Glossary

100% sample. See *census*.

balanced repeated replication. Balanced repeated replication (BRR) is a method of variance estimation for designs with two PSUs in every stratum. The BRR variance estimator tends to give more reasonable variance estimates for this design than does the linearized variance estimator, which can result in large values and undesirably wide confidence intervals. The BRR variance estimator is described in [SVY] **variance estimation**.

bootstrap. The bootstrap is a method of variance estimation. The bootstrap variance estimator for survey data is described in [SVY] **variance estimation**.

BRR. See *balanced repeated replication*.

census. When a census of the population is conducted, every individual in the population participates in the survey. Because of the time, cost, and other constraints, the data collected in a census are typically limited to items that can be quickly and easily determined, usually through a questionnaire.

cluster. A cluster is a collection of individuals that are sampled as a group. Although the cost in time and money can be greatly decreased, cluster sampling usually results in larger variance estimates when compared with designs in which individuals are sampled independently.

DEFF and **DEFT**. DEFF and DEFT are design effects. Design effects compare the sample-to-sample variability from a given survey dataset with a hypothetical SRS design with the same number of individuals sampled from the population.

DEFF is the ratio of two variance estimates. The design-based variance is in the numerator; the hypothetical SRS variance is in the denominator.

DEFT is the ratio of two standard-error estimates. The design-based standard error is in the numerator; the hypothetical SRS with-replacement standard error is in the denominator. If the given survey design is sampled with replacement, DEFT is the square root of DEFF.

delta method. See *linearization*.

design effects. See *DEFF* and *DEFT*.

direct standardization. Direct standardization is an estimation method that allows comparing rates that come from different frequency distributions.

Estimated rates (means, proportions, and ratios) are adjusted according to the frequency distribution from a standard population. The standard population is partitioned into categories called standard strata. The stratum frequencies for the standard population are called standard weights. The standardizing frequency distribution typically comes from census data, and the standard strata are most commonly identified by demographic information such as age, sex, and ethnicity.

finite population correction. Finite population correction (FPC) is an adjustment applied to the variance of a point estimator because of sampling without replacement, resulting in variance estimates that are smaller than the variance estimates from comparable with-replacement sampling designs.

FPC. See *finite population correction*.

Hadamard matrix. A Hadamard matrix is a square matrix with r rows and columns that has the property

$$H_r' H_r = r I_r$$

where I_r is the identity matrix of order r. Generating a Hadamard matrix with order $r = 2^p$ is easily accomplished. Start with a Hadamard matrix of order 2 (H_2), and build your H_r by repeatedly applying Kronecker products with H_2.

jackknife. The jackknife is a data-dependent way to estimate the variance of a statistic, such as a mean, ratio, or regression coefficient. Unlike BRR, the jackknife can be applied to practically any survey design. The jackknife variance estimator is described in [SVY] **variance estimation**.

linearization. Linearization is short for Taylor linearization. Also known as the delta method or the Huber/White/robust sandwich variance estimator, linearization is a method for deriving an approximation to the variance of a point estimator, such as a ratio or regression coefficient. The linearized variance estimator is described in [SVY] **variance estimation**.

MEFF and **MEFT**. MEFF and MEFT are misspecification effects. Misspecification effects compare the variance estimate from a given survey dataset with the variance from a misspecified model. In Stata, the misspecified model is fit without weighting, clustering, or stratification.

MEFF is the ratio of two variance estimates. The design-based variance is in the numerator; the misspecified variance is in the denominator.

MEFT is the ratio of two standard-error estimates. The design-based standard error is in the numerator; the misspecified standard error is in the denominator. MEFT is the square root of MEFF.

misspecification effects. See *MEFF* and *MEFT*.

point estimate. A point estimate is another name for a statistic, such as a mean or regression coefficient.

poststratification. Poststratification is a method for adjusting sampling weights, usually to account for underrepresented groups in the population. This usually results in decreased bias because of nonresponse and underrepresented groups in the population. Poststratification also tends to result in smaller variance estimates.

The population is partitioned into categories, called poststrata. The sampling weights are adjusted so that the sum of the weights within each poststratum is equal to the respective poststratum size. The poststratum size is the number of individuals in the population that are in the poststratum. The frequency distribution of the poststrata typically comes from census data, and the poststrata are most commonly identified by demographic information such as age, sex, and ethnicity.

predictive margins. Predictive margins provide a way of exploring the response surface of a fitted model in any response metric of interest—means, linear predictions, probabilities, marginal effects, risk differences, and so on. Predictive margins are estimates of responses (or outcomes) for the groups represented by the levels of a factor variable, controlling for the differing covariate distributions across the groups. They are the survey-data and nonlinear response analogue to what are often called estimated marginal means or least-squares means for linear models.

Because these margins are population-weighted averages over the estimation sample or subsamples, and because they take account of the sampling distribution of the covariates, they can be used to make inferences about treatment effects for the population.

primary sampling unit. Primary sampling unit (PSU) is a cluster that was sampled in the first sampling stage; see *cluster*.

probability weight. Probability weight is another term for sampling weight.

pseudolikelihood. A pseudolikelihood is a weighted likelihood that is used for point estimation. Pseudolikelihoods are not true likelihoods because they do not represent the distribution function for the sample data from a survey. The sampling distribution is instead determined by the survey design.

PSU. See *primary sampling unit*.

replicate-weight variable. A replicate-weight variable contains sampling weight values that were adjusted for resampling the data; see [SVY] **variance estimation** for more details.

resampling. Resampling refers to the process of sampling from the dataset. In the delete-one jackknife, the dataset is resampled by dropping one PSU and producing a replicate of the point estimates. In the BRR method, the dataset is resampled by dropping combinations of one PSU from each stratum. The resulting replicates of the point estimates are used to estimate their variances and covariances.

sample. A sample is the collection of individuals in the population that were chosen as part of the survey. Sample is also used to refer to the data, typically in the form of answered questions, collected from the sampled individuals.

sampling stage. Complex survey data are typically collected using multiple stages of clustered sampling. In the first stage, the PSUs are independently selected within each stratum. In the second stage, smaller sampling units are selected within the PSUs. In later stages, smaller and smaller sampling units are selected within the clusters from the previous stage.

sampling unit. A sampling unit is an individual or collection of individuals from the population that can be selected in a specific stage of a given survey design. Examples of sampling units include city blocks, high schools, hospitals, and houses.

sampling weight. Given a survey design, the sampling weight for an individual is the reciprocal of the probability of being sampled. The probability for being sampled is derived from stratification and clustering in the survey design. A sampling weight is typically considered to be the number of individuals in the population represented by the sampled individual.

sampling with and without replacement. Sampling units may be chosen more than once in designs that use sampling with replacement. Sampling units may be chosen at most once in designs that use sampling without replacement. Variance estimates from with-replacement designs tend to be larger than those from corresponding without-replacement designs.

SDR. See *successive difference replication*.

secondary sampling unit. Secondary sampling unit (SSU) is a cluster that was sampled from within a PSU in the second sampling stage. SSU is also used as a generic term unit to indicate any sampling unit that is not from the first sampling stage.

simple random sample. In a simple random sample (SRS), individuals are independently sampled—each with the same probability of being chosen.

SRS. See *simple random sample*.

SSU. See *secondary sampling unit*.

standard strata. See *direct standardization*.

standard weights. See *direct standardization*.

stratification. The population is partitioned into well-defined groups of individuals, called strata. In the first sampling stage, PSUs are independently sampled from within each stratum. In later sampling stages, SSUs are independently sampled from within each stratum for that stage.

Survey designs that use stratification typically result in smaller variance estimates than do similar designs that do not use stratification. Stratification is most effective in decreasing variability when sampling units are more similar within the strata than between them.

subpopulation estimation. Subpopulation estimation focuses on computing point and variance estimates for part of the population. The variance estimates measure the sample-to-sample variability, assuming that the same survey design is used to select individuals for observation from the population. This approach results in a different variance than measuring the sample-to-sample variability by restricting the samples to individuals within the subpopulation; see [SVY] **subpopulation estimation**.

successive difference replication. Successive difference replication (SDR) is a method of variance typically applied to systematic samples, where the observed sampling units are somehow ordered. The SDR variance estimator is described in [SVY] **variance estimation**.

survey data. Survey data consist of information about individuals that were sampled from a population according to a survey design. Survey data distinguishes itself from other forms of data by the complex nature under which individuals are selected from the population.

In survey data analysis, the sample is used to draw inferences about the population. Furthermore, the variance estimates measure the sample-to-sample variability that results from the survey design applied to the fixed population. This approach differs from standard statistical analysis, in which the sample is used to draw inferences about a physical process and the variance measures the sample-to-sample variability that results from independently collecting the same number of observations from the same process.

survey design. A survey design describes how to sample individuals from the population. Survey designs typically include stratification and cluster sampling at one or more stages.

Taylor linearization. See *linearization*.

variance estimation. Variance estimation refers to the collection of methods used to measure the amount of sample-to-sample variation of point estimates; see [SVY] **variance estimation**.