

➤ Correctly and Easily Compute Statistics for Complex Samples

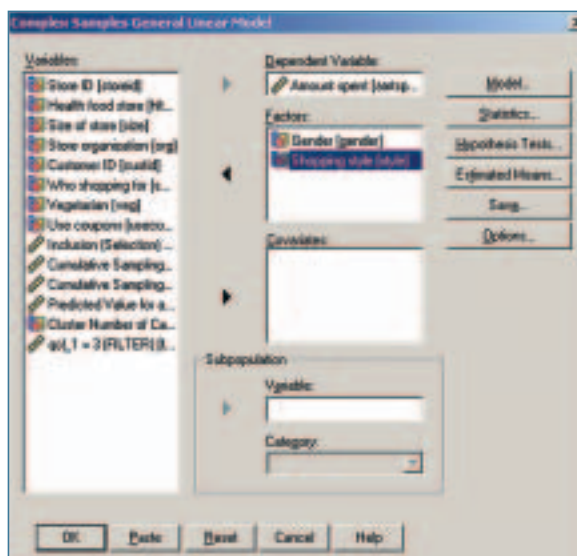
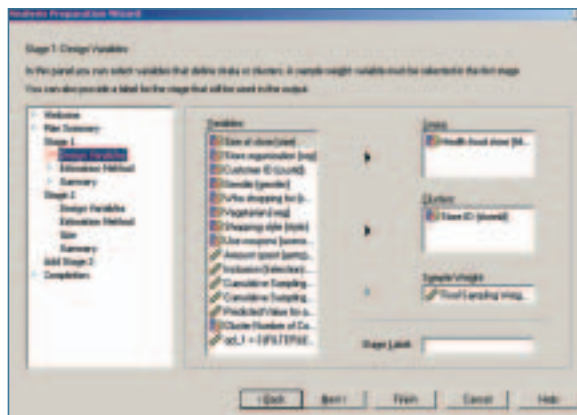
When you conduct sample surveys, use a statistics package dedicated to producing correct estimates for complex sample data. SPSS Complex Samples provides specialized statistics that enable you to correctly and easily compute statistics and their standard errors from complex sample designs. You can apply it to:

- Survey research—Obtain descriptive and inferential statistics for survey data
- Market research—Analyze customer satisfaction data
- Health research—Analyze large public-use datasets on public health topics such as health and nutrition or alcohol use and traffic fatalities
- Social science—Conduct secondary research on public survey datasets
- Public opinion research—Characterize attitudes on policy issues

SPSS Complex Samples provides you with:

- Everything you need for working with complex samples, from the planning stage and sampling through the analysis stage
- An intuitive Sampling Wizard that guides you step by step through the process of designing a scheme and drawing a sample
- An easy-to-use Analysis Preparation Wizard to help prepare public-use datasets that have been sampled, such as the National Health Inventory Survey data from the Centers for Disease Control and Prevention (CDC)
- Easier collaboration with colleagues through sharing of sampling and analysis plans
- More accurate analyses because you can take up to three stages into account when analyzing data from a multistage design
- The ability to assess your design's impact

- A more accurate picture of your data because, unlike traditional statistics, subpopulation assessments take other subpopulations into account
- Numerical outcome prediction through the complex samples general linear model (CSGLM)
- Categorical outcome prediction through complex samples logistic regression



A grocery store wants to determine if the frequency with which customers shop is related to the amount spent, controlling for gender of the customer and incorporating a sample design. First, the store specifies the sample design used in the Analysis Preparation Wizard (top). Next, the store sets up the model in the complex samples general linear model (CSGLM) (bottom).

You can use the following types of sample design information with SPSS Complex Samples:

- **Stratified sampling**—Increase the precision of your sample or ensure a representative sample from key groups by choosing to sample within subgroups of the survey population. For example, subgroups might be a specific number of males or females or contain people in certain job categories, people of a certain age group, and so on.
- **Clustered sampling**—Select clusters, which are groups of sampling units, for your survey. Clusters can include schools, hospitals, or geographic areas with sampling units that might be students, patients, or citizens. Clustering often helps make surveys more cost-effective.
- **Multistage sampling**—Select an initial or first-stage sample based on groups of elements in the population, then create a second-stage sample by drawing a sub-sample from each selected unit in the first-stage sample. By repeating this option, you can select a higher-stage sample. For example, in a face-to-face survey, you might sample individuals within households and city blocks.

More confidently reach results

As a researcher, you want to be confident about your results. Most conventional statistical software assumes your data arise from simple random sampling. Simple random sampling, however, is generally neither feasible nor cost-effective in most large-scale surveys. Analyzing such sample data with conventional statistics risks incorrect results. For example, estimated standard errors of statistics are often too small, giving you a false sense of precision. SPSS Complex Samples enables you to achieve more statistically valid inferences for populations measured in your complex sample data because it incorporates the sample design into survey analysis.

Work efficiently and easily

Only SPSS Complex Samples makes understanding and working with your complex sample survey results easy. Through the intuitive interface, you can analyze data and interpret results. When you're finished, you can publish datasets and include your sampling or analysis plans. The plan acts as a template and allows you to save all the decisions made when creating the plan—define it once and you're done. This saves time and improves accuracy for yourself and others who may want to plug your plans into the data to replicate results or pick up where you left off.

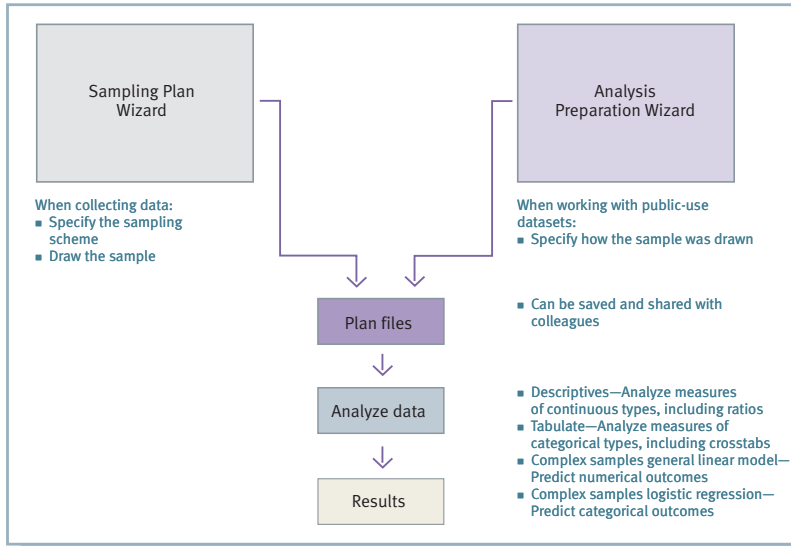
To begin your work in SPSS Complex Samples, use the wizards, which prompt you for the many factors you must consider. If you are creating your own samples, use the Sampling Wizard to define the sampling scheme. If you're using public-use datasets that have been sampled, such as those provided by the CDC, use the Analysis Preparation Wizard to specify how the samples were defined and how to estimate standard errors. Once you create a sample or specify standard errors, you can create plans, analyze your data, and produce results (see the diagram on the next page for workflow).

SPSS Complex Samples makes it easy to learn and work fast. Use the online help system, explore the interactive case studies, or run the online tutorial to learn more about using your data with the software. SPSS Complex Samples enables you to:

- Reach correct point estimates for statistics such as totals, means, and ratios
- Obtain the standard errors of these statistics
- Produce correct confidence intervals and hypothesis tests
- Predict numerical outcomes
- Predict categorical outcomes



■ Accurate analysis of survey data



Accurate analysis of survey data is easy in SPSS Complex Samples. Start with one of the wizards (which one depends on your data source) and then use the interactive interface to create plans, analyze data, and interpret results.

Features

Statistics

Complex Samples Plan (CSPLAN)

- Provides a common place to specify the sampling frame to create a complex sample design or analysis specification used by companion procedures in the SPSS Complex Samples add-on module. CSPLAN does not actually extract the sample or analyze data. To sample cases, use a sample design created by CSPLAN as input to the CSSELECT procedure (described on the next page). To analyze sample data, use an analysis design created by CSPLAN as input to the CSGLM, CSLOGISTIC, CSDESCRIPTIVES, or CSTABULATE procedures (described on the following pages).
 - Create a sample design: Use to extract sampling units from the active file
 - Create an analysis design: Use to analyze a complex sample
 - When you create a sample design, the procedure automatically saves an appropriate analysis design to the plan file. A plan file created for designing a sample, therefore, can be used for both sample selection and analysis.
 - Display a sample design or analysis design
 - Specify the plan in an external file
 - Name plan-wise variables to be created when you extract a sample or use it as input to the selection or estimation process with the PLANVARS subcommand
 - Specify final sample weights for each unit to be used by SPSS Complex Samples analysis procedures in the estimation process
 - Indicate overall sample weights that will be generated when the sample design is executed in the CSSELECT procedure
 - Select weights to be used when computing final sampling weights in a multistage design
 - Control output from the CSPLAN procedure with the PRINT subcommand
 - Display a plan specifications summary in which the output reflects your specifications at each stage of the design
 - Display a table showing MATRIX specifications
 - Signal stages of the design with the DESIGN subcommand. You can also use this subcommand to define stratification variables, cluster variables, or create descriptive labels for particular stages.
 - Specify the sample extraction method using the METHOD subcommand. Select from a variety of equal- and unequal-probability methods, including simple and systematic random sampling. Methods for sampling with probability proportionate to size (PPS) are also available. Units can be drawn with replacement (WR) or without replacement (WOR) from the population.
 - SIMPLE_WOR: Select units with equal probability. Extract units without replacement.
 - SIMPLE_WR: Select units with equal probability. Extract units with replacement.
 - SIMPLE_SYSTEMATIC: Select units at a fixed interval throughout the sampling frame or stratum. A random starting point is chosen within the first interval.
 - SIMPLE_CHROMY: Select units sequentially with equal probability. Extract units without replacement.
 - PPS_WOR: Select units with probability proportional to size. Extract units without replacement.
 - PPS_WR: Select units with probability proportional to size. Extract units with replacement.
 - PPS_SYSTEMATIC: Select units by systematic random sampling with probability proportional to size. Extract units without replacement.
 - PPS_CHROMY: Select units sequentially with probability proportional to size without replacement.
 - PPS_BREWER: Select two units from each stratum with probability proportional to size. Extract units without replacement.
 - PPS_MURTHY: Select two units from each stratum with probability proportional to size. Extract units without replacement.
 - PPS_SAMPFORD: Extends Brewer's method to select more than two units from each stratum with probability proportional to size. Extract units without replacement.

- Control for the number or percentage of units to be drawn: Set at each stage of the design. You can also choose output variables, such as stagewise sampling weights, which are created upon the sample design execution.
- Estimation methods: With replacement, equal probability without replacement in the first stage, and unequal probability without replacement
- Unequal probability estimation without replacement: Request in the first stage only
- Variable specification: Specify variables for input for the estimation process, including overall sample weights and inclusion probabilities
- Specify the number of sampling units drawn at the current stage using the SIZE subcommand
- Specify the percentage of units drawn at the current stage. For example, the sampling fraction, using the RATE subcommand
- Specify the minimum number of units drawn when you specify RATE. This is useful when the sampling rate for a particular stratum is very small due to rounding.
- Specify the maximum number of units to draw when you specify RATE. This is useful when the sampling rate for a particular stratum is larger than desired due to rounding.
- Specify the measure of size for population units in a PPS design. Specify a variable that contains the sizes or request that sizes be determined when the CSSELECT procedure scans the sample frame.
- Obtain stagewise sample information variables when you execute a sample design using the STAGEVARS subcommand. You can obtain:
 - The proportion of units drawn from the population at a particular stage using stagewise inclusion (selection) probabilities
 - Prior stages using cumulative sampling weight for a given stage
 - Uniquely identified units that have been selected more than once when your sample is done with replacement, with a duplication index for units selected in a given stage
 - Population size for a given stage
 - Number of units drawn at a given stage
 - Stagewise sampling rate
 - Sampling weight for a given stage

- Choose an estimation method for the current stage with the ESTIMATOR subcommand. You can indicate:
 - Equal selection probabilities without replacement
 - Unequal selection probabilities without replacement
 - Selection with replacement
- Specify the population size for each sample element with the POPSIZE subcommand
- Specify the proportion of units drawn from the population at a given stage with the INCLPROB subcommand

Complex Samples Selection (CSSELECT)

- Selects complex, probability-based samples from a population. CSSELECT chooses units according to a sample design created through the CSPLAN procedure.
 - Control the scope of execution and specify a seed value with the CRITERIA subcommand
 - Control whether or not user-missing values of classification (stratification and clustering) variables are treated as valid values with the CLASSMISSING subcommand
 - Use the most updated Mersenne Twister random number generator to select the sample
 - Specify general options concerning input and output files with the DATA subcommand
 - Opt to rename existing variables when the CSSELECT procedure writes sample weight variables and stagewise output variables requested in the plan file, such as inclusion probabilities
 - Write sampled units to an external file using an option to keep/drop specified variables
 - Automatically save first-stage joint inclusion probabilities to an external file when the plan file specifies a PPS_WR sampling method
 - Opt to generate text files containing a rule that describes characteristics of selected units
 - Control output display through the PRINT subcommand
 - Summarize the distribution of selected cases across strata. Information is reported per design stage.
 - Produce a case-processing summary

Complex Samples Descriptives (CSDESCRIPTIVES)

- Estimates means, sums, and ratios, and computes their standard errors, design effects, confidence intervals, and hypothesis tests for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSDESCRIPTIVES performs analyses for subpopulations.
 - Specify the name of a plan file, which is written by the CSPLAN procedure, containing analysis design specifications with the PLAN subcommand
 - Specify joint inclusion probabilities file names
 - Specify the analysis variables used by the MEAN and SUM subcommands using the SUMMARY subcommand
 - Request that means be estimated for variables specified on the SUMMARY subcommand through the MEAN subcommand
 - Request *t* tests of the population mean(s) and give the null hypothesis value(s) through the TTEST keyword. If you define subpopulations using the SUBPOP subcommand, then null hypothesis values are used in the test(s) for each subpopulation, as well as for the entire population.
 - Request that sums be estimated for variables specified on the SUMMARY subcommand through the SUM subcommand
 - Request *t* tests of the population sums and give the null hypothesis value(s) through the TTEST keyword. If you define subpopulations using the SUBPOP subcommand, then null hypothesis values are used in the test(s) for each subpopulation, as well as for the entire population.
 - Request that ratios be estimated for variables specified on the SUMMARY subcommand through the RATIO subcommand
 - Request *t* tests of the population ratios and give the null hypothesis value(s) through the TTEST keyword
 - Associate syntax with the mean, sum, or ratio estimates, including:
 - The number of valid observations in the dataset for each mean, sum, or ratio estimate

- The population size for each mean, sum, or ratio estimate
- The standard error for each mean, sum, or ratio estimate
- Coefficient of variation
- Design effects
- Square root of the design effects
- Confidence interval
- Specify subpopulations for which analyses are to be performed using the SUBPOP subcommand
 - Display results for all subpopulations in the same or a separate table
- Specify how to handle missing data
 - Base each statistic on all valid data for the analysis variable(s) used in computing the statistic. Compute ratios using all cases with valid data for both of the specified variables. You may base statistics for different variables on different sample sizes.
 - Base only cases with valid data for all analysis variables when computing statistics. Always base statistics for different variables on the same sample size.
 - Exclude user-missing values among the strata, cluster, and subpopulation variables
 - Include user-missing values among the strata, cluster, and subpopulation variables. Treat user-missing values for these variables as valid data.

Complex Samples Tabulate (CSTABULATE)

- Displays one-way frequency tables or two-way crosstabulations and associated standard errors, design effects, confidence intervals, and hypothesis tests for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSTABULATE creates tables for subpopulations.
 - Specify the name of an XML file, written by the CSPLAN procedure, containing analysis design using the PLAN subcommand
 - Specify the joint inclusion probabilities file name
 - Use the following statistics within the table:
 - Population size: Estimate the population size for each cell and marginal in a table
 - Standard error: Calculate the standard error for each population size estimate

- Row and column percentages: Express the population size estimate for each cell in a row or column as a percentage of the population size estimate for that row or column. This functionality is available for two-way crosstabulations.
 - Table percentages: Express the population size estimate in each cell of a table as a percentage of the population size estimate for that table
 - Coefficient of variation
 - Design effects
 - Square root of the design effects
 - Confidence interval: Specify any number between zero and 100 as the confidence interval
 - Unweighted counts: Use unweighted counts as the number of valid observations in the dataset for each population size estimate
 - Cumulative population size estimates: Use cumulative population size estimates for one-way frequency tables only
 - Cumulative percentages: Use cumulative percentages corresponding to the population size estimates for one-way frequency tables only
 - Expected population size estimates: Use expected population size estimates if the population size estimate of each cell in the two variables in the crosstabulation are statistically independent. This functionality is available for two-way crosstabulations only.
 - Residuals: Show the difference between the observed and expected population size estimates in each cell. This functionality is available for two-way crosstabulations only.
 - Pearson residuals: This functionality is available for two-way crosstabulations only
 - Adjusted Pearson residuals: This functionality is available for two-way crosstabulations only
- Use the following statistics and tests for the entire table:
 - Test of homogeneous proportions
 - Test of independence
 - Odds ratio
 - Relative risk
 - Risk difference
- Specify subpopulations for which analyses are to be performed using the SUBPOP subcommand
 - Display results for all subpopulations in the same or a separate table

- Specify how to handle missing data
 - Base each table on all valid data for the tabulation variable(s) used in creating the table. You may base tables for different variables on different sample sizes.
 - Use only cases with valid data for all tabulation variables in creating the tables. Always base tables for different variables on the same sample size.
 - Exclude user-missing values among the strata, cluster, and subpopulation variables
 - Include user-missing values among the strata, cluster, and subpopulation variables. Treat user-missing values for these variables as valid data.

Complex Samples General Linear Models (CSGLM)

- Enables you to build linear regression, analysis of variance (ANOVA), and analysis of covariance (ANCOVA) models for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSGLM performs analyses for subpopulations.
 - Models
 - Main effects
 - All n-way interactions
 - Fully crossed
 - Custom, including nested terms
 - Statistics
 - Model parameters: Coefficient estimates, standard error for each coefficient estimate, *t* test for each coefficient estimate, confidence interval for each coefficient estimate, design effect for each coefficient estimate, and square root of the design effect for each coefficient estimate
 - Population means of dependent variable and covariates
 - Model fit
 - Sample design information
 - Hypothesis tests
 - Test statistics: Wald F test, adjusted Wald F test, Wald Chi-square test, and adjusted Wald Chi-square test
 - Adjustment for multiple comparisons: Least significant difference, Bonferroni, sequential Bonferroni, Sidak, and sequential Sidak
 - Sampling degrees of freedom: Based on sample design or fixed by user

- Estimated means: Requests estimated marginal means for factors and interactions in the model
 - Contrasts: Simple, deviation, Helmert, repeated, or polynomial
- Model variables can be saved to the active file and/or exported to external files that contain parameter matrices
 - Variables: Predicted values and residuals
 - Parameter covariance matrix and its other statistics, as well as parameter correlation matrix and its other statistics, can be exported as an SPSS data file
 - Parameter estimates and/or the parameter covariance matrix can be exported to an XML file
- Output
 - Sample design information (such as strata and PSUs)
 - Regression coefficient estimates and *t* tests
 - Summary information about the dependent variable, covariates, and factors
 - Summary information about the sample, including the unweighted count and population size
 - Confidence limits for parameter estimates and user-specified confidence levels
 - Wald F test for model effects
 - Design effects
 - Multiple R-square
 - Set of contrast coefficients (L) matrices
 - Variance-covariance matrix of regression coefficient estimates
 - Root mean square error
 - Covariance and correlation matrices for regression coefficients
- Missing data handling
 - Listwise deletion of missing values
- Other
 - User-specified denominator, *df*, used in computing *p*-values for all test statistics
 - Collinearity diagnostics
 - Model can be fitted for subpopulations

Complex Samples Logistic Regression (CSLOGISTIC)

- Performs binary logistic regression analysis, as well as multinomial logistic regression (MLR) analysis, for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSLOGISTIC performs analyses for subpopulations.
 - Models
 - Main effects
 - All *n*-way interactions
 - Fully crossed
 - Custom, including nested terms
 - Statistics
 - Model parameters: Coefficient estimates, exponential estimates, standard error for each coefficient estimate, *t* test for each coefficient estimate, confidence interval for each coefficient estimate, design effect for each coefficient estimate, square root of the design effect for each coefficient estimate, covariances of parameter estimates, and correlations of the parameter estimates
 - Model fit: Pseudo R-square and classification table
 - Summary statistics for model variables
 - Sample design information
 - Hypothesis tests
 - Test statistics: Wald F test, adjusted Wald F test, Wald Chi-square test, and adjusted Wald Chi-square test
 - Adjustment for multiple comparisons: Least significant difference, Bonferroni, sequential Bonferroni, Sidak, and sequential Sidak
 - Sampling degrees of freedom: Based on sample design or fixed by user
 - Model variables can be saved to the active file and/or exported to external files that contain parameter matrices
 - Variables: Predicted category and predicted probabilities

- Parameter covariance matrix and its other statistics, as well as parameter correlation matrix and its other statistics, can be exported as an SPSS data file
- Parameter estimates and/or the parameter covariance matrix can be exported to an XML file
- Output
 - Sample design information (such as strata and PSUs)
 - Summary information about the dependent variable, covariates, and factors
 - Summary information about the sample, including the unweighted count and population size
 - Confidence limits for parameter estimates and user-specified confidence levels
 - Model summary statistics
 - Wald F test for model effects
 - Design effects
 - Classification table
 - Set of contrast coefficients (L) matrices
 - Variance-covariance matrix of regression coefficient estimates
 - Root mean square error
 - Covariance and correlation matrices for regression coefficients
- Missing data handling
 - Listwise deletion of missing values
- Other
 - User-specified denominator, *df*, used in computing *p*-values for all test statistics
 - Collinearity diagnostics
 - Model can be fitted for subpopulations

System requirements

- Software: SPSS Base 13.0
- Minimum free drive space: 1MB
- Other system requirements vary according to platform

