Title

stata.com

strate — Tabulate failure rates and rate ratios

Syntax

Tabulate failure rates

`strate` [*varlist*] [*if*] [*in*] [, *strate_options*]

Calculate rate ratios with the Mantel–Haenszel method

`stmh` *varname* [*varlist*] [*if*] [*in*] [, *options*]

Calculate rate ratios with the Mantel–Cox method

`stmc` *varname* [*varlist*] [*if*] [*in*] [, *options*]

strate_options	Description
Main	
`per(#)`	units to be used in reported rates
`smr(`*varname*`)`	use *varname* as reference-rate variable to calculate SMRs
`cluster(`*varname*`)`	cluster variable to be used by the jackknife
`jackknife`	report jackknife confidence intervals
`missing`	include missing values as extra categories
`level(#)`	set confidence level; default is `level(95)`
`output(`*filename*[, `replace`]`)`	save summary dataset as *filename*; use `replace` to overwrite existing *filename*
`nolist`	suppress listed output
`graph`	graph rates against exposure category
`nowhisker`	omit confidence intervals from the graph
Plot	
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
cline_options	affect rendition of the plotted points
CI plot	
`ciopts(`*rspike_options*`)`	affect rendition of the confidence intervals (whiskers)
Add plots	
`addplot(`*plot*`)`	add other plots to the generated graph
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***

options	Description
Main	
`by(`*varlist*`)`	tabulate rate ratio on *varlist*
`compare(`*num1*`,`*den2*`)`	compare categories of exposure variable
`missing`	include missing values as extra categories
`level(`*#*`)`	set confidence level; default is `level(95)`

You must `stset` your data before using `strate`, `stmh`, and `stmc`; see [ST] **stset**.
`by` is allowed with `stmh` and `stmc`; see [D] **by**.
`fweight`s, `iweight`s, and `pweight`s may be specified using `stset`; see [ST] **stset**.

Menu

strate

Statistics > Survival analysis > Summary statistics, tests, and tables > Tabulate failure rates and rate ratios

stmh

Statistics > Survival analysis > Summary statistics, tests, and tables > Tabulate Mantel-Haenszel rate ratios

stmc

Statistics > Survival analysis > Summary statistics, tests, and tables > Tabulate Mantel-Cox rate ratios

Description

`strate` tabulates rates by one or more categorical variables declared in *varlist*. You can also save an optional summary dataset, which includes event counts and rate denominators, for further analysis or display. The combination of the commands `stsplit` and `strate` implements most of, if not all, the functions of the special-purpose person-years programs in widespread use in epidemiology. See Clayton and Hills (1993) and [ST] **stsplit**. If your interest is solely in calculating person-years, see [ST] **stptime**.

`stmh` calculates stratified rate ratios and significance tests by using a Mantel–Haenszel-type method.

`stmc` calculates rate ratios that are stratified finely by time by using the Mantel–Cox method. The corresponding significance test (the log-rank test) is also calculated.

Both `stmh` and `stmc` can estimate the failure-rate ratio for two categories of the explanatory variable specified by the first argument of *varlist*. You can define categories to be compared by specifying them with the `compare()` option. The remaining variables in *varlist* before the comma are categorical variables, which are to be "controlled for" using stratification. Strata are defined by cross-classification of these variables.

You can also use `stmh` and `stmc` to carry out trend tests for a metric explanatory variable. Here a one-step Newton approximation to the log-linear Poisson regression coefficient is computed.

Options for strate

Main

per(#) specifies the units to be used in reported rates. For example, if the analysis time is in years, specifying per(1000) results in rates per 1,000 person-years.

smr(*varname*) specifies a reference-rate variable. strate then calculates SMRs rather than rates. This option will usually follow stsplit to separate the follow-up records by age bands and possibly calendar periods.

cluster(*varname*) defines a categorical variable that indicates clusters of data to be used by the jackknife. If the jackknife option is selected and this option is not specified, the cluster variable is taken as the id variable defined in the st data. Specifying cluster() implies jackknife.

jackknife specifies that jackknife confidence intervals be produced. This is the default if weights were specified when the dataset was stset.

missing specifies that missing values of the explanatory variables be treated as extra categories. The default is to exclude such observations.

level(#) specifies the confidence level, as a percentage, for confidence intervals. The default is level(95) or as set by set level; see [U] **20.7 Specifying the width of confidence intervals**.

output(*filename* [, replace]) saves a summary dataset in *filename*. The file contains counts of failures and person-time, rates (or SMRs), confidence limits, and all the categorical variables in the *varlist*. This dataset could be used for further calculations or simply as input to the table command; see [R] **table**.

replace specifies that *filename* be overwritten if it exists. This option is not shown in the dialog box.

nolist suppresses the output. This is used only when saving results to a file specified by output().

graph produces a graph of the rate against the numerical code used for the categories of *varname*.

nowhisker omits the confidence intervals from the graph.

Plot

marker_options affect the rendition of markers drawn at the plotted points, including their shape, size, color, and outline; see [G-3] ***marker_options***.

marker_label_options specify if and how the markers are to be labeled; see [G-3] ***marker_label_options***.

cline_options affect whether lines connect the plotted points and the rendition of those lines; see [G-3] ***cline_options***.

CI plot

ciopts(*rspike_options*) affects the rendition of the confidence intervals (whiskers); see [G-3] ***rspike_options***.

Add plots

addplot(*plot*) provides a way to add other plots to the generated graph; see [G-3] ***addplot_option***.

Y axis, X axis, Titles, Legend, Overall

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***).

Options for stmh and stmc

Main

by(*varlist*) specifies categorical variables by which the rate ratio is to be tabulated.

A separate rate ratio is produced for each category or combination of categories of *varlist*, and a test for unequal rate ratios (effect modification) is displayed.

compare(*num1*,*den2*) specifies the categories of the exposure variable to be compared. The first code defines the numerator categories, and the second code defines the denominator categories.

When compare is absent and there are only two categories, the larger is compared with the smaller; when there are more than two categories, compare analyzes log-linear trend.

missing specifies that missing values of the explanatory variables be treated as extra categories. The default is to exclude such observations.

level(*#*) specifies the confidence level, as a percentage, for confidence intervals. The default is level(95) or as set by set level; see [U] **20.7 Specifying the width of confidence intervals**.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Tabulation of rates by using strate

strate tabulates the rate, formed from the number of failures divided by the person-time, by different levels of one or more categorical explanatory variables specified by *varlist*. Confidence intervals for the rate are also given. By default, the confidence intervals are calculated using the quadratic approximation to the Poisson log likelihood for the log-rate parameter. However, whenever the Poisson assumption is questionable, jackknife confidence intervals can also be calculated. The jackknife option also allows for multiple records for the same cluster (usually subject).

strate can also calculate and report SMRs if the data have been merged with a suitable file of reference rates.

The summary dataset can be saved to a file specified with the output() option for further analysis or more elaborate graphical display.

If weights were specified when the dataset was stset, strate calculates jackknife confidence intervals by default.

▷ Example 1

Using the diet data (Clayton and Hills 1993) described in example 1 of [ST] **stsplit**, we will use strate to tabulate age-specific coronary heart disease (CHD). In this dataset, CHD has been coded as fail = 1, 3, or 13.

We first stset the data: failure codes for CHD are specified; origin is set to date of birth, making age analysis time; and the scale is set to 365.25, so analysis time is measured in years.

```
. use http://www.stata-press.com/data/r13/diet
(Diet data with dates)

. stset dox, origin(time doe) id(id) scale(365.25) fail(fail==1 3 13)

                id:  id
     failure event:  fail == 1 3 13
obs. time interval:  (dox[_n-1], dox]
 exit on or before:  failure
    t for analysis:  (time-origin)/365.25
            origin:  time doe

------------------------------------------------------------------------------
      337  total observations
        0  exclusions
------------------------------------------------------------------------------
      337  observations remaining, representing
      337  subjects
       46  failures in single-failure-per-subject data
 4603.669  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  20.04107
```

Now we `stsplit` the data into 10-year age bands.

```
. stsplit ageband, at(40(10)70) after(time=dob) trim
(26 + 0 obs. trimmed due to lower and upper bounds)
(418 observations (episodes) created)
```

`stsplit` added 418 observations to the dataset in memory and generated a new variable, `ageband`, which identifies each observation's age group.

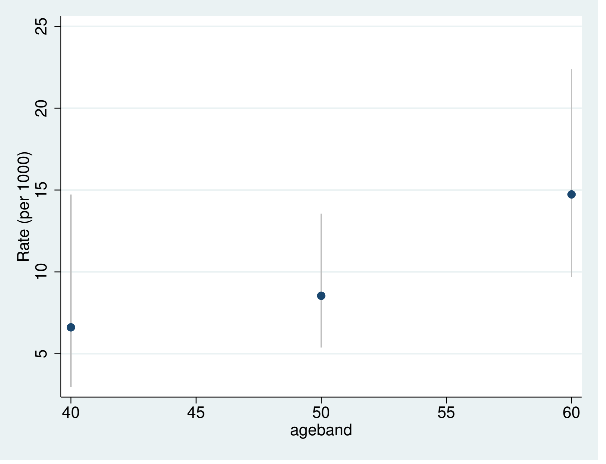
The CHD rate per 1,000 person-years can now be tabulated for categories of `ageband`:

```
. strate ageband, per(1000) graph

         failure _d:  fail == 1 3 13
   analysis time _t:  (dox-origin)/365.25
             origin:  time doe
                 id:  id
               note:  ageband<=40 trimmed

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(729 records included in the analysis)
```

ageband	D	Y	Rate	Lower	Upper
40	6	0.9070	6.6152	2.9719	14.7246
50	18	2.1070	8.5428	5.3823	13.5591
60	22	1.4933	14.7325	9.7007	22.3746



Because we specified the `graph` option, `strate` also generated a plot of the estimated rates and confidence intervals.

◁

The SMR for a cohort is the ratio of the total number of observed deaths to the number expected from age-specific reference rates. This expected number can be found by first expanding on age, using `stsplit`, and then multiplying the person-years in each age band by the reference rate for that band. `merge` (see [D] **merge**) can be used to add the reference rates to the dataset. Using the `smr` option to define the variable containing the reference rates, `strate` calculates SMRs and confidence intervals. You must specify the `per()` option. For example, if the reference rates were per 100,000, you would specify `per(100000)`. When reference rates are available by age and calendar period, you must call `stsplit` twice to expand on both time scales before merging the data with the reference-rate file.

▷ Example 2

In `smrchd.dta`, we have age-specific CHD rates per 1,000 person-years for a reference population. We can merge these data with our current data and use `strate` to obtain SMRs and confidence intervals.

```
. sort ageband
. merge m:1 ageband using http://www.stata-press.com/data/r13/smrchd
ageband was byte now float
    Result                           # of obs.
    -----------------------------------------
    not matched                            26
        from master                        26  (_merge==1)
        from using                          0  (_merge==2)
    matched                               729  (_merge==3)
    -----------------------------------------
. strate ageband, per(1000) smr(rate)
          failure _d:  fail == 1 3 13
    analysis time _t:  (dox-origin)/365.25
              origin:  time doe
                  id:  id
                note:  ageband<=40 trimmed
Estimated SMRs and lower/upper bounds of 95% confidence intervals
(729 records included in the analysis)
```

ageband	D	E	SMR	Lower	Upper
40	6	5.62	1.0670	0.4793	2.3749
50	18	18.75	0.9599	0.6048	1.5235
60	22	22.85	0.9629	0.6340	1.4624

◁

Stratified rate ratios using stmh

The `stmh` command is used for estimating rate ratios, controlled for confounding, using stratification. You can use it to estimate the ratio of the rates of failure for two categories of the explanatory variable. Categories to be compared may be defined by specifying the codes of the levels with `compare()`.

The first variable listed on the command line after `stmh` is the explanatory variable used in comparing rates, and any remaining variables, if any, are categorical variables, which are to be controlled for by using stratification.

▷ Example 3

To illustrate this command, let's return to the diet data. The variable `hienergy` is coded 1 if the total energy consumption is more than 2.75 Mcal and 0 otherwise. We want to compare the rate for `hienergy` level 1 with the rate for level 0, controlled for `ageband`.

To do this, we first `stset` and `stsplit` the data into age bands as before, and then we use `stmh`:

```
. use http://www.stata-press.com/data/r13/diet, clear
(Diet data with dates)
. stset dox, origin(time dob) enter(time doe) id(id) scale(365.25)
> fail(fail==1 3 13)
  (output omitted)
. stsplit ageband, at(40(10)70) after(time=dob) trim
(26 + 0 obs. trimmed due to lower and upper bounds)
(418 observations (episodes) created)
```

```
. stmh hienergy, by(ageband)
        failure _d:  fail == 1 3 13
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
 enter on or after:  time doe
                id:  id
              note:  ageband<=40 trimmed
Maximum likelihood estimate of the rate ratio
  comparing hienergy==1 vs. hienergy==0
  by ageband
RR estimate, and lower and upper 95% confidence limits
```

ageband	RR	Lower	Upper
40	1.24	0.23	6.76
50	0.43	0.16	1.16
60	0.50	0.21	1.20

```
Overall estimate controlling for ageband
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.534	4.36	0.0369	0.293	0.972

```
Approx test for unequal RRs (effect modification): chi2(2) =      1.19
                                                   Pr>chi2 =    0.5514
```

Because the RR estimates are approximate, the test for unequal rate ratios is also approximate.

We can also compare the effect of `hienergy` between jobs, controlling for `ageband`.

```
. stmh hienergy ageband, by(job)
        failure _d:  fail == 1 3 13
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
 enter on or after:  time doe
                id:  id
              note:  ageband<=40 trimmed
Mantel-Haenszel estimate of the rate ratio
  comparing hienergy==1 vs. hienergy==0
  controlling for ageband
  by job
RR estimate, and lower and upper 95% confidence limits
```

job	RR	Lower	Upper
0	0.42	0.13	1.33
1	0.64	0.22	1.87
2	0.51	0.21	1.26

```
Overall estimate controlling for ageband job
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.521	4.88	0.0271	0.289	0.939

```
Approx test for unequal RRs (effect modification): chi2(2) =      0.28
                                                   Pr>chi2 =    0.8695
```

◁

Log-linear trend test for metric explanatory variables using stmh

stmh may also be used to carry out trend tests for a metric explanatory variable. A one-step Newton approximation to the log-linear Poisson regression coefficient is also computed.

The diet dataset contains the height for each patient recorded in the variable height. We can test for a trend of heart disease rates with height controlling for ageband by typing

```
. stmh height ageband

         failure _d:  fail == 1 3 13
   analysis time _t:  (dox-origin)/365.25
             origin:  time dob
  enter on or after:  time doe
                 id:  id
               note:  ageband<=40 trimmed

Score test for trend of rates with height
  with an approximate estimate of the
  rate ratio for a one unit increase in height
  controlling for ageband

RR estimate, and lower and upper 95% confidence limits
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.906	18.60	0.0000	0.866	0.948

stmh tested for trend of heart disease rates with height within age bands and provided a rough estimate of the rate ratio for a 1-cm increase in height—this estimate is a one-step Newton approximation to the maximum likelihood estimate. It is not consistent, but it does provide a useful indication of the size of the effect.

The rate ratio is significantly less than 1, so there is clear evidence for a decreasing rate with increasing height (about 9% decrease in rate per centimeter increase in height).

Controlling for age with fine strata by using stmc

The stmc (Mantel–Cox) command is used to control for variation of rates on a time scale by breaking up time into short intervals, or *clicks*.

Usually this approach is used only to calculate significance tests, but the rate ratio estimated remains just as useful as in the coarsely stratified analysis from stmh. The method may be viewed as an approximate form of Cox regression.

The rate ratio produced by stmc is controlled for analysis time separately for each level of the variables specified with by() and then combined to give a rate ratio controlled for both time and the by() variables.

▷ Example 4

For example, to obtain the effect of high energy controlled for age by stratifying finely, we first stset the data specifying the date of birth, dob, as the origin (so analysis time is age), and then we use stmc:

```
. stset dox, origin(time dob) enter(time doe) id(id) scale(365.25)
> fail(fail==1 3 13)
  (output omitted)
```

```
. stmc hienergy
         failure _d:  fail == 1 3 13
   analysis time _t:  (dox-origin)/365.25
             origin:  time dob
  enter on or after:  time doe
                 id:  id
Mantel-Cox comparisons
Mantel-Haenszel estimates of the rate ratio
   comparing hienergy==1 vs. hienergy==0
   controlling for time (by clicks)
Overall Mantel-Haenszel estimate, controlling for time from dob
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.537	4.20	0.0403	0.293	0.982

The rate ratio of 0.537 is close to that obtained with stmh when controlling for age by using 10-year age bands.

◁

Stored results

stmh and stmc store the following in r():

Scalars
r(RR) overall rate ratio

Nathan Mantel (1919–2002) was an American biostatistician who grew up in New York. He worked at the National Cancer Institute from 1947 to 1974 on a wide range of medical problems and was also later affiliated with George Washington University and the American University in Washington.

William M. Haenszel (1910–1998) was an American biostatistician and epidemiologist who graduated from the University of Buffalo. He also worked at the National Cancer Institute and later at the University of Illinois.

Acknowledgments

The original versions of strate, stmh, and stmc were written by David Clayton of the Cambridge Institute for Medical Research and Michael Hills (retired) of the London School of Hygiene and Tropical Medicine.

References

Clayton, D. G., and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

———. 1995. ssa7: Analysis of follow-up studies. *Stata Technical Bulletin* 27: 19–26. Reprinted in *Stata Technical Bulletin Reprints*, vol. 5, pp. 219–227. College Station, TX: Stata Press.

——. 1997. ssa10: Analysis of follow-up studies with Stata 5.0. *Stata Technical Bulletin* 40: 27–39. Reprinted in *Stata Technical Bulletin Reprints*, vol. 7, pp. 253–268. College Station, TX: Stata Press.

Gail, M. H. 1997. A conversation with Nathan Mantel. *Statistical Science* 12: 88–97.

Hankey, B. 1997. A conversation with William M. Haenszel. *Statistical Science* 12: 108–112.

Also see

[ST] **stci** — Confidence intervals for means and percentiles of survival time

[ST] **stir** — Report incidence-rate comparison

[ST] **stptime** — Calculate person-time, incidence rates, and SMR

[ST] **stset** — Declare data to be survival-time data