Title

stata.com

stpower logrank — Sample size, power, and effect size for the log-rank test

Syntax

Sample-size determination

```
stpower logrank [ surv1 [ surv2 ] ] [ , options ]
```

Power determination

```
stpower logrank [ surv1 [ surv2 ] ] , n(numlist) [ options ]
```

Effect-size determination

```
stpower logrank [ surv1 ] , n(numlist) { power(numlist) | beta(numlist) } [ options ]
```

where

surv$_1$ is the survival probability in the control group at the end of the study t^*;

surv$_2$ is the survival probability in the experimental group at the end of the study t^*.

surv$_1$ and *surv*$_2$ may each be specified either as one number or as a list of values (see **[U] 11.1.8 numlist**) enclosed in parentheses.

options	Description
Main	
*`alpha(`*numlist*`)`	significance level; default is `alpha(0.05)`
*`power(`*numlist*`)`	power; default is `power(0.8)`
*`beta(`*numlist*`)`	probability of type II error; default is `beta(0.2)`
*`n(`*numlist*`)`	sample size; required to compute power or effect size
*`hratio(`*numlist*`)`	hazard ratio (effect size) of the experimental to the control group; default is `hratio(0.5)`
`onesided`	one-sided test; default is two sided
*`p1(`*numlist*`)`	proportion of subjects in the control group; default is `p1(0.5)`, meaning equal group sizes
*`nratio(`*numlist*`)`	ratio of sample sizes, N_2/N_1; default is `nratio(1)`, meaning equal group sizes
`schoenfeld`	use the formula based on the log hazard-ratio in calculations; default is to use the formula based on the hazard ratio
`parallel`	treat number lists in starred options as parallel (do not enumerate all possible combinations of values) when multiple values per option are specified

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(#)	the proportion of subjects anticipated to withdraw from the study; default is wdprob(0)

Reporting

table	display results in a table with default columns
columns(*colnames*)	display results in a table with specified *colnames* columns
notitle	suppress table title
nolegend	suppress table legend
colwidth(# [# ...])	column widths; default is colwidth(9)
separator(#)	draw a horizontal separator line every # lines; default is separator(0), meaning no separator lines
saving(*filename* [, replace])	save the table data to *filename*; use replace to overwrite existing *filename*
noheader	suppress table header; seldom used
continue	draw a continuation border in the table output; seldom used

*Starred options may be specified either as one number or as a list of values (see [U] 11.1.8 numlist). noheader and continue are not shown in the dialog box.

colnames	Description
alpha	significance level
power	power
beta	type II error probability
n	total number of subjects
n1	number of subjects in the control group
n2	number of subjects in the experimental group
e	total number of events (failures)
hr	hazard ratio
loghr	log of the hazard ratio
s1	survival probability in the control group
s2	survival probability in the experimental group
p1	proportion of subjects in the control group
nratio	ratio of sample sizes, experimental to control
w	proportion of withdrawals

By default, the following *colnames* are displayed:

`power`, `n`, `n1`, `n2`, `e`, and `alpha` are always displayed;

`hr` is displayed, unless the `schoenfeld` option is specified, in which case `loghr` is displayed;

`s1` and `s2` is displayed if survival probabilities are specified; and

`w` is displayed if withdrawal proportion (`wdprob()` option) is specified.

Menu

Statistics > Survival analysis > Power and sample size

Description

`stpower logrank` estimates required sample size, power, and effect size for survival analysis comparing survivor functions in two groups by using the log-rank test. It also reports the number of events (failures) required to be observed in a study. This command supports two methods to obtain the estimates, those according to Freedman (1982) and Schoenfeld (1981). The command provides options to take into account unequal allocation of subjects between the two groups and possible withdrawal of subjects from the study (loss to follow-up). Optionally, the estimates can be adjusted for uniform accrual of subjects into the study. Also the minimal effect size (minimal detectable value of the hazard ratio or the log hazard-ratio) may be obtained for given power and sample size.

You can use `stpower logrank` to

- calculate required number of events and sample size when you know power and effect size (expressed as a hazard ratio) for uncensored and censored survival data,
- calculate power when you know sample size (number of events) and effect size (expressed as a hazard ratio) for uncensored and censored survival data, and
- calculate effect size (hazard ratio or log hazard-ratio if the `schoenfeld` option is specified) when you know sample size (number of events) and power for uncensored and censored survival data.

`stpower logrank`'s input parameters, $surv_1$ and $surv_2$, are the values of survival probabilities in the control group (or the less favorable of the two groups), s_1, and in the experimental group, s_2, at the end of the study t^*.

Options

Main

`alpha(`*numlist*`)` sets the significance level of the test. The default is `alpha(0.05)`.

`power(`*numlist*`)` sets the power of the test. The default is `power(0.8)`. If `beta()` is specified, this value is set to be 1−`beta()`. Only one of `power()` or `beta()` may be specified.

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

`n(`*numlist*`)` specifies the number of subjects in the study to be used to compute the power of the test or the minimal effect size (minimal detectable value of the hazard ratio or log hazard-ratio) if `power()` or `beta()` is also specified.

`hratio(`*numlist*`)` specifies the hazard ratio (effect size) of the experimental group to the control group. The default is `hratio(0.5)`. This value defines the clinically significant improvement of the experimental procedure over the control desired to be detected by the log-rank test, with a certain power specified in `power()`. If both arguments *surv*$_1$ and *surv*$_2$ are specified, `hratio()` is not allowed and the hazard ratio is instead computed as $\ln(surv_2)/\ln(surv_1)$.

`onesided` indicates a one-sided test. The default is two sided.

`p1(`*numlist*`)` specifies the proportion of subjects in the control group. The default is `p1(0.5)`, meaning equal allocation of subjects to the control and the experimental groups. Only one of `p1()` or `nratio()` may be specified.

`nratio(`*numlist*`)` specifies the sample-size ratio of the experimental group relative to the control group, N_2/N_1. The default is `nratio(1)`, meaning equal allocation between the two groups. Only one of `nratio()` or `p1()` may be specified.

`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).

`parallel` reports results sequentially (in parallel) over the list of numbers supplied to options allowing *numlist*. By default, results are computed over all combinations of the number lists in the following order of nesting: `alpha()`; `p1()` or `nratio()`; list of arguments *surv*$_1$ and *surv*$_2$; `hratio()`; `power()` or `beta()`; and `n()`. This option requires that options with multiple values each contain the same number of elements.

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*. `simpson()` may not be combined with `st1()` and may not be used if arguments *surv*$_1$ or *surv*$_2$ are 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*. `st1()` may not be combined with `simpson()` and may not be used if arguments *surv*$_1$ or *surv*$_2$ are specified.

`wdprob(#)` specifies the proportion of subjects anticipated to withdraw from the study. The default is `wdprob(0)`. `wdprob()` may not be combined with `n()`.

Reporting

`table` displays results in a tabular format and is implied if any number list contains more than one element. This option is useful if you are producing results one case at a time and wish to construct your own custom table using a `forvalues` loop.

`columns(`*colnames*`)` specifies results in a table with specified *colnames* columns. The order of the columns in the output table is the same as the order of *colnames* specified in `columns()`. Column names in `columns()` must be space-separated.

`notitle` prevents the table title from displaying.

`nolegend` prevents the table legend from displaying and column headers from being marked.

`colwidth(#` [`# ...`]`)` specifies column widths. The default is 9 for all columns. The number of specified values may not exceed the number of columns in the table. A missing value (`.`) may be specified for any column to indicate the default width (9). If fewer widths are specified than the number of columns in the table, the last width specified is used for the remaining columns.

`separator(#)` specifies how often separator lines should be drawn between rows of the table. The default is `separator(0)`, meaning that no separator lines should be displayed.

`saving(`*filename* [`, replace`]`)` creates a Stata data file (`.dta` file) containing the table values with variable names corresponding to the displayed *colnames*. `replace` specifies that *filename* be overwritten if it exists. `saving()` is only appropriate with tabular output.

The following options are available with `stpower logrank` but are not shown in the dialog box:

`noheader` prevents the table header from displaying. This option is useful when the command is issued repeatedly, such as within a loop. `noheader` implies `notitle`.

`continue` draws a continuation border at the bottom of the table. This option is useful when the command is issued repeatedly within a loop.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Introduction

Consider a survival study comparing the 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 key assumption of the log-rank test is that the hazard functions are proportional. That is, $h_2(t) = \Delta h_1(t)$ for any t or, equivalently, $S_2(t) = \{S_1(t)\}^{\Delta}$, where Δ is the hazard ratio. If $\Delta < 1$, the survival in the experimental group is higher relative to the survival in the control group; the new treatment is superior to the standard treatment. If $\Delta > 1$, then the standard treatment is superior to the new 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 the test 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 `stpower logrank` for sample-size or power determination relate the power of the log-rank test directly to the number of events observed in the study. Depending on whether censoring occurs in a study, the required number of subjects is either equal to the number of events or is computed using the estimates of the number of events and the combined probability of an event (failure). Thus, in the presence of censoring, in addition to the number of events, the probability of a subject not being censored (failing) needs to be estimated to obtain the final estimate of the required number of subjects in the study.

To determine the required number of events, the investigator must specify the size or significance level, α, and the clinically significant difference between the two treatments (effect size) to be detected by the log-rank test, H_a: $\Delta = \Delta_a$, with prespecified power $1 - \beta$. The significance level, α, represents the probability of a type I error, a rejection of the null hypothesis when it is true. β represents the probability of a type II error, a failure to reject the null hypothesis when the alternative hypothesis is true. The significance level is often set to 0.05, and values for `power()` usually vary from 0.8 to 0.95. By default, `stpower logrank` uses `power(0.8)` (or, equivalently, `beta(0.2)`) and `alpha(0.05)`. The effect size, a difference between the two treatments, is usually expressed as a hazard ratio, Δ_a, using the `hratio()` option. Under an unequal allocation of subjects between the two groups, the proportion of subjects in the control group may be specified in `p1()`, or the ratio of sample sizes may be supplied to `nratio()`. Optionally, results for the one-sided log-rank test may be requested by using `onesided`.

When all subjects fail by the end of the study (no censoring), a type I study, the information above is sufficient to obtain the number of subjects required in the study. Often, in practice, not all subjects fail by the end of the study, in which case censoring of subjects occurs (a type II study). Here the estimates of the survival probabilities in the control and experimental groups are necessary to estimate the probability of an event and, then, the required sample size.

By default, `stpower logrank` performs computations for the uncensored data (a type I study). It uses the hazard ratio specified in `hratio()` or the default hazard ratio of 0.5 to obtain required sample size or power. For censored data (a type II study), under administrative censoring, the value of the survival probability in the control group (supplied as argument *surv*$_1$ or, in the presence of an accrual period, in the `simpson()` or `st1()`) option must be specified. If the value of the survival probability in the experimental group, *surv*$_2$, is omitted, *surv*$_1$ and the value of the hazard ratio in `hratio()` are used to compute the survival probability in the experimental group, s_2. If both arguments *surv*$_1$ and *surv*$_2$ are specified, the hazard ratio, Δ_a, is computed using these values and the `hratio()` option is not allowed.

If power determination is desired, sample size `n()` must be specified. If both `n()` and `power()` (or `beta()`) are specified, the minimal effect size (minimal value of the hazard ratio or log hazard-ratio) that can be detected by the log-rank test with requested power and fixed sample size is computed.

`stpower logrank` supports two methods, those of Freedman (1982) and Schoenfeld (1981), to obtain the estimates of the number of events or power (see also Marubini and Valsecchi [1997, 127, 134] and Collett [2003b, 301, 306]). The latter is used if option `schoenfeld` is specified. The final estimates of the sample size are based on the approximation of the probability of an event due to Freedman (1982), the default, or, for uniform accrual, due to Schoenfeld (1983) (see also Collett 2003b) if option `simpson()` is specified.

Optionally, the results may be displayed in a table by using `table` or `columns()`, as demonstrated in [ST] **stpower**. Refer to [ST] **stpower** and to example 7 in *Power and effect-size determination* to see how to obtain a graph of a power curve.

Computing sample size in the absence of censoring

We demonstrate several examples of how to use `stpower logrank` to obtain the estimates of sample size and number of events using Freedman (1982) and Schoenfeld (1981) methods for uncensored data, a type I study (when no censoring of subjects occurs).

▷ Example 1: Number of events (failures)

Consider a survival study to be conducted to compare the survivor function of subjects receiving a treatment (the experimental group) to the survivor function of those receiving a placebo or no treatment

(the control group) using the log-rank test. Suppose that the study continues until all subjects fail (no censoring). The investigator wants to know how many events need to be observed in the study to achieve a power of 80% of a two-sided log-rank test with $\alpha = 0.05$, to detect a 50% reduction in the hazard of the experimental group ($\Delta_a = 0.5$). Because the default settings of `stpower logrank` are `power(0.8)`, `alpha(0.05)`, and `hratio(0.5)`, to obtain the estimate of the required number of events for the above study using the Freedman method (the default), we simply type

```
. stpower logrank

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)

Input parameters:

      alpha =    0.0500  (two sided)
     hratio =    0.5000
      power =    0.8000
         p1 =    0.5000

Estimated number of events and sample sizes:

          E =        72
          N =        72
         N1 =        36
         N2 =        36
```

From the output, a total of 72 events (failures) must be observed to achieve the required power of 80%. Because all subjects experience an event by the end of the study, the number of subjects required to be recruited to the study is equal to the number of events. That is, the investigator needs to recruit a total of 72 subjects (36 per group) to the study.

We can request the Schoenfeld method by specifying the `schoenfeld` option:

```
. stpower logrank, schoenfeld

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Schoenfeld method
Ho: S1(t) = S2(t)

Input parameters:

      alpha =    0.0500  (two sided)
 ln(hratio) =   -0.6931
      power =    0.8000
         p1 =    0.5000

Estimated number of events and sample sizes:

          E =        66
          N =        66
         N1 =        33
         N2 =        33
```

We obtain a slightly smaller estimate, 66, of the total number of events.

◁

❑ Technical note

Freedman (1982) and Schoenfeld (1981) derive the formulas for the number of events based on the asymptotic distribution of the log-rank test statistic. Freedman (1982) uses the asymptotic mean and variance of the log-rank test statistic expressed as a function of the true hazard ratio, Δ, whereas Schoenfeld (1981) (see also Collett [2003b, 302]) bases the derivation on the asymptotic mean of the log-rank test statistic as a function of the true log hazard-ratio, $\ln(\Delta)$. We label the corresponding approaches as "Freedman method" and "Schoenfeld method" in the output.

For values of the hazard ratio close to one, the two methods tend to give similar results. In general, the Freedman method gives higher estimates than the Schoenfeld method. The performance of the Freedman method was studied by Lakatos and Lan (1992) and was found to slightly overestimate the sample size under the assumption of proportional hazards. Hsieh (1992) investigates the performance of the two methods under the unequal allocation and concludes that Freedman's formula predicts the highest power for the log-rank test when the sample-size ratio of the two groups equals the reciprocal of the hazard ratio. Schoenfeld's formula predicts highest powers when sample sizes in the two groups are equal.

❑

Computing sample size in the presence of censoring

Because of limited costs and time, it is often infeasible to continue the study until all subjects experience an event. Instead, the study terminates at some prespecified point in time. As a result, some subjects may not experience an event by the end of the study; that is, administrative censoring of subjects occurs. This increases the requirement on the number of subjects in the study to ensure that a certain number of events is observed.

In the presence of censoring (for a type II study), Freedman (1982) assumes the following. The analysis occurs at a fixed time, t^*, after the last patient was accrued, and all information about subject follow-up beyond time, t^*, is excluded. To minimize an overestimation of the sample size because of neglecting this information, the author suggests choosing t^* as the minimum follow-up time, f, beyond which the frequency of occurrence of events is low (the time at which, say, 85% of the total events expected are observed). Under this assumption, the number of required subjects does not depend on the rates of accrual and occurrence of events but only on the proportions of patients in the two treatment groups, s_1 and s_2, surviving after the minimum follow-up time, f. See *Including information about subject accrual* about how to compute the sample size in the presence of a long accrual.

If censoring of subjects occurs, the probability of a subject not being censored needs to be estimated to obtain an accurate estimate of the required sample size. The assumption above justifies a simple procedure, suggested by Freedman (1982) and used by default by `stpower logrank`, to compute this probability using the estimates of survival probabilities at the end of the study in the control and the experimental groups. Therefore, for a type II study (under administrative censoring), these probabilities must be supplied to `stpower logrank`.

There are three ways of providing the information about survival of subjects in two groups. The first way is to supply both survival probabilities as arguments $surv_1$ and $surv_2$. The second way is to specify the survival probability in the control group as $surv_1$ and a hazard ratio in `hratio()`. Finally, the third way is to supply survival in the control group $surv_1$ only and rely on the default `hratio(0.5)`. Below we demonstrate the first way.

▷ Example 2: Sample size in the presence of censoring

Consider an example from Machin et al. (2009, 91) of a study of patients with resectable colon cancer. The goal of the study was to compare the efficacy of the drug levamisole against a placebo with respect to relapse-free survival, using a one-sided log-rank test with a significance level of 5%. The investigators anticipated a 10% increase (from 50% to 60%, with a respective hazard ratio of 0.737) in the survival of the experimental group with respect to the survival of the control (placebo) group at the end of the study. They wanted to detect this increase with a power of 80%. To obtain the required sample size, we enter the survival probabilities 0.5 and 0.6 as arguments and specify the `onesided` option to request a one-sided test.

```
. stpower logrank 0.5 0.6, onesided

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)

Input parameters:

      alpha =    0.0500  (one sided)
         s1 =    0.5000
         s2 =    0.6000
     hratio =    0.7370
      power =    0.8000
         p1 =    0.5000

Estimated number of events and sample sizes:

          E =       270
          N =       600
         N1 =       300
         N2 =       300
```

From the above output, the investigators would have to observe a total of 270 events (relapses) to detect a 26% decrease in the hazard ($\Delta_a = 0.737$) of the experimental group relative to the hazard of the control group with a power of 80% using a one-sided log-rank test with $\alpha = 0.05$. They would have to recruit a total of 600 patients (300 per group) to observe that many events. In the absence of censoring, only 270 subjects would have been required to detect a decrease in hazard corresponding to $\Delta_a = 0.737$:

```
. stpower logrank, hratio(0.737) onesided

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)

Input parameters:

      alpha =    0.0500  (one sided)
     hratio =    0.7370
      power =    0.8000
         p1 =    0.5000

Estimated number of events and sample sizes:

          E =       270
          N =       270
         N1 =       135
         N2 =       135
```

Similarly, using the Schoenfeld method,

```
. stpower logrank 0.5 0.6, onesided schoenfeld
  (output omitted)
```

we find that 590 subjects are required in the study to observe a total of 266 events to ensure a power of a test of 80%.

◁

Although all examples demonstrated above assume equal group sizes, the information about the unequal allocation of subjects between the two groups may be provided by using `p1()` or `nratio()`.

Withdrawal of subjects from the study

Under administrative censoring, the subject is known to have experienced either of the two outcomes by the end of the study: survival or failure. Often, in practice, subjects may withdraw from the study before it terminates and therefore may not experience an event by the end of the study (or be censored), but for reasons other than administrative. Withdrawal of subjects from a study may greatly affect the estimate of the sample size and must be accounted for in the computations. Refer to [ST] **stpower** and [ST] **Glossary** for a formal definition of withdrawal.

Freedman (1982) suggests a conservative adjustment for the estimate of the sample size in the presence of withdrawal. Withdrawal is assumed to be independent of failure (event) times and administrative censoring.

The proportion of subjects anticipated to withdraw from a study may be specified by using `wdprob()`.

▷ Example 3: Withdrawal of subjects from the study

Continuing example 2, suppose that a withdrawal rate of 10% is expected in the study of colon cancer patients. To account for this, we also specify `wdprob(0.1)`:

```
. stpower logrank 0.5 0.6, onesided wdprob(0.1)

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)

Input parameters:
      alpha =    0.0500  (one sided)
         s1 =    0.5000
         s2 =    0.6000
     hratio =    0.7370
      power =    0.8000
         p1 =    0.5000
 withdrawal =     10.00%

Estimated number of events and sample sizes:
          E =       270
          N =       666
         N1 =       333
         N2 =       333
```

The estimate of the total sample size using the Freedman method increases from 600 to 666 when the withdrawal rate is assumed to be 10%. The adjustment of the estimate of the sample size for the withdrawal of subjects is conservative. It assumes equal withdrawals from each group; that is, 10% of subjects are lost by the end of the study in each group. This adjustment affects only the estimates of the sample sizes but not the number of events. The reasons for this are the following: withdrawal is assumed to be independent of event times, and the ratio of subjects surviving until the end of the study in the two groups does not change under equal withdrawals.

◁

We could use the alternative syntax and specify the survival probability in the control group, 0.5, with the value of the hazard ratio 0.737 in `hratio()` instead of supplying the two survival probabilities:

```
. stpower logrank 0.5, hratio(0.737) onesided wdprob(0.1)
```

(*output omitted*)

Including information about subject accrual

Many clinical studies have an accrual period of R, during which the subjects are recruited to the study, and a follow-up period of $f = T - R$, during which the subjects are followed up until the end of the study, T, and no new subjects enter the study. The information about the duration of an accrual and a follow-up period affects the probability of a subject failing during the study.

Freedman (1982) suggests approximating the combined event-free probability as an average of the survival probabilities in the control and the experimental groups at the minimum follow-up time, $t^* = f$ (the default approach used in `stpower logrank`). However, for a long accrual of subjects, this approach may overestimate the required number of subjects, often seriously, because it does not take into account the information about subject follow-up beyond time f. Here Freedman (1982) proposes to use the survival probabilities at the average follow-up time, defined as $t^* = (f+T)/2 = f+0.5R$, instead of the minimum follow-up time, f.

Alternatively, Schoenfeld (1983) (see also Collett [2003b, 306]) presents a formula for the required number of subjects allowing for uniform entry (accrual, recruitment) over $[0, R]$ and a follow-up period, f. This information is incorporated into the formula for the probability of a failure. The formula involves the integrals of the survivor functions of the control and the experimental groups. Schoenfeld (1983) suggests approximating the integral by using Simpson's rule, which requires the estimates of the survivor function at three specific time points, f, $0.5R + f$, and $T = R + f$. It is sufficient to provide the estimates of these three survival probabilities, $S_1(f)$, $S_1(0.5R+f)$, and $S_1(T)$, for the control group only. The corresponding survival probabilities of the experimental group are automatically computed using the value of the hazard ratio in `hratio()` and the proportional-hazards assumption.

The three estimates of the survival probabilities of the control group may be supplied by using the `simpson()` option to adjust the estimates of the sample size or power for uniform entry and a follow-up period. If the estimate of the survivor function over an array of values in the range $[f, T]$ is available from a previous study, it can be supplied using the `st1()` option to form a more accurate approximation of the probability of an event using numerical integration (see [R] **dydx**). Here the value of the length of the accrual period is needed for the computation. It is computed as the difference between the maximum and the minimum values of the time variable *varname*$_t$, supplied using `st1()`, that is, $R = T - f = \max(\textit{varname}_t) - \min(\textit{varname}_t)$.

For more information, see Cleves et al. (2010, sec. 16.2).

▷ Example 4: Sample size in the presence of accrual and follow-up periods

Consider an example described in Collett (2003b, 309) of a survival study of chronic active hepatitis. A new treatment is to be compared with a standard treatment with respect to the survival times of the patients with this disease. The investigators desire to detect a change in a hazard ratio of 0.57 with 90% power and a 5% two-sided significance level. Also subjects are to be entered into the study uniformly over a period of 18 months and then followed up for 24 months. From the Kaplan–Meier estimate of the survivor function available for the control group, the survival probabilities at $f = 24$, $0.5R + f = 33$, and $T = 42$ months are 0.70, 0.57, and 0.45, respectively.

```
. stpower logrank, hratio(0.57) power(0.9) schoenfeld simpson(0.7 0.57 0.45)
Note: probability of an event is computed using Simpson's rule with
      S1(f) = 0.70, S1(f+R/2) = 0.57, S1(T) = 0.45
      S2(f) = 0.82, S2(f+R/2) = 0.73, S2(T) = 0.63

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Schoenfeld method
Ho: S1(t) = S2(t)

Input parameters:

      alpha =    0.0500  (two sided)
 ln(hratio) =   -0.5621
      power =    0.9000
         p1 =    0.5000

Estimated number of events and sample sizes:

          E =       134
          N =       380
         N1 =       190
         N2 =       190
```

Collett (2003b, 305) reports the required number of events to be 133, which, apart from roundoff errors, agrees with our estimate of 134. In a later example, Collett (2003, 309) uses the number of events, rounded to 140, to compute the required sample size as $140/0.35 = 400$, where 0.35 is the estimate of the combined probability of an event. By hand, without rounding the number of events, we compute the required sample size as $133/0.35 = 380$ and obtain the same estimate of the total sample size as in the output.

Using the average follow-up time suggested by Freedman (1982), we obtain the following:

```
. stpower logrank 0.57, hratio(0.57) power(0.9) schoenfeld

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Schoenfeld method
Ho: S1(t) = S2(t)

Input parameters:

      alpha =    0.0500  (two sided)
         s1 =    0.5700
         s2 =    0.7259
 ln(hratio) =   -0.5621
      power =    0.9000
         p1 =    0.5000

Estimated number of events and sample sizes:

          E =       134
          N =       378
         N1 =       189
         N2 =       189
```

We specify the survival probability in the control group at $t^* = 0.5R + f = 0.5 \times 18 + 24 = 33$ as $S_1(33) = 0.57$ and the hazard ratio of 0.57 (coincidentally). The respective survival probability in the experimental group is $S_2(33) = S_1(33)^{\Delta} = 0.57^{0.57} = 0.726$. Here we obtain the estimate, 378, of the sample size, which is close to the estimate of 380 computed using the more complicated approximation. In this example, the two approximations produce similar results, but this may not always be the case.

◁

❑ Technical note

The approximation suggested by Schoenfeld (1983) and Collett (2003b) is considered to be more accurate because it takes into account information about the patient survival beyond the average follow-up time. In general, the Freedman and Schoenfeld approximations will tend to give similar results when $\{\widetilde{S}(f) + \widetilde{S}(T)\}/2 \approx \widetilde{S}(0.5R + f)$; see *Methods and formulas* for a formal definition of $\widetilde{S}(\cdot)$.

❑

If we use the survival probability in the control group, $S_1(24) = 0.7$, at a follow-up time $t^* = f = 24$ instead of the average follow-up time $t^* = 33$ in the presence of an accrual period,

```
. stpower logrank 0.7, hratio(0.57) power(0.9) schoenfeld
(output omitted)
```

we obtain the estimate of the total sample size of 550, which is significantly greater than the previously estimated sample sizes of 380 and 378.

Power and effect-size determination

Sometimes the number of subjects available for the enrollment into the study is limited. In such cases, the researchers may want to investigate with what power they can detect a desired treatment effect for a given sample size.

▷ Example 5: Using stpower logrank to compute power

Recall the colon cancer study described in example 2. Suppose that only 100 subjects are available to be recruited to the study. We find out how this affects the power to detect a hazard ratio of 0.737.

```
. stpower logrank 0.5, hratio(0.737) onesided n(100)
Estimated power for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)
Input parameters:
      alpha =    0.0500  (one sided)
         s1 =    0.5000
         s2 =    0.6000
     hratio =    0.7370
          N =       100
         p1 =    0.5000
Estimated number of events and power:
          E =        46
      power =    0.2646
```

The power to detect an alternative H_a: $\Delta = 0.737$ decreased from 0.8 to 0.2646 when the sample size decreased from 600 to 100 (the number of events decreased from 270 to 46).

◁

▷ Example 6: Using stpower logrank to compute effect size

Continuing the above example, we can find that the value of the hazard ratio that can be detected for a fixed sample size of 100 with 80% power is approximately 0.42, corresponding to an increase in survival probability from 0.5 to roughly 0.75.

```
. stpower logrank 0.5, n(100) power(0.8) onesided
Estimated hazard ratio for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)
Input parameters:
      alpha =    0.0500  (one sided)
         s1 =    0.5000
         s2 =    0.7455
          N =       100
      power =    0.8000
         p1 =    0.5000
Estimated number of events and hazard ratio:
          E =        38
     hratio =    0.4237
```

◁

▷ Example 7: Plotting power curves

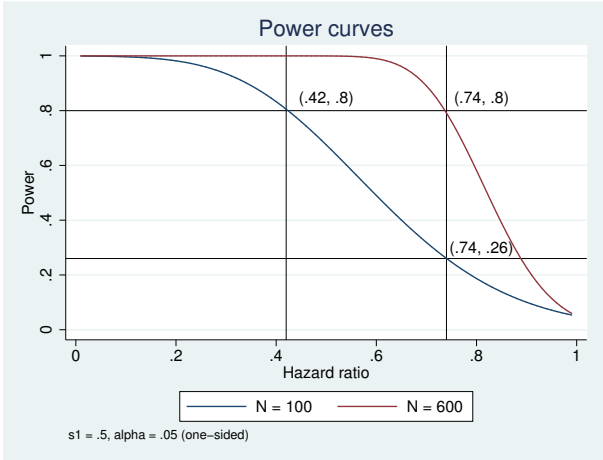
Here we demonstrate how to produce a graph of power curves over a range of hazard-ratio values. Continuing example 5, we visualize the effect of reducing the sample size from 600 to 100 on a power of the log-rank test to detect a hazard ratio of 0.737 by plotting two power curves for the sample sizes $N = 100$ and $N = 600$.

First, we generate a dataset named `mypower` containing the table values by using the `saving()` option. We request to compute the power for each of the two sample sizes over 100 values of the hazard ratio from 0.01 to 0.99 with 0.01 step size by supplying number lists 100, 600, and 0.01(0.01)0.99 to the `n()` and `hratio()` options, respectively. The values of hazard ratios, sample sizes, and powers are saved in variables `hr`, `n`, and `power`, respectively.

```
. quietly stpower logrank 0.5, hratio(0.01(0.01)0.99) n(100 600) onesided
> saving(mypower)
```

Next we generate the graph:

```
. use mypower
. twoway (line power hr if n==100) (line power hr if n==600),
>   yline( .8, lstyle(foreground) lwidth(vvthin))
>   xline(.42, lstyle(foreground) lwidth(vvthin))
>   yline(.26, lstyle(foreground) lwidth(vvthin))
>   xline(.74, lstyle(foreground) lwidth(vvthin))
>   legend(label(1 "N = 100") label(2 "N = 600"))
>   text(.85 .5 "(.42, .8)" .3 .81 "(.74, .26)" .85 .81 "(.74, .8)")
>   title("Power curves") note("s1 = .5, alpha = .05 (one sided)")
>   xtitle("Hazard ratio") ytitle("Power")
```



◁

❑ Technical note

The decrease in sample size reduces the number of events observed in the study and therefore changes the estimates of the power. If the number of events were fixed, power would have been independent of the sample size, provided that all other parameters were held constant, because the formulas relate power directly to the number of events and not the number of subjects.

❑

Examples 5 and 7 demonstrate that a significant reduction in a sample size (a number of events) greatly reduces the power of the log-rank test to detect a desired change in survival of the two groups. Indeed, we examine this further in the next section.

Performing the analysis using the log-rank test

▷ Example 8: Using the log-rank test to detect a change in survival for a fixed sample size

Continuing example 5, consider the generated dataset `drug.dta`, consisting of variables `drug`, a drug type, and `failtime`, a time to failure.

```
. use http://www.stata-press.com/data/r13/drug
(Patient Survival in Drug Trial)
. tabulate drug
```

Treatment type	Freq.	Percent	Cum.
Placebo	50	33.33	33.33
Drug A	50	33.33	66.67
Drug B	50	33.33	100.00
Total	150	100.00	

```
. by drug, sort: summarize failtime
```

```
-> drug = Placebo
```

Variable	Obs	Mean	Std. Dev.	Min	Max
failtime	50	1.03876	.5535538	.1687701	2.382302

```
-> drug = Drug A
```

Variable	Obs	Mean	Std. Dev.	Min	Max
failtime	50	1.191802	.5927507	.2366922	2.277536

```
-> drug = Drug B
```

Variable	Obs	Mean	Std. Dev.	Min	Max
failtime	50	1.717314	.8350659	.5511715	3.796102

Failure times of the control group (`Placebo`) were generated from the Weibull distribution with $\lambda_w = 0.693$ and $p = 2$ (see [ST] **streg**); failure times of the two experimental groups, `Drug A` and `Drug B`, were generated from Weibull distributions with hazard functions proportional to the hazard of the control group in ratios 0.737 and 0.42, respectively. The Weibull family of survival distributions is chosen arbitrarily, and the Weibull parameter, λ_w, is chosen such that the survival at 1 year, $t = 1$, is roughly equal to 0.5. Subjects are randomly allocated to one of the three groups in equal proportions. Subjects with failure times greater than $t = 1$ will be censored at $t = 1$.

Before analyzing these survival data, we need to set it up using `stset`. After that, we can use `sts test, logrank` to test the survivor functions separately for `Drug A` against `Placebo` and `Drug B` against `Placebo` by using the log-rank test. See [ST] **stset** and [ST] **sts test** for more information about these two commands.

```
. stset failtime, exit(time 1)
     failure event:  (assumed to fail at time=failtime)
obs. time interval:  (0, failtime]
 exit on or before:  time 1
```

```
      150  total observations
        0  exclusions
```

```
      150  observations remaining, representing
       59  failures in single-record/single-failure data
 128.9845  total analysis time at risk and under observation
                                         at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =         1
```

```
. sts test drug if drug!=2, logrank
         failure _d:  1 (meaning all fail)
   analysis time _t:  failtime
  exit on or before:  time 1
```

Log-rank test for equality of survivor functions

drug	Events observed	Events expected
Placebo	25	22.17
Drug A	21	23.83
Total	46	46.00

```
                chi2(1) =       0.70
                Pr>chi2 =     0.4028
```

```
. sts test drug if drug!=1, logrank
         failure _d:  1 (meaning all fail)
   analysis time _t:  failtime
  exit on or before:  time 1
```

Log-rank test for equality of survivor functions

drug	Events observed	Events expected
Placebo	25	16.61
Drug B	13	21.39
Total	38	38.00

```
                chi2(1) =       7.55
                Pr>chi2 =     0.0060
```

From the results from `sts test` for the `Drug A` group, we fail to reject the null hypothesis of no difference between the survivor functions in the two groups; the test made a type II error. On the other hand, for the `Drug B` group the one-sided p-value of 0.003, computed as $0.006/2 = 0.003$, suggests that the null hypothesis of nonsuperiority of the experimental treatment be rejected at the 0.005 significance level. We correctly conclude that the data provide the evidence that `Drug B` is superior to the `Placebo`.

Results from `sts test, logrank` for the two experimental groups agree with findings from examples 5 and 7. For the sample size of 100, the power of the log-rank test to detect the hazard ratio of 0.737 (10% increase in survival) is low (26%), whereas this sample size is sufficient for the test to detect a change in a hazard of 0.42 (25% increase in survival) with approximately 80% power.

Here we simulated our data from the alternative hypothesis and therefore can determine whether the correct decision or a type II error was made by the test. In practice, however, there is no way of determining the accuracy of the decision from the test. All we know is that in a long series of trials, there is a 5% chance that a particular test will incorrectly reject the null hypothesis and a 74% and a 20% chance that the test will miss the alternatives H_a: $\Delta = 0.737$ and H_a: $\Delta = 0.42$, respectively.

◁

Stored results

`stpower logrank` stores the following in `r()`:

Scalars

`r(E)`	total number of events (failures)
`r(power)`	power of test
`r(alpha)`	significance level of test
`r(hratio)`	hazard ratio
`r(onesided)`	type of test (`0` if two-sided test, `1` if one-sided test)
`r(s1)`	survival probability in the control group (if specified)
`r(s2)`	survival probability in the experimental group (if specified)
`r(p1)`	proportion of subjects in the control group
`r(w)`	proportion of withdrawals (if specified)
`r(Pr_E)`	probability of an event (failure) (when computed)

Macros

`r(method)`	type of method (`Freedman` or `Schoenfeld`)

Matrices

`r(N)`	1×3 matrix of required sample sizes

Methods and formulas

Let $S_1(t)$ and $S_2(t)$ denote the survivor functions of the control and the experimental groups and $\Delta(t) = \ln\{S_2(t)\}/\ln\{S_1(t)\}$ denote the hazard ratio at time t of the experimental to the control groups. Thus, for a given constant hazard ratio Δ, the survivor function of the experimental group at any time $t > 0$ may be computed as $S_2(t) = \{S_1(t)\}^{\Delta}$ under the assumption of proportional hazards. Define E and N to be the total number of events and the total number of subjects required for the study; w to be the proportion of subjects withdrawn from the study (lost to follow-up); $z_{(1-\alpha/k)}$ and $z_{(1-\beta)}$ to be the $(1-\alpha/k)$th and the $(1-\beta)$th quantiles of the standard normal distribution, with $k = 1$ for the one-sided test and $k = 2$ for the two-sided test. Let λ be the allocation ratio to the experimental group with respect to the control group, that is, $N_2 = \lambda N_1$. If π_1 is the proportion of subjects allocated to the control group, then $\lambda = (1-\pi_1)/\pi_1$.

The total number of events required to be observed in a study to ensure a power of $1-\beta$ of the log-rank test to detect the hazard ratio Δ with significance level α, according to Freedman (1982), is

$$E = \frac{1}{\lambda}(z_{1-\alpha/k} + z_{1-\beta})^2 \left(\frac{\lambda\Delta + 1}{\Delta - 1}\right)^2$$

and, according to Schoenfeld (1983) and Collett (2003a, 301), is

$$E = \frac{(z_{1-\alpha/k} + z_{1-\beta})^2}{\pi_1(1-\pi_1)\ln^2(\Delta)} = \frac{1}{\lambda}(z_{1-\alpha/k} + z_{1-\beta})^2 \left\{\frac{1+\lambda}{\ln(\Delta)}\right\}^2$$

Both formulas are approximations and rely on a set of assumptions such as distinct failure times, all subjects completing the course of the study (no withdrawal), and a constant ratio, λ, of subjects at risk in two groups at each failure time.

The total sample size required to observe the total number of events, E, is given by

$$N = \frac{E}{p_E}$$

The estimate of the sample size is rounded up to the nearest even integer, for an equal allocation, or rounded up to the nearest integer otherwise. The number of subjects required to be recruited in each group is obtained as $N_1 = \pi_1 N$ and $N_2 = N - N_1$, where N_1 is rounded down to the nearest integer.

By default, the probability of an event (failure), p_{E}, is approximated as suggested by Freedman (1982):

$$p_{\mathrm{E}} = 1 - \frac{S_1(t^*) + \lambda S_2(t^*)}{1 + \lambda}$$

where t^* is the minimum follow-up time, f, or, in the presence of an accrual period, the average follow-up time, $(f + T)/2 = f + 0.5R$.

If `simpson()` is specified, the probability of an event is approximated using Simpson's rule as suggested by Schoenfeld (1983):

$$p_{\mathrm{E}} = 1 - \frac{1}{6}\left\{\widetilde{S}(f) + 4\widetilde{S}(0.5R + f) + \widetilde{S}(T)\right\}$$

where $\widetilde{S}(t) = \{S_1(t) + \lambda S_2(t)\}/(1 + \lambda)$ and f, R, and $T = f + R$ are the follow-up period, the accrual period, and the total duration of the study, respectively.

The methods do not incorporate time explicitly but rather use it to determine values of the survival probabilities $S_1(t)$ and $S_2(t)$ used in the computations.

If `st1()` is used, the integral in the expression for the probability of an event

$$p_{\mathrm{E}} = 1 - \frac{1}{R}\int_f^T \widetilde{S}(t)dt$$

is computed numerically using cubic splines ([R] **dydx**). The value of R is computed as the difference between the maximum and the minimum values of *varname*$_t$ in `st1()`, $R = T - f = \max(\textit{varname}_t) - \min(\textit{varname}_t)$.

To account for the proportion of subjects, w, withdrawn from the study (lost to follow-up), a conservative adjustment to the total sample size is applied as follows:

$$N_w = \frac{N}{1 - w}$$

Equal withdrawal rates are assumed in the adjustment of the group sample sizes for the withdrawal of subjects. Equal withdrawals do not affect the estimates of the number of events, provided that withdrawal is independent of event times and the ratio of subjects at risk in two groups remains constant at each failure time.

The power for each method is estimated using the formula

$$1 - \beta = \Phi\{|\psi|^{-1}(\lambda N p_{\mathrm{E}})^{1/2} - z_{1-\alpha/k}\}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function; $\psi = (\lambda\Delta + 1)/(\Delta - 1)$ or $\psi = (1 + \lambda)/\ln(\Delta)$ if the `schoenfeld` option is specified.

The estimate of the hazard ratio (or log hazard-ratio) for fixed power and sample size is computed (iteratively for censoring) using the formulas for the sample size given above. The value of the hazard ratio (log hazard-ratio) corresponding to the reduction in a hazard of the experimental group relative to the control group is reported.

References

Cleves, M. A., W. W. Gould, R. G. Gutierrez, and Y. V. Marchenko. 2010. *An Introduction to Survival Analysis Using Stata*. 3rd ed. College Station, TX: Stata Press.

Collett, D. 2003a. *Modelling Binary Data*. 2nd ed. London: Chapman & Hall/CRC.

——. 2003b. *Modelling Survival Data in Medical Research*. 2nd ed. London: Chapman & Hall/CRC.

Freedman, L. S. 1982. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* 1: 121–129.

Hsieh, F. Y. 1992. Comparing sample size formulae for trials with unbalanced allocation using the logrank test. *Statistics in Medicine* 11: 1091–1098.

Lakatos, E., and K. K. G. Lan. 1992. A comparison of sample size methods for the logrank statistic. *Statistics in Medicine* 11: 179–191.

Machin, D., M. J. Campbell, S. B. Tan, and S. H. Tan. 2009. *Sample Size Tables for Clinical Studies*. 3rd ed. Chichester, UK: Wiley–Blackwell.

Marubini, E., and M. G. Valsecchi. 1997. *Analysing Survival Data from Clinical Trials and Observational Studies*. Chichester, UK: Wiley.

Schoenfeld, D. A. 1981. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 68: 316–319.

——. 1983. Sample-size formula for the proportional-hazards regression model. *Biometrics* 39: 499–503.

Also see [ST] **stpower** for more references.

Also see

[ST] **stpower** — Sample size, power, and effect size for survival analysis

[ST] **stpower cox** — Sample size, power, and effect size for the Cox proportional hazards model

[ST] **stpower exponential** — Sample size and power for the exponential test

[ST] **stcox** — Cox proportional hazards model

[ST] **sts test** — Test equality of survivor functions

[ST] **Glossary**

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

[R] **test** — Test linear hypotheses after estimation