# Title

stata.com

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



## 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 $\chi^2$ test of $H_0\colon \pi_1 = \pi_2$ versus $H_a\colon \pi_1 \neq \pi_2$, 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 p1 p2 [ , 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 $\chi^2$ 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 $\mathbf{g}\mathbf{H}^{-1}\mathbf{g}' <$ `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 $\Delta$ 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 $\rho$ 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 $\gamma$ 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

stata.com

Remarks are presented under the following headings:



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 $\chi^2$ 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 $\chi^2$ 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 $\chi^2$ 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
```

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

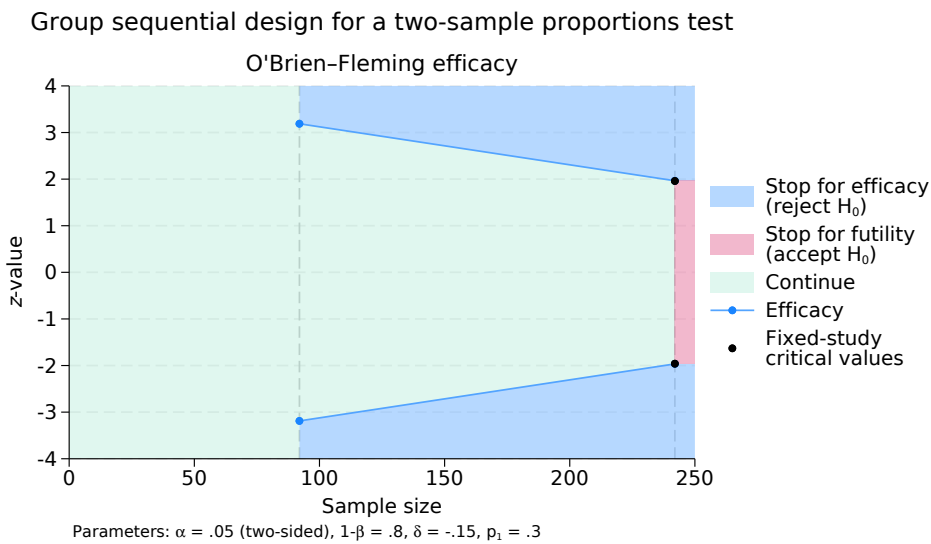


Figure 1. Two-sided test of the equality of two proportions with O'Brien–Fleming efficacy bounds

`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, $\pm 3.188$, sets a high bar for early stopping. Once data have been collected from 46 subjects in each group, Pearson's $\chi^2$ test is performed and the $z$ statistic, $z_1$, is compared with the efficacy critical values of $\pm 3.188$. To perform Pearson's $\chi^2$ 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 $\chi^2$ statistic (the $\chi^2$ 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 $\pm 1.965$, 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 $\chi^2$ 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
```

| Look | Info. frac. | Efficacy Lower | Efficacy Upper | p-value | Sample size N1 | Sample size N2 | Sample size 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 $\leq 0.0014$ 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
```

| row | col 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 $\Delta$, 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
```

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

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

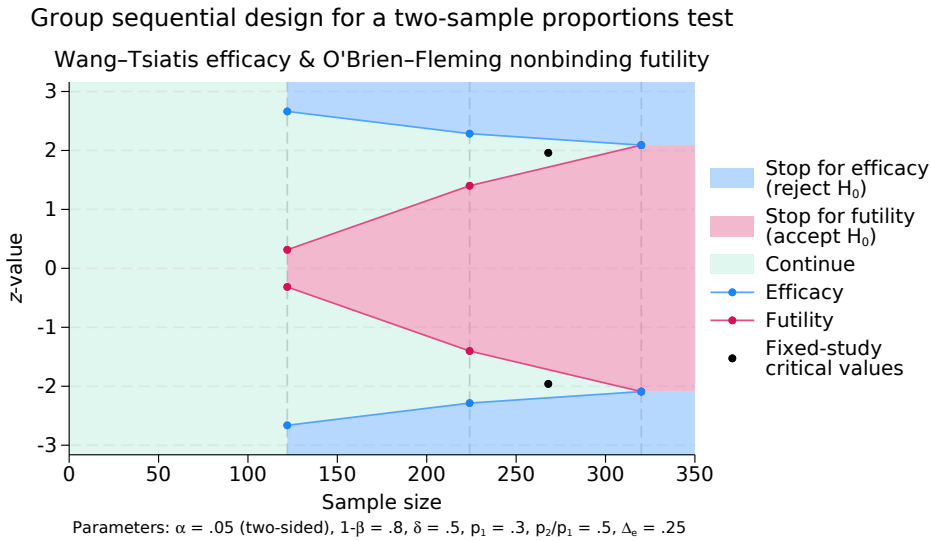


Figure 2. Two-sided test of the equality of two proportions with efficacy and futility bounds

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