Glossary

AIPW estimator. See *augmented inverse-probability-weighted estimator*.

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 $t = \textit{time} - \textit{origin}$.

ATE. See *average treatment effect*.

ATET. See *average treatment effect on the treated*.

augmented inverse-probability-weighted estimator. An augmented inverse-probability-weighted (AIPW) estimator is an inverse-probability-weighted estimator that includes an augmentation term that corrects the estimator when the treatment model is misspecified. When the treatment is correctly specified, the augmentation term vanishes as the sample size becomes large. An AIPW estimator uses both an outcome model and a treatment model and is a doubly robust estimator.

average treatment effect. The average treatment effect is the average effect of the treatment among all individuals in a population.

average treatment effect on the treated. The average treatment effect on the treated is the average effect of the treatment among those individuals who actually get the treatment.

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*.

CI assumption. See *conditional-independence assumption*.

conditional mean. The conditional mean expresses the average of one variable as a function of some other variables. More formally, the mean of y conditional on $\mathbf{x}$ is the mean of y for given values of $\mathbf{x}$; in other words, it is $E(y|\mathbf{x})$.

A conditional mean is also known as a regression or as a conditional expectation.

conditional-independence assumption. The conditional-independence assumption requires that the common variables that affect treatment assignment and treatment-specific outcomes be observable. The dependence between treatment assignment and treatment-specific outcomes can be removed by conditioning on these observable variables.

This assumption is also known as a selection-on-observables assumption because its central tenet is the observability of the common variables that generate the dependence.

counterfactual. A counterfactual is an outcome a subject would have obtained had that subject received a different level of treatment. In the binary-treatment case, the counterfactual outcome for a person who received treatment is the outcome that person would have obtained had the person instead not received treatment; similarly, the counterfactual outcome for a person who did not receive treatment is the outcome that person would have obtained had the person received treatment.

Also see *potential outcome*.

doubly robust estimator. A doubly robust estimator only needs one of two auxiliary models to be correctly specified to estimate a parameter of interest.

Doubly robust estimators for treatment effects are consistent when either the outcome model or the treatment model is correctly specified.

EE estimator. See *estimating-equation estimator*.

estimating-equation estimator. An estimating-equation (EE) estimator calculates parameters estimates by solving a system of equations. Each equation in this system is the sample average of a function that has mean zero.

These estimators are also known as M estimators or Z estimators in the statistics literature and as generalized method of moments (GMM) estimators in the econometrics literature.

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.

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.

i.i.d. sampling assumption. See *independent and identically distributed sampling assumption*.

independent and identically distributed sampling assumption. The independent and identically distributed (i.i.d.) sampling assumption specifies that each observation is unrelated to (independent of) all the other observations and that each observation is a draw from the same (identical) distribution.

individual-level treatment effect. An individual-level treatment effect is the difference in an individual's outcome that would occur because this individual is given one treatment instead of another. In other words, an individual-level treatment effect is the difference between two potential outcomes for an individual.

For example, the blood pressure an individual would obtain after taking a pill minus the blood pressure an individual would obtain had that person not taken the pill is the individual-level treatment effect of the pill on blood pressure.

inverse-probability-weighted estimators. Inverse-probability-weighted (IPW) estimators use weighted averages of the observed outcome variable to estimate the potential-outcome means. The weights are the reciprocals of the treatment probabilities estimated by a treatment model.

inverse-probability-weighted regression-adjustment estimators. Inverse-probability-weighted regression-adjustment (IPWRA) estimators use the reciprocals of the estimated treatment probability as weights to estimate missing-data-corrected regression coefficients that are subsequently used to compute the potential-outcome means.

IPW estimators. See *inverse-probability-weighted estimators*.

IPWRA estimators. See *inverse-probability-weighted regression-adjustment estimators*.

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

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

matching estimator. An estimator that compares differences between the outcomes of similar—that is, matched—individuals. Each individual that receives a treatment is matched to a similar individual that does not get the treatment, and the difference in their outcomes is used to estimate the individual-level treatment effect. Likewise, each individual that does not receive a treatment is matched to a similar individual that does get the treatment, and the difference in their outcomes is used to estimate the individual-level treatment effect.

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

multivalued treatment effect. A multivalued treatment refers to a treatment that has more than two values. For example, a person could have taken a 20 mg dose of a drug, a 40 mg dose of the drug, or not taken the drug at all.

nearest-neighbor matching. Nearest-neighbor matching uses the distance between observed variables to find similar individuals.

observational data. In observational data, treatment assignment is not controlled by those who collected the data; thus some common variables affect treatment assignment and treatment-specific outcomes.

outcome model. An outcome model is a model used to predict the outcome as a function of covariates and parameters.

overlap assumption. The overlap assumption requires that each individual have a positive probability of each possible treatment level.

POMs. See *potential-outcome means*.

potential outcome. The potential outcome is the outcome an individual would obtain if given a specific treatment.

For example, an individual has one potential blood pressure after taking a pill and another potential blood pressure had that person not taken the pill.

potential-outcome means. The potential-outcome means refers to the means of the potential outcomes for a specific treatment level.

The mean blood pressure if everyone takes a pill and the mean blood pressure if no one takes a pill are two examples.

The average treatment effect is the difference between potential-outcome mean for the treated and the potential-outcome mean for the not treated.

propensity score. The propensity score is the probability that an individual receives a treatment.

propensity-score matching. Propensity-score matching uses the distance between estimated propensity scores to find similar individuals.

regression-adjustment estimators. Regression-adjustment estimators use means of predicted outcomes for each treatment level to estimate each potential-outcome mean.

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

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

selection-on-observables. See *conditional-independence assumption*.

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*.

smooth treatment-effects estimator. A smooth treatment-effects estimator is a smooth function of the data so that standard methods approximate the distribution of the estimator. The RA, IPW, AIPW, and IPWRA estimators are all smooth treatment-effects estimators while the nearest-neighbor matching estimator and the propensity-score matching estimator are not.

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**.

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, cumulative hazard, and hazard ratio*.

treatment model. A treatment model is a model used to predict treatment-assignment probabilities as a function of covariates and parameters.

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.

unconfoundedness. See *conditional-independence assumption*.

weighted-regression-adjustment estimator. Weighted-regression-adjustment estimators use means of predicted outcomes for each treatment level to estimate each potential-outcome mean. The weights are used to estimate censoring-adjusted regression coefficients.