A trial that crosses a binding futility bound must be stopped for futility; if it continues, the familywise type I error will not be controlled at the specified significance level.

The number of looks, or analyses of the trial data, is specified with nlooks(). Alternatively, the information() option can be used to specify the spacing of the looks as a *numlist* of increasing information levels. In this case, values of the numlist are automatically rescaled so that the final look has the maximum information required by the design. If neither nlooks() nor information() is specified, the default is two looks.

By default, the sample sizes in each arm are rounded up to whole numbers at each look, but the nfractional option can be used to report fractional sample sizes. If nlooks() is specified, the default behavior is to divide information evenly among looks before rounding. Rounding can cause slight differences in the amount of information collected at each look, and nlooks() suboption equal can be specified to enforce equal information increments by requiring the same number of new observations per arm at each look.

Background for examples 1 and 2

Alzheimer's disease (AD) is an incurable neurodegenerative disease characterized by memory loss and progressive cognitive decline. Historically, the only sure way to diagnose AD was through autopsy, but recent research has identified biomarkers that can be used to diagnose AD and track disease progression in living patients.

One of the most promising biomarkers for AD is glucose metabolism in the brain, which can be measured by a type of imaging known as fluorodeoxyglucose positron emission tomography (FDG PET). Mosconi (2005) writes that FDG PET "has revealed glucose metabolic reductions in the parieto-temporal, frontal and posterior cingulate cortices to be the hallmark of AD". FDG PET measures glucose metabolism as standardized uptake value ratios (SUVRs), which can be used to track disease progression, with SUVR levels falling as the disease becomes more severe.

Matthews et al. (2021) conducted a phase 2 clinical trial of the neuroprotective agent riluzole versus a placebo for the treatment of mild AD. They used FDG PET to measure the SUVR of each subject at baseline and again after six months of treatment, and they compared the average change in SUVR in the control arm against that of the treatment arm. The results of their study were encouraging, with smaller declines in SUVR observed in the experimental arm than in the control arm.

Computing sample size and stopping boundaries with known standard deviation

▷ Example 1: Pocock efficacy bounds for a test of two sample means

Suppose that we want design a follow-up study focusing on a target population of Alzheimer's patients suffering from clinical depression. We will consider a placebo-controlled clinical trial, with participants randomized to the treatment and placebo arms at a 1:1 ratio. The SUVR in the posterior cingulate will be measured for each participant at baseline and again after six months of treatment, and the mean change in SUVR from the control arm (μ_1) and experimental arm (μ_2) will be calculated. We will test the null hypothesis $H_0: \mu_1 = \mu_2$ versus the two-sided alternative hypothesis $H_a: \mu_1 \neq \mu_2$.

Based on previous studies, we anticipate SUVR will decrease by an average of 0.05 in the control arm and 0.01 in the experimental arm, giving $\mu_1 = -0.05$ and $\mu_2 = -0.01$. We will assume both arms have a known standard deviation of 0.035, but this assumption is likely unrealistic and is relaxed in the next example.

We require 80% power to detect the specified difference in means, and we will conduct a two-sided trial with familywise significance level of 5%, using Pocock efficacy bounds with two evenly spaced looks. Except for the efficacy boundary, these design specifications correspond to the default values of the respective options in `gsdesign twomeans`, so they are not specified.

```
. gsdesign twomeans -0.05 -0.01, sd(0.035) knownsds efficacy(pocock)

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

Efficacy: Pocock

Study parameters:
      alpha =  0.0500  (two-sided)
      power =  0.8000
      delta =  0.0400
         m1 = -0.0500
         m2 = -0.0100
         sd =  0.0350

Expected sample size:
         H0 =   27.59
         Ha =   21.22

Info. ratio =  1.1104
    N fixed =      26
      N max =      28
     N1 max =      14
     N2 max =      14

Fixed-study crit. values = ±1.9600

Critical values, p-values, and sample sizes for a group sequential design
---------------------------------------------------------------------------
      |  Info.            Efficacy         |           Sample size
 Look |  frac.     Lower     Upper  p-value |      N1         N2          N
------+------------------------------------+--------------------------------
    1 |   0.50   -2.1783    2.1783   0.0294 |       7          7         14
    2 |   1.00   -2.1783    2.1783   0.0294 |      14         14         28
---------------------------------------------------------------------------

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

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

The next section of the output displays the expected sample size, which is the average sample size if the group sequential trial were to be repeated many times. The following section reports 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 can compare the expected sample sizes with the sample size for a fixed study and the maximum sample size of the GSD. If the null hypothesis of equal change in SUVR between the control and the experimental arms were true, the average trial would require 27.59 participants, nearly the full sample size of 28. This is because the efficacy bounds do not allow for early stopping to accept H_0, so if the null hypothesis is true, the trial will usually proceed to the final look. If H_a is true, the average trial will require 21.22 participants, which is a savings over the 26 participants required by the fixed trial.

We also see the critical value for a fixed study with an equivalent significance level. The critical values of ±1.96 would be used to reject H_0 at the 0.05 level if a fixed study design were conducted instead of a GSD.

Finally, `gsdesign twomeans` displays a table with the critical values and p-values for the efficacy stopping boundaries as well as the sample sizes at each look. Pocock efficacy bounds use the same critical value at all looks, and to maintain a familywise type I error of 0.05, the z statistic must meet or exceed ±2.178 at any look to reject H_0.

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

```
. gsdesign twomeans -0.05 -0.01, sd(0.035) knownsds efficacy(pocock) graphbounds
  (output omitted)
```

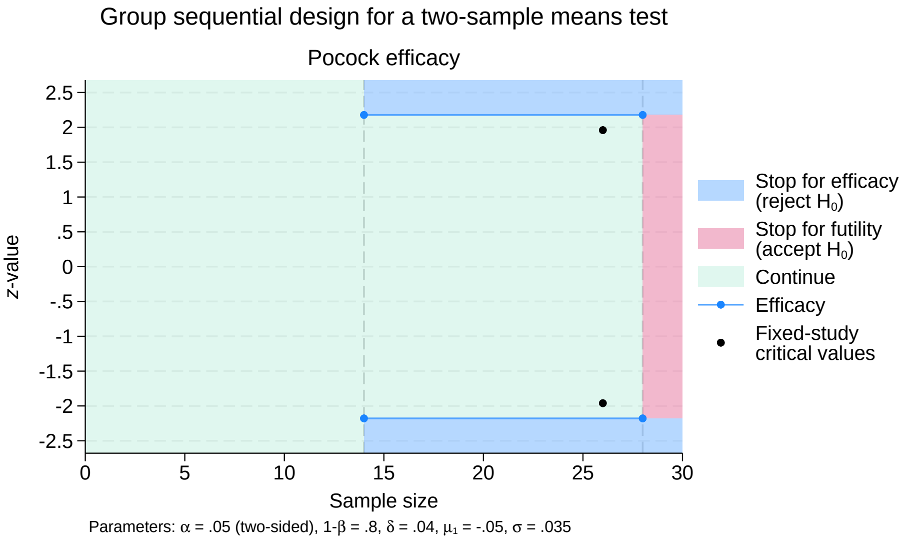


Figure 1. Two-sided Pocock efficacy bounds for a test of the equality of two means

On the graph, the horizontal axis is the total sample size (the sum of the sample sizes of both arms), and the vertical axis is the z-value of the test statistic. The efficacy bounds are marked as blue lines, with the location of looks indicated by blue dots. For comparison, the critical values of an equivalently powered fixed study are marked with black dots.

The rejection region is shaded blue, the acceptance region red, and the continuation region green. Before the first look, which occurs once results have been collected from 14 participants (7 in the control arm and 7 in the experimental arm), it is impossible to reject H_0 because no test has been conducted, so the entire range of z-values is in the continuation region. Beginning at the first look, z-values equal to or more extreme than ± 2.1783 are in the rejection region. The efficacy-only design does not permit early stopping to accept H_0, so the acceptance region begins at the second and final look, and it encompasses z-values less extreme than ± 2.1783.

◁

Unknown standard deviation and hypothesis tests on means

▷ Example 2: Unknown standard deviation, specifying difference between means

In the previous example, we relied on the assumption that the population standard deviation was known to be 0.035 in both arms, which led to sample-size calculations based on a two-sample z test. Here we relax that assumption and assume that the standard deviation will be estimated from the sample. We anticipate the standard deviation of the control group will be 0.05, while the standard deviation of the experimental group will be 0.035. This yields sample sizes for a t test, which is demonstrated below.

Additionally, instead of specifying μ_2 directly, here we use the `diff()` option to specify the difference in means between the two arms. We omit the `graphbounds` option because the graph is minimally changed from the previous example.

```
. gsdesign twomeans -0.05, diff(0.04) sd1(0.05) sd2(0.035) efficacy(pocock)

Group sequential design for a two-sample means test
Satterthwaite's t test assuming unequal variances
H0: m2 = m1 versus Ha: m2 != m1

Efficacy: Pocock

Study parameters:
      alpha =  0.0500  (two-sided)
      power =  0.8000
      delta =  0.0400
         m1 = -0.0500
         m2 = -0.0100
       diff =  0.0400
        sd1 =  0.0500
        sd2 =  0.0350

Expected sample size:
         H0 =   43.35
         Ha =   33.60

Info. ratio =  1.1104
    N fixed =      40
      N max =      44
     N1 max =      22
     N2 max =      22

Fixed-study crit. values = ±1.9600

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

      Info.            Efficacy                    Sample size
 Look  frac.      Lower     Upper   p-value       N1        N2          N

    1   0.50    -2.1783    2.1783    0.0294       11        11         22
    2   1.00    -2.1783    2.1783    0.0294       22        22         44

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

Specifying the difference in means instead of the experimental group mean has not changed the study parameters, but changing our assumptions about the standard deviations has increased the fixed-study sample size from 26 in the previous example to 40 here. The information ratio is unchanged, but the sample sizes required by the GSD have increased correspondingly, as have the expected sample sizes under H_0 and H_a.

The testing procedure has also changed. Instead of comparing the z statistic directly with the efficacy critical values, a t test is performed, and we use the significance level approach described in [ADAPT] **gsbounds**. The table at the bottom of the output provides the p-value corresponding to each critical value. We can compare the p-value for the t test with these p-value boundaries.

Suppose the first look is conducted with 11 observations from each arm. From the data we collect, we have a mean change in SUVR of −0.014 in the experimental arm with standard deviation 0.038 and a mean change in SUVR of −0.062 in the control arm with standard deviation 0.057. We conduct a t test using `ttesti`, the immediate form of the [R] **ttest** command. We type `ttesti 11 -0.014 0.038 11 -0.062 0.057, unequal`, with the first three arguments specifying the experimental group sample size, mean, and standard deviation, respectively, and the following three arguments specifying the control group sample size, mean, and standard deviation. Option `unequal` indicates that we do not assume that the population standard deviations of the two groups are equal and instructs `ttesti` to use Satterthwaite's method to estimate the degrees of freedom for the t test.

```
. ttesti 11 -0.014 0.038 11 -0.062 0.057, unequal

Two-sample t test with unequal variances
─────────────────────────────────────────────────────────────────────────────
         │     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
─────────┼───────────────────────────────────────────────────────────────────
       x │      11       -.014    .0114574        .038   -.0395287    .0115287
       y │      11       -.062    .0171861        .057   -.1002931   -.0237069
─────────┼───────────────────────────────────────────────────────────────────
Combined │      22       -.038    .0113582    .0532747   -.0616207   -.0143793
─────────┼───────────────────────────────────────────────────────────────────
    diff │                .048    .0206552                .0045018    .0914982
─────────────────────────────────────────────────────────────────────────────
    diff = mean(x) - mean(y)                                      t =   2.3239
H0: diff = 0                     Satterthwaite's degrees of freedom =  17.4227

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9838         Pr(|T| > |t|) = 0.0325          Pr(T > t) = 0.0162
```

The t statistic of 2.324 cannot be compared directly with the efficacy critical values, but the p-value for the two-sided test, $p_1 = 0.0325$, can be compared with the p-value equivalent of the critical value at the first look. Because $p_1 > 0.0294$, we cannot reject H_0 at this look, and the trial continues to the second and final look. At the second look, we conduct another t test and calculate the p-value, p_2. If $p_2 \leq 0.0294$, then H_0 is rejected; otherwise, H_0 is accepted.

◁

Background for example 3

Wilkinson et al. (2011) published the results of a clinical trial of levalbuterol used as a continuous nebulization for the treatment of acute pediatric asthma exacerbations (asthma attacks). Levalbuterol was found to be inferior to the standard of care, nebulization with racemic albuterol, which was an unexpected result. To determine whether levalbuterol is more effective at higher doses, we wish to conduct a similar study using 7.5 mg of levalbuterol instead of the 3.75 mg dose used by Wilkinson et al.

Study participants are children aged 6 to 17 who have previously been diagnosed with asthma by a physician and who present to the emergency department with acute asthma exacerbation of moderate severity. Participants are randomly assigned to either the treatment or the control group. Participants in the treatment group receive 7.5 mg of levalbuterol administered via nebulizer over the course of one hour, while participants in the control group receive the standard of care, which is a one-hour nebulization with 7.5 mg of racemic albuterol.

Upon hospital admission, each participant's one-second forced expiratory volume is assessed. This is a measurement of how much air the participant can exhale in one second, and higher values indicate better lung function. A second measurement of expiratory volume is conducted two hours after treatment, and the change in one-second forced expiratory volume (ΔFEV1) is calculated as the percent improvement (or percent decline, for negative ΔFEV1) compared with the participant's baseline value.

Efficacy and futility stopping

▷ Example 3: Error-spending efficacy and futility bounds

Suppose we wish to design a clinical trial that will compare the average ΔFEV1 in the control arm, μ_1, against the average ΔFEV1 in the experimental arm, μ_2. We will test the null hypothesis $H_0 : \mu_1 = \mu_2$ versus the one-sided alternative $H_a : \mu_1 < \mu_2$ with a familywise significance level of 2.5%. We require

90% power to detect the difference between a 50% increase in mean ΔFEV1 in the control arm and a 60% mean increase in the experimental arm, with a common standard deviation of 35. Suppose that we are particularly concerned about adverse events in the group receiving high-dose levalbuterol, so we will randomize participants to the experimental and control arms in a 2:1 ratio, ensuring a larger sample size (and more power) to detect adverse events in the experimental arm.

Depending on the recruitment rate, this clinical trial could take months or even years to complete. There is an ethical imperative not to expose participants to inferior treatments, so if high-dose levalbuterol is more effective than racemic albuterol, we would want to know as soon as possible. To this end, we employ an error-spending O'Brien–Fleming-style efficacy bound. If high-dose levalbuterol is not superior to racemic albuterol, we want to terminate the trial early for futility, so we also specify a nonbinding Kim–DeMets futility bound with parameter $\rho_f = 2$. If a nonbinding futility bound is crossed, the trial can be stopped for futility, but if the trial is continued, the familywise type I error is still controlled at the desired level. We specify a four-look design and graph the bounds for inspection.

```
. gsdesign twomeans 50 60, sd(35) nratio(2) alpha(0.025) power(0.9) onesided
> efficacy(errobfleming) futility(kdemets(2)) nlooks(4) graphbounds

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

Efficacy: Error-spending O'Brien-Fleming style
Futility: Error-spending Kim-DeMets, nonbinding, rho = 2.0000

Study parameters:
      alpha =  0.0250  (upper one-sided)
      power =  0.9000
     nratio =  2.0000
      delta = 10.0000
         m1 = 50.0000
         m2 = 60.0000
         sd = 35.0000

Expected sample size:
         H0 =  353.85
         Ha =  470.42

Info. ratio =  1.0859
    N fixed =     582
      N max =     632
     N1 max =     211
     N2 max =     421

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 |   0.25     4.3326    0.0000    -0.8088    0.7907
    2 |   0.50     2.9631    0.0015     0.3702    0.3556
    3 |   0.75     2.3590    0.0092     1.2438    0.1068
    4 |   1.00     2.0141    0.0220     2.0141    0.0220
------+------------------------------------------------

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

```
----------------------------------------------
         |           Sample size
    Look |       N1          N2              N
---------+------------------------------------
       1 |       53         106            159
       2 |      106         211            317
       3 |      158         316            474
       4 |      211         421            632
----------------------------------------------
```

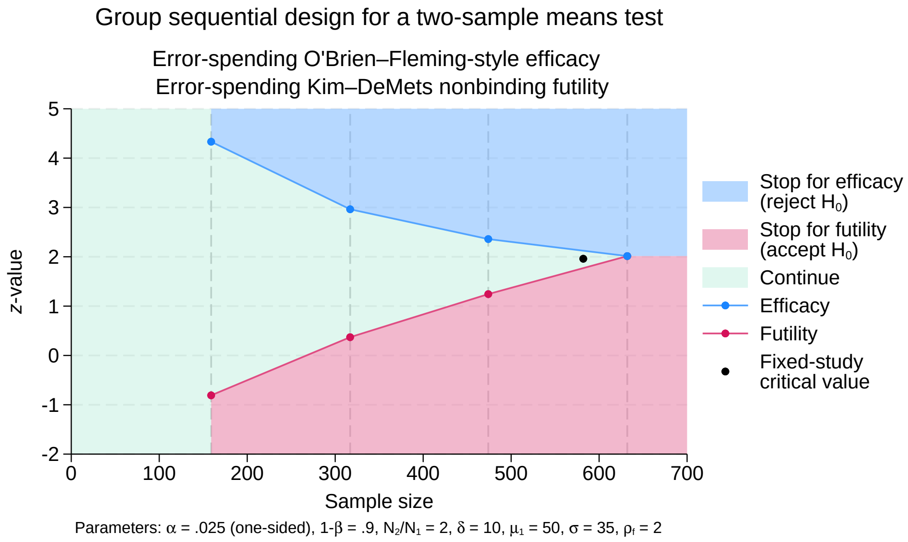


Figure 2. Sample size for a test of the equality of two means with efficacy and futility bounds

The output begins with a description of the test being performed, the types of boundaries, and a summary of the parameters used in the design.

Next it displays the expected sample sizes under the null and alternative hypotheses, the information ratio, the sample size that would be required for an equivalently powered fixed study, the maximum sample size for a GSD, and the critical value for a fixed study.

The sample size for a corresponding fixed study is the same sample size that would be calculated had we run `power twomeans 50 60, sd(35) nratio(2) power(0.9) alpha(0.025) onesided`. The fixed-study critical value of 1.96 would be used to reject H_0 at the 0.025 level using a fixed study design.

The expect sample sizes under the null and alternative hypotheses are both smaller than the fixed-study sample size. This is not surprising, because this design incorporates both efficacy and futility stopping.

At the bottom of the output is a table with the critical values and p-values for the stopping boundaries as well as the sample sizes at each look. O'Brien–Fleming boundaries are very conservative at early looks, with final critical values only slightly larger than those of an equivalent fixed-sample design. The error-spending approximation of the classical O'Brien–Fleming bounds shares this property, yielding efficacy critical values at the first look of 4.333, but only 2.014 at the final look, a minor increase over the fixed-study critical values.

While the population standard deviation was not assumed to be known when designing this trial, the large sample sizes involved enable the use of a large-sample z test. The first look is conducted when we have data from 53 controls and 106 experimental participants, and the test statistic, z_1, is compared with the boundary critical values. If $z_1 \geq 4.333$, we reject H_0 and terminate the trial early for efficacy. Even if we terminate the trial after the first look, we will have data about adverse events for over 100 experimental participants because we randomized twice as many participants to the experimental arm as the control arm. If $z_1 < -0.809$, we may accept H_0 and terminate the trial for futility, but if the trial is continued, the familywise type I error is still controlled. If $z_1 \in [-0.809, 4.333)$, then the trial must continue to the next look.

At the second look, the testing procedure is the same, but the critical values of the efficacy bounds and the futility bounds have narrowed, shrinking the continuation region to $z_2 \in [0.37, 2.963)$. If the trial continues to the third look, the continuation region is further reduced to $z_3 \in [1.244, 2.359)$. At the fourth and final look, the futility critical values equal the efficacy critical values and there is no continuation region: If $z_4 \geq 2.014$, then H_0 is rejected; otherwise, H_0 is accepted. The boundaries are displayed on the graph, and the critical value for a fixed study with equivalent significance level and power is marked with a black dot.

◁

Stored results

`gsdesign twomeans` 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(delta)`	effect size
`r(diff)`	difference between the experimental- and control-group means
`r(effparam)`	efficacy parameter (if `wtsiatis()`, `kdemets()`, or `hsdecani()` specified)
`r(ESS0)`	expected sample size under null hypothesis
`r(ESS1)`	expected sample size under alternative hypothesis
`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(knownsds)`	1 if option `knownsds` is specified, 0 otherwise
`r(m1)`	control-group mean
`r(m2)`	experimental-group mean
`r(N_fixed)`	sample size of a fixed study design
`r(N_fixedfrac)`	fractional sample size of a fixed study design
`r(N_max)`	maximum sample size if the study continues to completion
`r(N1_fixed)`	sample size of the control group in a fixed study design
`r(N1_fixedfrac)`	fractional sample size of the control group in a fixed study design
`r(N1_max)`	maximum sample size of the control group if the study continues to completion
`r(N2_fixed)`	sample size of the experimental group in a fixed study design
`r(N2_fixedfrac)`	fractional sample size of the experimental group in a fixed study design
`r(N2_max)`	maximum sample size of the experimental group if the study continues to completion
`r(nfractional)`	1 if `nfractional` is specified, 0 otherwise
`r(nlooks)`	number of analyses
`r(nratio)`	specified ratio of sample sizes, $N2/N1$
`r(nratio_a)`	attained ratio of sample sizes
`r(onesided)`	1 for a one-sided test, 0 otherwise
`r(pow_converged)`	1 if power calculation iteration algorithm converged, 0 otherwise
`r(pow_deltax)`	final parameter tolerance achieved for power calculation
`r(pow_ftolerance)`	requested distance of power calculation objective function from 0
`r(pow_function)`	final distance of power calculation objective function from 0

r(pow_init)	initial value for power calculation sample size
r(pow_iter)	number of iterations performed for power calculation
r(pow_maxiter)	maximum number of iterations for power calculation
r(pow_tolerance)	requested parameter tolerance for power calculation
r(power)	specified overall power
r(power_a)	attained overall power
r(sd)	common standard deviation of both groups (if sd1() and sd2() not specified)
r(sd1)	standard deviation of the control group
r(sd2)	standard deviation of the experimental group
r(stop)	0 for futility bounds, 1 for efficacy bounds, 2 for both
r(unequal)	0 if sd1 = sd2, 1 otherwise
r(z_fixed)	critical value for an equivalent fixed study design

Macros

r(cmd)	gsdesign
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
r(method)	twomeans

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(bspent_a)	attained cumulative beta spent per look (when futility bounds are specified)
r(design)	sample size and stopping boundaries at interim looks
r(info_frac)	specified information fraction
r(info_frac_a)	fraction of attained information
r(info_level)	specified information level
r(p_crit)	*p*-values corresponding to boundary critical values
r(sampsize)	sample size at interim looks

Methods and formulas

Sample sizes at interim analyses are calculated as the product of the information fraction, the information ratio, and the sample size of a fixed-sample study.

See *Methods and formulas* in [ADAPT] **gsbounds** for the formulas used to calculate the stopping boundaries, information fraction, and information ratio. See *Methods and formulas* in [PSS-2] **power twomeans** for the formulas used to calculate the sample size for a fixed study. See *Methods and formulas* in [ADAPT] **gsdesign** for the formulas used to calculate the expected sample size.

References

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. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70: 659–663. https://doi.org/10.1093/biomet/70.3.659.

Matthews, D. C., X. Mao, K. Dowd, D. Tsakanikas, C. S. Jiang, C. Meuser, R. D. Andrews, A. S. Lukic, J. Lee, N. Hampilos, N. Shafiian, M. Sano, P. D. Mozley, H. Fillit, B. S. McEwen, D. C. Shungu, and A. C. Pereira. 2021. Riluzole, a glutamate modulator, slows cerebral glucose metabolism decline in patients with Alzheimer's disease. *Brain* 144: 3742–3755. https://doi.org/10.1093/brain/awab222.

Mosconi, L. 2005. Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. *European Journal of Nuclear Medicine and Molecular Imaging* 32: 486–510. https://doi.org/10.1007/s00259-005-1762-7.

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.

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.

Wilkinson, M., B. Bulloch, P. Garcia-Filion, and L. Keahey. 2011. Efficacy of racemic albuterol versus levalbuterol used as a continuous nebulization for the treatment of acute asthma exacerbations: A randomized, double-blind, clinical trial. *Journal of Asthma* 48: 188–193. https://doi.org/10.3109/02770903.2011.554939.

Also see

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

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

[ADAPT] **gsbounds** — Boundaries for group sequential trials

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

[ADAPT] **gsdesign onemean** — Group sequential design for a one-sample mean test

[ADAPT] **Glossary**

[PSS-2] **power twomeans** — Power analysis for a two-sample means test

[R] **ttest** — t tests (mean-comparison tests)

[R] **ztest** — z tests (mean-comparison tests, known variance)

gsdesign oneproportion — Group sequential design for a one-sample proportion test

Description Quick start Menu Syntax Options
Remarks and examples Stored results Methods and formulas References Also see

Description

gsdesign oneproportion computes stopping boundaries and sample sizes for interim analyses of clinical trials using a one-sample proportion test with a group sequential design (GSD). Stopping can be for efficacy, futility, or both. For stopping boundary calculations without sample sizes, see [ADAPT] **gsbounds**. For sample-size calculations for a fixed-sample test of one proportion, see [PSS-2] **power oneproportion**.

Quick start

Sample size and stopping boundaries for a two-sided score test of $H_0: \pi = 0.3$ versus $H_a: \pi \neq 0.3$, with default significance level $\alpha = 0.05$ and power of 0.8 to detect the difference between the proportion under the null hypothesis, $p_0 = 0.3$, and an observed proportion of $p_a = 0.4$, using default group sequential specifications of O'Brien–Fleming efficacy boundaries with two analyses (one interim, one final)

```
gsdesign oneproportion 0.3 0.4
```

Same as above, but use Hwang–Shih–de Cani error-spending efficacy bounds with parameter $\gamma = -2$ and four looks

```
gsdesign oneproportion 0.3 0.4, efficacy(hsdecani(-2)) nlooks(4)
```

Same as above, but specified as p_0 and difference $p_a - p_0 = 0.1$

```
gsdesign oneproportion 0.3, diff(0.1) efficacy(hsdecani(-2)) nlooks(4)
```

Same as above, but add nonbinding O'Brien–Fleming-style futility bounds and graph the boundaries

```
gsdesign oneproportion 0.3, diff(0.1) efficacy(hsdecani(-2))     ///
    futility(errobfleming) nlooks(4) graphbounds
```

Same as above, but time the looks to occur with 40%, 60%, 80%, and 100% of the data

```
gsdesign oneproportion 0.3, diff(0.1) efficacy(hsdecani(-2))     ///
    futility(errobfleming) information(4 6 8 10) graphbounds
```

Same as above, but remove the efficacy bound and make the futility bound binding

```
gsdesign oneproportion 0.3, diff(0.1) futility(errobfleming, binding) ///
    information(4 6 8 10) graphbounds
```

Menu

Statistics > Power, precision, and sample size

Syntax

`gsdesign oneproportion` p_0 p_a [, *onepropopts boundopts*]

where p_0 is the null (hypothesized) proportion or the value of the proportion under the null hypothesis, and p_a is the alternative (target) proportion or the value of the proportion under the alternative hypothesis.

onepropopts	Description
Main	
`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)`
`onesided`	request a one-sided test; default is two-sided
`nfractional`	report fractional sample size
`diff(#)`	difference between the alternative proportion and the null proportion, $p_a - p_0$; specify instead of the alternative proportion p_a
`test(`*test*`)`	specify the type of test; options are `score` (the default) and `wald`
`continuity`	apply continuity correction to the normal approximation of the discrete distribution
`force`	allow calculation with unsupported `power oneproportion` options
`poweriteration(`*powiteropts*`)`	iteration options for the calculation of fixed-study sample size; seldom used

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

`force` and `poweriteration()` do not appear in the dialog box.

powiteropts	Description
`init(#)`	initial value for fixed-study sample size
`iterate(#)`	maximum number of iterations; default is `iterate(500)`
`tolerance(#)`	parameter tolerance; default is `tolerance(1e-12)`
`ftolerance(#)`	function tolerance; default is `ftolerance(1e-12)`

boundopts	Description
Bounds	
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(#[, equal])	total number of analyses (nlooks() − 1 interim analyses and one final analysis); use equal to enforce equal information increments; if neither nlooks() nor information() is specified, the default is nlooks(2)
information(*numlist*)	sequence of information levels for analyses; default is evenly spaced
nopvalues	suppress p-values
Graph	
graphbounds[(*graphopts*)]	graph boundaries
matlistopts(*general_options*)	control the display of boundaries and sample size; seldom used
optimopts	optimization options for boundary calculations; seldom used

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

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

graphopts	Description
xdimsampsize	label the x axis with the sample size collected (default)
xdiminformation	label the x axis with the information fraction; 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

`alpha(#)` sets the overall significance level, which is the familywise type I error rate for all analyses (interim and final). `alpha()` must be in $(0, 0.5)$. The default is `alpha(0.05)`.

`power(#)` sets the overall power for all analyses. `power()` must be in $(0.5, 1)$. 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. `beta()` must be in $(0, 0.5)$. 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.

`onesided` requests a study design for a one-sided test. The direction of the test is inferred from the effect size.

`nfractional` specifies that fractional sample sizes be reported.

`diff(#)` specifies the difference between the alternative proportion and the null proportion, $p_a - p_0$. You can either specify the alternative proportion p_a as a command argument or specify the difference between the two proportions in `diff()`. If you specify `diff(#)`, the alternative proportion is computed as $p_a = p_0 + \#$.

`test(`*test*`)` specifies the type of test that will be used for data analysis. Sample-size calculations depend on the test that will be conducted. *test* is either `score` or `wald`.

`score` requests computations for the score test, which uses the value of the null proportion in the formula for the standard error of the estimator of the proportion. This is the default test, and this test can be performed with command `prtest`; see [R] **prtest**.

`wald` requests computations for the Wald test, which uses the value of the alternative proportion in the formula for the standard error of the estimator of the proportion.

Note that `power oneproportion` option `test(binomial)` cannot be used for sample-size calculations and is not compatible with `gsdesign oneproportion`. However, option `continuity` implements a continuity correction that yields an estimate of the sample size that would be required by the exact binomial test at the specified significance level and power. The exact binomial test can be performed with command `bitest`; see [R] **bitest**. When the exact binomial test is performed, you can use the significance level approach and compare the p-value from the exact test to the p-value boundaries reported by `gsdesign oneproportion`.

`continuity` requests that the continuity correction of Levin and Chen (1999) be applied to the normal approximation of the discrete distribution. This yields an estimate of the sample size that would be required by the exact binomial test at the specified significance level and power.

Bounds

`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(# [, equal]) 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).

equal indicates that equal information increments be enforced, which is to say that the same number of new observations will be collected at each look. The default behavior is to start by dividing information evenly among looks, then proceed by rounding up to a whole number of observations at each look. This can cause slight differences in the information collected at each look.

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.

Graph

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

graphopts are the following:

xdimsampsize labels the *x* axis with the sample size collected (the default).

xdiminformation labels the *x* axis with the information fraction unless information() is specified, in which case information levels will be used.

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 gsdesign oneproportion but are not shown in the dialog box:

force indicates that gsdesign oneproportion should allow unsupported power oneproportion options, such as options specifying a cluster randomized design. Even with option force, the power oneproportion options specified must be compatible with sample-size determination, not effect size or power calculation. In addition, *numlist*s are not supported in options or in arguments as they are with power, even when force is specified.

poweriteration(*powiteropts*) controls the iterative algorithm used to calculate the fixed-study sample size. This is seldom used.

powiteropts are the following:

init(#) specifies an initial value for the sample size when iteration is used to compute the fixed-study sample size. The default is to use a closed-form normal approximation to compute an initial sample size.

iterate(#) specifies the maximum number of iterations for the Newton method during calculation of the fixed-study sample size. The default is iterate(500).

tolerance(#) specifies the tolerance used to determine whether successive parameter estimates have converged when calculating the fixed-study sample size. The default is tolerance(1e-12). See *Convergence criteria* in [M-5] **solvenl()** for details.

ftolerance(#) specifies the tolerance used when calculating the fixed-study sample size to determine whether the proposed solution of a nonlinear equation is sufficiently close to 0 based on the squared Euclidean distance. The default is ftolerance(1e-12). See *Convergence criteria* in [M-5] **solvenl()** for details.

matlistopts(*general_options*) affects the display of the matrix of boundaries and sample sizes. *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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**. 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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

`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 $\mathbf{g}\mathbf{H}^{-1}\mathbf{g}' <$ 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:

Introduction
Using gsdesign oneproportion
Background for examples
Computing sample size and stopping boundaries

This entry describes the use of the gsdesign oneproportion command for designing a group sequential analysis for a one-sample proportion test. See [ADAPT] **GSD intro** for a general introduction to GSDs for clinical trials; see [ADAPT] **gsbounds** for information about group sequential bounds; and see [ADAPT] **gsdesign** for information about designing group sequential clinical trials with the gsdesign command. Also see [PSS-2] **Intro (power)** for a general introduction to power and sample-size analysis, and see [PSS-2] **power oneproportion** for details about study design for a one-sample proportion test.

Introduction

The analysis of proportions is carried out in clinical trials where the response variable, or endpoint, is binary. Each observation is a Bernoulli outcome with a fixed probability p of observing an event of interest in a population. We assume the outcome is observed a fixed number of times and that individual observations are independent with shared probability of success p. As an example, in a clinical trial of a drug for tuberculosis treatment, the endpoint of interest might be culture status after eight weeks. Each observation is the binary indicator of whether tuberculosis was present or absent from the culture taken from one participant.

Sometimes an endpoint that can take several values is discretized into a binary endpoint. For example, the World Health Organization defines a low birthweight as below 2,500 grams. In a clinical trial examining the effect of a prenatal nutrition program on the proportion of newborns with low birthweight, each observation is the binary indicator of whether an infant weighs less than 2,500 grams at birth.

Some clinical trials combine multiple endpoints into a single composite endpoint, which can be binary. A heart failure trial might use a composite endpoint such as "cardiovascular death or heart failure hospitalization". In this case, each observation is an indicator of whether heart failure caused a participant's death or hospitalization. The outcome from participants who died due to heart failure, were hospitalized due to heart failure, or died following hospitalization for heart failure would be recorded as 1, while the outcome from participants who neither died nor were hospitalized due to heart failure would be recorded as 0.

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 the 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.

`gsdesign oneproportion` calculates sample size and stopping boundaries for a group sequential trial comparing the population proportion of one group against a prespecified reference value. Specifically, we consider the null hypothesis $H_0 : p = p_0$ versus the two-sided alternative hypothesis $H_a : p \neq p_0$, the upper one-sided alternative $H_a : p > p_0$, or the lower one-sided alternative $H_a : p < p_0$.

Two common hypothesis tests for a one-sample proportion are the small-sample binomial test and the asymptotic (large-sample) normal test. `gsdesign oneproportion` allows sample-size calculations for tests using the large-sample normal approximation of the sampling distribution of the test statistic. This test, implemented with command `prtest`, yields a z statistic that can be compared with the boundary critical values calculated by `gsdesign oneproportion`. For small samples, the exact binomial test can be performed by using the `bitest` command, but the significance level approach must be used to compare the p-value from the test statistic to the boundary.

Using gsdesign oneproportion

`gsdesign oneproportion` calculates sample size and stopping boundaries for a group sequential trial comparing a population proportion against a hypothesized value. `gsdesign oneproportion` can be thought of as a combination of `power oneproportion` for sample-size calculations and `gsbounds` for stopping boundary calculations.

To compute sample size, you must specify the effect size. There are two ways to do this: by specifying p_0 and p_a, the proportions under the null and alternative hypotheses, respectively, or by specifying p_0 and the difference $p_a - p_0$ in the `diff()` option. There is no default value for `diff()`, so either p_0 and p_a or p_0 and `diff()` must be included as part of the command specification. By default, sample sizes are calculated assuming that a score test will be conducted. To perform sample-size calculations for a Wald test, specify the `test(wald)` option.

Options alpha(), power(), beta(), and onesided are used for both sample-size and stopping-boundary calculations. The default significance level, known as the familywise type I error rate, is 0.05 and can be changed by specifying the alpha() option. The default power is 0.8, which corresponds to a type II error rate of 0.2. This can be modified either by specifying the power in the power() option or by specifying the type II error in the beta() option. The default test is two-sided, and the onesided option requests a one-sided test, the direction of which is indicated by the sign of the effect size.

The group sequential stopping rule is determined by the efficacy() and futility() options. Stopping can be for efficacy, futility, or both, and if no stopping rule is specified, the default is to use an O'Brien–Fleming efficacy bound. If futility bounds are requested, the default behavior is to treat them as nonbinding. A trial that crosses a nonbinding futility bound can be stopped for futility, but the familywise type I error is controlled even if the trial continues. Binding futility bounds can be requested with futility() suboption binding. A trial that crosses a binding futility bound must be stopped for futility. If it continues, the familywise type I error will not be controlled at the specified significance level.

The number of looks, or analyses of the trial data, is specified with nlooks(). Alternatively, the information() option can be used to specify the spacing of the looks as a *numlist* of increasing information levels. In this case, values of the numlist are automatically rescaled so that the final look has the maximum information required by the design. If neither nlooks() nor information() is specified, the default is two looks.

By default, the sample size is rounded up to a whole number at each look, but the nfractional option can be used to report fractional sample sizes. If nlooks() is specified, the default behavior is to divide information evenly among each look before rounding. Rounding can cause slight differences in the amount of information collected at each look, and nlooks() suboption equal can be specified to enforce equal information increments by requiring the same number of new observations at each look.

Background for examples

Oncology is an area where single-arm trials are becoming increasingly common, and some have even led to approval by regulators. This trend is studied by Tenhunen et al. (2020), who note that regulatory approval is most common in trials where the response rate is compared with a prespecified threshold for success, and many of the approved single-arm trials are lung cancer trials and trials of late-line treatments.

In example 1, we used gsdesign onemean to calculate sample sizes and bounds for a clinical trial with a continuous endpoint for sunitinib malate as a salvage therapy for lung cancer. Salvage therapy, also known as rescue therapy, is a term for treatments that are considered when all standard treatment protocols have failed because they were ineffective or because they caused the patient intolerable side effects. Here we consider the design of a clinical trial of sunitinib as a salvage therapy for advanced unresectable non–small cell lung cancer, where "unresectable" describes tumors that cannot be removed surgically.

The trial's outcome of interest is the objective response rate (ORR), defined as the proportion of participants that exhibit at least a partial response to therapy (Delgado and Guddati 2021). Each participant's response to therapy can be considered a Bernoulli trial, with a response of 1 indicating clinical improvement and a response of 0 indicating lack of improvement.

Socinski et al. (2008) report the results of a phase 2 clinical trial of sunitinib for participants with advanced non–small cell lung cancer that had progressed despite treatment with the standard of care, a platinum-based chemotherapy regimen. They found an ORR of 11.1%, which might seem small but can be considered a victory because the probability of clinical improvement without treatment is effectively 0% (Shatola et al. 2020).

In this trial, there can be no active control because all standard treatment protocols have failed, and it is not ethical to recruit participants to a placebo control arm knowing that they stand no chance of improvement. Fortunately, a control arm is not necessary to compare the ORR of the experimental arm with a prespecified clinically relevant ORR, so a single-arm trial can be used.

Computing sample size and stopping boundaries

▷ Example 1: Efficacy bounds for a large-sample test of one proportion

We use `gsdesign oneproportion` to calculate efficacy bounds and sample sizes for this situation. The ORR of untreated patients with advanced unresectable non–small cell lung cancer is 0%, and we define a clinically meaningful threshold for success to be an ORR of 5%. Typically, the null hypothesis in a clinical trial is "no treatment effect", but we are uninterested in clinically irrelevant improvements in ORR, so we modify our null hypothesis to be "no meaningful treatment effect".

We will test $H_0 : p \leq 0.05$ against $H_a : p > 0.05$, which is identical to testing whether sunitinib is substantially superior to no treatment, with a superiority margin of $\delta = 0.05$ (Chow et al. 2018, chap. 4.1.2). We require 90% power to detect the difference between $p_0 = 0.05$ and $p_a = 0.111$ at a familywise significance level of 2.5%. We plan on conducting a large-sample score test, and we employ group sequential specifications of Pocock efficacy boundaries with two analyses (one interim, one final).

```
. gsdesign oneproportion 0.05 0.111, alpha(0.025) power(0.9) onesided
> efficacy(pocock)

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

Efficacy: Pocock

Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.9000
      delta = 0.0610
         p0 = 0.0500
         pa = 0.1110

Expected sample size:
         H0 = 202.50
         Ha = 143.78

Info. ratio = 1.1001
    N fixed =    186
      N max =    204

Fixed-study crit. value = 1.9600

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

────────────────────────────────────────────────
      │  Info.        Efficacy     │ Sample size
 Look │  frac.     Upper   p-value │           N
──────┼────────────────────────────┼────────────
    1 │   0.50    2.1783    0.0147 │         102
    2 │   1.00    2.1783    0.0147 │         204
────────────────────────────────────────────────

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

gsdesign oneproportion displays the specified study parameters, including p0, the proportion under the null hypothesis; pa, the proportion under the alternative hypothesis; and delta, the difference between pa and p0.

The next section of the output displays the expected sample size, which is the average sample size if the group sequential trial were to be repeated many times. The following section reports 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 sample size at the final look of the GSD to the sample size from a fixed study design.

Expected sample size is calculated under both the null and the alternative hypotheses. Because this design does not include futility bounds that would allow stopping to accept H_0, the expected sample size under the null hypothesis is 202.5 participants, nearly the full sample size of 204. If the alternative hypothesis is true, the ability to stop early for treatment efficacy yields an expected sample size of 143.78, a savings over the 186 subjects required by the fixed design.

The table at the end of the output displays the stopping boundaries and sample sizes at each look, but it is informative to examine the bounds visually as well. We rerun the previous command with the graphbounds option to produce a graph of the stopping boundaries.

```
. gsdesign oneproportion 0.05 0.111, alpha(0.025) power(0.9) onesided
> efficacy(pocock) graphbounds
```

(output omitted)

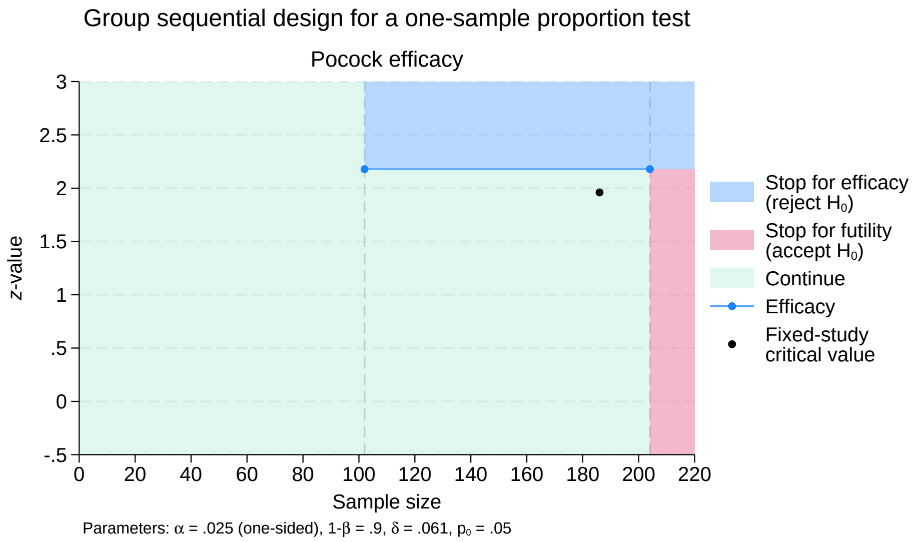


Figure 1. Pocock efficacy bounds for a test of one proportion

On the graph, the horizontal axis marks the sample size collected, and the vertical axis represents the z value of the test statistic. No tests are conducted until 102 observations have been collected, so for the first 101 observations, all z-values lie within the continuation region. Once data from 102 participants have been collected, a score test will be performed.

The z statistic from that test, z_1, will be compared with the efficacy critical value, marked with a blue dot. If $z_1 \geq 2.178$, it lies in the blue rejection region, so we will reject H_0 and the trial will be stopped for efficacy; but if $z_1 < 2.178$, it lies in the green continuation region, and the trial will continue to the second and final look.

At the final look, the ORR will be calculated using data from all 204 participants, and a score test will be conducted. As before, if $z_2 \geq 2.178$, then H_0 will be rejected. But this time, if $z_2 < 2.178$, it will lie within the red acceptance region because there is no continuation region at the final look; so if $z_2 < 2.178$, we will accept H_0. The concept of accepting the null hypothesis has a long history in the context of sequential hypothesis testing (see [ADAPT] **GSD intro** for details).

◁

▷ Example 2: Efficacy and futility bounds with uneven information increments

We continue the scenario from example 1, but we adjust the design by adding futility boundaries and additional interim looks at the data. Previously, we used a classical Pocock efficacy bound, which is characterized by having the same critical values at all looks. The classical Wang–Tsiatis efficacy bound offers an alternative that is known for having very conservative critical values at early looks but uses a final critical value that is close to the critical value of an equivalently powered fixed-sample test. Here we choose an error-spending approximation of the O'Brien–Fleming efficacy bound, which is similar in shape to the classical O'Brien–Fleming bound but is constructed using the error-spending method. See *Methods and formulas* in [ADAPT] **gsbounds** for more information about the error-spending approximation of the classical O'Brien–Fleming bound.

In addition to the error-spending O'Brien–Fleming-style efficacy bound, we add a nonbinding error-spending Hwang–Shih–de Cani futility bound with parameter $\gamma_f = -3$. Nonbinding futility bounds offer the option of stopping early to accept H_0 if the futility bound is crossed, but if the trial continues after crossing a nonbinding futility bound, the familywise type I error is still controlled. We also add two more looks in between the first look (with half the data) and the final look (with the complete dataset). Option `information()` allows us to schedule those looks to occur with 75% and 90% of the data.

```
. gsdesign oneproportion 0.05 0.111, alpha(0.025) power(0.9) onesided
> efficacy(errobfleming) futility(hsdecani(-3))
> information(50 75 90 100) graphbounds

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

Efficacy: Error-spending O'Brien-Fleming style
Futility: Error-spending Hwang-Shih-de Cani, nonbinding, gamma = -3.0000

Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.9000
      delta = 0.0610
         p0 = 0.0500
         pa = 0.1110

Expected sample size:
         H0 = 124.28
         Ha = 147.09

Info. ratio = 1.0852
    N fixed =    186
      N max =    201

Fixed-study crit. value = 1.9600

Critical values, p-values, and sample sizes for a group sequential design
─────────────────────────────────────────────────────────────────────
      │ Info.        Efficacy              Futility        │ Sample size
 Look │ frac.     Upper   p-value      Lower   p-value     │           N
──────┼────────────────────────────────────────────────────┼─────────────
    1 │  0.50    2.9626    0.0015     0.2963    0.3835     │         101
    2 │  0.75    2.3590    0.0092     1.1477    0.1255     │         151
    3 │  0.90    2.1649    0.0152     1.6551    0.0490     │         181
    4 │  1.00    2.0731    0.0191     2.0731    0.0191     │         201
─────────────────────────────────────────────────────────────────────

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

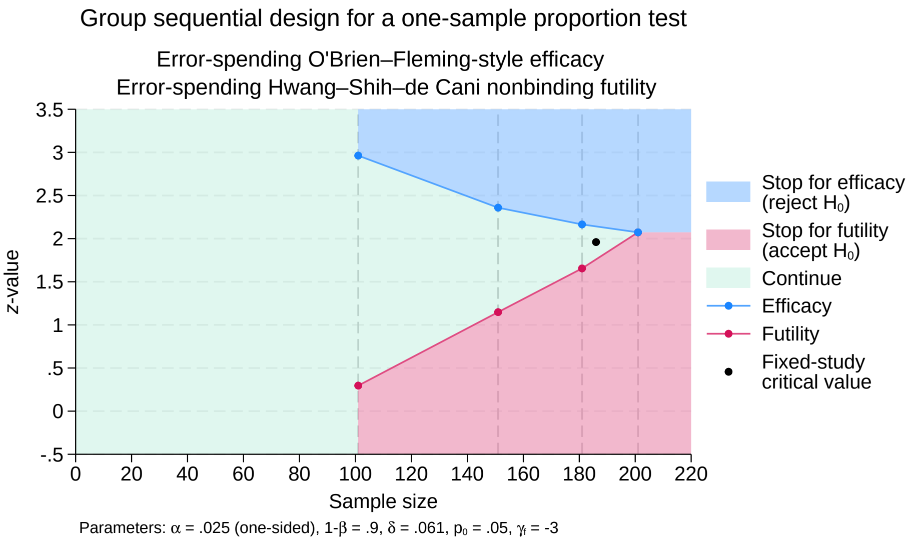


Figure 2. One proportion test with error-spending efficacy and futility bounds

Compared with the design in example 1, the design with futility stopping has nearly the same maximum sample size (201 versus 204 observations) and expected sample size under H_a (147.09 versus 143.78 observations) but a much smaller expected sample size under H_0 (124.28 versus 202.5 observations).

The improved efficiency when H_0 is true is due to the ability to accept the null hypothesis at interim analyses. If $z_1 < 0.296$, the trial can be ended for futility with only 101 observations. As the study proceeds from one look to the next, the continuation region shrinks as the efficacy and futility bounds get closer together. By the third look, with 90% of the data (181 observations), the continuation region has shrunk to $z_3 \in [1.655, 2.165)$. At the final look, there is no continuation region because the efficacy and futility critical values are equal. If $z_4 \geq 2.073$, then we reject H_0; otherwise, we accept H_0.

◁

▷ Example 3: Futility-only stopping

We continue the scenario from example 2. In that example, we allowed for early stopping for efficacy (to reject H_0) as well as for futility (to accept H_0). This enabled a substantial reduction in sample size compared with a fixed study design, but sometimes there are reasons for intentionally choosing a less efficient design.

One argument against early stopping for efficacy is that a larger sample size allows for a better characterization of adverse events, which are harmful side effects of the treatment and negative medical outcomes not associated with an underlying disease. Socinski et al. (2008) report the incidence of several adverse events among participants taking sunitinib, including fatigue, pain, hypertension, and pulmonary hemorrhage. If sunitinib is effective at treating non–small cell lung cancer, it will be important to fully understand its side effects before using it to treat the general population. But if sunitinib is not an effective treatment, it would be both wasteful and unethical not to stop the trial as soon as the lack of efficacy is apparent.

To avoid stopping for efficacy before side effects can be characterized, we modify the design from example 2 by removing the efficacy bound, but the rest of the design remains the same.

```
. gsdesign oneproportion 0.05 0.111, alpha(0.025) power(0.9) onesided
> futility(hsdecani(-3)) information(50 75 90 100) graphbounds
Group sequential design for a one-sample proportion test
Score z test
H0: p = p0 versus Ha: p > p0
Futility: Error-spending Hwang-Shih-de Cani, nonbinding, gamma = -3.0000
Study parameters:
      alpha = 0.0250  (upper one-sided)
      power = 0.9000
      delta = 0.0610
         p0 = 0.0500
         pa = 0.1110
Expected sample size:
         H0 = 123.48
         Ha = 197.78
Info. ratio = 1.0658
    N fixed =    186
      N max =    198
Fixed-study crit. value = 1.9600
Critical values, p-values, and sample sizes for a group sequential design
──────┬─────────────────────────────────────────────────┬────────────
      │ Info.        Efficacy             Futility      │ Sample size
 Look │ frac.      Upper   p-value      Lower   p-value │           N
──────┼─────────────────────────────────────────────────┼────────────
    1 │  0.50                          0.2748    0.3917 │          99
    2 │  0.75                          1.1214    0.1311 │         148
    3 │  0.90                          1.6221    0.0524 │         178
    4 │  1.00     1.9600    0.0250     1.9600    0.0250 │         198
──────┴─────────────────────────────────────────────────┴────────────
Note: Critical values are for z statistics; otherwise, use p-value
      boundaries.
```

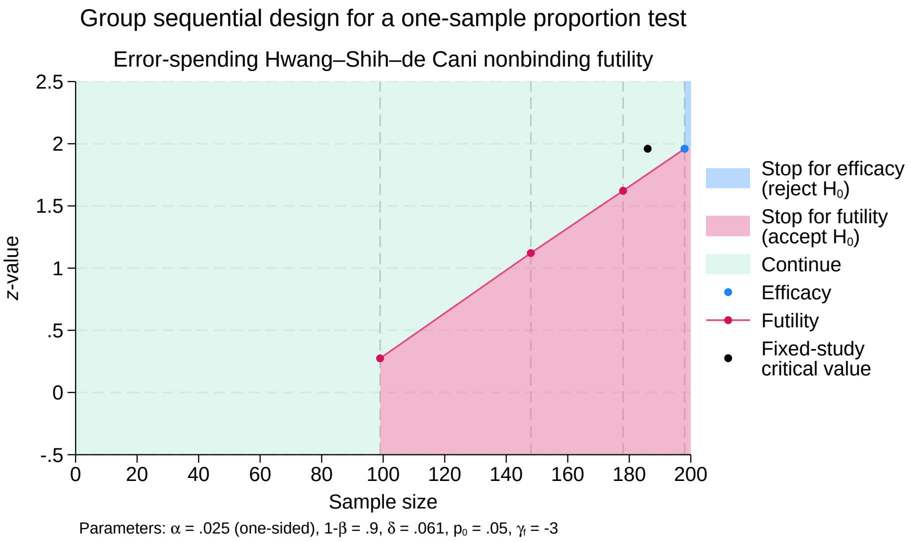


Figure 3. One proportion test with futility bounds

The expected sample size under H_a is nearly the maximum sample size (197.78 versus 198 observations), but the expected sample size under H_0 is a third less than the fixed-study sample size (123.48 versus 186 observations). This achieves our goal of collecting data from as many patients as possible if sunitinib is effective but saving participants from unnecessarily receiving the treatment if it is ineffective.

Efficacy stopping is not an option during the first three looks, so the efficacy critical value for these looks is reported as missing. At each of these looks, the trial can be terminated for futility if the test statistic is below the futility boundary, but because the futility bound is nonbinding, the familywise type I error will be controlled even if the trial is continued.

At the final look, it is at last possible to reject H_0. If $z_4 \geq 1.96$, then we reject the null hypothesis; otherwise, we accept H_0. On the graph, the efficacy bound is displayed as a single blue dot at the final look because that is the only time efficacy stopping is allowed. Interestingly, the nonbinding futility-only design uses the same critical value as the fixed design. To understand why, we display the contents of matrix r(aspent_nofstop), the cumulative type I error spent per look if the trial does not stop for futility (even if it were to cross the futility bound at an interim look).

```
. matlist r(aspent_nofstop), format(%50.3g)

             | alpha spent
             | assuming no futility stopping
-------------+-------------------------------
      Look 1 |                             0
      Look 2 |                             0
      Look 3 |                             0
      Look 4 |                          .025
```

Just like a fixed-sample trial with a one-sided significance level of 2.5%, this design spends its entire allotment of type I error during a single analysis. Practically speaking, this means that if H_0 is true, it is impossible to reject the null (and commit a type I error) during the first three looks. If H_0 is true and this trial is repeated many times (each time continuing until the final look even if the futility bound is crossed during an interim analysis), then 2.5% of the time we will erroneously reject the null hypothesis at the final look. Viewed from this perspective, this error-spending regimen is essentially the same as that of a fixed-sample design, which is why it uses the same critical value.

◁

Stored results

gsdesign oneproportion 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(continuity)	1 if continuity correction is used, 0 otherwise
r(delta)	effect size
r(diff)	difference between the alternative and null proportions
r(effparam)	efficacy parameter (if wtsiatis(), kdemets(), or hsdecani() specified)
r(ESS0)	expected sample size under null hypothesis
r(ESS1)	expected sample size under alternative hypothesis
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(N_fixed)	sample size of a fixed study design
r(N_fixedfrac)	fractional sample size of a fixed study design
r(N_max)	maximum sample size if the study continues to completion
r(nfractional)	1 if nfractional is specified, 0 otherwise
r(nlooks)	number of analyses
r(onesided)	1 for a one-sided test, 0 otherwise
r(p0)	proportion under the null hypothesis
r(pa)	proportion under the alternative hypothesis
r(pow_converged)	1 if power calculation iteration algorithm converged, 0 otherwise
r(pow_deltax)	final parameter tolerance achieved for power calculation
r(pow_ftolerance)	requested distance of power calculation objective function from 0
r(pow_function)	final distance of power calculation objective function from 0
r(pow_init)	initial value for power calculation sample size
r(pow_iter)	number of iterations performed for power calculation
r(pow_maxiter)	maximum number of iterations for power calculation
r(pow_tolerance)	requested parameter tolerance for power calculation
r(power)	specified overall power
r(power_a)	attained 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)	gsdesign
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
r(method)	oneproportion
r(test)	score or wald

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(bspent_a)`	attained cumulative beta spent per look (when futility bounds are specified)
`r(design)`	sample size and stopping boundaries at interim looks
`r(info_frac)`	specified information fraction
`r(info_frac_a)`	fraction of attained information
`r(info_level)`	specified information level
`r(p_crit)`	p-values corresponding to boundary critical values
`r(sampsize)`	sample size at interim looks

Methods and formulas

Sample sizes at interim analyses are calculated as the product of the information fraction, the information ratio, and the sample size of a fixed-sample study.

See *Methods and formulas* in [ADAPT] **gsbounds** for the formulas used to calculate the stopping boundaries, information fraction, and information ratio. See *Methods and formulas* in [PSS-2] **power oneproportion** for the formulas used to calculate the sample size for a fixed study. See *Methods and formulas* in [ADAPT] **gsdesign** for the formulas used to calculate the expected sample size.

References

Chow, S.-C., J. Shao, H. Wang, and Y. Lokhnygina. 2018. *Sample Size Calculations in Clinical Research*. 3rd ed. Boca Raton, FL: CRC Press.

Delgado, A., and A. K. Guddati. 2021. Clinical endpoints in oncology—a primer. *American Journal of Cancer Research* 11: 1121–1131.

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. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70: 659–663. https://doi.org/10.1093/biomet/70.3.659.

Levin, B., and X. Chen. 1999. Is the one-half continuity correction used once or twice to derive a well-known approximate sample size formula to compare two independent binomial distributions? *American Statistician* 53: 62–66. https://doi.org/10.1080/00031305.1999.10474431.

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.

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.

Shatola, A., K. N. Nguyen, E. Kamangar, and M. E. Daly. 2020. Spontaneous regression of non-small cell lung cancer: A case report and literature review. *Cureus* 12: e6639. https://doi.org/10.7759/cureus.6639.

Socinski, M. A., S. Novello, J. R. Brahmer, R. Rosell, J. M. Sanchez, C. P. Belani, R. Govindan, J. N. Atkins, H. H. Gillenwater, C. Pallares, L. Tye, P. Selaru, R. C. Chao, and G. V. Scagliotti. 2008. Multicenter, phase II trial of sunitinib in previously treated, advanced non–small-cell lung cancer. *Journal of Clinical Oncology* 26: 650–656. https://doi.org/10.1200/JCO.2007.13.9303.

Tenhunen, O., F. Lasch, A. Schiel, and M. Turpeinen. 2020. Single-arm clinical trials as pivotal evidence for cancer drug approval: A retrospective cohort study of centralized European marketing authorizations between 2010 and 2019. *Clinical Pharmacology and Therapeutics* 108: 653–660. https://doi.org/10.1002/cpt.1965.

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

[ADAPT] **gsbounds** — Boundaries for group sequential trials

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

[ADAPT] **gsdesign twoproportions** — Group sequential design for a two-sample proportions test

[ADAPT] **Glossary**

[PSS-2] **power oneproportion** — Power analysis for a one-sample proportion test

[R] **bitest** — Binomial probability test

[R] **prtest** — Tests of proportions

gsdesign twoproportions — Group sequential design for a two-sample proportions test

Description Quick start Menu Syntax Options
Remarks and examples Stored results Methods and formulas References Also see

Description

gsdesign twoproportions computes stopping boundaries and sample sizes for interim analyses of clinical trials using a two-sample proportions test with a group sequential design (GSD). Stopping can be for efficacy, futility, or both. For stopping boundary calculations without sample sizes, see [ADAPT] **gsbounds**. For sample-size calculations for a fixed-sample test of two proportions, see [PSS-2] **power twoproportions**.

Quick start

Sample size and stopping boundaries for a two-sided χ^2 test of $H_0: \pi_1 = \pi_2$ versus $H_a: \pi_1 \neq \pi_2$, with default familywise significance level $\alpha = 0.05$ and power of 0.8 to detect the difference between a control-group proportion of $p_1 = 0.7$ and an experimental-group proportion of $p_2 = 0.55$, using default group sequential specifications of O'Brien–Fleming efficacy boundaries with two analyses (one interim, one final)

```
gsdesign twoproportions 0.7 0.55
```

Same as above, but specified as $p_1 = 0.7$ and difference between proportions $p_2 - p_1 = -0.15$

```
gsdesign twoproportions 0.7, diff(-0.15)
```

Same as above, but specified as $p_1 = 0.7$ and ratio $p_2/p_1 = 0.7857$

```
gsdesign twoproportions 0.7, ratio(0.7857)
```

Same as above, but specified as $p_1 = 0.7$ and odds ratio $\{p_2/(1-p_2)\}/\{p_1/(1-p_1)\} = 0.5238$

```
gsdesign twoproportions 0.7, oratio(0.5238)
```

Same as above, but use a Wang–Tsiatis efficacy bound with parameter $\Delta_e = 0.25$ and conduct four looks

```
gsdesign twoproportions 0.7, oratio(0.5238) efficacy(wtsiatis(0.25)) ///
      nlooks(4)
```

Same as above, but calculate sample size for a likelihood-ratio test and add a binding O'Brien–Fleming futility bound

```
gsdesign twoproportions 0.7, oratio(0.5238) test(lrchi2)        ///
      efficacy(wtsiatis(0.25)) futility(obfleming, binding) nlooks(4)
```

Same as above, but allocate twice as many participants to the experimental group as the control group and graph the boundaries

```
gsdesign twoproportions 0.7, oratio(0.5238) test(lrchi2) nratio(2)  ///
      efficacy(wtsiatis(0.25)) futility(obfleming, binding)    ///
      nlooks(4) graphbounds
```

Menu

Statistics > Power, precision, and sample size

Syntax

`gsdesign twoproportions` p_1 p_2 [, *twopropopts boundopts*]

where p_1 is the proportion in the control (reference) group, and p_2 is the proportion in the experimental (treatment) group.

twopropopts	Description
Main	
`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)`
`onesided`	request a one-sided test; default is two-sided
`nfractional`	report fractional sample size
`nratio(#)`	ratio of sample sizes of experimental to control groups; default is `nratio(1)`, meaning equal group sizes
`diff(#)`	difference between the experimental-group and the control-group proportions, $p_2 - p_1$; specify instead of the experimental-group proportion p_2
`rdiff(#)`	risk difference, $p_2 - p_1$; synonym for `diff()`
`ratio(#)`	ratio of the experimental-group proportion to the control-group proportion, p_2/p_1; specify instead of the experimental-group proportion p_2
`rrisk(#)`	relative risk, p_2/p_1; synonym for `ratio()`
`oratio(#)`	odds ratio, $\{p_2/(1-p_2)\}/\{p_1/(1-p_1)\}$; specify instead of the experimental-group proportion p_2
`effect(`*effect*`)`	specify the type of effect to display; default is `effect(diff)`
`test(`*test*`)`	specify the type of test; options are `chi2` (the default) and `lrchi2`
`continuity`	apply continuity correction to the normal approximation of the discrete distribution
`force`	allow calculation with unsupported `power twoproportions` options
`poweriteration(`*powiteropts*`)`	iteration options for the calculation of fixed-study sample size; seldom used

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

`force` and `poweriteration()` do not appear in the dialog box.

effect	Description
`diff`	difference between proportions, $p_2 - p_1$; the default
`rdiff`	risk difference, $p_2 - p_1$; synonym for `diff`
`ratio`	ratio of proportions, p_2/p_1
`rrisk`	relative risk, p_2/p_1; synonym for `ratio`
`oratio`	odds ratio, $\{p_2/(1-p_2)\}/\{p_1/(1-p_1)\}$

powiteropts	Description
`init(#)`	initial value for fixed-study sample size
`iterate(#)`	maximum number of iterations; default is `iterate(500)`
`tolerance(#)`	parameter tolerance; default is `tolerance(1e-12)`
`ftolerance(#)`	function tolerance; default is `ftolerance(1e-12)`

boundopts	Description
Bounds	
`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(#`[`, equal`]`)`	total number of analyses (`nlooks()` $-$ 1 interim analyses and one final analysis); use `equal` to enforce equal information increments; if neither `nlooks()` nor `information()` is specified, the default is `nlooks(2)`
`information(`*numlist*`)`	sequence of information levels for analyses; default is evenly spaced
`nopvalues`	suppress p-values
Graph	
`graphbounds`[`(`*graphopts*`)`]	graph boundaries
`matlistopts(`*general_options*`)`	control the display of boundaries and sample size; seldom used
optimopts	optimization options for boundary calculations; seldom used

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

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

graphopts	Description
`xdimsampsize`	label the x axis with the sample size collected (default)
`xdiminformation`	label the x axis with the information fraction; 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

`alpha(#)` sets the overall significance level, which is the familywise type I error rate for all analyses (interim and final). `alpha()` must be in $(0, 0.5)$. The default is `alpha(0.05)`.

`power(#)` sets the overall power for all analyses. `power()` must be in $(0.5, 1)$. 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. `beta()` must be in $(0, 0.5)$. 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.

`onesided` requests a study design for a one-sided test. The direction of the test is inferred from the effect size.

`nfractional` specifies that fractional sample sizes be reported.

`nratio(#)` specifies the sample-size ratio of the experimental group relative to the control group, $N2/N1$. The default is `nratio(1)`, meaning equal allocation between the two groups.

`diff(#)` specifies the difference between the experimental-group proportion and the control-group proportion, $p_2 - p_1$. You can either specify the experimental-group proportion p_2 as a command argument or specify the difference between the two proportions in `diff()`. If you specify `diff(#)`, the experimental-group proportion is computed as $p_2 = p_1 + \#$. This option may not be combined with `rdiff()`, `ratio()`, `rrisk()`, or `oratio()`.

`rdiff(#)` specifies the risk difference $p_2 - p_1$. This is a synonym for option `diff()`. `rdiff()` may not be combined with `diff()`, `ratio()`, `rrisk()`, or `oratio()`.

`ratio(#)` specifies the ratio of the experimental-group proportion to the control-group proportion, p_2/p_1. You can either specify the experimental-group proportion p_2 as a command argument or specify the ratio of the two proportions in `ratio()`. If you specify `ratio(#)`, the experimental-group proportion is computed as $p_2 = p_1 \times \#$. This option may not be combined with `diff()`, `rdiff()`, `rrisk()`, or `oratio()`.

`rrisk(#)` specifies the relative risk or risk ratio, p_2/p_1. This is a synonym for option `ratio()`. `rrisk()` may not be combined with `diff()`, `rdiff()`, `ratio()`, or `oratio()`.

`oratio(#)` specifies the odds ratio $\{p_2/(1 - p_2)\}/\{p_1/(1 - p_1)\}$. You can either specify the experimental-group proportion p_2 as a command argument or specify the odds ratio in `oratio()`. If you specify `oratio(#)`, the experimental-group proportion is computed as $p_2 = 1/\{1 + (1 - p_1)/(p_1 \times \#)\}$. This option may not be combined with `diff()`, `rdiff()`, `ratio()`, or `rrisk()`.

`effect(`*effect*`)` specifies the parameterization of the effect size to be reported in the output as `delta`. *effect* is one of `diff`, `rdiff`, `ratio`, `rrisk`, or `oratio`. If the effect size is specified with option `diff()` or as p_1 and p_2, the default is to parameterize `delta` as the difference between proportions, equivalent to specifying `effect(diff)`. If the effect size is specified using option `rdiff()`, `ratio()`, `rrisk()`, or `oratio()`, then `delta` defaults to using the corresponding parameterization. `effect(`*effect*`)`, however, requests an alternative parameterization of effect size `delta`—one that corresponds to *effect*.

`test(`*test*`)` specifies the type of test that will be used for data analysis. Sample-size calculations depend on the test that will be conducted. *test* is either `chi2` or `lrchi2`.

`chi2` requests computations for Pearson's χ^2 test. This is the default test, and this test can be performed with command `prtest` or command `tabulate twoway`; see [R] **prtest** and [R] **tabulate twoway**, respectively.

`lrchi2` requests computations for the likelihood-ratio test. This test can be performed with command `tabulate twoway`; see [R] **tabulate twoway**.

Note that `power twoproportions` option `test(fisher)` cannot be used to calculate sample size and is therefore not compatible with `gsdesign twoproportions`. However, option `continuity` implements a continuity correction that yields an estimate of the sample size that would be required by Fisher's exact test at the specified significance level and power. Fisher's exact test can be performed with command `tabulate twoway`; see [R] **tabulate twoway**. When Fisher's exact test is performed, you can use the significance level approach and compare the p-value from the t test to the p-value boundaries reported by `gsdesign twoproportions`, as demonstrated in example 2.

`continuity` requests that the continuity correction of Casagrande, Pike, and Smith (1978) be applied to the normal approximation of the discrete distribution. This yields an estimate of the sample size that would be required by Fisher's exact test at the specified significance level and power. `continuity` cannot be specified with `test(lrchi2)`.

Bounds

`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(`#[`, equal`]`)` 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)`.

`equal` indicates that equal information increments be enforced, which is to say that the same number of new observations will be collected at each look. The default behavior is to start by dividing information evenly among looks, then proceed by rounding up to a whole number of observations at each look. This can cause slight differences in the information collected at each look.

`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.

Graph

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

graphopts are the following:

xdimsampsize labels the x axis with the sample size collected (the default).

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

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 gsdesign twoproportions but are not shown in the dialog box:

force indicates that gsdesign twoproportions should allow unsupported power twoproportions options, such as options specifying a cluster randomized design. Even with option force, the power twoproportions options specified must be compatible with sample-size determination, not effect size or power calculation. In addition, *numlist*s are not supported in options or in arguments as they are with power, even when force is specified.

poweriteration(*powiteropts*) controls the iterative algorithm used to calculate the fixed-study sample size. This is seldom used.

powiteropts are the following:

init(#) specifies an initial value for the sample size when iteration is used to compute the fixed-study sample size. The default is to use a closed-form normal approximation to compute an initial sample size.

iterate(#) specifies the maximum number of iterations for the Newton method during calculation of the fixed-study sample size. The default is iterate(500).

tolerance(#) specifies the tolerance used to determine whether successive parameter estimates have converged when calculating the fixed-study sample size. The default is tolerance(1e-12). See *Convergence criteria* in [M-5] **solvenl()** for details.

ftolerance(#) specifies the tolerance used when calculating the fixed-study sample size to determine whether the proposed solution of a nonlinear equation is sufficiently close to 0 based on the squared Euclidean distance. The default is ftolerance(1e-12). See *Convergence criteria* in [M-5] **solvenl()** for details.

matlistopts(*general_options*) affects the display of the matrix of boundaries and sample sizes. *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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**. 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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

`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:

Introduction
Using gsdesign twoproportions
Background for examples
Computing sample size and stopping boundaries

This entry describes the use of the `gsdesign twoproportions` command for designing a group sequential analysis for a two-sample proportions test. See [ADAPT] **GSD intro** for a general introduction to GSDs for clinical trials; see [ADAPT] **gsbounds** for information about group sequential bounds; and see [ADAPT] **gsdesign** for information about designing group sequential clinical trials with the `gsdesign` command. Also see [PSS-2] **Intro (power)** for a general introduction to power and sample-size analysis, and see [PSS-2] **power twoproportions** for details about study design for a two-sample proportions test.

Introduction

The comparison of two independent proportions is carried out in clinical trials with two groups of participants (known as two-arm trials), where the response variable, or endpoint, is binary. We use the term "success" to indicate observing the outcome of interest, but the outcome of interest could be something that nobody would consider a success in the traditional sense of the word, such as hospitalization or even death.

As an example, in a clinical trial of a drug to treat chronic HIV infection, the endpoint of interest might be whether the disease progresses to AIDS during a two-year course of treatment. Each observation is the binary indicator of whether one participant's HIV progresses to AIDS.

Sometimes an endpoint that can take several values is discretized into a binary endpoint. For instance, the Apgar score of newborn health can range from 0 to 10, and scores below 4 are considered low by the American Academy of Pediatrics (2015). A clinical trial investigating the effect of labor support by a lay doula on newborn health might discretize the Apgar score taken five minutes after birth to determine the proportion of newborns with low five-minute Apgar scores. In this case, each observation is the binary indicator of whether an infant had a five-minute Apgar score below 4.

Some clinical trials combine multiple endpoints into a single composite endpoint, which can be binary. A clinical trial of a treatment for COVID-19 might use a composite endpoint, such as "death or intubation". In this case, each observation is an indicator of whether a participant died or was intubated. The outcome from participants who died, were intubated, or died following intubation would be recorded as 1, while the outcome from participants who neither died nor were intubated would be recorded as 0.

To conduct hypothesis tests, we view each observation as a Bernoulli outcome, and within each arm, we assume that the probability of success is constant for all participants in that arm. We use the notation p_1 to denote the probability of success in the control arm and p_2 to denote the probability of success in the experimental arm. We assume the outcome is observed a fixed number of times in each arm and that each Bernoulli outcome is independent of all other observations.

gsdesign twoproportions calculates sample size and stopping boundaries for a group sequential trial comparing the population proportion of a reference (control) group against the population proportion of an experimental (treatment) group. Specifically, we consider the null hypothesis $H_0: p_2 = p_1$ versus the two-sided alternative hypothesis $H_a: p_2 \neq p_1$, the upper one-sided alternative $H_a: p_2 > p_1$, or the lower one-sided alternative $H_a: p_2 < p_1$.

When the sample size is large, Pearson's χ^2 test can be used to test the null hypothesis. Command prtest implements this test and reports an asymptotically normal z statistic that can be compared directly with the boundary critical values reported by gsdesign twoproportions (the square of the z statistic has an asymptotic χ^2 distribution, hence the name of the test). If H_0 is tested using a method that does not produce a normally distributed test statistic, the significance level approach must be used to compare the p-value from the test statistic to the boundary.

Using gsdesign twoproportions

gsdesign twoproportions calculates sample size and stopping boundaries for a group sequential trial comparing the proportion of successes in two different populations. gsdesign twoproportions can be thought of as a combination of power twoproportions for sample-size calculations and gsbounds for stopping boundary calculations. By default, sample sizes are calculated assuming that Pearson's χ^2 test will be conducted. To perform sample-size calculations for a likelihood-ratio test, specify the test(lrchi2) option.

To compute sample size, you must provide the effect size. There are several ways to do this: by specifying p_1 and p_2, the proportions of the control and experimental groups, respectively; by specifying p_1 and the difference between the experimental-group proportion and the control-group proportion (diff $= p_1 - p_2$); by specifying p_1 and the risk difference (rdiff $= p_1 - p_2$); by specifying p_1 and the ratio of the experimental-group proportion to the control-group proportion (ratio $= p_2/p_1$); by specifying p_1 and the relative risk (rrisk $= p_2/p_1$); or by specifying p_1 and the odds ratio (oratio $= \{p_2/(1-p_2)\}/\{p_1/(1-p_1)\}$). There is no default value for the effect size, so it must be specified in one of these formats.

Options alpha(), power(), beta(), and onesided are used for both sample-size and stopping-boundary calculations. The default significance level, known as the familywise type I error rate, is 0.05 and can be changed by specifying the alpha() option. The default power is 0.8, which corresponds to a type II error rate of 0.2. This can be modified either by specifying the power in the power() option or by specifying the type II error in the beta() option. The default test is two-sided, and the onesided option requests a one-sided test, the direction of which is indicated by the sign of the effect size.

The group sequential stopping rule is determined by the efficacy() and futility() options. Stopping can be for efficacy, futility, or both, and if no stopping rule is specified, the default is to use an O'Brien–Fleming efficacy bound. If futility bounds are requested, the default behavior is to treat them as nonbinding. A trial that crosses a nonbinding futility bound can be stopped for futility, but the familywise type I error is controlled even if the trial continues. Binding futility bounds can be requested with futility() suboption binding. A trial that crosses a binding futility bound must be stopped for futility. If it continues, the familywise type I error will not be controlled at the specified significance level.

The number of looks, or analyses of the trial data, is specified with nlooks(). Alternatively, the information() option can be used to specify the spacing of the looks as a *numlist* of increasing information levels. In this case, values of the numlist are automatically rescaled so that the final look has the maximum information required by the design. If neither nlooks() nor information() is specified, the default is two looks.

By default, the sample size is rounded up to a whole number at each look, but the nfractional option can be used to report fractional sample sizes. If nlooks() is specified, the default behavior is to divide information evenly among each look before rounding. Rounding can cause slight differences in the amount of information collected at each look, and nlooks() suboption equal can be specified to enforce equal information increments by requiring the same number of new observations at each look.

Background for examples

Beta blockers are a class of drugs that are used to reduce the risk of myocardial infarctions (MI), known colloquially as heart attacks. In example 3 in [ADAPT] **gsdesign**, we re-created the experimental design of the landmark Beta-Blocker Heart Attack Trial, which examined the effect of the beta blocker propranolol on participant survival. Here we consider a clinical trial of beta blockers conducted by the Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography (DECREASE) Study Group.

The DECREASE Study Group reported the results of a multicenter, randomized clinical trial to evaluate the use of beta blockers to reduce the incidence of MI within 30 days of major vascular surgery (Poldermans et al. 1999). The target population consisted of patients with cardiac risk factors who were undergoing major vascular surgery. Participants who were randomly assigned to the experimental arm began taking a daily dose of the beta blocker bisoprolol at least one week before their scheduled surgery, and continued taking daily bisoprolol for at least 30 days after surgery, during which time they also received standard perioperative care. Participants randomized to the control arm only received standard perioperative care.

A composite endpoint was used, with the outcomes of interest being death from cardiac causes and nonfatal MI. The outcome of a participant was recorded as 1 if, in the 30 days after surgery, the participant died due to cardiac causes or suffered a nonfatal MI. The outcome was recorded as 0 if the participant survived for 30 days postoperatively without MI.

Computing sample size and stopping boundaries

▷ Example 1: Sample size and efficacy bounds for a large-sample test of two proportions

Suppose that we are interested in designing a study that follows Poldermans et al. (1999). They assumed that the incidence of the primary endpoint would be 30% in the control arm and 15% in the experimental arm. They planned for a familywise two-sided significance level of 5%, power of 80%, and one interim look at approximately 38% of the sample size using an O'Brien–Fleming efficacy boundary. Below, we use gsdesign twoproportions to design and graph a study with these parameters, and we leave test() at its default value of chi2.

```
. gsdesign twoproportions 0.3 0.15, efficacy(obfleming) information(0.38 1)
> graphbounds
Group sequential design for a two-sample proportions test
Pearson's chi-squared test
H0: p2 = p1 versus Ha: p2 != p1
Efficacy: O'Brien-Fleming
Study parameters:
      alpha =  0.0500  (two-sided)
      power =  0.8000
      delta = -0.1500  (difference)
         p1 =  0.3000
         p2 =  0.1500
Expected sample size:
         H0 =  241.78
         Ha =  231.11
Info. ratio =  1.0024
    N fixed =     242
      N max =     242
     N1 max =     121
     N2 max =     121
Fixed-study crit. values = ±1.9600
Critical values, p-values, and sample sizes for a group sequential design
------+--------------------------------------+-----------------------------
      | Info.             Efficacy           |          Sample size
 Look | frac.      Lower     Upper   p-value |       N1        N2         N
------+--------------------------------------+-----------------------------
    1 |  0.38    -3.1878    3.1878    0.0014 |       46        46        92
    2 |  1.00    -1.9651    1.9651    0.0494 |      121       121       242
------+--------------------------------------+-----------------------------
Note: Critical values are for z statistics; otherwise, use p-value
      boundaries.
```

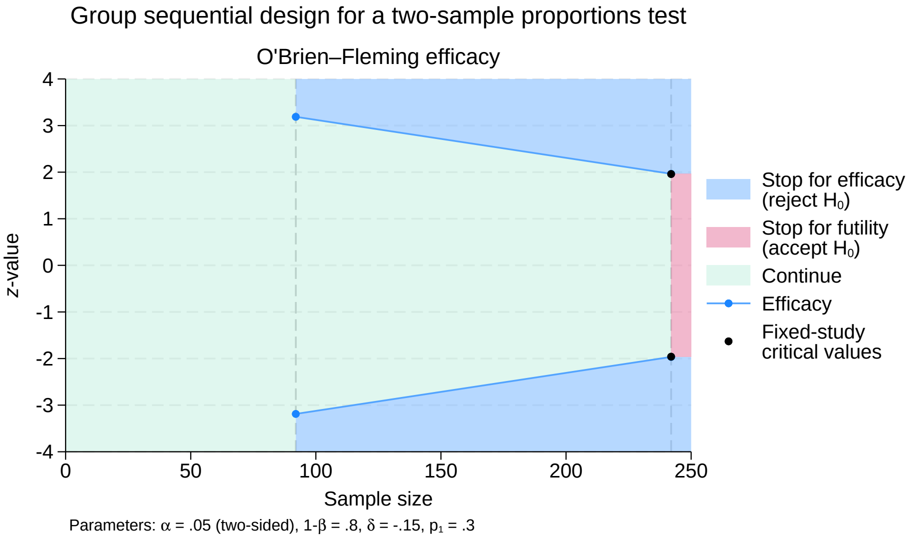


Figure 1. Two-sided test of the equality of two proportions with O'Brien–Fleming efficacy bounds

gsdesign twoproportions displays the specified study parameters, including the control group proportion p1, the experimental group proportion p2, and the difference in proportions.

The next section of the output displays the expected sample size (ESS), which is the average sample size if the group sequential trial were to be repeated many times. The following section reports 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.

Without futility bounds, we cannot stop the trial early to accept H_0, so if the null hypothesis is true, it is not surprising that the ESS of 241.78 is nearly equal to the maximum sample size of 242. If H_a is true, the ESS is 231.11, a modest savings over the maximum sample size.

Examining the boundary critical values, we see the reason that the ESS under H_a was not lower: there is only one interim look, and the critical value at that look, ±3.188, sets a high bar for early stopping. Once data have been collected from 46 subjects in each group, Pearson's χ^2 test is performed and the z statistic, z_1, is compared with the efficacy critical values of ±3.188. To perform Pearson's χ^2 test with a two-sided alternative, we could use command prtest, which reports a z statistic that can be compared directly with the boundary critical values, or command tabulate, which performs the same test but reports a χ^2 statistic (the χ^2 statistic is the square of the z statistic, and the p-values reported by the two tests are identical).

On the graph, we see that if $|z_1| \geq 3.188$, then it lies in the blue rejection region and the trial will be stopped early for efficacy due to the early rejection of H_0. If $|z_1| < 3.188$, then it lies in the green continuation region and the trial will continue on to the final look. At the final look, the critical values are ±1.965, and there is no continuation region. If $|z_2| \geq 1.965$, then H_0 is rejected; otherwise, z_2 lies in the red acceptance region, which indicates that H_0 is accepted.

O'Brien–Fleming efficacy bounds are known for being very conservative at early looks, but the final look of an O'Brien–Fleming design uses a critical value that is only slightly larger than the fixed-study critical value and requires a sample size only slightly larger than the fixed-study sample size. These traits are exaggerated in this example with a single interim analysis, which explains why the fixed-study critical values, marked on the plot as black dots, overlie the critical values for the final look of the GSD. The information ratio of 1.0024 indicates that the GSD needs only 0.24% more information than a fixed study design; after rounding the sample size up to a whole number in each arm, both designs require a total of 242 participants.

◁

▷ Example 2: Sample size and efficacy bounds for an exact test of two proportions

In the previous example, we calculated sample sizes and bounds for a group sequential trial inspired by the DECREASE study, and we assumed that the researchers would analyze the results of the trial using the large-sample Pearson's χ^2 test. In reality, Poldermans et al. (1999, 1791) state that "differences between the groups in the rates of occurrence of the primary end point were evaluated by Fisher's exact test".

Sample sizes for Fisher's exact test can be estimated using the continuity correction of Casagrande, Pike, and Smith (1978), implemented in the continuity option. The rest of the study parameters remain the same, but to add variety, we specify the effect size in terms of the control-group proportion of 0.3 and the relative risk (p_2/p_1) of 0.5.

```
. gsdesign twoproportions 0.3, rrisk(0.5) continuity efficacy(obfleming)
> information(0.38 1)

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

Efficacy: O'Brien-Fleming

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.8000
      delta = 0.5000  (relative risk)
         p1 = 0.3000
         p2 = 0.1500
      rrisk = 0.5000

Expected sample size:
         H0 = 267.76
         Ha = 255.93

Info. ratio = 1.0024
    N fixed =    268
      N max =    268
     N1 max =    134
     N2 max =    134

Fixed-study crit. values = ±1.9600

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

──────────────────────────────────────────────────────────────────────────────────────
      │  Info.            Efficacy            │        Sample size
 Look │  frac.     Lower     Upper   p-value  │      N1         N2          N
──────┼───────────────────────────────────────┼───────────────────────────────────────
    1 │   0.38   -3.1878    3.1878    0.0014  │      51         51        102
    2 │   1.00   -1.9651    1.9651    0.0494  │     134        134        268
──────────────────────────────────────────────────────────────────────────────────────

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

The boundary critical values are the same as in example 1, but the continuity correction requires a slightly larger sample. Poldermans et al. (1999) report that the first look was conducted when data had been recorded from 53 participants in the control arm and 59 participants in the experimental arm. In practice, it is rare to conduct an analysis with exactly the desired sample size for that look, but type I error control is robust to minor deviations in attained sample size (DeMets et al. 1984).

At the time of the first look, 9 participants in the control arm had died due to postoperative cardiac causes and 9 more had nonfatal heart attacks, for a total of 18 participants who experienced the endpoint and 35 who did not. In the experimental arm, there were only 2 deaths from cardiac causes and no nonfatal heart attacks, giving a total of 2 participants who experienced the endpoint and 57 who did not.

We will repeat the analysis of the DECREASE trial using Fisher's exact test, but because the exact test does not produce a z statistic, we must use the significance level approach described in [ADAPT] **gsbounds**. We will compare the p-value from the exact test against the p-value reported in the table above. The rejection region at the first look is $|z_1| \geq 3.188$, which corresponds to a p-value ≤ 0.0014 using the significance level approach. We conduct Fisher's exact test using the immediate form of the `tabulate` command with the `exact` option.

```
. tabi 18 35 \ 2 57, exact

                      col
       row          1          2       Total

         1         18         35          53
         2          2         57          59

     Total         20         92         112
            Fisher's exact =                 0.000
    1-sided Fisher's exact =                 0.000
```

The two-sided p-value from the exact test was too small to be displayed in the output from `tabi`, but the value is saved as `r(p_exact)`.

```
. display r(p_exact)
.00002983
```

The p-value from the exact test is less than 0.0014, so we would reject H_0 at this look and terminate the trial early for treatment efficacy. This is the same action taken by the independent safety committee that performed the interim analysis of the DECREASE trial (Montori et al. 2005).

◁

▷ Example 3: Sample size, efficacy bounds, and futility bounds for a test of two proportions

In the previous example, we used O'Brien–Fleming efficacy bounds to re-create the design of the DECREASE clinical trial. Our design called for a maximum sample of 268 participants, the same size as the sample required by a fixed design with equivalent power and significance level. The actual DECREASE trial was terminated for efficacy at the first look, but a careful examination of the ESS from the design in example 2 reveals modest reductions in ESS over the fixed study design, suggesting room for improvement.

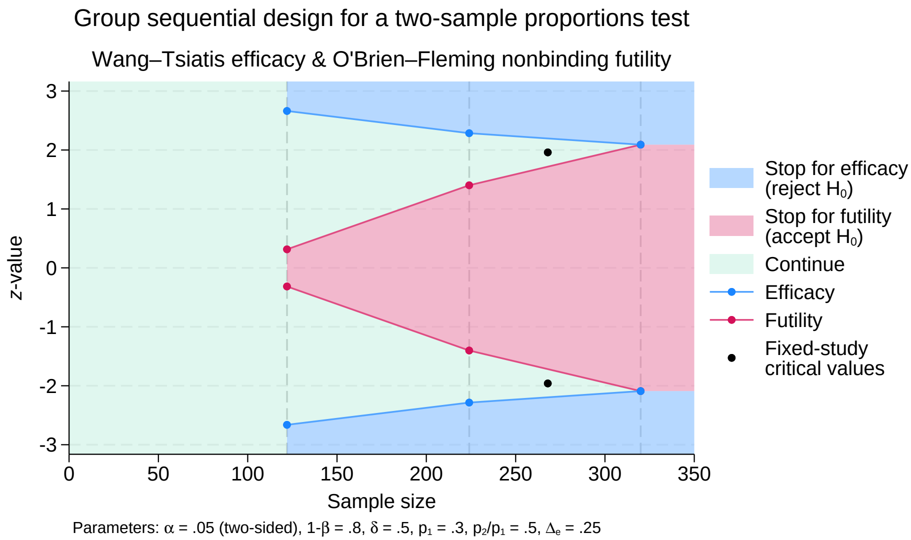
Here we modify the design of the DECREASE trial with the goal of lowering the ESS under both the null and alternative hypotheses without dramatically increasing the maximum sample size. To start, we will change the O'Brien–Fleming efficacy bound to a boundary that is somewhat less conservative at early looks, increasing the probability of early stopping for efficacy if H_a is true. One option is to use Pocock efficacy boundaries, which use the same critical value at all looks and are very effective at rejecting H_0 at early analyses. Unfortunately, the critical value at the final look of a Pocock design is much larger than the fixed-study critical value, and if the test statistic at the final look exceeds the fixed-study critical value but not the Pocock critical value, we will be unable to reject H_0 and will regret having chosen Pocock bounds.

Both O'Brien–Fleming and Pocock designs are members of the Wang–Tsiatis family of boundaries indexed by power parameter Δ, with $\Delta = 0$ for O'Brien–Fleming bounds and $\Delta = 0.5$ for Pocock bounds. We can split the difference between the two by using a Wang–Tsiatis bound with $\Delta = 0.25$ for a boundary that is somewhat less conservative at early looks but not dramatically larger than the fixed-study critical value at the final look.

The second change we make is adding nonbinding O'Brien–Fleming futility bounds to allow the trial to stop early if there is strong evidence that the treatment is not meaningfully different from the control. Nonbinding futility bounds give the independent Data Monitoring Committee the option of stopping the trial if a futility bound is crossed, but the trial is not required to stop; if it continues after crossing a nonbinding futility bound, the type I error is still controlled at the desired familywise significance level.

Our final change is to add another interim analysis approximately halfway between the first look (with 38% of the data) and the final data analysis. We modify the *numlist* provided to the `information()` option to include a second interim look with 70% of the data. Adding additional interim analyses provides more opportunities to stop the trial early, but conducting more hypothesis tests requires larger efficacy critical values to control type I error, so there is a tradeoff.

```
. gsdesign twoproportions 0.3, rrisk(0.5) continuity efficacy(wtsiatis(0.25))
> futility(obfleming) information(0.38 0.7 1) graphbounds

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

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

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.8000
      delta = 0.5000  (relative risk)
         p1 = 0.3000
         p2 = 0.1500
      rrisk = 0.5000

Expected sample size:
         H0 = 212.07
         Ha = 234.64

Info. ratio = 1.1915
    N fixed =    268
      N max =    320
     N1 max =    160
     N2 max =    160

Fixed-study crit. values = ±1.9600

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

      |  Info.              Efficacy                           Futility
 Look |  frac.      Lower     Upper   p-value       Lower     Upper   p-value
------+---------------------------------------------------------------------
    1 |   0.38    -2.6622    2.6622    0.0078     -0.3150    0.3150    0.7528
    2 |   0.70    -2.2851    2.2851    0.0223     -1.4017    1.4017    0.1610
    3 |   1.00    -2.0902    2.0902    0.0366     -2.0902    2.0902    0.0366

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

      |          Sample size
 Look |       N1         N2          N
------+--------------------------------
    1 |       61         61        122
    2 |      112        112        224
    3 |      160        160        320
```


Group sequential design for a two-sample proportions test

Wang–Tsiatis efficacy & O'Brien–Fleming nonbinding futility

Parameters: α = .05 (two-sided), 1-β = .8, δ = .5, p_1 = .3, p_2/p_1 = .5, Δ_e = .25

Figure 2. Two-sided test of the equality of two proportions with efficacy and futility bounds

As anticipated, the modified design has a smaller ESS under both the null and alternative hypotheses, with ESSs of 212.07 and 234.64, respectively. The maximum sample size, required if the trial continues to the final look, is 320 participants, approximately 19% more than the fixed-study sample size of 268.

The addition of nonbinding futility bounds raises the possibility of terminating the trial early to accept H_0. If the result of an interim analysis lies in the acceptance region, drawn on the graph in red, the Data Monitoring Committee is able to stop the trial for futility. If the committee decides to continue collecting data, the familywise type I error of the trial is still controlled at the desired 5% significance level.

`gsdesign twoproportions` displays the boundary critical values as z values and displays the corresponding p-values. When conducting Fisher's exact test, we must use the significance level approach to compare p-values from the tests against p-values corresponding to the boundary critical values.

Once data have been collected from 61 participants in each arm, the first interim analysis occurs and Fisher's exact test is conducted. If the two-sided p-value from the test, which we will denote p_1, is greater than 0.753, then it lies in the futility region and H_0 can be accepted, terminating the trial. If $p_1 \leq 0.008$, then it lies in the rejection region and we reject H_0, terminating the trial due to treatment efficacy. If $0.008 < p_1 \leq 0.753$, then p_1 is in the continuation region and the trial continues recruiting participants.

The testing procedure at the second look is similar, but the rejection and acceptance regions have grown and the continuation region has shrunk to $(0.022, 0.161]$. At the final look, the futility bound meets the efficacy bound, and there is no continuation region; if $p_3 \leq 0.037$, then we reject H_0, and if $p_3 > 0.037$, then we accept H_0.

◁

Stored results

gsdesign twoproportions 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(continuity)	1 if continuity correction is used, 0 otherwise
r(delta)	effect size
r(diff)	difference between the experimental- and control-group proportions (if diff() specified)
r(effparam)	efficacy parameter (if wtsiatis(), kdemets(), or hsdecani() specified)
r(ESS0)	expected sample size under null hypothesis
r(ESS1)	expected sample size under alternative hypothesis
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(N_fixed)	sample size of a fixed study design
r(N_fixedfrac)	fractional sample size of a fixed study design
r(N_max)	maximum sample size if the study continues to completion
r(N1_fixed)	sample size of the control group in a fixed study design
r(N1_fixedfrac)	fractional sample size of the control group in a fixed study design
r(N1_max)	maximum sample size of the control group if the study continues to completion
r(N2_fixed)	sample size of the experimental group in a fixed study design
r(N2_fixedfrac)	fractional sample size of the experimental group in a fixed study design
r(N2_max)	maximum sample size of the experimental group if the study continues to completion
r(nfractional)	1 if nfractional is specified, 0 otherwise
r(nlooks)	number of analyses
r(nratio)	specified ratio of sample sizes, $N2/N1$
r(nratio_a)	attained ratio of sample sizes
r(onesided)	1 for a one-sided test, 0 otherwise
r(oratio)	odds ratio (if oratio() specified)
r(p1)	control-group proportion
r(p2)	experimental-group proportion
r(pow_converged)	1 if power calculation iteration algorithm converged, 0 otherwise
r(pow_deltax)	final parameter tolerance achieved for power calculation
r(pow_ftolerance)	requested distance of power calculation objective function from 0
r(pow_function)	final distance of power calculation objective function from 0
r(pow_init)	initial value for power calculation sample size
r(pow_iter)	number of iterations performed for power calculation
r(pow_maxiter)	maximum number of iterations for power calculation
r(pow_tolerance)	requested parameter tolerance for power calculation
r(power)	specified overall power
r(power_a)	attained overall power
r(ratio)	ratio of the experimental-group proportion to the control-group proportion (if ratio() specified)
r(rdiff)	risk difference (if rdiff() specified)
r(rrisk)	relative risk (if rrisk() specified)
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)	gsdesign
r(cmdline)	command as typed
r(direction)	upper, lower, or two-sided
r(effbnd)	pocock, obfleming, wtsiatis, errpocock, errobfleming, kdemets, or hsdecani
r(effect)	specified effect: diff, ratio, etc.
r(futbnd)	pocock, obfleming, wtsiatis, errpocock, errobfleming, kdemets, or hsdecani
r(method)	twoproportions
r(test)	chi2 or lrchi2

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(bspent_a)	attained cumulative beta spent per look (when futility bounds are specified)
r(design)	sample size and stopping boundaries at interim looks
r(info_frac)	specified information fraction
r(info_frac_a)	fraction of attained information
r(info_level)	specified information level
r(p_crit)	p-values corresponding to boundary critical values
r(sampsize)	sample size at interim looks

Methods and formulas

Sample sizes at interim analyses are calculated as the product of the information fraction, the information ratio, and the sample size of a fixed-sample study.

See *Methods and formulas* in [ADAPT] **gsbounds** for the formulas used to calculate the stopping boundaries, information fraction, and information ratio. See *Methods and formulas* in [PSS-2] **power twoproportions** for the formulas used to calculate the sample size for a fixed study. See *Methods and formulas* in [ADAPT] **gsdesign** for the formulas used to calculate the ESS.

References

American Academy of Pediatrics Committee on Fetus and Newborn, American College of Obstetricians and Gynecologists Committee on Obstetric Practice, K. L. Watterberg, S. Aucott, W. E. Benitz, J. J. Cummings, E. C. Eichenwald, J. Goldsmith, B. B. Poindexter, K. Puopolo, D. L. Stewart, K. S. Wang, J. L. Ecker, J. R. Wax, A. E. B. Borders, Y. Y. El-Sayed, R. P. Heine, D. J. Jamieson, M. A. Mascola, H. L. Minkoff, A. M. Stuebe, J. E. Sumners, M. G. Tuuli, and K. R. Wharton. 2015. The Apgar score. *Pediatrics* 136: 819–822. https://doi.org/10.1542/peds.2015-2651.

Casagrande, J. T., M. C. Pike, and P. G. Smith. 1978. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics* 34: 483–486. https://doi.org/10.2307/2530613.

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.

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. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70: 659–663. https://doi.org/10.1093/biomet/70.3.659.

Montori, V. M., P. J. Devereaux, N. K. J. Adhikari, K. E. A. Burns, C. H. Eggert, M. Briel, C. Lacchetti, T. W. Leung, E. Darling, D. M. Bryant, H. C. Bucher, H. J. Schünemann, M. O. Meade, D. J. Cook, P. J. Erwin, A. Sood, R. Sood, B. Lo, C. A. Thompson, Q. Zhou, E. Mills, and G. H. Guyatt. 2005. Randomized trials stopped early for benefit: A systematic review. *Journal of the American Medical Association* 294: 2203–2209. https://doi.org/10.1001/jama.294.17.2203.

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.

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.

Poldermans, D., E. Boersma, J. J. Bax, I. R. Thomson, L. L. M. van de Ven, J. D. Blankensteijn, H. F. Baars, T.-I. Yo, G. Trocino, C. Vigna, J. R. T. C. Roelandt, P. M. Fioretti, B. Paelinck, and H. van Urk. 1999. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. *New England Journal of Medicine* 341: 1789–1794. https://doi.org/10.1056/NEJM199912093412402.

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

[ADAPT] **gsbounds** — Boundaries for group sequential trials

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

[ADAPT] **gsdesign oneproportion** — Group sequential design for a one-sample proportion test

[ADAPT] **Glossary**

[PSS-2] **power twoproportions** — Power analysis for a two-sample proportions test

[R] **prtest** — Tests of proportions

[R] **tabulate twoway** — Two-way table of frequencies

gsdesign logrank — Group sequential design for a log-rank test

Description	Quick start	Menu	Syntax	Options
Remarks and examples	Stored results	Methods and formulas	References	Also see

Description

gsdesign logrank computes stopping boundaries and sample sizes for interim analyses of group sequential clinical trials performing survival analysis by using a log-rank test to compare survivor functions. Stopping can be for efficacy, futility, or both. For stopping boundary calculations without sample sizes, see [ADAPT] **gsbounds**. For sample-size calculations for a fixed-sample study using a log-rank test, see [PSS-2] **power logrank**.

Quick start

Sample size and stopping boundaries for the log-rank test of H_0 : hazard ratio $\Delta = 1$ versus two-sided alternative $H_a : \Delta \neq 1$, with default familywise significance level $\alpha = 0.05$ and power of 0.8 to detect a hazard ratio of $\Delta_a = 0.737$ without censoring, using default group sequential specifications of O'Brien–Fleming efficacy boundaries with two analyses (one interim, one final)

```
gsdesign logrank, hratio(0.737)
```

Same as above, specified as a log hazard-ratio of −0.305

```
gsdesign logrank, lnhratio(-0.305)
```

Same as above, but use Schoenfeld's sample-size calculation instead of the default Freedman method

```
gsdesign logrank, lnhratio(-0.305) schoenfeld
```

Sample size for censored design with survival probabilities $surv_1 = 0.5$ and $surv_2 = 0.6$, using a Kim–DeMets efficacy boundary with parameter $\rho_e = 3$ and analyses at 50%, 75%, and 100% of the total data

```
gsdesign logrank 0.5 0.6, efficacy(kdemets(3)) information(50 75 100)
```

Same as above, specified as $surv_1 = 0.5$ and hazard ratio of 0.737

```
gsdesign logrank 0.5, hratio(0.737) efficacy(kdemets(3))        ///
     information(50 75 100)
```

Same as above, but add a binding O'Brien–Fleming-style futility bound

```
gsdesign logrank 0.5, hratio(0.737) efficacy(kdemets(3))        ///
     futility(errobfleming, binding) information(50 75 100)
```

Same as above, but report fractional sample sizes and graph the boundaries

```
gsdesign logrank 0.5, hratio(0.737) nfractional efficacy(kdemets(3)) ///
     futility(errobfleming, binding) information(50 75 100)     ///
     graphbounds
```

Menu

Statistics > Power, precision, and sample size

Syntax

`gsdesign logrank` [*surv*$_1$ [*surv*$_2$]] [, *logrankopts boundopts*]

where *surv*$_1$ is the survival probability in the control (reference) group at the end of the study, and *surv*$_2$ is the survival probability in the experimental (treatment) group at the end of the study.

logrankopts	Description
Main	
`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)`
`onesided`	request a one-sided test; default is two-sided
`nfractional`	report fractional sample size
`nratio(#)`	ratio of sample sizes of experimental to control groups; default is `nratio(1)`, meaning equal group sizes
`hratio(#)`	hazard ratio of the experimental to the control group; default is `hratio(0.5)`; may not be combined with `lnhratio()`
`lnhratio(#)`	log hazard-ratio of the experimental to the control group; may not be combined with `hratio()`
`schoenfeld`	use the formula based on the log hazard-ratio in calculations; default is to use the formula based on the hazard ratio
`effect(`*effect*`)`	type of effect to display; default is `effect(hratio)` unless option `schoenfeld` is specified, in which case it is `effect(lnhratio)`
Censoring	
`simpson(# # #`|*matname*`)`	survival probabilities in the control group at three specific time points to compute the probability of an event (failure), using Simpson's rule under uniform accrual
`st1(`*varname*$_s$ *varname*$_t$`)`	variables *varname*$_s$, containing survival probabilities in the control group, and *varname*$_t$, containing respective time points, to compute the probability of an event (failure), using numerical integration under uniform accrual
`wdprob(#)`	proportion of subjects anticipated to withdraw from the study; default is `wdprob(0)`
`force`	allow calculation with unsupported `power logrank` options

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

`force` does not appear in the dialog box.

effect	Description
hratio	hazard ratio
lnhratio	log hazard-ratio

boundopts	Description
Bounds	
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(#[, equal])	total number of analyses (nlooks() − 1 interim analyses and one final analysis); use equal to enforce equal information increments; if neither nlooks() nor information() is specified, the default is nlooks(2)
information(*numlist*)	sequence of information levels for analyses; default is evenly spaced
nopvalues	suppress p-values
Graph	
graphbounds[(*graphopts*)]	graph boundaries
matlistopts(*general_options*)	control the display of boundaries and sample size; seldom used
optimopts	optimization options for boundary calculations; seldom used

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

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

graphopts	Description
xdimsampsize	label the x axis with the sample size collected (default)
xdiminformation	label the x axis with the information fraction; 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

alpha(#) sets the overall significance level, which is the familywise type I error rate for all analyses (interim and final). alpha() must be in $(0, 0.5)$. The default is alpha(0.05).

power(#) sets the overall power for all analyses. power() must be in $(0.5, 1)$. 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. beta() must be in $(0, 0.5)$. 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.

onesided requests a study design for a one-sided test. The direction of the test is inferred from the effect size.

nfractional specifies that fractional sample sizes be reported.

nratio(#) specifies the sample-size ratio of the experimental group relative to the control group, $N2/N1$. The default is nratio(1), meaning equal allocation between the two groups.

hratio(#) specifies the hazard ratio (effect size) of the experimental group to the control group. The default is hratio(0.5). This value typically defines the clinically significant improvement of the experimental procedure over the control procedure desired to be detected by the log-rank test with a certain power.

You can specify an effect size either as a hazard ratio in hratio() or as a log hazard-ratio in lnhratio(). The default is hratio(0.5). If both arguments $surv_1$ and $surv_2$ are specified, hratio() is not allowed and the hazard ratio is instead computed as $\ln(surv_2)/\ln(surv_1)$.

This option may not be combined with lnhratio().

lnhratio(#) specifies the log hazard-ratio (effect size) of the experimental group to the control group. This value typically defines the clinically significant improvement of the experimental procedure over the control procedure desired to be detected by the log-rank test with a certain power.

You can specify an effect size either as a hazard ratio in hratio() or as a log hazard-ratio in lnhratio(). The default is hratio(0.5). If both arguments $surv_1$ and $surv_2$ are specified, lnhratio() is not allowed and the log hazard-ratio is computed as $\ln\{\ln(surv_2)/\ln(surv_1)\}$.

This option may not be combined with hratio().

schoenfeld requests calculations using the formula based on the log hazard-ratio, according to Schoenfeld (1981). The default is to use the formula based on the hazard ratio, according to Freedman (1982). See the technical note in [PSS-2] **power logrank** for a comparison of the two formulas.

effect(*effect*) specifies the type of effect size to be reported in the output as delta. *effect* is one of hratio or lnhratio. By default, the effect size delta is a hazard ratio, effect(hratio), for a hazard-ratio test and a log hazard-ratio, effect(lnhratio), for a log hazard-ratio test (when schoenfeld is specified).

Censoring

simpson(# # #|*matname*) specifies survival probabilities in the control group at three specific time points to compute the probability of an event (failure) using Simpson's rule under the assumption of uniform accrual. Either the actual values or a 1×3 matrix, *matname*, containing these values can be specified. By default, the probability of an event is approximated as an average of the failure probabilities $1-s_1$ and $1-s_2$; see *Methods and formulas* in [PSS-2] **power logrank**. `simpson()` may not be combined with `st1()` and may not be used if command argument *surv*$_1$ or *surv*$_2$ is specified.

`st1(`*varname*$_s$ *varname*$_t$`)` specifies variables *varname*$_s$, containing survival probabilities in the control group, and *varname*$_t$, containing respective time points, to compute the probability of an event (failure) using numerical integration under the assumption of uniform accrual; see [R] **dydx**. The minimum and the maximum values of *varname*$_t$ must be the length of the follow-up period and the duration of the study, respectively. By default, the probability of an event is approximated as an average of the failure probabilities $1-s_1$ and $1-s_2$; see *Methods and formulas* in [PSS-2] **power logrank**. `st1()` may not be combined with `simpson()` and may not be used if command argument *surv*$_1$ or *surv*$_2$ is specified.

`wdprob(#)` specifies the proportion of subjects anticipated to withdraw from the study. The default is `wdprob(0)`.

Bounds

`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(#` [`, equal`]`)` 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)`.

`equal` indicates that equal information increments be enforced, which is to say that the same number of new observations will be collected at each look. The default behavior is to start by dividing information evenly among looks, then proceed by rounding up to a whole number of observations at each look. This can cause slight differences in the information collected at each look.

`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.

Graph

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

graphopts are the following:

`xdimsampsize` labels the x axis with the sample size collected (the default).

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

`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 `gsdesign logrank` but are not shown in the dialog box:

`force` indicates that `gsdesign logrank` should allow unsupported `power logrank` options, such as options specifying a cluster randomized design. Even with option `force`, the `power logrank` options specified must be compatible with sample-size determination, not effect size or power calculation. In addition, *numlist*s are not supported in options or in arguments as they are with `power`, even when `force` is specified.

`matlistopts(`*general_options*`)` affects the display of the matrix of boundaries and sample sizes. *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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**. 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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

`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 $\mathbf{g}\mathbf{H}^{-1}\mathbf{g}' <$ 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:

Introduction
Using gsdesign logrank
Background for examples
Computing sample size and boundaries in the absence of censoring
Computing sample size and boundaries in the presence of censoring
Computing sample size and boundaries with uniform accrual

This entry describes the use of the gsdesign logrank command for designing a group sequential analysis for a two-sample comparison of survivor functions using the log-rank test. See [ADAPT] **GSD intro** for a general introduction to group sequential designs (GSDs) for clinical trials; see [ADAPT] **gsbounds** for information about group sequential bounds; and see [ADAPT] **gsdesign** for information about designing group sequential clinical trials with the gsdesign command. Also see [PSS-2] **Intro (power)** for a general introduction to power and sample-size analysis, and see [PSS-2] **power logrank** for details about fixed-sample study designs for a log-rank test of two survivor functions.

Introduction

When analyzing time-to-event data, we frequently use the failure function, the survivor function, and the hazard function, denoted $F(t)$, $S(t)$, and $h(t)$, respectively. The failure function is the probability of experiencing a failure event at or before time t. If we denote the time of failure as T, we can define the failure function as the cumulative distribution function of T, where $F(t) = \Pr(T \leq t)$. The probability density function of T is the derivative of the failure function with respect to time, written as $f(t) = \partial F(t)/\partial t$. The survivor function is defined as the probability of surviving beyond time t, expressed mathematically as $S(t) = \Pr(T > t) = 1 - F(t)$. The hazard function at time t is the instantaneous rate of failure at time t, conditional on survival until time t, written as $h(t) = f(t)/S(t)$.

Consider a survival study comparing survivor functions in two groups using the log-rank test. 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 the most powerful nonparametric test of $S_1(\cdot) = S_2(\cdot)$ if the hazard functions are proportional. That is, $h_2(t) = \Delta h_1(t)$ for all t or, equivalently, $S_2(t) = \{S_1(t)\}^{\Delta}$, where Δ is the hazard ratio. If $\Delta < 1$, 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. Under the proportional-hazards assumption, the test of the equality of the two survivor functions $H_0: S_1(t) = S_2(t)$ versus $H_a: S_1(t) \neq S_2(t)$ is equivalent to testing $H_0: \Delta = 1$ versus $H_a: \Delta \neq 1$ or $H_0: \ln(\Delta) = 0$ versus $H_a: \ln(\Delta) \neq 0$.

The methods implemented in `gsdesign logrank` for boundary and sample-size calculations relate the power of the log-rank test directly to 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. In the presence of administrative censoring or withdrawal, `gsdesign logrank` requires additional information to estimate the probability that a participant's failure time will be observed.

Using gsdesign logrank

`gsdesign logrank` computes stopping boundaries and sample size for a log-rank test comparing the survivor functions in two groups. `gsdesign logrank` can be thought of as a combination of `power logrank` for sample-size calculations and `gsbounds` for stopping boundary calculations. `gsdesign logrank` supports two methods of estimating the required sample size: the method of Freedman (1982), which uses a formula based on the hazard ratio and is the default, and the method of Schoenfeld (1981), which uses a formula based on the log hazard-ratio.

To determine the required number of events, the investigator must specify the effect size. Effect size is usually expressed as a hazard ratio, Δ_a, by using the `hratio()` option. Alternatively, you may specify the effect size as a log hazard-ratio, $\ln(\Delta_a)$, with the `lnhratio()` option. When administrative censoring is anticipated, the survival probabilities of the two groups, $surv_1$ and $surv_2$, may be specified and the effect size is calculated from the survival probabilities. If the effect size is not specified, a hazard ratio of 0.5 is assumed.

By default, all computations assume no censoring. In the presence of administrative censoring, you must specify a survival probability at the end of the study in the control group as the first command argument, *surv*$_1$. You can also specify a survival probability at the end of the study in the experimental group as the second command argument, *surv*$_2$. Otherwise, it will be computed using the specified hazard ratio or log hazard-ratio and the control-group survival probability. To accommodate an accrual period under the assumption of uniform accrual, survival information may instead be supplied in the `simpson()` option or the `st1()` option; see *Including information about subject accrual* in [PSS-2] **power logrank** for details. To adjust the sample-size calculation for withdrawal from the trial, specify the anticipated proportion of withdrawals in the `wdprob()` option.

By default, `gsdesign logrank` assumes that the control and experimental arms will be the same size. If participants are not allocated equally between the two arms, the `nratio()` option is used to specify the ratio of participants in the experimental arm to the control arm.

Options `alpha()`, `power()`, `beta()`, and `onesided` are used for both sample-size and stopping-boundary calculations. The default significance level, known as the familywise type I error rate, is 0.05 and can be changed by specifying the `alpha()` option. The default power is 0.8, which corresponds to a type II error rate of 0.2. This can be modified either by specifying the power in the `power()` option or by specifying the type II error in the `beta()` option. The default test is two-sided, and the `onesided` option requests a one-sided test, the direction of which is indicated by the effect size.

The group sequential stopping rule is determined by the `efficacy()` and `futility()` options. Stopping can be for efficacy, futility, or both, and if no stopping rule is specified, the default is to use an O'Brien–Fleming efficacy bound. If futility bounds are requested, the default behavior is to treat them as nonbinding. A trial that crosses a nonbinding futility bound can be stopped for futility, but the familywise type I error is controlled even if the trial continues. Binding futility bounds can be requested with `futility()` suboption `binding`. A trial that crosses a binding futility bound must be stopped for futility; if it continues, the familywise type I error will not be controlled at the specified significance level.

The number of looks, or analyses of the trial data, is specified with `nlooks()`. Alternatively, the `information()` option can be used to specify the spacing of the looks as a *numlist* of increasing information levels. In this case, values of the numlist are automatically rescaled so that the final look has the maximum information required by the design. If neither `nlooks()` nor `information()` is specified, the default is two looks.

By default, the sample sizes in each arm are rounded up to whole numbers at each look, but the `nfractional` option can be used to report fractional sample sizes. If `nlooks()` is specified, the default behavior is to divide information evenly among each look before rounding. Rounding can cause slight differences in the amount of information collected at each look, and `nlooks()` suboption `equal` can be specified to enforce equal information increments by requiring the same number of new observations per arm at each look.

Background for examples

In the following examples, we consider designing a clinical trial of a treatment for hepatocellular carcinoma, the most common type of primary liver cancer. In 2023, Peng et al. described the results of the LAUNCH trial, a phase 3 randomized controlled trial comparing lenvatinib monotherapy (the control arm) against lenvatinib plus transarterial chemoembolization (the experimental arm) as a treatment for primary advanced hepatocellular carcinoma. The primary endpoint of the trial was overall survival, the time from randomization to death from any cause.

Lenvatinib is an anti-cancer medication that can be taken orally and is used to treat some thyroid, kidney, and liver cancers, including hepatocellular carcinoma. Transarterial chemoembolization is a procedure where a catheter is inserted in the artery supplying blood to the tumor, and small particles with injectable anti-cancer drugs are introduced directly into the area of the tumor, blocking off the tumor's blood supply and providing a concentrated dose of chemotherapeutic medication.

Computing sample size and boundaries in the absence of censoring

▷ Example 1: GSD for a log-rank test with O'Brien–Fleming efficacy bounds

Peng et al. (2023) randomized participants to the control and experimental arms in a 1:1 ratio and conducted a log-rank test of $H_0: \Delta = 1$ versus the two-sided alternative $H_a: \Delta \neq 1$ with a familywise significance level of 5%. They required 90% power to detect a hazard ratio of $\Delta_a = 0.67$ and planned a single interim analysis using classical O'Brien–Fleming efficacy bounds once two-thirds of the data had been collected. We use `gsdesign logrank` to design and graph the boundaries of a clinical trial with these parameters.

```
. gsdesign logrank, hratio(0.67) power(0.9) efficacy(obfleming)
> information(0.667 1) graphbounds
Group sequential design for two-sample comparison of survivor functions
Log-rank test, Freedman method
H0: HR = 1 versus Ha: HR != 1
Efficacy: O’Brien–Fleming
Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
      delta = 0.6700  (hazard ratio)
     hratio = 0.6700
Censoring:
       Pr_E = 1.0000
Expected number of events:
         H0 = 272.71
         Ha = 220.55
Info. ratio = 1.0155
    E fixed =    270
    N fixed =    270
      N max =    274
     N1 max =    137
     N2 max =    137
Fixed-study crit. values = ±1.9600
Critical values, p-values, and sample sizes for a group sequential design

           Info.              Efficacy                   Events
     Look  frac.       Lower     Upper   p-value              E

        1   0.67     -2.4524    2.4524    0.0142            183
        2   1.00     -2.0028    2.0028    0.0452            274

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

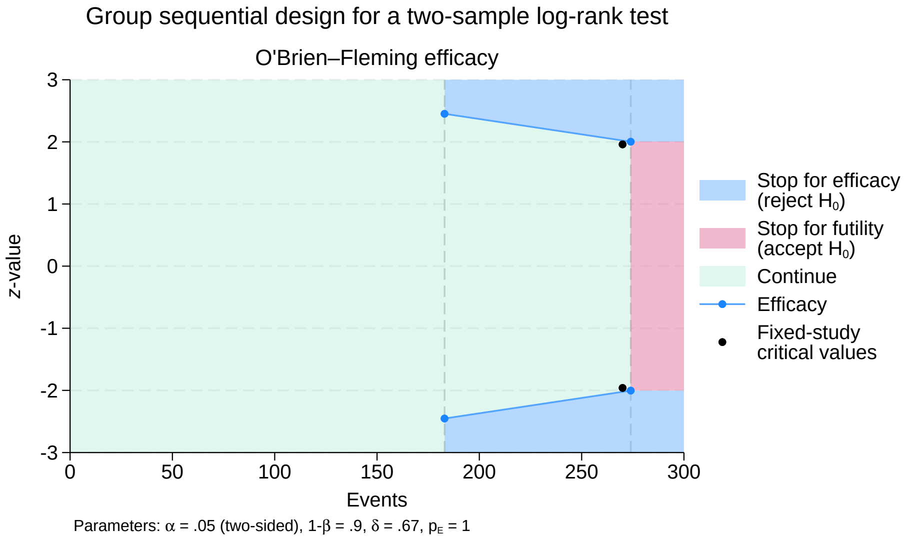


Figure 1. GSD with O’Brien–Fleming efficacy bounds for a two-sided log-rank test

gsdesign logrank displays the specified study parameters, including hratio, the hazard ratio under the alternative hypothesis, and Pr_E, the probability that a participant will die by the end of the study.

The next section of the output displays the expected number of events, which is the average number of events if the group sequential trial were to be repeated many times. The following section reports 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.

If the null hypothesis of $H_0 : \Delta = 1$ were true, the control and experimental arms of the trial would have equal hazards. In this case, the average trial would require 272.71 events, nearly the full sample size of 274. This is because the efficacy bounds do not allow for early stopping to accept H_0, so if the null hypothesis is true, the trial will usually proceed to the final look. If H_a is true, the average trial will require 220.55 events, which is a savings over the 270 events required by the fixed trial.

The table at the end of the output displays the critical values for the stopping boundaries and the corresponding p-values as well as 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.

We planned the first look to occur with 66.7% of the data, which corresponds to 183 events. Examining the graph, we see that the entire region from 0 to 182 events is shaded green, the color of the continuation region. This is because the data have not yet been analyzed, so the trial cannot be stopped. The first look will be conducted once 183 deaths have occurred, and a log-rank test will be performed. We denote the square root of the χ^2 test statistic from the first look as z_1 and 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 $|z_1| \geq 2.452$, 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 $|z_1| < 2.452$, it lies in the continuation region, and we proceed to the final look.

At the final look, there is no continuation region; H_0 must be rejected or accepted. While accepting the null hypothesis is taboo in many disciplines, it has a long history in the context of sequential trials (see *Origins of GSD* in [ADAPT] **GSD intro**). As before, we take the square root of the χ^2 test statistic and label it z_2. If $|z_2| \geq 2.003$, then we reject H_0 and conclude that $\Delta \neq 1$, while if $|z_2| < 2.003$, then we accept H_0.

◁

Computing sample size and boundaries in the presence of censoring

▷ Example 2: GSD for a log-rank test with censoring

In the previous example, we assumed no censoring would occur, so the failure time of all participants would be observed. That is often an unrealistic expectation, and here we adjust the sample size to account for censoring. We divide censoring into two types: administrative censoring, which occurs when the trial ends before all participants have experienced a failure event, and withdrawal, which occurs when a participant withdraws from the study or is lost to follow-up. gsdesign logrank takes a conservative stance on withdrawal, assuming that participants who withdraw do so as soon as the study begins, before they can contribute meaningful data to the trial. For more information about censoring, see *Computing sample size in the presence of censoring* in [PSS-2] **power logrank**.

Peng et al. (2023) describe an anticipated withdrawal rate of 10%, which we will incorporate using the `wdprob()` option. Based on a previous study of lenvatinib as a treatment for hepatocellular carcinoma (Kudo et al. 2018), we anticipate that 5% of the participants in the control arm will be alive at the end of the trial. We include the control-group survival probability as command argument $surv_1$.

```
. gsdesign logrank 0.05, hratio(0.67) wdprob(0.1) power(0.9)
> efficacy(obfleming) information(0.667 1)

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

Efficacy: O'Brien-Fleming

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
      delta = 0.6700  (hazard ratio)
     hratio = 0.6700

Censoring and withdrawal:
         s1 = 0.0500
         s2 = 0.1344
       Pr_E = 0.9078
       Pr_w = 0.1000

Expected number of events:
         H0 = 272.71
         Ha = 220.55

Info. ratio = 1.0155
    E fixed =    270
    N fixed =    330
      N max =    336
     N1 max =    168
     N2 max =    168

Fixed-study crit. values = ±1.9600

Critical values, p-values, and sample sizes for a group sequential design
----------------------------------------------------------
      |  Info.            Efficacy          |       Events
 Look |  frac.      Lower     Upper  p-value |            E
------+-------------------------------------+-------------
    1 |   0.67    -2.4524    2.4524   0.0142 |          183
    2 |   1.00    -2.0028    2.0028   0.0452 |          274
----------------------------------------------------------

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

In addition to `Pr_E`, the probability that an event is observed by the end of the study, the output of `gsdesign logrank` now includes additional information about censoring and withdrawal. The survival probabilities of the control and experimental arms are labeled `s1` and `s2`, respectively, and the probability of withdrawal is labeled `Pr_w`.

Now that censoring is incorporated into the design, we must recruit a larger sample of 168 participants in each arm, but the number of events at each look is unchanged, as are the critical values of the efficacy bound. This is to be expected because the power of the log-rank test is calculated in terms of the number of events observed, which is not affected by censoring.

◁

Computing sample size and boundaries with uniform accrual

▷ Example 3: GSD for a log-rank test with uniform accrual

In example 2, we considered the effect of censoring on the study design, but we did not account for the fact that the first participants recruited to the study would be observed for longer than the last participants to join. Peng et al. (2023) describe the recruitment period as lasting 24 months and the follow-up period as lasting an additional 24 months. The first participants to join the study would be monitored for up to 48 months (or until they died or withdrew from the study), while the last participants would only be monitored for 24 months.

If participants are recruited to the study at a uniform rate, Schoenfeld (1983) recommends a sample-size calculation that involves estimating the integral of the survivor function using Simpson's rule. `gsdesign logrank` implements this calculation with the `simpson()` option, which requires estimates of the survival probability in the control group at three points: at the minimum follow-up time, halfway between the minimum and maximum follow-up times, and at the maximum follow-up time. This corresponds to 24 months, 36 months, and 48 months in the LAUNCH study.

Based on the previous work of Kudo et al. (2018) and the assumptions made by Peng et al. (2023) about the tumor burden in their population of interest, we predict that 20% of participants in the control arm will be alive 24 months after they join the study, 10% will be alive after 36 months, and 5% will be alive after 48 months.

```
. gsdesign logrank, hratio(0.67) simpson(0.2 0.1 .05) wdprob(0.1) power(0.9)
> efficacy(obfleming) information(0.667 1)

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

Efficacy: O'Brien-Fleming

Study parameters:
      alpha = 0.0500  (two-sided)
      power = 0.9000
      delta = 0.6700  (hazard ratio)
     hratio = 0.6700

Censoring and withdrawal:
       Pr_E = 0.8350
       Pr_w = 0.1000

Expected number of events:
         H0 = 272.71
         Ha = 220.55

Info. ratio = 1.0155
    E fixed =    270
    N fixed =    360
      N max =    364
     N1 max =    182
     N2 max =    182

Fixed-study crit. values = ±1.9600
```

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

────────┬──────────────────────────────────────────┬───────────────
        │  Info.             Efficacy              │    Events
  Look  │  frac.      Lower      Upper    p-value  │          E
────────┼──────────────────────────────────────────┼───────────────
     1  │   0.67    -2.4524     2.4524     0.0142  │        183
     2  │   1.00    -2.0028     2.0028     0.0452  │        274
────────┴──────────────────────────────────────────┴───────────────
Note: Critical values are for z statistics; otherwise, use
      p-value boundaries.
```

In example 2, we calculated the overall probability of observing a failure event, `Pr_E`, to be 0.908. After taking accrual into account, we now estimate `Pr_E` to be 0.835. The reduced chance of observing a failure event corresponds to a larger required sample size: 360 participants for a fixed-sample trial (up from 330) and a maximum of 364 participants for the GSD (up from 336).

Altering the assumptions about participant accrual affected the required sample size but not the number of failures needed to attain 90% power or the critical values required to achieve a 5% familywise significance level.

◁

▷ Example 4: GSD for a one-sided log-rank test with efficacy and futility stopping

In the previous examples, we endeavored to design a clinical trial modeled after the LAUNCH study with increasingly sophisticated estimates of the probability of observing a failure event. We did not modify the details of the interim analysis, a relatively simple design with an O'Brien–Fleming efficacy bound and a a single look at 66.7% of the total number of failure events. Here we depart from the design of Peng et al. (2023) and calculate stopping bounds and sample sizes for a study with both efficacy and futility stopping.

Futility stopping is the complement of efficacy stopping, and it allows the trial to end early if interim results are overwhelmingly unfavorable. This is done by calculating a futility bound, which we will draw on the graph in red. Much as the efficacy bound shown in figure 1 divided the range of interim test statistics into continuation and rejection regions, the futility bound further partitions the range of test statistics by defining an acceptance region. With efficacy-only stopping, the acceptance region only existed at the final look, but with futility stopping, it is possible to accept H_0 before the scheduled end of the trial.

There are two types of futility boundaries: nonbinding (the default) and binding. In a trial with binding futility bounds, if an interim test statistic lies in the acceptance region, the trial must be terminated for futility; if it continues, the familywise type I error will not be controlled at the desired significance level. In contrast, if an interim test statistic crosses a nonbinding futility bound, the Data Monitoring Committee can decide to halt the trial or allow it to continue without risk of overrunning the specified alpha level.

We choose to implement a nonbinding Hwang–Shih–de Cani futility boundary with parameter $\gamma_f = -4$. Hwang–Shih–de Cani bounds are calculated using the error-spending approach, which makes them incompatible with classical O'Brien–Fleming bounds. Fortunately, there is an error-spending approximation of the O'Brien–Fleming bound that we can use instead (see *Methods and formulas* in [ADAPT] **gsbounds** for more information about classical and error-spending bounds). We specify error-spending O'Brien–Fleming-style efficacy bounds with the `efficacy(errobfleming)` option.

We anticipate that the hazard ratio will be less than 1, indicating that the experimental treatment of lenvatinib plus transarterial chemoembolization is superior to the control treatment of lenvatinib alone. As such, we request a one-sided test and reduce the significance level to half of what it was with a two-sided alternative hypothesis. Finally, we add an additional interim analysis once half of the total number of failure events have been observed.

```
. gsdesign logrank, hratio(0.67) simpson(0.2 0.1 0.05) wdprob(0.1) onesided
> power(0.9) alpha(0.025) efficacy(errobfleming)
> futility(hsdecani(-4)) information(0.5 0.667 1) graphbounds

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 Hwang-Shih-de Cani, nonbinding, gamma = -4.0000

Study parameters:
      alpha = 0.0250  (lower one-sided)
      power = 0.9000
      delta = 0.6700  (hazard ratio)
     hratio = 0.6700

Censoring and withdrawal:
       Pr_E = 0.8350
       Pr_w = 0.1000

Expected number of events:
         H0 = 183.14
         Ha = 211.81

Info. ratio = 1.0306
    E fixed =    270
    N fixed =    360
      N max =    370
     N1 max =    185
     N2 max =    185

Fixed-study crit. value = -1.9600

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

──────┬─────────────────────────────────────────────────┬────────────
      │ Info.        Efficacy             Futility      │    Events
 Look │ frac.      Lower   p-value      Upper   p-value │           E
──────┼─────────────────────────────────────────────────┼────────────
    1 │  0.50    -2.9626    0.0015    -0.0672    0.4732 │         139
    2 │  0.67    -2.5374    0.0056    -0.6491    0.2581 │         185
    3 │  1.00    -1.9945    0.0230    -1.9945    0.0230 │         278
──────┴─────────────────────────────────────────────────┴────────────

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

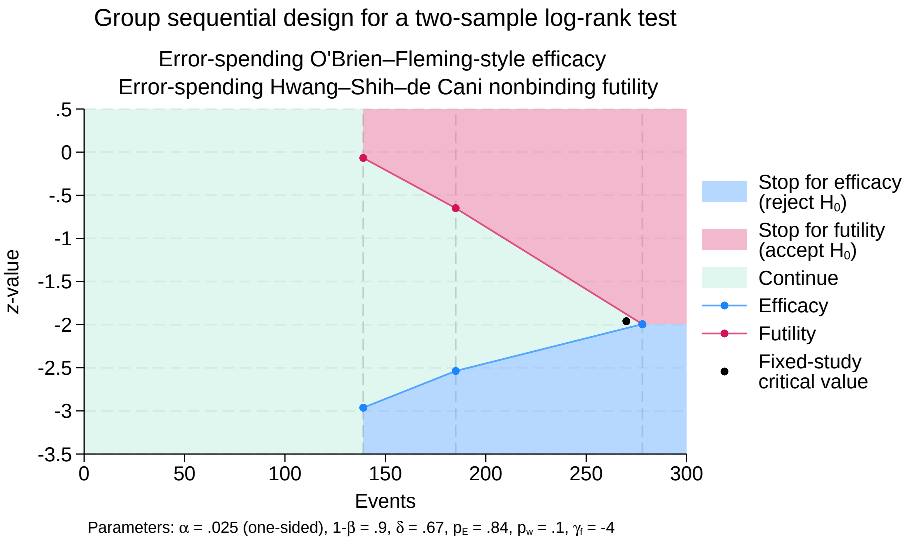


Figure 2. Error-spending efficacy and futility bounds for a one-sided log-rank test

The output of gsdesign logrank starts off quite similar to that of example 3, but the alternative hypothesis is now reported as Ha: HR < 1. By halving the significance level when transitioning to a one-sided test, we have kept the number of participants and events required by a fixed study unchanged.

Compared with the design in example 3, the maximum sample size has increased from 364 to 370, but the expected number of events has changed much more dramatically. Without futility stopping, 272.71 events were expected under H_0, but that has decreased to 183.14 events now that the trial can be stopped early to accept the null hypothesis. The expected number of events under H_a has decreased as well, from 220.55 to 211.81, due to the additional opportunity to stop the trial for efficacy once half of the data have been collected.

Once 139 events have been recorded, the log-rank test is conducted and z_1, the square root of the χ^2 statistic, is calculated. If the hazard ratio is less than 1, then z_1 is negative; if the hazard ratio is greater than 1 (meaning the control is outperforming the experimental treatment), then z_1 is positive. If $z_1 \leq -2.963$, then the test statistic lies within the rejection region, so we reject H_0 and terminate the trial early due to treatment efficacy. If $z_1 > -0.067$, then it lies within the acceptance region and we have the option of terminating the trial due to futility. If z_1 lies in the continuation region of $(-2.963, -0.067]$, then the trial must continue.

The second look occurs once 185 failures have been observed, and the testing procedure is similar except the continuation region has shrunk to $(-2.537, -0.649]$. At the final look, the efficacy and futility critical values are the same, leaving no continuation region. If $z_3 \leq -1.995$, then we reject H_0; otherwise, H_0 is accepted.

◁

Stored results

gsdesign logrank 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(delta)	effect size
r(E_fixed)	number of events in a fixed study design
r(E_fixedfrac)	fractional number of events in a fixed study design
r(E_max)	maximum number of events if the study continues to completion
r(effparam)	efficacy parameter (if wtsiatis(), kdemets(), or hsdecani() specified)
r(ESS0)	expected sample size under null hypothesis
r(ESS1)	expected sample size under alternative hypothesis
r(futparam)	futility parameter (if wtsiatis(), kdemets(), or hsdecani() specified)
r(hratio)	hazard ratio (unless lnhratio() specified)
r(info_ratio)	ratio of maximum information required to that of a fixed study design
r(lnhratio)	log hazard-ratio (if lnhratio() specified)
r(N_fixed)	sample size of a fixed study design
r(N_fixedfrac)	fractional sample size of a fixed study design
r(N_max)	maximum sample size if the study continues to completion
r(N1_fixed)	sample size of the control group in a fixed study design
r(N1_fixedfrac)	fractional sample size of the control group in a fixed study design
r(N1_max)	maximum sample size of the control group if the study continues to completion
r(N2_fixed)	sample size of the experimental group in a fixed study design
r(N2_fixedfrac)	fractional sample size of the experimental group in a fixed study design
r(N2_max)	maximum sample size of the experimental group if the study continues to completion
r(nfractional)	1 if nfractional is specified, 0 otherwise
r(nlooks)	number of analyses
r(nratio)	specified ratio of sample sizes, $N2/N1$
r(nratio_a)	attained ratio of sample sizes
r(onesided)	1 for a one-sided test, 0 otherwise
r(power)	specified overall power
r(power_a)	attained overall power
r(Pr_E)	probability of an event (failure)
r(Pr_w)	proportion of withdrawals
r(s1)	survival probability in the control group (if specified)
r(s2)	survival probability in the experimental group (if specified)
r(stop)	0 for futility bounds, 1 for efficacy bounds, 2 for both
r(t_min)	minimum time (if st1() specified)
r(t_max)	maximum time (if st1() specified)
r(z_fixed)	critical value for an equivalent fixed study design

Macros

r(cmd)	gsdesign
r(cmdline)	command as typed
r(direction)	upper, lower, or two-sided
r(effbnd)	pocock, obfleming, wtsiatis, errpocock, errobfleming, kdemets, or hsdecani
r(effect)	hratio or lnhratio
r(futbnd)	pocock, obfleming, wtsiatis, errpocock, errobfleming, kdemets, or hsdecani
r(method)	logrank
r(survvar)	name of the variable containing survival probabilities (if st1() specified)
r(test)	Freedman or Schoenfeld
r(timevar)	name of the variable containing time points (if st1() specified)

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(bspent_a)`	attained cumulative beta spent per look (when futility bounds are specified)
`r(design)`	sample size and stopping boundaries at interim looks
`r(info_frac)`	specified information fraction
`r(info_frac_a)`	fraction of attained information
`r(info_level)`	specified information level
`r(p_crit)`	*p*-values corresponding to boundary critical values
`r(sampsize)`	sample size at interim looks
`r(simpmat)`	control-group survival probabilities (if `simpson()` is specified)

Methods and formulas

Sample sizes at interim analyses are calculated as the product of the information fraction, the information ratio, and the sample size of a fixed-sample study.

See *Methods and formulas* in [ADAPT] **gsbounds** for the formulas used to calculate the stopping boundaries, information fraction, and information ratio. See *Methods and formulas* in [PSS-2] **power logrank** for the formulas used to calculate the sample size for a fixed study. See *Methods and formulas* in [ADAPT] **gsdesign** for the formulas used to calculate the expected sample size.

References

Freedman, L. S. 1982. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* 1: 121–129. https://doi.org/10.1002/sim.4780010204.

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.

Kudo, M., R. S. Finn, S. Qin, K.-H. Han, K. Ikeda, F. Piscaglia, A. Baron, J.-W. Park, G. Han, J. Jassem, J. F. Blanc, A. Vogel, D. Komov, T. R. J. Evans, C. Lopez, C. Dutcus, M. Guo, K. Saito, S. Kraljevic, T. Tamai, M. Ren, and A.-L. Cheng. 2018. Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: A randomised phase 3 non-inferiority trial. *Lancet* 391: 1163–1173. https://doi.org/10.1016/s0140-6736(18)30207-1.

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.

Peng, Z., W. Fan, B. Zhu, G. Wang, J. Sun, C. Xiao, F. Huang, R. Tang, Y. Cheng, Z. Huang, Y. Liang, H. Fan, L. Qiao, F. Li, W. Zhuang, B. Peng, J. Wang, J. Li, and M. Kuang. 2023. Lenvatinib combined with transarterial chemoembolization as first-line treatment for advanced hepatocellular carcinoma: A phase III, randomized clinical trial (LAUNCH). *Journal of Clinical Oncology* 41: 117–127. https://doi.org/10.1200/JCO.22.00392.

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.

Schoenfeld, D. A. 1981. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 68: 316–319. https://doi.org/10.2307/2335833.

———. 1983. Sample-size formula for the proportional-hazards regression model. *Biometrics* 39: 499–503. https://doi.org/10.2307/2531021.

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

[ADAPT] **gsbounds** — Boundaries for group sequential trials

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

[ADAPT] **Glossary**

[PSS-2] **power logrank** — Power analysis for the log-rank test

[ST] **sts test** — Test equality of survivor functions

gsdesign usermethod — Add your own methods to the gsdesign command

Description Syntax Options Remarks and examples Stored results
References Also see

Description

The `gsdesign` *usermethod* command allows you to add your own methods to create a group sequential design (GSD) and produce tables and graphs of the stopping boundaries.

Syntax

`gsdesign` *usermethod* `...` [, *designopts boundopts*]

where *usermethod* is the name of the method you would like to add to the `gsdesign` command, *designopts* are options controlling the sample-size calculation, and *boundopts* are options controlling the calculation of the stopping boundaries.

When naming your `gsdesign` methods, you should follow the same convention as for naming the programs you add to Stata—do not pick "nice" names that may later be used by Stata's built-in methods. The length of *usermethod* may not exceed 16 characters.

designopts	Description
Main	
usermethodopts	method-specific options for user-defined method
* `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)`
`onesided`	request a one-sided test; default is two-sided
* `nfractional`	report fractional sample size

*User-written sample-size evaluators must allow options `alpha()`, `power()`, and `nfractional`.

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

boundopts	Description
Bounds	
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(#[, equal])	total number of analyses (nlooks() − 1 interim analyses and one final analysis); use equal to enforce equal information increments; if neither nlooks() nor information() is specified, the default is nlooks(2)
information(*numlist*)	sequence of information levels for analyses; default is evenly spaced
nopvalues	suppress p-values
Graph	
graphbounds[(*graphopts*)]	graph boundaries
matlistopts(*general_options*)	control the display of boundaries and sample size; seldom used
optimopts	optimization options for boundary calculations; seldom used

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

graphopts	Description
`xdimsampsize`	label the x axis with the sample size collected (default)
`xdiminformation`	label the x axis with the information fraction; 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

alpha(#) sets the overall significance level, which is the familywise type I error rate for all analyses (interim and final). alpha() must be in $(0, 0.5)$. The default is alpha(0.05).

power(#) sets the overall power for all analyses. power() must be in $(0.5, 1)$. 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. beta() must be in $(0, 0.5)$. 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.

onesided requests a study design for a one-sided test. The direction of the test is inferred from the effect size.

nfractional specifies that fractional sample sizes be reported.

nratio(#) specifies the sample-size ratio of the experimental group relative to the control group, $N2/N1$. The default is nratio(1), meaning equal allocation between the two groups.

Bounds

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(# [, equal]) 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).

equal indicates that equal information increments be enforced, which is to say that the same number of new observations will be collected at each look. The default behavior is to start by dividing information evenly among looks, then proceed by rounding up to a whole number of observations at each look. This can cause slight differences in the information collected at each look.

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.

Graph

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

graphopts are the following:

xdimsampsize labels the x axis with the sample size collected (the default).

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

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***).

matlistopts(*general_options*) affects the display of the matrix of boundaries and sample sizes. *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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**.

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* in [ADAPT] **gsbounds**. 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* in [ADAPT] **gsbounds**.

`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* in [ADAPT] **gsbounds**.

`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 $\mathbf{g}\mathbf{H}^{-1}\mathbf{g}' <$ 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:

Steps for adding a new method to the gsdesign command
A quick example
Convention for naming options and storing results
Example: A log-rank test for substantial superiority
    Graphing boundaries
Initializer and parser
    Using an initializer and parser
    Initializer's s() return settings

This entry describes the use of the `gsdesign` command with a user-defined sample-size evaluator. See [ADAPT] **GSD intro** for a general introduction to GSDs for clinical trials; see [ADAPT] **gsbounds** for information about group sequential bounds; and see [ADAPT] **gsdesign** for information about designing group sequential clinical trials with the `gsdesign` command. Also see [PSS-2] **Intro (power)** for a general introduction to power and sample-size analysis, and see [PSS-2] **power usermethod** for additional details about how to write your own sample-size evaluator.

Steps for adding a new method to the gsdesign command

gsdesign works by combining stopping boundaries calculated by gsbounds with the fixed-design sample size calculated by power. If the sample-size calculation you want does not exist as a built-in power method, you can write your own sample-size evaluator and use it with gsdesign.

Adding your own methods to gsdesign is easy. Suppose you want to add your own method, *usermethod*, to gsdesign:

1. Create the evaluator, an r-class program called power_cmd_*usermethod* that computes the sample size that would be required for a fixed study design. Save the program as ado-file power_cmd_*usermethod*.ado.

   A. Be sure your program accepts the nfractional option. This is necessary because gsdesign uses fractional sample sizes when calculating the sample size required at each look.

   B. 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 local macro r(endpoint) as "survival".

   C. If your method allows one-sided tests, store local macro r(direction) as "upper" for an upper one-sided test and as "lower" for a lower one-sided test.

2. Optionally, create an initializer or a parser, s-class programs called, respectively, power_cmd_*usermethod*_init (defined by ado-file power_cmd_*usermethod*_init.ado) and power_cmd_*usermethod*_parse (defined by ado-file power_cmd_*usermethod*_parse.ado). This step is not necessary but can be used to customize the titles and parameters displayed by gsdesign. See *Initializer and parser* for more details.

3. Place all of your programs where Stata can find them.

You are done. You can now use gsdesign *usermethod* like any other gsdesign method.

All user-defined methods for gsdesign are, by construct, also user-defined methods for the power command. This means that your evaluator can be used to calculate the sample size for a fixed study design by running command power *usermethod*. This ability can be exploited, as we do in our second example. However, it bears mentioning that the power command allows user-defined evaluators to calculate power and effect size in addition to sample size, but gsdesign only supports sample-size calculations.

A quick example

Before we discuss the technical details in the following sections, let's try an example to show how easy this all is. We will 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 and save it as power_cmd_myztest.ado.

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

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


        /* calculate sample size for a fixed study */
        tempname N
        scalar `N' = ((invnormal(`power') + invnormal(1 - `alpha' / 2)) / `stddiff')^2
        if ("`nfractional'" == "") {
                scalar `N' = ceil(`N')
        }


        /* return stored results */
        return scalar N       = `N'
        return scalar alpha   = `alpha'
        return scalar power   = `power'
        return scalar stddiff = `stddiff'
end
```

Our program consists of three sections: the syntax command for parsing options, the sample-size computation, and returning the stored results. The three sections work as follows:

Parse: The power_cmd_myztest program accepts three of gsdesign's *designopts*: alpha() for significance level, power() for power, and nfractional to compute fractional sample size. It also has its own option, stddiff(), to specify a standardized difference.

Compute: After parsing options, sample size is computed and stored in temporary scalar `N'.

Return: Finally, the resulting sample size and other results are returned as scalars. Following power's convention for naming commonly returned results, the computed sample size is stored in r(N), the significance level in r(alpha), and the power in r(power). The program additionally stores the standardized difference in r(stddiff).

We save our program as power_cmd_myztest.ado and place the program where Stata can find it. Now we can use myztest within gsdesign as we would any other existing method of gsdesign.

To design a group sequential trial using myztest with a standardized difference of 0.7 and default specifications of O'Brien–Fleming efficacy bounds with two evenly spaced looks, power of 0.8, and two-sided significance level of 0.05, we run

```
. 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

------+----------------------------------------+------------
      | Info.             Efficacy             | Sample size
 Look | frac.       Lower     Upper    p-value |           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.
```

We can use any type of boundary allowed by gsdesign, and we can even display the bounds on a graph. For a four-look design with Wang–Tsiatis efficacy bounds with efficacy parameter $\Delta_e = 0.25$ and O'Brien–Fleming nonbinding futility bounds, we run

```
. gsdesign myztest, stddiff(0.7) efficacy(wtsiatis(0.25))  futility(obfleming)
> nlooks(4) graphbounds

Group sequential design for myztest
Two-sided test

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

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

Expected sample size:
         H0 =  12.56
         Ha =  14.09

Info. ratio = 1.2141
    N fixed =     17
      N max =     20

Fixed-study crit. values = ±1.9600
```

```
Critical values, p-values, and sample sizes for a group sequential design
─────────────────────────────────────────────────────────────────────────
     │ Info.            Efficacy                      Futility
Look │ frac.      Lower     Upper  p-value     Lower     Upper  p-value
─────┼───────────────────────────────────────────────────────────────────
   1 │  0.25    -2.9887    2.9887   0.0028         .         .        .
   2 │  0.50    -2.5132    2.5132   0.0120   -0.8059    0.8059   0.4203
   3 │  0.75    -2.2709    2.2709   0.0232   -1.5492    1.5492   0.1213
   4 │  1.00    -2.1133    2.1133   0.0346   -2.1133    2.1133   0.0346
─────────────────────────────────────────────────────────────────────────
Note: Critical values are for z statistics; otherwise, use p-value
      boundaries.

──────────────────
     │ Sample size
Look │           N
─────┼────────────
   1 │           5
   2 │          10
   3 │          15
   4 │          20
──────────────────
```

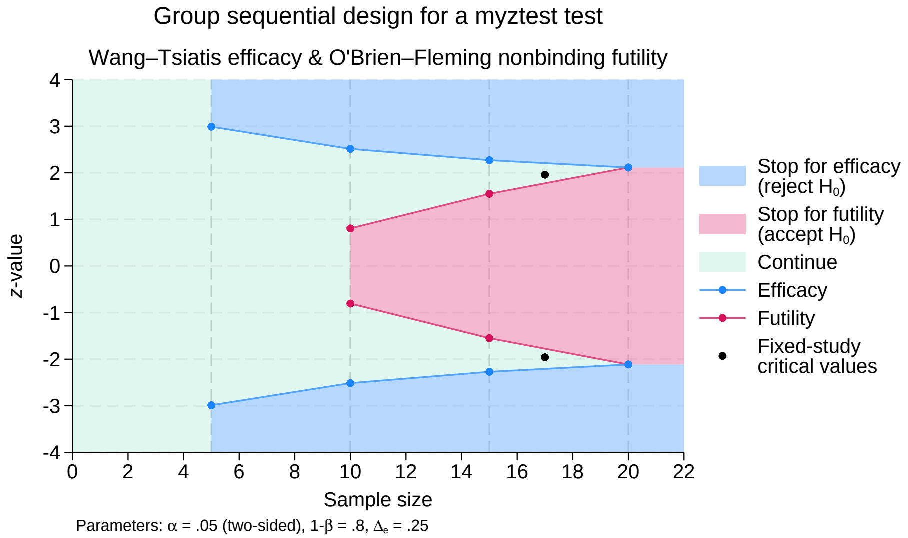


Figure 1. User-written one-sample z test with efficacy and futility bounds

The above is just a simple example. Your program can be as complicated as you would like; you can even use simulations to compute your results. You can also customize your output and graphs with an initializer or parser.

Convention for naming options and storing results

You can specify any method-specific options you want, but for the gsdesign command to automatically recognize its common design options, you must ensure that you follow gsdesign's naming convention for *designopts* in your program. For example, gsdesign specifies the significance level in the `alpha()` option with minimum abbreviation of `a()`. You need to ensure that you use the same option name with the same abbreviation in your evaluator to specify the significance level. The same applies to all the *designopts* described in the Syntax section.

To be compatible with `gsdesign`, you must ensure that your sample-size evaluator stores the total sample size in scalar `r(N)`. For two-sample methods, you must additionally store the control-group sample size in scalar `r(N1)` and the experimental-group sample size in scalar `r(N2)`.

For time-to-event methods, your evaluator must store local macro `r(endpoint)` as "`survival`" and store the number of events in scalar `r(E)`. If your method allows for censoring, store the survival probability of the control group in scalar `r(s1)` and the survival probability of the experimental group in scalar `r(s2)`, store the overall probability of experiencing a failure event in scalar `r(Pr_E)`, and store the probability of withdrawal in scalar `r(Pr_w)`.

If your method allows one-sided tests, it should store local macro `r(direction)` as "`upper`" when an upper one-sided test is specified and as "`lower`" when a lower one-sided test is specified.

If you want to display additional parameters in the `gsdesign` output, you must store them as scalars and let `gsdesign` know to display them through the use of an initializer or parser. However, the full functionality of the `gsdesign` command is available without the use of an initializer or parser.

Example: A log-rank test for substantial superiority

Many aspects of the COVID-19 pandemic were unprecedented, including the speed with which vaccines were developed. Unlike the yearslong development process of previous vaccines, the first COVID-19 vaccines began phase 3 clinical trials for efficacy less than half a year after COVID-19 was declared a global pandemic. One of these vaccines was produced by the company Sinovac, and Palacios et al. (2020) describe the PROFISCOV phase 3 clinical trial of the Sinovac COVID-19 vaccine among healthcare workers in Brazil.

The primary endpoint, or outcome of interest, was the incidence of symptomatic COVID-19. Rather than merely recording whether study participants caught COVID-19, the researchers monitored how long it took each participant to catch COVID-19, making this a survival study. Participants who had not experienced symptomatic COVID-19 by the end of the study's one-year follow-up period were considered to have been censored. This type of censoring is known as administrative censoring.

The PROFISCOV study measured vaccine efficacy as $1 - \mathrm{HR}$, where HR is the hazard ratio of the experimental to the control participants (Palacios et al. 2020, Study Protocol). The alternative hypothesis of the PROFISCOV study was a vaccine efficacy of 60%, corresponding to a hazard ratio of 0.4. However, the null hypothesis was not a vaccine efficacy of 0% (which would correspond to a hazard ratio of 1 and indicate no treatment effect); instead, the null hypothesis was a vaccine efficacy of 30% (corresponding to a hazard ratio of 0.7). To declare the Sinovac COVID-19 vaccine effective, the planners of the PROFISCOV study required it to beat the control by more than 30%. This type of study is known as a superiority trial or, more specifically, a substantial superiority trial with a superiority margin of 30%.

`gsdesign` does not have a built-in method for calculating sample size for a substantial superiority test of two survivor functions, so we will write our own. We assume a log-rank test will be used to compare the two survivor functions, so we model our command after `power logrank`. We write a sample-size evaluator based on the *Methods and formulas* described in [PSS-2] **power logrank**, but we follow the example of Julious (2010, 264) to modify the formulas to accommodate a superiority margin, provided in the form of a hazard ratio under the null hypothesis.

We will call our new method superlogrank. It will compute the number of events and sample size for a fixed-design substantial superiority trial using a log-rank test to compare two survivor functions. Sample-size evaluator power_cmd_superlogrank accepts the standard gsdesign *designopts* of alpha(), power(), nfractional, and onesided, but we decide to make onesided a required option because we are only interested in testing a one-sided alternative hypothesis: that the Sinovac COVID-19 vaccine is substantially better than the placebo.

Like power logrank, our command power_cmd_superlogrank performs sample-size calculations for a test of the hazard ratio using the Freedman method (the default) or for a test of the log hazard-ratio using the Schoenfeld method (with option schoenfeld). Most of the remaining syntax for power_cmd_superlogrank is akin to a simplified version of the power logrank syntax: the survival probability in the control group is provided as an optional argument to the command (specified before the comma), the hazard ratio under the alternative hypothesis is specified with option hratio(), the probability of withdrawal is specified with option wdprob(), and the ratio of experimental-group sample size to control-group sample size is specified with option nratio().

Command power_cmd_superlogrank accepts the additional option hr0(), the hazard ratio under the null hypothesis. We set the default to be hr0(1), which corresponds to a superiority margin of 0 (that is, the vaccine efficacy under the null hypothesis is 0%). If hr0() is left at its default, the substantial superiority test reduces to a standard log-rank test.

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

        /* parse syntax and check for valid options */
        syntax [anything(name=s1)]  /// P(survival) of control group (optional)
               , ONESIDed           /// one-sided test (required option)
                 [ HRatio(real 0.5) /// hazard ratio under Ha
                   hr0(real 1)      /// hazard ratio under H0
                   WDProb(real 0)   /// P(nonadministrative censoring)
                   NRATio(real 1)   /// ratio of experimental/controls
                   SCHoenfeld       /// use Schoenfeld calculation
                   Alpha(real 0.05) /// significance level
                   Power(real 0.8)  /// power
                   NFRACtional      /// report fractional sample size
                 ]


        /* assume 0% survival if s1 is not specified */
        if (`"`s1'"' != "") {
                confirm number `s1'
                assert (`s1' >= 0) & (`s1' < 1)
        }
        else {
                local s1 = 0
        }
        assert (`hratio'  > 0)
        assert (`hr0'     > 0)
        assert (`nratio'  > 0)
        assert (`wdprob' >= 0) & (`wdprob' < 1)
        assert (`alpha'   > 0) & (`alpha'  < 1)
        assert (`power'   > 0) & (`power'  < 1)
        /* calculate number of failures (events) & fixed-study sample size */
        tempname zalpha zbeta Dratio lhs rhs E s2 prE N N1 N2
        scalar ‘zalpha’ = invnormal(1 - ‘alpha’)
        scalar ‘zbeta’  = invnormal(‘power’)
        scalar ‘Dratio’ = ‘hratio’ / ‘hr0’
        scalar ‘lhs’    = (‘zalpha’ + ‘zbeta’)^2 / ‘nratio’
        if ("‘schoenfeld’" != "") {
                /* Schoenfeld calculation */
                scalar ‘rhs’ =  ((‘nratio’ + 1) / log(‘Dratio’))^2
        }
        else {
                /* Freedman calculation */
                scalar ‘rhs’ = ((‘nratio’ * ‘Dratio’ + 1) / (‘Dratio’ - 1))^2
        }
        scalar ‘E’   = ‘lhs’ * ‘rhs’
        scalar ‘s2’  = ‘s1’^‘hratio’
        scalar ‘prE’ = 1 - (‘s1’ + ‘nratio’ * ‘s2’) / (‘nratio’ + 1)
        scalar ‘N’   = ‘E’ / (‘prE’ * (1 - ‘wdprob’))
        scalar ‘N1’  = ‘N’ * ‘nratio’ / (‘nratio’ + 1)
        scalar ‘N2’  = ‘N’ / (‘nratio’ + 1)


        if ("‘nfractional’" == "") {
                /* round up to a whole number */
                scalar ‘E’  = ceil(‘E’)
                scalar ‘N1’ = ceil(‘N1’)
                scalar ‘N2’ = ceil(‘N2’)
                scalar ‘N’  = ‘N1’ + ‘N2’
        }


        /* return stored results */
        return scalar E           = ‘E’
        return scalar N           = ‘N’
        return scalar N1          = ‘N1’
        return scalar N2          = ‘N2’
        return scalar hratio      = ‘hratio’
        return scalar hr0         = ‘hr0’
        return scalar nratio      = ‘nratio’
        return scalar s1          = ‘s1’
        return scalar s2          = ‘s2’
        return scalar Pr_E        = ‘prE’
        return scalar Pr_w        = ‘wdprob’
        return scalar alpha       = ‘alpha’
        return scalar power       = ‘power’
        return scalar nfractional = ("‘nfractional’" != "")
        return local  direction   = cond(‘Dratio’ > 1, "upper", "lower")
        return local  endpoint    = "survival"
end
```

While this program is considerably more complicated than our previous program, `power_cmd_myztest`, it contains the same three basic parts: it starts by parsing the syntax, then it calculates the sample size, and finally it returns the stored results. The three sections work as follows:

Parse: The `power_cmd_superlogrank` program accepts four common `gsdesign` *designopts* (`onesided`, `alpha()`, `power()`, and `nfractional`), as well as several of its own options. To match the syntax of `power logrank`, program `power_cmd_superlogrank` reads the survival probability of the control group as an argument (before the comma) rather than as an option. The syntax is parsed with the `syntax` command and checked for validity.

Compute: The required number of events (failures) is calculated and stored in temporary scalar ‘E’, and the control-group sample size, experimental-group sample size, and total sample size are calculated and stored in temporary scalars ‘N1’, ‘N2’, and ‘N’, respectively. Additional temporary scalars hold the probability of survival in the experimental group (‘s2’) and the overall probability of failure (‘prE’).

Return: The design parameters specified to power_cmd_superlogrank are returned as scalars, as are indicators that a one-sided test was conducted and that the fractional sample size was calculated. The overall sample size is returned as r(N). By returning the control- and experimental-group sample sizes as r(N1) and r(N2), power_cmd_superlogrank tells gsdesign that method superlogrank performs a two-sample test.

Because local macro r(endpoint) is returned as "survival", gsdesign will recognize superlogrank as a survival method and know to look for returned results r(E), r(Pr_E), r(s1), r(s2), and r(Pr_w). Additionally, gsdesign will know the direction of the one-sided test because power_cmd_superlogrank stores local macro r(direction) as either "upper" or "lower".

Any user-defined method for gsdesign is, by design, also a user-defined method for the power command. This enables us to perform a simple sanity check of our new program: if superlogrank is used as a power method and option hr0() is left at its default value of 1, it should yield the same sample size as power logrank with the same options. For this sanity check, we arbitrarily choose a control-group survival probability of 83%, hazard ratio of 0.8, withdrawal probability of 12%, significance level of 2.5% for a one-sided test using the Schoenfeld method, and power of 90%, and we allocate 1.5 times as many participants to the experimental arm as to the control arm. We verify:

```
. power logrank 0.83, hratio(0.8) wdprob(0.12) nratio(1.5) schoenfeld
> onesided alpha(0.025) power(0.9)
Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Schoenfeld method
H0: ln(HR) = 0  versus  Ha: ln(HR) < 0
Study parameters:
        alpha =    0.0250
        power =    0.9000
        delta =   -0.2231  (log hazard-ratio)
       hratio =    0.8000
        N2/N1 =    1.5000
Censoring and withdrawal:
           s1 =    0.8300
           s2 =    0.8615
         Pr_E =    0.1511
         Pr_w =    0.1200
Estimated number of events and sample sizes:
            E =       880
            N =     6,614
           N1 =     2,646
           N2 =     3,968
        N2/N1 =    1.4996
. power superlogrank 0.83, hratio(0.8) wdprob(0.12) nratio(1.5) schoenfeld
> onesided alpha(0.025) power(0.9)
Estimated sample sizes
One-sided test

  ┌─────────────────────────┐
  │   alpha    power      N │
  ├─────────────────────────┤
  │    .025       .9  6,614 │
  └─────────────────────────┘
```

The output of power superlogrank is stark compared with the detailed output of power logrank, but the sample-size calculation is identical. The output of power superlogrank can be improved through the addition of an initializer or parser, but the functionality of gsdesign superlogrank does not require an initializer or parser.

Returning to the design of the PROFISCOV trial, Palacios et al. (2020) report that the study was designed to have 90% power with a one-sided significance level of 2.5%, and it used error-spending Hwang–Shih–de Cani efficacy and futility bounds with parameter $\gamma_e = \gamma_f = -4$ and a single interim look once 40% of the total number of events had been observed. We assume that all participants in the clinical trial will be followed until they develop symptomatic COVID-19, so we omit the command argument specifying the control-group survival probability. Using a hazard ratio of 0.7 under the null hypothesis and 0.4 under the alternative hypothesis, we calculate the stopping boundaries and sample sizes using gsdesign superlogrank.

```
. gsdesign superlogrank, hratio(0.4) hr0(0.7) onesided alpha(0.025) power(0.9)
> efficacy(hsdecani(-4)) futility(hsdecani(-4)) information(0.4 1)
Group sequential design for superlogrank
One-sided test
Efficacy: Error-spending Hwang-Shih-de Cani, gamma = -4.0000
Futility: Error-spending Hwang-Shih-de Cani, nonbinding, gamma = -4.0000
Study parameters:
      alpha = 0.0250  (lower one-sided)
      power = 0.9000
Censoring:
         s1 = 0.0000
         s2 = 0.0000
       Pr_E = 1.0000
Expected number of events:
         H0 = 113.41
         Ha = 126.11
Info. ratio = 1.0142
    E fixed =    142
    N fixed =    142
      N max =    144
     N1 max =     72
     N2 max =     72
Fixed-study crit. value = -1.9600
Critical values, p-values, and sample sizes for a group sequential design

      | Info.        Efficacy           Futility      |    Events
 Look | frac.     Lower   p-value    Upper   p-value |         E
------+---------------------------------------------+-----------
    1 |  0.40   -2.9037    0.0018   0.3739    0.6457 |        58
    2 |  1.00   -1.9753    0.0241  -1.9753    0.0241 |       144

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

gsdesign begins by displaying the study parameters and, because it knows that superlogrank is a survival method, details about censoring.

The next section of the output displays the expected number of events, which is the average number of events if the group sequential trial were to be repeated many times. The following section reports 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 number of failures at the final look of the GSD to the number of failures in a fixed study design.

In this case, the maximum sample size is the same as the maximum number of events because we omitted information about censoring, so gsdesign superlogrank assumes that all participants are followed until they contract symptomatic COVID-19.

The table at the end of the output displays the stopping boundaries and 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 for substantial superiority.

Graphing boundaries

It is unrealistic to assume, as we did above, that all participants in the clinical trial will be followed until they develop symptomatic COVID-19. Here we assume that only 1% of participants in the control group develop symptomatic COVID-19 during the follow-up period, and we assume that 10% of all participants withdraw from the study before contracting COVID-19. We leave the rest of the design parameters at their previous values, but we add gsdesign option graphbounds to display the boundaries visually.

```
. gsdesign superlogrank 0.99, hratio(0.4) hr0(0.7) wdprob(0.1) onesided
> alpha(0.025) power(0.9) efficacy(hsdecani(-4))
> futility(hsdecani(-4)) information(0.4 1) graphbounds

Group sequential design for superlogrank
One-sided test

Efficacy: Error-spending Hwang-Shih-de Cani, gamma = -4.0000
Futility: Error-spending Hwang-Shih-de Cani, nonbinding, gamma = -4.0000

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

Censoring and withdrawal:
         s1 = 0.9900
         s2 = 0.9960
       Pr_E = 0.0070
       Pr_w = 0.1000

Expected number of events:
         H0 = 113.41
         Ha = 126.11

Info. ratio = 1.0142
    E fixed =    142
    N fixed = 22,404
      N max = 22,722
     N1 max = 11,361
     N2 max = 11,361

Fixed-study crit. value = -1.9600

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

-------------------------------------------------------------------
      | Info.        Efficacy             Futility     |    Events
 Look | frac.     Lower   p-value     Upper   p-value  |         E
------+------------------------------------------------+-----------
    1 |  0.40   -2.9037    0.0018    0.3739    0.6457  |        58
    2 |  1.00   -1.9753    0.0241   -1.9753    0.0241  |       144
-------------------------------------------------------------------

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

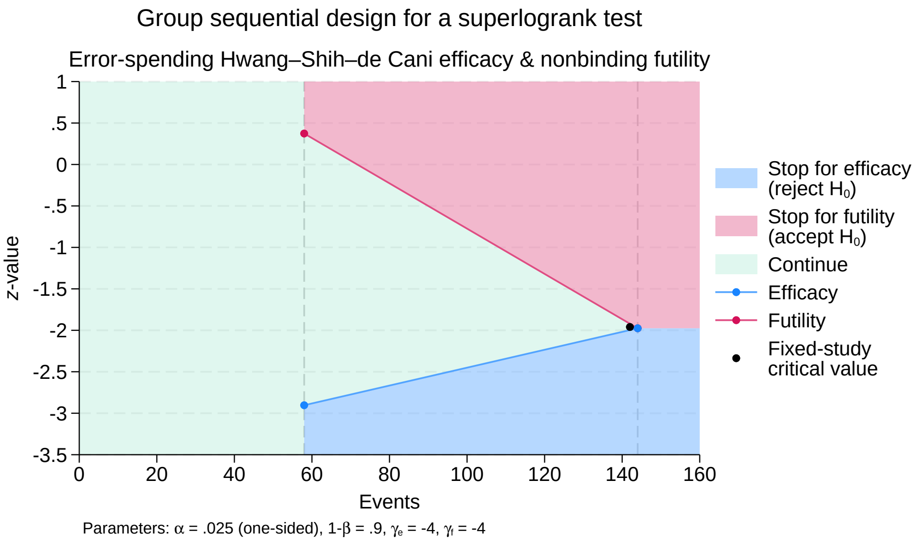


Figure 2. Log-rank test for substantial superiority with efficacy and futility bounds

The required number of events is unchanged from its previous value, but incorporating information about censoring has increased the number of participants we need in order to observe those failures. After taking into account participant withdrawal as well as administrative censoring, we anticipate requiring 22,722 participants to observe 144 failures.

Examining the graph, we see that the entire region from 0 to 58 events is shaded green, the color of the continuation region. This is because the data have not yet been analyzed, so the trial cannot be stopped. The first look will be conducted once 58 participants have contracted symptomatic COVID-19, and a log-rank test for substantial superiority will be performed. If the test statistic, z_1, is ≤ -2.904, 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 $z_1 > 0.374$, it lies in the acceptance region and we can accept H_0, terminating the trial early for futility. Because the futility bound is nonbinding, if we continue the trial despite z_1 crossing the futility bound, the familywise type I error is still controlled. If $-2.904 < z_1 \leq 0.374$, we say that z_1 lies in the continuation region, and the trial must proceed to the second and final look.

At the final look, there is no continuation region; the futility critical value equals the efficacy critical value of -1.975, so H_0 must be rejected or accepted. While accepting the null hypothesis is taboo in many disciplines, it has a long history in the context of sequential trials (see Origins of GSD for a history of GSDs). If test statistic $z_2 \leq -1.975$, we reject H_0; if $z_2 > -1.975$, we accept H_0.

Initializer and parser

The initializer and parser are optional s-class programs named power_cmd_*usermethod*_init and power_cmd_*usermethod*_parse, respectively. Initializers and parsers are more important for user-defined power commands than for user-defined gsdesign commands, but they can still be useful tools to customize the output and graphs produced by gsdesign.

The option to provide both an initializer and a parser is provided as a convenience to the user, but in practice only one is ever needed because the s() returned values can be set by either an initializer or a parser. In fact, it is generally counterproductive to use both an initializer and a parser because the s() returned values are collected by gsdesign (or by power, in the case of power *usermethod*) after first running the parser and then the initializer. This means that if the initializer executes sreturn clear, it will clear any s() returned values set by the parser.

The difference between the initializer and the parser is that the parser is executed with all the arguments and options specified to gsdesign (or to power, in the case of power *usermethod*), while only options are passed to the initializer, not arguments. This is done to enable the parser to parse the full command specification (instead of the evaluator program), should you so desire. A side effect is that a parser can be more useful than an initializer if your user-defined method accepts arguments as well as options.

Using an initializer and parser

Using our user-defined method superlogrank as an example, we define a parser, power_cmd_superlogrank_parse, to set s() results and customize the output and graph produced by gsdesign superlogrank. We choose a parser over an initializer because program power_cmd_superlogrank accepts the control-group survival probability as an argument (before the comma), not an option, so it will only be passed to a parser, not an initializer. We write our parser and save it as power_cmd_superlogrank_parse.ado.

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

        /* parse relevant syntax */
        syntax [anything(name=s1)] ///
               , [WDProb(string)   ///
                  NRATio(string)   ///
                  SCHoenfeld       ///
                  *                /// asterisk (*) captures all other options
                 ]


        /* identify parameters to display */
        local diparam hratio hr0
        local grparam HR{sub:a} HR{sub:0}
        if (‘"‘nratio’"’ != "") {
                local diparam ‘diparam’ nratio
                local grparam ‘grparam’ N{sub:2}/N{sub:1}
        }
        if (‘"‘s1’"’ != "") {
                local diparam ‘diparam’ s1 s2 Pr_E
                local grparam ‘grparam’ S{sub:1}(T) S{sub:2}(T) p{sub:E}
        }
        if (‘"‘wdprob’"’ != "") {
                local diparam ‘diparam’ Pr_w
                local grparam ‘grparam’ p{sub:w}
        }
        /* return stored results */
        sreturn clear
        local suptest  = "Log-rank test for substantial superiority"
        local testtype = cond("‘schoenfeld’" == "", "Freedman", "Schoenfeld")
        sreturn local pss_subtitle = "‘suptest’, ‘testtype’ method"
        sreturn local pss_title "for two-sample comparison of survivor functions"
        sreturn local pss_colnames ‘diparam’
        sreturn local pss_colgrsymbols ‘grparam’
end
```

We rerun the same gsdesign superlogrank command specification as before, but this time the parser sets s-class returned values to customize the output and graph.

```
. gsdesign superlogrank 0.99, hratio(0.4) hr0(0.7) wdprob(0.1) onesided
> alpha(0.025) power(0.9) efficacy(hsdecani(-4))
> futility(hsdecani(-4)) information(0.4 1) graphbounds
Group sequential design for two-sample comparison of survivor functions
Log-rank test for substantial superiority, Freedman method

Efficacy: Error-spending Hwang-Shih-de Cani, gamma = -4.0000
Futility: Error-spending Hwang-Shih-de Cani, nonbinding, gamma = -4.0000

Study parameters:
      alpha = 0.0250  (lower one-sided)
      power = 0.9000
     hratio = 0.4000
        hr0 = 0.7000

Censoring and withdrawal:
         s1 = 0.9900
         s2 = 0.9960
       Pr_E = 0.0070
       Pr_w = 0.1000

Expected number of events:
         H0 = 113.41
         Ha = 126.11

Info. ratio = 1.0142
    E fixed =    142
    N fixed = 22,404
      N max = 22,722
     N1 max = 11,361
     N2 max = 11,361

Fixed-study crit. value = -1.9600

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

------+-------------------------------------------------+------------
      |  Info.        Efficacy             Futility      |    Events
 Look |  frac.     Lower   p-value      Upper   p-value |           E
------+-------------------------------------------------+------------
    1 |   0.40    -2.9037    0.0018     0.3739    0.6457 |          58
    2 |   1.00    -1.9753    0.0241    -1.9753    0.0241 |         144
------+-------------------------------------------------+------------

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

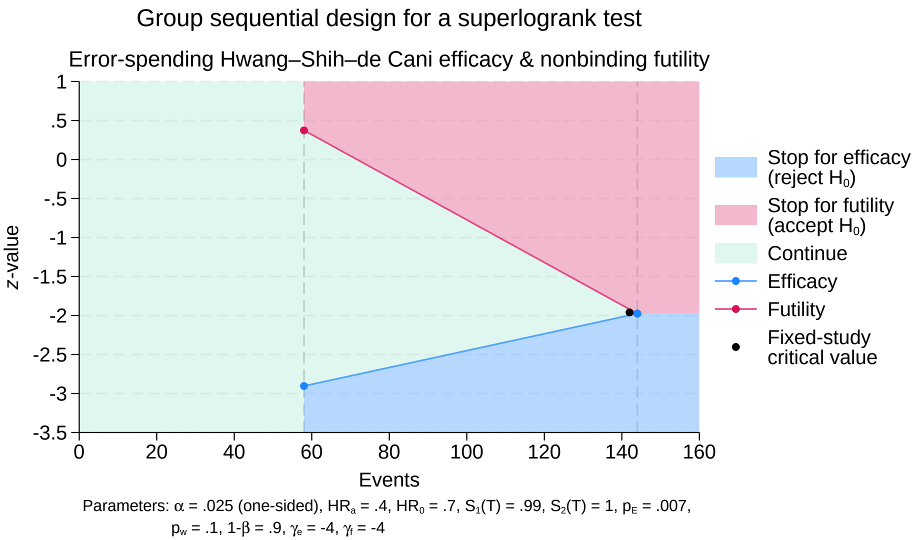


Figure 3. Customized graph of log-rank test for substantial superiority

With the addition of the parser, `gsdesign superlogrank` displays the values of study parameters `hratio` and `hr0`. Also, our additional parameters and their custom symbols now appear in the "Parameters:" note on the graph.

Initializer's s() return settings

The following `s()` results may be set by the initializer or parser. See [PSS-2] **power usermethod** for more details.

Macros

`s(pss_samples)`	`onesample` for a one-sample test or `twosample` for a two-sample test
`s(pss_colnames)`	columns to be added to the default supported columns
`s(pss_allcolnames)`	all supported columns
`s(pss_tabcolnames)`	columns to be added to the default table
`s(pss_alltabcolnames)`	all columns to be displayed in the default table
`s(pss_collabels)`	labels for the specified columns
`s(pss_colformats)`	formats for the specified columns
`s(pss_colwidths)`	widths for the specified columns
`s(pss_colgrlabels)`	labels to be used to label columns on the graph
`s(pss_colgrsymbols)`	symbols to be used to label columns on the graph
`s(pss_delta)`	column name containing the effect-size parameter
`s(pss_target)`	column name containing the target parameter
`s(pss_targetlabel)`	label for the target parameter
`s(pss_argnames)`	column names containing command arguments
`s(pss_title)`	method-specific title
`s(pss_subtitle)`	subtitle
`s(pss_hyp_lhs)`	left-hand-side parameter or value for the hypothesis
`s(pss_hyp_rhs)`	right-hand-side parameter or value for the hypothesis
`s(pss_grhyp_lhs)`	left-hand-side parameter or value for the hypothesis on the graph
`s(pss_grhyp_rhs)`	right-hand-side parameter or value for the hypothesis on the graph

Stored results

Stored results include those stored by the user-defined method and the standard results from `gsdesign`; see *Stored results* in [ADAPT] **gsdesign**.

References

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.

Julious, S. A. 2010. *Sample Sizes for Clinical Trials*. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781584887409.

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.

Palacios, R., E. G. Patiño, R. de Oliveira Piorelli, M. T. R. P. Conde, A. P. Batista, G. Zeng, Q. Xin, E. G. Kallas, J. Flores, C. F. Ockenhouse, and C. Gast. 2020. Double-blind, randomized, placebo-controlled phase III clinical trial to evaluate the efficacy and safety of treating healthcare professionals with the adsorbed COVID-19 (inactivated) vaccine manufactured by Sinovac—PROFISCOV: A structured summary of a study protocol for a randomised controlled trial. *Trials* 21: 853. https://doi.org/10.1186/s13063-020-04775-4.

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.

Also see

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

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

[ADAPT] **gsbounds** — Boundaries for group sequential trials

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

[ADAPT] **Glossary**

[PSS-2] ***power usermethod*** — Add your own methods to the power command

Glossary

2 × 2 contingency table. A 2 × 2 contingency table is used to describe the association between a binary independent variable and a binary response variable. See [ADAPT] **gsdesign twoproportions**.

acceptance region. In classical hypothesis testing, an acceptance region is the complement of the rejection region and is defined as a set of values of a test statistic for which the null hypothesis cannot be rejected. Group sequential designs further differentiate between the acceptance region, where the null hypothesis is accepted and the trial is terminated early for futility, and the continuation region, where the trial is continued due to insufficient evidence to accept or reject the null hypothesis. Also see *rejection region* and *continuation region*.

accrual period or **recruitment period** or **accrual**. The accrual period (or recruitment period) is the period during which participants are being enrolled (recruited) into a study. Also see *follow-up period*.

active control or **active comparator**. In a clinical trial of an experimental treatment for a condition where there is an existing standard of care, there is often an ethical argument against giving study participants a placebo, so the control group is given the standard of care and the experimental treatment is compared with the active control. Also see *placebo control*.

adaptive design. As defined by the US Food and Drug Administration (2019), an adaptive design is a "clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial."

administrative censoring. Administrative censoring is the right-censoring that occurs when the study observation period ends. All participants complete the course of the study and are known to have experienced one of two outcomes at the end of the study: survival or failure. This type of censoring should not be confused with withdrawal or loss to follow-up. Also see *censored, uncensored, left-censored, and right-censored*.

adverse event. Adverse events are harmful side effects of a treatment and negative medical outcomes not associated with an underlying disease. Clinical trials must closely track the incidence and severity of adverse events to ensure that a treatment is safe as well as effective.

allocation ratio. The allocation ratio, n_2/n_1, is the number of study participants in the experimental (treatment) group divided by the number of participants in the control (reference) group.

alpha. Alpha, α, denotes the significance level. Also see *familywise significance level*.

alternative hypothesis. In hypothesis testing, the alternative hypothesis represents the counterpoint to which the null hypothesis is compared. When the parameter being tested is a scalar, the alternative hypothesis can be either one-sided or two-sided. Also see *null hypothesis*.

arm. In the context of a clinical trial, groups of study participants given the same treatment are often called arms. In a classic two-arm randomized controlled trial, the experimental arm is given the experimental treatment and the control arm is given the control treatment. Also see *single-arm trial* and *two-arm trial*.

attained power. When calculating the required sample size for a specified significance level and power, the resulting sample size is often fractional and must be rounded up to a whole number. This causes the attained power to be slightly greater than the requested power. Also see *power*.

attained sample-size ratio. When specifying a sample-size ratio that results in noninteger sample sizes, gsdesign will round up the computed sample sizes to the nearest integers. The attained sample-size ratio is computed using the rounded sample sizes. Also see *sample-size ratio*.

balanced design. A balanced design represents an experiment in which the numbers of treated and untreated study participants are equal. For many types of two-sample hypothesis tests, the power of the test is maximized with balanced designs. Balanced designs may also be called equal-allocation designs.

Bernoulli trial. A Bernoulli trial is an experiment with only two possible outcomes, "success" or "failure", recorded as 0 and 1, respectively. In a clinical trial with a binary outcome, each participant's response is viewed as an independent Bernoulli trial with a fixed probability of success. See [ADAPT] **gsdesign oneproportion** and [ADAPT] **gsdesign twoproportions**.

beta. Beta, β, denotes the probability of committing a type II error, namely, failing to reject the null hypothesis even though it is false. Also see *type II error*.

binding futility boundaries or **binding futility bounds**. In a group sequential clinical trial with binding futility bounds, if the test statistic at an interim analysis crosses the futility boundary, the trial must be stopped for futility; otherwise, it risks overrunning the specified significance level. Group sequential designs with binding futility bounds require smaller efficacy critical values than equivalent group sequential designs with nonbinding futility boundaries. Also see *nonbinding futility boundaries*.

binary outcome. When the response of each participant in a clinical trial is either "success" or "failure", we say the trial has a binary outcome. Analysis of binary clinical trial data treats each response as a Bernoulli trial with a fixed probability of success. A test of proportions, such as a binomial test or Pearson's χ^2 test, is conducted to determine if the data are compatible with the null hypothesis. See [ADAPT] **gsdesign oneproportion** and [ADAPT] **gsdesign twoproportions**. Also see *composite endpoint*.

binomial test. A binomial test is a test for which the exact sampling distribution of the test statistic is binomial. See [R] **bitest**. Also see [ADAPT] **gsdesign oneproportion**.

biomarker. A biomarker is a characteristic of the body that can be measured objectively and that serves as an indicator of healthy biological processes, disease status, or response to a therapeutic treatment.

Response biomarkers are frequently used as surrogate endpoints for clinical trials where the clinical outcome of interest is too difficult, time consuming, or expensive to measure. For example, instead of relying on autopsy to diagnose Alzheimer's disease, we can use medical imaging to measure brain glucose metabolism as a response biomarker.

Biomarkers can also serve as risk factors used to define a population of interest. For example, the APOE $\epsilon 4$ gene is a known risk factor for Alzheimer's disease, and it can be used as a biomarker to define the target population of a clinical trial.

bisection method. This method finds a root x of a function $f(x)$ such that $f(x) = 0$ by repeatedly subdividing an interval on which $f(x)$ is defined until the change in successive root estimates is within the requested tolerance and function $f(\cdot)$ evaluated at the current estimate is sufficiently close to 0.

blinding. Blinding refers to clinical trials where the identity of the treatment is hidden. In an open-label trial, participants are told which treatment they are receiving. In a single-blinded trial, participants do not know which treatment they receive, but the researchers administering the treatments and the data analysts are unblinded, meaning they know which treatment each participant receives. If the study design of a blinded trial calls for the experimental treatment to be compared with no intervention, then the control group is given a placebo so that they do not know they are members of the control group. In a double-blinded trial, both the participants and the researchers administering the treatments are blinded to the identity of the treatments, and in a triple-blinded trial, even the data analysts are blinded.

boundary or **bound**. See *stopping boundary*.

boundary-calculation procedure or **boundary-calculation method**. In the context of a group sequential design, the boundary-calculation procedure refers to the method used to create a stopping boundary. Boundary-calculation procedures fall into two broad categories: classical stopping bounds and error-spending bounds. Classical stopping bounds calculate boundary critical values directly, while error-spending bounds define an error-spending function that partitions type I or type II error between the planned looks. Also see *classical Wang–Tsiatis bounds*, *classical Pocock bounds*, *classical O'Brien–Fleming bounds*, *error-spending Pocock bounds*, *error-spending O'Brien–Fleming bounds*, *error-spending Kim–DeMets bounds*, and *error-spending Hwang–Shih–de Cani bounds*.

censored, **uncensored**, **left-censored**, and **right-censored**. An observation is censored when the exact time of failure is not known, and it is uncensored when the exact time of failure is known.

An observation is left-censored when the exact time of failure is not known; it is merely known that the failure occurred before t_l. Suppose that the event of interest is becoming employed. If a subject is already employed when first interviewed, his outcome is left-censored.

An observation is right-censored when the time of failure is not known; it is merely known that the failure occurred after t_r. If a patient survives until the end of a study, the patient's time of death is right-censored.

Also see *administrative censoring*.

clinical trial. A clinical trial is an experiment testing the effect of a treatment or procedure on human participants.

clinically meaningful difference or **clinically meaningful effect** or **clinically significant difference**. Clinically meaningful difference represents the magnitude of an effect of interest that is of clinical importance. What is meant by "clinically meaningful" may vary from study to study.

clinical outcome. The clinical outcome is an outcome that confers direct clinical benefit, such as overall survival. In practice, clinical outcomes are often expensive or time consuming to measure, so surrogate endpoints are frequently measured instead. Also see *endpoint* and *target parameter*.

composite endpoint. Sometimes, when designing a clinical trial, there are multiple endpoints of interest. One solution is to combine multiple endpoints into a single composite endpoint. For example, a clinical trial of a treatment for COVID-19 might use a composite endpoint such as "death or intubation", where each participant's response is an indicator of whether they died or were intubated. Also see *binary outcome*.

continuation region. In group sequential designs, a continuation region is defined as a set of values of a test statistic that provide insufficient evidence to accept or reject the null hypothesis. If the test statistic from an interim analysis of clinical trial data lies within the continuation region, the trial will continue as planned (as opposed to stopping early if the test statistic lies within the acceptance region or the rejection region). There is no continuation region at the final analysis, because at this stage, the null hypothesis must be either accepted or rejected. Also see *acceptance region* and *rejection region*.

control arm. See *control group*.

control group. A control group (or arm) comprises study participants who are randomly assigned to a group where they receive the control treatment, which is either no treatment or a standard treatment. In hypothesis testing, this is usually the reference group. Also see *experimental group*.

control treatment. In a clinical trial, the control treatment is the reference treatment against which an experimental treatment is judged. If there are no existing treatments that are comparable with the experimental treatment, then the control group will typically receive a placebo. When a standard of care exists, there is often an ethical argument against using a placebo; in this case, an active control is used, in which control-group participants receive the existing standard of care. Also see *experimental treatment*.

critical value. In classical hypothesis testing, a critical value is a boundary of the rejection region. In the context of a group sequential design, there are two types of critical values: efficacy critical values, which are boundaries of the rejection region, and futility critical values, which are boundaries of the acceptance region. Also see *efficacy critical values* and *futility critical values*.

Data Monitoring Committee (DMC) or **Data and Safety Monitoring Committee (DSMC)** or **Data and Safety Monitoring Board (DSMB)**. In the context of a clinical trial, a DMC is a panel of experts that is tasked with periodically reviewing data collected by the trial. The DMC will analyze data on safety concerns, such as adverse events suffered by study participants, and the DMC will advise the sponsor of the trial if the study is believed to pose unnecessary risk to participants. In adaptive clinical trials that allow stopping for efficacy or futility, the DMC will perform interim analyses of incomplete trial data to evaluate the effectiveness of the experimental treatment. Not all clinical trials require the use of a DMC.

delta. Delta, δ, in the context of power and sample-size calculations, denotes the effect size. In the context of a Wang–Tsiatis efficacy or futility boundary, capital Greek letter Delta, Δ, represents the parameter of the boundary calculation. See *Classical (Wang–Tsiatis) bounds* in *Methods and formulas* of [ADAPT] **gsbounds** for the formula. Also see *effect size*.

directional test. See *one-sided test*.

dropout. Dropout is the withdrawal of participants before the end of a study and leads to incomplete or missing data. Also see *withdrawal*.

effect size. The effect size is the size of the clinically meaningful difference between the treatments being compared, typically expressed as a quantity that is independent of the unit of measure. For example, in a one-sample mean test, the effect size is a standardized difference between the mean and its reference value. In other cases, the effect size may be measured as an odds ratio or a risk ratio. Also see *delta*.

efficacy or **clinical efficacy**. Efficacy, the capacity to produce a desired result, is an important concept in clinical trials. In the context of a clinical trial, efficacy is quantified by measuring one or more endpoints. The efficacy of an experimental treatment is most commonly established by demonstrating that the experimental treatment compares favorably against a control treatment, but in the case of single-arm clinical trials, the endpoint from the group receiving the experimental treatment is compared against a prespecified reference value. In a clinical trial designed to demonstrate efficacy, the null hypothesis is that the experimental treatment lacks efficacy, and efficacy is established by rejecting H_0. Also see *efficacy boundaries*, *efficacy stopping*, and *futility*.

efficacy boundaries or **efficacy bounds**. In the context of group sequential designs for clinical trials, efficacy bounds are boundaries of the rejection region. If a test statistic is equal to or more extreme than the efficacy critical value, the test statistic is within the rejection region and the null hypothesis is rejected, allowing the trial to be terminated for treatment efficacy. Also see *futility boundaries*, *efficacy*, and *efficacy critical values*.

efficacy critical values. Efficacy critical values define efficacy boundaries in a group sequential design. At each look, a hypothesis test is conducted. If the test statistic is a z statistic, it is compared directly with the efficacy critical value; if not, the significance level approach is used to compare the significance level of the test statistic with the significance level of the efficacy critical value. Also see *efficacy boundaries* and *futility critical values*.

efficacy stopping. In the context of group sequential designs for clinical trials, efficacy stopping refers to the early termination of a clinical trial due to treatment efficacy. This occurs when the test statistic calculated at an interim analysis lies within the rejection region, so the null hypothesis is rejected. Also see *efficacy* and *futility stopping*.

endpoint. The endpoint of a clinical trial is the target parameter that is used for hypothesis testing. Often, the clinical outcome of interest is difficult, time consuming, or expensive to measure, so a surrogate endpoint is measured instead. If there are multiple endpoints of interest, it is common to combine them into a single composite endpoint or to designate a primary endpoint that is used for sample-size determination. Also see *surrogate endpoint* and *composite endpoint*.

equal-allocation design. See *balanced design*.

error-spending approach or **error-spending function**. Instead of calculating boundary critical values directly, the error-spending approach to group sequential designs defines an error-spending function that partitions the alpha (for efficacy bounds) or beta (for futility bounds) into per-look probabilities of committing a type I or type II error. The critical value at each look is calculated based on the error spent, and the critical value at a look does not depend on critical values of future looks.

error-spending O'Brien–Fleming-style bound. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an error-spending O'Brien–Fleming-style bound. Boundary critical values from an error-spending O'Brien–Fleming-style bound are very similar to those of classical O'Brien–Fleming bounds, but they are obtained using an error-spending function. Also see *O'Brien–Fleming bounds* and *error-spending approach*.

error-spending Pocock-style bound. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an error-spending Pocock-style bound. Boundary critical values from an error-spending Pocock-style bound are very similar to those of classical Pocock bounds, but they are obtained using an error-spending function. Also see *Pocock bounds* and *error-spending approach*.

ESS. See *expected sample size*.

exact test. An exact test is one for which the probability of observing the data under the null hypothesis is calculated directly, often by enumeration. Exact tests do not rely on any asymptotic approximations and are therefore widely used with small datasets. See [ADAPT] **gsdesign oneproportion** and [ADAPT] **gsdesign twoproportions**.

expected sample size (ESS) or **average sample number**. In the context of a group sequential design, the ESS is the average sample size that would be required if the trial were to be repeated many times with the same design and with a given effect size. The ESSs under the null and alternative hypotheses are denoted as ESS_0 and ESS_1, respectively. Also see *maximum sample size*.

experimental arm. See *experimental group*.

experimental group. An experimental group (or arm) is a group of participants that receives a treatment or procedure of interest defined in a controlled experiment. In hypothesis testing, this is usually a comparison group. Also see *control group*.

experimental study. In an experimental study, as opposed to an observational study, the assignment of participants to treatments is controlled by investigators. For example, a study that compares a new treatment with a standard treatment by assigning each treatment to a group of participants is an experimental study. Also see *observational study*.

experimental treatment. In a clinical trial, an experimental treatment is a new treatment, such as a drug, medical device, or medical procedure, that is being tested. Typically, the experimental treatment is compared with a control treatment, but in the case of single-arm clinical trials, the endpoint from the group receiving the experimental treatment is compared with a prespecified reference value. Also see *control treatment*.

failure function. When analyzing time-to-event data, the failure function is the probability of experiencing a failure event at or before time t. If we denote the time of failure as T, we can define the failure function as the cumulative distribution function of T, where $F(t) = \Pr(T \leq t)$. The probability density function of T is the derivative of the failure function with respect to time, written as $f(t) = \partial F(t)/\partial t$. Also see *hazard function* and *survivor function*.

familywise error rate or **familywise type I error**. When multiple hypothesis tests are conducted, the familywise error rate is the probability of committing a type I error during at least one test. Also see *type I error* and *familywise significance level*.

familywise significance level. When multiple hypothesis tests are conducted, the familywise significance level is an upper bound to the familywise error rate. Also see *significance level* and *familywise error rate*.

finite population correction. When sampling is performed without replacement from a finite population, a finite population correction is applied to the standard error of the estimator to reduce sampling variance.

Fisher–Irwin exact test. See *Fisher's exact test*.

Fisher's exact test. Fisher's exact test is an exact small-sample test of independence between rows and columns in a 2×2 contingency table. Conditional on the marginal totals, the test statistic has a hypergeometric distribution under the null hypothesis. See [ADAPT] **gsdesign twoproportions** and [R] **tabulate twoway**.

Fisher information or **information**. When estimating parameters from data, the Fisher information for those parameters is a matrix that quantifies the precision with which the parameters can be estimated from the data. Technically, the Fisher information is the expected value of the negative Hessian matrix of the log likelihood. In the context of clinical trials, it is common to conduct a hypothesis test of a single parameter, in which case the Fisher information is a scalar. In this case, a larger value of the Fisher information indicates that more is known about the parameter (typically due to a larger sample size). Also see *information ratio*.

fixed-sample design (FSD) or **fixed study design** or **fixed-sample study design** or **fixed design**. An FSD is an experimental design where the sample size is fixed. See [ADAPT] **GSD intro** for a comparison of FSDs versus group sequential designs.

follow-up period or **follow-up**. The (minimum) follow-up period is the period after the last participant entered the study until the end of the study. During the follow-up period, existing participants are under observation and no new participants enter the study. If T is the total duration of a study and r is the accrual period of the study, then follow-up period f is equal to $T - r$. Also see *accrual period*.

fractional sample size. Sample-size calculations that compute sample size as a continuous quantity will often produce noninteger sample sizes. In practice, a fractional sample size must be rounded up to a whole number of participants. This rounding can cause the attained power to exceed the requested power. Also see *sample size*.

futility. Futility, defined as a lack of the ability to produce a desired result, has particular importance in the context of a clinical trial designed to demonstrate treatment efficacy. In this case, futility refers to the inability of the clinical trial to reject the null hypothesis and demonstrate efficacy. Clinical trials allowing for futility stopping may be terminated early for futility if the result of an interim analysis supports accepting the null hypothesis. Also see *futility boundaries*, *futility stopping*, and *efficacy*.

futility boundaries or **futility bounds**. In the context of group sequential designs for clinical trials, futility bounds are boundaries of the acceptance region. If a test statistic is less extreme than the futility critical value, the test statistic is within the acceptance region and the null hypothesis can be accepted, allowing the trial to be terminated for treatment futility.

There are two types of futility boundaries, binding futility boundaries and nonbinding futility boundaries. If the test statistic at an interim analysis crosses a binding futility boundary, the trial must be stopped for futility; otherwise, it risks overrunning the specified significance level. If a nonbinding futility boundary is used, the familywise type I error is controlled even if the trial continues after crossing the futility boundary. Also see *efficacy boundaries*, *futility*, and *futility critical values*.

futility critical values. Futility critical values define futility boundaries in a group sequential design. At each look, a hypothesis test is conducted. If the test statistic is a z statistic, it is compared directly with the futility critical value; if not, the significance level approach is used to compare the significance level of the test statistic to the significance level of the futility critical value. Also see *futility boundaries* and *efficacy critical values*.

futility stopping. In the context of group sequential designs for clinical trials, futility stopping refers to the early termination of a clinical trial due to treatment futility, often described as “abandoning a lost cause”. This occurs when the test statistic calculated at an interim analysis lies within the acceptance region and the null hypothesis is accepted. Also see *futility* and *efficacy stopping*.

group sequential clinical trial or **group sequential trial**. A group sequential clinical trial is a clinical trial that uses a group sequential design. Also see *group sequential design (GSD)*.

group sequential design (GSD). A GSD is an experimental design where the sample size is not fixed in advance, and preplanned interim analyses of the partial dataset are conducted (typically during the accrual period) to allow early stopping for efficacy or futility. GSDs are frequently used in clinical trials.

GSD. See *group sequential design (GSD)*.

hazard function. When analyzing time-to-event data, the hazard function at time t is the instantaneous rate of failure at time t, conditional on survival until time t. The hazard function is written as $h(t) = f(t)/S(t)$, where $f(t)$ is the derivative of the failure function with respect to time, written as $f(t) = \partial F(t)/\partial t$, and $S(t)$ is the survivor function. Also see *failure function* and *survivor function*.

hazard ratio and **log hazard-ratio**. The hazard ratio is the ratio of the hazard functions of two different populations. If the hazard functions are proportional, then $h_2(t) = \Delta h_1(t)$ for all t or, equivalently, $S_2(t) = \{S_1(t)\}^{\Delta}$. Here $h_1(t)$ and $h_2(t)$ are the hazard functions for the control group and the experimental group, respectively; Δ is the hazard ratio; and $S_1(t)$ and $S_2(t)$ are the survivor functions of the control and the experimental groups, respectively.

The log hazard-ratio is the natural logarithm of the hazard ratio. If a log-rank test is used to compare the survivor functions of the two populations, under the proportional-hazards assumption the null hypothesis is $H_0\colon \Delta = 1$ or, equivalently, $H_0\colon \ln(\Delta) = 0$. See [ADAPT] **gsdesign logrank**.

Also see *hazard function* and *time-to-event data*.

Hwang–Shih–de Cani bound or **error-spending Hwang–Shih–de Cani bound** or **Hwang–Shih–de Cani design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an error-spending Hwang–Shih–de Cani design. Hwang–Shih–de Cani bounds are defined by an error-spending function indexed by parameter γ, and smaller values of γ yield bounds that are more conservative at early looks. Also see *error-spending approach*.

hypothesis. A hypothesis is a statement about a population parameter of interest.

hypothesis testing or **hypothesis test**. This method of inference evaluates the validity of a hypothesis based on a sample from the population. See *Hypothesis testing* in *Remarks and examples* of [PSS-2] **Intro (power)**.

information fraction. In a group sequential clinical trial, the information fraction is the proportion of the maximum information that has been collected at the time of a scheduled look at the clinical trial data. In most cases, the information fraction is the proportion of the maximum sample size that has been collected. For time-to-event data, the information fraction is the proportion of the total number of failure events that have been observed, not the total number of study participants.

information ratio. In the context of a group sequential clinical trial, the information ratio is the ratio of the maximum information of the group sequential trial to the Fisher information of an equivalent fixed study design. In most cases, this is the ratio of the maximum sample size of the group sequential trial to the sample size of the fixed design, but for trials with time-to-event endpoints, the information ratio corresponds to the ratio of the maximum number of failure events observed in the group sequential trial to the number of failures observed in a fixed-design trial.

interim analysis or **interim look**. In the context of an adaptive clinical trial, an interim look is an analysis of trial data that is conducted while the trial is still under way and before the maximum sample size has been reached.

Kim–DeMets bound or **error-spending Kim–DeMets bound** or **Kim–DeMets design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an error-spending Kim–DeMets design. Kim–DeMets bounds are defined by an error-spending function indexed by parameter ρ, and larger values of ρ yield bounds that are more conservative at early looks. Also see *error-spending approach*.

likelihood-ratio test. The likelihood-ratio test is one of the three classical testing procedures used to compare the fit of two models, one of which, the constrained model, is nested within the full (unconstrained) model. Under the null hypothesis, the constrained model fits the data as well as the full model. The likelihood-ratio test requires one to determine the maximal value of the log-likelihood function for both the constrained and the full models. See [ADAPT] **gsdesign twoproportions** and [R] **lrtest**.

look. In the context of a group sequential clinical trial, a look is an analysis of the clinical trial data that has been collected up to that point. Looks conducted while the trial is still collecting data are called interim looks, and the final look is performed when data from the maximum sample size have been collected. Also see *interim analysis*.

loss to follow-up. Participants are lost to follow-up if they do not complete the course of the study for reasons unrelated to the event of interest. For example, loss to follow-up occurs if participants move to a different area or decide to no longer participate in a study. Loss to follow-up should not be confused with administrative censoring. If participants are lost to follow-up, the information about the outcome those participants would have experienced at the end of the study, had they completed the study, is unavailable. Also see *withdrawal*, *administrative censoring*, and *follow-up period*.

lower one-sided test or **lower one-tailed test**. A lower one-sided test is a one-sided test of a scalar parameter in which the alternative hypothesis is lower one-sided, meaning that the alternative hypothesis states that the parameter is less than the value conjectured under the null hypothesis. Also see *One-sided test versus two-sided test* in *Remarks and examples* of [PSS-2] **Intro (power)**.

maximum information. In a group sequential clinical trial, the maximum information is the Fisher information of the parameter estimated during the hypothesis test, calculated at the maximum sample size. Also see *information fraction*.

maximum sample size. In a clinical trial following an adaptive design, the sample size of the trial is often not fixed in advance. However, in many adaptive designs, such as group sequential designs, the maximum possible sample size can be calculated before the study begins. Also see *expected sample size* and *sample size*.

nominal alpha or **nominal significance level**. This is a desired or requested significance level. Also see *familywise significance level*.

nonbinding futility boundaries or **nonbinding futility bounds**. In a group sequential clinical trial with nonbinding futility bounds, if the test statistic at an interim analysis crosses the futility boundary, the trial may be stopped for futility or continued without risk of overrunning the specified significance level. Group sequential designs with nonbinding futility bounds use the same efficacy critical values as equivalent group sequential designs without futility stopping. Also see *binding futility boundaries*.

noninferiority trial. A noninferiority trial is a clinical trial where the goal is to determine whether the experimental treatment is unacceptably inferior to the control (or comparator) treatment, which is almost always an active control. If the experimental treatment has some advantageous characteristics (for example, it produces fewer side effects than the control, is less expensive, or is easier to administer), practitioners might prefer the experimental treatment even if it is not more efficacious than the control.

When designing a noninferiority trial, researchers define a noninferiority margin, denoted as δ, to quantify an acceptable reduction in efficacy. The null hypothesis in a noninferiority trial is that the effect of the control treatment beats the effect of the experimental treatment by a margin of δ or more; the one-sided alternative hypothesis is that the effect of the control treatment does not beat the effect of the experimental treatment by a margin of at least δ. For example, if the endpoint is a population mean and an upper one-sided test is desired, δ will be < 0 and the null and alternative hypotheses can be written as $H_0 : \mu_e - \mu_c \leq \delta$ and $H_a : \mu_e - \mu_c > \delta$, where μ_e is the mean response of the experimental group and μ_c is the mean response of the control group. Also see the related *substantial superiority trial*.

null hypothesis. In hypothesis testing, the null hypothesis typically represents the conjecture that one is attempting to disprove. Often, the null hypothesis is that a treatment has no effect or that a statistic is equal across populations.

O'Brien–Fleming bounds or **classical O'Brien–Fleming bounds** or **O'Brien–Fleming design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use an O'Brien–Fleming design. O'Brien–Fleming efficacy bounds are characterized by being extremely conservative at early looks. O'Brien–Fleming bounds are a special case of classical Wang–Tsiatis bounds with parameter $\Delta = 0$. Also see *Wang–Tsiatis bounds*.

observational study. In an observational study, as opposed to an experimental study, the assignment of participants to treatments happens naturally and is thus beyond the control of investigators. Investigators can only observe participants and measure their characteristics. For example, a study that evaluates the effect of exposure of children to household pesticides is an observational study. Also see *experimental study*.

observed level of significance. See *p-value*.

odds and **odds ratio**. The odds in favor of an event are $\text{Odds} = p/(1-p)$, where p is the probability of the event. Thus, if $p = 0.2$, the odds are 0.25, and if $p = 0.8$, the odds are 4.

The log of the odds is $\ln(\text{Odds}) = \text{logit}(p) = \ln\{p/(1-p)\}$, and logistic regression models, for instance, fit $\ln(\text{Odds})$ as a linear function of the covariates.

The odds ratio is a ratio of two odds: $\text{Odds}_2/\text{Odds}_1$. The individual odds that appear in the ratio are usually for an experimental group and a control group or for two different demographic groups.

one-sample test. A one-sample test compares a parameter of interest from one sample to a reference value. For example, a one-sample mean test compares a mean of the sample against a reference value.

one-sided test or **one-tailed test**. A one-sided test is a hypothesis test of a scalar parameter in which the alternative hypothesis is one-sided, meaning that the alternative hypothesis states that the parameter is either less than or greater than the value conjectured under the null hypothesis, but not both. Also see *One-sided test versus two-sided test* in *Remarks and examples* of [PSS-2] **Intro (power)**.

overall significance level. See *familywise significance level*.

Pearson's χ^2 test. In the context of a clinical trial, Pearson's χ^2 test is commonly used to test whether the observed event counts in a contingency table are consistent with the null hypothesis. See [ADAPT] **gsdesign twoproportions**. Also see *2×2 contingency tables*.

placebo or **sham treatment**. In a clinical trial, a placebo is an inactive treatment, such as a sugar pill, that is designed to look like the experimental treatment. In studies of medical procedures, the term *sham treatment* is often used. Typically, study participants receiving a placebo are blinded, meaning that they are not told whether they are receiving the placebo or the experimental treatment. Also see *standard of care*.

placebo control. In a clinical trial, a placebo control is a control group that receives a placebo instead of an active control.

Pocock bounds or **classical Pocock bounds** or **Pocock design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use a Pocock design. Pocock efficacy bounds are characterized by using the same critical value at all looks. Pocock bounds are a special case of classical Wang–Tsiatis bounds with parameter $\Delta = 0.5$. Also see *Wang–Tsiatis bounds*.

population parameter. See *target parameter*.

power. The power of a test is the probability of correctly rejecting the null hypothesis when it is false. It is often denoted as $1-\beta$ in the statistical literature, where β is the type-II-error probability. Commonly used values for power are 80% and 90%. See [PSS-2] **Intro (power)** for more details about power.

power and sample-size (PSS) analysis. Power and sample-size analysis investigates the optimal allocation of study resources to increase the likelihood of the successful achievement of a study objective. The focus of power and sample-size analysis is on studies that use hypothesis testing for inference. Power and sample-size analysis provides an estimate of the sample size required to achieve the desired power of a test in a future study. See [PSS-2] **Intro (power)**.

probability of a type I error. This is the probability of committing a type I error and incorrectly rejecting the null hypothesis. Also see *type I error* and *significance level*.

probability of a type II error. This is the probability of committing a type II error and incorrectly accepting the null hypothesis. Common values for the probability of a type II error are 0.1 and 0.2 or, equivalently, 10% and 20%. Also see *type II error*, *beta*, and *power*.

PSS analysis. See *power and sample-size (PSS) analysis*.

p-value. The p-value is the probability of obtaining a test statistic as extreme as or more extreme than the one observed in a sample assuming the null hypothesis is true.

randomized controlled trial (RCT). In this experimental study, treatments are randomly assigned to two or more groups of participants, one of which is a control group.

recruitment period or **recruitment**. See *accrual period*.

rejection region. In hypothesis testing, a rejection region is a set of values of a test statistic for which the null hypothesis can be rejected. In the context of a group sequential design, a trial can be terminated early for efficacy if the test statistic falls within the rejection region during an interim analysis. Also see *acceptance region* and *continuation region*.

relative risk. See *risk ratio*.

risk difference. A risk difference is defined as the probability of an event occurring when a risk factor is increased by one unit minus the probability of the event occurring without the increase in the risk factor.

When the risk factor is binary, the risk difference is the probability of the outcome when the risk factor is present minus the probability when the risk factor is not present.

When one compares two populations, a risk difference is defined as a difference between the probabilities of an event in the two groups. It is typically a difference between the probability in the comparison group or experimental group and the probability in the reference group or control group.

risk factor. A risk factor is a variable that is associated with an increased or decreased probability of an outcome.

risk ratio or **relative risk**. A risk ratio, also called a relative risk, measures the increase in the likelihood of an event occurring when a risk factor is increased by one unit. It is the ratio of the probability of the event when the risk factor is increased by one unit over the probability without that increase.

When the risk factor is binary, the risk ratio is the ratio of the probability of the event when the risk factor occurs over the probability when the risk factor does not occur.

When one compares two populations, a risk ratio is defined as a ratio of the probabilities of an event in the two groups. It is typically a ratio of the probability in the comparison group or experimental group to the probability in the reference group or control group.

sample size. This is the number of participants in a sample. In a clinical trial with time-to-event data, the effective sample size is the number of events observed. In this case, sample-size calculations will determine the number of events that must be observed to achieve the specified power. If administrative censoring, loss to follow-up, or withdrawal are expected, the total required sample size can be estimated and will be larger than the number of events observed. Also see *expected sample size*, *fractional sample size*, and *maximum sample size*.

sample-size determination. This pertains to the computation of a sample size given power and effect size and any other study parameters.

sample-size ratio. The ratio of the experimental-group sample size relative to the control-group sample size, n_2/n_1.

Satterthwaite's t test. Satterthwaite's t test is a modification of the two-sample t test to account for unequal variances in the two populations. See [ADAPT] **gsdesign twomeans** for an example and see *Methods and formulas* of [PSS-2] **power twomeans** for formulas.

score test. A score test, also known as a Lagrange multiplier test, is one of the three classical testing procedures used to compare the fit of two models, one of which, the constrained model, is nested within the full (unconstrained) model. The null hypothesis is that the constrained model fits the data as well as the full model. The score test only requires one to fit the constrained model. See [ADAPT] **gsdesign oneproportion** and [R] **prtest**.

sensitivity analysis. Sensitivity analysis investigates the effect of varying study parameters on power, sample size, and other components of a study. The true values of study parameters are usually unknown, and analyses of power and sample size use best guesses for these values. It is therefore important to evaluate the sensitivity of the computed power or sample size in response to changes in study parameters.

significance level. In hypothesis testing, the significance level α is an upper bound for the probability of a type I error. Also see *alpha*, *probability of a type I error*, and *familywise significance level*.

significance level approach. The efficacy and futility critical values from a group sequential design are intended to be compared with z statistics. If the test statistic used does not follow a standard normal distribution under the null hypothesis, the significance level approach is used to compare the significance level of the test statistic against the significance level of the efficacy critical value. This is done by comparing the p-value of the test statistic against the p-value corresponding to the efficacy or futility critical value. See *Significance level approach* in *Methods and formulas* of [ADAPT] **gsbounds** for details.

single-arm trial. A single-arm clinical trial is a trial where all study participants receive the experimental treatment. Because there is no control group, the endpoint is compared with a prespecified reference value. Also see *two-arm trial*.

size of test. See *significance level*.

standard of care. The standard of care is the medically accepted first-line treatment for a disease or condition. In a clinical trial of a treatment for a condition where there is a recognized standard of care, it is common to compare the experimental treatment to an active control consisting of participants who receive the standard of care. Also see *active control* and *placebo*.

stopping boundary. A stopping boundary is a set of critical values that define an efficacy or futility boundary. Also see *stopping rule*, *efficacy boundaries*, and *futility boundaries*.

stopping rule. In the context of a group sequential clinical trial, a stopping rule refers to an efficacy or futility boundary that allows the trial to be terminated before data from the maximum sample size have been collected. This occurs when the test statistic at an interim analysis crosses the efficacy or futility boundary, leading to the rejection or acceptance of the null hypothesis. Also see *efficacy stopping* and *futility stopping*.

study participant. Human subjects who volunteer to join a clinical trial are known as study participants.

substantial superiority trial or **superiority trial**. A substantial superiority trial is a clinical trial where the goal is to determine whether the experimental treatment is substantially superior to the control treatment. This is done by defining a clinically relevant superiority margin, denoted as δ, before the trial begins. The one-sided alternative hypothesis is that the effect of the experimental treatment beats the effect of the control treatment by a margin greater than δ; the null hypothesis is that it does not. For example, if the endpoint is a population mean and an upper one-sided test is desired, δ will be > 0 and the null and alternative hypotheses can be written as $H_0 : \mu_e - \mu_c \leq \delta$ and $H_a : \mu_e - \mu_c > \delta$, where μ_e is the mean response of the experimental group and μ_c is the mean response of the control (or comparator) group. Also see related concept *noninferiority trial*.

surrogate endpoint. When the clinical outcome of interest is too difficult, time consuming, or expensive to measure, clinical trials often use a surrogate endpoint as their target parameter. A surrogate endpoint is an endpoint that is known to be associated with the clinical outcome of interest but is easier to measure. Many clinical trials use biomarkers as surrogate endpoints. Also see *endpoint*.

survivor function. When analyzing time-to-event data, the survivor function is defined as the probability of surviving beyond time t. If we denote the time of failure as T, we can define the survivor function as $S(t) = \Pr(T > t) = 1 - F(t)$, where $F(t)$ is the failure function. Also see *hazard function* and *failure function*.

t test. A t test is a test for which the sampling distribution of the test statistic is a Student's t distribution.

A one-sample t test is used to test whether the mean of a population is equal to a specified value when the variance must also be estimated. The test statistic follows Student's t distribution with $N - 1$ degrees of freedom, where N is the sample size.

A two-sample t test is used to test whether the means of two populations are equal when the variances of the populations must also be estimated. When the two populations' variances are unequal, a modification to the standard two-sample t test is used; see *Satterthwaite's t test*.

target parameter. In power and sample-size analysis, the target parameter is the parameter of interest or the parameter in the study about which hypothesis tests are conducted. Also see *endpoint*.

test statistic. In hypothesis testing, a test statistic is a function of the sample that does not depend on any unknown parameters.

time-to-event data or **survival data**. Time-to-event data, also known as survival data, are collected from clinical trials where the endpoint is the amount of time elapsed before a participant experiences a failure event. See [ADAPT] **gsdesign logrank**.

two-arm trial. A two-arm clinical trial is a trial where participants are assigned to one of two treatment groups. Typically, one group is an experimental group and the other is a control group. Also see *single-arm trial*.

two-sample test. A two-sample test is used to test whether the parameters of interest of the two independent populations are equal, for example, a two-sample test of means, proportions, or hazard ratios. See [ADAPT] **gsdesign twomeans**, [ADAPT] **gsdesign twoproportions**, and [ADAPT] **gsdesign logrank**.

two-sided test or **two-tailed test**. A two-sided test is a hypothesis test of a parameter in which the alternative hypothesis is the complement of the null hypothesis. In the context of a test of a scalar parameter, the alternative hypothesis states that the parameter is less than or greater than the value conjectured under the null hypothesis.

type I error. The type I error of a test is the error of rejecting the null hypothesis when it is true. Also see *probability of a type I error* and *familywise type I error*.

type II error. The type II error of a test is the error of not rejecting the null hypothesis when it is false. Also see *probability of a type II error*.

unbalanced design or **unequal-allocation design**. An unbalanced design indicates an experiment in which the numbers of treated and untreated participants differ. Also see [PSS-4] **Unbalanced designs**.

upper one-sided test or **upper one-tailed test**. An upper one-sided test is a one-sided test of a scalar parameter in which the alternative hypothesis is upper one-sided, meaning that the alternative hypothesis states that the parameter is greater than the value conjectured under the null hypothesis. Also see *One-sided test versus two-sided test* in *Remarks and examples* of [PSS-2] **Intro (power)**.

Wald test. A Wald test is one of the three classical testing procedures used to compare the fit of two models, one of which, the constrained model, is nested within the full (unconstrained) model. Under the null hypothesis, the constrained model fits the data as well as the full model. The Wald test requires one to fit the full model but does not require one to fit the constrained model. Also see [ADAPT] **gsdesign oneproportion** and [R] **test**.

Wang–Tsiatis bounds or **classical Wang–Tsiatis bounds** or **Wang–Tsiatis design**. In a group sequential clinical trial, one technique for calculating efficacy or futility boundaries is to use a Wang–Tsiatis design. Wang–Tsiatis bounds are indexed by parameter Δ, and smaller values of Δ yield bounds that are more conservative at early looks. Classical Pocock bounds and classical O'Brien–Fleming bounds are both special cases of the Wang–Tsiatis family of bounds. Also see *Pocock bounds* and *O'Brien–Fleming bounds*.

withdrawal. Withdrawal is the process under which participants withdraw from a study for reasons unrelated to the event of interest. For example, withdrawal occurs if participants move to a different area or decide to no longer participate in a study. Withdrawal should not be confused with administrative censoring. If participants withdraw from the study, the information about the outcome those participants would have experienced at the end of the study, had they completed the study, is unavailable. Also see *loss to follow-up* and *administrative censoring*.

z **statistic**. A z statistic is a test statistic that follows the standard normal distribution under the null hypothesis. Also see *z test*.

z **test**. A z test is a test for which a potentially asymptotic sampling distribution of the test statistic is a normal distribution. For example, a one-sample z test of means is used to test whether the mean of a population is equal to a specified value when the variance is assumed to be known. The distribution of its test statistic is normal. See [ADAPT] **gsdesign onemean** and [ADAPT] **gsdesign twomeans**.

Reference

US Food and Drug Administration. 2019. Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. https://www.fda.gov/media/78495/download.

Subject and author index

See the combined subject index and the combined author index in the *Stata Index*.