Glossary

accelerated failure-time model. A model in which everyone has, in a sense, the same survivor function, $S(\tau)$, and an individual's τ_j is a function of his or her characteristics and of time, such as $\tau_j = t * \exp(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j})$.

AFT, accelerated failure time. See *accelerated failure-time model*.

analysis time. Analysis time is like time, except that 0 has a special meaning: $t = 0$ is the time of onset of risk, the time when failure first became possible.

Analysis time is usually not what is recorded in a dataset. A dataset of patients might record calendar time. Calendar time must then be mapped to analysis time.

The letter t is reserved for time in analysis-time units. The term *time* is used for time measured in other units.

The *origin* is the *time* corresponding to $t = 0$, which can vary subject to subject. Thus $t = \textit{time} - \textit{origin}$.

at risk. A subject is at risk from the instant the first failure event becomes possible and usually stays that way until failure, but a subject can have periods of being at risk and not at risk.

attributable fraction. An attributable fraction is the reduction in the risk of a disease or other condition of interest when a particular risk factor is removed.

baseline. In survival analysis, baseline is the state at which the covariates, usually denoted by the row vector $\mathbf{x}$, are zero. For example, if the only measured covariate is systolic blood pressure, the baseline survivor function would be the survivor function for someone with zero systolic blood pressure. This may seem ridiculous, but covariates are usually centered so that the mathematical definition of baseline (covariate is zero) translates into something meaningful (mean systolic blood pressure).

boundary kernel. A boundary kernel is a special kernel used to smooth hazard functions in the boundaries of the data range. Boundary kernels are applied when the `epan2`, `biweight`, or `rectangle` `kernel()` is specified with `stcurve, hazard` or `sts graph, hazard`.

cause-specific hazard. In a competing-risks analysis, the cause-specific hazard is the hazard function that generates the events of a given type. For example, if heart attack and stroke are competing events, then the cause-specific hazard for heart attacks describes the biological mechanism behind heart attacks independently of that for strokes. Cause-specific hazards can be modeled using Cox regression, treating the other events as censored.

censored, **censoring**, **left-censoring**, and **right-censoring**. An observation is left-censored when the exact time of failure is not known; it is merely known that the failure occurred before t_l. Suppose that the event of interest is becoming employed. If a subject is already employed when first interviewed, his outcome is left-censored.

An observation is right-censored when the time of failure is not known; it is merely known that the failure occurred after t_r. If a patient survives until the end of a study, the patient's time of death is right-censored.

In common usage, *censored* without a modifier means right-censoring.

Also see *truncation, left-truncation, and right-truncation*.

CIF. See *cumulative incidence function*.

competing risks. Competing risks models are survival-data models in which the failures are generated by more than one underlying process. For example, death may be caused by either heart attack or stroke. There are various methods for dealing with competing risks. One direct way is to duplicate failures for one competing risk as censored observations for the other risk and stratify on the risk type. Another is to directly model the cumulative incidence of the event of interest in the presence of competing risks. The former method uses `stcox` and the latter, `stcrreg`.

confounding. In the analysis of contingency tables, factor or interaction effects are said to be confounded when the effect of one factor is combined with that of another. For example, the effect of alcohol consumption on esophageal cancer may be confounded with the effects of age, smoking, or both. In the presence of confounding, it is often useful to stratify on the confounded factors that are not of primary interest, in the above example, age and smoking.

count-time data. See *ct data*.

covariates. Covariates are the explanatory variables that appear in a model. For instance, if survival time were to be explained by age, sex, and treatment, then those variables would be the covariates. Also see *time-varying covariates*.

crude estimate. A crude estimate has not been adjusted for the effects of other variables. Disregarding a stratification variable, for example, yields a crude estimate.

ct data. ct stands for count time. ct data are an aggregate organized like a life table. Each observation records a time, the number known to fail at that time, the number censored, and the number of new entries. See [ST] **ctset**.

cumulative hazard. See *hazard, cumulative hazard, and hazard ratio*.

cumulative incidence estimator. In a competing-risks analysis, the cumulative incidence estimator estimates the cumulative incidence function (CIF). Assume for now that you have one event of interest (type 1) and one competing event (type 2). The cumulative incidence estimator for type 1 failures is then obtained by

$$\widehat{\text{CIF}}_1(t) = \sum_{j:t_j \le t} \widehat{h}_1(t_j)\widehat{S}(t_{j-1})$$

with

$$\widehat{S}(t) = \prod_{j:t_j \le t} \left\{1 - \widehat{h}_1(t_j) - \widehat{h}_2(t_j)\right\}$$

The t_j index the times at which events (of any type) occur, and $\widehat{h}_1(t_j)$ and $\widehat{h}_2(t_j)$ are the cause-specific hazard contributions for type 1 and type 2, respectively. $\widehat{S}(t)$ estimates the probability that you are event free at time t.

The above generalizes to multiple competing events in the obvious way.

cumulative incidence function. In a competing-risks analysis, the cumulative incidence function, or CIF, is the probability that you will observe the event of primary interest before a given time. Formally,

$$\text{CIF}(t) = P(T \le t \text{ and event type of interest})$$

for time-to-failure, T.

cumulative subhazard. See *subhazard, cumulative subhazard, and subhazard ratio*.

DFBETA. A DFBETA measures the change in the regressor's coefficient because of deletion of that subject. Also see *partial DFBETA*.

effect size. The effect size is the size of the clinically significant difference between the treatments being compared, often expressed as the hazard ratio (or the log of the hazard ratio) in survival analysis.

event. An event is something that happens at an instant in time, such as being exposed to an environmental hazard, being diagnosed as myopic, or becoming employed.

The failure event is of special interest in survival analysis, but there are other equally important events, such as the exposure event, from which analysis time is defined.

In st data, events occur at the end of the recorded time span.

event of interest. In a competing-risks analysis, the event of interest is the event that is the focus of the analysis, that for which the cumulative incidence in the presence of competing risks is estimated.

failure event. Survival analysis is really time-to-failure analysis, and the failure event is the event under analysis. The failure event can be death, heart attack, myopia, or finding employment. Many authors—including Stata—write as if the failure event can occur only once per subject, but when we do, we are being sloppy. Survival analysis encompasses repeated failures, and all of Stata's survival analysis features can be used with repeated-failure data.

frailty. In survival analysis, it is often assumed that subjects are alike—homogeneous—except for their observed differences. The probability that subject j fails at time t may be a function of j's covariates and random chance. Subjects j and k, if they have equal covariate values, are equally likely to fail.

Frailty relaxes that assumption. The probability that subject j fails at time t becomes a function of j's covariates and j's unobserved frailty value, ν_j. Frailty ν is assumed to be a random variable. Parametric survival models can be fit even in the presence of such heterogeneity.

Shared frailty refers to the case in which groups of subjects share the same frailty value. For instance, subjects 1 and 2 may share frailty value ν because they are genetically related. Both parametric and semiparametric models can be fit under the shared-frailty assumption.

future history. Future history is information recorded after a subject is no longer at risk. The word *history* is often dropped, and the term becomes simply *future*. Perhaps the failure event is cardiac infarction, and you want to know whether the subject died soon in the *future*, in which case you might exclude the subject from analysis.

Also see *past history*.

gaps. Gaps refers to gaps in observation between entry time and exit time; see *under observation*.

hazard, **cumulative hazard**, and **hazard ratio**. The hazard or hazard rate at time t, $h(t)$, is the instantaneous rate of failure at time t conditional on survival until time t. Hazard rates can exceed 1. Say that the hazard rate were 3. If an individual faced a constant hazard of 3 over a unit interval and if the failure event could be repeated, the individual would be expected to experience three failures during the time span.

The cumulative hazard, $H(t)$, is the integral of the hazard function $h(t)$, from 0 (the onset of risk) to t. It is the total number of failures that would be expected to occur up until time t, if the failure event could be repeated. The relationship between the cumulative hazard function, $H(t)$, and the survivor function, $S(t)$, is

$$S(t) = \exp\{-H(t)\}$$
$$H(t) = -\ln\{S(t)\}$$

The hazard ratio is the ratio of the hazard function evaluated at two different values of the covariates: $h(t\,|\,\mathbf{x})/h(t\,|\,\mathbf{x}_0)$. The hazard ratio is often called the relative hazard, especially when $h(t\,|\,\mathbf{x}_0)$ is the baseline hazard function.

hazard contributions. Hazard contributions are the increments of the estimated cumulative hazard function obtained through either a nonparametric or semiparametric analysis. For these analysis types, the estimated cumulative hazard is a step function that increases every time a failure occurs. The hazard contribution for that time is the magnitude of that increase.

Because the time between failures usually varies from failure to failure, hazard contributions do not directly estimate the hazard. However, one can use the hazard contributions to formulate an estimate of the hazard function based on the method of smoothing.

ID variable. An ID variable identifies groups; equal values of an ID variable indicate that the observations are for the same group. For instance, a stratification ID variable would indicate the strata to which each observation belongs.

When an ID variable is referred to without modification, it means subjects, and usually this occurs in multiple-record st data. In multiple-record data, each physical observation in the dataset represents a time span, and the ID variable ties the separate observations together:

idvar	*t*0	*t*
1	0	5
1	5	7

ID variables are usually numbered 1, 2, ..., but that is not required. An ID variable might be numbered 1, 3, 7, 22, ..., or -5, -4, ..., or even 1, 1.1, 1.2, ...

incidence and **incidence rate**. Incidence is the number of new failures (for example, number of new cases of a disease) that occur during a specified period in a population at risk (for example, of the disease).

Incidence rate is incidence divided by the sum of the length of time each individual was exposed to the risk.

Do not confuse incidence with prevalence. Prevalence is the fraction of a population that has the disease. Incidence refers to the rate at which people contract a disease, whereas prevalence is the total number actually sick at a given time.

Kaplan–Meier product-limit estimate. This is an estimate of the survivor function, which is the product of conditional survival to each time at which an event occurs. The simple form of the calculation, which requires tallying the number at risk and the number who die and at each time, makes accounting for censoring easy. The resulting estimate is a step function with jumps at the event times.

left-censoring. See *censored, censoring, left-censoring, and right-censoring*.

left-truncation. See *truncation, left-truncation, and right-truncation*.

life table. Also known as a mortality table or actuarial table, a life table is a table that shows for each analysis time the fraction that survive to that time. In mortality tables, analysis time is often age.

likelihood displacement value. A likelihood displacement value is an influence measure of the effect of deleting a subject on the overall coefficient vector. Also see *partial likelihood displacement value*.

LMAX value. An LMAX value is an influence measure of the effect of deleting a subject on the overall coefficient vector and is based on an eigensystem analysis of efficient score residuals. Also see *partial LMAX value*.

multiarm trial. A multiarm trial is a trial comparing survivor functions of more than two groups.

multiple-record st data. See *st data*.

odds and **odds ratio**. The odds in favor of an event are $o = p/(1-p)$, where p is the probability of the event. Thus if $p = 0.2$, the odds are 0.25, and if $p = 0.8$, the odds are 4.

The log of the odds is $\ln(o) = \text{logit}(p) = \ln\{p/(1-p)\}$, and logistic-regression models, for instance, fit $\ln(o)$ as a linear function of the covariates.

The odds ratio is a ratio of two odds: o_1/o_0. The individual odds that appear in the ratio are usually for an experimental group and a control group, or two different demographic groups.

offset variable and **exposure variable**. An offset variable is a variable that is to appear on the right-hand side of a model with coefficient 1:

$$y_j = \text{offset}_j + b_0 + b_1 x_j + \cdots$$

In the above, b_0 and b_1 are to be estimated. The offset is not constant. Offset variables are often included to account for the amount of exposure. Consider a model where the number of events observed over a period is the length of the period multiplied by the number of events expected in a unit of time:

$$n_j = T_j e(X_j)$$

When we take logs, this becomes

$$\log(n_j) = \log(T_j) + \log\{e(X_j)\}$$

$\ln(T_j)$ is an offset variable in this model.

When the log of a variable is an offset variable, the variable is said to be an exposure variable. In the above, T_j is an exposure variable.

partial DFBETA. A partial DFBETA measures the change in the regressor's coefficient because of deletion of that individual record. In single-record data, the partial DFBETA is equal to the DFBETA. Also see *DFBETA*.

partial likelihood displacement value. A partial likelihood displacement value is an influence measure of the effect of deleting an individual record on the coefficient vector. For single-record data, the partial likelihood displacement value is equal to the likelihood displacement value. Also see *likelihood displacement value*.

partial LMAX value. A partial LMAX value is an influence measure of the effect of deleting an individual record on the overall coefficient vector and is based on an eigensystem analysis of efficient score residuals. In single-record data, the partial LMAX value is equal to the LMAX value. Also see *LMAX value*.

past history. Past history is information recorded about a subject before the subject was both *at risk* and *under observation*. Consider a dataset that contains information on subjects from birth to death and an analysis in which subjects became at risk once diagnosed with a particular kind of cancer. The past history on the subject would then refer to records before the subjects were diagnosed.

The word *history* is often dropped, and the term becomes simply *past*. For instance, we might want to know whether a subject smoked in the past.

Also see *future history*.

penalized log-likelihood function. This is a log-likelihood function that contains an added term, usually referred to as a roughness penalty, that reduces its value when the model overfits the data. In Cox models with frailty, such functions are used to prevent the variance of the frailty from growing too large, which would allow the individual frailty values to perfectly fit the data.

power. The power of a test is the probability of correctly rejecting the null hypothesis when it is false. It is often denoted as $1 - \beta$ in statistical literature, where β is the type II error probability. Commonly used values for power are 80% and 90%. Also see *type I error* and *type II error*.

proportional hazards model. This is a model in which, between individuals, the ratio of the instantaneous failure rates (the hazards) is constant over time.

right-censoring. See *censored, censoring, left-censoring, and right-censoring*.

right-truncation. See *truncation, left-truncation, and right-truncation*.

risk factor. This is a variable associated with an increased or decreased risk of failure.

risk pool. At a particular point in time, this is the subjects at risk of failure.

semiparametric model. This is a model that is not fully parameterized. The Cox proportional hazards model is such a model:

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \cdots + \beta_k x_k)$$

In the Cox model, $h_o(t)$ is left unparameterized and not even estimated. Meanwhile, the relative effects of covariates are parameterized as $\exp(\beta_1 x_1 + \cdots + \beta_k x_k)$.

shape parameter. A shape parameter governs the shape of a probability distribution. One example is the parameter p of the Weibull model.

single-record st data. See *st data*.

singleton-group data. A singleton is a frailty group that contains only 1 observation. A dataset containing only singletons is known as singleton-group data.

SMR. See *standardized mortality (morbidity) ratio*.

snapshot data. Snapshot data are those in which each record contains the values of a set of variables for a subject at an instant in time. The name arises because each observation is like a snapshot of the subject.

In snapshot datasets, one usually has a group of observations (snapshots) for each subject.

Snapshot data must be converted to st data before they can be analyzed. This requires making assumptions about what happened between the snapshots. See [ST] **snapspan**.

spell data. Spell data are survival data in which each record represents a fixed period, consisting of a begin time, an end time, possibly a censoring/failure indicator, and other measurements (covariates) taken during that specific period.

st data. st stands for survival time. In survival-time data, each observation represents a span of survival, recorded in variables $t0$ and t. For instance, if in an observation $t0$ were 3 and t were 5, the span would be $(t0, t]$, meaning from just after $t0$ up to and including t.

Sometimes variable $t0$ is not recorded; $t0$ is then assumed to be 0. In such a dataset, an observation that had $t = 5$ would record the span $(0, 5]$.

Each observation also includes a variable d, called the failure variable, which contains 0 or nonzero (typically, 1). The failure variable records what happened at the end of the span: 0, the subject was still alive (had not yet failed) or 1, the subject died (failed).

Sometimes variable d is not recorded; d is then assumed to be 1. In such a dataset, all time-span observations would be assumed to end in failure.

Finally, each observation in an st dataset can record the entire history of a subject or each can record a part of the history. In the latter case, groups of observations record the full history. One observation might record the period $(0, 5]$ and the next, $(5, 8]$. In such cases, there is a variable ID that records the subject for which the observation records a time span. Such data are called multiple-record st data. When each observation records the entire history of a subject, the data are called single-record st data. In the single-record case, the ID variable is optional.

See [ST] **stset**.

standardized mortality (morbidity) ratio. Standardized mortality (morbidity) ratio (SMR) is the observed number of deaths divided by the expected number of deaths. It is calculated using indirect standardization: you take the population of the group of interest—say, by age, sex, and other factors—and calculate the expected number of deaths in each cell (expected being defined as the number of deaths that would have been observed if those in the cell had the same mortality as some other population). You then take the ratio to compare the observed with the expected number of deaths. For instance,

Age	(1) Population of group	(2) Deaths per 100,000 in general pop.	(1)×(2) Expected # of deaths	(4) Observed deaths
25–34	95,965	105.2	100.9	92
34–44	78,280	203.6	159.4	180
44–54	52,393	428.9	224.7	242
55–64	28,914	964.6	278.9	312
Total			763.9	826

$$\text{SMR} = 826/763.9 = 1.08$$

stratified model. A stratified survival model constrains regression coefficients to be equal across levels of the stratification variable, while allowing other features of the model to vary across strata.

stratified test. A stratified test is performed separately for each stratum. The stratum-specific results are then combined into an overall test statistic.

subhazard, **cumulative subhazard**, and **subhazard ratio**. In a competing-risks analysis, the hazard of the subdistribution (or subhazard for short) for the event of interest (type 1) is defined formally as

$$\overline{h}_1(t) = \lim_{\delta \to 0} \left\{ \frac{P(t < T \leq t + \delta \text{ and event type 1 }) \mid T > t \text{ or } (T \leq t \text{ and not event type 1})}{\delta} \right\}$$

Less formally, think of this hazard as that which generates failure events of interest while keeping subjects who experience competing events "at risk" so that they can be adequately counted as not having any chance of failing.

The cumulative subhazard $\overline{H}_1(t)$ is the integral of the subhazard function $\overline{h}_1(t)$, from 0 (the onset of risk) to t. The cumulative subhazard plays a very important role in competing-risks analysis. The cumulative incidence function (CIF) is a direct function of the cumulative subhazard:

$$\text{CIF}_1(t) = 1 - \exp\{-\overline{H}_1(t)\}$$

The subhazard ratio is the ratio of the subhazard function evaluated at two different values of the covariates: $\overline{h}_1(t|\mathbf{x})/\overline{h}_1(t|\mathbf{x}_0)$. The subhazard ratio is often called the relative subhazard, especially when $\overline{h}_1(t|\mathbf{x}_0)$ is the baseline subhazard function.

survival-time data. See *st data*.

survivor function. Also known as the survivorship function and the survival function, the survivor function, $S(t)$, is 1) the probability of surviving beyond time t, or equivalently, 2) the probability that there is no failure event prior to t, 3) the proportion of the population surviving to time t, or equivalently, 4) the reverse cumulative distribution function of T, the time to the failure event: $S(t) = \Pr(T > t)$. Also see *hazard*.

thrashing. Subjects are said to thrash when they are censored and immediately reenter with different covariates.

time-varying covariates. Time-varying covariates appear in a survival model whose values vary over time. The values of the covariates vary, not the effect. For instance, in a proportional hazards model, the log hazard at time t might be $b \times \text{age}_t + c \times \text{treatment}_t$. Variable age might be time varying, meaning that as the subject ages, the value of age changes, which correspondingly causes the hazard to change. The effect b, however, remains constant.

Time-varying variables are either continuously varying or discretely varying.

In the continuously varying case, the value of the variable x at time t is $x_t = x_o + f(t)$, where $f()$ is some function and often is the identity function, so that $x_t = x_o + t$.

In the discretely varying case, the value of x changes at certain times and often in no particular pattern:

idvar	*t0*	*t*	*bp*
1	0	5	150
1	5	7	130
1	7	9	135

In the above data, the value of *bp* is 150 over the period $(0, 5]$, then 130 over $(5, 7]$, and 135 over $(7, 9]$.

truncation, left-truncation, and right-truncation. In survival analysis, truncation occurs when subjects are observed only if their failure times fall within a certain observational period of a study. Censoring, on the other hand, occurs when subjects are observed for the whole duration of a study, but the exact times of their failures are not known; it is known only that their failures occurred within a certain time span.

Left-truncation occurs when subjects come under observation only if their failure times exceed some time t_l. It is only because they did not fail before t_l that we even knew about their existence. Left-truncation differs from left-censoring in that, in the censored case, we know that the subject failed before time t_l, but we just do not know exactly when.

Imagine a study of patient survival after surgery, where patients cannot enter the sample until they have had a post-surgical test. The patients' survival times will be left-truncated. This is a "delayed entry" problem, one common type of left-truncation.

Right-truncation occurs when subjects come under observation only if their failure times do not exceed some time t_r. Right-truncated data typically occur in registries. For example, a cancer registry includes only subjects who developed a cancer by a certain time, and thus survival data from this registry will be right-truncated.

type I error or **false-positive result**. The type I error of a test is the error of rejecting the null hypothesis when it is true. The probability of committing a type I error, significance level of a test, is often denoted as α in statistical literature. One traditionally used value for α is 5%. Also see *type II error* and *power*.

type II error or **false-negative result**. The type II error of a test is the error of not rejecting the null hypothesis when it is false. The probability of committing a type II error is often denoted as β in statistical literature. Commonly used values for β are 20% or 10%. Also see *type I error* and *power*.

under observation. A subject is under observation when failure events, should they occur, would be observed (and so recorded in the dataset). Being under observation does not mean that a subject is necessarily at risk. Subjects usually come under observation before they are at risk. The statistical concern is with periods when subjects are at risk but not under observation, even when the subject is (later) known not to have failed during the hiatus.

In such cases, since failure events would not have been observed, the subject necessarily had to survive the observational hiatus, and that leads to bias in statistical results unless the hiatus is accounted for properly.

Entry time and exit time record when a subject first and last comes under observation, between which there may be observational gaps, but usually there are not. There is only one entry time and one exit time for each subject. Often, entry time corresponds to analysis time $t = 0$, or before, and exit time corresponds to the time of failure.

Delayed entry means that the entry time occurred after $t = 0$.