The Stata Journal (2022)
22, Number 4, pp. 969–974

DOI: 10.1177/1536867X221141057

Estimating the risk of events with stprisk

Matteo Bottai
Division of Biostatistics
Institute of Environmental Medicine
Karolinska Institutet
Stockholm, Sweden
matteo.bottai@ki.se

Abstract. Incidence rates are popular summary measures of the occurrence over time of events of interest. They are also called mortality rates or failure rates, depending on the context. The incidence rate is defined as the ratio between the total number of events and total follow-up time and can be estimated with the `strate` command. When the event of interest can occur multiple times on any given subject over a time period, like infections, the incidence rate represents an average count per unit of time, such as the average number of infections per year. When the event of interest can occur only once, such as death, an alternative summary measure is the risk, or probability, of occurrence per unit time, such as the risk of dying in one year. In this article, I present the `stprisk` command, which estimates risks, and illustrate its use and interpretation through a data example.

Keywords: st0698, stprisk, incidence rates, mortality rates, survival analysis

1 Introduction

Scientific research often entails analyzing the occurrence over time of events of interest, such as death or a cancer diagnosis. This type of analysis is generally known as survival analysis, although different terms are used across different research areas. Stata has a comprehensive suite of commands for the analysis of survival data, and the help documentation offers an excellent description of the relevant methods and available commands (`help st`).

The `strate` command can estimate incidence rates, which are popular summary measures of the occurrence over time of events of interest. The incidence rate is defined as the ratio between the total number of events and total follow-up time. When the event of interest can occur multiple times on any given subject over a time period, like infections, the incidence rate represents an average count per unit of time, such as average number of infections per year. When the event of interest can occur only once, such as death, an alternative summary measure is the risk, or probability, of occurrence per unit of time, such as the risk of dying in one year.

The following section introduces the new `stprisk` command, which estimates the risk of occurrence of events of interest over time, and its relation to the popular incidence rate through a data example; section 3 shows the syntax of `stprisk`; section 4 provides some technical details about risks and the implementation of the `stprisk` command; and section 5 contains some final remarks.

2 Incidence rates and incidence risks

I illustrate the basis of the new `stprisk` command with an example. I use the data from a fictitious clinical trial on survival in cancer patients available in Stata.

```
. sysuse cancer
(Patient survival in drug trial)
```

I summarize the content of the dataset with the `describe` command.

```
. describe
Contains data from /usr/local/stata17/ado/base/c/cancer.dta
 Observations:            48                  Patient survival in drug trial
    Variables:             8                  3 Mar 2020 16:09
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
studytime       byte    %8.0g                 Months to death or end of exp.
died            byte    %8.0g      diedlbl    Patient died
drug            byte    %8.0g      type       Drug type
age             byte    %8.0g                 Patient's age at start of exp.
_st             byte    %8.0g                 1 if record is to be used; 0
                                                otherwise
_d              byte    %8.0g                 1 if failure; 0 if censored
_t              byte    %10.0g                Analysis time when record ends
_t0             byte    %10.0g                Analysis time when record begins
-------------------------------------------------------------------------------
Sorted by:
```

The interest of this study is in comparing survival with the three different treatment groups, which comprise two active drugs and a placebo. To help present the arguments contained in the remainder of this section, I set the unit of measurement of the time-to-death variable to be years with the `stset` command.

```
. stset studytime, failure(died) scale(12)
Survival-time data settings
         Failure event: died!=0 & died<.
Observed time interval: (0, studytime]
     Exit on or before: failure
     Time for analysis: time/12
--------------------------------------------------------------------------
         48  total observations
          0  exclusions
--------------------------------------------------------------------------
         48  observations remaining, representing
         31  failures in single-record/single-failure data
         62  total analysis time at risk and under observation
                                             At risk from t =         0
                                  Earliest observed entry t =         0
                                       Last observed exit t =      3.25
```

Henceforth, I refer to the incidence rate as mortality rate, considering the event of interest is death. I estimate the mortality rate by treatment group with the `strate` command.

```
. strate drug
        Failure _d: died
  Analysis time _t: studytime/12
Estimated failure rates
Number of records = 48
```

drug	D	Y	Rate	Lower	Upper
Placebo	19	15.0000	1.266667	0.807948	1.985827
Other	6	17.4167	0.344498	0.154769	0.766810
NA	6	29.5833	0.202817	0.091118	0.451446

```
Notes: Rate = D/Y = failures/person-time.
       Lower and Upper are bounds of 95% confidence intervals.
```

From the above output, the mortality rate in the placebo group is 1.27. On its website, the Centers for Disease Control and Prevention defines the incidence rate as "a measure of the frequency with which new cases of illness, injury, or other health condition occur, expressed explicitly per a time frame [...]." A rate is not a risk. If one interpreted the above rate naïvely as a risk, one might conclude that any given patient is expected to die 1.27 times every year or, alternatively, that 100 patients are expected to report 127 deaths every year.

I now estimate the mortality risk with the `stprisk` command. Section 4 gives more details on its definition and interpretation.

```
. stprisk drug
    Incidence risk
```

drug	Risk	[95% Conf.	Int.]
Placebo	0.830235	0.578028	0.973872
Other	0.381161	0.163559	0.724608
NA	0.206644	0.090824	0.430367

The interpretation of the above risks is simple. For example, in the placebo group, the probability for any given subject to die in a year is 0.83, or alternatively, we expect 83 deaths out of 100 subjects every year.

The mortality rate converges to the mortality risk as the latter tends to zero. This limit behavior is analogous to that of the mean of a Poisson distribution, which converges to the mean of the binomial distribution as the latter tends to zero. This explains why the mortality rate and the mortality risk are numerically closer in drug group 3 than they are in the placebo group 1.

3 The stprisk command

The syntax of the `stprisk` command, similar to that of `strate`, is

`stprisk` [*varlist*] [*if*] [*in*] [, `level(#) graph nowhisker`]

`stprisk` tabulates rates by one or more categorical variables declared in *varlist*. When *varlist* is omitted, `stprisk` estimates the mortality risk for the entire dataset. The `level()` option specifies the level of the confidence intervals. The default is `level(95)` or as set by `set level`. The `graph` option plots rates against the groups defined by *varlist* when a *varlist* is specified. The `nowhisker` option omits the confidence intervals from the graph.

You must `stset` your data before using `stprisk`; see [ST] **stset**.

4 Incidence risks

This section provides the basic definition and interpretation of incidence risks. Its content consists of slightly edited excerpts from published articles (Bottai 2017; Discacciati and Bottai 2017; Lagergren, Bottai, and Santoni 2021; and Bottai, Discacciati, and Santoni 2021). Let T represent a continuous time-to-event variable with support on the positive real half-line. Let $S(t) = \Pr(T > t)$ and $H(t) = -\log\{S(t)\}$ indicate the survival function and the cumulative hazard function, respectively. The function $S(t)$ is defined over the entire real line, $\mathbb{R}$, while $H(t)$ is defined over the set $\{t \in \mathbb{R} : S(t) > 0\}$.

The probability of occurrence of an event over the time interval $[t_0, t_1]$, with $t_0 < t_1$, conditional on $T > t_0$, is

$$\Pr(T \leq t_1 \mid T > t_0) = 1 - S(t_1)/S(t_0)$$

defined over the set $\{t \in \mathbb{R} : S(t) > 0\}$. Bottai (2017) defined the geometric rate of the event over the interval $[t_0, t_1]$, conditional on $T > t_0$, as

$$G(t_0, t_1) = 1 - \{S(t_1)/S(t_0)\}^{1/(t_1 - t_0)} \tag{1}$$

Bottai, Discacciati, and Santoni (2021) later referred to the above as the average probability of occurrence of the event. As explained in the articles, the word average indicates the geometric mean. The following example may help interpret $G(t_0, t_1)$. Suppose the event of interest is death and $t_0 = 0$ and $t_1 = 3$. We split the time interval $[0, 3]$ into the three one-unit disjoint intervals $[0, 1]$, $[1, 2]$, and $[2, 3]$. The mortality rate is the complement of the probability of surviving all three intervals conditional on being alive at time 0, which is $S(3)/S(0)$. This is algebraically equal to the product of the probabilities of surviving each interval conditional on being alive at its start, $S(3)/S(0) = \{S(1)/S(0)\}\{S(2)/S(1)\}\{S(3)/S(2)\}$. The average probability per unit of time, therefore, is the geometric mean of the probabilities in the three intervals, $\{S(3)/S(0)\}^{1/3}$. Applying the geometric mean at each interval yields the probability of surviving the entire period $[0, 3]$.

For example, the probability of surviving one year in the placebo group is 0.225, as evinced in the following output:

```
. sts list if drug==1, at(1)
        Failure _d: died
  Analysis time _t: studytime/12
Kaplan--Meier survivor function
              Beg.              Survivor      Std.
    Time      total      Fail    function     error     [95% conf. int.]
-------------------------------------------------------------------------
       0          0         0      1.0000         .          .          .
       1          6        15      0.2250    0.0971     0.0721     0.4290
       2          1         4           .         .          .          .
-------------------------------------------------------------------------
Note: Survivor function is calculated over full data and evaluated at
      indicated times; it is not calculated from aggregates shown at left.
```

From (1), the monthly mortality risk over the first year is $G(0, 12) = 1 - 0.225^{1/12} = 0.117$.

The definition of incidence risk given in (1) can also be written as

$$G(t_0, t_1) = 1 - \exp\left\{\frac{H(t_0) - H(t_1)}{t_1 - t_0}\right\} \tag{2}$$

The `stprisk` command uses (2), with t_0 set equal to the start of the follow-up time and t_1 set equal to the largest observed time. The cumulative hazard $H(t_0)$ is equal to zero. The cumulative hazard $H(t_1)$ and its confidence interval are obtained from the `sts list` command with the Nelson–Aalen option; see [ST] **sts list**.

5 Conclusions

The new `stprisk` command provides nonparametric estimates and confidence intervals for the risk of occurrence of an event of interest over time. The command is computationally efficient, and its syntax is patterned on that of the `strate` command.

The risk of occurrence of events is applicable to any event that can occur only once. In the example given in section 2, the event of interest is death, but risks can be assessed for other once-only events, such as first cancer diagnosis, hospital discharge, and first employment.

As the time period (t_0, t_1) tends to zero, the risk $G(t_0, t_1)$ defined in (1) tends to the instantaneous risk, and the incidence rate tends to the instantaneous incidence rate, which is also known as the hazard. The similarities and differences between these quantities are expounded in the article by Bottai, Discacciati, and Santoni (2021).

6 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 22-4
. net install st0698     (to install program files, if available)
. net get st0698         (to install ancillary files, if available)
```

7 References

Bottai, M. 2017. A regression method for modelling geometric rates. *Statistical Methods in Medical Research* 26: 2700–2707. https://doi.org/10.1177/0962280215606474.

Bottai, M., A. Discacciati, and G. Santoni. 2021. Modeling the probability of occurrence of events. *Statistical Methods in Medical Research* 30: 1976–1987. https://doi.org/10.1177/09622802211022403.

Discacciati, A., and M. Bottai. 2017. Instantaneous geometric rates via generalized linear models. *Stata Journal* 17: 358–371. https://doi.org/10.1177/1536867X1701700207.

Lagergren, J., M. Bottai, and G. Santoni. 2021. Patient age and survival after surgery for esophageal cancer. *Annals of Surgical Oncology* 28: 159–166. https://doi.org/10.1245/s10434-020-08653-w.

About the author

Matteo Bottai is a professor of biostatistics at the Division of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.