SPSS Complex Samples™ 15.0

For more information about SPSS® software products, please visit our Web site at *http://www.spss.com* or contact

SPSS Inc.
233 South Wacker Drive, 11th Floor
Chicago, IL 60606-6412
Tel: (312) 651-3000
Fax: (312) 651-3668

1 2 3 4 5 6 7 8 9 0   09 08 07 06

ISBN-13: 978-1-56827-383-9
ISBN-10: 1-56827-383-5

Preface

SPSS 15.0 is a comprehensive system for analyzing data. The SPSS Complex Samples optional add-on module provides the additional analytic techniques described in this manual. The Complex Samples add-on module must be used with the SPSS 15.0 Base system and is completely integrated into that system.

Installation

To install the SPSS Complex Samples add-on module, run the License Authorization Wizard using the authorization code that you received from SPSS Inc. For more information, see the installation instructions supplied with the SPSS Complex Samples add-on module.

Compatibility

SPSS is designed to run on many computer systems. See the installation instructions that came with your system for specific information on minimum and recommended requirements.

Serial Numbers

Your serial number is your identification number with SPSS Inc. You will need this serial number when you contact SPSS Inc. for information regarding support, payment, or an upgraded system. The serial number was provided with your Base system.

Customer Service

If you have any questions concerning your shipment or account, contact your local office, listed on the SPSS Web site at *http://www.spss.com/worldwide*. Please have your serial number ready for identification.

Training Seminars

SPSS Inc. provides both public and onsite training seminars. All seminars feature hands-on workshops. Seminars will be offered in major cities on a regular basis. For more information on these seminars, contact your local office, listed on the SPSS Web site at *http://www.spss.com/worldwide*.

Technical Support

The services of SPSS Technical Support are available to maintenance customers. Customers may contact Technical Support for assistance in using SPSS or for installation help for one of the supported hardware environments. To reach Technical Support, see the SPSS Web site at *http://www.spss.com*, or contact your local office, listed on the SPSS Web site at *http://www.spss.com/worldwide*. Be prepared to identify yourself, your organization, and the serial number of your system.

Additional Publications

Additional copies of SPSS product manuals may be purchased directly from SPSS Inc. Visit the SPSS Web Store at *http://www.spss.com/estore*, or contact your local SPSS office, listed on the SPSS Web site at *http://www.spss.com/worldwide*. For telephone orders in the United States and Canada, call SPSS Inc. at 800-543-2185. For telephone orders outside of North America, contact your local office, listed on the SPSS Web site.

The *SPSS Statistical Procedures Companion*, by Marija Norušis, has been published by Prentice Hall. A new version of this book, updated for SPSS 15.0, is planned. The *SPSS Advanced Statistical Procedures Companion*, also based on SPSS 15.0, is forthcoming. The *SPSS Guide to Data Analysis* for SPSS 15.0 is also in development. Announcements of publications available exclusively through Prentice Hall will be available on the SPSS Web site at *http://www.spss.com/estore* (select your home country, and then click Books).

Tell Us Your Thoughts

Your comments are important. Please let us know about your experiences with SPSS products. We especially like to hear about new and interesting applications using the SPSS Complex Samples add-on module. Please send e-mail to *suggest@spss.com* or write to SPSS Inc., Attn.: Director of Product Planning, 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.

About This Manual

This manual documents the graphical user interface for the procedures included in the SPSS Complex Samples add-on module. Illustrations of dialog boxes are taken from SPSS for Windows. Dialog boxes in other operating systems are similar. Detailed information about the command syntax for features in the SPSS Complex Samples add-on module is available in two forms: integrated into the overall Help system and as a separate document in PDF form in the *SPSS 15.0 Command Syntax Reference*, available from the Help menu.

Contacting SPSS

If you would like to be on our mailing list, contact one of our offices, listed on our Web site at *http://www.spss.com/worldwide*.

Contents

Part I: User's Guide

1 Introduction to SPSS Complex Samples Procedures 1

Properties of Complex Samples 1
Usage of Complex Samples Procedures 2
Plan Files. 3
Further Readings 3

2 Sampling from a Complex Design 4

Creating a New Sample Plan 5
Sampling Wizard: Design Variables 6
Tree Controls for Navigating the Sampling Wizard 7
Sampling Wizard: Sampling Method 8
Sampling Wizard: Sample Size 10
Define Unequal Sizes. 11
Sampling Wizard: Output Variables. 12
Sampling Wizard: Plan Summary 14
Sampling Wizard: Draw Sample Selection Options 15
Sampling Wizard: Draw Sample Output Files. 16
Sampling Wizard: Finish 17
Modifying an Existing Sample Plan 18
Sampling Wizard: Plan Summary 19

Running an Existing Sample Plan 20
CSPLAN and CSSELECT Commands Additional Features 20

3 Preparing a Complex Sample for Analysis 21

Creating a New Analysis Plan 22
Analysis Preparation Wizard: Design Variables 23
Tree Controls for Navigating the Analysis Wizard 24
Analysis Preparation Wizard: Estimation Method 25
Analysis Preparation Wizard: Size 26
Define Unequal Sizes 27
Analysis Preparation Wizard: Plan Summary 28
Analysis Preparation Wizard: Finish 29
Modifying an Existing Analysis Plan 30
Analysis Preparation Wizard: Plan Summary 31

4 Complex Samples Plan 32

5 Complex Samples Frequencies 34

Complex Samples Frequencies Statistics 36
Complex Samples Missing Values 37
Complex Samples Options 37

6 Complex Samples Descriptives 39

Complex Samples Descriptives Statistics 41
Complex Samples Descriptives Missing Values. 42
Complex Samples Options 42

7 Complex Samples Crosstabs 44

Complex Samples Crosstabs Statistics 46
Complex Samples Missing Values. 48
Complex Samples Options 48

8 Complex Samples Ratios 50

Complex Samples Ratios Statistics 52
Complex Samples Ratios Missing Values 53
Complex Samples Options 53

9 Complex Samples General Linear Model 55

Complex Samples General Linear Model Statistics 59
Complex Samples Hypothesis Tests 60
Complex Samples General Linear Model Estimated Means 62
Complex Samples General Linear Model Save 63
Complex Samples General Linear Model Options 64
CSGLM Command Additional Features 65

10 Complex Samples Logistic Regression 66

Complex Samples Logistic Regression Reference Category 68
Complex Samples Logistic Regression Model 69
Complex Samples Logistic Regression Statistics. 71
Complex Samples Hypothesis Tests 73
Complex Samples Logistic Regression Odds Ratios. 74
Complex Samples Logistic Regression Save 75
Complex Samples Logistic Regression Options 77
CSLOGISTIC Command Additional Features 78

11 Complex Samples Ordinal Regression 79

Complex Samples Ordinal Regression Response Probabilities. 81
Complex Samples Ordinal Regression Model 82
Complex Samples Ordinal Regression Statistics 84
Complex Samples Hypothesis Tests 86
Complex Samples Ordinal Regression Odds Ratios 87
Complex Samples Ordinal Regression Save 89
Complex Samples Ordinal Regression Options 90
CSORDINAL Command Additional Features. 92

Part II: Examples

12 Complex Samples Sampling Wizard 94

Obtaining a Sample from a Full Sampling Frame 94
Using the Wizard 94
Plan Summary 104
Sampling Summary 105
Sample Results 106
Obtaining a Sample from a Partial Sampling Frame 107
Using the Wizard to Sample from the First Partial Frame 107
Sample Results 120
Using the Wizard to Sample from the Second Partial Frame 121
Sample Results 125
Sampling With Probability Proportional to Size (PPS) 126
Using the Wizard 126
Plan Summary 138
Sampling Summary 138
Sample Results 140
Related Procedures 143

13 Complex Samples Analysis Preparation Wizard 144

Using the Complex Samples Analysis Preparation Wizard to Ready NHIS Public Data. 144
Using the Wizard 144
Summary. 147
Preparing for Analysis When Sampling Weights Are Not in the Data File. . . 147
Computing Inclusion Probabilities and Sampling Weights 148

Using the Wizard . . . 151
Summary . . . 158
Related Procedures . . . 158

14 Complex Samples Frequencies 159

Using Complex Samples Frequencies to Analyze Nutritional Supplement Usage . . . 159
Running the Analysis . . . 159
Frequency Table . . . 162
Frequency by Subpopulation . . . 163
Summary . . . 163
Related Procedures . . . 164

15 Complex Samples Descriptives 165

Using Complex Samples Descriptives to Analyze Activity Levels . . . 165
Running the Analysis . . . 165
Univariate Statistics . . . 168
Univariate Statistics by Subpopulation . . . 169
Summary . . . 170
Related Procedures . . . 170

16 Complex Samples Crosstabs 171

Using Complex Samples Crosstabs to Measure the Relative Risk of an Event . . . 171
Running the Analysis . . . 171
Crosstabulation . . . 175

Risk Estimate 175
Risk Estimate by Subpopulation 177
Summary 177
Related Procedures 177

17 Complex Samples Ratios 179

Using Complex Samples Ratios to Aid Property Value Assessment 179
Running the Analysis 179
Ratios 182
Pivoted Ratios Table 183
Summary 184
Related Procedures 184

18 Complex Samples General Linear Model 186

Using Complex Samples General Linear Model to Fit a Two-Factor ANOVA . 186
Running the Analysis 186
Model Summary 191
Tests of Model Effects 192
Parameter Estimates 193
Estimated Marginal Means 194
Summary 197
Related Procedures 197

19 Complex Samples Logistic Regression 199

Using Complex Samples Logistic Regression to Assess Credit Risk 199
Running the Analysis 199

Pseudo R-Squares 204
Classification 205
Tests of Model Effects 206
Parameter Estimates 207
Odds Ratios 208
Summary 209
Related Procedures 210

20 Complex Samples Ordinal Regression 211

Using Complex Samples Ordinal Regression to Analyze Survey Results. . . . 211
Running the Analysis 211
Pseudo R-Squares 217
Tests of Model Effects 218
Parameter Estimates 218
Classification 220
Odds Ratios 221
Generalized Cumulative Model 223
Dropping Non-Significant Predictors 224
Warnings 227
Comparing Models 228
Summary 228
Related Procedures 229

Appendices

Bibliography ***230***

Index ***231***

Part I:
User's Guide

Chapter

1

Introduction to SPSS Complex Samples Procedures

An inherent assumption of analytical procedures in traditional software packages is that the observations in a data file represent a simple random sample from the population of interest. This assumption is untenable for an increasing number of companies and researchers who find it both cost-effective and convenient to obtain samples in a more structured way.

The SPSS Complex Samples option allows you to select a sample according to a complex design and incorporate the design specifications into the data analysis, thus ensuring that your results are valid.

Properties of Complex Samples

A complex sample can differ from a simple random sample in many ways. In a simple random sample, individual sampling units are selected at random with equal probability and without replacement (WOR) directly from the entire population. By contrast, a given complex sample can have some or all of the following features:

Stratification. Stratified sampling involves selecting samples independently within non-overlapping subgroups of the population, or strata. For example, strata may be socioeconomic groups, job categories, age groups, or ethnic groups. With stratification, you can ensure adequate sample sizes for subgroups of interest, improve the precision of overall estimates, and use different sampling methods from stratum to stratum.

Clustering. Cluster sampling involves the selection of groups of sampling units, or clusters. For example, clusters may be schools, hospitals, or geographical areas, and sampling units may be students, patients, or citizens. Clustering is common in multistage designs and area (geographic) samples.

Multiple stages. In multistage sampling, you select a first-stage sample based on clusters. Then you create a second-stage sample by drawing subsamples from the selected clusters. If the second-stage sample is based on subclusters, you can then add a third stage to the sample. For example, in the first stage of a survey, a sample of cities could be drawn. Then, from the selected cities, households could be sampled. Finally, from the selected households, individuals could be polled. The Sampling and Analysis Preparation wizards allow you to specify three stages in a design.

Nonrandom sampling. When selection at random is difficult to obtain, units can be sampled systematically (at a fixed interval) or sequentially.

Unequal selection probabilities. When sampling clusters that contain unequal numbers of units, you can use probability-proportional-to-size (PPS) sampling to make a cluster's selection probability equal to the proportion of units it contains. PPS sampling can also use more general weighting schemes to select units.

Unrestricted sampling. Unrestricted sampling selects units with replacement (WR). Thus, an individual unit can be selected for the sample more than once.

Sampling weights. Sampling weights are automatically computed while drawing a complex sample and ideally correspond to the "frequency" that each sampling unit represents in the target population. Therefore, the sum of the weights over the sample should estimate the population size. Complex Samples analysis procedures require sampling weights in order to properly analyze a complex sample. Note that these weights should be used entirely within the Complex Samples option and should not be used with other analytical procedures via the Weight Cases procedure, which treats weights as case replications.

Usage of Complex Samples Procedures

Your usage of Complex Samples procedures depends on your particular needs. The primary types of users are those who:

- Plan and carry out surveys according to complex designs, possibly analyzing the sample later. The primary tool for surveyors is the Sampling Wizard.
- Analyze sample data files previously obtained according to complex designs. Before using the Complex Samples analysis procedures, you may need to use the Analysis Preparation Wizard.

Regardless of which type of user you are, you need to supply design information to Complex Samples procedures. This information is stored in a **plan file** for easy reuse.

Plan Files

A plan file contains complex sample specifications. There are two types of plan files:

Sampling plan. The specifications given in the Sampling Wizard define a sample design that is used to draw a complex sample. The sampling plan file contains those specifications. The sampling plan file also contains a default analysis plan that uses estimation methods suitable for the specified sample design.

Analysis plan. This plan file contains information needed by Complex Samples analysis procedures to properly compute variance estimates for a complex sample. The plan includes the sample structure, estimation methods for each stage, and references to required variables, such as sample weights. The Analysis Preparation Wizard allows you to create and edit analysis plans.

There are several advantages to saving your specifications in a plan file, including:

- A surveyor can specify the first stage of a multistage sampling plan and draw first-stage units now, collect information on sampling units for the second stage, and then modify the sampling plan to include the second stage.
- An analyst who doesn't have access to the sampling plan file can specify an analysis plan and refer to that plan from each Complex Samples analysis procedure.
- A designer of large-scale public use samples can publish the sampling plan file, which simplifies the instructions for analysts and avoids the need for each analyst to specify his or her own analysis plans.

Further Readings

For more information on sampling techniques, see the following texts:

Cochran, W. G. 1977. *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley and Sons.

Kish, L. 1987. *Statistical Design for Research*. New York: John Wiley and Sons.

Murthy, M. N. 1967. *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

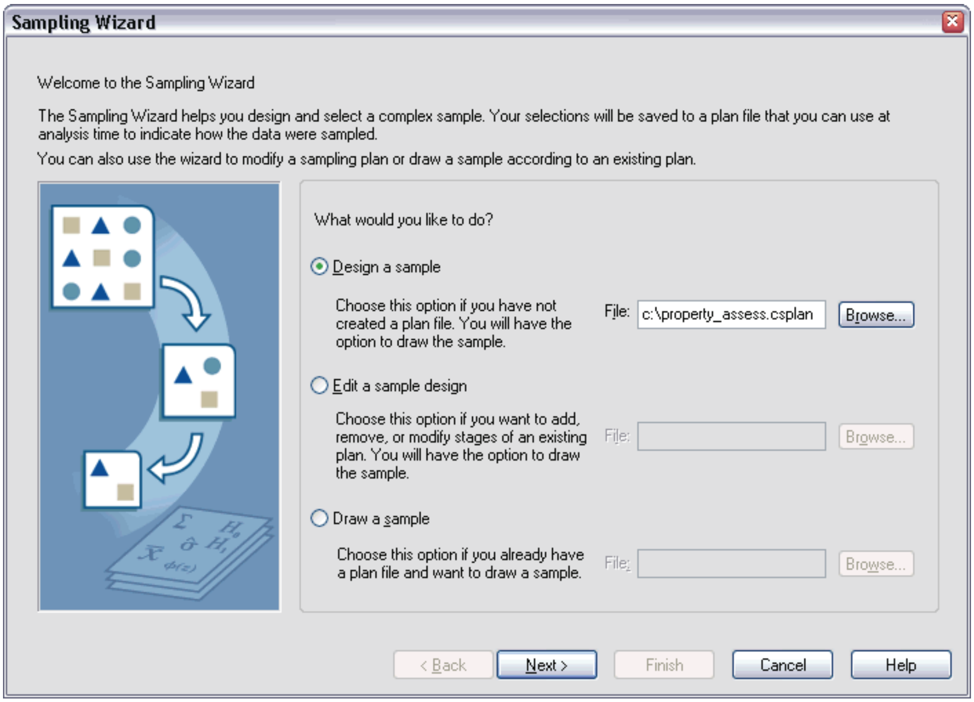
Särndal, C., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Chapter

2

Sampling from a Complex Design

Figure 2-1
Sampling Wizard, Welcome step



The Sampling Wizard guides you through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, you should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

Creating a New Sample Plan

▶ From the menus choose:

Analyze
  Complex Samples
    Select a Sample...

▶ Select Design a sample and choose a plan filename to save the sample plan.

▶ Click Next to continue through the Wizard.

▶ Optionally, in the Define Variables step, you can define strata, clusters, and input sample weights. After you define these, click Next.

▶ Optionally, in the Sampling Method step, you can choose a method for selecting items.

If you select PPS Brewer or PPS Murthy, you can click Finish to draw the sample. Otherwise, click Next and then:

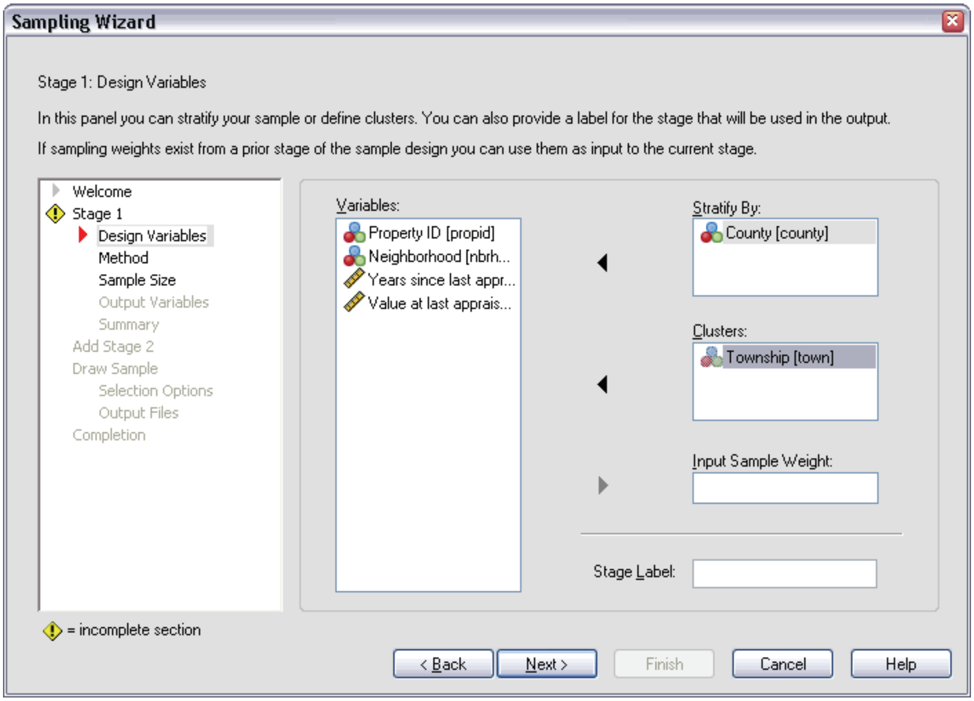
▶ In the Sample Size step, specify the number or proportion of units to sample.

You can now click Finish to draw the sample. Optionally, in further steps, you can:

- Choose output variables to save.
- Add a second or third stage to the design.
- Set various selection options, including which stages to draw samples from, the random number seed, and whether to treat user-missing values as valid values of design variables.
- Choose where to save output data.
- Paste your selections as command syntax.

Sampling Wizard: Design Variables

Figure 2-2
Sampling Wizard, Design Variables step



This step allows you to select stratification and clustering variables and to define input sample weights. You can also specify a label for the stage.

Stratify By. The cross-classification of stratification variables defines distinct subpopulations, or strata. Separate samples are obtained for each stratum. To improve the precision of your estimates, units within strata should be as homogeneous as possible for the characteristics of interest.

Clusters. Cluster variables define groups of observational units, or clusters. Clusters are useful when directly sampling observational units from the population is expensive or impossible; instead, you can sample clusters from the population and then sample observational units from the selected clusters. However, the use of clusters can introduce correlations among sampling units, resulting in a loss of precision. To minimize this effect, units within clusters should be as heterogeneous as possible for the characteristics of interest. You must define at least one cluster variable in order to plan a multistage design. Clusters are also necessary in the use of several different sampling methods. For more information, see Sampling Wizard: Sampling Method on p. 8.

Input Sample Weight. If the current sample design is part of a larger sample design, you may have sample weights from a previous stage of the larger design. You can specify a numeric variable containing these weights in the first stage of the current design. Sample weights are computed automatically for subsequent stages of the current design.

Stage Label. You can specify an optional string label for each stage. This is used in the output to help identify stagewise information.

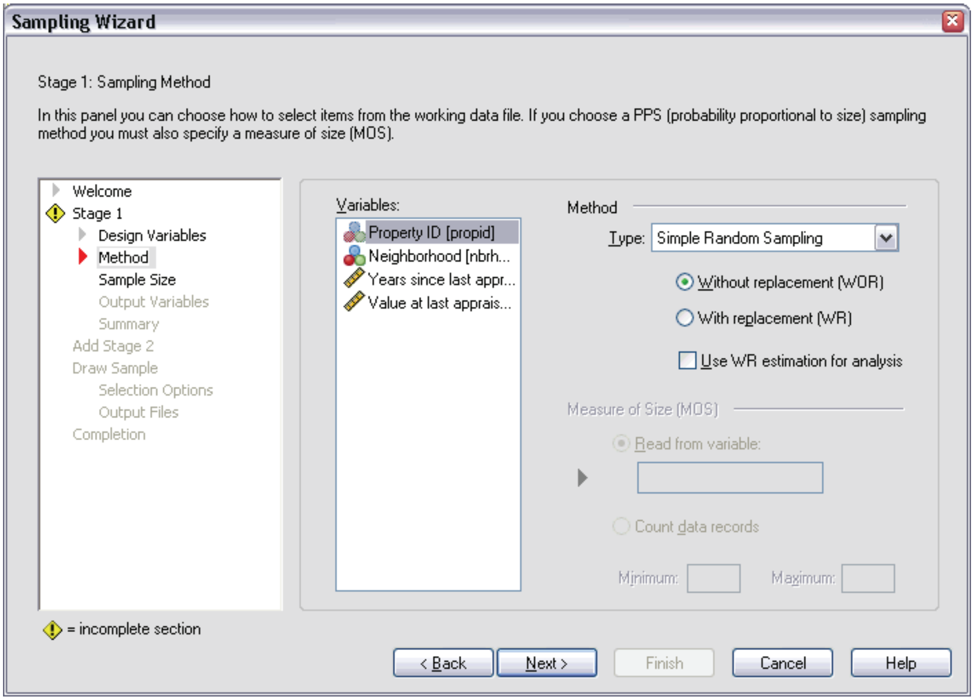
Note: The source variable list has the same content across steps of the Wizard. In other words, variables removed from the source list in a particular step are removed from the list in all steps. Variables returned to the source list appear in the list in all steps.

Tree Controls for Navigating the Sampling Wizard

On the left side of each step in the Sampling Wizard is an outline of all the steps. You can navigate the Wizard by clicking on the name of an enabled step in the outline. Steps are enabled as long as all previous steps are valid—that is, if each previous step has been given the minimum required specifications for that step. See the Help for individual steps for more information on why a given step may be invalid.

Sampling Wizard: Sampling Method

Figure 2-3
Sampling Wizard, Method step



This step allows you to specify how to select cases from the active dataset.

Method. Controls in this group are used to choose a selection method. Some sampling types allow you to choose whether to sample with replacement (WR) or without replacement (WOR). See the type descriptions for more information. Note that some probability-proportional-to-size (PPS) types are available only when clusters have been defined and that all PPS types are available only in the first stage of a design. Moreover, WR methods are available only in the last stage of a design.

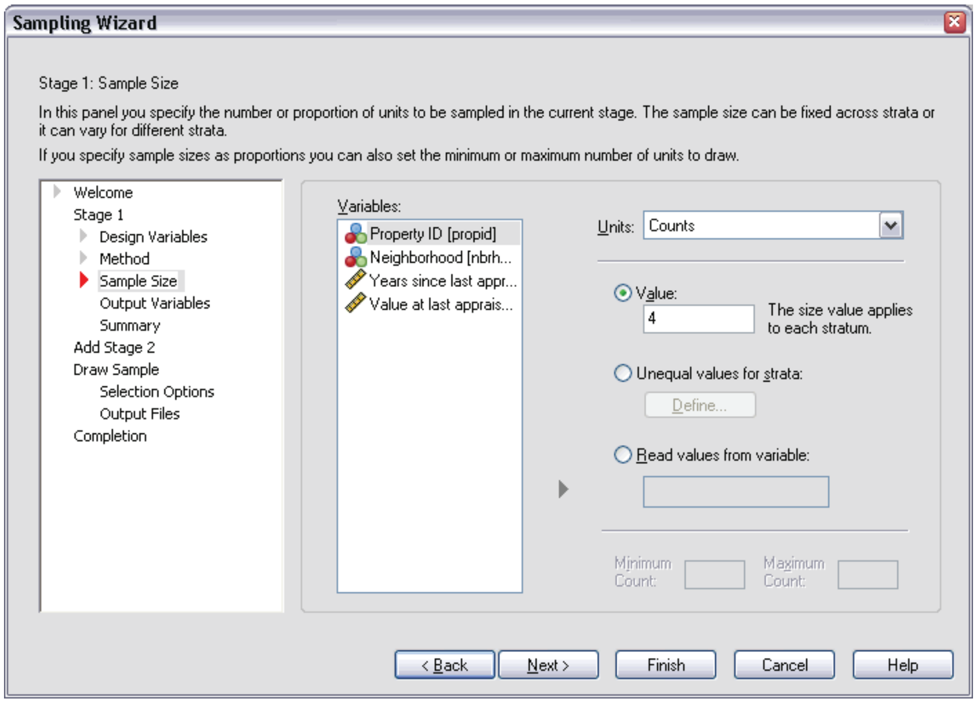
- **Simple Random Sampling.** Units are selected with equal probability. They can be selected with or without replacement.
- **Simple Systematic.** Units are selected at a fixed interval throughout the sampling frame (or strata, if they have been specified) and extracted without replacement. A randomly selected unit within the first interval is chosen as the starting point.

- **Simple Sequential.** Units are selected sequentially with equal probability and without replacement.
- **PPS.** This is a first-stage method that selects units at random with probability proportional to size. Any units can be selected with replacement; only clusters can be sampled without replacement.
- **PPS Systematic.** This is a first-stage method that systematically selects units with probability proportional to size. They are selected without replacement.
- **PPS Sequential.** This is a first-stage method that sequentially selects units with probability proportional to cluster size and without replacement.
- **PPS Brewer.** This is a first-stage method that selects two clusters from each stratum with probability proportional to cluster size and without replacement. A cluster variable must be specified to use this method.
- **PPS Murthy.** This is a first-stage method that selects two clusters from each stratum with probability proportional to cluster size and without replacement. A cluster variable must be specified to use this method.
- **PPS Sampford.** This is a first-stage method that selects more than two clusters from each stratum with probability proportional to cluster size and without replacement. It is an extension of Brewer's method. A cluster variable must be specified to use this method.
- **Use WR estimation for analysis.** By default, an estimation method is specified in the plan file that is consistent with the selected sampling method. This allows you to use with-replacement estimation even if the sampling method implies WOR estimation. This option is available only in stage 1.

Measure of Size (MOS). If a PPS method is selected, you must specify a measure of size that defines the size of each unit. These sizes can be explicitly defined in a variable or they can be computed from the data. Optionally, you can set lower and upper bounds on the MOS, overriding any values found in the MOS variable or computed from the data. These options are available only in stage 1.

Sampling Wizard: Sample Size

Figure 2-4
Sampling Wizard, Sample Size step



This step allows you to specify the number or proportion of units to sample within the current stage. The sample size can be fixed or it can vary across strata. For the purpose of specifying sample size, clusters chosen in previous stages can be used to define strata.

Units. You can specify an exact sample size or a proportion of units to sample.

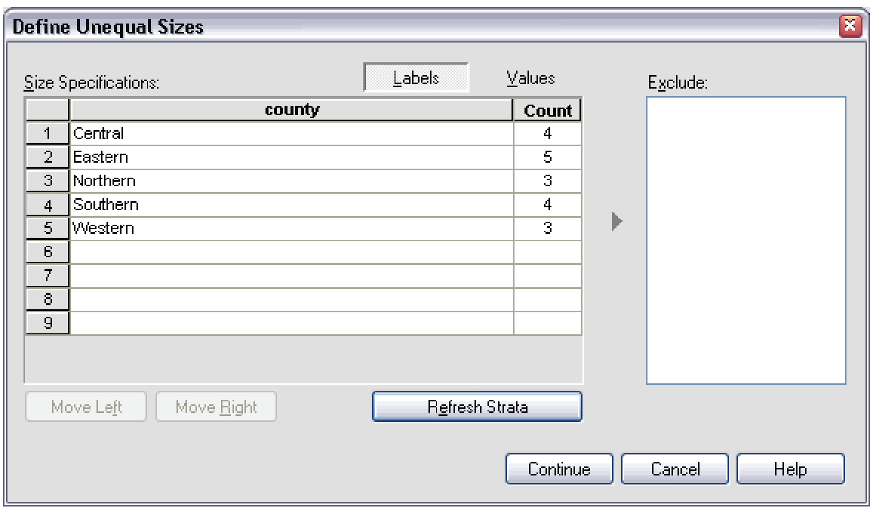
- **Value.** A single value is applied to all strata. If Counts is selected as the unit metric, you should enter a positive integer. If Proportions is selected, you should enter a non-negative value. Unless sampling with replacement, proportion values should also be no greater than 1.

- **Unequal values for strata.** Allows you to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **Read values from variable.** Allows you to select a numeric variable that contains size values for strata.

If Proportions is selected, you have the option to set lower and upper bounds on the number of units sampled.

Define Unequal Sizes

Figure 2-5
Define Unequal Sizes dialog box



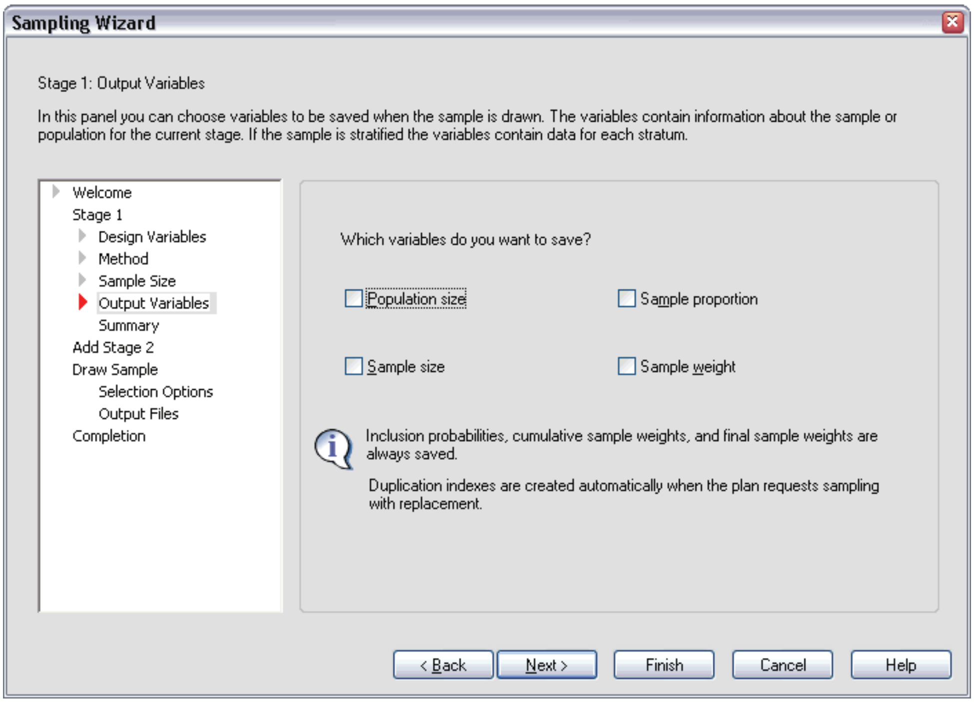
The Define Unequal Sizes dialog box allows you to enter sizes on a per-stratum basis.

Size Specifications grid. The grid displays the cross-classifications of up to five strata or cluster variables—one stratum/cluster combination per row. Eligible grid variables include all stratification variables from the current and previous stages and all cluster variables from previous stages. Variables can be reordered within the grid or moved to the Exclude list. Enter sizes in the rightmost column. Click Labels or Values to toggle the display of value labels and data values for stratification and cluster variables in the grid cells. Cells that contain unlabeled values always show values. Click Refresh Strata to repopulate the grid with each combination of labeled data values for variables in the grid.

Exclude. To specify sizes for a subset of stratum/cluster combinations, move one or more variables to the Exclude list. These variables are not used to define sample sizes.

Sampling Wizard: Output Variables

Figure 2-6
Sampling Wizard, Output Variables step



This step allows you to choose variables to save when the sample is drawn.

Population size. The estimated number of units in the population for a given stage. The root name for the saved variable is *PopulationSize_*.

Sample proportion. The sampling rate at a given stage. The root name for the saved variable is *SamplingRate_*.

Sample size. The number of units drawn at a given stage. The root name for the saved variable is *SampleSize_*.

Sample weight. The inverse of the inclusion probabilities. The root name for the saved variable is *SampleWeight_*.

Some stagewise variables are generated automatically. These include:

Inclusion probabilities. The proportion of units drawn at a given stage. The root name for the saved variable is *InclusionProbability_*.

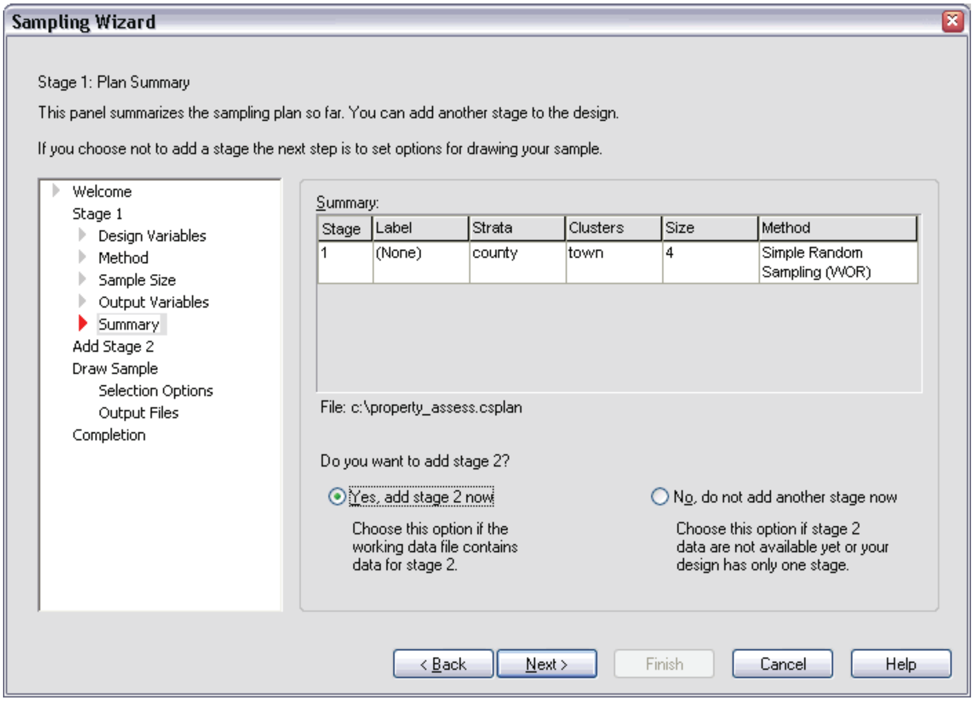
Cumulative weight. The cumulative sample weight over stages previous to and including the current one. The root name for the saved variable is *SampleWeightCumulative_*.

Index. Identifies units selected multiple times within a given stage. The root name for the saved variable is *Index_*.

Note: Saved variable root names include an integer suffix that reflects the stage number—for example, *PopulationSize_1_* for the saved population size for stage 1.

Sampling Wizard: Plan Summary

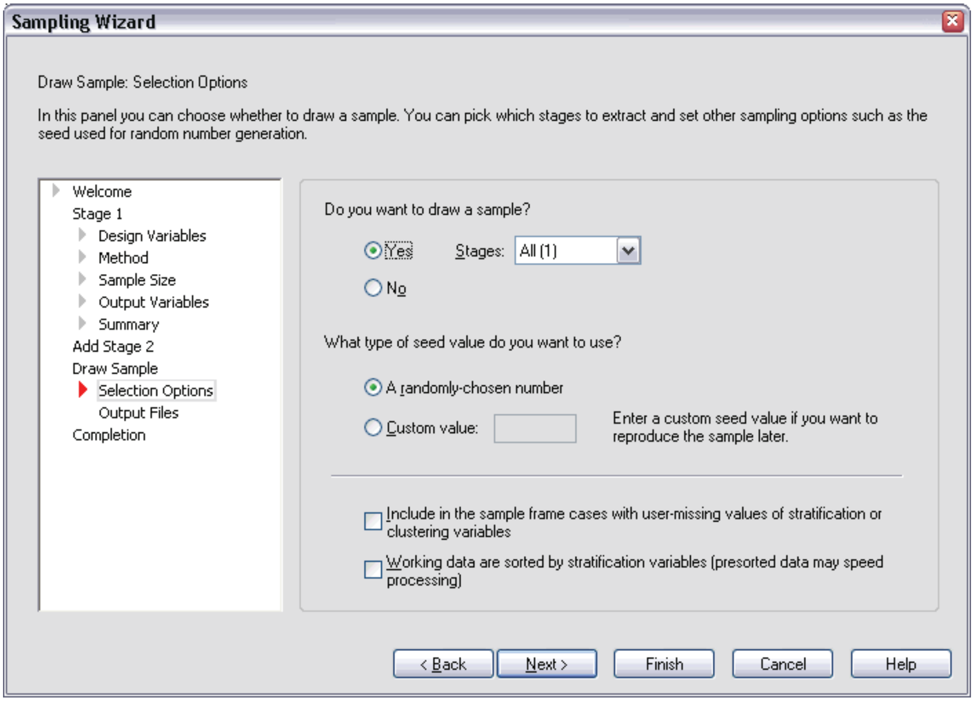
Figure 2-7
Sampling Wizard, Plan Summary step



This is the last step within each stage, providing a summary of the sample design specifications through the current stage. From here, you can either proceed to the next stage (creating it, if necessary) or set options for drawing the sample.

Sampling Wizard: Draw Sample Selection Options

Figure 2-8
Sampling Wizard, Draw Sample Selection Options step



This step allows you to choose whether to draw a sample. You can also control other sampling options, such as the random seed and missing-value handling.

Draw sample. In addition to choosing whether to draw a sample, you can also choose to execute part of the sampling design. Stages must be drawn in order—that is, stage 2 cannot be drawn unless stage 1 is also drawn. When editing or executing a plan, you cannot resample locked stages.

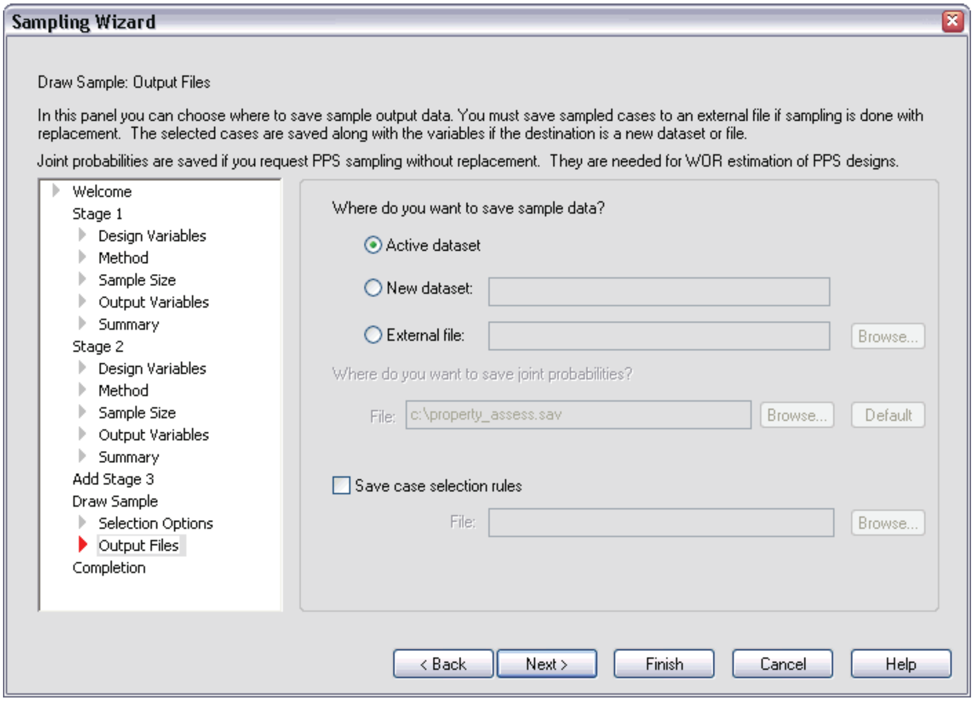
Seed. This allows you to choose a seed value for random number generation.

Include user-missing values. This determines whether user-missing values are valid. If so, user-missing values are treated as a separate category.

Data already sorted. If your sample frame is presorted by the values of the stratification variables, this option allows you to speed the selection process.

Sampling Wizard: Draw Sample Output Files

Figure 2-9
Sampling Wizard, Draw Sample Output Files step



This step allows you to choose where to direct sampled cases, weight variables, joint probabilities, and case selection rules.

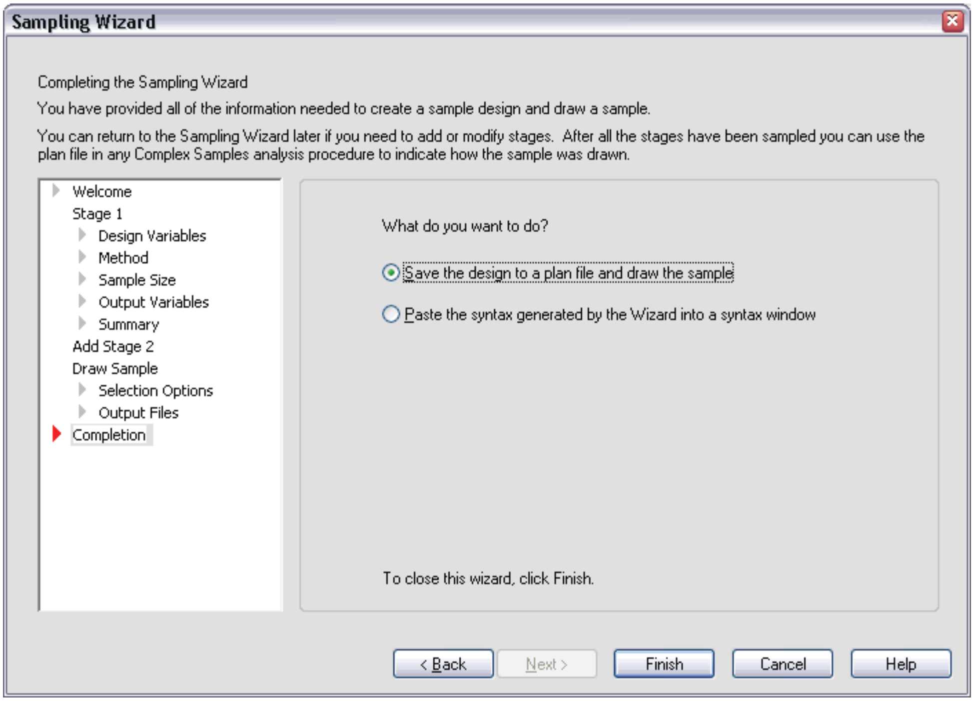
Sample data. These options let you determine where sample output is written. It can be added to the active dataset, written to a new dataset, or saved to an external SPSS-format data file. Datasets are available during the current session but are not available in subsequent sessions unless you explicitly save them as data files. Dataset names must adhere to SPSS variable naming rules. If an external file or new dataset is specified, the sampling output variables and variables in the active dataset for the selected cases are written.

Joint probabilities. These options let you determine where joint probabilities are written. They are saved to an external SPSS-format data file. Joint probabilities are produced if the PPS WOR, PPS Brewer, PPS Sampford, or PPS Murthy method is selected and WR estimation is not specified.

Case selection rules. If you are constructing your sample one stage at a time, you may want to save the case selection rules to a text file. They are useful for constructing the subframe for subsequent stages.

Sampling Wizard: Finish

Figure 2-10
Sampling Wizard, Finish step



This is the final step. You can save the plan file and draw the sample now or paste your selections into a syntax window.

When making changes to stages in the existing plan file, you can save the edited plan to a new file or overwrite the existing file. When adding stages without making changes to existing stages, the Wizard automatically overwrites the existing plan file. If you want to save the plan to a new file, select Paste the syntax generated by the Wizard into a syntax window and change the filename in the syntax commands.

Modifying an Existing Sample Plan

▶ From the menus choose:

Analyze
  Complex Samples
    Select a Sample...

▶ Select Edit a sample design and choose a plan file to edit.

▶ Click Next to continue through the Wizard.

▶ Review the sampling plan in the Plan Summary step, and then click Next.

Subsequent steps are largely the same as for a new design. See the Help for individual steps for more information.

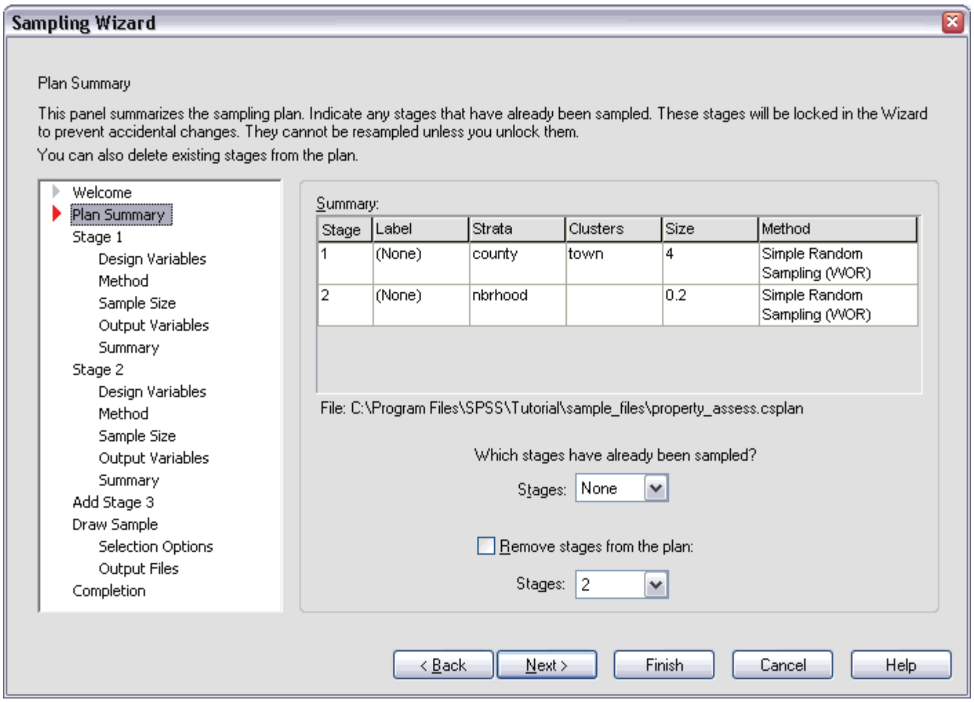
▶ Navigate to the Finish step, and specify a new name for the edited plan file or choose to overwrite the existing plan file.

Optionally, you can:

- Specify stages that have already been sampled.
- Remove stages from the plan.

Sampling Wizard: Plan Summary

Figure 2-11
Sampling Wizard, Plan Summary step



This step allows you to review the sampling plan and indicate stages that have already been sampled. If editing a plan, you can also remove stages from the plan.

Previously sampled stages. If an extended sampling frame is not available, you will have to execute a multistage sampling design one stage at a time. Select which stages have already been sampled from the drop-down list. Any stages that have been executed are locked; they are not available in the Draw Sample Selection Options step, and they cannot be altered when editing a plan.

Remove stages. You can remove stages 2 and 3 from a multistage design.

Running an Existing Sample Plan

- From the menus choose:

  Analyze
    Complex Samples
      Select a Sample...

- Select Draw a sample and choose a plan file to run.
- Click Next to continue through the Wizard.
- Review the sampling plan in the Plan Summary step, and then click Next.
- The individual steps containing stage information are skipped when executing a sample plan. You can now go on to the Finish step at any time.

Optionally, you can:

- Specify stages that have already been sampled.

CSPLAN and CSSELECT Commands Additional Features

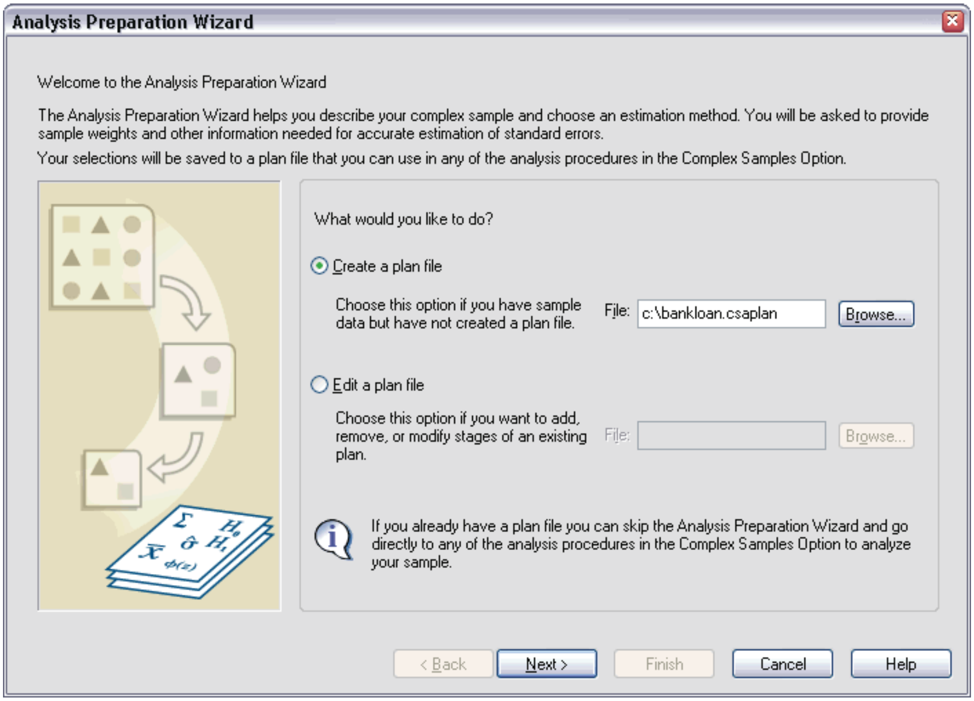
The SPSS command language also allows you to:

- Specify custom names for output variables.
- Control the output in the Viewer. For example, you can suppress the stagewise summary of the plan that is displayed if a sample is designed or modified, suppress the summary of the distribution of sampled cases by strata that is shown if the sample design is executed, and request a case processing summary.
- Choose a subset of variables in the active dataset to write to an external sample file or to a different dataset.

See the *SPSS Command Syntax Reference* for complete syntax information.

Chapter
3
Preparing a Complex Sample for Analysis

Figure 3-1
Analysis Preparation Wizard, Welcome step



The Analysis Preparation Wizard guides you through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. Before using the Wizard, you should have a sample drawn according to a complex design.

Creating a new plan is most useful when you do not have access to the sampling plan file used to draw the sample (recall that the sampling plan contains a default analysis plan). If you do have access to the sampling plan file used to draw the sample, you can use the default analysis plan contained in the sampling plan file or override the default analysis specifications and save your changes to a new file.

Creating a New Analysis Plan

▶ From the menus choose:

Analyze
  Complex Samples
    Prepare for Analysis...

▶ Select Create a plan file, and choose a plan filename to which you will save the analysis plan.

▶ Click Next to continue through the Wizard.

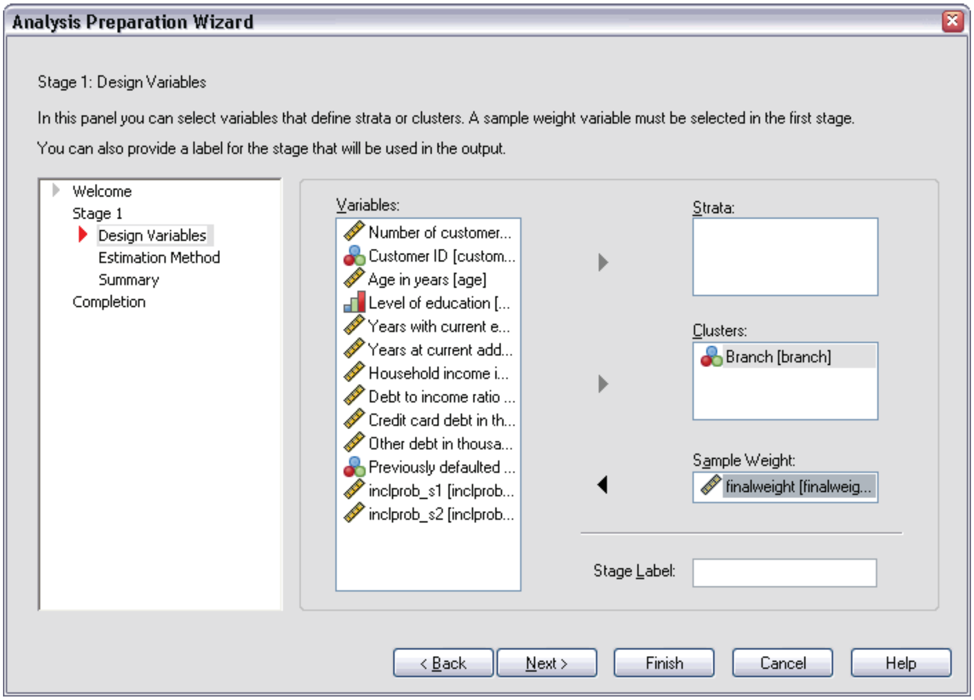
▶ Specify the variable containing sample weights in the Design Variables step, optionally defining strata and clusters.

You can now click Finish to save the plan. Optionally, in further steps you can:

- Select the method for estimating standard errors in the Estimation Method step.
- Specify the number of units sampled or the inclusion probability per unit in the Size step.
- Add a second or third stage to the design.
- Paste your selections as command syntax.

Analysis Preparation Wizard: Design Variables

Figure 3-2
Analysis Preparation Wizard, Design Variables step



This step allows you to identify the stratification and clustering variables and define sample weights. You can also provide a label for the stage.

Strata. The cross-classification of stratification variables defines distinct subpopulations, or strata. Your total sample represents the combination of independent samples from each stratum.

Clusters. Cluster variables define groups of observational units, or clusters. Samples drawn in multiple stages select clusters in the earlier stages and then subsample units from the selected clusters. When analyzing a data file obtained by sampling clusters with replacement, you should include the duplication index as a cluster variable.

Sample Weight. You must provide sample weights in the first stage. Sample weights are computed automatically for subsequent stages of the current design.

Stage Label. You can specify an optional string label for each stage. This is used in the output to help identify stagewise information.

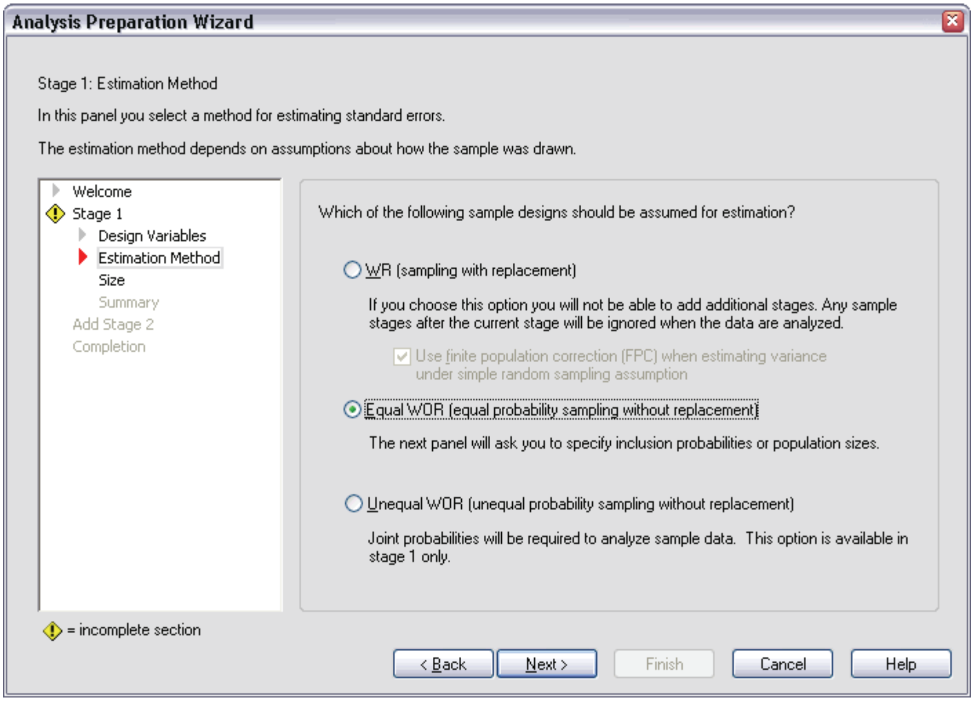
Note: The source variable list has the same contents across steps of the Wizard. In other words, variables removed from the source list in a particular step are removed from the list in all steps. Variables returned to the source list show up in all steps.

Tree Controls for Navigating the Analysis Wizard

At the left side of each step of the Analysis Wizard is an outline of all the steps. You can navigate the Wizard by clicking on the name of an enabled step in the outline. Steps are enabled as long as all previous steps are valid—that is, as long as each previous step has been given the minimum required specifications for that step. For more information on why a given step may be invalid, see the Help for individual steps.

Analysis Preparation Wizard: Estimation Method

Figure 3-3
Analysis Preparation Wizard, Estimation Method step



This step allows you to specify an estimation method for the stage.

WR (sampling with replacement). WR estimation does not include a correction for sampling from a finite population (FPC) when estimating the variance under the complex sampling design. You can choose to include or exclude the FPC when estimating the variance under simple random sampling (SRS).

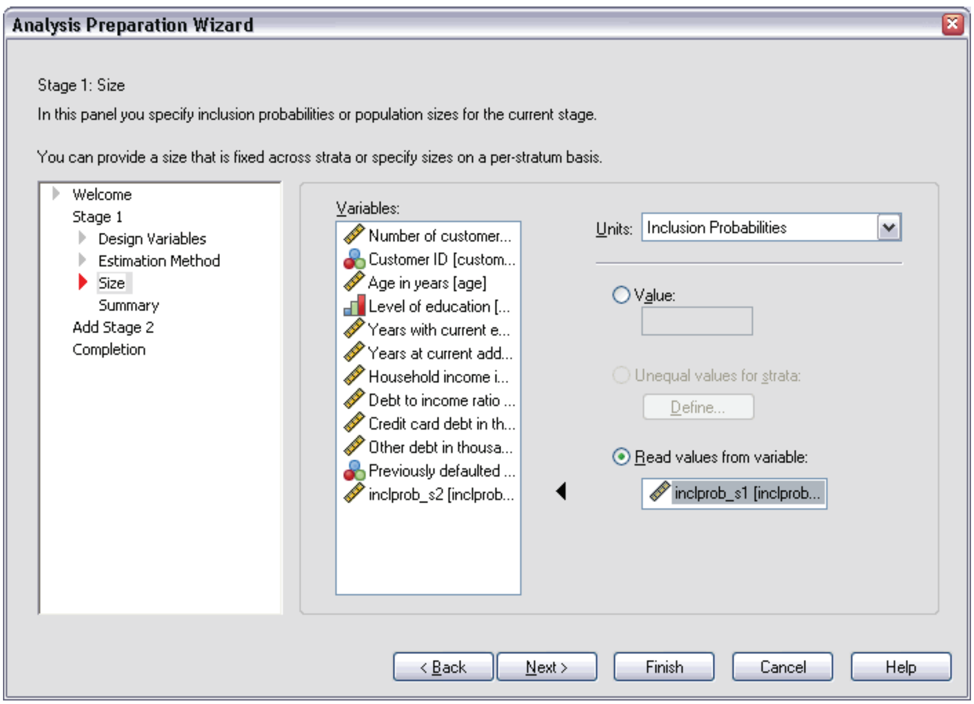
Choosing not to include the FPC for SRS variance estimation is recommended when the analysis weights have been scaled so that they do not add up to the population size. The SRS variance estimate is used in computing statistics like the design effect. WR estimation can be specified only in the final stage of a design; the Wizard will not allow you to add another stage if you select WR estimation.

Equal WOR (equal probability sampling without replacement). Equal WOR estimation includes the finite population correction and assumes that units are sampled with equal probability. Equal WOR can be specified in any stage of a design.

Unequal WOR (unequal probability sampling without replacement). In addition to using the finite population correction, Unequal WOR accounts for sampling units (usually clusters) selected with unequal probability. This estimation method is available only in the first stage.

Analysis Preparation Wizard: Size

Figure 3-4
Analysis Preparation Wizard, Size step



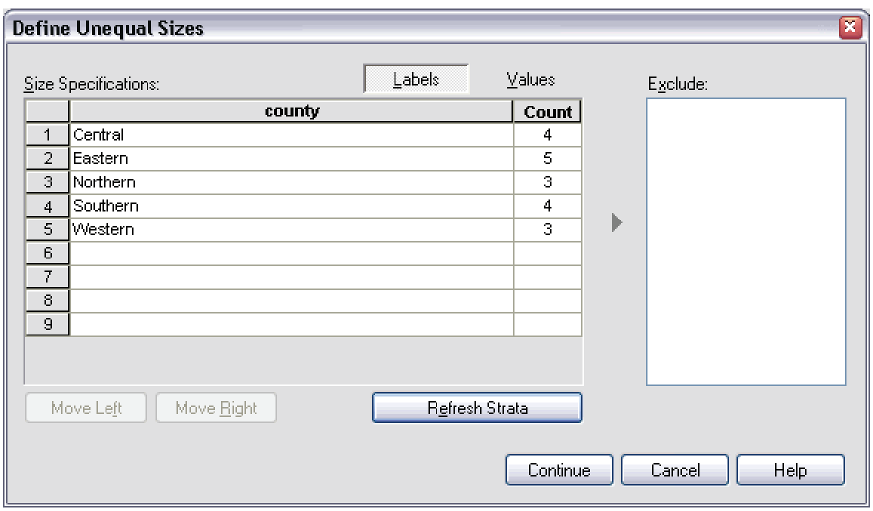
This step is used to specify inclusion probabilities or population sizes for the current stage. Sizes can be fixed or can vary across strata. For the purpose of specifying sizes, clusters specified in previous stages can be used to define strata. Note that this step is necessary only when Equal WOR is chosen as the Estimation Method.

Units. You can specify exact population sizes or the probabilities with which units were sampled.

- **Value.** A single value is applied to all strata. If Population Sizes is selected as the unit metric, you should enter a non-negative integer. If Inclusion Probabilities is selected, you should enter a value between 0 and 1, inclusive.
- **Unequal values for strata.** Allows you to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **Read values from variable.** Allows you to select a numeric variable that contains size values for strata.

Define Unequal Sizes

Figure 3-5
Define Unequal Sizes dialog box



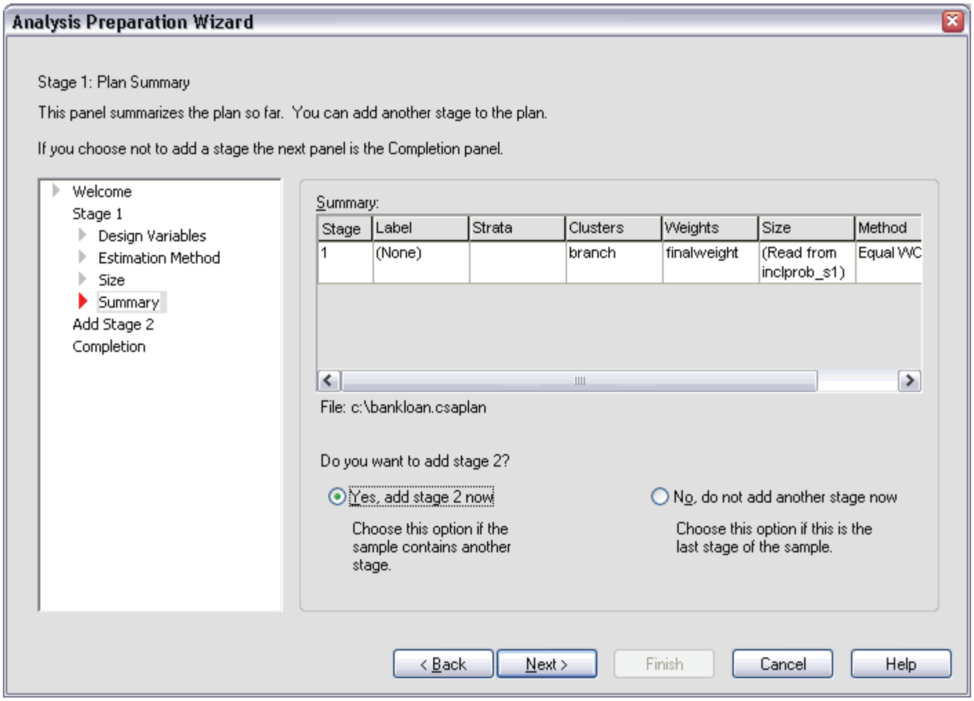
The Define Unequal Sizes dialog box allows you to enter sizes on a per-stratum basis.

Size Specifications grid. The grid displays the cross-classifications of up to five strata or cluster variables—one stratum/cluster combination per row. Eligible grid variables include all stratification variables from the current and previous stages and all cluster variables from previous stages. Variables can be reordered within the grid or moved to the Exclude list. Enter sizes in the rightmost column. Click Labels or Values to toggle the display of value labels and data values for stratification and cluster variables in the grid cells. Cells that contain unlabeled values always show values. Click **Refresh Strata** to repopulate the grid with each combination of labeled data values for variables in the grid.

Exclude. To specify sizes for a subset of stratum/cluster combinations, move one or more variables to the Exclude list. These variables are not used to define sample sizes.

Analysis Preparation Wizard: Plan Summary

Figure 3-6
Analysis Preparation Wizard, Plan Summary step



This is the last step within each stage, providing a summary of the analysis design specifications through the current stage. From here, you can either proceed to the next stage (creating it if necessary) or save the analysis specifications.

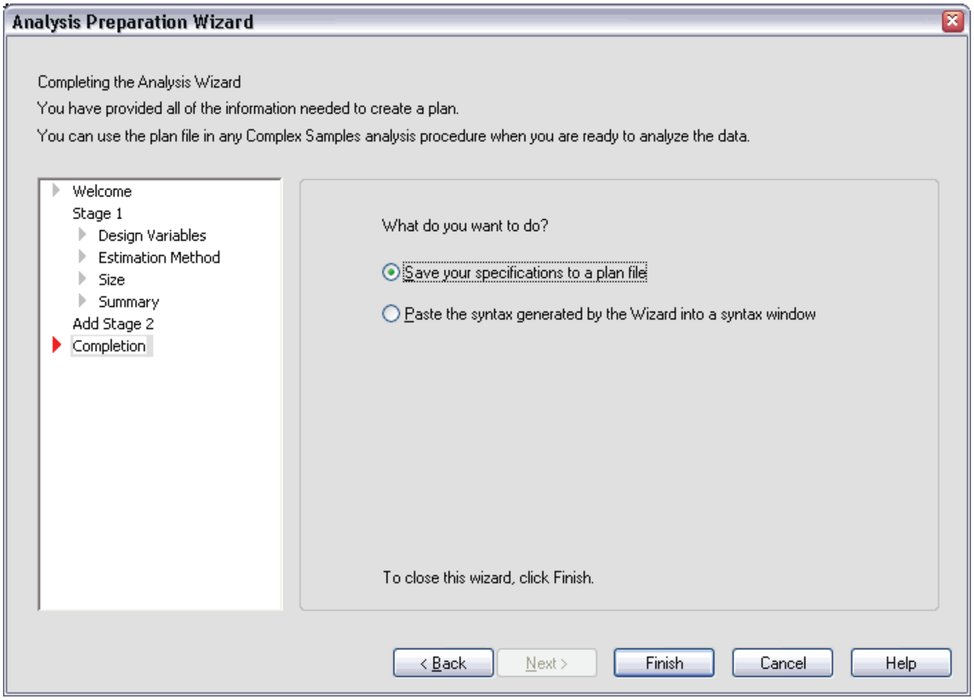
If you cannot add another stage, it is likely because:

- No cluster variable was specified in the Design Variables step.

- You selected WR estimation in the Estimation Method step.
- This is the third stage of the analysis, and the Wizard supports a maximum of three stages.

Analysis Preparation Wizard: Finish

Figure 3-7
Analysis Preparation Wizard, Finish step



This is the final step. You can save the plan file now or paste your selections to a syntax window.

When making changes to stages in the existing plan file, you can save the edited plan to a new file or overwrite the existing file. When adding stages without making changes to existing stages, the Wizard automatically overwrites the existing plan file. If you want to save the plan to a new file, choose to Paste the syntax generated by the Wizard into a syntax window and change the filename in the syntax commands.

Modifying an Existing Analysis Plan

▶ From the menus choose:

Analyze
  Complex Samples
    Prepare for Analysis...

▶ Select Edit a plan file, and choose a plan filename to which you will save the analysis plan.

▶ Click Next to continue through the Wizard.

▶ Review the analysis plan in the Plan Summary step, and then click Next.

Subsequent steps are largely the same as for a new design. For more information, see the Help for individual steps.

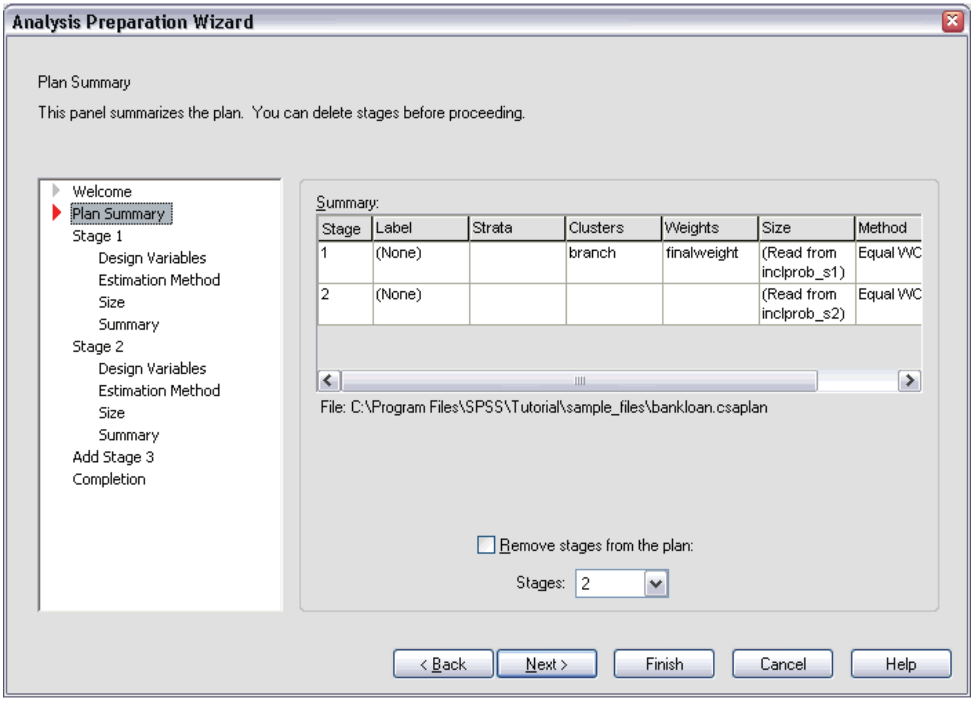
▶ Navigate to the Finish step, and specify a new name for the edited plan file, or choose to overwrite the existing plan file.

Optionally, you can:

- Remove stages from the plan.

Analysis Preparation Wizard: Plan Summary

Figure 3-8
Analysis Preparation Wizard, Plan Summary step



This step allows you to review the analysis plan and remove stages from the plan.

Remove Stages. You can remove stages 2 and 3 from a multistage design. Since a plan must have at least one stage, you can edit but not remove stage 1 from the design.

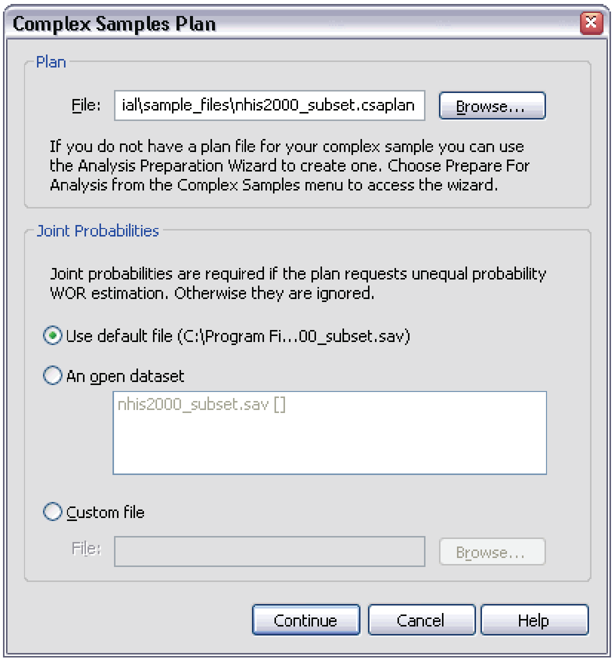
Chapter

4

Complex Samples Plan

Complex Samples analysis procedures require analysis specifications from an analysis or sample plan file in order to provide valid results.

Figure 4-1
Complex Samples Plan dialog box



Plan. Specify the path of an analysis or sample plan file.

Joint Probabilities. In order to use Unequal WOR estimation for clusters drawn using a PPS WOR method, you need to specify a separate file or an open dataset containing the joint probabilities. This file or dataset is created by the Sampling Wizard during sampling.

Chapter

5

Complex Samples Frequencies

The Complex Samples Frequencies procedure produces frequency tables for selected variables and displays univariate statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Example. Using the Complex Samples Frequencies procedure, you can obtain univariate tabular statistics for vitamin usage among U.S. citizens, based on the results of the National Health Interview Survey (NHIS) and with an appropriate analysis plan for this public-use data.

Statistics. The procedure produces estimates of cell population sizes and table percentages, plus standard errors, confidence intervals, coefficients of variation, design effects, square roots of design effects, cumulative values, and unweighted counts for each estimate. Additionally, chi-square and likelihood-ratio statistics are computed for the test of equal cell proportions.

Data. Variables for which frequency tables are produced should be categorical. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Frequencies

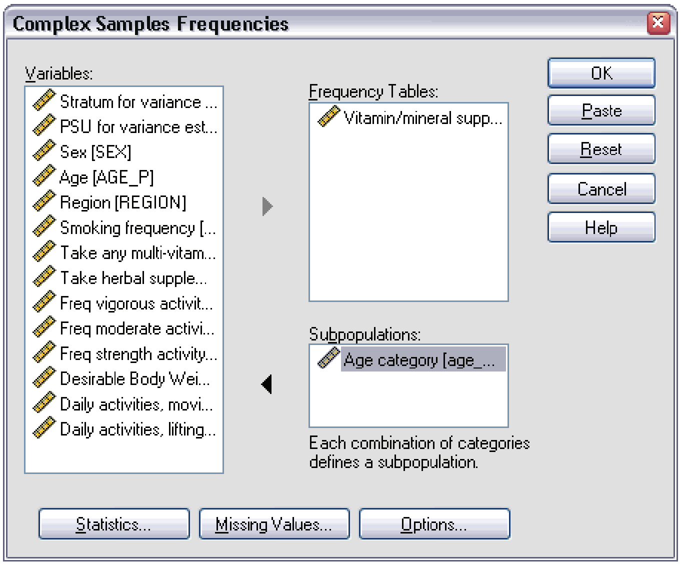
▶ From the menus choose:

Analyze
  Complex Samples
    Frequencies...

▶ Select a plan file and optionally select a custom joint probabilities file.

▶ Click Continue.

Figure 5-1
Frequencies dialog box



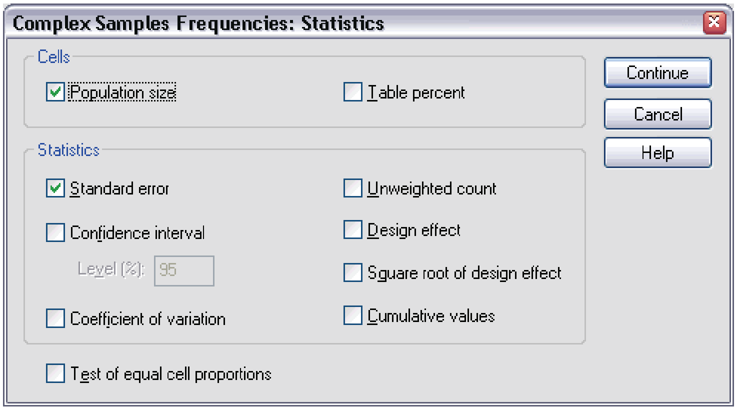
▶ Select at least one frequency variable.

Optionally, you can:

- Specify variables to define subpopulations. Statistics are computed separately for each subpopulation.

Complex Samples Frequencies Statistics

Figure 5-2
Frequencies Statistics dialog box



Cells. This group allows you to request estimates of the cell population sizes and table percentages.

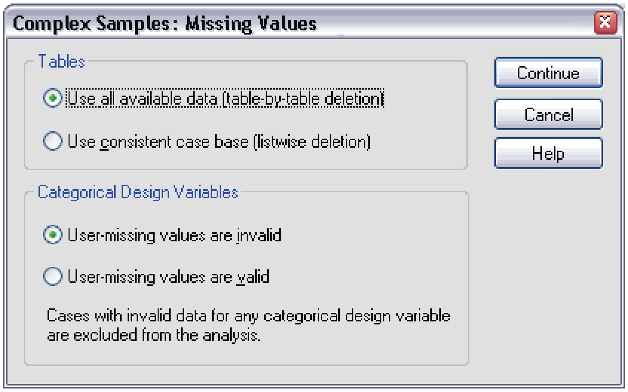
Statistics. This group produces statistics associated with the population size or table percentage.

- **Standard error.** The standard error of the estimate.
- **Confidence interval.** A confidence interval for the estimate, using the specified level.
- **Coefficient of variation.** The ratio of the standard error of the estimate to the estimate.
- **Unweighted count.** The number of units used to compute the estimate.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Cumulative values.** The cumulative estimate through each value of the variable.

Test of equal cell proportions. This produces chi-square and likelihood-ratio tests of the hypothesis that the categories of a variable have equal frequencies. Separate tests are performed for each variable.

Complex Samples Missing Values

Figure 5-3
Missing Values dialog box



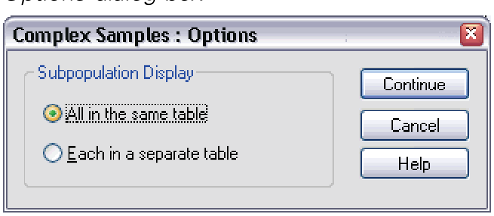
Tables. This group determines which cases are used in the analysis.

- **Use all available data.** Missing values are determined on a table-by-table basis. Thus, the cases used to compute statistics may vary across frequency or crosstabulation tables.
- **Use consistent case base.** Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent across tables.

Categorical Design Variables. This group determines whether user-missing values are valid or invalid.

Complex Samples Options

Figure 5-4
Options dialog box



Subpopulation Display. You can choose to have subpopulations displayed in the same table or in separate tables.

Chapter

6

Complex Samples Descriptives

The Complex Samples Descriptives procedure displays univariate summary statistics for several variables. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Example. Using the Complex Samples Descriptives procedure, you can obtain univariate descriptive statistics for the activity levels of U.S. citizens, based on the results of the National Health Interview Survey (NHIS) and with an appropriate analysis plan for this public use data.

Statistics. The procedure produces means and sums, plus *t* tests, standard errors, confidence intervals, coefficients of variation, unweighted counts, population sizes, design effects, and square roots of design effects for each estimate.

Data. Measures should be scale variables. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Descriptives

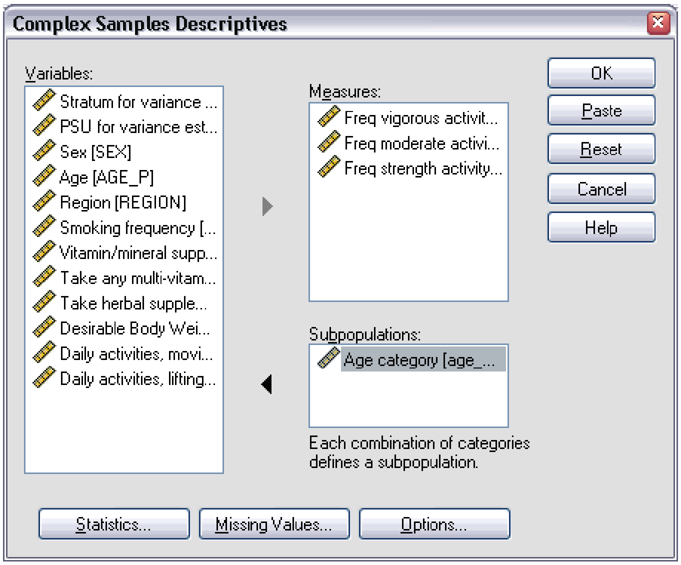
▶ From the menus choose:

Analyze
  Complex Samples
    Descriptives...

▶ Select a plan file, and optionally select a custom joint probabilities file.

▶ Click Continue.

Figure 6-1
Descriptives dialog box



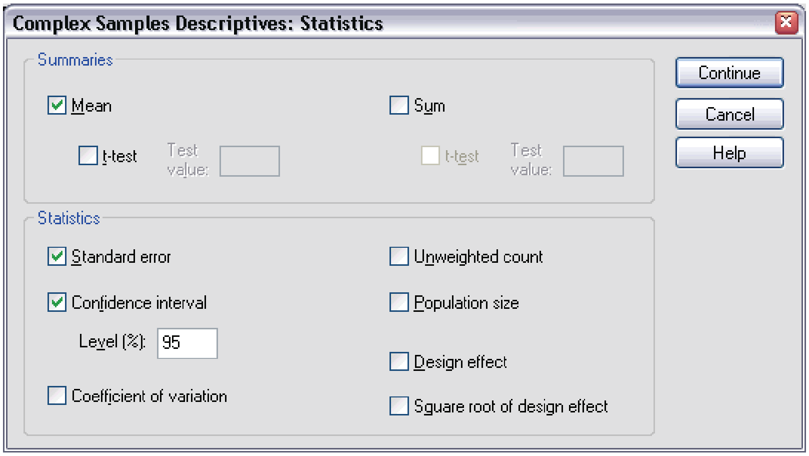
▶ Select at least one measure variable.

Optionally, you can:

- Specify variables to define subpopulations. Statistics are computed separately for each subpopulation.

Complex Samples Descriptives Statistics

Figure 6-2
Descriptives Statistics dialog box



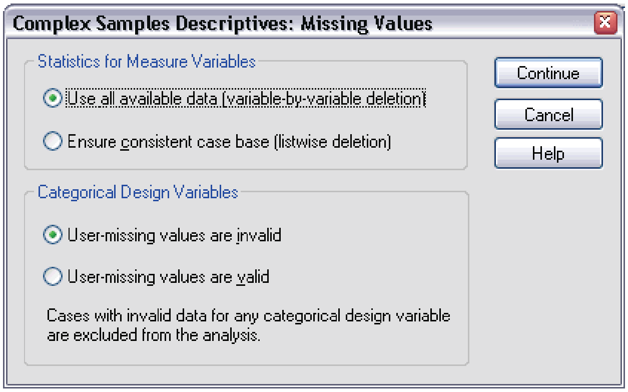
Summaries. This group allows you to request estimates of the means and sums of the measure variables. Additionally, you can request *t* tests of the estimates against a specified value.

Statistics. This group produces statistics associated with the mean or sum.

- **Standard error.** The standard error of the estimate.
- **Confidence interval.** A confidence interval for the estimate, using the specified level.
- **Coefficient of variation.** The ratio of the standard error of the estimate to the estimate.
- **Unweighted count.** The number of units used to compute the estimate.
- **Population size.** The estimated number of units in the population.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

Complex Samples Descriptives Missing Values

Figure 6-3
Descriptives Missing Values dialog box



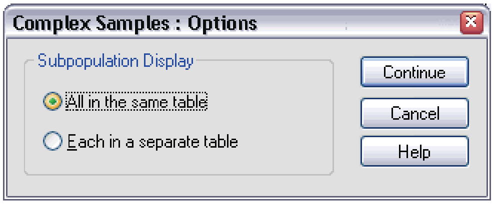
Statistics for Measure Variables. This group determines which cases are used in the analysis.

- **Use all available data.** Missing values are determined on a variable-by-variable basis, thus the cases used to compute statistics may vary across measure variables.
- **Ensure consistent case base.** Missing values are determined across all variables, thus the cases used to compute statistics are consistent.

Categorical Design Variables. This group determines whether user-missing values are valid or invalid.

Complex Samples Options

Figure 6-4
Options dialog box



Subpopulation Display. You can choose to have subpopulations displayed in the same table or in separate tables.

Chapter

7

Complex Samples Crosstabs

The Complex Samples Crosstabs procedure produces crosstabulation tables for pairs of selected variables and displays two-way statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Example. Using the Complex Samples Crosstabs procedure, you can obtain cross-classification statistics for smoking frequency by vitamin usage of U.S. citizens, based on the results of the National Health Interview Survey (NHIS) and with an appropriate analysis plan for this public-use data.

Statistics. The procedure produces estimates of cell population sizes and row, column, and table percentages, plus standard errors, confidence intervals, coefficients of variation, expected values, design effects, square roots of design effects, residuals, adjusted residuals, and unweighted counts for each estimate. The odds ratio, relative risk, and risk difference are computed for 2-by-2 tables. Additionally, Pearson and likelihood-ratio statistics are computed for the test of independence of the row and column variables.

Data. Row and column variables should be categorical. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Crosstabs

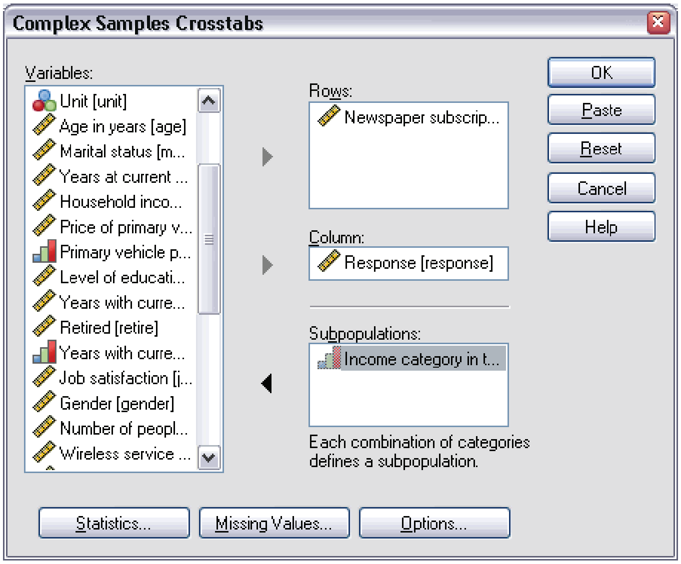
▶ From the menus choose:

Analyze
  Complex Samples
    Crosstabs...

▶ Select a plan file and, optionally, select a custom joint probabilities file.

▶ Click Continue.

Figure 7-1
Crosstabs dialog box



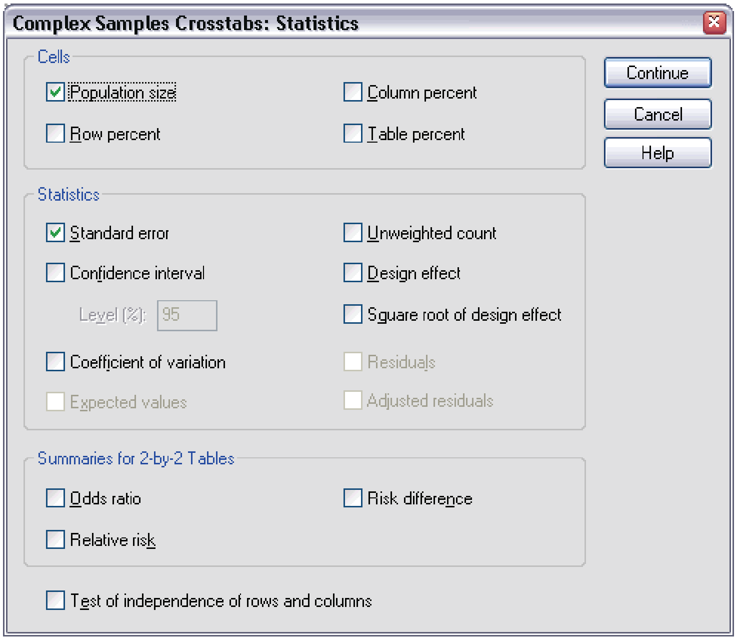
▶ Select at least one row variable and one column variable.

Optionally, you can:

- Specify variables to define subpopulations. Statistics are computed separately for each subpopulation.

Complex Samples Crosstabs Statistics

Figure 7-2
Crosstabs Statistics dialog box



Cells. This group allows you to request estimates of the cell population size and row, column, and table percentages.

Statistics. This group produces statistics associated with the population size and row, column, and table percentages.

- **Standard error.** The standard error of the estimate.
- **Confidence interval.** A confidence interval for the estimate, using the specified level.
- **Coefficient of variation.** The ratio of the standard error of the estimate to the estimate.
- **Expected values.** The expected value of the estimate, under the hypothesis of independence of the row and column variable.
- **Unweighted count.** The number of units used to compute the estimate.

- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Residuals.** The expected value is the number of cases that you would expect in the cell if there were no relationship between the two variables. A positive residual indicates that there are more cases in the cell than there would be if the row and column variables were independent.
- **Adjusted residuals.** The residual for a cell (observed minus expected value) divided by an estimate of its standard error. The resulting standardized residual is expressed in standard deviation units above or below the mean.

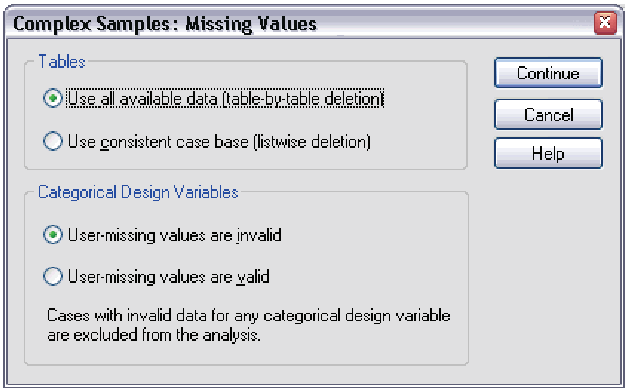
Summaries for 2-by-2 Tables. This group produces statistics for tables in which the row and column variable each have two categories. Each is a measure of the strength of the association between the presence of a factor and the occurrence of an event.

- **Odds ratio.** The odds ratio can be used as an estimate of relative risk when the occurrence of the factor is rare.
- **Relative risk.** The ratio of the risk of an event in the presence of the factor to the risk of the event in the absence of the factor.
- **Risk difference.** The difference between the risk of an event in the presence of the factor and the risk of the event in the absence of the factor.

Test of independence of rows and columns. This produces chi-square and likelihood-ratio tests of the hypothesis that a row and column variable are independent. Separate tests are performed for each pair of variables.

Complex Samples Missing Values

Figure 7-3
Missing Values dialog box



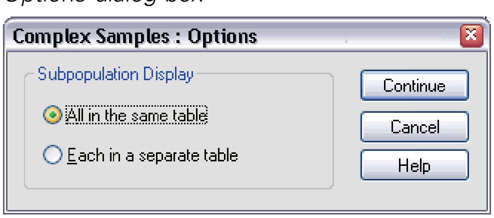
Tables. This group determines which cases are used in the analysis.

- **Use all available data.** Missing values are determined on a table-by-table basis. Thus, the cases used to compute statistics may vary across frequency or crosstabulation tables.
- **Use consistent case base.** Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent across tables.

Categorical Design Variables. This group determines whether user-missing values are valid or invalid.

Complex Samples Options

Figure 7-4
Options dialog box



Subpopulation Display. You can choose to have subpopulations displayed in the same table or in separate tables.

Chapter

8

Complex Samples Ratios

The Complex Samples Ratios procedure displays univariate summary statistics for ratios of variables. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Example. Using the Complex Samples Ratios procedure, you can obtain descriptive statistics for the ratio of current property value to last assessed value, based on the results of a statewide survey carried out according to a complex design and with an appropriate analysis plan for the data.

Statistics. The procedure produces ratio estimates, *t* tests, standard errors, confidence intervals, coefficients of variation, unweighted counts, population sizes, design effects, and square roots of design effects.

Data. Numerators and denominators should be positive-valued scale variables. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Ratios

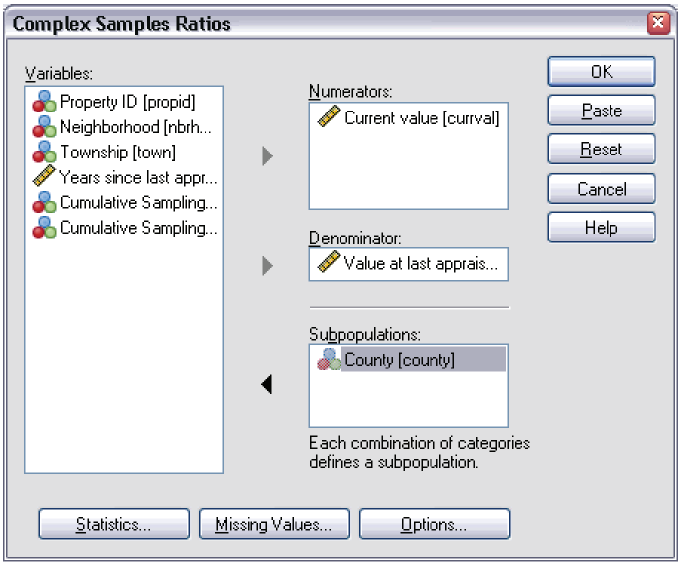
- From the menus choose:

  Analyze
    Complex Samples
      Ratios...

- Select a plan file and, optionally, select a custom joint probabilities file.

- Click Continue.

Figure 8-1
Ratios dialog box



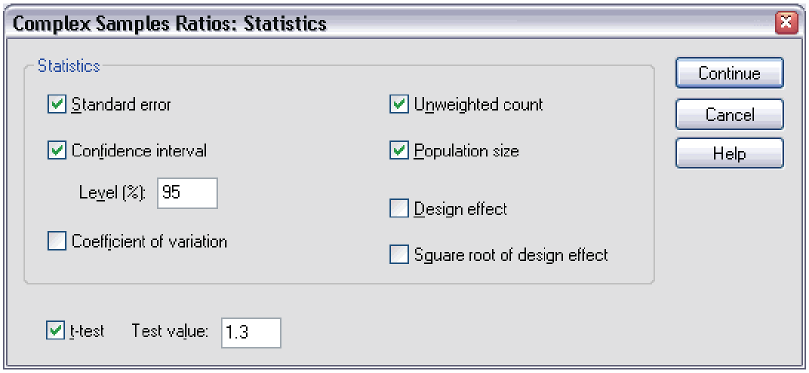
▶ Select at least one numerator variable and denominator variable.

Optionally, you can:

- Specify variables to define subgroups for which statistics are produced.

Complex Samples Ratios Statistics

Figure 8-2
Ratios Statistics dialog box



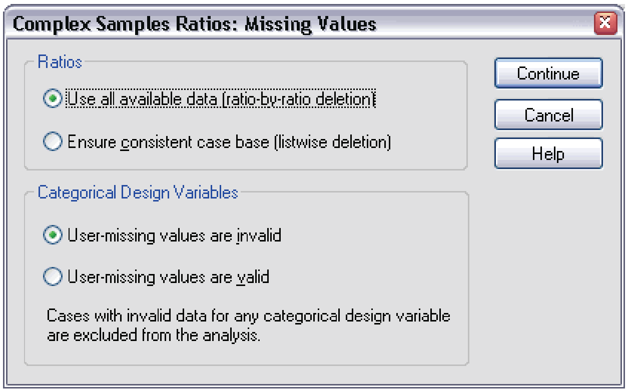
Statistics. This group produces statistics associated with the ratio estimate.

- **Standard error.** The standard error of the estimate.
- **Confidence interval.** A confidence interval for the estimate, using the specified level.
- **Coefficient of variation.** The ratio of the standard error of the estimate to the estimate.
- **Unweighted count.** The number of units used to compute the estimate.
- **Population size.** The estimated number of units in the population.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

t-test. You can request *t* tests of the estimates against a specified value.

Complex Samples Ratios Missing Values

Figure 8-3
Ratios Missing Values dialog box



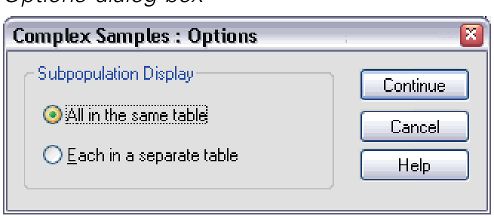
Ratios. This group determines which cases are used in the analysis.

- **Use all available data.** Missing values are determined on a ratio-by-ratio basis. Thus, the cases used to compute statistics may vary across numerator-denominator pairs.
- **Ensure consistent case base.** Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent.

Categorical Design Variables. This group determines whether user-missing values are valid or invalid.

Complex Samples Options

Figure 8-4
Options dialog box



Subpopulation Display. You can choose to have subpopulations displayed in the same table or in separate tables.

Chapter 9

Complex Samples General Linear Model

The Complex Samples General Linear Model (CSGLM) procedure performs linear regression analysis, as well as analysis of variance and covariance, for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Example. A grocery store chain surveyed a set of customers concerning their purchasing habits, according to a complex design. Given the survey results and how much each customer spent in the previous month, the store wants to see if the frequency with which customers shop is related to the amount they spend in a month, controlling for the gender of the customer and incorporating the sampling design.

Statistics. The procedure produces estimates, standard errors, confidence intervals, *t* tests, design effects, and square roots of design effects for model parameters, as well as the correlations and covariances between parameter estimates. Measures of model fit and descriptive statistics for the dependent and independent variables are also available. Additionally, you can request estimated marginal means for levels of model factors and factor interactions.

Data. The dependent variable is quantitative. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

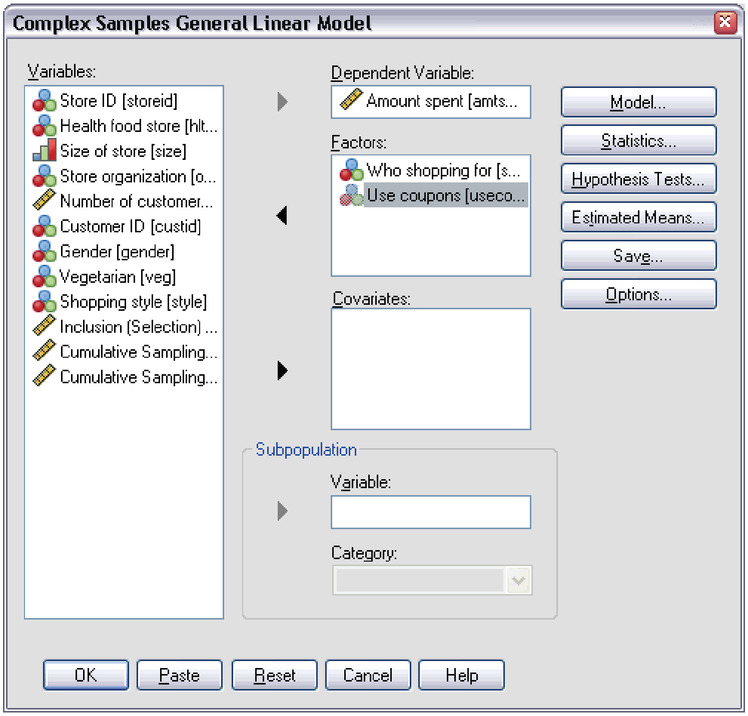
Obtaining a Complex Samples General Linear Model

From the menus choose:

Analyze
  Complex Samples
    General Linear Model...

- ▶ Select a plan file and, optionally, select a custom joint probabilities file.
- ▶ Click Continue.

Figure 9-1
General Linear Model dialog box

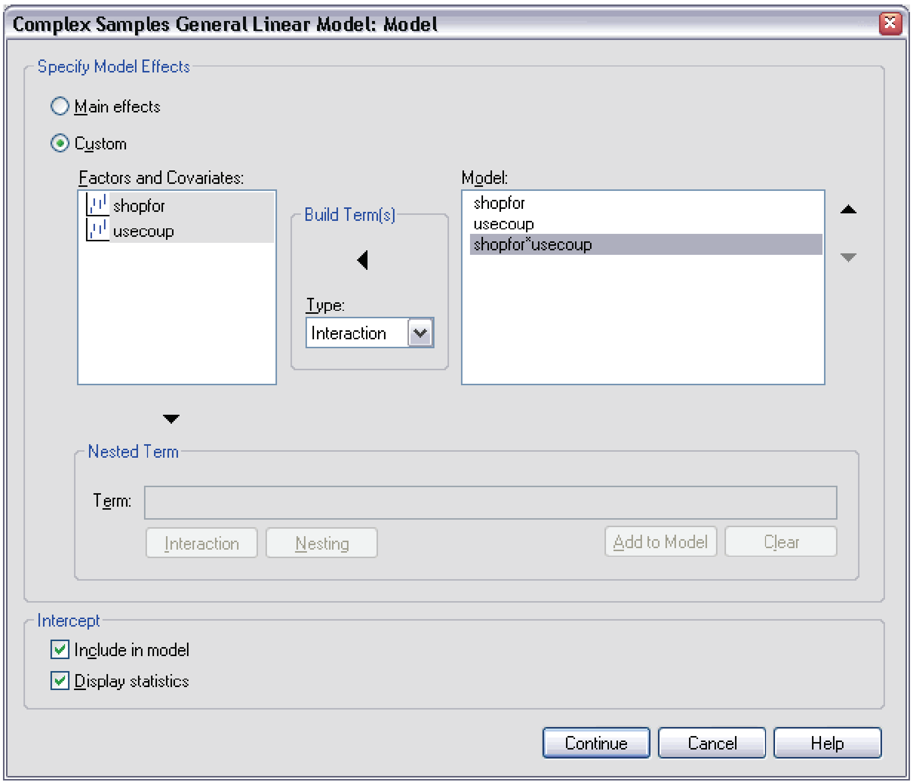


- ▶ Select a dependent variable.

Optionally, you can:

- Select variables for factors and covariates, as appropriate for your data.
- Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable.

Figure 9-2
Model dialog box



Specify Model Effects. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term for all selected variables.

Main effects. Creates a main-effects term for each variable selected.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *Customer* effect can be said to be **nested within** the *Store location* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

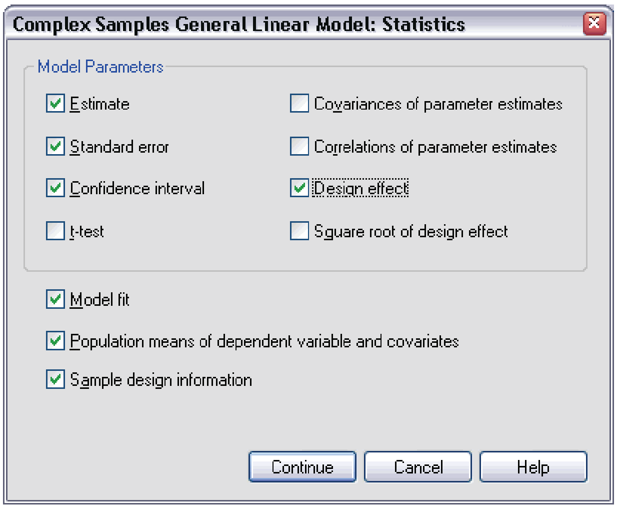
Limitations. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A**A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

Intercept. The intercept is usually included in the model. If you can assume the data pass through the origin, you can exclude the intercept. Even if you include the intercept in the model, you can choose to suppress statistics related to it.

Complex Samples General Linear Model Statistics

Figure 9-3
General Linear Model Statistics dialog box



Model Parameters. This group allows you to control the display of statistics related to the model parameters.

- **Estimate.** Displays estimates of the coefficients.
- **Standard error.** Displays the standard error for each coefficient estimate.
- **Confidence interval.** Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.
- **t-test.** Displays a *t* test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the model coefficients.
- **Correlations of parameter estimates.** Displays an estimate of the correlation matrix for the model coefficients.

- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

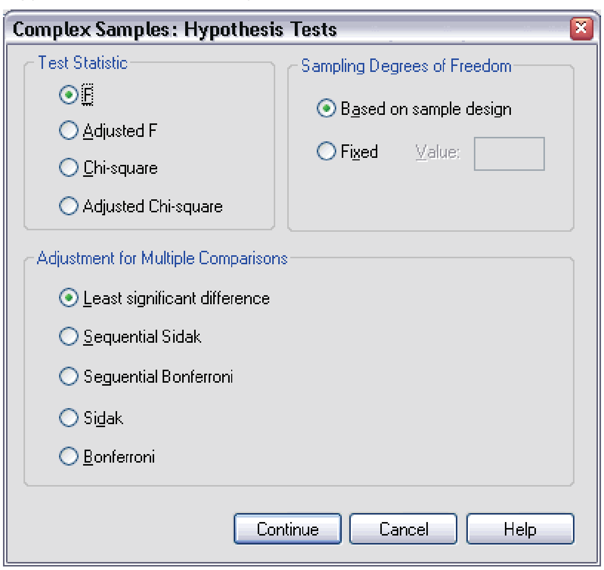
Model fit. Displays R^2 and root mean squared error statistics.

Population means of dependent variable and covariates. Displays summary information about the dependent variable, covariates, and factors.

Sample design information. Displays summary information about the sample, including the unweighted count and the population size.

Complex Samples Hypothesis Tests

Figure 9-4
Hypothesis Tests dialog box



Test Statistic. This group allows you to select the type of statistic used for testing hypotheses. You can choose between *F*, adjusted *F*, chi-square, and adjusted chi-square.

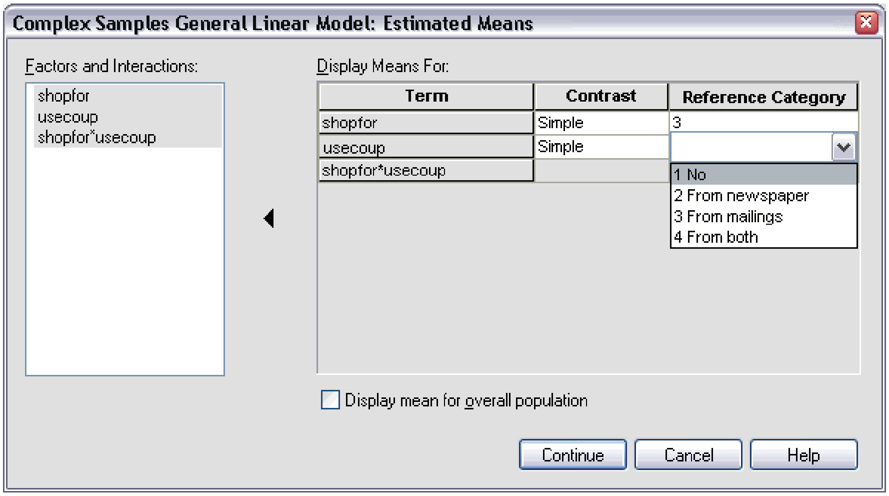
Sampling Degrees of Freedom. This group gives you control over the sampling design degrees of freedom used to compute *p* values for all test statistics. If based on the sampling design, the value is the difference between the number of primary sampling units and the number of strata in the first stage of sampling. Alternatively, you can set a custom degrees of freedom by specifying a positive integer.

Adjustment for Multiple Comparisons. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **Least significant difference.** This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.
- **Sequential Sidak.** This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- **Sequential Bonferroni.** This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- **Sidak.** This method provides tighter bounds than the Bonferroni approach.
- **Bonferroni.** This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

Complex Samples General Linear Model Estimated Means

Figure 9-5
General Linear Model Estimated Means dialog box



The Estimated Means dialog box allows you to display the model-estimated marginal means for levels of factors and factor interactions specified in the Model subdialog box. You can also request that the overall population mean be displayed.

Term. Estimated means are computed for the selected factors and factor interactions.

Contrast. The contrast determines how hypothesis tests are set up to compare the estimated means.

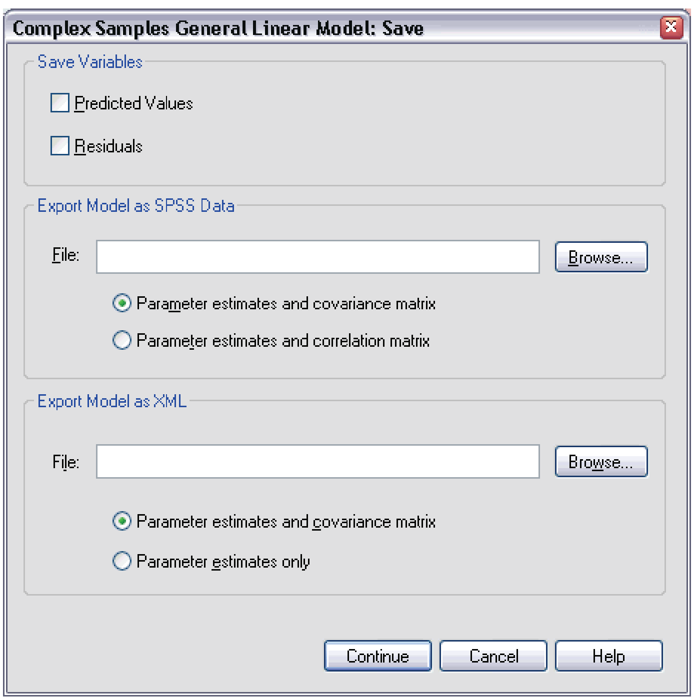
- **Simple.** Compares the mean of each level to the mean of a specified level. This type of contrast is useful when there is a control group.
- **Deviation.** Compares the mean of each level (except a reference category) to the mean of all of the levels (grand mean). The levels of the factor can be in any order.
- **Difference.** Compares the mean of each level (except the first) to the mean of previous levels. They are sometimes called reverse Helmert contrasts.
- **Helmert.** Compares the mean of each level of the factor (except the last) to the mean of subsequent levels.

- **Repeated.** Compares the mean of each level (except the last) to the mean of the subsequent level.
- **Polynomial.** Compares the linear effect, quadratic effect, cubic effect, and so on. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are often used to estimate polynomial trends.

Reference Category. The simple and deviation contrasts require a reference category, or factor level against which the others are compared.

Complex Samples General Linear Model Save

Figure 9-6
General Linear Model Save dialog box



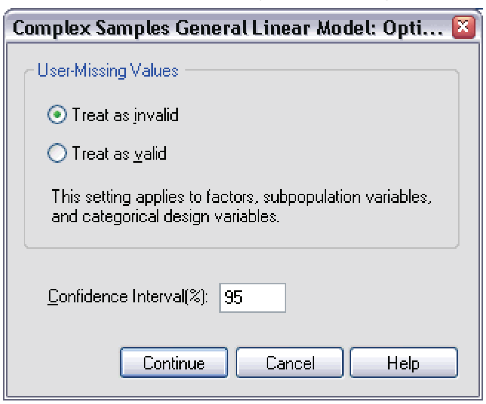
Save Variables. This group allows you to save the model predicted values and residuals as new variables in the working file.

Export Model as SPSS data. Writes an SPSS data file containing a covariance (or correlation, if selected) matrix of the parameter estimates in the model. Also, for each dependent variable, there will be a row of parameter estimates, a row of standard errors, a row of significance values for the *t* statistics corresponding to the parameter estimates, and a row of sampling design degrees of freedom. You can use this matrix file in other procedures that read an SPSS matrix file.

Export Model as XML. Saves the parameter estimates and the parameter covariance matrix, if selected, in XML (PMML) format. SmartScore and the server version of SPSS (a separate product) can use this model file to apply the model information to other data files for scoring purposes.

Complex Samples General Linear Model Options

Figure 9-7
General Linear Model Options dialog box



User-Missing Values. All design variables, as well as the dependent variable and any covariates, must have valid data. Cases with invalid data for any of these variables are deleted from the analysis. These controls allow you to decide whether user-missing values are treated as valid among the strata, cluster, subpopulation, and factor variables.

Confidence Interval. This is the confidence interval level for coefficient estimates and estimated marginal means. Specify a value greater than or equal to 50 and less than 100.

CSGLM Command Additional Features

The SPSS command language also allows you to:

- Specify custom tests of effects versus a linear combination of effects or a value (using the `CUSTOM` subcommand).
- Fix covariates at values other than their means when computing estimated marginal means (using the `EMMEANS` subcommand).
- Specify a metric for polynomial contrasts (using the `EMMEANS` subcommand).
- Specify a tolerance value for checking singularity (using the `CRITERIA` subcommand).
- Create user-specified names for saved variables (using the `SAVE` subcommand).
- Produce a general estimable function table (using the `PRINT` subcommand).

See the *SPSS Command Syntax Reference* for complete syntax information.

Chapter

10

Complex Samples Logistic Regression

The Complex Samples Logistic Regression procedure performs logistic regression analysis on a binary or multinomial dependent variable for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Example. A loan officer has collected past records of customers given loans at several different branches, according to a complex design. While incorporating the sample design, the officer wants to see if the probability with which a customer defaults is related to age, employment history, and amount of credit debt.

Statistics. The procedure produces estimates, exponentiated estimates, standard errors, confidence intervals, t tests, design effects, and square roots of design effects for model parameters, as well as the correlations and covariances between parameter estimates. Pseudo R^2 statistics, classification tables, and descriptive statistics for the dependent and independent variables are also available.

Data. The dependent variable is categorical. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

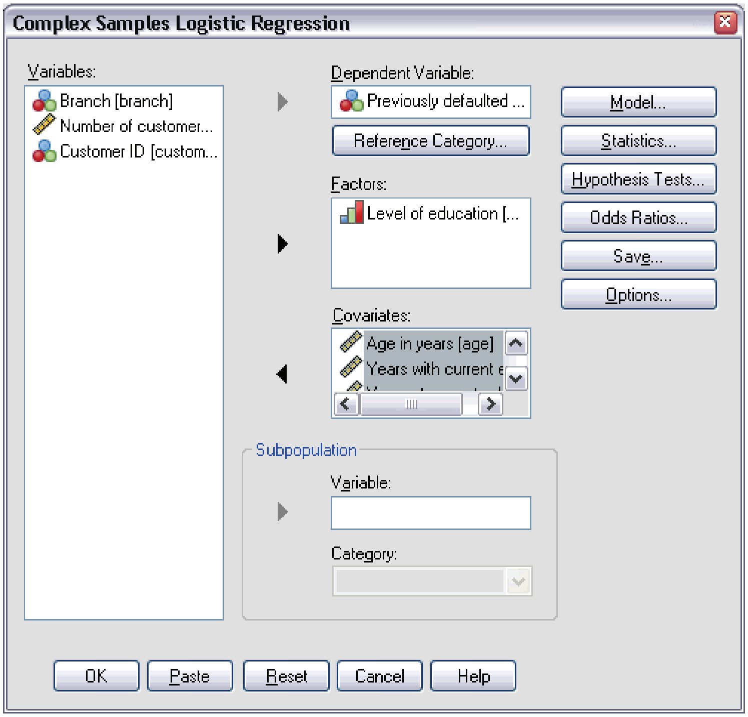
Obtaining Complex Samples Logistic Regression

From the menus choose:

Analyze
  Complex Samples
    Logistic Regression...

- ▶ Select a plan file and, optionally, select a custom joint probabilities file.
- ▶ Click Continue.

Figure 10-1
Logistic Regression dialog box



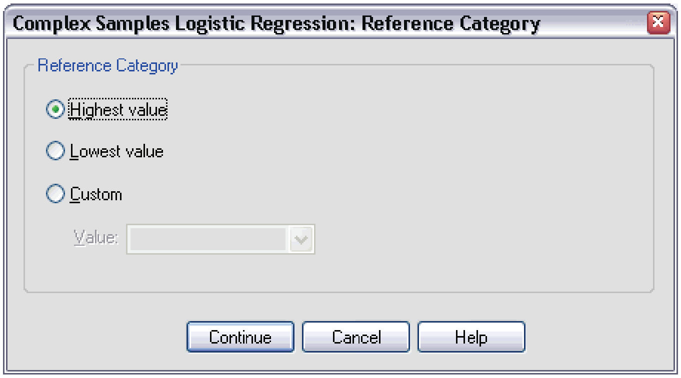
- ▶ Select a dependent variable.

Optionally, you can:

- Select variables for factors and covariates, as appropriate for your data.
- Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable.

Complex Samples Logistic Regression Reference Category

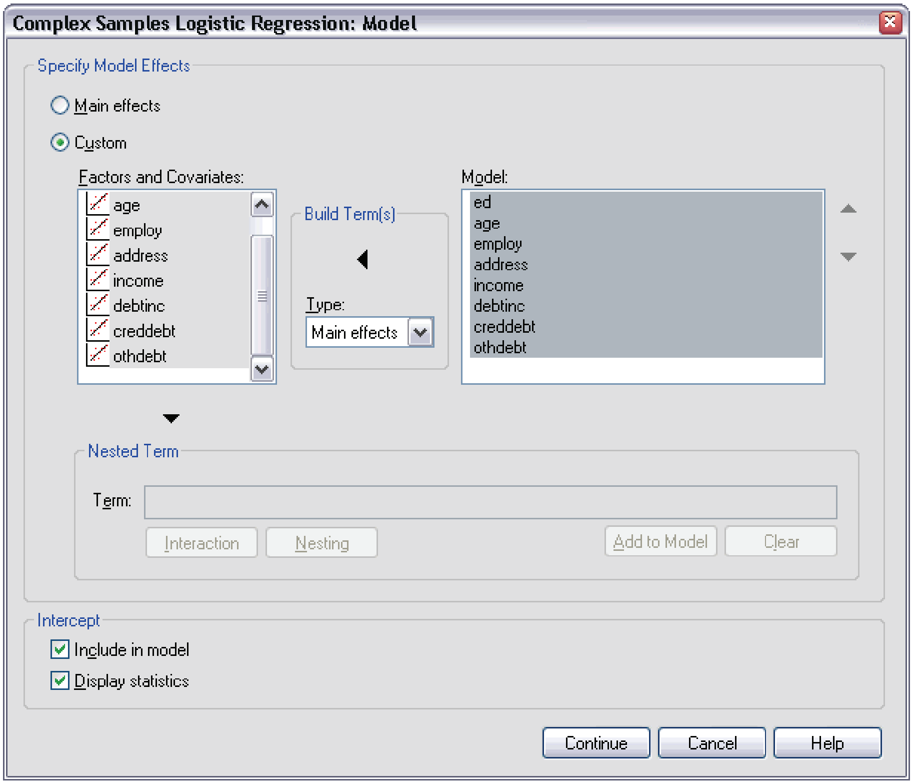
Figure 10-2
Logistic Regression Reference Category dialog box



By default, the Complex Samples Logistic Regression procedure makes the highest-valued category the reference category. This dialog box allows you to specify the highest, lowest, or a custom category as the reference category.

Complex Samples Logistic Regression Model

Figure 10-3
Logistic Regression Model dialog box



Specify Model Effects. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term for all selected variables.

Main effects. Creates a main-effects term for each variable selected.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *Customer* effect can be said to be **nested within** the *Store location* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

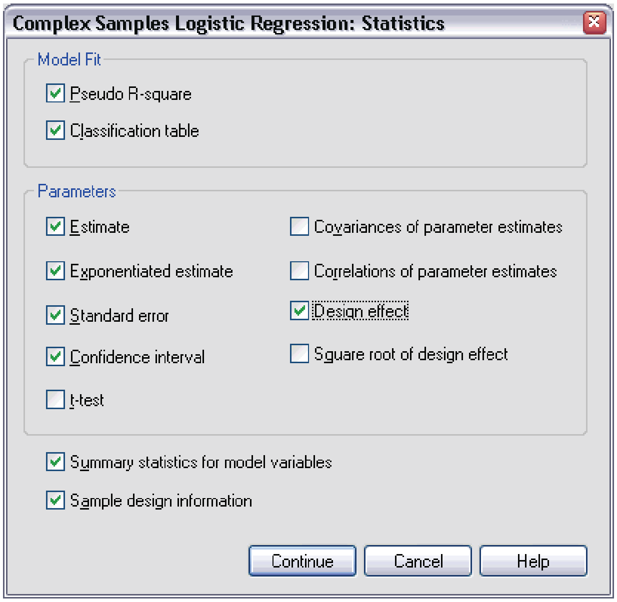
Limitations. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

Intercept. The intercept is usually included in the model. If you can assume the data pass through the origin, you can exclude the intercept. Even if you include the intercept in the model, you can choose to suppress statistics related to it.

Complex Samples Logistic Regression Statistics

Figure 10-4
Logistic Regression Statistics dialog box



Model Fit. Controls the displays of statistics that measure the overall model performance.

- **Pseudo R-square.** The R^2 statistic from linear regression does not have an exact counterpart among logistic regression models. There are, instead, multiple measures that attempt to mimic the properties of the R^2 statistic.
- **Classification table.** Displays the tabulated cross-classifications of the observed category by the model-predicted category on the dependent variable.

Parameters. This group allows you to control the display of statistics related to the model parameters.

- **Estimate.** Displays estimates of the coefficients.
- **Exponentiated estimate.** Displays the base of the natural logarithm raised to the power of the estimates of the coefficients. While the estimate has nice properties for statistical testing, the exponentiated estimate, or exp(B), is easier to interpret.
- **Standard error.** Displays the standard error for each coefficient estimate.

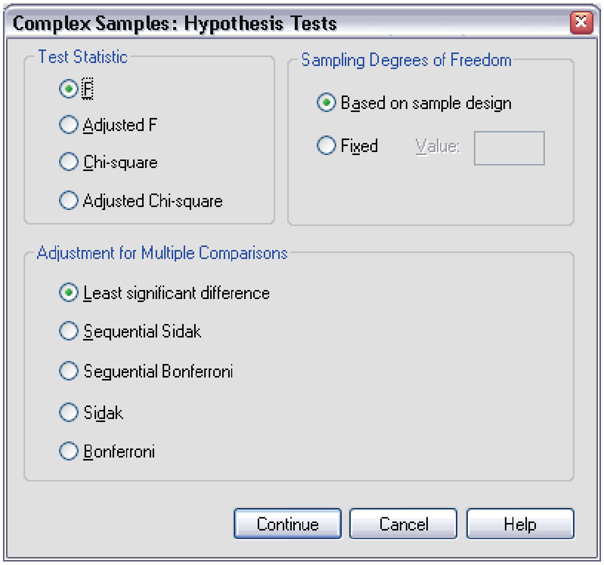
- **Confidence interval.** Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.
- **t-test.** Displays a *t* test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the model coefficients.
- **Correlations of parameter estimates.** Displays an estimate of the correlation matrix for the model coefficients.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

Summary statistics for model variables. Displays summary information about the dependent variable, covariates, and factors.

Sample design information. Displays summary information about the sample, including the unweighted count and the population size.

Complex Samples Hypothesis Tests

Figure 10-5
Hypothesis Tests dialog box



Test Statistic. This group allows you to select the type of statistic used for testing hypotheses. You can choose between F, adjusted F, chi-square, and adjusted chi-square.

Sampling Degrees of Freedom. This group gives you control over the sampling design degrees of freedom used to compute p values for all test statistics. If based on the sampling design, the value is the difference between the number of primary sampling units and the number of strata in the first stage of sampling. Alternatively, you can set a custom degrees of freedom by specifying a positive integer.

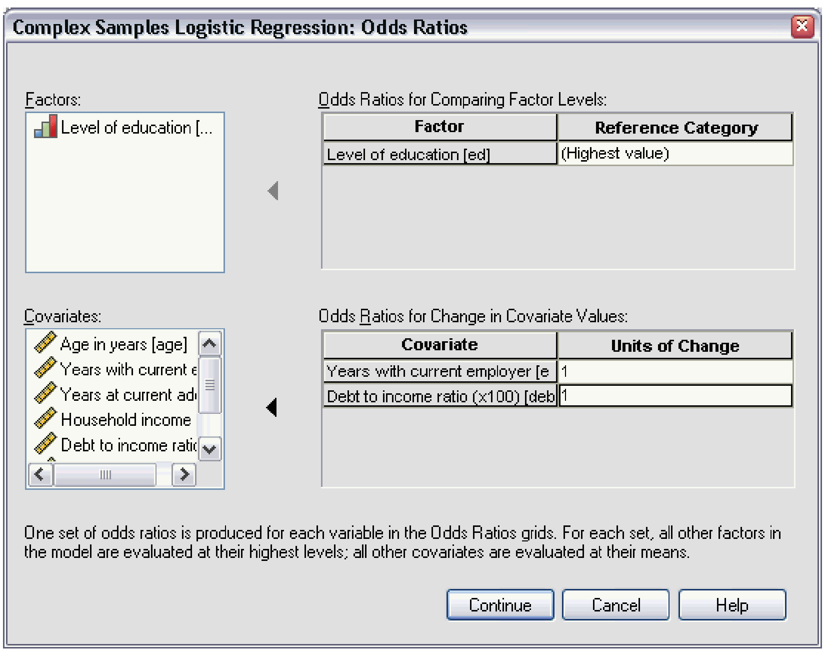
Adjustment for Multiple Comparisons. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **Least significant difference.** This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.

- **Sequential Sidak.** This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- **Sequential Bonferroni.** This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- **Sidak.** This method provides tighter bounds than the Bonferroni approach.
- **Bonferroni.** This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

Complex Samples Logistic Regression Odds Ratios

Figure 10-6
Logistic Regression Odds Ratios dialog box



The Odds Ratios dialog box allows you to display the model-estimated odds ratios for specified factors and covariates. A separate set of odds ratios is computed for each category of the dependent variable except the reference category.

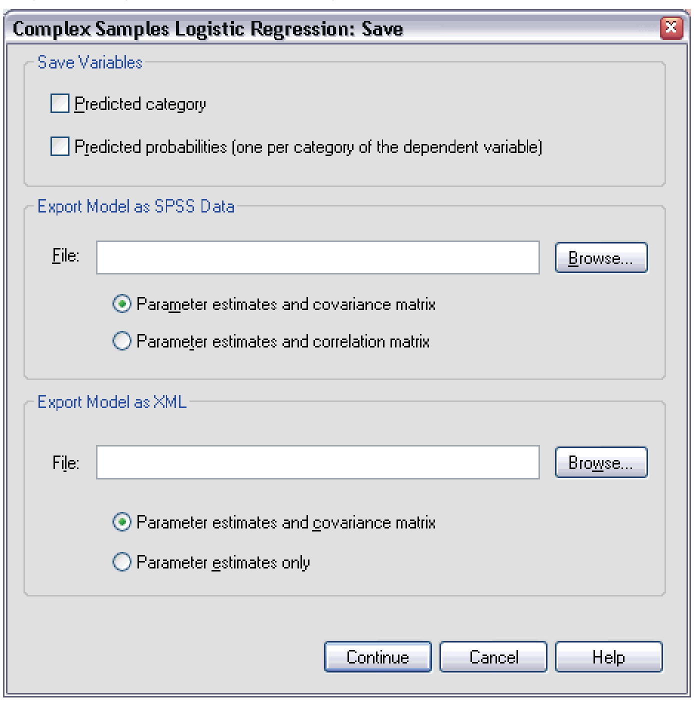
Factors. For each selected factor, displays the ratio of the odds at each category of the factor to the odds at the specified reference category.

Covariates. For each selected covariate, displays the ratio of the odds at the covariate's mean value plus the specified units of change to the odds at the mean.

When computing odds ratios for a factor or covariate, the procedure fixes all other factors at their highest levels and all other covariates at their means. If a factor or covariate interacts with other predictors in the model, then the odds ratios depend not only on the change in the specified variable but also on the values of the variables with which it interacts. If a specified covariate interacts with itself in the model (for example, *age***age*), then the odds ratios depend on both the change in the covariate and the value of the covariate.

Complex Samples Logistic Regression Save

Figure 10-7
Logistic Regression Save dialog box



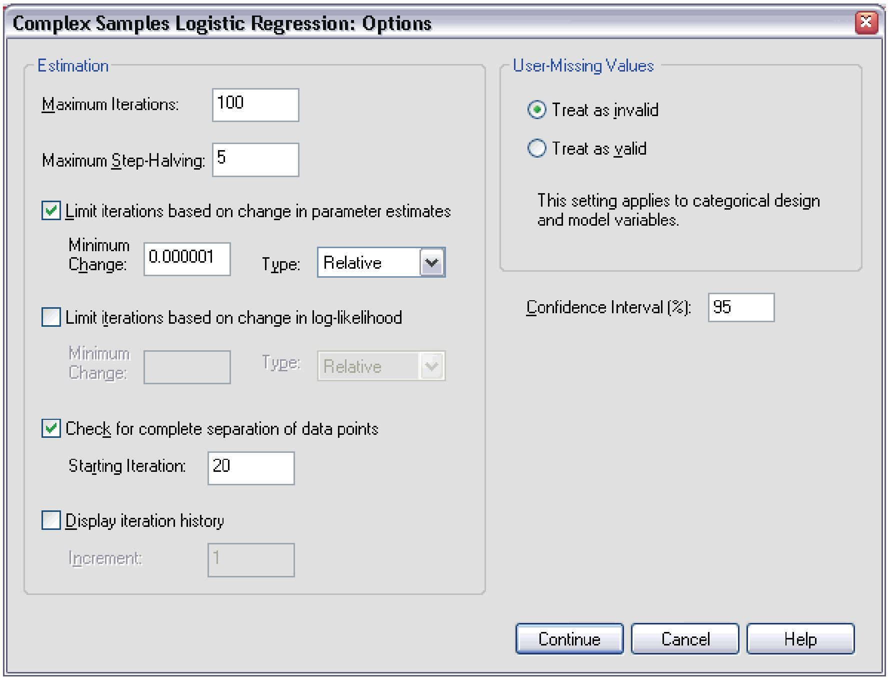
Save Variables. This group allows you to save the model-predicted category and predicted probabilities as new variables in the active dataset.

Export Model as SPSS data. Writes an SPSS data file containing a covariance (or correlation, if selected) matrix of the parameter estimates in the model. Also, for each dependent variable, there will be a row of parameter estimates, a row of standard errors, a row of significance values for the *t* statistics corresponding to the parameter estimates, and a row of sampling design degrees of freedom. You can use this matrix file in other procedures that read an SPSS matrix file.

Export Model as XML. Saves the parameter estimates and the parameter covariance matrix, if selected, in XML (PMML) format. SmartScore and the server version of SPSS (a separate product) can use this model file to apply the model information to other data files for scoring purposes.

Complex Samples Logistic Regression Options

Figure 10-8
Logistic Regression Options dialog box



Estimation. This group gives you control of various criteria used in the model estimation.

- **Maximum Iterations.** The maximum number of iterations the algorithm will execute. Specify a non-negative integer.
- **Maximum Step-Halving.** At each iteration, the step size is reduced by a factor of 0.5 until the log-likelihood increases or maximum step-halving is reached. Specify a positive integer.
- **Limit iterations based on change in parameter estimates.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the parameter estimates is less than the value specified, which must be non-negative.
- **Limit iterations based on change in log-likelihood.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the log-likelihood function is less than the value specified, which must be non-negative.

- **Check for complete separation of data points.** When selected, the algorithm performs tests to ensure that the parameter estimates have unique values. Separation occurs when the procedure can produce a model that correctly classifies every case.
- **Display iteration history.** Displays parameter estimates and statistics at every *n* iterations beginning with the 0^{th} iteration (the initial estimates). If you choose to print the iteration history, the last iteration is always printed regardless of the value of *n*.

User-Missing Values. All design variables, as well as the dependent variable and any covariates, must have valid data. Cases with invalid data for any of these variables are deleted from the analysis. These controls allow you to decide whether user-missing values are treated as valid among the strata, cluster, subpopulation, and factor variables.

Confidence Interval. This is the confidence interval level for coefficient estimates, exponentiated coefficient estimates, and odds ratios. Specify a value greater than or equal to 50 and less than 100.

CSLOGISTIC Command Additional Features

The SPSS command language also allows you to:

- Specify custom tests of effects versus a linear combination of effects or a value (using the `CUSTOM` subcommand).
- Fix values of other model variables when computing odds ratios for factors and covariates (using the `ODDSRATIOS` subcommand).
- Specify a tolerance value for checking singularity (using the `CRITERIA` subcommand).
- Create user-specified names for saved variables (using the `SAVE` subcommand).
- Produce a general estimable function table (using the `PRINT` subcommand).

See the *SPSS Command Syntax Reference* for complete syntax information.

Chapter

11

Complex Samples Ordinal Regression

The Complex Samples Ordinal Regression procedure performs regression analysis on a binary or ordinal dependent variable for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Example. Representatives considering a bill before the legislature are interested in whether there is public support for the bill and how support for the bill is related to voter demographics. Pollsters design and conduct interviews according to a complex sampling design. Using Complex Samples Ordinal Regression, you can fit a model for the level of support for the bill based upon voter demographics.

Data. The dependent variable is ordinal. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Ordinal Regression

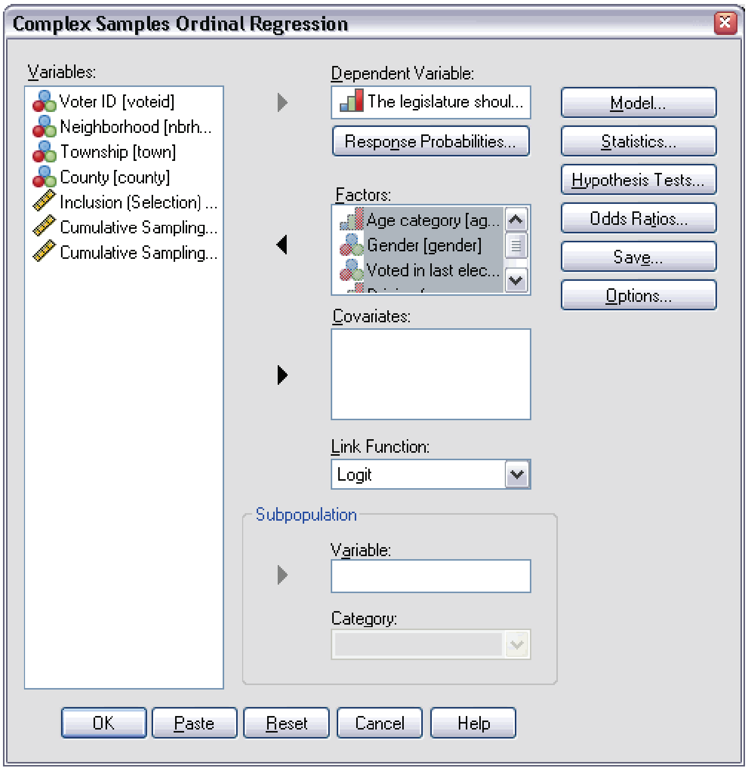
From the menus choose:

Analyze
  Complex Samples
    Ordinal Regression...

▶ Select a plan file and, optionally, select a custom joint probabilities file.

▶ Click Continue.

Figure 11-1
Complex Samples Ordinal Regression dialog box



▶ Select a dependent variable.

Optionally, you can:

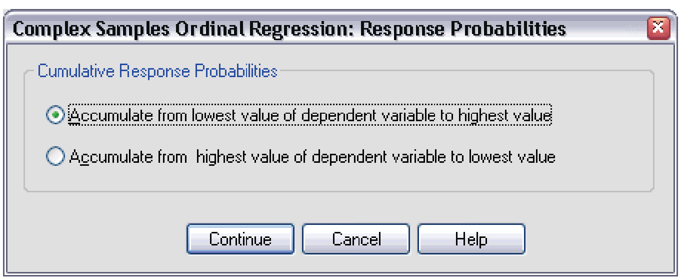
- Select variables for factors and covariates, as appropriate for your data.
- Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable, although variances are still properly estimated based on the entire dataset.
- Select a link function.

Link Function. The link function is a transformation of the cumulative probabilities that allows estimation of the model. Five link functions are available, summarized in the following table.

Function	**Form**	**Typical application**
Logit	$\log(\ \xi\ /\ (1-\xi)\ )$	Evenly distributed categories
Complementary log-log	$\log(-\log(1-\xi))$	Higher categories more probable
Negative log-log	$-\log(-\log(\xi))$	Lower categories more probable
Probit	$\Phi^{-1}(\xi)$	Latent variable is normally distributed
Cauchit (inverse Cauchy)	$\tan(\pi(\xi-0.5))$	Latent variable has many extreme values

Complex Samples Ordinal Regression Response Probabilities

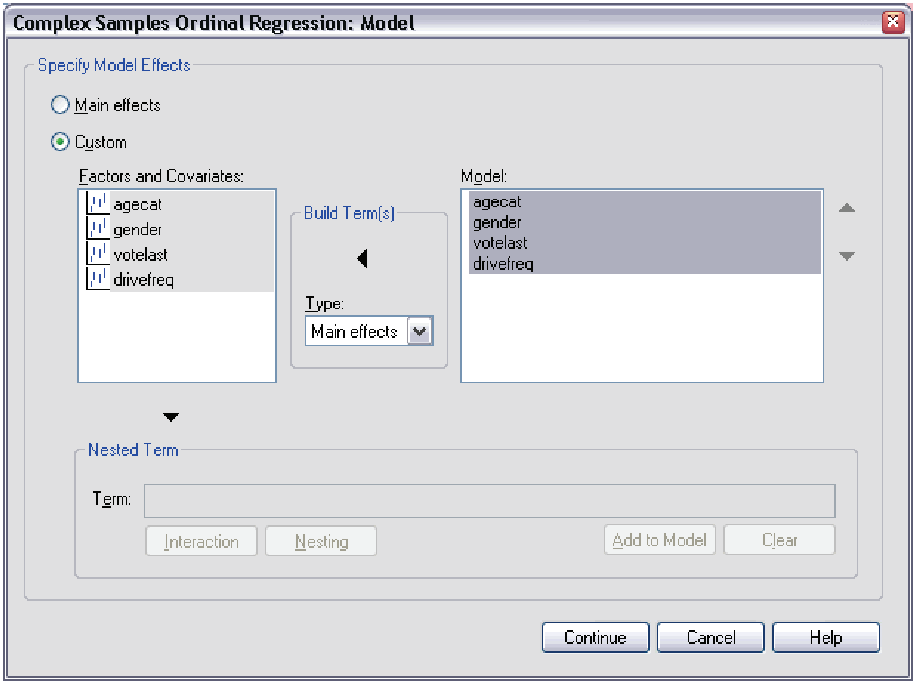
Figure 11-2
Complex Samples Ordinal Regression Response Probabilities dialog box



The Response Probabilities dialog box allows you to specify whether the cumulative probability of a response (that is, the probability of belonging up to and including a particular category of the dependent variable) increases with increasing or decreasing values of the dependent variable.

Complex Samples Ordinal Regression Model

Figure 11-3
Complex Samples Ordinal Regression Model dialog box



Specify Model Effects. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term for all selected variables.

Main effects. Creates a main-effects term for each variable selected.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *Customer* effect can be said to be **nested within** the *Store location* effect.

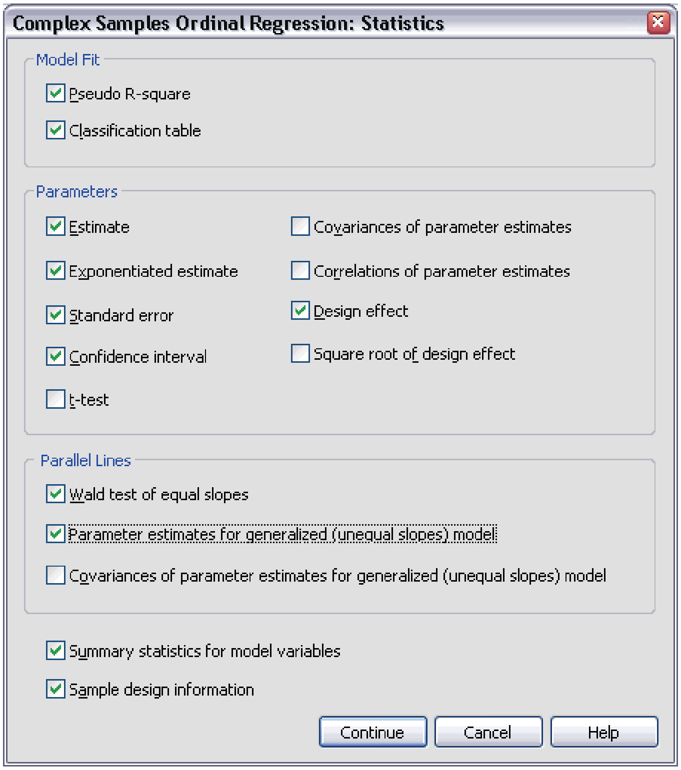
Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

Limitations. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

Complex Samples Ordinal Regression Statistics

Figure 11-4
Complex Samples Ordinal Regression Statistics dialog box



Model Fit. Controls the displays of statistics that measure the overall model performance.

- **Pseudo R-square.** The R^2 statistic from linear regression does not have an exact counterpart among ordinal regression models. There are, instead, multiple measures that attempt to mimic the properties of the R^2 statistic.
- **Classification table.** Displays the tabulated cross-classifications of the observed category by the model-predicted category on the dependent variable.

Parameters. This group allows you to control the display of statistics related to the model parameters.

- **Estimate.** Displays estimates of the coefficients.
- **Exponentiated estimate.** Displays the base of the natural logarithm raised to the power of the estimates of the coefficients. While the estimate has nice properties for statistical testing, the exponentiated estimate, or exp(B), is easier to interpret.
- **Standard error.** Displays the standard error for each coefficient estimate.
- **Confidence interval.** Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.
- **t-test.** Displays a *t* test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the model coefficients.
- **Correlations of parameter estimates.** Displays an estimate of the correlation matrix for the model coefficients.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure, expressed in units comparable to those of the standard error, of the effect of specifying a complex design, where values further from 1 indicate greater effects.

Parallel Lines. This group allows you to request statistics associated with a model with non-parallel lines where a separate regression line is fitted for each response category (except the last).

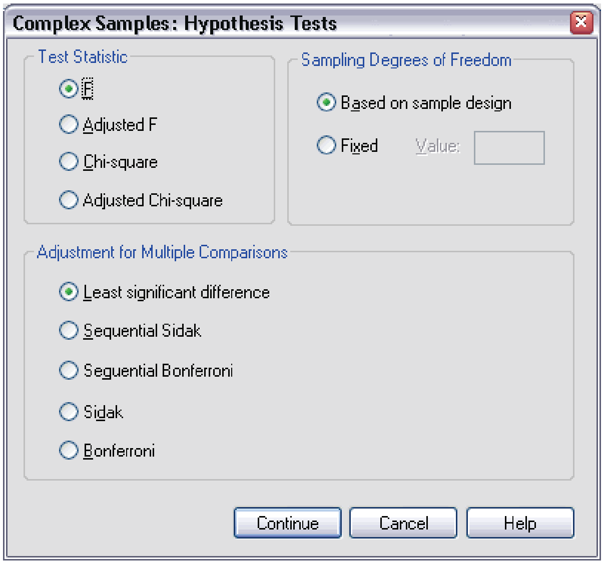
- **Wald test.** Produces a test of the null hypothesis that regression parameters are equal for all cumulative responses. The model with non-parallel lines is estimated and the Wald test of equal parameters is applied.
- **Parameter estimates.** Displays estimates of the coefficients and standard errors for the model with non-parallel lines.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the coefficients of the model with non-parallel lines.

Summary statistics for model variables. Displays summary information about the dependent variable, covariates, and factors.

Sample design information. Displays summary information about the sample, including the unweighted count and the population size.

Complex Samples Hypothesis Tests

Figure 11-5
Hypothesis Tests dialog box



Test Statistic. This group allows you to select the type of statistic used for testing hypotheses. You can choose between *F*, adjusted *F*, chi-square, and adjusted chi-square.

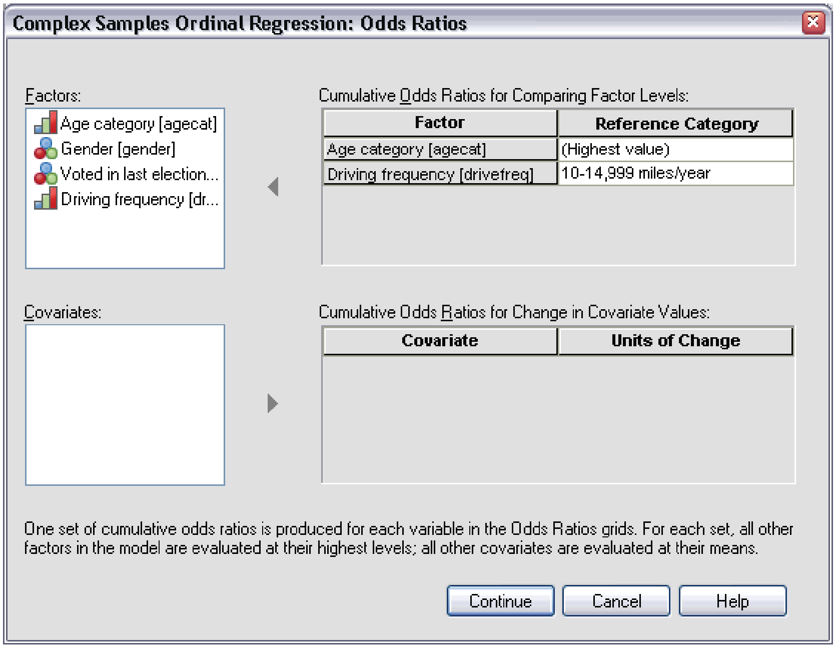
Sampling Degrees of Freedom. This group gives you control over the sampling design degrees of freedom used to compute *p* values for all test statistics. If based on the sampling design, the value is the difference between the number of primary sampling units and the number of strata in the first stage of sampling. Alternatively, you can set a custom degrees of freedom by specifying a positive integer.

Adjustment for Multiple Comparisons. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **Least significant difference.** This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.
- **Sequential Sidak.** This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- **Sequential Bonferroni.** This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- **Sidak.** This method provides tighter bounds than the Bonferroni approach.
- **Bonferroni.** This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

Complex Samples Ordinal Regression Odds Ratios

Figure 11-6
Complex Samples Ordinal Regression Odds Ratios dialog box



The Odds Ratios dialog box allows you to display the model-estimated cumulative odds ratios for specified factors and covariates. This feature is only available for models using the Logit link function. A single cumulative odds ratio is computed for all categories of the dependent variable except the last; the proportional odds model postulates that they are all equal.

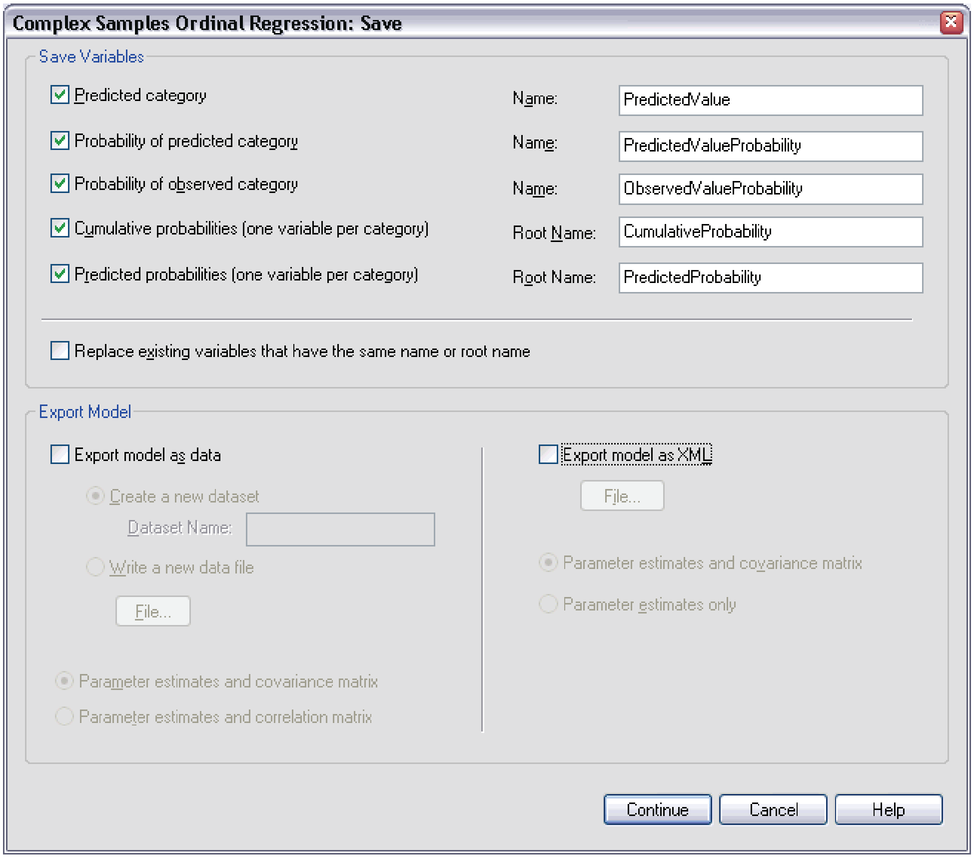
Factors. For each selected factor, displays the ratio of the cumulative odds at each category of the factor to the odds at the specified reference category.

Covariates. For each selected covariate, displays the ratio of the cumulative odds at the covariate's mean value plus the specified units of change to the odds at the mean.

When computing odds ratios for a factor or covariate, the procedure fixes all other factors at their highest levels and all other covariates at their means. If a factor or covariate interacts with other predictors in the model, then the odds ratios depend not only on the change in the specified variable but also on the values of the variables with which it interacts. If a specified covariate interacts with itself in the model (for example, *age*age*), then the odds ratios depend on both the change in the covariate and the value of the covariate.

Complex Samples Ordinal Regression Save

Figure 11-7
Complex Samples Ordinal Regression Save dialog box



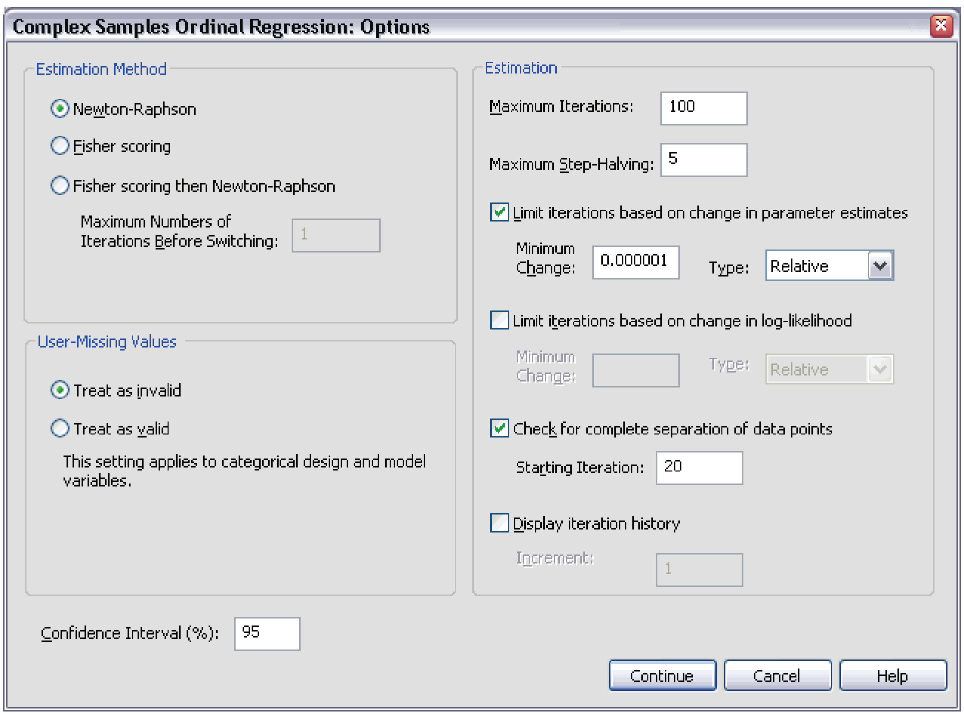
Save Variables. This group allows you to save the model-predicted category, probability of predicted category, probability of observed category, cumulative probabilities, and predicted probabilities as new variables in the active dataset.

Export model as data. Writes an SPSS data file containing a covariance (or correlation, if selected) matrix of the parameter estimates in the model. Also, for each dependent variable, there will be a row of parameter estimates, a row of standard errors, a row of significance values for the *t* statistics corresponding to the parameter estimates, and a row of sampling design degrees of freedom.

Export model as XML. Saves the parameter estimates and the parameter covariance matrix, if selected, in XML (PMML) format. SmartScore and the server version of SPSS (a separate product) can use this model file to apply the model information to other data files for scoring purposes.

Complex Samples Ordinal Regression Options

Figure 11-8
Complex Samples Ordinal Regression Options dialog box



Estimation Method. You can select a parameter estimation method; choose between Newton-Raphson, Fisher scoring, or a hybrid method in which Fisher scoring iterations are performed before switching to the Newton-Raphson method. If convergence is achieved during the Fisher scoring phase of the hybrid method before the maximum number of Fisher iterations is reached, the algorithm continues with the Newton-Raphson method.

Estimation. This group gives you control of various criteria used in the model estimation.

- **Maximum Iterations.** The maximum number of iterations the algorithm will execute. Specify a non-negative integer.
- **Maximum Step-Halving.** At each iteration, the step size is reduced by a factor of 0.5 until the log-likelihood increases or maximum step-halving is reached. Specify a positive integer.
- **Limit iterations based on change in parameter estimates.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the parameter estimates is less than the value specified, which must be non-negative.
- **Limit iterations based on change in log-likelihood.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the log-likelihood function is less than the value specified, which must be non-negative.
- **Check for complete separation of data points.** When selected, the algorithm performs tests to ensure that the parameter estimates have unique values. Separation occurs when the procedure can produce a model that correctly classifies every case.
- **Display iteration history.** Displays parameter estimates and statistics at every *n* iterations beginning with the 0^{th} iteration (the initial estimates). If you choose to print the iteration history, the last iteration is always printed regardless of the value of *n*.

User-Missing Values. Scale design variables, as well as the dependent variable and any covariates, should have valid data. Cases with invalid data for any of these variables are deleted from the analysis. These controls allow you to decide whether user-missing values are treated as valid among the strata, cluster, subpopulation, and factor variables.

Confidence Interval. This is the confidence interval level for coefficient estimates, exponentiated coefficient estimates, and odds ratios. Specify a value greater than or equal to 50 and less than 100.

CSORDINAL Command Additional Features

The SPSS command language also allows you to:

- Specify custom tests of effects versus a linear combination of effects or a value (using the CUSTOM subcommand).
- Fix values of other model variables at values other than their means when computing cumulative odds ratios for factors and covariates (using the ODDSRATIOS subcommand).
- Use unlabeled values as custom reference categories for factors when odds ratios are requested (using the ODDSRATIOS subcommand).
- Specify a tolerance value for checking singularity (using the CRITERIA subcommand).
- Produce a general estimable function table (using the PRINT subcommand).
- Save more than 25 probability variables (using the SAVE subcommand).

See the *SPSS Command Syntax Reference* for complete syntax information.

Part II: Examples

Chapter

12

Complex Samples Sampling Wizard

The Sampling Wizard guides you through the steps for creating, modifying, or executing a sampling plan file. Before using the wizard, you should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

Obtaining a Sample from a Full Sampling Frame

A state agency is charged with ensuring fair property taxes from county to county. Taxes are based on the appraised value of the property, so the agency wants to survey a sample of properties by county to be sure that each county's records are equally up to date. However, resources for obtaining current appraisals are limited, so it's important that what is available is used wisely. The agency decides to employ complex sampling methodology to select a sample of properties.

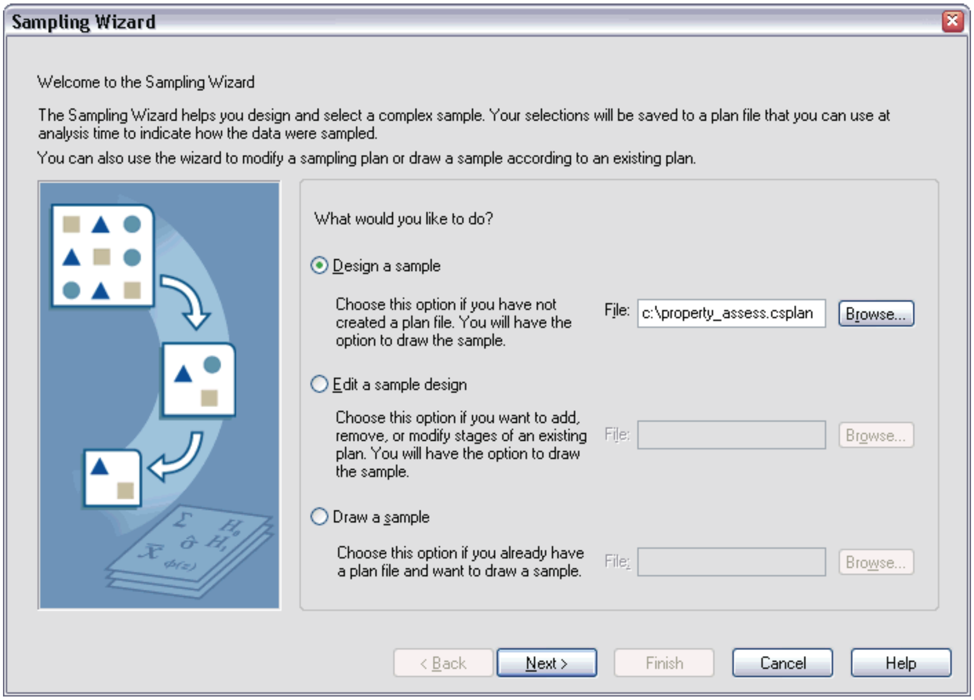
A listing of properties is collected in *property_assess_cs.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. Use the Complex Samples Sampling Wizard to select a sample.

Using the Wizard

▶ To run the Complex Samples Sampling Wizard, from the menus choose:

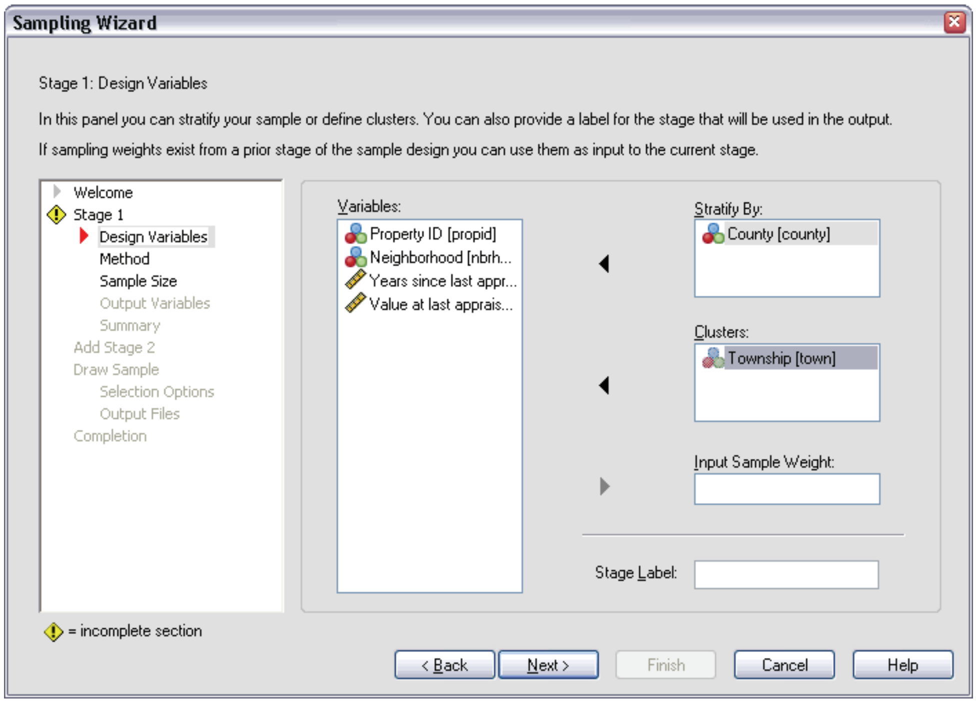
Analyze
  Complex Samples
    Select a Sample...

Figure 12-1
Sampling Wizard, Welcome step



- Select Design a sample, and type c:\property_assess.csplan as the name of the plan file.
- Click Next.

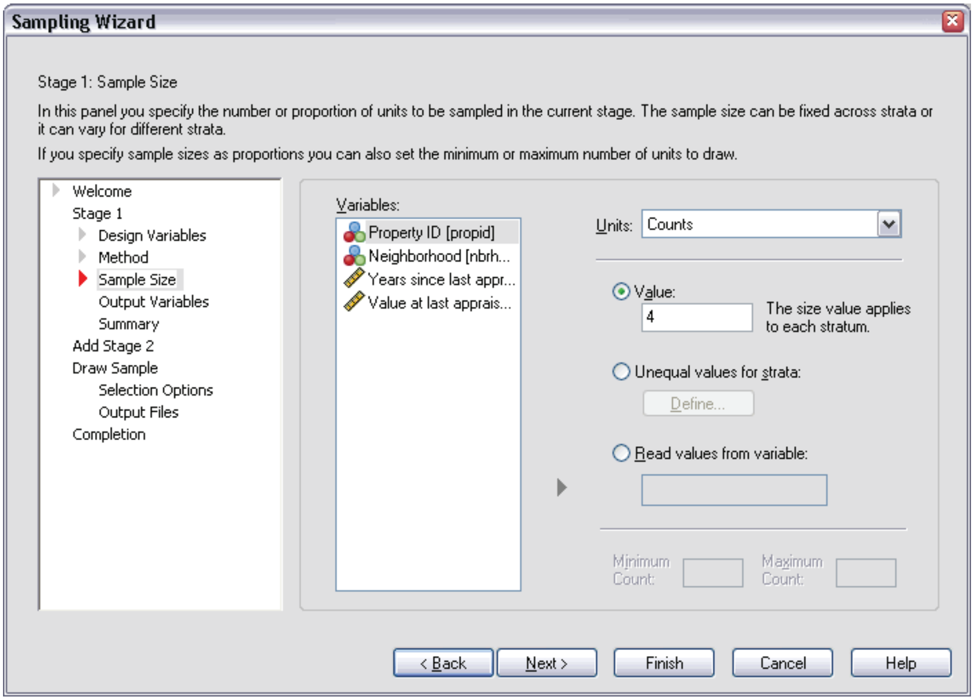
Figure 12-2
Sampling Wizard, Design Variables step (stage 1)



- Select *County* as a stratification variable.
- Select *Township* as a cluster variable.
- Click **Next**, and then click **Next** in the Method step.

This design structure means that independent samples are drawn for each county. In this stage, townships are drawn as the primary sampling unit using the default method, simple random sampling.

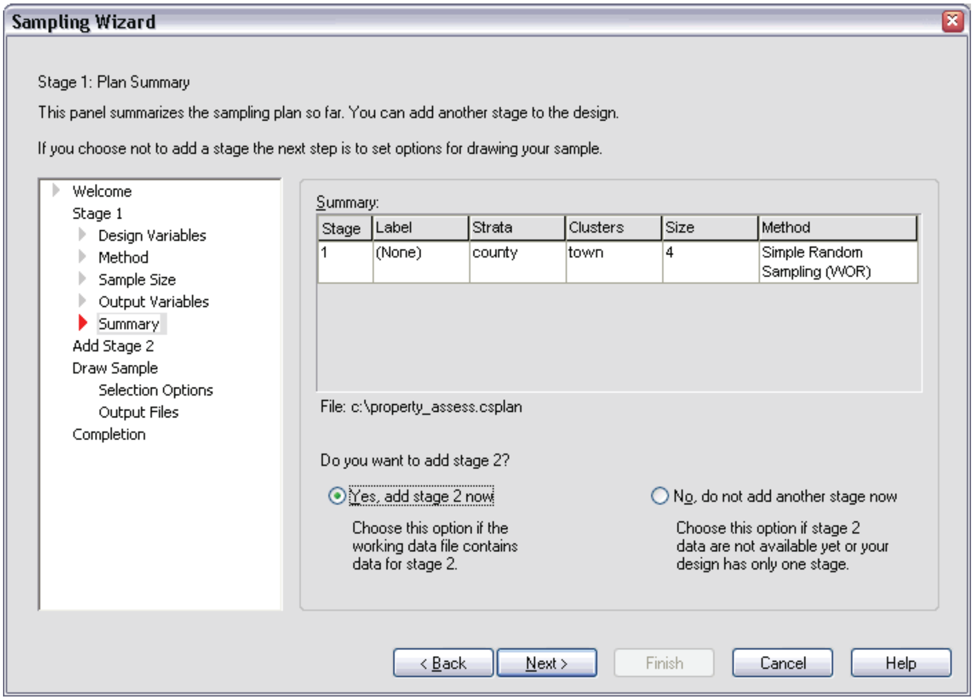
Figure 12-3
Sampling Wizard, Sample Size step (stage 1)



▶ Type 4 as the value for the number of clusters to select in this stage.

▶ Click **Next**, and then click **Next** in the Output Variables step.

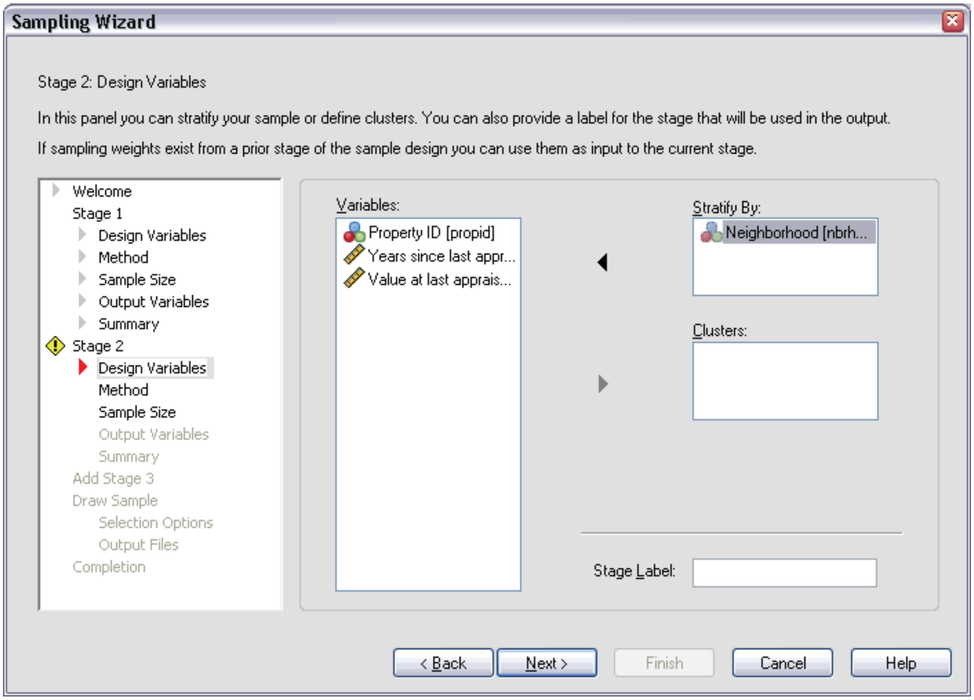
Figure 12-4
Sampling Wizard, Plan Summary step (stage 1)



▶ Select Yes, add stage 2 now.

▶ Click Next.

Figure 12-5
Sampling Wizard, Design Variables step (stage 2)

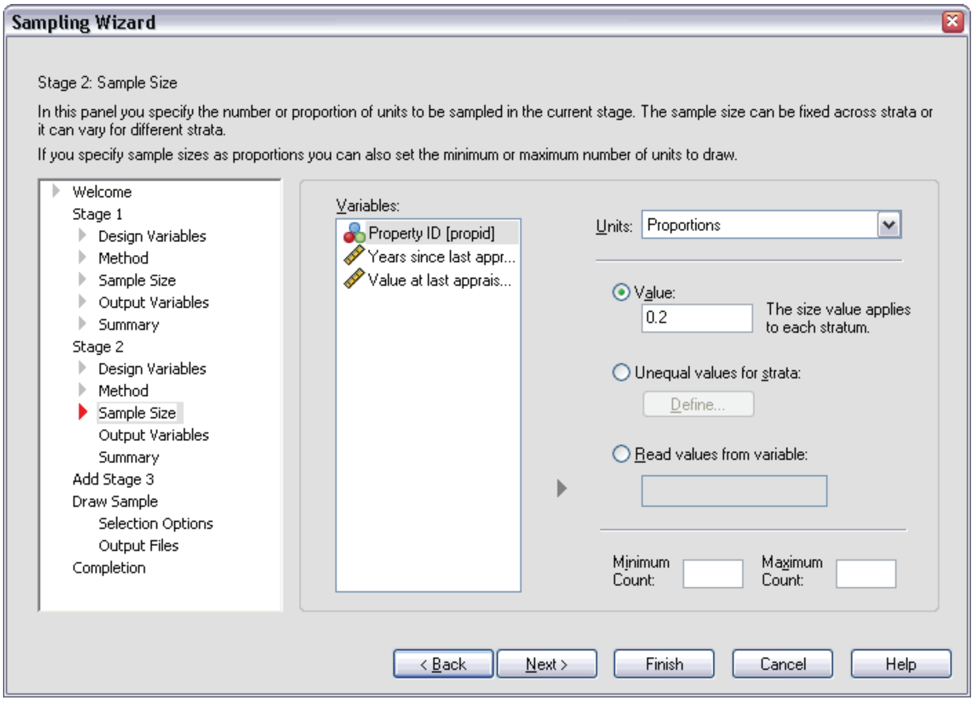


▶ Select *Neighborhood* as a stratification variable.

▶ Click **Next**, and then click **Next** in the Method step.

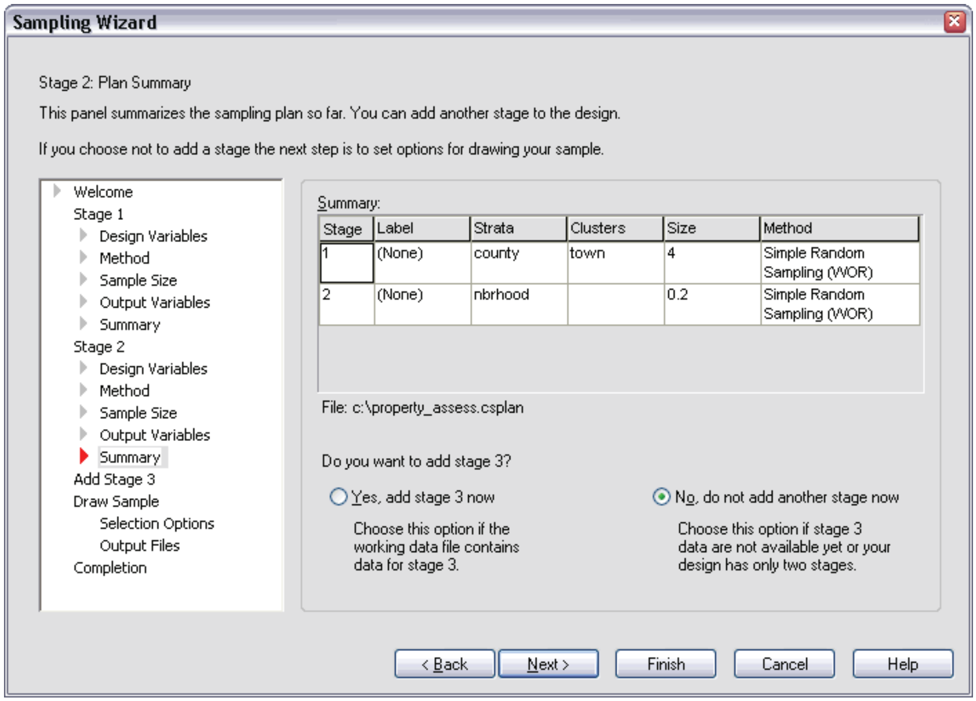
This design structure means that independent samples are drawn for each neighborhood of the townships drawn in stage 1. In this stage, properties are drawn as the primary sampling unit using simple random sampling.

Figure 12-6
Sampling Wizard, Sample Size step (stage 2)



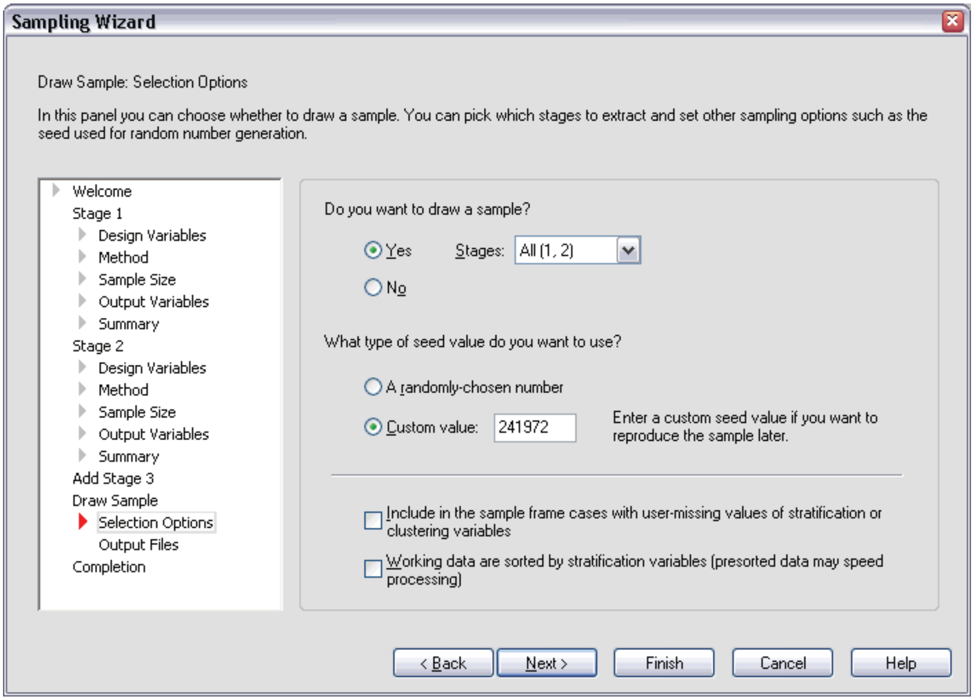
- Select **Proportions** from the Units drop-down list.
- Type **0.2** as the value of the proportion of units to sample from each stratum.
- Click **Next**, and then click **Next** in the Output Variables step.

Figure 12-7
Sampling Wizard, Plan Summary step (stage 2)



▶ Look over the sampling design, and then click Next.

Figure 12-8
Sampling Wizard, Draw Sample, Selection Options step

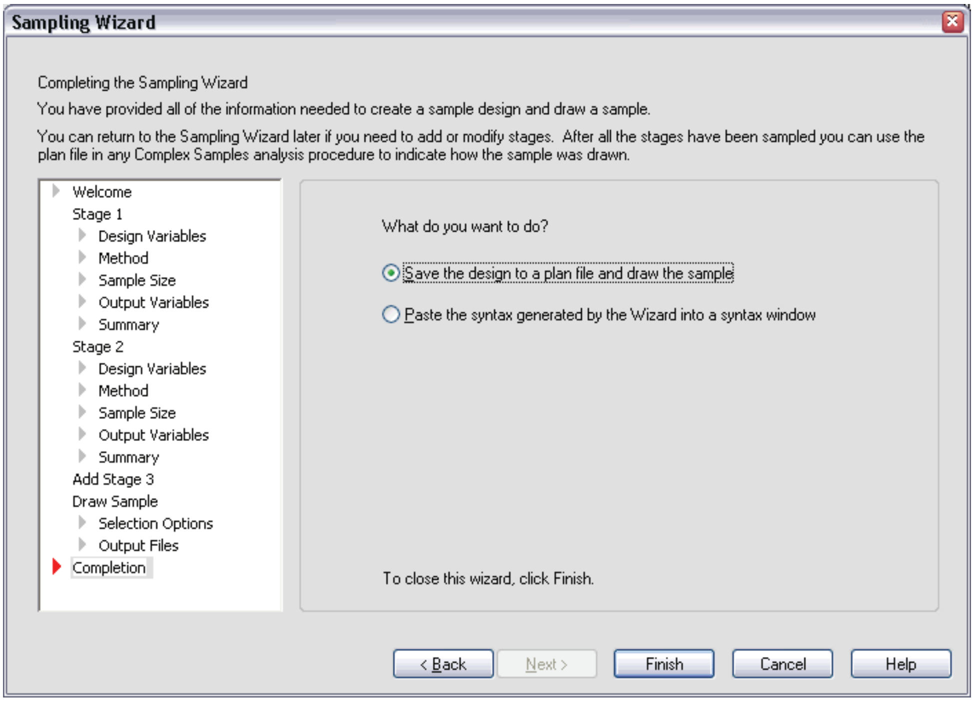


- Select **Custom value** for the type of random seed to use, and type **241972** as the value.

  Using a custom value allows you to replicate the results of this example exactly.

- Click **Next**, and then click **Next** in the Draw Sample, Output Files step.

Figure 12-9
Sampling Wizard, Finish step



▶ Click Finish.

These selections produce the sampling plan file *property_assess.csplan* and draw a sample according to that plan.

Plan Summary

Figure 12-10
Plan summary

			Stage 1	Stage 2
Design Variables	Stratification	1	County	Neighborhood
	Cluster	1	Township	
Sample Information	Selection Method		Simple random sampling without replacement	Simple random sampling without replacement
	Number of Units Sampled		4	
	Variables Created or Modified	Stagewise Inclusion (Selection) Probability	InclusionProbability_1_	InclusionProbability_2_
		Stagewise Cumulative Sample Weight	SampleWeightCumulative_1_	SampleWeightCumulative_2_
	Proportion of Units Sampled			.2
Analysis Information	Estimator Assumption		Equal probability sampling without replacement	Equal probability sampling without replacement
	Inclusion Probability		Obtained from variable InclusionProbability_1_	Obtained from variable InclusionProbability_2_

Plan File: c:\property_assess.csplan
Weight Variable: SampleWeight_Final_

The summary table reviews your sampling plan, and it is useful for making sure that the plan represents your intentions.

Sampling Summary

Figure 12-11
Stage summary

County	Number of Units Sampled		Proportion of Units Sampled	
	Requested	Actual	Requested	Actual
Eastern	4	4	44.4%	44.4%
Central	4	4	57.1%	57.1%
Western	4	4	25.0%	25.0%
Northern	4	4	44.4%	44.4%
Southern	4	4	50.0%	50.0%

Plan File: c:\property_assess.csplan

This summary table reviews the first stage of sampling, and it is useful for checking that the sampling went according to plan. Four townships were sampled from each county, as requested.

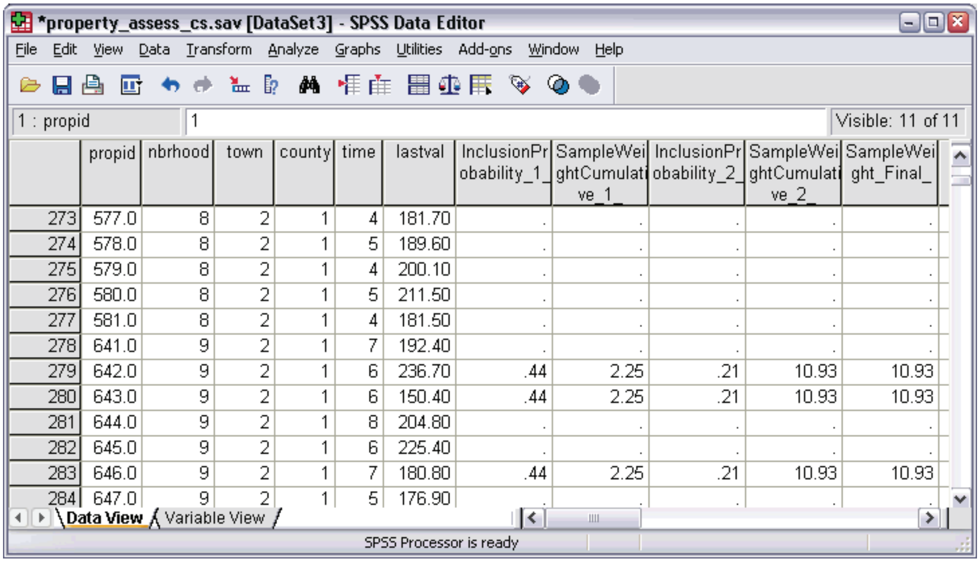
Figure 12-12
Stage summary

County	Township	Neighborhood	Number of Units Sampled		Proportion of Units Sampled	
			Requested	Actual	Requested	Actual
Eastern	2	8	4	4	20.0%	19.0%
		9	14	14	20.0%	20.6%
		10	7	7	20.0%	18.9%
		11	14	14	20.0%	20.0%
	6	36	13	13	20.0%	20.3%
		37	14	14	20.0%	20.6%
		38	13	13	20.0%	20.6%
	7	43	12	12	20.0%	20.7%
		44	11	11	20.0%	19.6%
		45	11	11	20.0%	20.8%
		46	13	13	20.0%	20.0%
	9	57	13	13	20.0%	20.6%
		58	5	5	20.0%	18.5%
		59	11	11	20.0%	19.3%
		60	13	13	20.0%	19.4%
Central	22	148	9	9	20.0%	19.6%
		149	8	8	20.0%	20.0%

This summary table (the top part of which is shown here) reviews the second stage of sampling. It is also useful for checking that the sampling went according to plan. Approximately 20% of the properties were sampled from each neighborhood from each township sampled in the first stage, as requested.

Sample Results

Figure 12-13
Data Editor with sample results



	propid	nbrhood	town	county	time	lastval	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	SampleWeight_Final_
273	577.0	8	2	1	4	181.70	.	.	.	.	.
274	578.0	8	2	1	5	189.60	.	.	.	.	.
275	579.0	8	2	1	4	200.10	.	.	.	.	.
276	580.0	8	2	1	5	211.50	.	.	.	.	.
277	581.0	8	2	1	4	181.50	.	.	.	.	.
278	641.0	9	2	1	7	192.40	.	.	.	.	.
279	642.0	9	2	1	6	236.70	.44	2.25	.21	10.93	10.93
280	643.0	9	2	1	6	150.40	.44	2.25	.21	10.93	10.93
281	644.0	9	2	1	8	204.80	.	.	.	.	.
282	645.0	9	2	1	6	225.40	.	.	.	.	.
283	646.0	9	2	1	7	180.80	.44	2.25	.21	10.93	10.93
284	647.0	9	2	1	5	176.90	.	.	.	.	.

You can see the sampling results in the Data Editor. Five new variables were saved to the working file, representing the inclusion probabilities and cumulative sampling weights for each stage, plus the final sampling weights.

- Cases with values for these variables were selected to the sample.
- Cases with system-missing values for the variables were not selected.

The agency will now use its resources to collect current valuations for the properties selected in the sample. Once those valuations are available, you can process the sample with Complex Samples analysis procedures, using the sampling plan *property_assess.csplan* to provide the sampling specifications.

Obtaining a Sample from a Partial Sampling Frame

A company is interested in compiling and selling a database of high-quality survey information. The survey sample should be representative but efficiently carried out, so complex sampling methods are used. The full sampling design calls for the following structure:

Stage	Strata	Clusters
1	Region	Province
2	District	City
3	Subdivision	

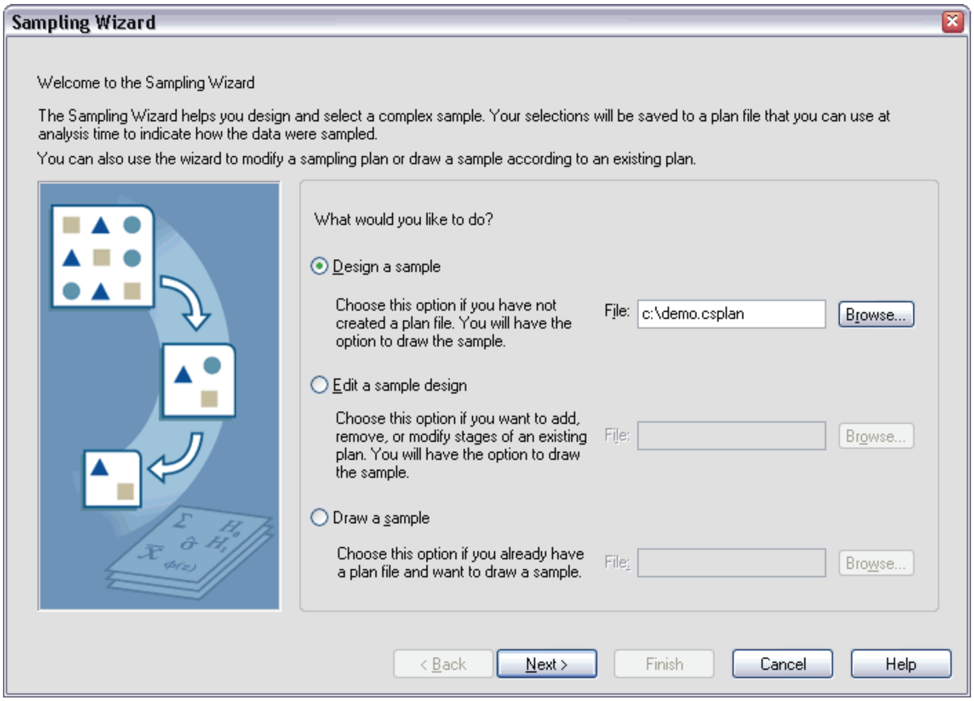
In the third stage, households are the primary sampling unit, and selected households will be surveyed. However, since information is easily available only to the city level, the company plans to execute the first two stages of the design now and then collect information on the numbers of subdivisions and households from the sampled cities. The available information to the city level is collected in *demo_cs_1.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. Note that this file contains a variable *Subdivision* that contains all 1's. This is a placeholder for the "true" variable, whose values will be collected after the first two stages of the design are executed, that allows you to specify the full three-stage sampling design now. Use the Complex Samples Sampling Wizard to specify the full complex sampling design, and then draw the first two stages.

Using the Wizard to Sample from the First Partial Frame

▶ To run the Complex Samples Sampling Wizard, from the menus choose:

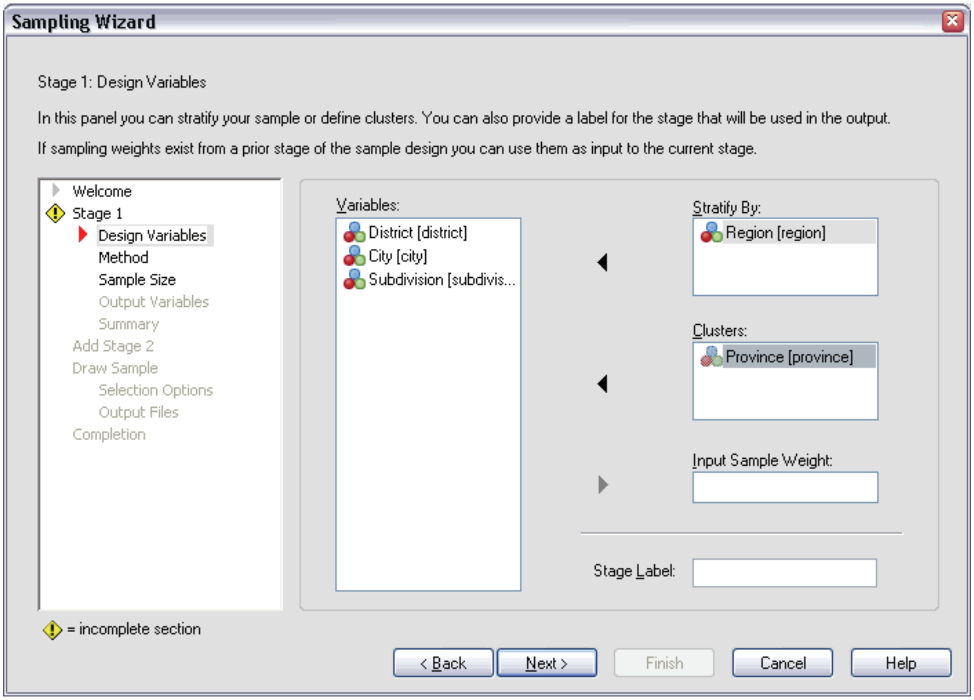
Analyze
  Complex Samples
    Select a Sample...

Figure 12-14
Sampling Wizard, Welcome step



- Select Design a sample, and type c:\demo.csplan as the name of the plan file.
- Click Next.

Figure 12-15
Sampling Wizard, Design Variables step (stage 1)



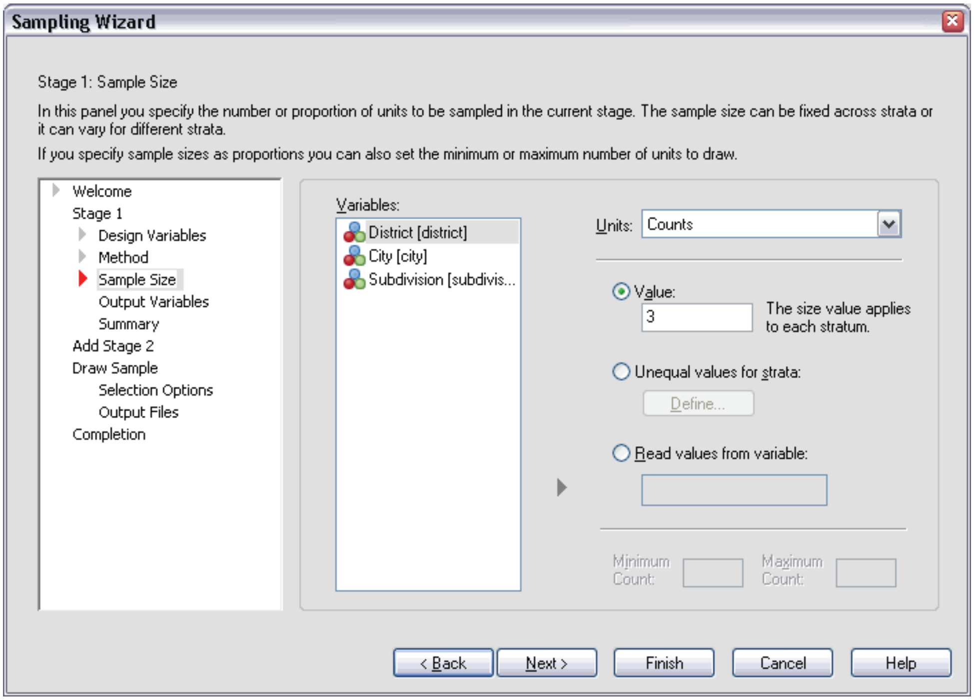
▶ Select *Region* as a stratification variable.

▶ Select *Province* as a cluster variable.

▶ Click Next, and then click Next in the Method step.

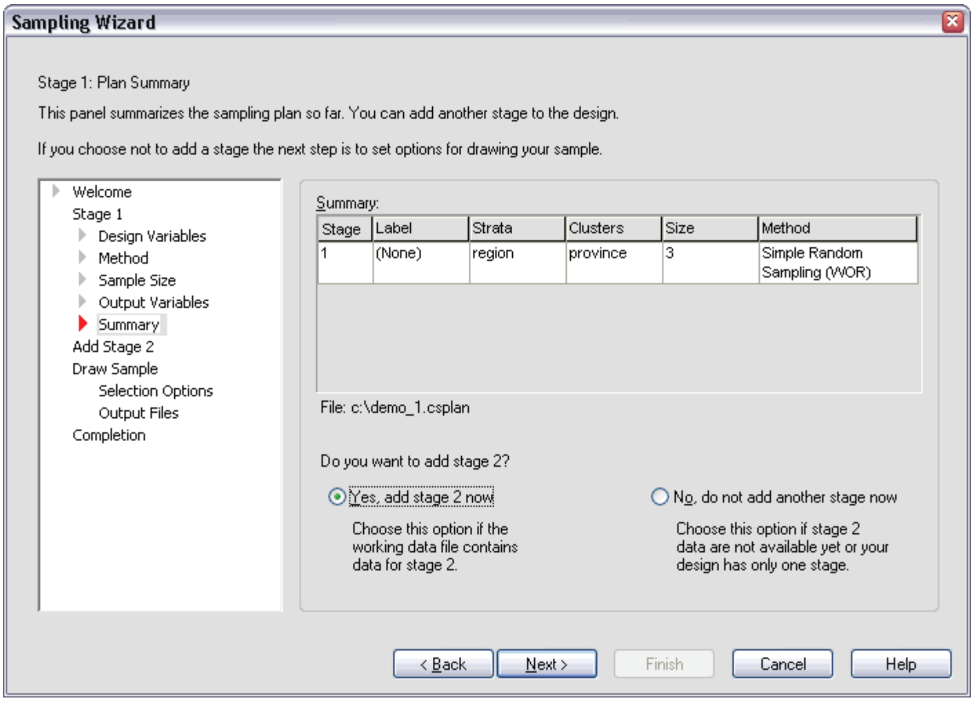
This design structure means that independent samples are drawn for each region. In this stage, provinces are drawn as the primary sampling unit using the default method, simple random sampling.

Figure 12-16
Sampling Wizard, Sample Size step (stage 1)



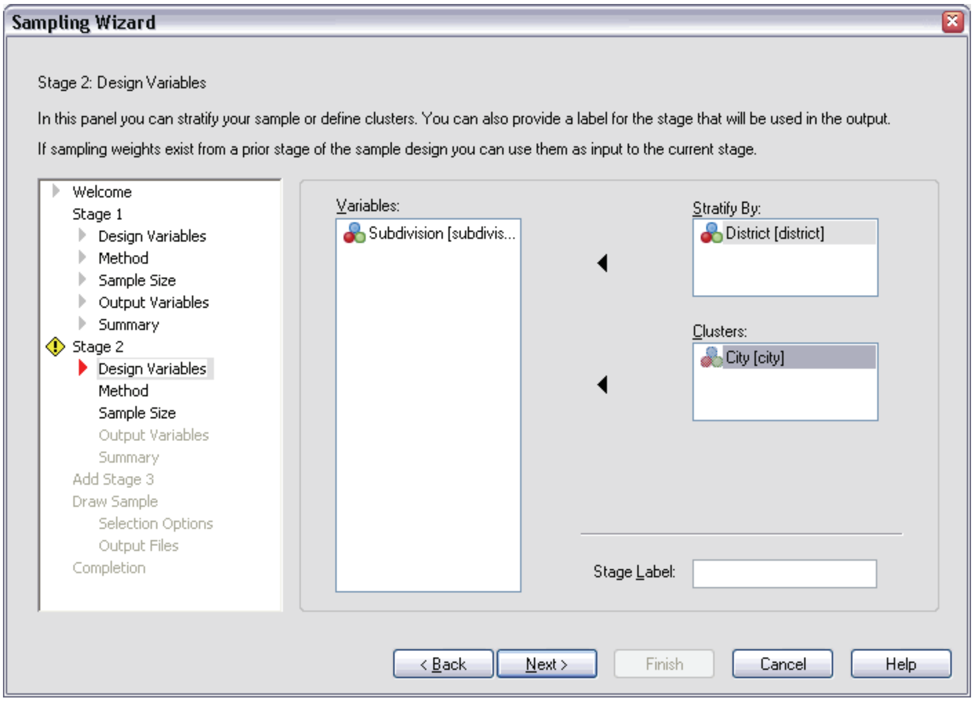
- Select **Counts** from the Units drop-down list.
- Type **3** as the value for the number of clusters to select in this stage.
- Click **Next**, and then click **Next** in the Output Variables step.

Figure 12-17
Sampling Wizard, Plan Summary step (stage 1)



- Select Yes, add stage 2 now.
- Click Next.

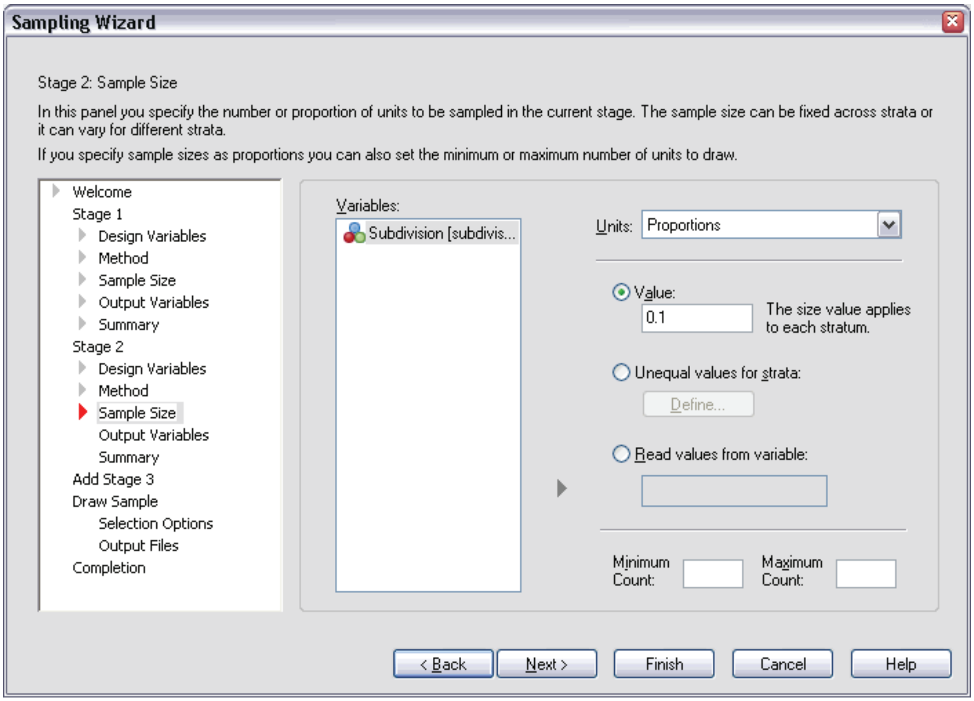
Figure 12-18
Sampling Wizard, Design Variables step (stage 2)



- Select *District* as a stratification variable.
- Select *City* as a cluster variable.
- Click Next, and then click Next in the Method step.

This design structure means that independent samples are drawn for each district. In this stage, cities are drawn as the primary sampling unit using the default method, simple random sampling.

Figure 12-19
Sampling Wizard, Sample Size step (stage 2)

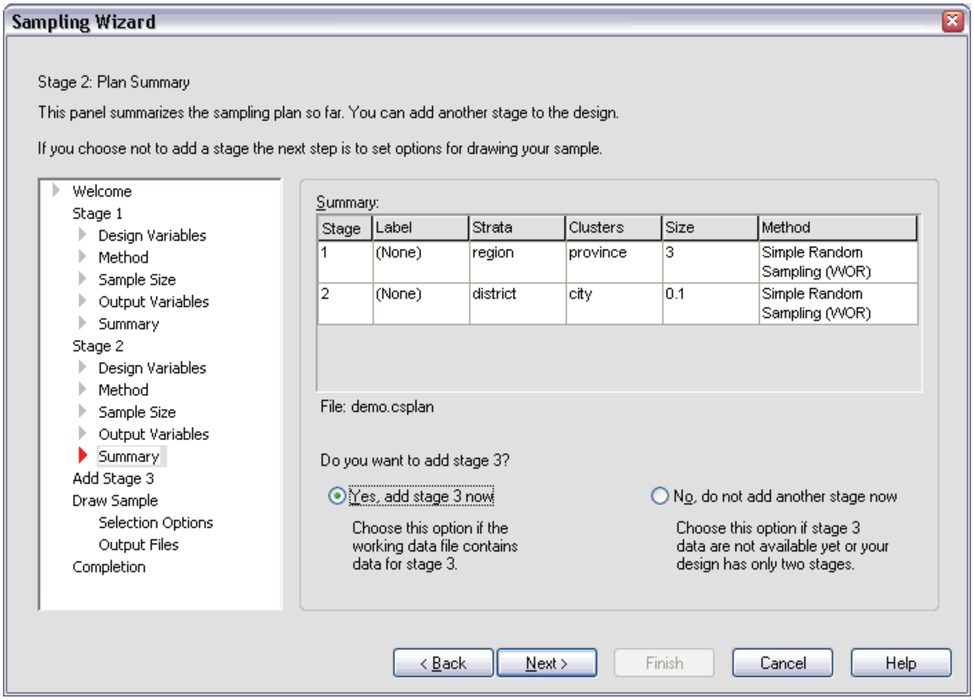


▶ Select Proportions from the Units drop-down list.

▶ Type 0.1 as the value of the proportion of units to sample from each strata.

▶ Click Next, and then click Next in the Output Variables step.

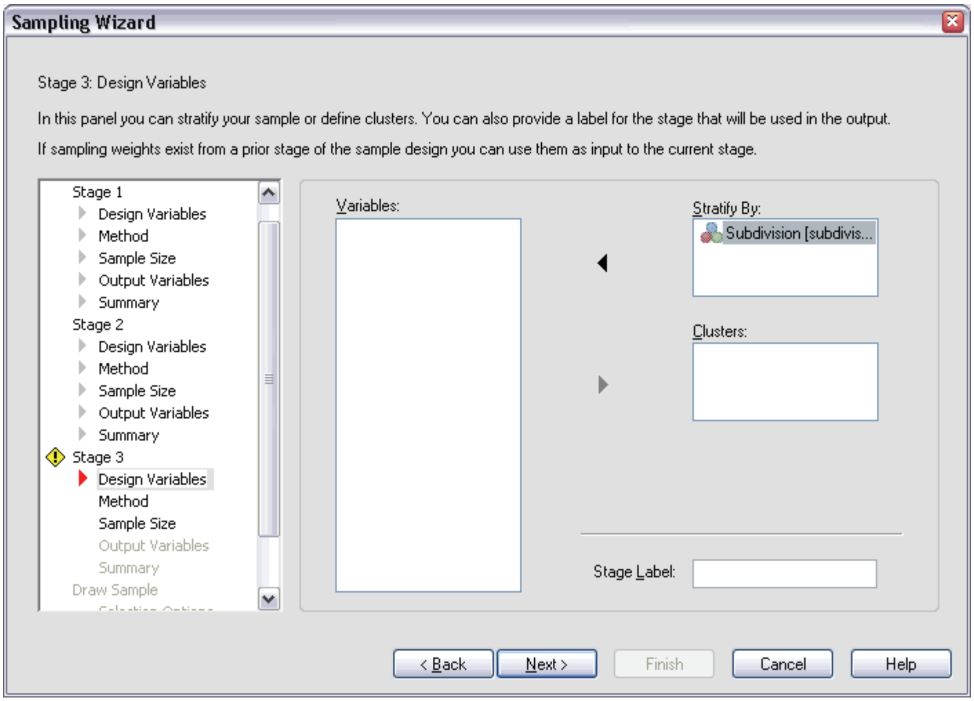
Figure 12-20
Sampling Wizard, Plan Summary step (stage 2)



▶ Select Yes, add stage 3 now.

▶ Click Next.

Figure 12-21
Sampling Wizard, Design Variables step (stage 3)

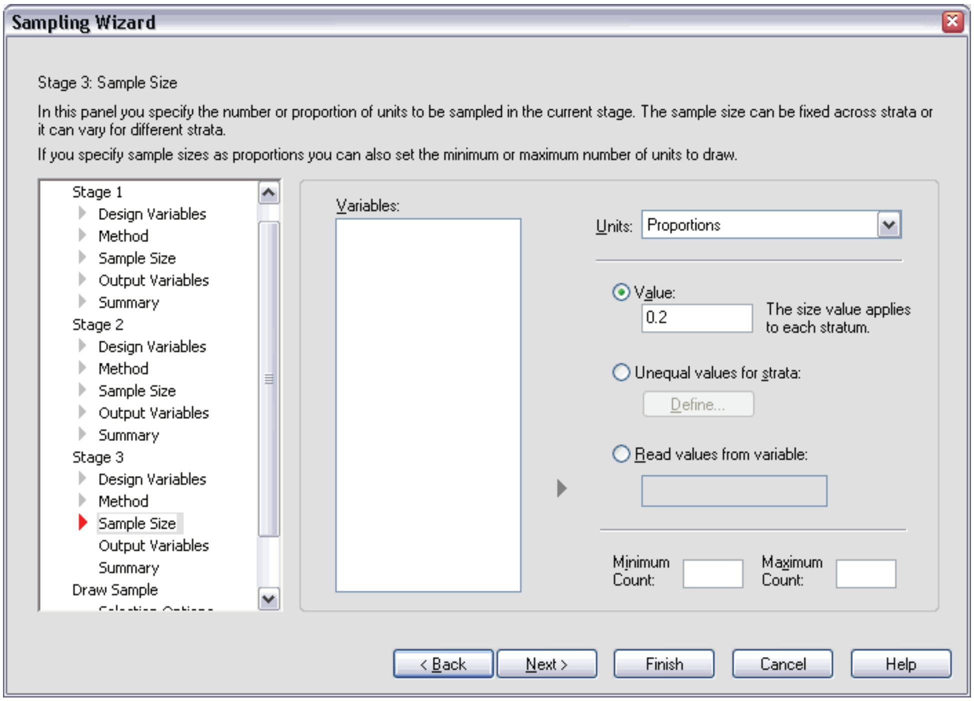


▶ Select *Subdivision* as a stratification variable.

▶ Click Next, and then click Next in the Method step.

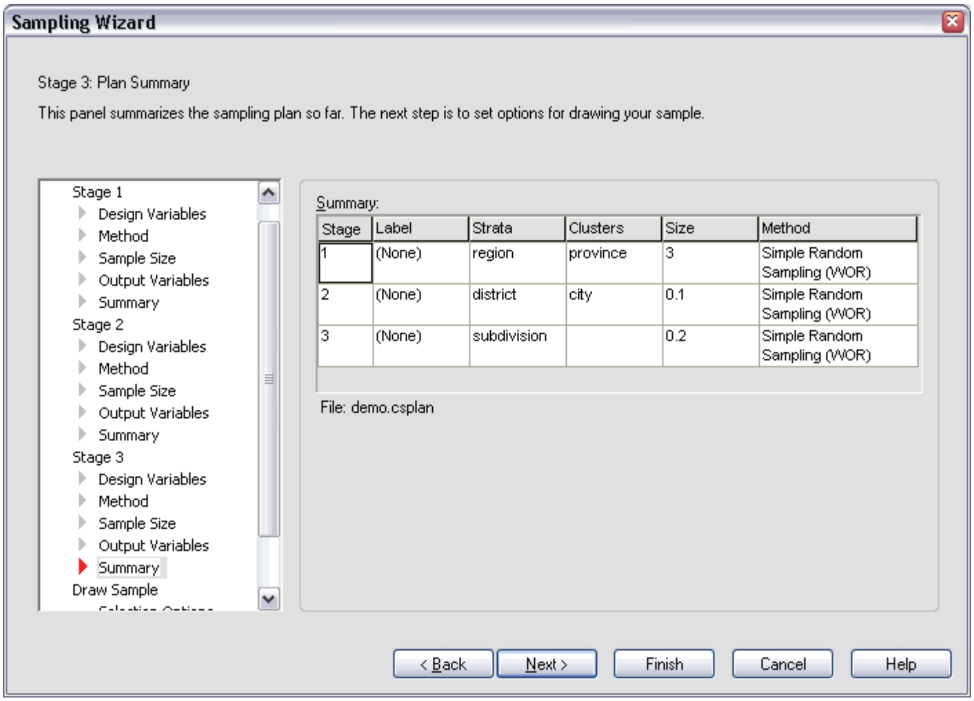
This design structure means that independent samples are drawn for each subdivision. In this stage, household units are drawn as the primary sampling unit using the default method, simple random sampling.

Figure 12-22
Sampling Wizard, Sample Size step (stage 3)



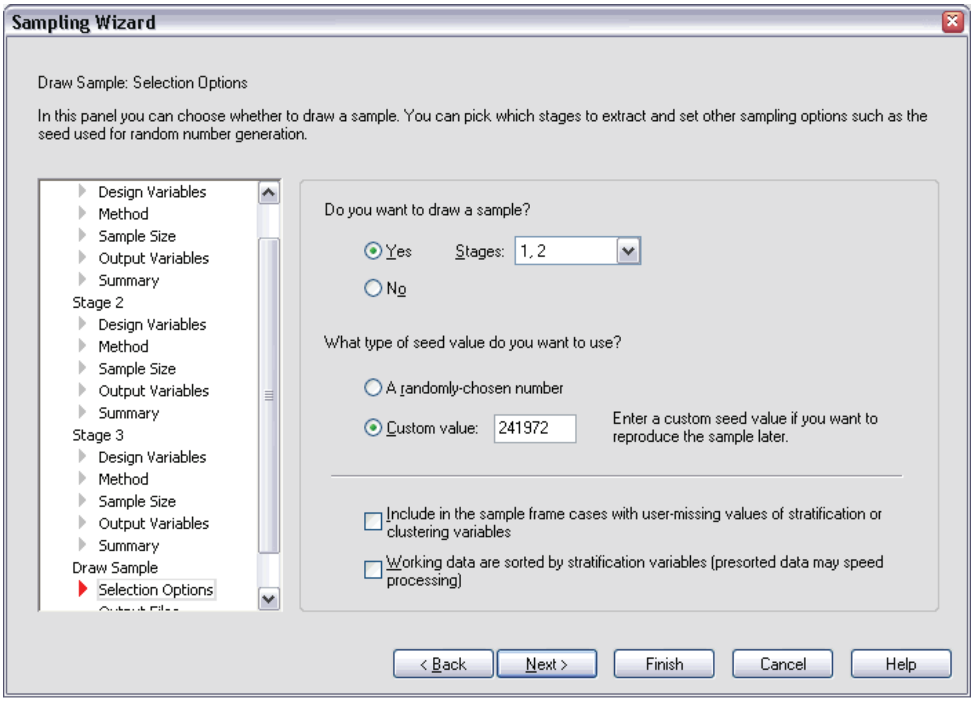
- Select Proportions from the Units drop-down list.
- Type 0.2 as the value for the proportion of units to select in this stage.
- Click Next, and then click Next in the Output Variables step.

Figure 12-23
Sampling Wizard, Plan Summary step (stage 3)



▶ Look over the sampling design, and then click **Next**.

Figure 12-24
Sampling Wizard, Draw Sample, Selection Options step



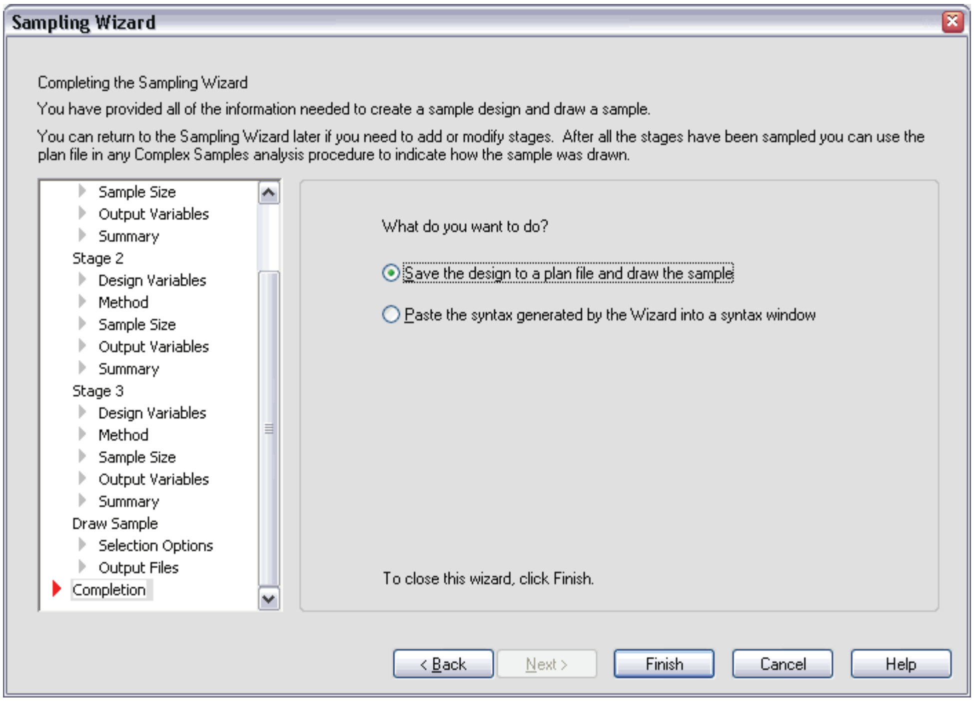
▶ Select 1, 2 as the stages to sample now.

▶ Select Custom value for the type of random seed to use, and type 241972 as the value.

Using a custom value allows you to replicate the results of this example exactly.

▶ Click Next, and then click Next in the Draw Sample, Output Files step.

Figure 12-25
Sampling Wizard, Finish step

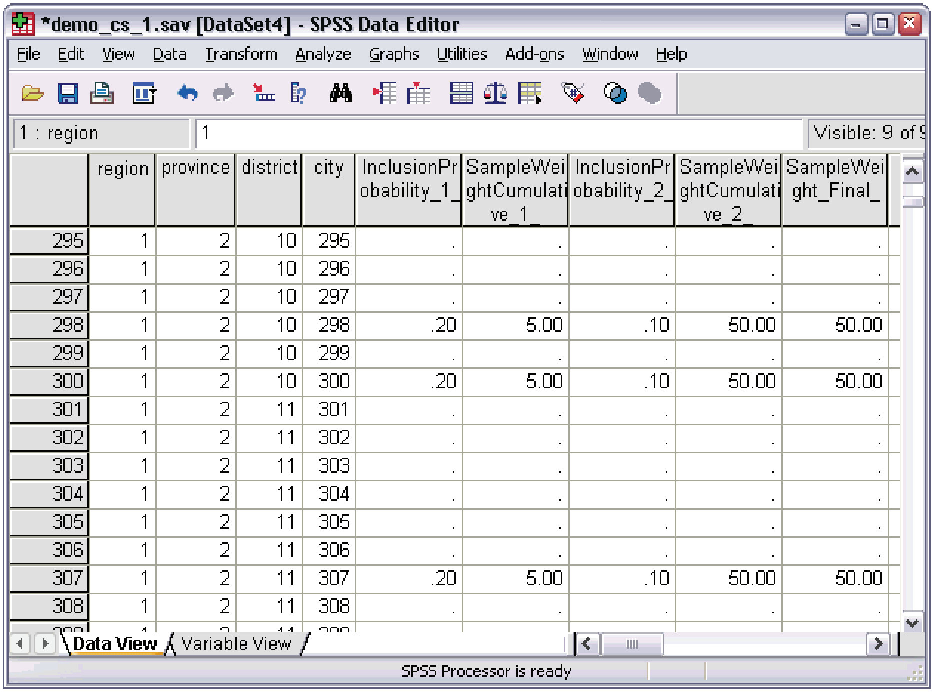


▶ Click Finish.

These selections produce the sampling plan file *demo.csplan* and draw a sample according to the first two stages of that plan.

Sample Results

Figure 12-26
Data Editor with sample results



	region	province	district	city	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	SampleWeight_Final_
295	1	2	10	295	.	.	.	.	.
296	1	2	10	296	.	.	.	.	.
297	1	2	10	297	.	.	.	.	.
298	1	2	10	298	.20	5.00	.10	50.00	50.00
299	1	2	10	299	.	.	.	.	.
300	1	2	10	300	.20	5.00	.10	50.00	50.00
301	1	2	11	301	.	.	.	.	.
302	1	2	11	302	.	.	.	.	.
303	1	2	11	303	.	.	.	.	.
304	1	2	11	304	.	.	.	.	.
305	1	2	11	305	.	.	.	.	.
306	1	2	11	306	.	.	.	.	.
307	1	2	11	307	.20	5.00	.10	50.00	50.00
308	1	2	11	308	.	.	.	.	.

You can see the sampling results in the Data Editor. Five new variables were saved to the working file, representing the inclusion probabilities and cumulative sampling weights for each stage, plus the "final" sampling weights for the first two stages.

- Cities with values for these variables were selected to the sample.
- Cities with system-missing values for the variables were not selected.

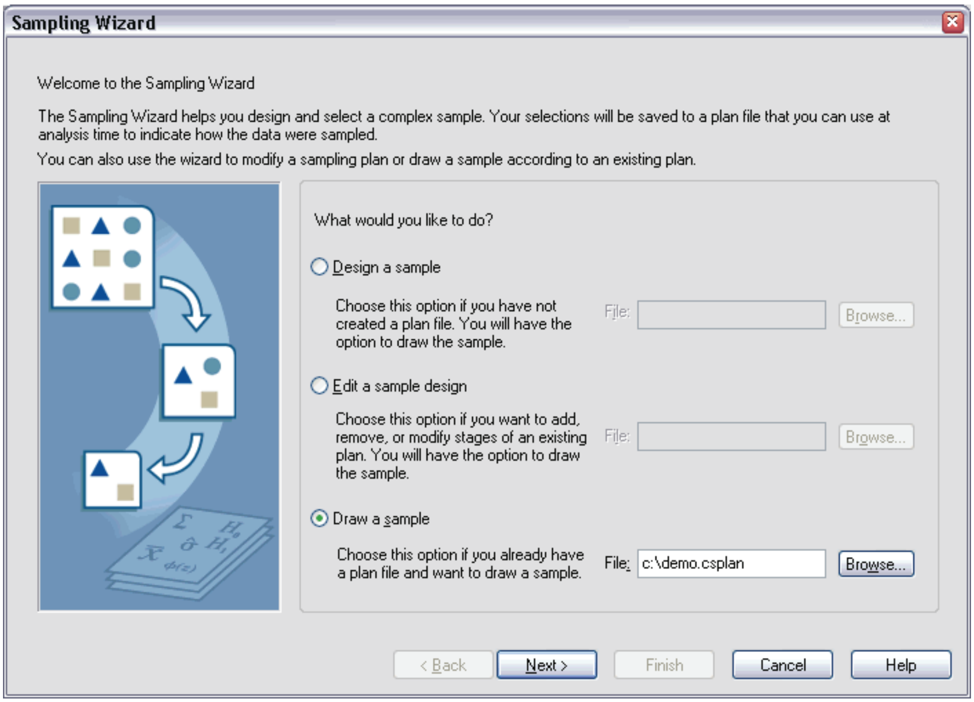
For each city selected, the company acquired subdivision and household unit information and placed it in *demo_cs_2.sav*. Use this file and the Sampling Wizard to sample the third stage of this design.

Using the Wizard to Sample from the Second Partial Frame

▶ To run the Complex Samples Sampling Wizard, from the menus choose:

Analyze
  Complex Samples
    Select a Sample...

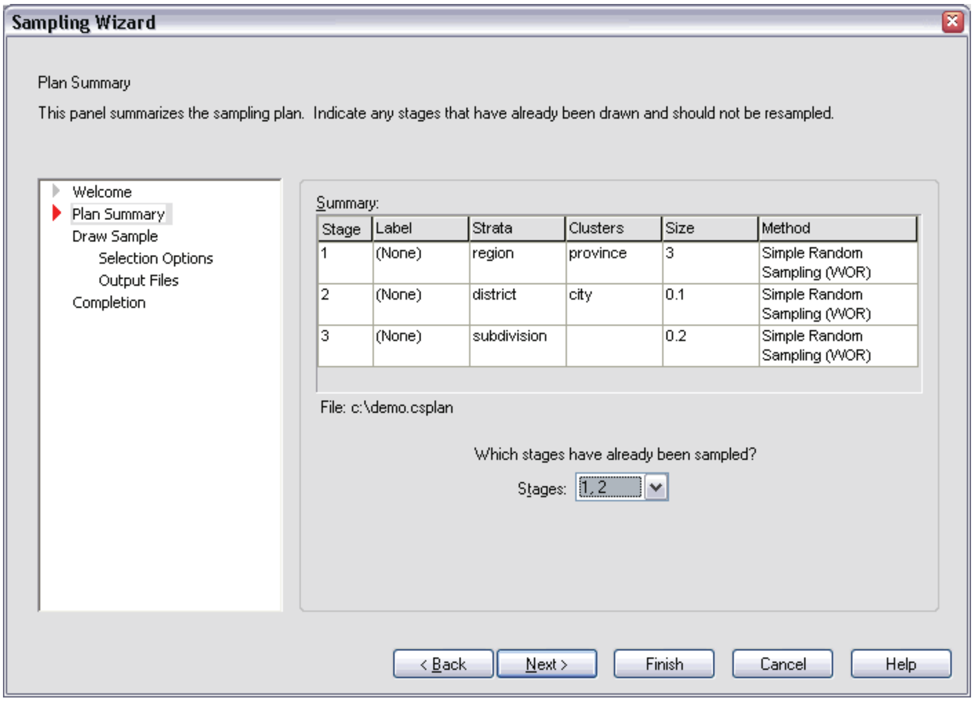
Figure 12-27
Sampling Wizard, Welcome step



▶ Select Draw a sample, and select the c:\demo.csplan plan file that you created.

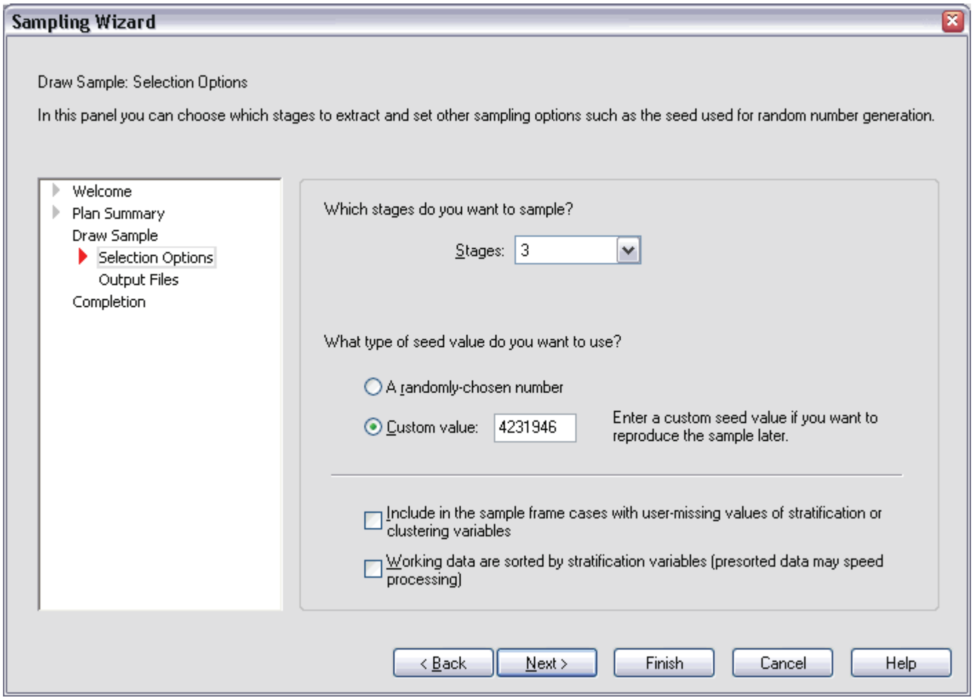
▶ Click Next.

Figure 12-28
Sampling Wizard, Plan Summary step (stage 3)



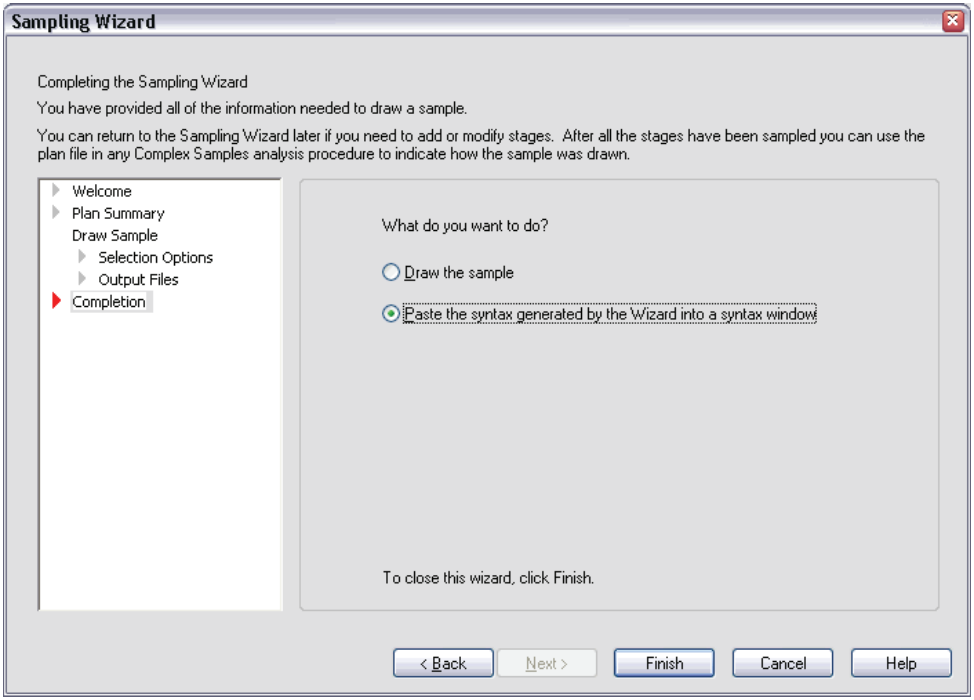
- ▶ Select 1, 2 as stages already sampled.
- ▶ Click Next.

Figure 12-29
Sampling Wizard, Draw Sample, Selection Options step



- Select **Custom value** for the type of random seed to use and type **4231946** as the value.
- Click **Next**, and then click **Next** in the Draw Sample, Output Files step.

Figure 12-30
Sampling Wizard, Finish step



▶ Select Paste the syntax generated by the Wizard into a syntax window.

▶ Click Finish.

The following syntax is generated:

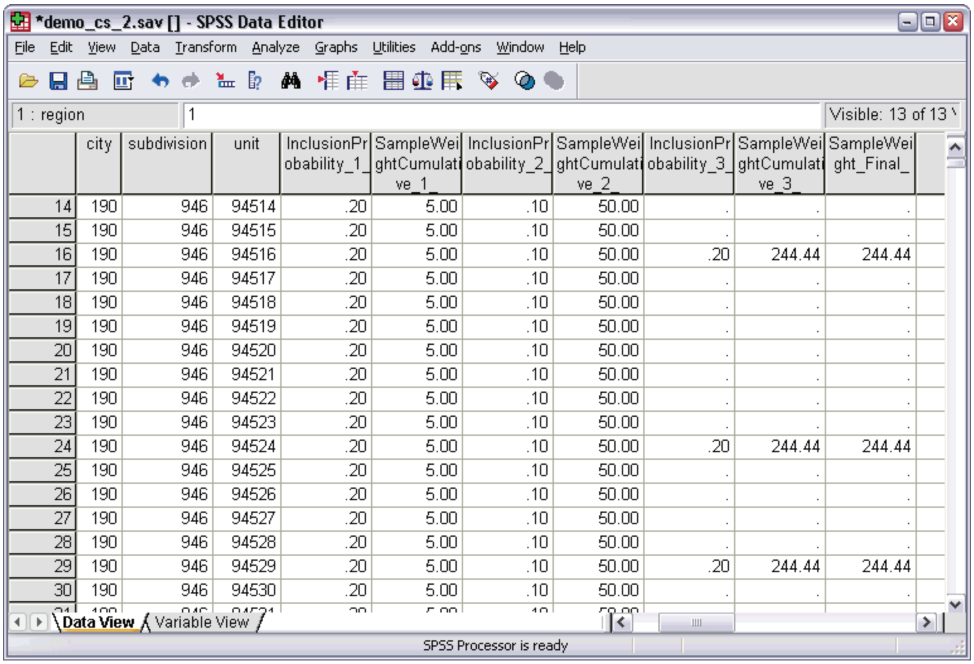
```
* Sampling Wizard.
CSSELECT
 /PLAN FILE='c:\demo.csplan'
 /CRITERIA STAGES = 3 SEED = 4231946
 /CLASSMISSING EXCLUDE
 /DATA RENAMEVARS
 /PRINT SELECTION.
```

Printing the sampling summary in this case produces a cumbersome table that causes problems in the Output Viewer. To turn off display of the sampling summary, replace `SELECTION` with `CPS` in the `PRINT` subcommand. Then run the syntax within the syntax window.

These selections draw a sample according to the third stage of the *demo.csplan* sampling plan.

Sample Results

Figure 12-31
Data Editor with sample results



	city	subdivision	unit	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	InclusionProbability_3_	SampleWeightCumulative_3_	SampleWeight_Final_
14	190	946	94514	.20	5.00	.10	50.00	.	.	.
15	190	946	94515	.20	5.00	.10	50.00	.	.	.
16	190	946	94516	.20	5.00	.10	50.00	.20	244.44	244.44
17	190	946	94517	.20	5.00	.10	50.00	.	.	.
18	190	946	94518	.20	5.00	.10	50.00	.	.	.
19	190	946	94519	.20	5.00	.10	50.00	.	.	.
20	190	946	94520	.20	5.00	.10	50.00	.	.	.
21	190	946	94521	.20	5.00	.10	50.00	.	.	.
22	190	946	94522	.20	5.00	.10	50.00	.	.	.
23	190	946	94523	.20	5.00	.10	50.00	.	.	.
24	190	946	94524	.20	5.00	.10	50.00	.20	244.44	244.44
25	190	946	94525	.20	5.00	.10	50.00	.	.	.
26	190	946	94526	.20	5.00	.10	50.00	.	.	.
27	190	946	94527	.20	5.00	.10	50.00	.	.	.
28	190	946	94528	.20	5.00	.10	50.00	.	.	.
29	190	946	94529	.20	5.00	.10	50.00	.20	244.44	244.44
30	190	946	94530	.20	5.00	.10	50.00	.	.	.

You can see the sampling results in the Data Editor. Three new variables were saved to the working file, representing the inclusion probabilities and cumulative sampling weights for the third stage, plus the final sampling weights. These new weights take into account the weights computed during the sampling of the first two stages.

- Units with values for these variables were selected to the sample.
- Units with system-missing values for these variables were not selected.

The company will now use its resources to obtain survey information for the housing units selected in the sample. Once the surveys are collected, you can process the sample with Complex Samples analysis procedures, using the sampling plan *demo.csplan* to provide the sampling specifications.

Sampling With Probability Proportional to Size (PPS)

Representatives considering a bill before the legislature are interested in whether there is public support for the bill and how support for the bill is related to voter demographics. Pollsters design and conduct interviews according to a complex sampling design.

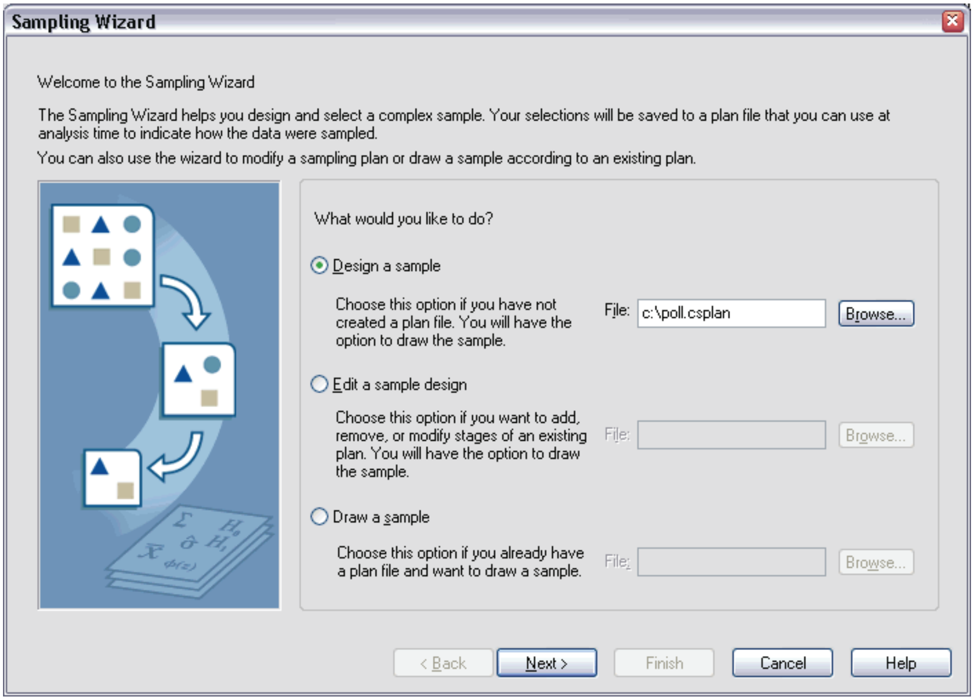
A list of registered voters is collected in *poll_cs.sav*. Use the Complex Samples Sampling Wizard to select a sample for further analysis.

Using the Wizard

▶ To run the Complex Samples Sampling Wizard, from the menus choose:

Analyze
  Complex Samples
    Select a Sample...

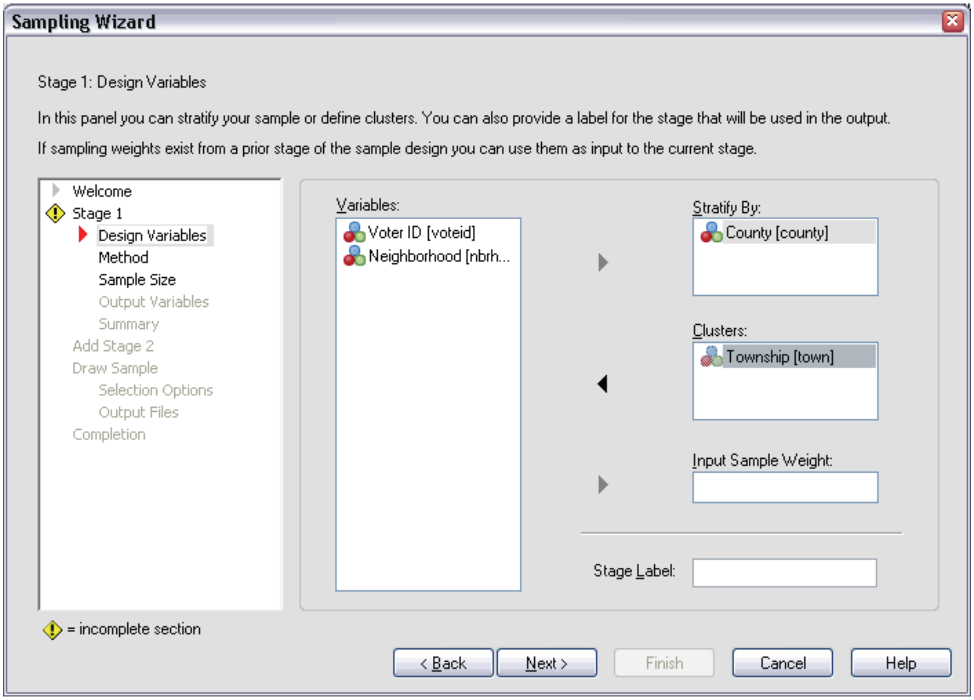
Figure 12-32
Sampling Wizard, Welcome step



▶ Select Design a sample, and type c:\poll.csplan as the name of the plan file.

▶ Click Next.

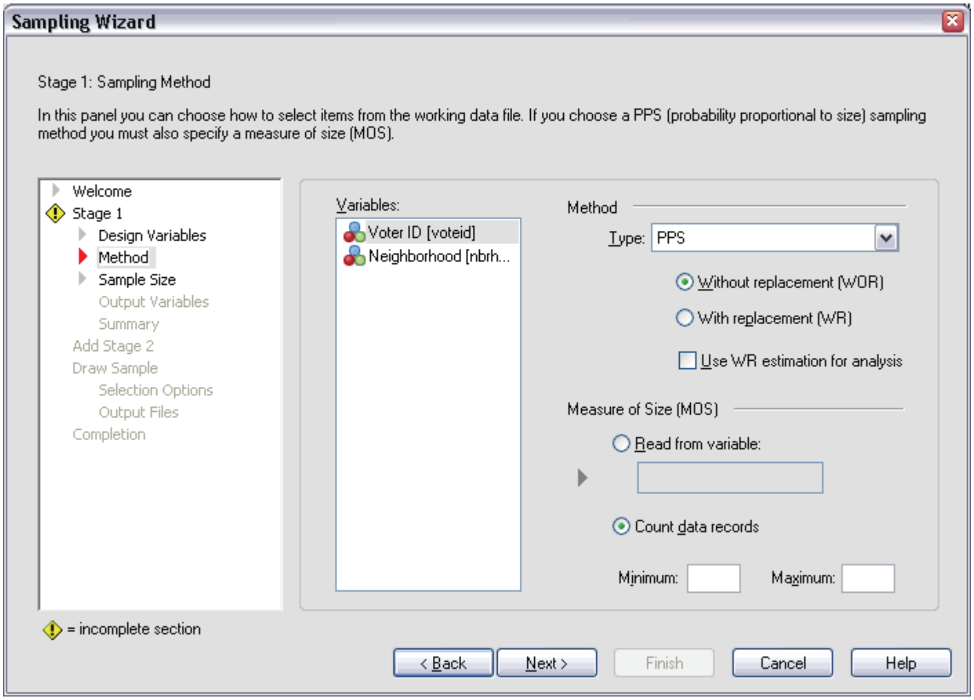
Figure 12-33
Sampling Wizard, Design Variables step (stage 1)



- Select *County* as a stratification variable.
- Select *Township* as a cluster variable.
- Click Next.

This design structure means that independent samples are drawn for each county. In this stage, townships are drawn as the primary sampling unit.

Figure 12-34
Sampling Wizard, Method step (stage 1)



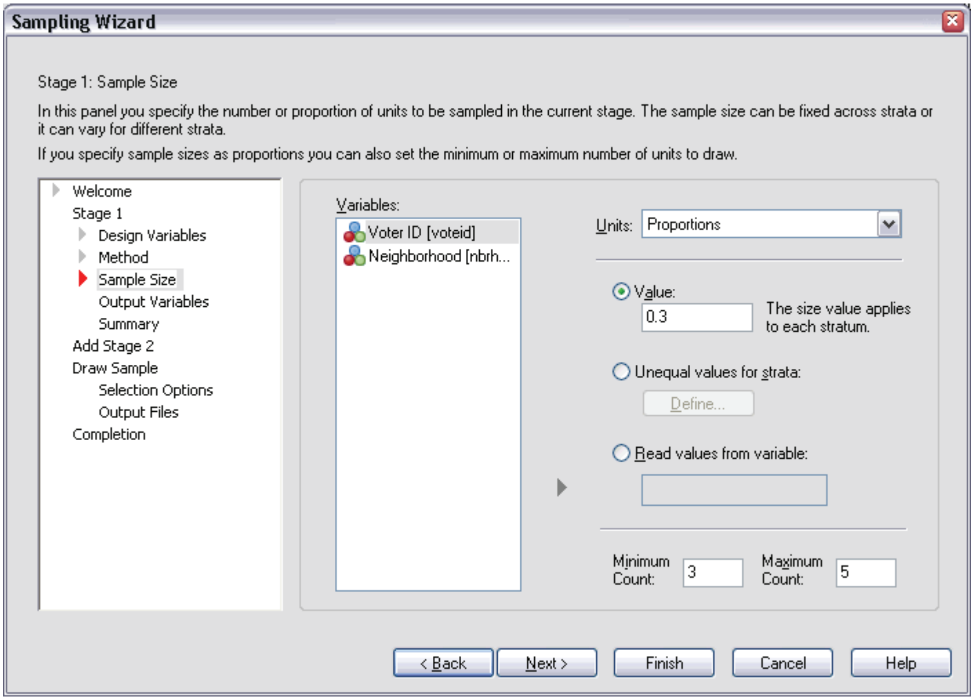
▶ Select PPS as the sampling method.

▶ Select Count data records as the measure of size.

▶ Click Next.

Within each county, townships are drawn without replacement with probability proportional to the number of records for each township. Using a PPS method generates joint sampling probabilities for the townships; you will specify where to save these values in the Output Files step.

Figure 12-35
Sampling Wizard, Sample Size step (stage 1)

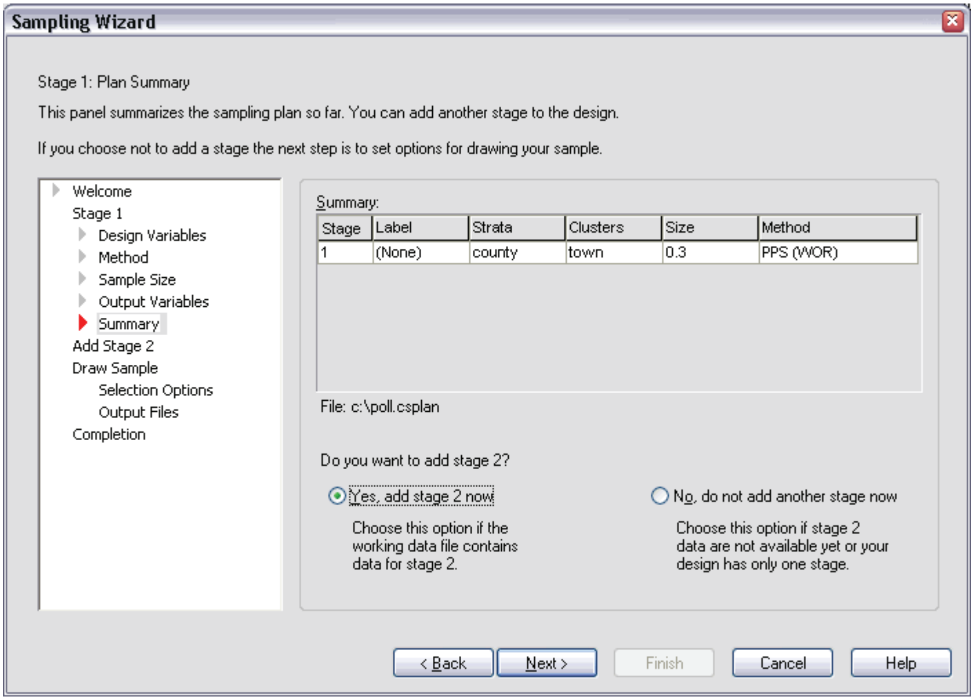


- ▶ Select Proportions from the Units drop-down list.
- ▶ Type 0.3 as the value for the proportion of townships to select per county in this stage.

  Legislators from the Western county point out that there are fewer townships in their county than in others. In order to ensure adequate representation, they would like to establish a minimum of 3 townships sampled from each county.

- ▶ Type 3 as the minimum number of townships to select and 5 as the maximum.
- ▶ Click Next, and then click Next in the Output Variables step.

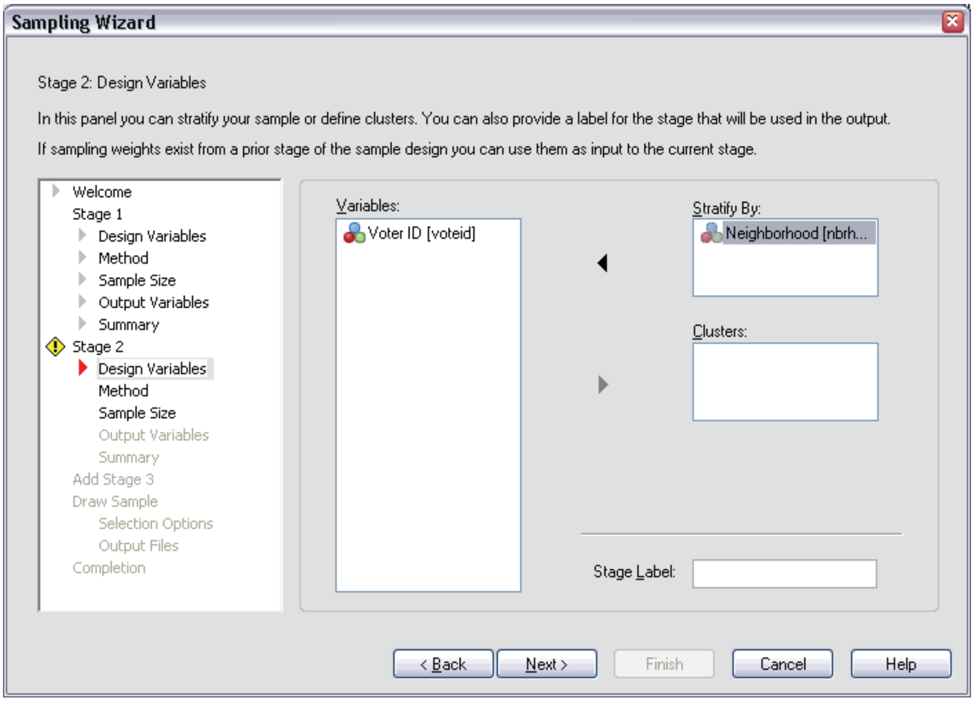
Figure 12-36
Sampling Wizard, Plan Summary step (stage 1)



▶ Select Yes, add stage 2 now.

▶ Click Next.

Figure 12-37
Sampling Wizard, Design Variables step (stage 2)

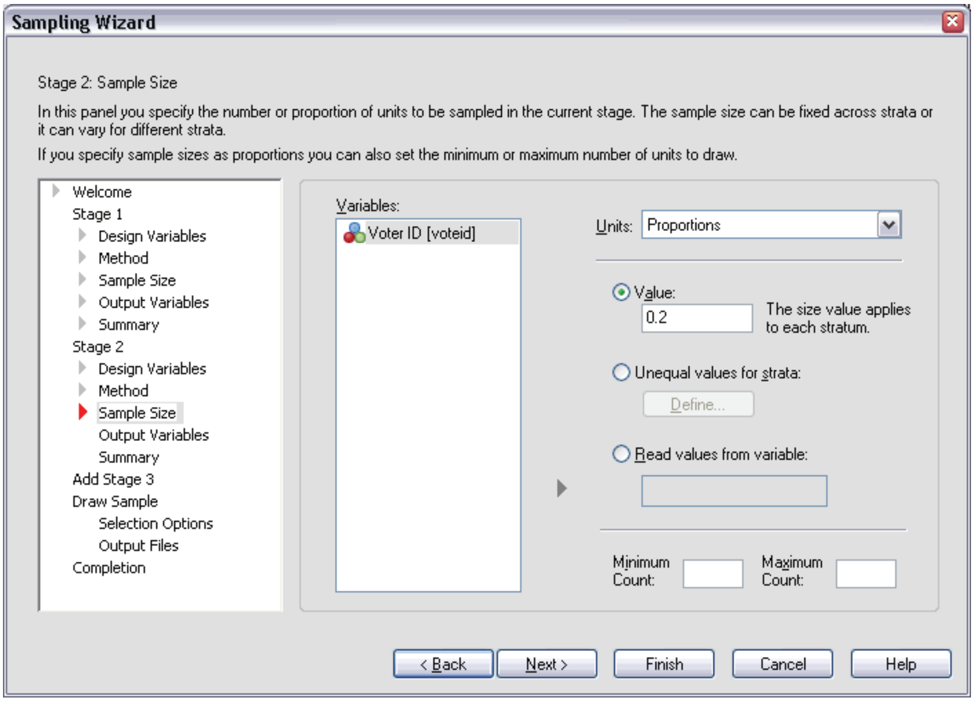


▶ Select *Neighborhood* as a stratification variable.

▶ Click **Next**, and then click **Next** in the Method step.

This design structure means that independent samples are drawn for each neighborhood of the townships drawn in stage 1. In this stage, voters are drawn as the primary sampling unit using simple random sampling without replacement.

Figure 12-38
Sampling Wizard, Sample Size step (stage 2)

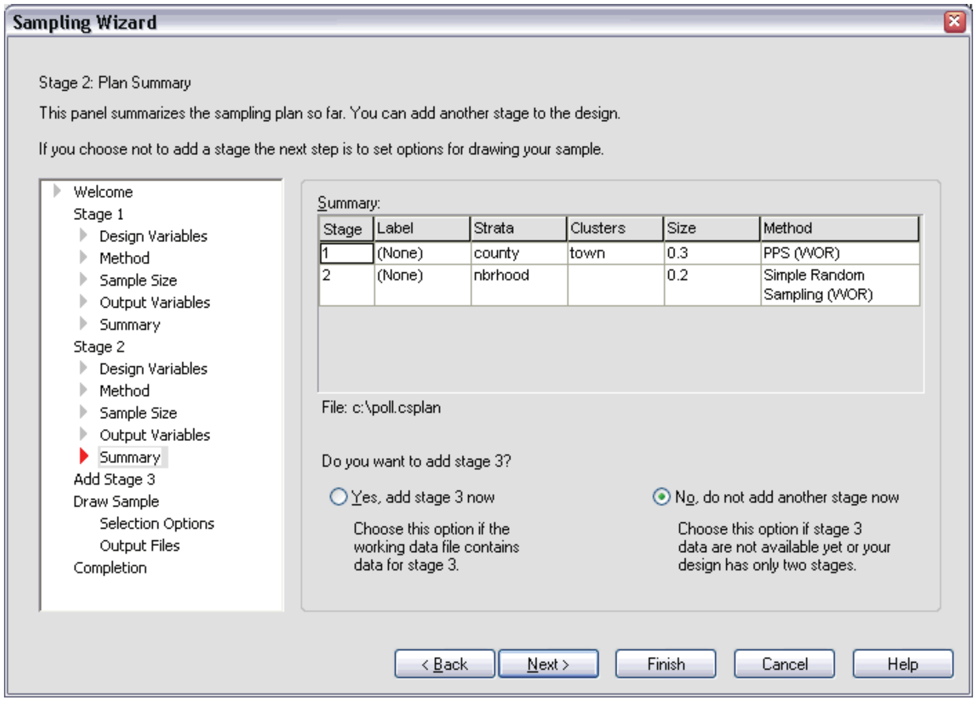


▶ Select Proportions from the Units drop-down list.

▶ Type 0.2 as the value of the proportion of units to sample from each strata.

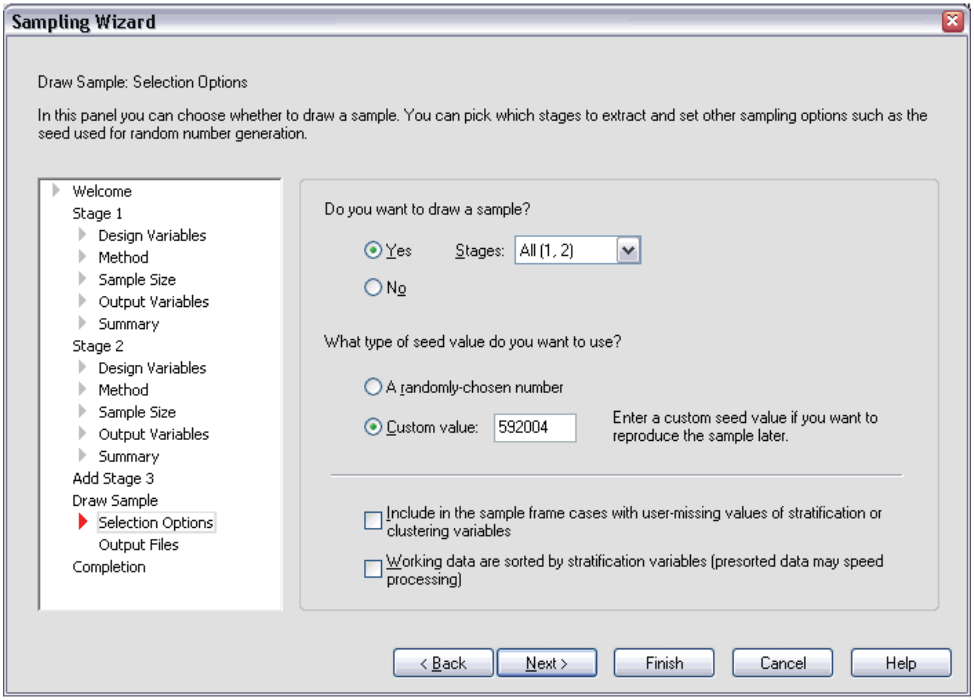
▶ Click Next, and then click Next in the Output Variables step.

Figure 12-39
Sampling Wizard, Plan Summary step (stage 2)



▶ Look over the sampling design, and then click **Next**.

Figure 12-40
Sampling Wizard, Draw Sample, Selection Options step

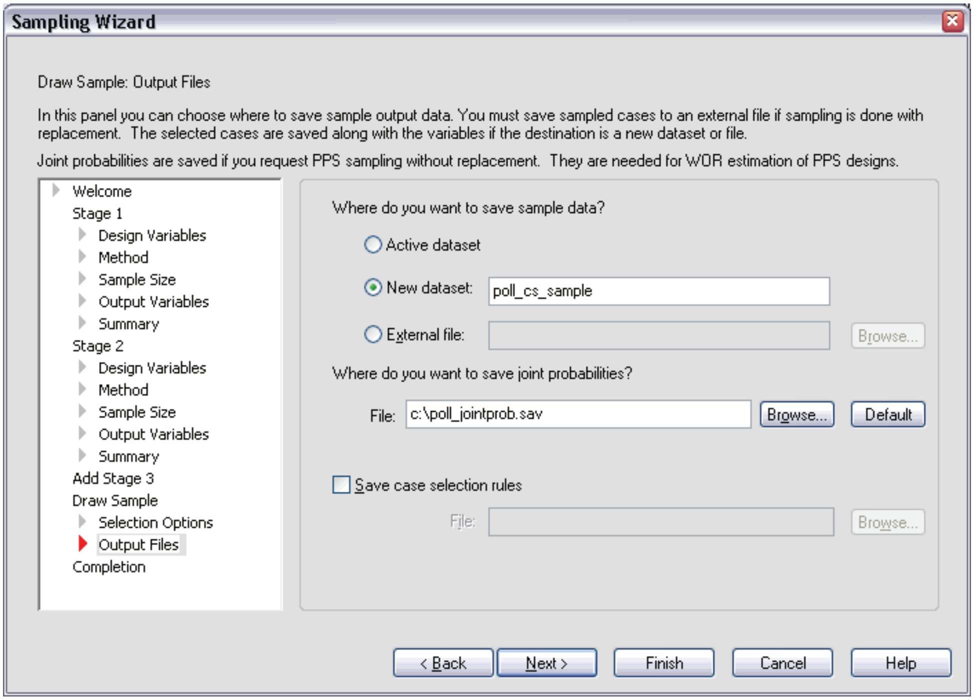


- ▶ Select Custom value for the type of random seed to use, and type 592004 as the value.

  Using a custom value allows you to replicate the results of this example exactly.

- ▶ Click Next.

Figure 12-41
Sampling Wizard, Draw Sample, Selection Options step

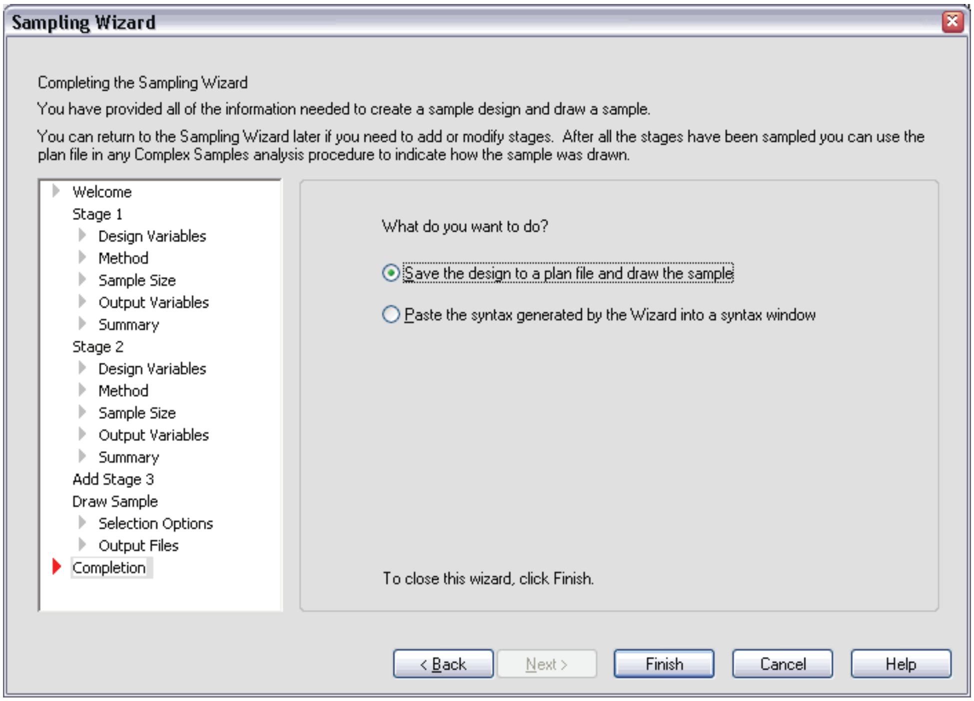


▶ Choose to save the sample to a new dataset, and type c:\poll_cs_sample as the name of the dataset.

▶ Type c:\poll_jointprob.sav as the name of the joint probabilities file.

▶ Click Next.

Figure 12-42
Sampling Wizard, Finish step



▶ Click Finish.

These selections produce the sampling plan file *poll.csplan* and draw a sample according to that plan, save the sample results to the new dataset *poll_cs_sample*, and save the joint probabilities file to the external data file *poll_jointprob.sav*.

Plan Summary

Figure 12-43
Plan summary

			Stage 1	Stage 2
Design Variables	Stratification	1	County	Neighborhood
	Cluster	1	Township	
Sample Information	Selection Method		PPS sampling without replacement	Simple random sampling without replacement
	Measure of Size		Obtained from data	
	Proportion of Units Sampled		.3	.2
	Minimum Number of Units Sampled		3	
	Maximum Number of Units Sampled		5	
	Variables Created or Modified	Stagewise Inclusion (Selection) Probability	InclusionProbability_1_	InclusionProbability_2_
		Stagewise Cumulative Sample Weight	SampleWeightCumulative_1_	SampleWeightCumulative_2_
Analysis Information	Estimator Assumption		Unequal probability sampling without replacement (using joint inclusion probabilities)	Equal probability sampling without replacement
	Inclusion Probability		Obtained from variable InclusionProbability_1_	Obtained from variable InclusionProbability_2_

Plan File: c:\poll.csplan
Weight Variable: SampleWeight_Final_

The summary table reviews your sampling plan, and it is useful for making sure that the plan represents your intentions.

Sampling Summary

Figure 12-44
Stage summary

	Number of Units Sampled		Proportion of Units Sampled	
County	Requested	Actual	Requested	Actual
Eastern	4	4	30.0%	30.8%
Central	4	4	30.0%	30.8%
Western	3	3	30.0%	50.0%
Northern	5	5	30.0%	33.3%
Southern	3	3	30.0%	50.0%

Plan File: c:\poll.csplan

This summary table reviews the first stage of sampling, and it is useful for checking that the sampling went according to plan. Recall that you requested a 30% sample of townships by county; the actual proportions sampled are close to 30%, except in the Western and Southern counties. This is because these counties each have only six townships, and you also specified that a minimum of three townships should be selected per county.

Figure 12-45
Stage summary

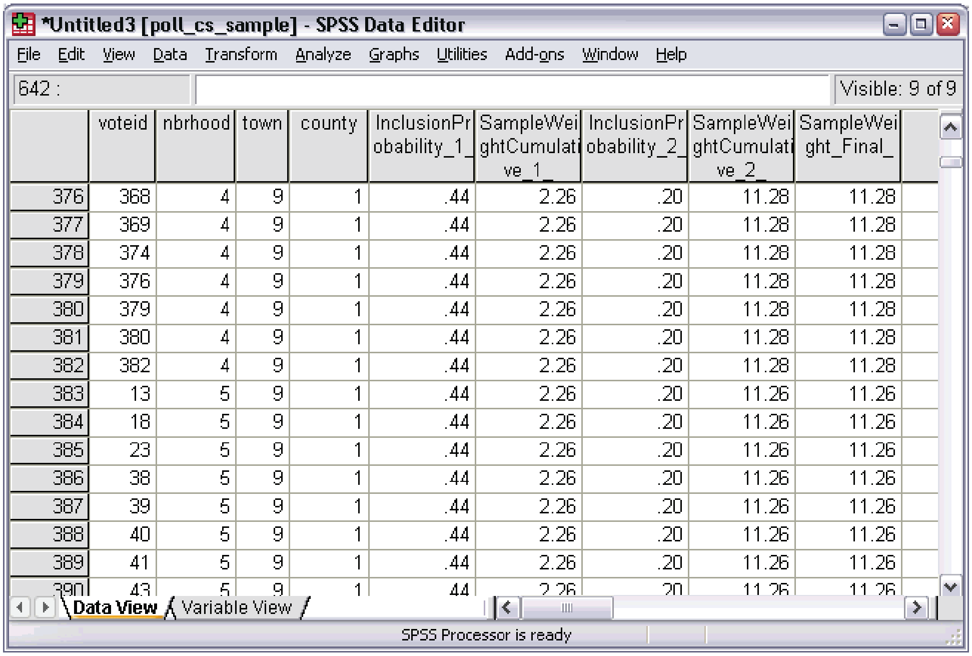
			Number of Units Sampled		Proportion of Units Sampled	
County	Township	Neighborhood	Requested	Actual	Requested	Actual
Eastern	9	1	49	49	20.0%	19.9%
		2	143	143	20.0%	20.0%
		3	113	113	20.0%	20.0%
		4	77	77	20.0%	20.0%
		5	139	139	20.0%	20.0%
		6	120	120	20.0%	20.0%
	10	1	149	149	20.0%	20.1%
		2	117	117	20.0%	20.0%
		3	116	116	20.0%	20.0%
		4	69	69	20.0%	19.9%
	11	1	65	65	20.0%	19.9%
		2	72	72	20.0%	19.9%
		3	109	109	20.0%	20.0%
		4	140	140	20.0%	20.0%
		5	42	42	20.0%	19.8%
		6	142	142	20.0%	20.0%
	12	1	145	145	20.0%	20.1%
		2	69	69	20.0%	20.1%
		3	98	98	20.0%	20.1%
		4	134	134	20.0%	20.0%
		5	114	114	20.0%	20.0%
		6	137	137	20.0%	19.9%
Central	2	1	119	119	20.0%	20.1%
		2	153	153	20.0%	19.9%
		3	101	101	20.0%	20.0%
		4	52	52	20.0%	19.8%
		5	144	144	20.0%	20.0%

Plan File: c:\poll.csplan

This summary table (the top part of which is shown here) reviews the second stage of sampling. It is also useful for checking that the sampling went according to plan. Approximately 20% of the voters were sampled from each neighborhood from each township sampled in the first stage, as requested.

Sample Results

Figure 12-46
Data Editor with sample results

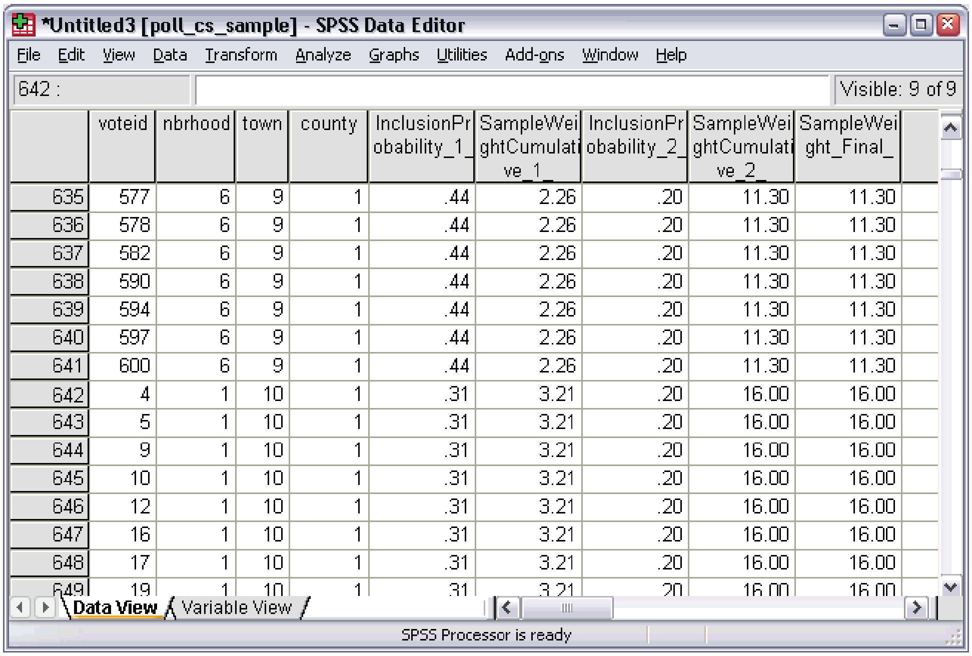


	voteid	nbrhood	town	county	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	SampleWeight_Final_
376	368	4	9	1	.44	2.26	.20	11.28	11.28
377	369	4	9	1	.44	2.26	.20	11.28	11.28
378	374	4	9	1	.44	2.26	.20	11.28	11.28
379	376	4	9	1	.44	2.26	.20	11.28	11.28
380	379	4	9	1	.44	2.26	.20	11.28	11.28
381	380	4	9	1	.44	2.26	.20	11.28	11.28
382	382	4	9	1	.44	2.26	.20	11.28	11.28
383	13	5	9	1	.44	2.26	.20	11.26	11.26
384	18	5	9	1	.44	2.26	.20	11.26	11.26
385	23	5	9	1	.44	2.26	.20	11.26	11.26
386	38	5	9	1	.44	2.26	.20	11.26	11.26
387	39	5	9	1	.44	2.26	.20	11.26	11.26
388	40	5	9	1	.44	2.26	.20	11.26	11.26
389	41	5	9	1	.44	2.26	.20	11.26	11.26
390	43	5	9	1	.44	2.26	.20	11.26	11.26

You can see the sampling results in the newly created dataset. Five new variables were saved to the working file, representing the inclusion probabilities and cumulative sampling weights for each stage, plus the final sampling weights. Voters who were not selected to the sample are excluded from this dataset.

The final sampling weights are identical for voters within the same neighborhood because they are selected according to a simple random sampling method within neighborhoods. However, they are different across neighborhoods within the same township because the sampled proportions are not exactly 20% in all neighborhoods.

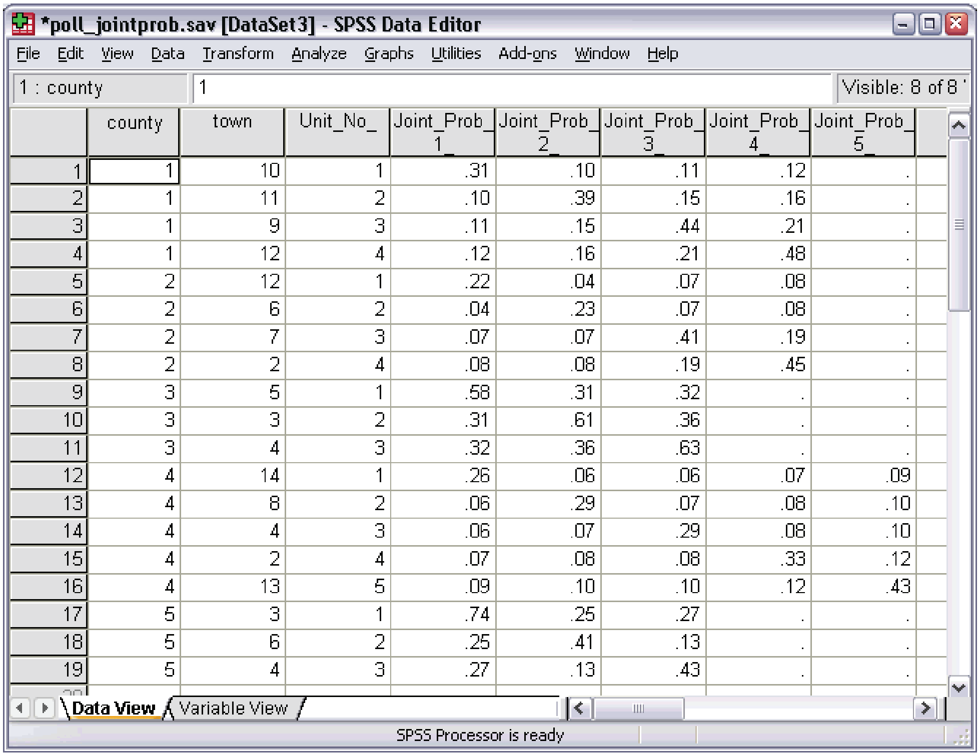
Figure 12-47
Data Editor with sample results



	voteid	nbrhood	town	county	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	SampleWeight_Final_
635	577	6	9	1	.44	2.26	.20	11.30	11.30
636	578	6	9	1	.44	2.26	.20	11.30	11.30
637	582	6	9	1	.44	2.26	.20	11.30	11.30
638	590	6	9	1	.44	2.26	.20	11.30	11.30
639	594	6	9	1	.44	2.26	.20	11.30	11.30
640	597	6	9	1	.44	2.26	.20	11.30	11.30
641	600	6	9	1	.44	2.26	.20	11.30	11.30
642	4	1	10	1	.31	3.21	.20	16.00	16.00
643	5	1	10	1	.31	3.21	.20	16.00	16.00
644	9	1	10	1	.31	3.21	.20	16.00	16.00
645	10	1	10	1	.31	3.21	.20	16.00	16.00
646	12	1	10	1	.31	3.21	.20	16.00	16.00
647	16	1	10	1	.31	3.21	.20	16.00	16.00
648	17	1	10	1	.31	3.21	.20	16.00	16.00
649	19	1	10	1	.31	3.21	.20	16.00	16.00

Unlike voters in the second stage, the first-stage sampling weights are not identical for townships within the same county because they are selected with probability proportional to size.

Figure 12-48
Joint probabilities file



	county	town	Unit_No_	Joint_Prob_1_	Joint_Prob_2_	Joint_Prob_3_	Joint_Prob_4_	Joint_Prob_5_
1	1	10	1	.31	.10	.11	.12	.
2	1	11	2	.10	.39	.15	.16	.
3	1	9	3	.11	.15	.44	.21	.
4	1	12	4	.12	.16	.21	.48	.
5	2	12	1	.22	.04	.07	.08	.
6	2	6	2	.04	.23	.07	.08	.
7	2	7	3	.07	.07	.41	.19	.
8	2	2	4	.08	.08	.19	.45	.
9	3	5	1	.58	.31	.32	.	.
10	3	3	2	.31	.61	.36	.	.
11	3	4	3	.32	.36	.63	.	.
12	4	14	1	.26	.06	.06	.07	.09
13	4	8	2	.06	.29	.07	.08	.10
14	4	4	3	.06	.07	.29	.08	.10
15	4	2	4	.07	.08	.08	.33	.12
16	4	13	5	.09	.10	.10	.12	.43
17	5	3	1	.74	.25	.27	.	.
18	5	6	2	.25	.41	.13	.	.
19	5	4	3	.27	.13	.43	.	.

The file *poll_jointprob.sav* contains first-stage joint probabilities for selected townships within counties. *County* is a first-stage stratification variable, and *Township* is a cluster variable. Combinations of these variables identify all first-stage PSUs uniquely. *Unit_No_* labels PSUs within each stratum and is used to match up with *Joint_Prob_1_*, *Joint_Prob_2_*, *Joint_Prob_3_*, *Joint_Prob_4_*, and *Joint_Prob_5_*. The first two strata each have 4 PSUs; therefore, the joint inclusion probability matrices are 4×4 for these strata, and the *Joint_Prob_5_* column is left empty for these rows. Similarly, strata 3 and 5 have 3×3 joint inclusion probability matrices, and stratum 4 has a 5×5 joint inclusion probability matrix.

The need for a joint probabilities file is seen by perusing the values of the joint inclusion probability matrices. When the sampling method is not a PPS WOR method, the selection of a PSU is independent of the selection of another PSU, and their joint inclusion probability is simply the product of their inclusion probabilities. In contrast, the joint inclusion probability for Townships 9 and 10 of County 1 is approximately 0.11 (see the first case of *Joint_Prob_3_* or the third case of *Joint_Prob_1_*), or less than the product of their individual inclusion probabilities (the product of the first case of *Joint_Prob_1_* and the third case of *Joint_Prob_3_* is 0.31×0.44=0.1364).

The pollsters will now conduct interviews for the selected sample. Once the results are available, you can process the sample with Complex Samples analysis procedures, using the sampling plan *poll.csplan* to provide the sampling specifications and *poll_jointprob.sav* to provide the needed joint inclusion probabilities.

Related Procedures

The Complex Samples Sampling Wizard procedure is a useful tool for creating a sampling plan file and drawing a sample.

- To ready a sample for analysis when you do not have access to the sampling plan file, use the Analysis Preparation Wizard.

Chapter

13

Complex Samples Analysis Preparation Wizard

The Analysis Preparation Wizard guides you through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. It is most useful when you do not have access to the sampling plan file used to draw the sample.

Using the Complex Samples Analysis Preparation Wizard to Ready NHIS Public Data

The National Health Interview Survey (NHIS) is a large, population-based survey of the U.S. civilian population. Interviews are carried out face-to-face in a nationally representative sample of households. Demographic information and observations about health behavior and status are obtained for members of each household.

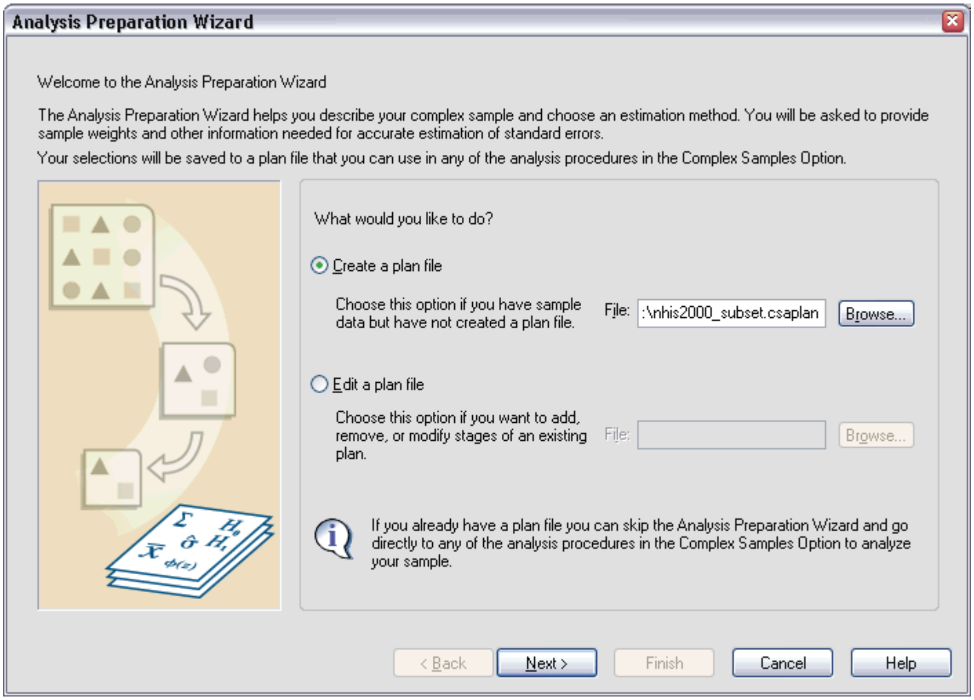
A subset of the 2000 survey is collected in *nhis2000_subset.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. Use the Complex Samples Analysis Preparation Wizard to create an analysis plan for this data file so that it can be processed by Complex Samples analysis procedures.

Using the Wizard

▶ To prepare a sample using the Complex Samples Analysis Preparation Wizard, from the menus choose:

Analyze
  Complex Samples
    Prepare for Analysis...

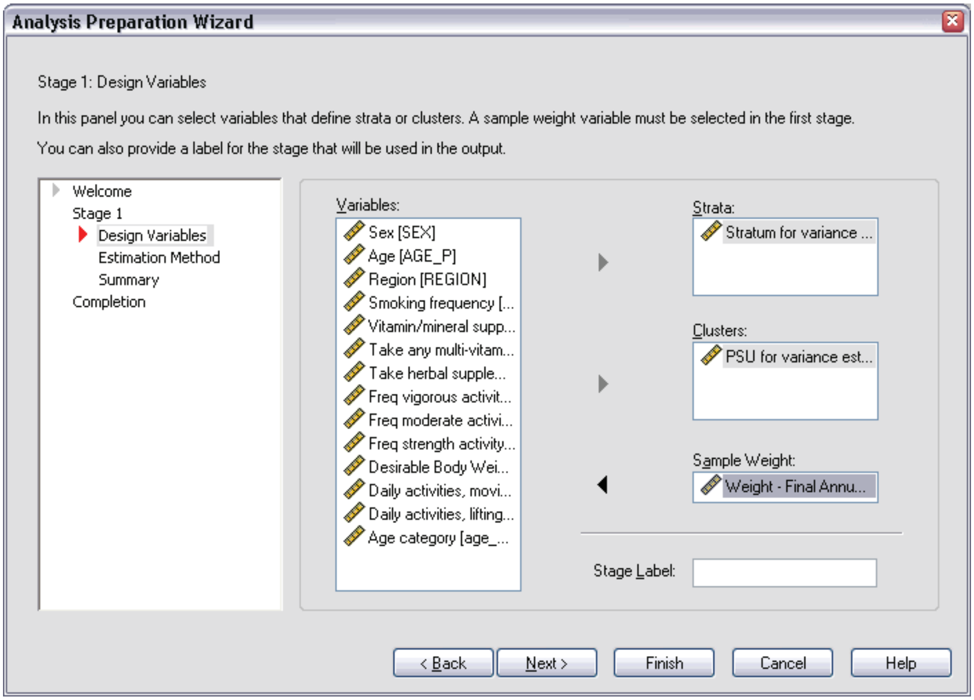
Figure 13-1
Analysis Preparation Wizard, Welcome step



▶ Enter c:\nhis2000_subset.csaplan as the name for the analysis plan file.

▶ Click Next.

Figure 13-2
Analysis Preparation Wizard, Design Variables step (stage 1)



The data are obtained using a complex multistage sample. However, for end users, the original NHIS design variables were transformed to a simplified set of design and weight variables whose results approximate those of the original design structures.

- Select *Stratum for variance estimation* as a strata variable.
- Select *PSU for variance estimation* as a cluster variable.
- Select *Weight - Final Annual* as the sample weight variable.
- Click Finish.

Summary

Figure 13-3
Summary

			Stage 1
Design Variables	Stratification	1	Stratum for variance estimation
	Cluster	1	PSU for variance estimation
Analysis Information	Estimator Assumption		Sampling with replacement

Plan File: c:\nhis2000_subset.csaplan
Weight Variable: Weight - Final Annual
SRS Estimator: Sampling without replacement

The summary table reviews your analysis plan. The plan consists of one stage with a design of one stratification variable and one cluster variable. With-replacement (WR) estimation is used, and the plan is saved to *c:/nhis2000_subset.csaplan*. You can now use this plan file to process *nhis2000_subset.sav* with Complex Samples analysis procedures.

Preparing for Analysis When Sampling Weights Are Not in the Data File

A loan officer has a collection of customer records, taken according to a complex design; however, the sampling weights are not included in the file. This information is contained in *bankloan_cs_noweights.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. Starting with what she knows about the sampling design, the officer wants to use the Complex Samples Analysis Preparation Wizard to create an analysis plan for this data file so that it can be processed by Complex Samples analysis procedures.

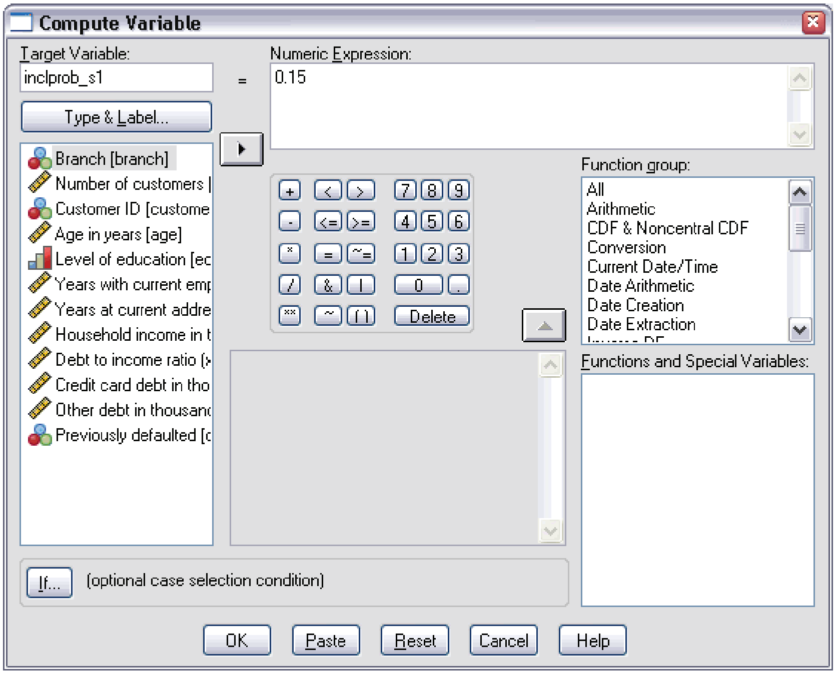
The loan officer knows that the records were selected in two stages, with 15 out of 100 bank branches selected with equal probability and without replacement in the first stage. One hundred customers were then selected from each of those banks with equal probability and without replacement in the second stage, and information on the number of customers at each bank is included in the data file. The first step to creating an analysis plan is to compute the stagewise inclusion probabilities and final sampling weights.

Computing Inclusion Probabilities and Sampling Weights

▶ To compute the inclusion probabilities for the first stage, from the menus choose:

Transform
  Compute Variable...

Figure 13-4
Compute Variable dialog box



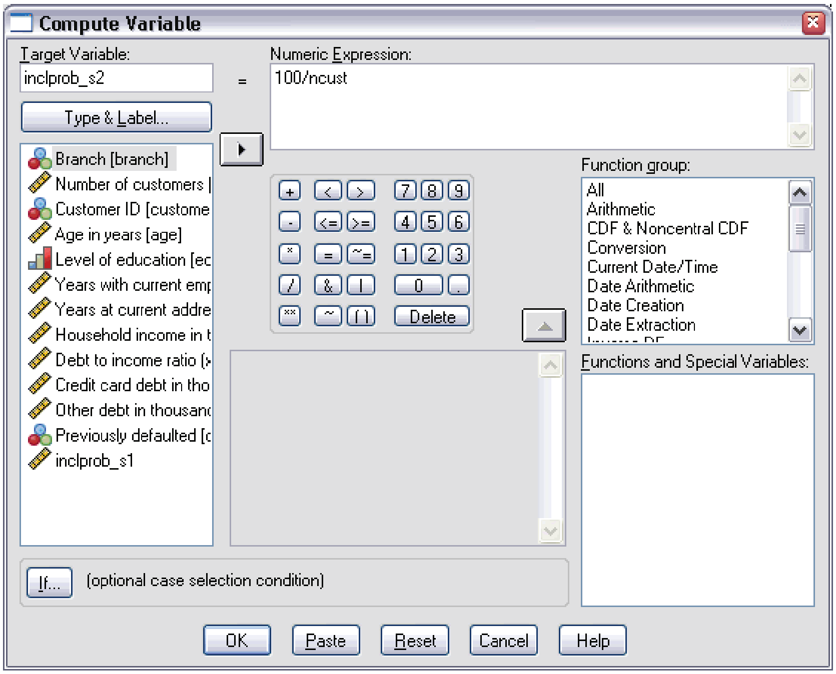
Fifteen out of one hundred bank branches were selected without replacement in the first stage; thus, the probability that a given bank was selected is 15/100 = 0.15.

▶ Type inclprob_s1 as the target variable.

▶ Type 0.15 as the numeric expression.

▶ Click OK.

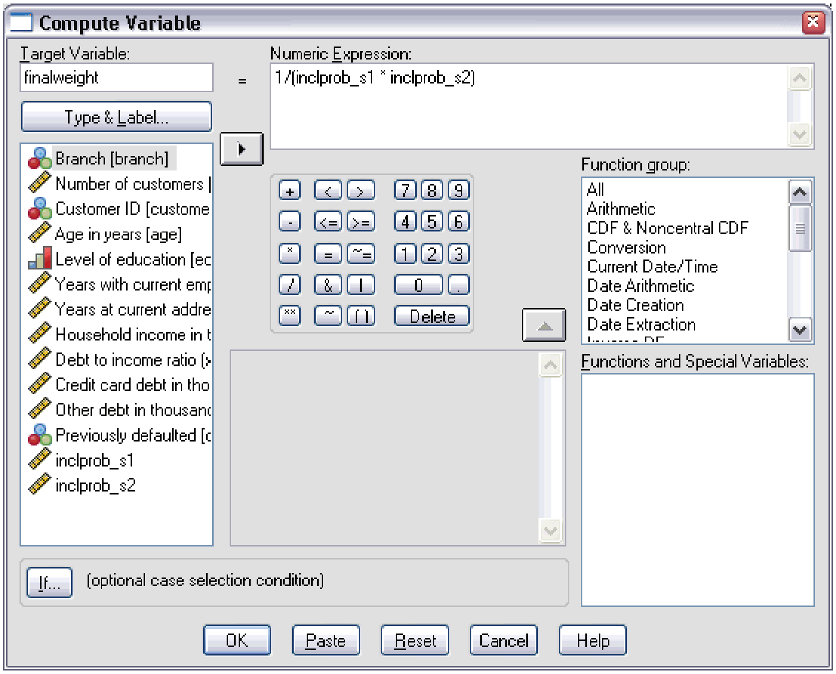
Figure 13-5
Compute Variable dialog box



One hundred customers were selected from each branch in the second stage; thus, the stage 2 inclusion probability for a given customer at a given bank is 100/the number of customers at that bank.

- Recall the Compute Variable dialog box.
- Type inclprob_s2 as the target variable.
- Type 100/ncust as the numeric expression.
- Click OK.

Figure 13-6
Compute Variable dialog box



Now that you have the inclusion probabilities for each stage, it's easy to compute the final sampling weights.

- ▶ Recall the Compute Variable dialog box.
- ▶ Type finalweight as the target variable.
- ▶ Type 1/(inclprob_s1 * inclprob_s2) as the numeric expression.
- ▶ Click OK.

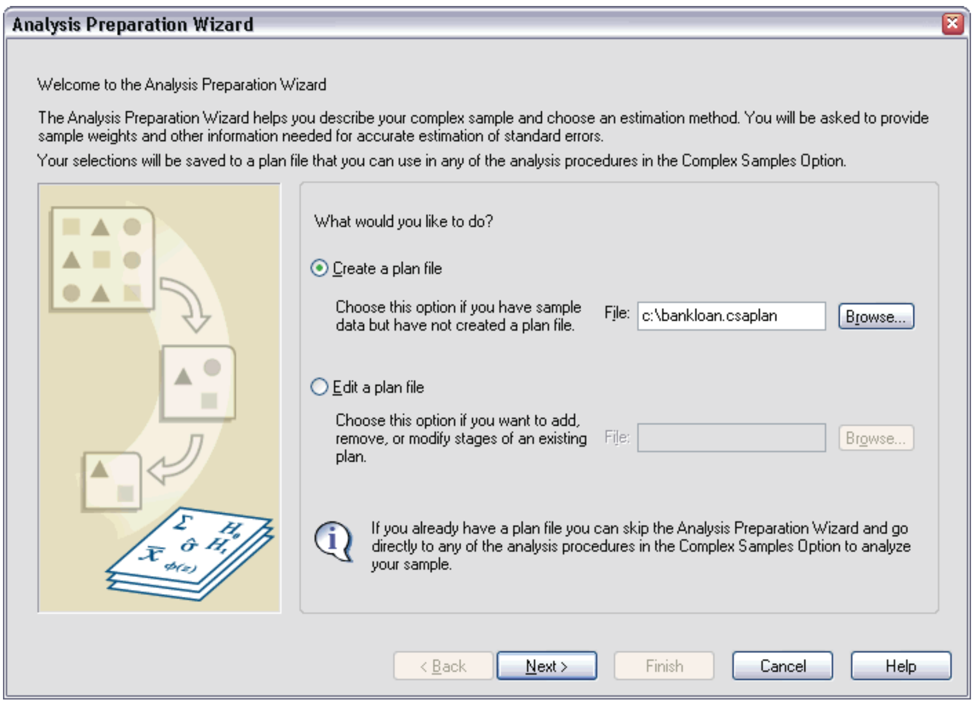
You are now ready to create the analysis plan.

Using the Wizard

▶ To prepare a sample using the Complex Samples Analysis Preparation Wizard, from the menus choose:

Analyze
  Complex Samples
    Prepare for Analysis...

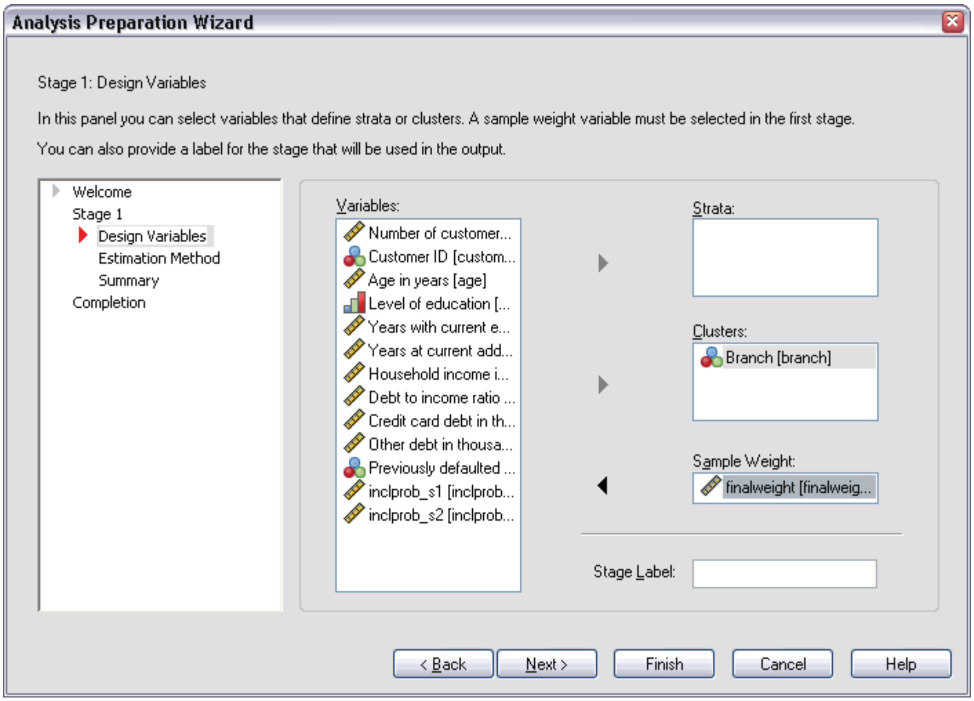
Figure 13-7
Analysis Preparation Wizard, Welcome step



▶ Enter c:\bankloan.csaplan as the name for the analysis plan file.

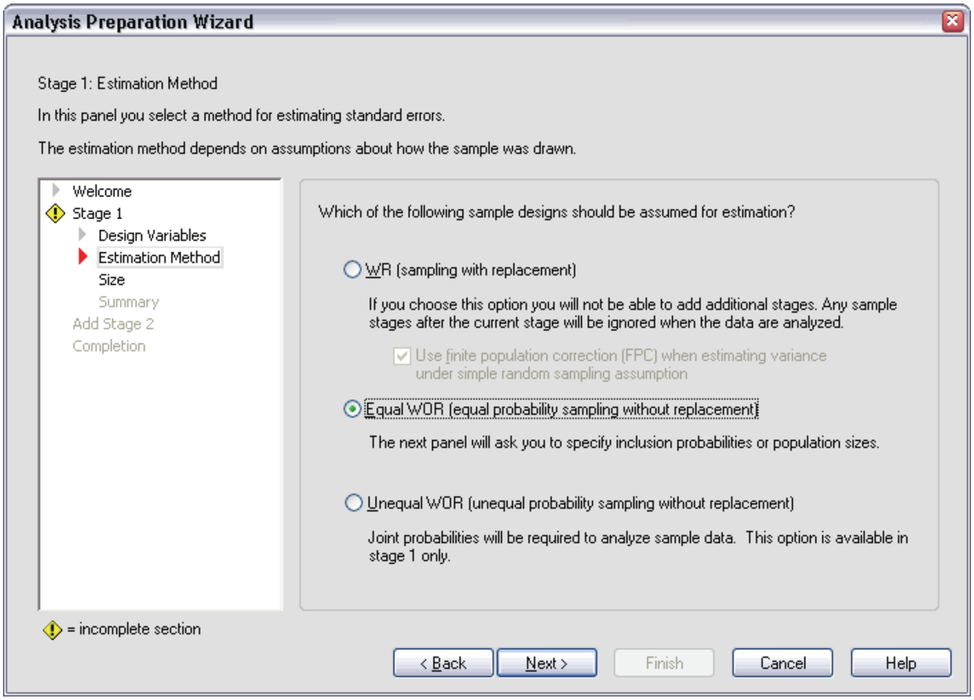
▶ Click Next.

Figure 13-8
Analysis Preparation Wizard, Design Variables step (stage 1)



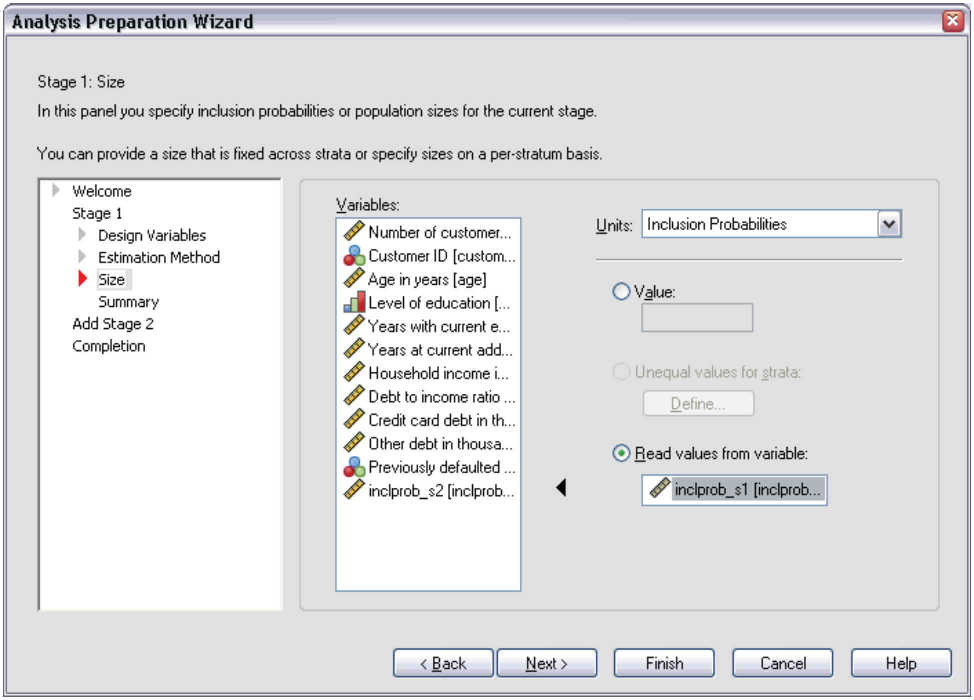
- Select *Branch* as a cluster variable.
- Select *finalweight* as the sample weight variable.
- Click Next.

Figure 13-9
Analysis Preparation Wizard, Estimation Method step (stage 1)



- ▶ Select Equal WOR as the first stage estimation method.
- ▶ Click Next.

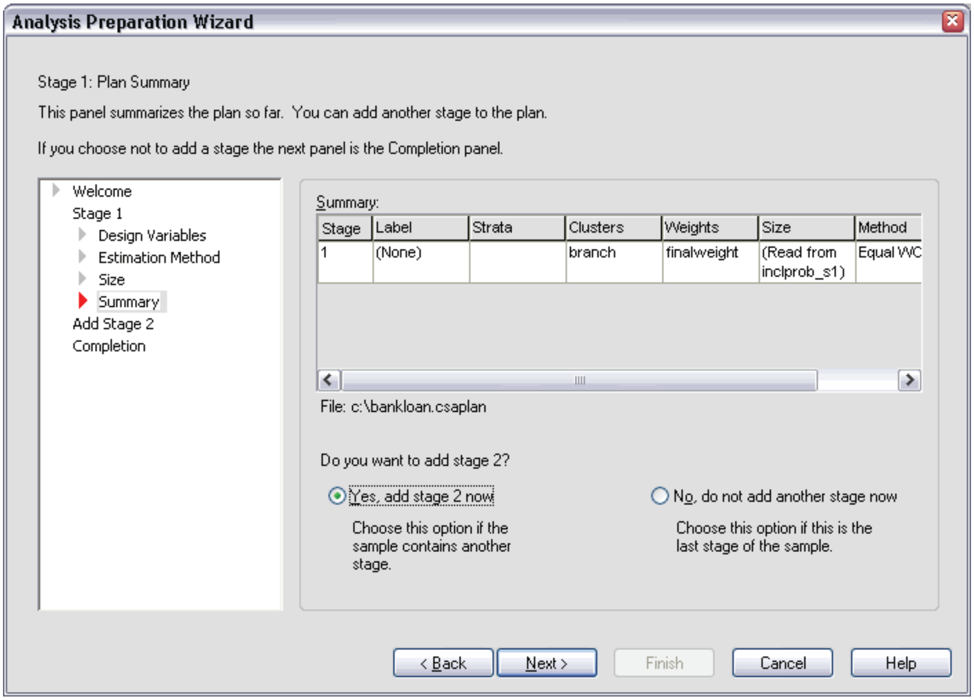
Figure 13-10
Analysis Preparation Wizard, Size step (stage 1)



▶ Select Read values from variable and select *inclprob_s1* as the variable containing the first stage inclusion probabilities.

▶ Click Next.

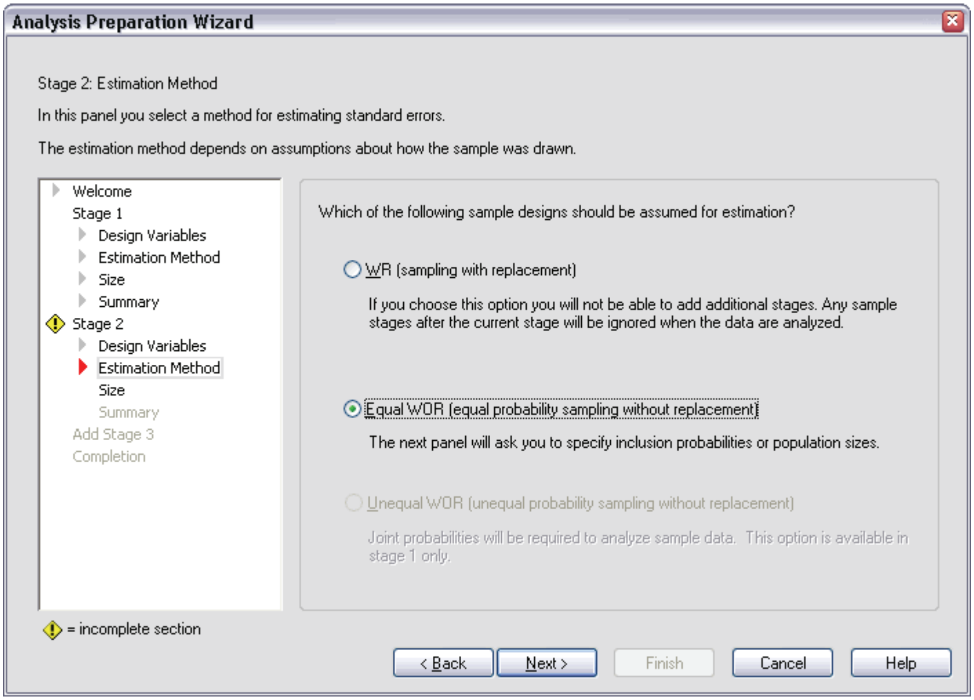
Figure 13-11
Analysis Preparation Wizard, Plan Summary step (stage 1)



▶ Select Yes, add stage 2 now.

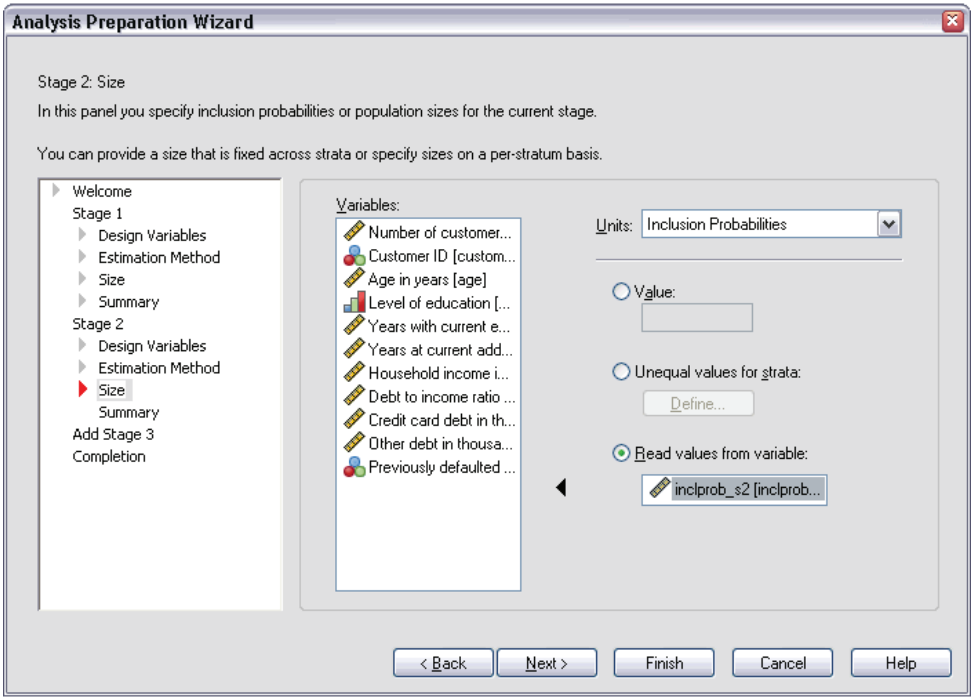
▶ Click Next, and then click Next in the Design step.

Figure 13-12
Analysis Preparation Wizard, Estimation Method step (stage 2)



- ▶ Select Equal WOR as the second stage estimation method.
- ▶ Click Next.

Figure 13-13
Analysis Preparation Wizard, Size step (stage 2)



▶ Select Read values from variable and select *inclprob_s2* as the variable containing the second stage inclusion probabilities.

▶ Click Finish.

Summary

Figure 13-14
Summary table

			Stage 1	Stage 2
Design Variables	Cluster	1	Branch	
Analysis Information	Estimator Assumption		Equal probability sampling without replacement	Equal probability sampling without replacement
	Inclusion Probability		Obtained from variable inclprob_s1	Obtained from variable inclprob_s2

Plan File: c:\bankloan.csaplan
Weight Variable: finalweight
SRS Estimator: Sampling without replacement

The summary table reviews your analysis plan. The plan consists of two stages with a design of one cluster variable. Equal probability without replacement (WOR) estimation is used, and the plan is saved to *c:/bankloan.csaplan*. You can now use this plan file to process *bankloan_noweights.sav* (with the inclusion probabilities and sampling weights you've computed) with Complex Samples analysis procedures.

Related Procedures

The Complex Samples Analysis Preparation Wizard procedure is a useful tool for readying a sample for analysis when you do not have access to the sampling plan file.

- To create a sampling plan file and draw a sample, use the Sampling Wizard.

Chapter

14

Complex Samples Frequencies

The Complex Samples Frequencies procedure produces frequency tables for selected variables and displays univariate statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Using Complex Samples Frequencies to Analyze Nutritional Supplement Usage

A researcher wants to study the use of nutritional supplements among U.S. citizens, using the results of the National Health Interview Survey (NHIS) and a previously created analysis plan. For more information, see Using the Complex Samples Analysis Preparation Wizard to Ready NHIS Public Data in Chapter 13 on p. 144.

A subset of the 2000 survey is collected in *nhis2000_subset.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. The analysis plan is stored in *nhis2000_subset.csaplan*. Use Complex Samples Frequencies to produce statistics for nutritional supplement usage.

Running the Analysis

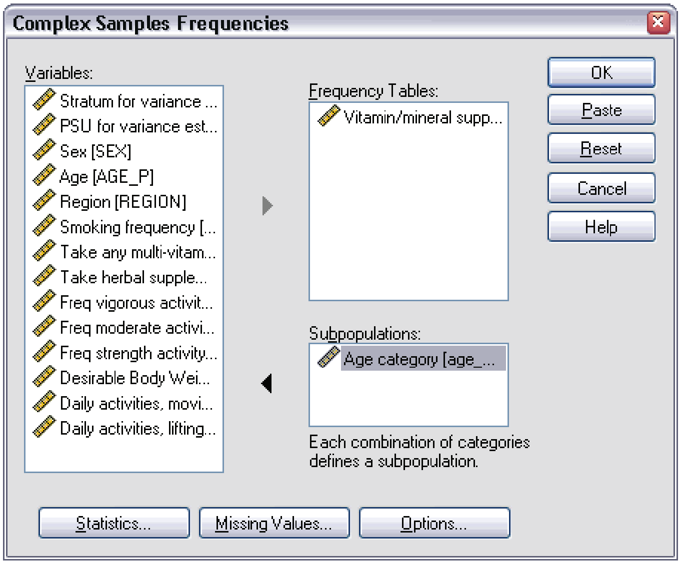
▶ To run a Complex Samples Frequencies analysis, from the menus choose:

Analyze
  Complex Samples
    Frequencies...

Figure 14-1
Complex Samples Plan dialog box

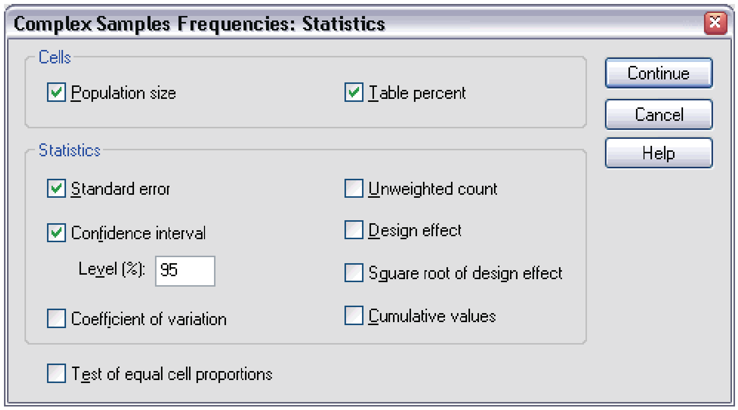
- Browse to the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS and select *nhis2000_subset.csaplan*.
- Click Continue.

Figure 14-2
Frequencies dialog box



- Select *Vitamin/mineral supplmnts-past 12 m* as a frequency variable.
- Select *Age category* as a subpopulation variable.
- Click Statistics.

Figure 14-3
Frequencies Statistics dialog box



- Select Table percent in the Cells group.
- Select Confidence interval in the Statistics group.
- Click Continue.
- Click OK in the Frequencies dialog box.

Frequency Table

Figure 14-4
Frequency table for variable/situation

		Estimate	Standard Error	95% Confidence Interval Lower	95% Confidence Interval Upper
Population Size	Yes	102767095	1185126.709	100435967	105098223
	No	90794234	1094401.949	88641560	92946908
	Total	193561329	1789098.713	190042196	197080462
% of Total	Yes	53.1%	.4%	52.4%	53.8%
	No	46.9%	.4%	46.2%	47.6%
	Total	100.0%	.0%	100.0%	100.0%

Each selected statistic is computed for each selected cell measure. The first column contains estimates of the number and percentage of the population that do or do not take vitamin/mineral supplements. The confidence intervals are nonoverlapping; thus, you can conclude that, overall, more Americans take vitamin/mineral supplements than not.

Frequency by Subpopulation

Figure 14-5
Frequency table by subpopulation

Age category			Estimate	Standard Error	95% Confidence Interval Lower	95% Confidence Interval Upper
18-24	Population Size	Yes	10018312	350602.352	9328681.9	10707942
		No	15472368	499182.391	14490483	16454253
		Total	25490680	680732.812	24151688	26829672
	% of Total	Yes	39.3%	1.0%	37.4%	41.2%
		No	60.7%	1.0%	58.8%	62.6%
		Total	100.0%	.0%	100.0%	100.0%
25-44	Population Size	Yes	39163840	660855.719	37863946	40463734
		No	39503150	645934.187	38232606	40773694
		Total	78666990	961114.325	76776491	80557489
	% of Total	Yes	49.8%	.6%	48.7%	50.9%
		No	50.2%	.6%	49.1%	51.3%
		Total	100.0%	.0%	100.0%	100.0%
45-64	Population Size	Yes	34154952	598603.728	32977507	35332397
		No	24005512	497723.833	23026496	24984528
		Total	58160464	814680.415	56557999	59762929
	% of Total	Yes	58.7%	.6%	57.5%	60.0%
		No	41.3%	.6%	40.0%	42.5%
		Total	100.0%	.0%	100.0%	100.0%
65+	Population Size	Yes	19429991	439459.793	18565580	20294402
		No	11813204	314238.078	11195102	12431306
		Total	31243195	587623.439	30087348	32399042
	% of Total	Yes	62.2%	.7%	60.7%	63.6%
		No	37.8%	.7%	36.4%	39.3%
		Total	100.0%	.0%	100.0%	100.0%

When computing statistics by subpopulation, each selected statistic is computed for each selected cell measure by value of *Age category*. The first column contains estimates of the number and percentage of the population of each category that do or do not take vitamin/mineral supplements. The confidence intervals for the table percentages are all nonoverlapping; thus, you can conclude that the use of vitamin/mineral supplements increases with age.

Summary

Using the Complex Samples Frequencies procedure, you have obtained statistics for the use of nutritional supplements among U.S. citizens.

- Overall, more Americans take vitamin/mineral supplements than not.
- When broken down by age category, greater proportions of Americans take vitamin/mineral supplements with increasing age.

Related Procedures

The Complex Samples Frequencies procedure is a useful tool for obtaining univariate descriptive statistics of categorical variables for observations obtained via a complex sampling design.

- The Complex Samples Sampling Wizard is used to specify complex sampling design specifications and obtain a sample. The sampling plan file created by the Sampling Wizard contains a default analysis plan and can be specified in the Plan dialog box when you are analyzing the sample obtained according to that plan.
- The Complex Samples Analysis Preparation Wizard is used to set analysis specifications for an existing complex sample. The analysis plan file created by the Sampling Wizard can be specified in the Plan dialog box when you are analyzing the sample corresponding to that plan.
- The Complex Samples Crosstabs procedure provides descriptive statistics for the crosstabulation of categorical variables.
- The Complex Samples Descriptives procedure provides univariate descriptive statistics for scale variables.

Chapter
15
Complex Samples Descriptives

The Complex Samples Descriptives procedure displays univariate summary statistics for several variables. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Using Complex Samples Descriptives to Analyze Activity Levels

A researcher wants to study the activity levels of U.S. citizens, using the results of the National Health Interview Survey (NHIS) and a previously created analysis plan. For more information, see Using the Complex Samples Analysis Preparation Wizard to Ready NHIS Public Data in Chapter 13 on p. 144.

A subset of the 2000 survey is collected in *nhis2000_subset.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. The analysis plan is stored in *nhis2000_subset.csaplan*. Use Complex Samples Descriptives to produce univariate descriptive statistics for activity levels.

Running the Analysis

▶ To run a Complex Samples Descriptives analysis, from the menus choose:

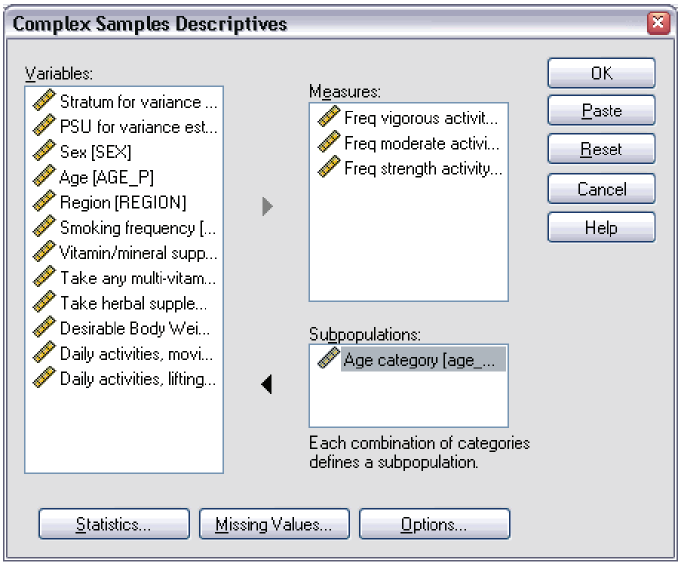
Analyze
  Complex Samples
    Descriptives...

Figure 15-1
Complex Samples Plan dialog box

▶ Browse to the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS, and select *nhis2000_subset.csaplan*.

▶ Click Continue.

Figure 15-2
Descriptives dialog box

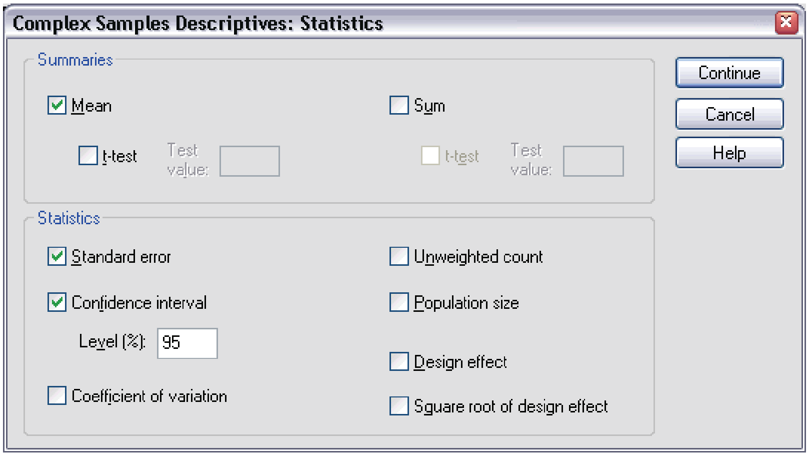


- Select *Freq vigorous activity (times per wk)* through *Freq strength activity (times per wk)* as measure variables.

- Select *Age category* as a subpopulation variable.

- Click Statistics.

Figure 15-3
Descriptives Statistics dialog box



- Select Confidence interval in the Statistics group.
- Click Continue.
- Click OK in the Complex Samples Descriptives dialog box.

Univariate Statistics

Figure 15-4
Univariate statistics

		Estimate	Standard Error	95% Confidence Interval	
				Lower	Upper
Mean	Freq vigorous activity (times per wk)	3.73	.033	3.66	3.79
	Freq moderate activity (times per wk)	4.90	.041	4.82	4.98
	Freq strength activity (times per wk)	3.52	.042	3.43	3.60

Each selected statistic is computed for each measure variable. The first column contains estimates of the average number of times per week that a person engages in a particular type of activity. The confidence intervals for the means are nonoverlapping. Thus, you can conclude that, overall, Americans engage in a strength activity less often than vigorous activity, and they engage in vigorous activity less often than moderate activity.

Univariate Statistics by Subpopulation

Figure 15-5
Univariate statistics by subpopulation

Age category			Estimate	Standard Error	95% Confidence Interval	
					Lower	Upper
18-24	Mean	Freq vigorous activity (times per wk)	3.92	.087	3.75	4.09
		Freq moderate activity (times per wk)	5.18	.137	4.91	5.45
		Freq strength activity (times per wk)	3.45	.085	3.28	3.62
25-44	Mean	Freq vigorous activity (times per wk)	3.55	.048	3.46	3.65
		Freq moderate activity (times per wk)	4.73	.056	4.62	4.84
		Freq strength activity (times per wk)	3.28	.052	3.18	3.38
45-64	Mean	Freq vigorous activity (times per wk)	3.79	.063	3.66	3.91
		Freq moderate activity (times per wk)	4.88	.070	4.74	5.02
		Freq strength activity (times per wk)	3.65	.092	3.47	3.84
65+	Mean	Freq vigorous activity (times per wk)	4.18	.111	3.96	4.39
		Freq moderate activity (times per wk)	5.22	.084	5.06	5.39
		Freq strength activity (times per wk)	4.66	.155	4.36	4.97

Each selected statistic is computed for each measure variable by values of *Age category*. The first column contains estimates of the average number of times per week that people of each category engage in a particular type of activity. The confidence intervals for the means allow you to make some interesting conclusions.

- In terms of vigorous and moderate activities, 25–44-year-olds are less active than those 18–24 and 45–64, and 45–64-year-olds are less active than those 65 or older.
- In terms of strength activity, 25–44-year-olds are less active than those 45–64, and 18–24 and 45–64-year-olds are less active than those 65 or older.

Summary

Using the Complex Samples Descriptives procedure, you have obtained statistics for the activity levels of U.S. citizens.

- Overall, Americans spend varying amounts of time at different types of activities.
- When broken down by age, it roughly appears that post-collegiate Americans are initially less active than they were while in school but become more conscientious about exercising as they age.

Related Procedures

The Complex Samples Descriptives procedure is a useful tool for obtaining univariate descriptive statistics of scale measures for observations obtained via a complex sampling design.

- The Complex Samples Sampling Wizard is used to specify complex sampling design specifications and obtain a sample. The sampling plan file created by the Sampling Wizard contains a default analysis plan and can be specified in the Plan dialog box when you are analyzing the sample obtained according to that plan.
- The Complex Samples Analysis Preparation Wizard is used to set analysis specifications for an existing complex sample. The analysis plan file created by the Sampling Wizard can be specified in the Plan dialog box when you are analyzing the sample corresponding to that plan.
- The Complex Samples Ratios procedure provides descriptive statistics for ratios of scale measures.
- The Complex Samples Frequencies procedure provides univariate descriptive statistics of categorical variables.

Chapter
16

Complex Samples Crosstabs

The Complex Samples Crosstabs procedure produces crosstabulation tables for pairs of selected variables and displays two-way statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Using Complex Samples Crosstabs to Measure the Relative Risk of an Event

A company that sells magazine subscriptions traditionally sends monthly mailings to a purchased database of names. The response rate is typically low, so you need to find a way to better target prospective customers. One suggestion is to focus mailings on people with newspaper subscriptions, on the assumption that people who read newspapers are more likely to subscribe to magazines.

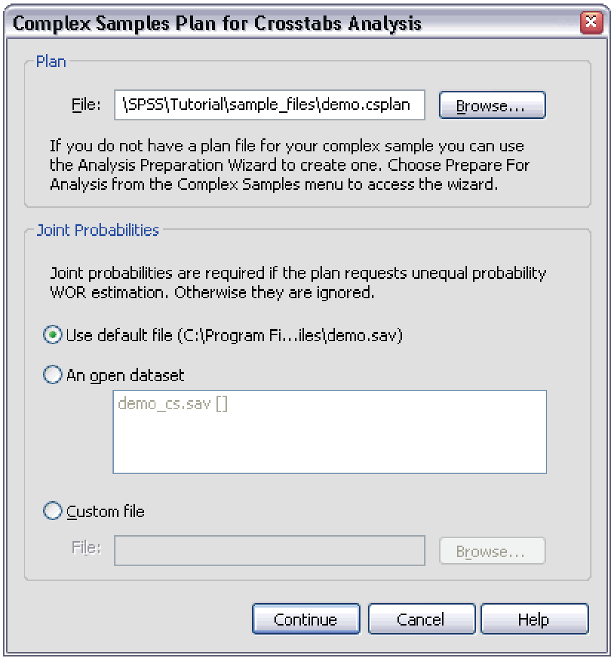
Use the Complex Samples Crosstabs procedure to test this theory by constructing a two-by-two table of *Newspaper subscription* by *Response* and computing the relative risk that a person with a newspaper subscription will respond to the mailing. This information is collected in *demo_cs.sav* and should be analyzed using the sampling plan file *demo.csplan*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS.

Running the Analysis

▶ To run a Complex Samples Crosstabs analysis, from the menus choose:

Analyze
  Complex Samples
    Crosstabs...

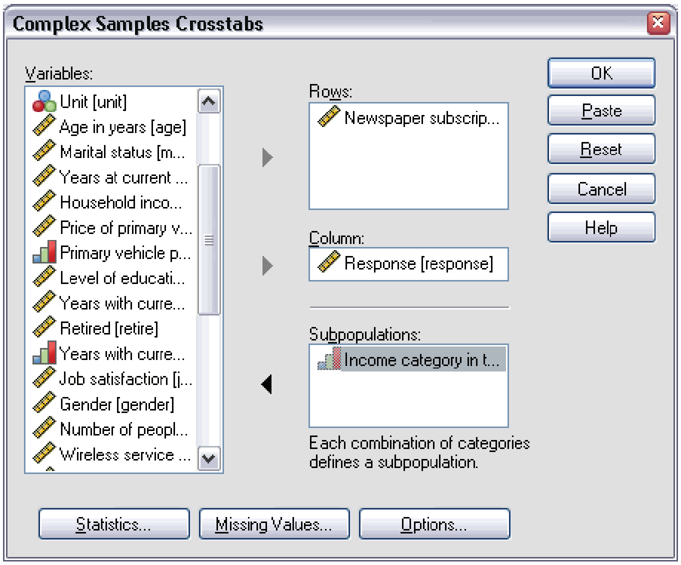
Figure 16-1
Complex Samples Plan dialog box



▶ Browse to the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS and select *demo.csplan*.

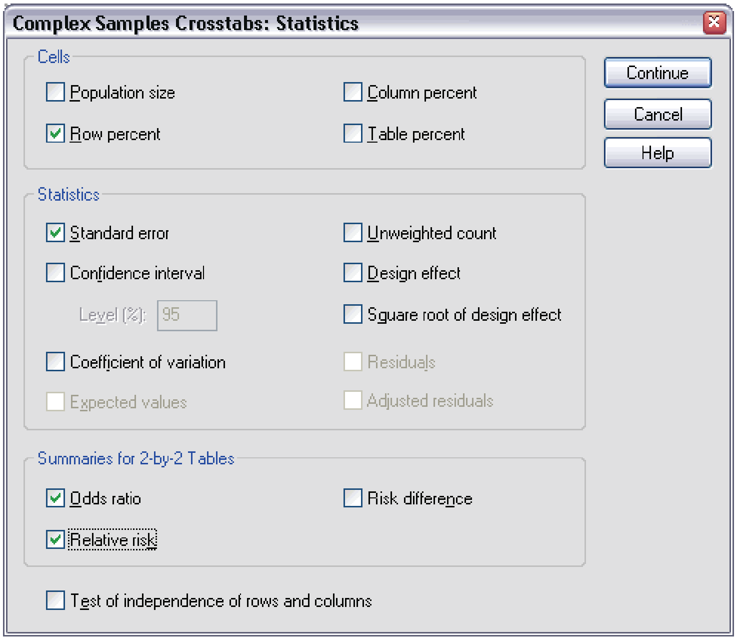
▶ Click Continue.

Figure 16-2
Crosstabs dialog box



- Select *Newspaper subscription* as a row variable.
- Select *Response* as a column variable.
- There is also some interest in seeing the results broken down by income categories, so select *Income category in thousands* as a subpopulation variable.
- Click Statistics.

Figure 16-3
Crosstabs Statistics dialog box



- ▶ Deselect Population size and select Row percent in the Cells group.
- ▶ Select Odds ratio and Relative risk in the Summaries for 2-by-2 Tables group.
- ▶ Click Continue.
- ▶ Click OK in the Complex Samples Crosstabs dialog box.

These selections produce a crosstabulation table and risk estimate for *Newspaper subscription* by *Response*. Separate tables with results split by *Income category in thousands* are also created.

Crosstabulation

Figure 16-4
Crosstabulation for newspaper subscription by response

Newspaper subscription			Response		
			Yes	No	Total
Yes	% within Newspaper subscription	Estimate	17.2%	82.8%	100.0%
		Standard Error	1.0%	1.0%	.0%
No	% within Newspaper subscription	Estimate	10.3%	89.7%	100.0%
		Standard Error	.7%	.7%	.0%
Total	% within Newspaper subscription	Estimate	12.8%	87.2%	100.0%
		Standard Error	.7%	.7%	.0%

The crosstabulation shows that, overall, few people responded to the mailing. However, a greater proportion of newspaper subscribers responded.

Risk Estimate

Figure 16-5
Risk estimate for newspaper subscription by response

			Estimate
Newspaper subscription * Response	Odds Ratio		1.812
	Relative Risk	For cohort Response = Yes	1.673
		For cohort Response = No	.923

Statistics are computed only for 2-by-2 tables with all cells observed.

The relative risk is a ratio of event probabilities. The relative risk of a response to the mailing is the ratio of the probability that a newspaper subscriber responds to the probability that a nonsubscriber responds. Thus, the estimate of the relative risk is simply 17.2%/10.3% = 1.673. Likewise, the relative risk of nonresponse is the ratio of the probability that a subscriber does not respond to the probability that a nonsubscriber does not respond. Your estimate of this relative risk is 0.923. Given these results, you can estimate that a newspaper subscriber is 1.673 times as likely to respond to the mailing as a nonsubscriber, or 0.923 times as likely as a nonsubscriber not to respond.

The odds ratio is a ratio of event odds. The odds of an event is the ratio of the probability that the event occurs to the probability that the event does not occur. Thus, the estimate of the odds that a newspaper subscriber responds to the mailing is 17.2%/82.8% = 0.208. Likewise, the estimate of the odds that a nonsubscriber responds is 10.3%/89.7% = 0.115. The estimate of the odds ratio is therefore 0.208/0.115 = 1.812 (note there is some rounding error in the intervening steps). The odds ratio is also the ratio of the relative risk of responding to the relative risk of not responding, or 1.673/0.923 = 1.812.

Odds Ratio versus Relative Risk

Since it is a ratio of ratios, the odds ratio is very difficult to interpret. The relative risk is easier to interpret, so the odds ratio alone is not very helpful. However, there are certain commonly occurring situations in which the estimate of the relative risk is not very good, and the odds ratio can be used to approximate the relative risk of the event of interest. The odds ratio should be used as an approximation of the relative risk of the event of interest when both of the following conditions are met:

- The probability of the event of interest is small (< 0.1). This condition guarantees that the odds ratio will make a good approximation to the relative risk. In this example, the event of interest is a response to the mailing.
- The design of the study is case control. This condition signals that the usual estimate of the relative risk will likely not be good. A case-control study is retrospective, most often used when the event of interest is unlikely or when the design of a prospective experiment is impractical or unethical.

Neither condition is met in this example, since the overall proportion of respondents was 12.8% and the design of the study was not case control, so it's safer to report 1.673 as the relative risk, rather than the value of the odds ratio.

Risk Estimate by Subpopulation

Figure 16-6
Risk estimate for newspaper subscription by response, controlling for income category

Income category				Estimate
Under $25	Newspaper subscription * Response	Odds Ratio		2.712
		Relative Risk	For cohort Response = Yes	2.241
			For cohort Response = No	.826
$25 - $49	Newspaper subscription * Response	Odds Ratio		1.794
		Relative Risk	For cohort Response = Yes	1.645
			For cohort Response = No	.917
$50 - $74	Newspaper subscription * Response	Odds Ratio		1.168
		Relative Risk	For cohort Response = Yes	1.152
			For cohort Response = No	.986
$75+	Newspaper subscription * Response	Odds Ratio		1.242
		Relative Risk	For cohort Response = Yes	1.227
			For cohort Response = No	.988

Statistics are computed only for 2-by-2 tables with all cells observed.

Relative risk estimates are computed separately for each income category. Note that the relative risk of a positive response for newspaper subscribers appears to gradually decrease with increasing income, which indicates that you may be able to further target the mailings.

Summary

Using Complex Samples Crosstabs risk estimates, you found that you can increase your response rate to direct mailings by targeting newspaper subscribers. Further, you found some evidence that the risk estimates may not be constant across *Income category*, so you may be able to increase your response rate even more by targeting lower-income newspaper subscribers.

Related Procedures

The Complex Samples Crosstabs procedure is a useful tool for obtaining descriptive statistics of the crosstabulation of categorical variables for observations obtained via a complex sampling design.

- The Complex Samples Sampling Wizard is used to specify complex sampling design specifications and obtain a sample. The sampling plan file created by the Sampling Wizard contains a default analysis plan and can be specified in the Plan dialog box when you are analyzing the sample obtained according to that plan.
- The Complex Samples Analysis Preparation Wizard is used to set analysis specifications for an existing complex sample. The analysis plan file created by the Sampling Wizard can be specified in the Plan dialog box when you are analyzing the sample corresponding to that plan.
- The Complex Samples Frequencies procedure provides univariate descriptive statistics of categorical variables.

Chapter

17

Complex Samples Ratios

The Complex Samples Ratios procedure displays univariate summary statistics for ratios of variables. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Using Complex Samples Ratios to Aid Property Value Assessment

A state agency is charged with ensuring that property taxes are fairly assessed from county to county. Taxes are based on the appraised value of the property, so the agency wants to track property values across counties to be sure that each county's records are equally up-to-date. Since resources for obtaining current appraisals are limited, the agency chose to employ complex sampling methodology to select properties.

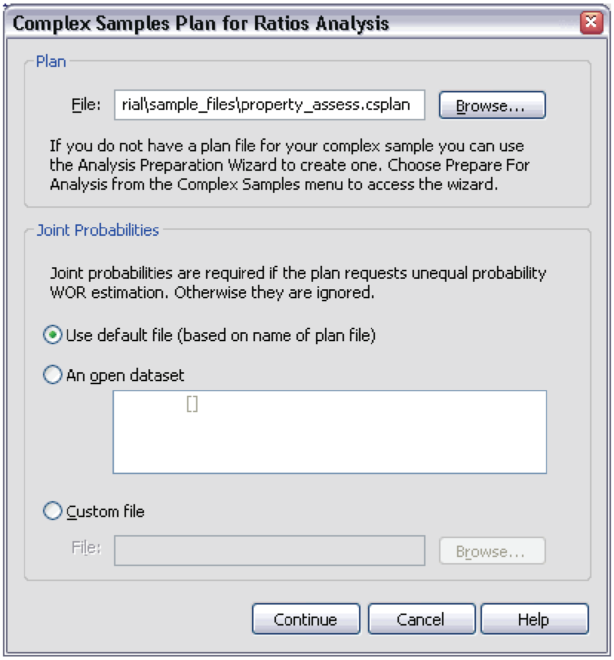
The sample of properties selected and their current appraisal information is collected in *property_assess_cs_sample.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. Use Complex Samples Ratios to assess the change in property values across the five counties since the last appraisal.

Running the Analysis

▶ To run a Complex Samples Ratios analysis, from the menus choose:

Analyze
  Complex Samples
    Ratios...

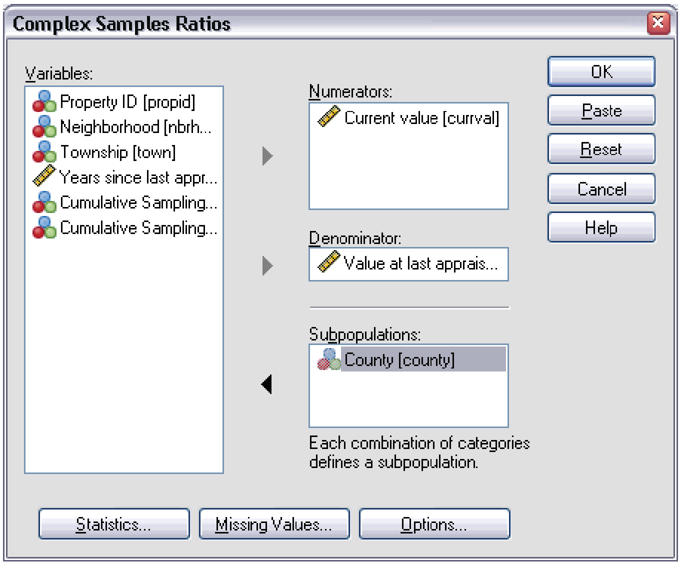
Figure 17-1
Complex Samples Plan dialog box



▶ Browse to the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS and select *property_assess.csplan*.

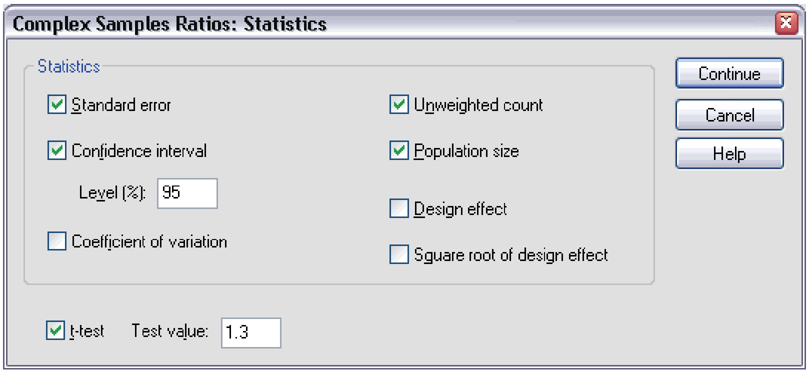
▶ Click Continue.

Figure 17-2
Complex Samples Ratios dialog box



- Select *Current value* as a numerator variable.
- Select *Value at last appraisal* as the denominator variable.
- Select *County* as a subpopulation variable.
- Click Statistics.

Figure 17-3
Ratios Statistics dialog box



- Select Confidence interval, Unweighted count, and Population size in the Statistics group.
- Select t-test and enter 1.3 as the test value.
- Click Continue.
- Click OK in the Complex Samples Ratios dialog box.

Ratios

Figure 17-4
Ratios table

County	Numerator	Denominator	Ratio Estimate	Standard Error	95% Confidence Interval Lower	95% Confidence Interval Upper	Te
Eastern	Current value	Value at last appraisal	1.381	.068	1.236	1.525	
Central	Current value	Value at last appraisal	1.364	.064	1.227	1.502	
Western	Current value	Value at last appraisal	1.524	.053	1.410	1.638	
Northern	Current value	Value at last appraisal	1.277	.032	1.208	1.346	
Southern	Current value	Value at last appraisal	1.195	.029	1.134	1.256	

The default display of the table is very wide, so you will need to pivot it for a better view.

Pivoting the Ratios Table

- Double-click the table to activate it.

▶ From the Viewer menus choose:

Pivot
  Pivoting Trays

▶ Drag *Numerator* and then *Denominator* from the row to the layer.

▶ Drag *County* from the row to the column.

▶ Drag *Statistics* from the column to the row.

▶ Close the pivoting trays window.

Pivoted Ratios Table

Figure 17-5
Pivoted ratios table

Numerator: Current value
Denominator: Value at last appraisal

		County				
		Eastern	Central	Western	Northern	Southern
Ratio Estimate		1.381	1.364	1.524	1.277	1.195
Standard Error		.068	.064	.053	.032	.029
95% Confidence Interval	Lower	1.236	1.227	1.410	1.208	1.134
	Upper	1.525	1.502	1.638	1.346	1.256
Hypothesis Test	Test Value	1.3	1.3	1.3	1.3	1.3
	t	1.191	.997	4.201	-.702	-3.646
	df	15	15	15	15	15
	Sig.	.252	.334	.001	.493	.002
Unweighted Count		168	179	202	205	220

The ratios table is now pivoted so that statistics are easier to compare across counties.

- The ratio estimates range from a low of 1.195 in the Southern county to a high of 1.524 in the Western county.
- There is also quite a bit of variability in the standard errors, which range from a low of 0.029 in the Southern county to 0.068 in the Eastern county.
- Some of the confidence intervals do not overlap; thus, you can conclude that the ratios for the Western county are higher than the ratios for the Northern and Southern counties.

- Finally, as a more objective measure, note that the significance values of the *t* tests for the Western and Southern counties are less than 0.05. Thus, you can conclude that the ratio for the Western county is greater than 1.3 and the ratio for the Southern county is less than 1.3.

Summary

Using the Complex Samples Ratios procedure, you have obtained various statistics for the ratios of *Current value* to *Value at last appraisal*. The results suggest that there may be certain inequities in the assessment of property taxes from county to county, namely:

- The ratios for the Western county are high, indicating that their records are not as up-to-date as other counties with respect to the appreciation of property values. Property taxes are probably too low in this county.
- The ratios for the Southern county are low, indicating that their records are more up-to-date than the other counties with respect to the appreciation of property values. Property taxes are probably too high in this county.
- The ratios for the Southern county are lower than those of the Western county but are still within the objective goal of 1.3.

Resources used to track property values in the Southern county will be reassigned to the Western county to bring these counties' ratios in line with the others and with the goal of 1.3.

Related Procedures

The Complex Samples Ratios procedure is a useful tool for obtaining univariate descriptive statistics of the ratio of scale measures for observations obtained via a complex sampling design.

- The Complex Samples Sampling Wizard is used to specify complex sampling design specifications and obtain a sample. The sampling plan file created by the Sampling Wizard contains a default analysis plan and can be specified in the Plan dialog box when you are analyzing the sample obtained according to that plan.

- The Complex Samples Analysis Preparation Wizard is used to set analysis specifications for an existing complex sample. The analysis plan file created by the Sampling Wizard can be specified in the Plan dialog box when you are analyzing the sample corresponding to that plan.
- The Complex Samples Descriptives procedure provides descriptive statistics for scale variables.

Chapter

18

Complex Samples General Linear Model

The Complex Samples General Linear Model (CSGLM) procedure performs linear regression analysis, as well as analysis of variance and covariance, for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Using Complex Samples General Linear Model to Fit a Two-Factor ANOVA

A grocery store chain surveyed a set of customers concerning their purchasing habits, according to a complex design. Given the survey results and how much each customer spent in the previous month, the store wants to see if the frequency with which customers shop is related to the amount they spend in a month, controlling for the gender of the customer and incorporating the sampling design.

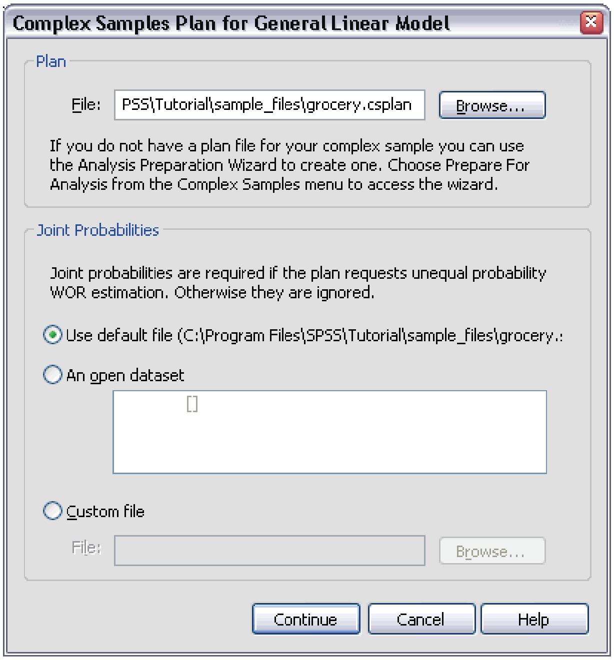
This information is collected in *grocery_1month_sample.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. Use the Complex Samples General Linear Model procedure to perform a two-factor (or two-way) ANOVA on the amounts spent.

Running the Analysis

▶ To run a Complex Samples General Linear Model analysis, from the menus choose:

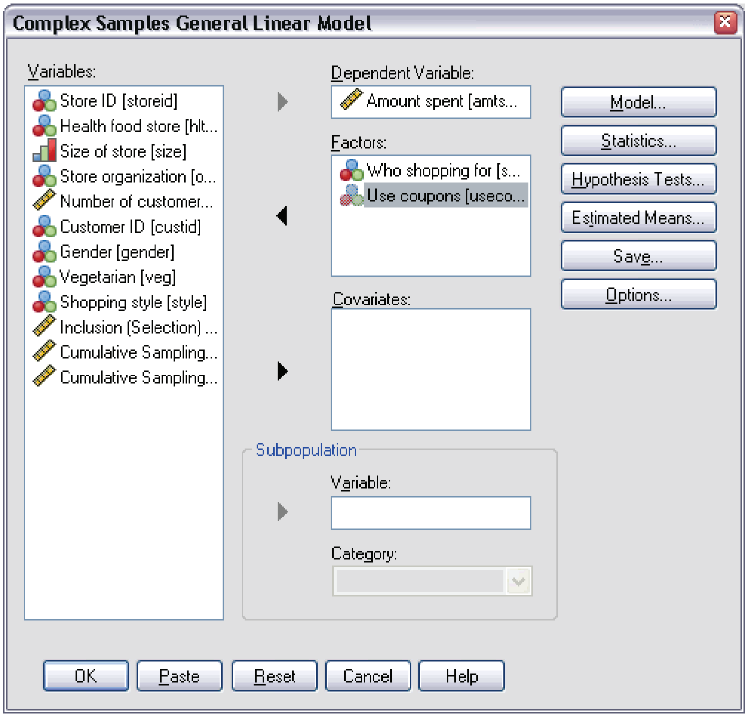
Analyze
  Complex Samples
    General Linear Model...

Figure 18-1
Complex Samples Plan dialog box



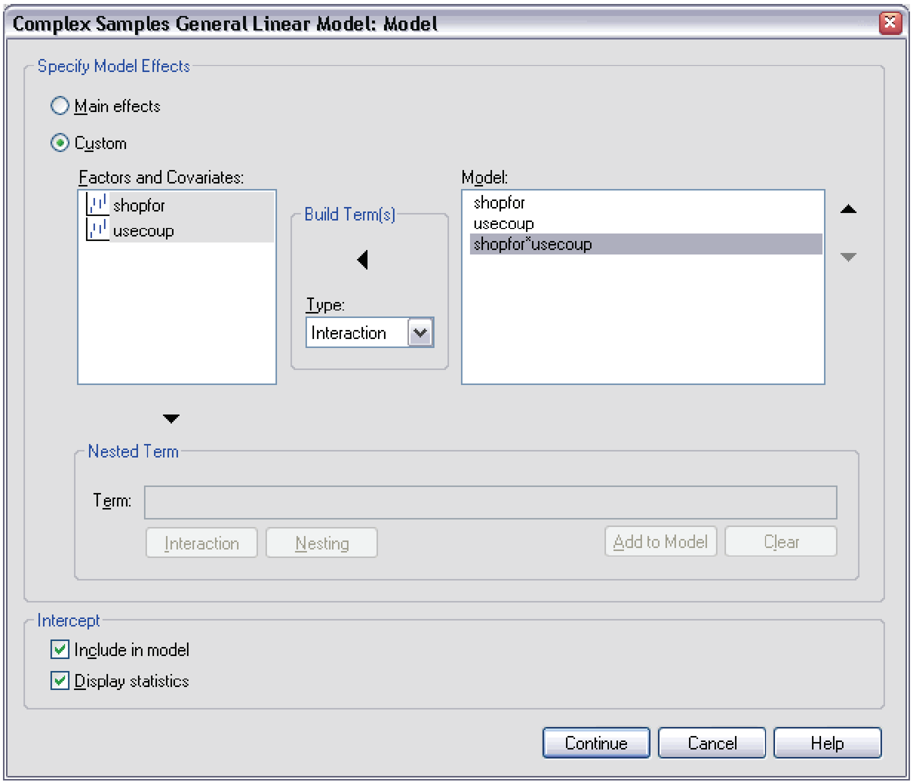
- Browse to the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS and select *grocery.csplan*.
- Click Continue.

Figure 18-2
General Linear Model dialog box



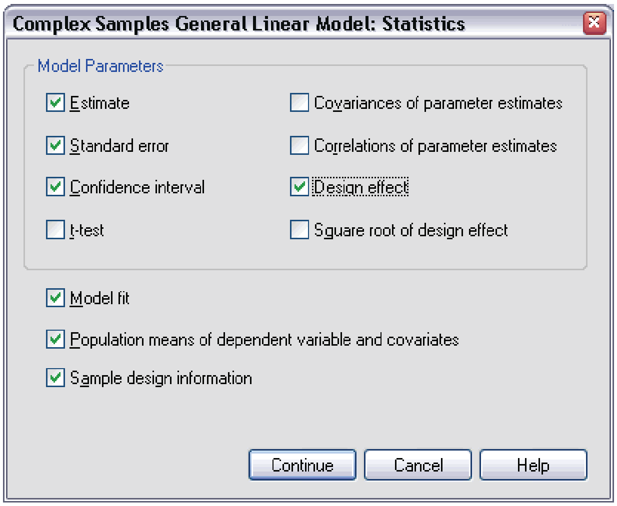
- ▶ Select *Amount spent* as the dependent variable.
- ▶ Select *Who shopping for* and *Use coupons* as factors.
- ▶ Click Model.

Figure 18-3
Model dialog box



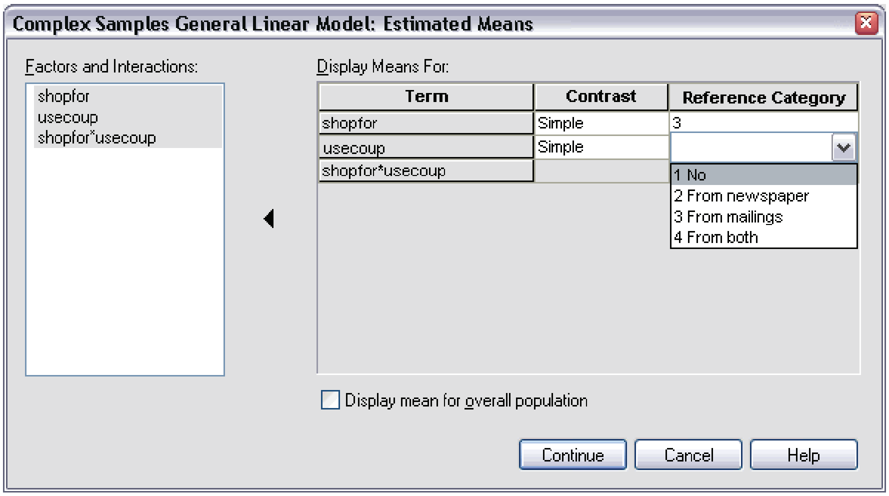
- Choose to build a Custom model.
- Select Main effects as the type of term to build and select *shopfor* and *usecoup* as model terms.
- Select Interaction as the type of term to build and add the *shopfor**usecoup* interaction as a model term.
- Click Continue.
- Click Statistics in the General Linear Model dialog box.

Figure 18-4
General Linear Model Statistics dialog box



- Select Estimate, Standard error, Confidence interval, and Design effect in the Model Parameters group.
- Click Continue.
- Click Estimated Means in the General Linear Model dialog box.

Figure 18-5
General Linear Model Estimated Means dialog box



- Choose to display means for *shopfor*, *usecoup*, and the *shopfor*usecoup* interaction.
- Select a Simple contrast and 3 Self and family as the reference category for *shopfor*. Note that, once selected, the category appears as "3" in the dialog box.
- Select a Simple contrast and 1 No as the reference category for *usecoup*.
- Click Continue.
- Click OK in the General Linear Model dialog box.

Model Summary

Figure 18-6
R-square statistic

R Square	.601

a. Model: Amount spent = (Intercept) + shopfor + usecoup + shopfor * usecoup

R Square, the coefficient of determination, is a measure of the strength of the model fit. It shows that about 60% of the variation in *Amount spent* is explained by the model, which gives you good explanatory ability. You may still want to add other predictors to the model to further improve the fit.

Tests of Model Effects

Figure 18-7
Tests of between-subjects effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	11.000	3.000	127.231	.001
(Intercept)	1.000	13.000	6321.597	.000
shopfor	2.000	12.000	643.593	.000
usecoup	3.000	11.000	87.453	.000
shopfor * usecoup	6.000	8.000	10.688	.002

a. Model: Amount spent = (Intercept) + shopfor + usecoup + shopfor * usecoup

Each term in the model, plus the model as a whole, is tested for whether the value of its effect equals 0. Terms with significance values of less than 0.05 have some discernible effect. Thus, all model terms contribute to the model.

Parameter Estimates

Figure 18-8
Parameter estimates

Parameter	Estimate	Std. Error	95% Confidence Interval		Design Effect
			Lower	Upper	
(Intercept)	518.249	11.731	492.905	543.592	1.387
[shopfor=1]	-174.757	10.762	-198.0	-151.51	.950
[shopfor=2]	-129.443	11.455	-154.2	-104.70	.925
[shopfor=3]	.000[a]	.	.	.	.
[usecoup=1]	-140.838	10.180	-162.8	-118.85	.649
[usecoup=2]	-63.026	13.195	-91.531	-34.520	.940
[usecoup=3]	-31.375	9.726	-52.387	-10.363	.564
[usecoup=4]	.000[a]	.	.	.	.
[shopfor=1] * [usecoup=1]	41.693	11.170	17.562	65.824	.606
[shopfor=1] * [usecoup=2]	44.505	18.068	5.471	83.539	1.413
[shopfor=1] * [usecoup=3]	9.204	11.057	-14.684	33.092	.594
[shopfor=1] * [usecoup=4]	.000[a]	.	.	.	.
[shopfor=2] * [usecoup=1]	89.211	10.967	65.518	112.903	.533
[shopfor=2] * [usecoup=2]	54.267	14.949	21.972	86.562	.836
[shopfor=2] * [usecoup=3]	17.884	13.753	-11.828	47.595	.797
[shopfor=2] * [usecoup=4]	.000[a]	.	.	.	.
[shopfor=3] * [usecoup=1]	.000[a]	.	.	.	.
[shopfor=3] * [usecoup=2]	.000[a]	.	.	.	.
[shopfor=3] * [usecoup=3]	.000[a]	.	.	.	.
[shopfor=3] * [usecoup=4]	.000[a]	.	.	.	.

a. Set to zero because this parameter is redundant.

b. Model: Amount spent = (Intercept) + shopfor + usecoup + shopfor * usecoup

The parameter estimates show the effect of each predictor on *Amount spent*. The value of 518.249 for the intercept term indicates that the grocery chain can expect a shopper with a family who uses coupons from the newspaper and targeted mailings to spend $518.25, on average. You can tell that the intercept is associated with these factor levels because those are the factor levels whose parameters are redundant.

- The *shopfor* coefficients suggest that among customers who use both mailed coupons and newspaper coupons, those without family tend to spend less than those with spouses, who in turn spend less than those with dependents at home. Since the tests of model effects showed that this term contributes to the model, these differences are not due to chance.
- The *usecoup* coefficients suggest that spending among customers with dependents at home decreases with decreased coupon usage. There is a moderate amount of uncertainty in the estimates, but the confidence intervals do not include 0.

- The interaction coefficients suggest that customers who do not use coupons or only clip from the newspaper and do not have dependents tend to spend more than you would otherwise expect. If any portion of an interaction parameter is redundant, the interaction parameter is redundant.
- The deviation in the values of the design effects from 1 indicate that some of the standard errors computed for these parameter estimates are larger than those you would obtain if you assumed that these observations came from a simple random sample, while others are smaller. It is vitally important to incorporate the sampling design information in your analysis because you might otherwise infer, for example, that the *usecoup=3* coefficient is not different from 0!

The parameter estimates are useful for quantifying the effect of each model term, but the estimated marginal means tables can make it easier to interpret the model results.

Estimated Marginal Means

Figure 18-9
Estimated marginal means by levels of Who shopping for

Who shopping for	Mean	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper
Self	308.5326	3.94286	300.0145	317.0506
Self and spouse	370.3361	4.87908	359.7955	380.8767
Self and family	459.4392	7.19769	443.8895	474.9888

This table displays the model-estimated marginal means and standard errors of *Amount spent* at the factor levels of *Who shopping for*. This table is useful for exploring the differences between the levels of this factor. In this example, a customer who shops for him- or herself is expected to spend about $308.53, while a customer with a spouse is expected to spend $370.33, and a customer with dependents will spend $459.44. To see whether this represents a real difference or is due to chance variation, look at the test results.

Figure 18-10
Individual test results for estimated marginal means of gender

Who shopping for Simple Contrast[a]	Contrast Estimate	Hypothesized Value	Difference (Estimate - Hypothesized)	Std. Error	df1	df2	Wald F	Sig.
Level Self vs. Level Self and family	-150.907	.000	-150.907	4.903	1.000	13.00	947.41	.000
Level Self and spouse vs. Level Self and family	-89.103	.000	-89.103	5.903	1.000	13.00	227.84	.000

a. Reference Category = Self and family

The individual tests table displays two simple contrasts in spending.

- The contrast estimate is the difference in spending for the listed levels of *Who shopping for*.
- The hypothesized value of 0.00 represents the belief that there is no difference in spending.
- The Wald *F* statistic, with the displayed degrees of freedom, is used to test whether the difference between a contrast estimate and hypothesized value is due to chance variation.
- Since the significance values are less than 0.05, you can conclude that there are differences in spending.

The values of the contrast estimates are different from the parameter estimates. This is because there is an interaction term containing the *Who shopping for* effect. As a result, the parameter estimate for *shopfor=1* is a simple contrast between the levels *Self* and *Self and Family* at the level *From both* of the variable *Use coupons*. The contrast estimate in this table is averaged over the levels of *Use coupons*.

Figure 18-11
Overall test results for estimated marginal means of gender

df1	df2	Wald F	Sig.
2.000	12.000	643.593	.000

The overall test table reports the results of a test of all of the contrasts in the individual test table. Its significance value of less than 0.05 confirms that there is a difference in spending among the levels of *Who shopping for*.

Figure 18-12
Estimated marginal means by levels of shopping style

Use coupons	Mean	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper
No	319.6455	6.51429	305.5722	333.7188
From newspaper	386.7469	4.32295	377.4077	396.0861
From mailings	394.5028	5.54218	382.5297	406.4760
From both	416.8486	6.51260	402.7790	430.9182

This table displays the model-estimated marginal means and standard errors of *Amount spent* at the factor levels of *Use coupons*. This table is useful for exploring the differences between the levels of this factor. In this example, a customer who does not use coupons is expected to spend about $319.65, and those who do use coupons are expected to spend considerably more.

Figure 18-13
Individual test results for estimated marginal means of shopping style

Use coupons Simple Contrast[a]	Contrast Estimate	Hypothesized Value	Difference (Estimate - Hypothesized)	Std. Error	df1	df2	Wald F	Sig.
Level From newspaper vs. Level No	67.101	.000	67.101	6.537	1.000	13.000	105.35	.000
Level From mailings vs. Level No	74.857	.000	74.857	5.875	1.000	13.000	162.33	.000
Level From both vs. Level No	97.203	.000	97.203	5.603	1.000	13.000	300.92	.000

a. Reference Category = No

The individual tests table displays three simple contrasts, comparing the spending of customers who do not use coupons to those who do.

Since the significance values of the tests are less than 0.05, you can conclude that customers who use coupons tend to spend more than those who don’t.

Figure 18-14
Overall test results for estimated marginal means of shopping style

df1	df2	Wald F	Sig.
3.000	11.000	87.453	.000

The overall test table reports the results of a test of all the contrasts in the individual test table. Its significance value of less than 0.05 confirms that there is a difference in spending among the levels of *Use coupons*. Note that the overall tests for *Use* *coupons* and *Who shopping for* are equivalent to the tests of model effects because the hypothesized contrast values are equal to 0.

Figure 18-15
Estimated marginal means by levels of gender by shopping style

Who shopping for	Use coupons	Mean	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper
Self	No	244.3471	6.00949	231.3644	257.3298
	From newspaper	324.9708	5.94134	312.1353	337.8063
	From mailings	321.3207	4.11028	312.4410	330.2005
	From both	343.4916	6.57845	329.2797	357.7034
Self and spouse	No	337.1783	7.12181	321.7925	352.5640
	From newspaper	380.0468	7.91038	362.9574	397.1361
	From mailings	375.3141	6.22468	361.8665	388.7617
	From both	388.8054	7.12101	373.4214	404.1894
Self and family	No	377.4111	11.58215	352.3894	402.4328
	From newspaper	455.2232	6.14420	441.9494	468.4969
	From mailings	486.8736	10.76529	463.6166	510.1306
	From both	518.2488	11.73120	492.9050	543.5925

This table displays the model-estimated marginal means, standard errors, and confidence intervals of *Amount spent* at the factor combinations of *Who shopping for* and *Use coupons*. This table is useful for exploring the interaction effect between these two factors that was found in the tests of model effects.

Summary

In this example, the estimated marginal means revealed differences in spending between customers at varying levels of *Who shopping for* and *Use coupons*. The tests of model effects confirmed this, as well as the fact that there appears to be a *Who shopping for**Use coupons* interaction effect. The model summary table revealed that the present model explains somewhat more than half of the variation in the data and could likely be improved by adding more predictors.

Related Procedures

The Complex Samples General Linear Model procedure is a useful tool for modeling a scale variable when the cases have been drawn according to a complex sampling scheme.

- The Complex Samples Sampling Wizard is used to specify complex sampling design specifications and obtain a sample. The sampling plan file created by the Sampling Wizard contains a default analysis plan and can be specified in the Plan dialog box when you are analyzing the sample obtained according to that plan.
- The Complex Samples Analysis Preparation Wizard is used to specify analysis specifications for an existing complex sample. The analysis plan file created by the Sampling Wizard can be specified in the Plan dialog box when you are analyzing the sample corresponding to that plan.
- The Complex Samples Logistic Regression procedure allows you to model a categorical response.
- The Complex Samples Ordinal Regression procedure allows you to model an ordinal response.

Chapter

19

Complex Samples Logistic Regression

The Complex Samples Logistic Regression procedure performs logistic regression analysis on a binary or multinomial dependent variable for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Using Complex Samples Logistic Regression to Assess Credit Risk

If you are a loan officer at a bank, you want to be able to identify characteristics that are indicative of people who are likely to default on loans and then use those characteristics to identify good and bad credit risks.

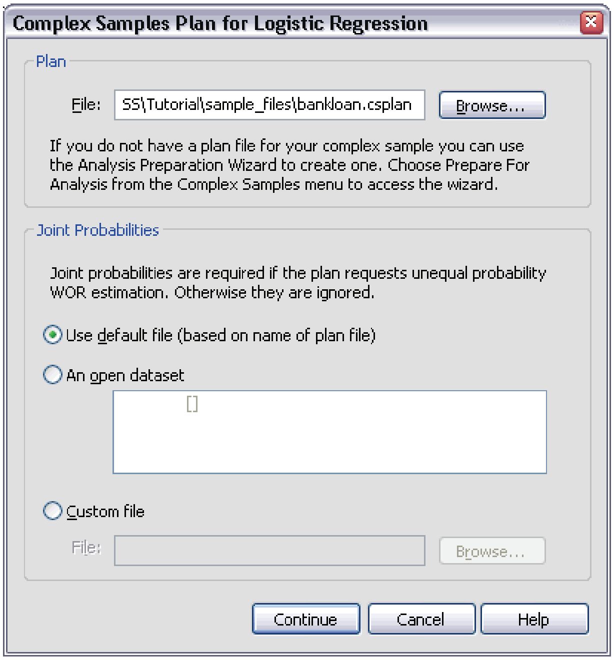
Suppose that a loan officer has collected past records of customers given loans at several different branches, according to a complex design. This information is contained in *bankloan_cs.sav*, found in the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS. The officer wants to see if the probability with which a customer defaults is related to age, employment history, and amount of credit debt, incorporating the sampling design.

Running the Analysis

▶ To create the logistic regression model, from the menus choose:

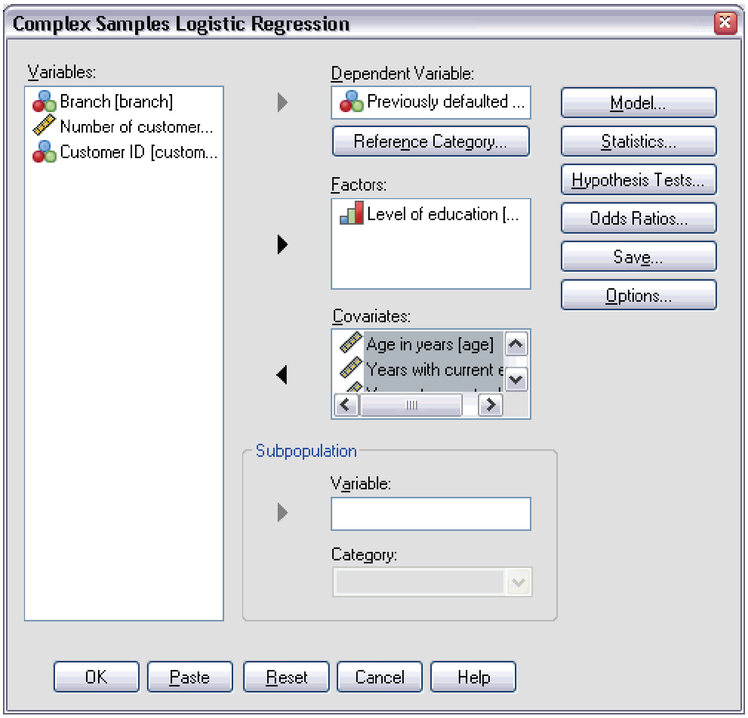
Analyze
  Complex Samples
    Logistic Regression...

Figure 19-1
Complex Samples Plan dialog box



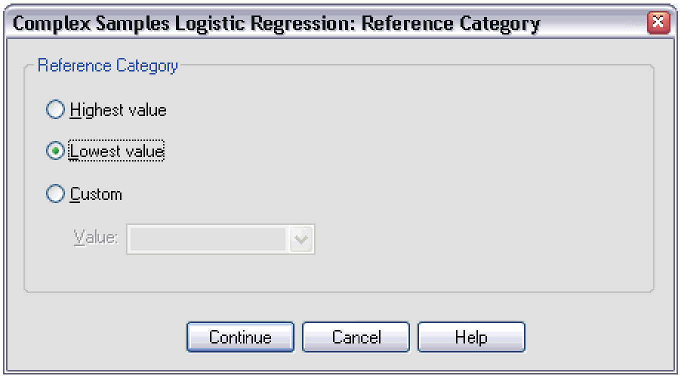
- Browse to the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS and select *bankloan.csaplan*.
- Click Continue.

Figure 19-2
Logistic Regression dialog box



- ▶ Select *Previously defaulted* as the dependent variable.
- ▶ Select *Level of education* as a factor.
- ▶ Select *Age in years* through *Other debt in thousands* as covariates.
- ▶ Select *Previously defaulted* and click Reference Category.

Figure 19-3
Logistic Regression Reference Category dialog box



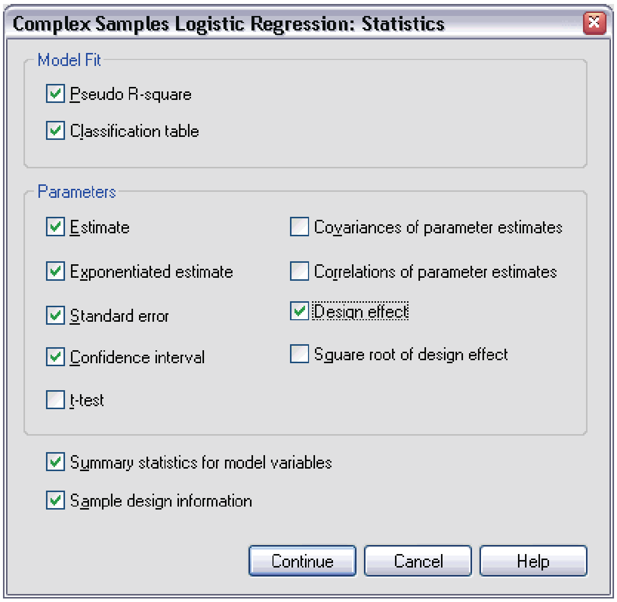
▶ Select Lowest value as the reference category.

This sets the "did not default" category as the reference category; thus, the odds ratios reported in the output will have the property that increasing odds ratios correspond to increasing probability of default.

▶ Click Continue.

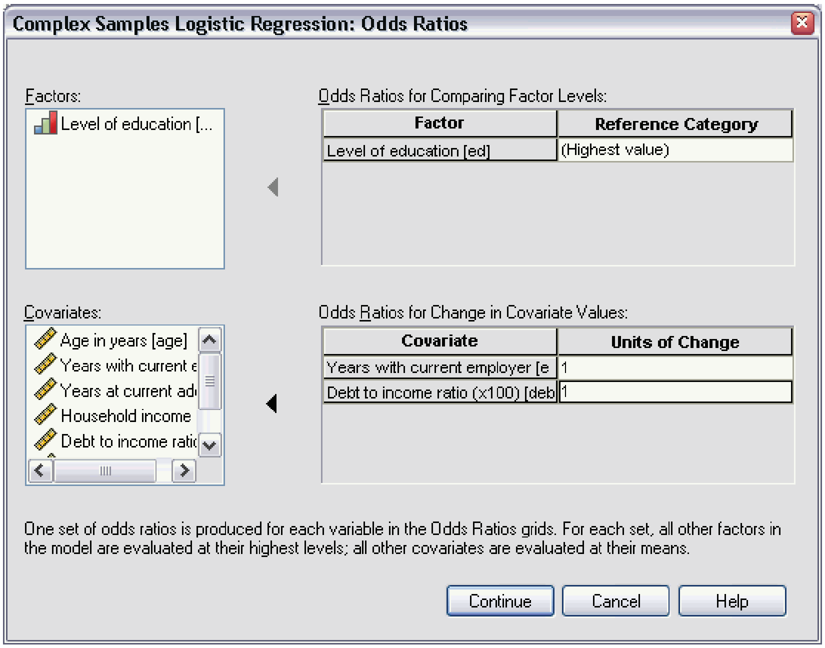
▶ Click Statistics in the Logistic Regression dialog box.

Figure 19-4
Logistic Regression Statistics dialog box



- Select Classification table in the Model Fit group
- Select Estimate, Exponentiated estimate, Standard error, Confidence interval, and Design effect in the Parameters group.
- Click Continue.
- Click Odds Ratios in the Logistic Regression dialog box.

Figure 19-5
Logistic Regression Odds Ratios dialog box



- ▶ Choose to create odds ratios for the factor *ed* and the covariates *employ* and *debtinc*.
- ▶ Click Continue.
- ▶ Click OK in the Logistic Regression dialog box.

Pseudo R-Squares

Figure 19-6
Pseudo R-square statistics

Cox and Snell	.330
Nagelkerke	.451
McFadden	.304

Dependent Variable: Previously defaulted (reference category = No)
Model: (Intercept), ed, age, employ, address, income, debtinc, creddebt, othdebt

In the linear regression model, the coefficient of determination, R^2, summarizes the proportion of variance in the dependent variable associated with the predictor (independent) variables, with larger R^2 values indicating that more of the variation is explained by the model, to a maximum of 1. For regression models with a categorical dependent variable, it is not possible to compute a single R^2 statistic that has all of the characteristics of R^2 in the linear regression model, so these approximations are computed instead. The following methods are used to estimate the coefficient of determination.

- Cox and Snell's R^2(Cox and Snell, 1989) is based on the log likelihood for the model compared to the log likelihood for a baseline model. However, with categorical outcomes, it has a theoretical maximum value of less than 1, even for a "perfect" model.
- Nagelkerke's R^2(Nagelkerke, 1991) is an adjusted version of the Cox & Snell *R*-square that adjusts the scale of the statistic to cover the full range from 0 to 1.
- McFadden's R^2(McFadden, 1974) is another version, based on the log-likelihood kernels for the intercept-only model and the full estimated model.

What constitutes a "good" R^2 value varies between different areas of application. While these statistics can be suggestive on their own, they are most useful when comparing competing models for the same data. The model with the largest R^2 statistic is "best" according to this measure.

Classification

Figure 19-7
Classification table

	Predicted		
Observed	No	Yes	Percent Correct
No	188289.667	31871.267	85.5%
Yes	49970.600	77675.133	60.9%
Overall Percent	68.5%	31.5%	76.5%

Dependent Variable: Previously defaulted (reference category = No)
Model: (Intercept), ed, age, employ, address, income, debtinc, creddebt, othdebt

The classification table shows the practical results of using the logistic regression model. For each case, the predicted response is *Yes* if that case's model-predicted logit is greater than 0. Cases are weighted by *finalweight*, so that the classification table reports the expected model performance in the population.

- Cells on the diagonal are correct predictions.
- Cells off the diagonal are incorrect predictions.

Based upon the cases used to create the model, you can expect to correctly classify 85.5% of the non-defaulters in the population using this model. Likewise, you can expect to correctly classify 60.9% of the defaulters. Overall, you can expect to classify 76.5% of the cases are classified correctly; however, because this table was constructed with the cases used to create the model, these estimates are likely to be overly optimistic.

Tests of Model Effects

Figure 19-8
Tests of between-subjects effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	11.000	4.000	14.669	.010
(Intercept)	1.000	14.000	5.777	.031
ed	4.000	11.000	1.683	.224
age	1.000	14.000	5.352	.036
employ	1.000	14.000	88.244	.000
address	1.000	14.000	1.123	.307
income	1.000	14.000	.007	.932
debtinc	1.000	14.000	27.632	.000
creddebt	1.000	14.000	33.402	.000
othdebt	1.000	14.000	.709	.414

Dependent Variable: Previously defaulted (reference category = No)
Model: (Intercept), ed, age, employ, address, income, debtinc, creddebt, othdebt

Each term in the model, plus the model as a whole, is tested for whether its effect equals 0. Terms with significance values less than 0.05 have some discernable effect. Thus, *age*, *employ*, *debtinc*, and *creddebt* contribute to the model, while the other main effects do not. In a further analysis of the data, you would probably remove *ed*, *address*, *income*, and *othdebt* from model consideration.

Parameter Estimates

Figure 19-9
Parameter estimates

Previously defaulted	Parameter	B	Std. Error	95% Confidence Interval		Design Effect	Exp(B)	95% Confidence Interval for Exp(B)	
				Lower	Upper			Lower	Upper
Yes	(Intercept)	-1.140	.399	-1.995	-.284	.665	.320	.136	.753
	[ed=1]	.720	.340	-.010	1.449	.862	2.054	.990	4.259
	[ed=2]	.684	.371	-.112	1.481	1.247	1.983	.894	4.397
	[ed=3]	.518	.307	-.140	1.177	.813	1.679	.869	3.244
	[ed=4]	.789	.302	.142	1.437	.817	2.202	1.152	4.208
	[ed=5]	.000[a]	.	.	.	.	1.000	.	.
	age	-.023	.010	-.043	-.002	.418	.978	.958	.998
	employ	-.225	.024	-.277	-.174	1.200	.798	.758	.840
	address	-.028	.026	-.085	.029	.651	.972	.919	1.029
	income	.000	.003	-.007	.006	1.410	1.000	.993	1.006
	debtinc	.095	.018	.056	.134	1.222	1.100	1.058	1.143
	creddebt	.493	.085	.310	.676	1.373	1.637	1.363	1.966
	othdebt	.026	.031	-.041	.094	1.219	1.027	.960	1.098

Dependent Variable: Previously defaulted (reference category = No)
Model: (Intercept), ed, age, employ, address, income, debtinc, creddebt, othdebt
a. Set to zero because this parameter is redundant.

The parameter estimates table summarizes the effect of each predictor. Note that parameter values affect the likelihood of the "did default" category relative to the "did not default" category. Thus, parameters with positive coefficients increase the likelihood of default, while parameters with negative coefficients decrease the likelihood of default.

The meaning of a logistic regression coefficient is not as straightforward as that of a linear regression coefficient. While *B* is convenient for testing the model effects, *Exp(B)* is easier to interpret. *Exp(B)* represents the ratio change in the odds of the event of interest attributable to a one-unit increase in the predictor for predictors that are not part of interaction terms. For example, *Exp(B)* for *employ* is equal to 0.798, which means that the odds of default for people who have been with their current employer for two years are 0.798 times the odds of default for those who have been with their current employer for one year, all other things being equal.

The design effects indicate that some of the standard errors computed for these parameter estimates are larger than those you would obtain if you assumed that these observations came from a simple random sample, while others are smaller. It is vitally important to incorporate the sampling design information in your analysis because you might otherwise infer, for example, that the age coefficient is no different from 0!

Odds Ratios

Figure 19-10
Odds ratios for level of education

		Previously defaulted	Odds Ratio	95% Confidence Interval Lower	95% Confidence Interval Upper
Level of education	Did not complete high school vs. Post-undergraduate degree	Yes	2.054	.990	4.259
	High school degree vs.	Yes	1.983	.894	4.397
	Some college vs.	Yes	1.679	.869	3.244
	College degree vs.	Yes	2.202	1.152	4.208

Dependent Variable: Previously defaulted (reference category = No)
Model: (Intercept), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. Factors and covariates used in the computation are fixed at the following values: Level of education=Post-undergraduate degree; Age in years=34.19; Years with current employer=6.99; Years at current address=6.32; Household income in thousands=60.1581; Debt to income ratio (x100)=9.9341; Credit card debt in thousands=1.9764; Other debt in thousands=3.9164

This table displays the odds ratios of *Previously defaulted* at the factor levels of *Level of education*. The reported values are the ratios of the odds of default for *Did not complete high school* through *College degree*, compared to the odds of default for *Post-undergraduate degree*. Thus, the odds ratio of 2.054 in the first row of the table means that the odds of default for a person who did not complete high school are 2.054 times the odds of default for a person who has a post-undergraduate degree.

Figure 19-11
Odds ratios for years with current employer

Units of Change		Previously defaulted	Odds Ratio	95% Confidence Interval Lower	95% Confidence Interval Upper
Years with current employer	1.000	Yes	.798	.758	.840

Dependent Variable: Previously defaulted (reference category = No)
Model: (Intercept), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. Factors and covariates used in the computation are fixed at the following values: Level of education=Post-undergraduate degree; Age in years=34.19; Years with current employer=6.99; Years at current address=6.32; Household income in thousands=60.1581; Debt to income ratio (x100)=9.9341; Credit card debt in thousands=1.9764; Other debt in thousands=3.9164

This table displays the odds ratio of *Previously defaulted* for a unit change in the covariate *Years with current employer*. The reported value is the ratio of the odds of default for a person with 7.99 years at their current job compared to the odds of default for a person with 6.99 years (the mean).

Figure 19-12
Odds ratios for debt to income ratio

Units of Change		Previously defaulted	Odds Ratio	95% Confidence Interval	
				Lower	Upper
Debt to income ratio (x100)	1.000	Yes	1.100	1.058	1.143

Dependent Variable: Previously defaulted (reference category = No)
Model: (Intercept), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. Factors and covariates used in the computation are fixed at the following values: Level of education=Post-undergraduate degree; Age in years=34.19; Years with current employer=6.99; Years at current address=6.32; Household income in thousands=60.1581; Debt to income ratio (x100)=9.9341; Credit card debt in thousands=1.9764; Other debt in thousands=3.9164

This table displays the odds ratio of *Previously defaulted* for a unit change in the covariate *Debt to income ratio*. The reported value is the ratio of the odds of default for a person with a debt/income ratio of 10.9341 compared to the odds of default for a person with 9.9341 (the mean).

Note that because none of these predictors are part of interaction terms, the values of the odds ratios reported in these tables are equal to the values of the exponentiated parameter estimates. When a predictor is part of an interaction term, its odds ratio as reported in these tables will also depend on the values of the other predictors that make up the interaction.

Summary

Using the Complex Samples Logistic Regression Procedure, you have constructed a model for predicting the probability that a given customer will default on a loan.

A critical issue for loan officers is the cost of Type I and Type II errors. That is, what is the cost of classifying a defaulter as a non-defaulter (Type I)? What is the cost of classifying a non-defaulter as a defaulter (Type II)? If bad debt is the primary concern, then you want to lower your Type I error and maximize your **sensitivity**. If growing your customer base is the priority, then you want to lower your Type II error and maximize your **specificity**. Usually, both are major concerns, so you have to choose a decision rule for classifying customers that gives the best mix of sensitivity and specificity.

Related Procedures

The Complex Samples Logistic Regression procedure is a useful tool for modeling a categorical variable when the cases have been drawn according to a complex sampling scheme.

- The Complex Samples Sampling Wizard is used to specify complex sampling design specifications and obtain a sample. The sampling plan file created by the Sampling Wizard contains a default analysis plan and can be specified in the Plan dialog box when you are analyzing the sample obtained according to that plan.
- The Complex Samples Analysis Preparation Wizard is used to specify analysis specifications for an existing complex sample. The analysis plan file created by the Sampling Wizard can be specified in the Plan dialog box when you are analyzing the sample corresponding to that plan.
- The Complex Samples General Linear Model procedure allows you to model a scale response.
- The Complex Samples Ordinal Regression procedure allows you to model an ordinal response.

Chapter

20

Complex Samples Ordinal Regression

The Complex Samples Ordinal Regression procedure creates a predictive model for an ordinal dependent variable for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Using Complex Samples Ordinal Regression to Analyze Survey Results

Representatives considering a bill before the legislature are interested in whether there is public support for the bill and how support for the bill is related to voter demographics. Pollsters design and conduct interviews according to a complex sampling design.

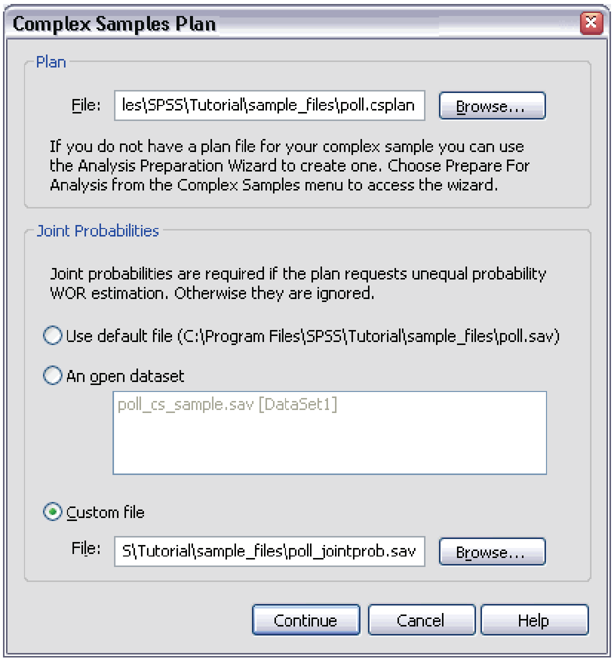
The survey results are collected in *poll_cs_sample.sav*. The sampling plan used by the pollsters is contained in *poll.csplan*; because it makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*poll_jointprob.sav*). Use Complex Samples Ordinal Regression to fit a model for the level of support for the bill based upon voter demographics.

Running the Analysis

▶ To run a Complex Samples Ordinal Regression analysis, from the menus choose:

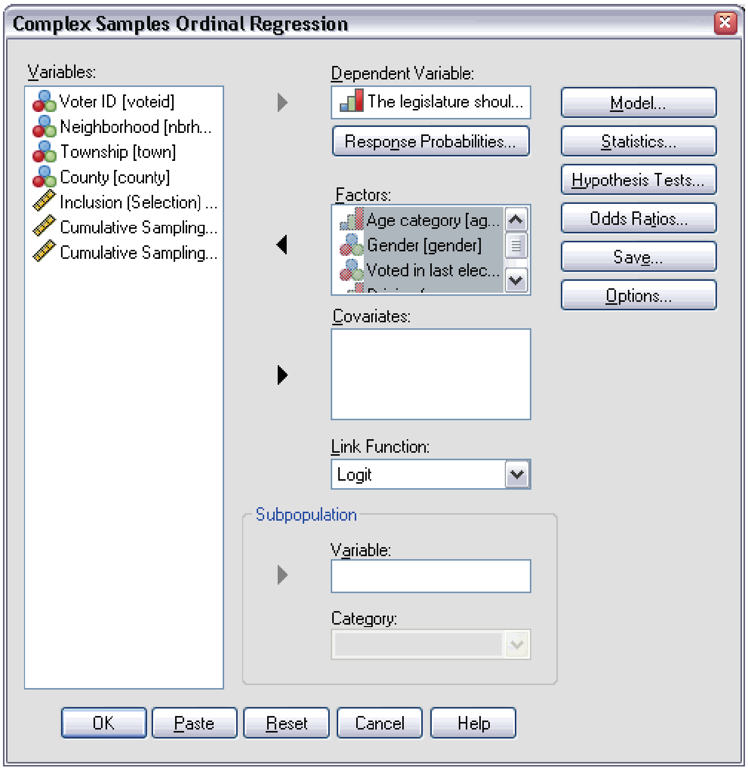
Analyze
  Complex Samples
    Ordinal Regression...

Figure 20-1
Complex Samples Plan dialog box



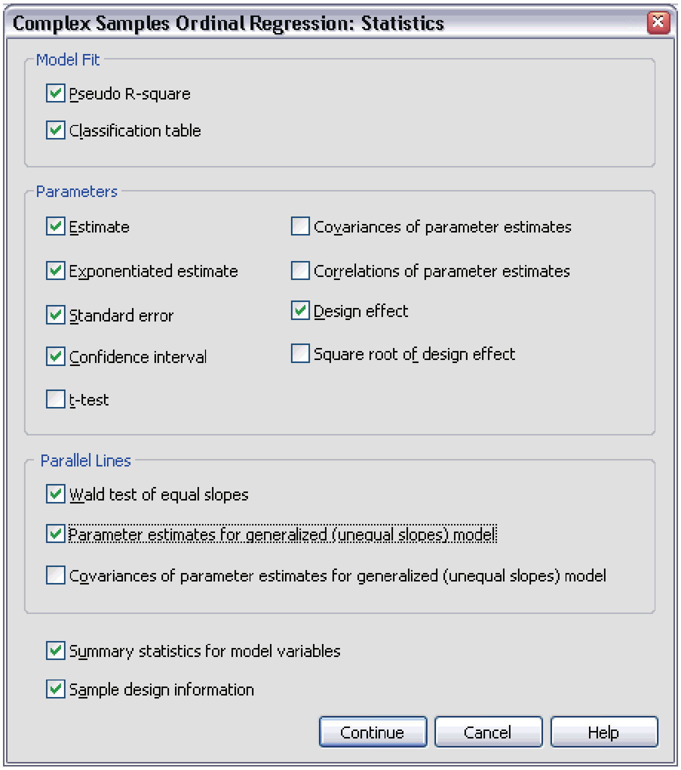
- ▶ Browse to the *\tutorial\sample_files* subdirectory of the directory in which you installed SPSS and select *poll.csplan* as the plan file.
- ▶ Select *poll_jointprob.sav* as the joint probabilities file.
- ▶ Click Continue.

Figure 20-2
Complex Samples Ordinal Regression main dialog box



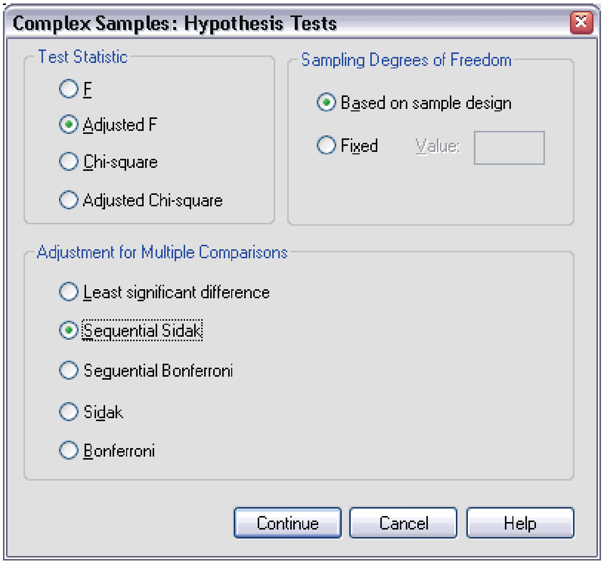
- Select *The legislature should enact a gas tax* as the dependent variable.
- Select *Age category* through *Driving frequency* as factors.
- Click Statistics.

Figure 20-3
Statistics dialog box



- Select Classification table in the Model Fit group.
- Select Estimate, Exponentiated estimate, Standard error, Confidence interval, and Design effect in the Parameters group.
- Select Wald test of equal slopes and Parameter estimates for generalized (unequal slopes) model.
- Click Continue.
- Click Hypothesis Tests in the Complex Samples Ordinal Regression dialog box.

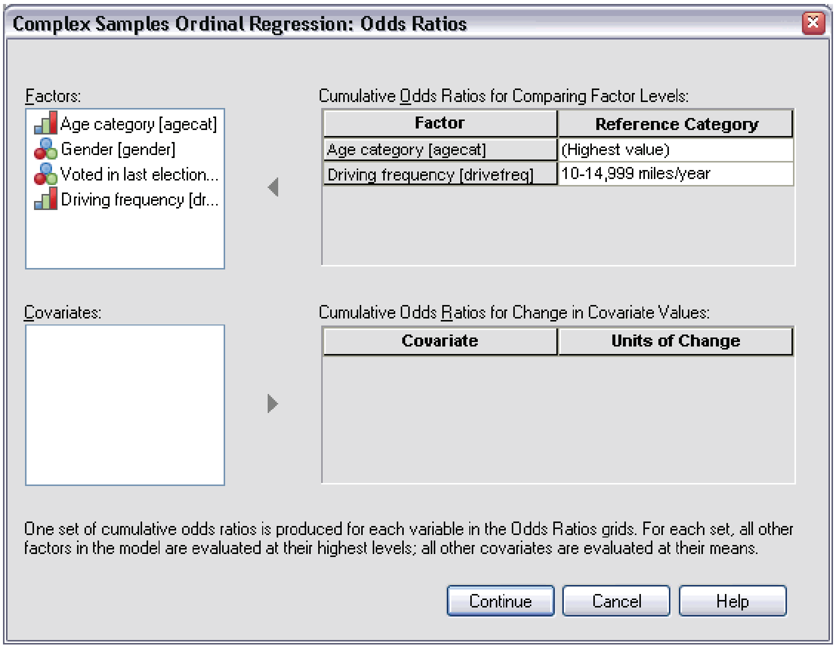
Figure 20-4
Hypothesis Tests dialog box



Even for a moderate number of predictors and response categories, the Wald *F* test statistic can be inestimable for the test of parallel lines.

- ▶ Select Adjusted F in the Test Statistic group.
- ▶ Select Sequential Sidak as the adjustment method for multiple comparisons.
- ▶ Click Continue.
- ▶ Click Odds Ratios in the Complex Samples Ordinal Regression dialog box.

Figure 20-5
Odds Ratios dialog box



- ▶ Choose to produce cumulative odds ratios for *Age category* and *Driving frequency*.
- ▶ Select 10-14,999 miles/year, a more “typical” yearly mileage than the maximum, as the reference category for *Driving frequency*.
- ▶ Click Continue.
- ▶ Click OK in the Complex Samples Ordinal Regression dialog box.

Pseudo R-Squares

Figure 20-6
Pseudo R-Squares

Cox and Snell	.179
Nagelkerke	.191
McFadden	.071

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, gender, votelast, drivefreq
Link function: Logit

In the linear regression model, the coefficient of determination, R^2, summarizes the proportion of variance in the dependent variable associated with the predictor (independent) variables, with larger R^2 values indicating that more of the variation is explained by the model, to a maximum of 1. For regression models with a categorical dependent variable, it is not possible to compute a single R^2 statistic that has all of the characteristics of R^2 in the linear regression model, so these approximations are computed instead. The following methods are used to estimate the coefficient of determination.

- Cox and Snell's R^2(Cox and Snell, 1989) is based on the log likelihood for the model compared to the log likelihood for a baseline model. However, with categorical outcomes, it has a theoretical maximum value of less than 1, even for a "perfect" model.
- Nagelkerke's R^2(Nagelkerke, 1991) is an adjusted version of the Cox & Snell *R*-square that adjusts the scale of the statistic to cover the full range from 0 to 1.
- McFadden's R^2(McFadden, 1974) is another version, based on the log-likelihood kernels for the intercept-only model and the full estimated model.

What constitutes a "good" R^2 value varies between different areas of application. While these statistics can be suggestive on their own, they are most useful when comparing competing models for the same data. The model with the largest R^2 statistic is "best" according to this measure.

Tests of Model Effects

Figure 20-7
Tests of model effects

Source	df1	df2	Adjusted Wald F	Sig.	Sequential Sidak Sig.
agecat	2.283	31.966	6.215	.004	.003
gender	1.000	14.000	.046	.834	.834
votelast	1.000	14.000	.076	.787	.787
drivefreq	3.785	52.987	228.015	.000	.000

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, gender, votelast, drivefreq
Link function: Logit

Each term in the model is tested for whether its effect equals 0. Terms with significance values less than 0.05 have some discernable effect. Thus, *agecat* and *drivefreq* contribute to the model, while the other main effects do not. In a further analysis of the data, you would consider removing *gender* and *votelast* from the model.

Parameter Estimates

The parameter estimates table summarizes the effect of each predictor. While interpretation of the coefficients in this model is difficult due to the nature of the link function, the signs of the coefficients for covariates and relative values of the coefficients for factor levels can give important insights into the effects of the predictors in the model.

- For covariates, positive (negative) coefficients indicate positive (inverse) relationships between predictors and outcome. An increasing value of a covariate with a positive coefficient corresponds to an increasing probability of being in one of the “higher” cumulative outcome categories.
- For factors, a factor level with a greater coefficient indicates a greater probability of being in one of the “higher” cumulative outcome categories. The sign of a coefficient for a factor level is dependent upon that factor level’s effect relative to the reference category.

Figure 20-8
Parameter estimates

Parameter		B	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper	Design Effect	Exp(B)	95% Confidence Interval for Exp(B) Lower	95% Confidence Interval for Exp(B) Upper
Threshold	[opinion_gastax=1]	-3.343	.104	-3.566	-3.120	1.132	.035	.028	.044
	[opinion_gastax=2]	-1.910	.098	-2.120	-1.700	1.058	.148	.120	.183
	[opinion_gastax=3]	-.674	.090	-.866	-.482	.915	.510	.421	.618
Regression	[agecat=1]	-.324	.079	-.494	-.154	1.793	.723	.610	.858
	[agecat=2]	-.138	.054	-.255	-.022	1.158	.871	.775	.978
	[agecat=3]	-.095	.076	-.257	.068	2.206	.909	.773	1.070
	[agecat=4]	.000[a]	.	.	.	.	1.000	.	.
	[gender=0]	-.008	.035	-.084	.068	.949	.992	.920	1.071
	[gender=1]	.000[a]	.	.	.	.	1.000	.	.
	[votelast=0]	-.011	.039	-.095	.073	1.103	.989	.909	1.076
	[votelast=1]	.000[a]	.	.	.	.	1.000	.	.
	[drivefreq=1]	-3.751	.153	-4.079	-3.423	1.117	.023	.017	.033
	[drivefreq=2]	-3.003	.116	-3.251	-2.755	1.226	.050	.039	.064
	[drivefreq=3]	-2.295	.114	-2.540	-2.050	1.585	.101	.079	.129
	[drivefreq=4]	-1.570	.092	-1.769	-1.372	1.078	.208	.171	.254
	[drivefreq=5]	-.812	.089	-1.003	-.621	.941	.444	.367	.537
	[drivefreq=6]	.000[a]	.	.	.	.	1.000	.	.

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, gender, votelast, drivefreq
Link function: Logit

a. Set to zero because this parameter is redundant.

You can make the following interpretations based on the parameter estimates:

- Those in lower age categories show greater support for the bill than those in the highest age category.
- Those who drive less frequently show greater support for the bill than those who drive more frequently.
- The coefficients for the variables *gender* and *votelast*, in addition to not being statistically significant, appear to be small compared to other coefficients.

The design effects indicate that some of the standard errors computed for these parameter estimates are larger than those you would obtain if you used a simple random sample, while others are smaller. It is vitally important to incorporate the sampling design information in your analysis because you might otherwise infer, for example, that the coefficient for the third level of *Age category*, *[agecat=3]*, is significantly different from 0!

Classification

Figure 20-9
Categorical variable information

		Weighted Count	Weighted Percent
The legislature should enact a gas tax[a]	Strongly agree	25132.955	21.3%
	Agree	32261.425	27.3%
	Disagree	29477.417	24.9%
	Strongly disagree	31314.203	26.5%
Age category	18-30	20509.504	17.4%
	31-45	35380.506	29.9%
	46-60	34865.792	29.5%
	>60	27430.198	23.2%
Gender	Male	61424.547	52.0%
	Female	56761.453	48.0%
Voted in last election	No	70607.216	59.7%
	Yes	47578.784	40.3%
Driving frequency	Do not own car	3437.137	2.9%
	<10,000 miles/year	10816.349	9.2%
	10-14,999 miles/year	32539.364	27.5%
	15-19,999 miles/year	39179.814	33.2%
	20-29,999 miles/year	25617.804	21.7%
	>=30,000 miles/year	6595.532	5.6%
Population Size		118186.000	100.0%

a. Dependent variable values are sorted in ascending order.

Given the observed data, the "null" model (that is, one without predictors) would classify all customers into the modal group, *Agree*. Thus, the null model would be correct 27.3% of the time.

Figure 20-10
Classification table

Observed	Predicted				
	Strongly agree	Agree	Disagree	Strongly disagree	Percent Correct
Strongly agree	7067.567	12130.814	3875.825	2058.750	28.1%
Agree	4271.234	14464.286	7320.767	6205.137	44.8%
Disagree	2024.816	11703.368	7108.487	8640.746	24.1%
Strongly disagree	889.869	8169.109	6946.522	15308.703	48.9%
Overall Percent	12.1%	39.3%	21.4%	27.3%	37.2%

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, gender, votelast, drivefreq
Link function: Logit

The classification table shows the practical results of using the model. For each case, the predicted response is the response category with the highest model-predicted probability. Cases are weighted by *Final Sampling Weight*, so that the classification table reports the expected model performance in the population.

- Cells on the diagonal are correct predictions.
- Cells off the diagonal are incorrect predictions.

The model correctly classifies 9.9% more, or 37.2% of the cases. In particular, the model does considerably better at classifying those who *Agree* or *Strongly disagree*, and slightly worse with those who *Disagree*.

Odds Ratios

Cumulative odds are defined as the ratio of the probability that the dependent variable takes a value less than or equal to a given response category to the probability that it takes a value greater than that response category. The **cumulative odds ratio** is the ratio of cumulative odds for different predictor values, and is closely related to the exponentiated parameter estimates. Interestingly, the cumulative odds ratio itself does not depend upon the response category.

Figure 20-11
Cumulative odds ratios for Age category

		Cumulative Odds Ratio	95% Confidence Interval		Design Effect	Square Root Design Effect
			Lower	Upper		
Age category	18-30 vs. >60	1.383	1.166	1.639	1.793	1.339
	31-45 vs. >60	1.148	1.022	1.290	1.158	1.076
	46-60 vs. >60	1.100	.935	1.294	2.206	1.485

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, gender, votelast, drivefreq
Link function: Logit

a. Factors and covariates used in the computation are fixed at the following values: Age category=>60; Gender=Female; Voted in last election=Yes; Driving frequency=>=30,000 miles/year

This table displays cumulative odds ratios for the factor levels of *Age category*. The reported values are the ratios of the cumulative odds for *18–30* through *46–60*, compared to the cumulative odds for *>60*. Thus, the odds ratio of 1.383 in the first row of the table means that the cumulative odds for a person aged 18–30 are 1.383 times the cumulative odds for a person older than 60. Note that because *Age category* is not involved in any interaction terms, the odds ratios are merely the ratios of the exponentiated parameter estimates. For example, the cumulative odds ratio for *18–30* vs. *>60* is 1.00 / 0.723 = 1.383.

Figure 20-12
Odds ratios for driving frequency

		Cumulative Odds Ratio	95% Confidence Interval		Design Effect	Square Root Design Effect
			Lower	Upper		
Driving frequency	Do not own car vs. 10-14,999 miles/year	4.288	2.878	6.390	2.345	1.531
	<10,000 miles/year vs. 10-14,999 miles/year	2.030	1.656	2.488	1.838	1.356
	15-19,999 miles/year vs. 10-14,999 miles/year	.484	.430	.546	1.450	1.204
	20-29,999 miles/year vs. 10-14,999 miles/year	.227	.193	.267	2.095	1.448
	>=30,000 miles/year vs. 10-14,999 miles/year	.101	.079	.129	1.585	1.259

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, gender, votelast, drivefreq
Link function: Logit

a. Factors and covariates used in the computation are fixed at the following values: Age category=>60; Gender=Female; Voted in last election=Yes; Driving frequency=>=30,000 miles/year

This table displays the cumulative odds ratios for the factor levels of *Driving frequency*, using *10–14,999 miles/year* as the reference category. Since *Driving frequency* is not involved in any interaction terms, the odds ratios are merely the ratios of the exponentiated parameter estimates. For example, the cumulative odds ratio for *20–29,999 miles/year* vs. *10–14,999 miles/year* is 0.101 / 0.444 = 0.227.

Generalized Cumulative Model

Figure 20-13
Test of parallel lines

df1	df2	Adjusted Wald F	Sig.	Sequential Sidak Sig.
8.769	122.767	1.894	.061	.392

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, gender, votelast, drivefreq
Link function: Logit

The test of parallel lines can help you assess whether the assumption that the parameters are the same for all response categories is reasonable. This test compares the estimated model with one set of coefficients for all categories to a generalized model with a separate set of coefficients for each category.

The Wald *F* test is an omnibus test of the contrast matrix for the parallel lines assumption that provides asymptotically correct *p* values; for small to mid-sized samples, the adjusted Wald *F* statistic performs well. The significance value is near 0.05, suggesting that the generalized model may give an improvement in the model fit; however, the Sequential Sidak adjusted test reports a significance value high enough (0.392) that, overall, there is no clear evidence for rejecting the parallel lines assumption. The Sequential Sidak test starts with individual contrast Wald tests to provide an overall *p* value, and these results should be comparable to the omnibus Wald test result. The fact that they are so different in this example is somewhat surprising but could be due to the existence of many contrasts in the test and a relatively small design degrees of freedom.

Figure 20-14
Parameter estimates for generalized cumulative model (shown in part)

The legislature should enact a gas tax	Parameter	B	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper
Strongly agree	(Threshold)	-3.681	.221	-4.155	-3.207
	[agecat=1]	-.320	.096	-.525	-.115
	[agecat=2]	-.075	.071	-.227	.077
	[agecat=3]	-.022	.073	-.180	.135
	[agecat=4]	.000[a]	.		
	[gender=0]	-.082	.054	-.197	.033
	[gender=1]	.000[a]	.		
	[votelast=0]	.008	.052	-.104	.120
	[votelast=1]	.000[a]	.		
	[drivefreq=1]	-4.096	.267	-4.669	-3.523
	[drivefreq=2]	-3.367	.237	-3.876	-2.857
	[drivefreq=3]	-2.678	.224	-3.158	-2.199
	[drivefreq=4]	-1.928	.213	-2.384	-1.471
	[drivefreq=5]	-1.015	.252	-1.555	-.476
	[drivefreq=6]	.000[a]	.		
Agree	(Threshold)	-1.963	.153	-2.291	-1.635
	[agecat=1]	-.385	.095	-.587	-.182
	[agecat=2]	-.130	.069	-.279	.018
	[agecat=3]	-.139	.101	-.356	.077
	[agecat=4]	.000[a]	.		
	[gender=0]	-.004	.040	-.090	.082
	[gender=1]	.000[a]	.		
	[votelast=0]	.009	.059	-.117	.135
	[votelast=1]	.000[a]	.		
	[drivefreq=1]	-3.867	.318	-4.549	-3.185
	[drivefreq=2]	-3.005	.175	-3.380	-2.630
	[drivefreq=3]	-2.290	.187	-2.691	-1.888
	[drivefreq=4]	-1.633	.166	-1.988	-1.278
	[drivefreq=5]	-.909	.137	-1.204	-.615
	[drivefreq=6]	.000[a]	.		

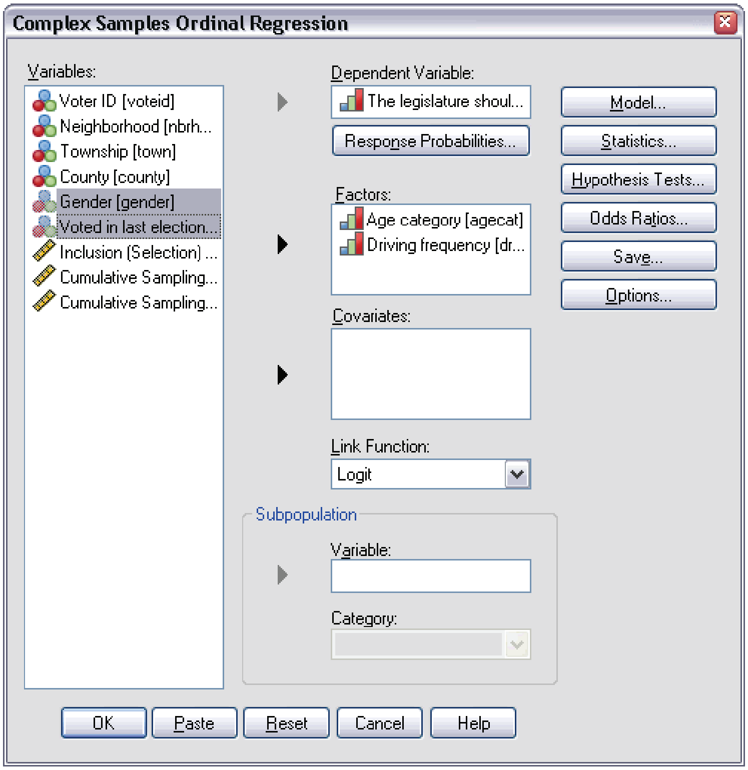
Moreover, the estimated values of the generalized model coefficients don't appear to differ much from the estimates under the parallel lines assumption.

Dropping Non-Significant Predictors

The tests of model effects showed that the model coefficients for *Gender* and *Voted in last election* are not statistically significantly different from 0. To produce a reduced model, recall the Complex Samples Ordinal Regression dialog box.

▶ Click Continue in the Plan dialog box.

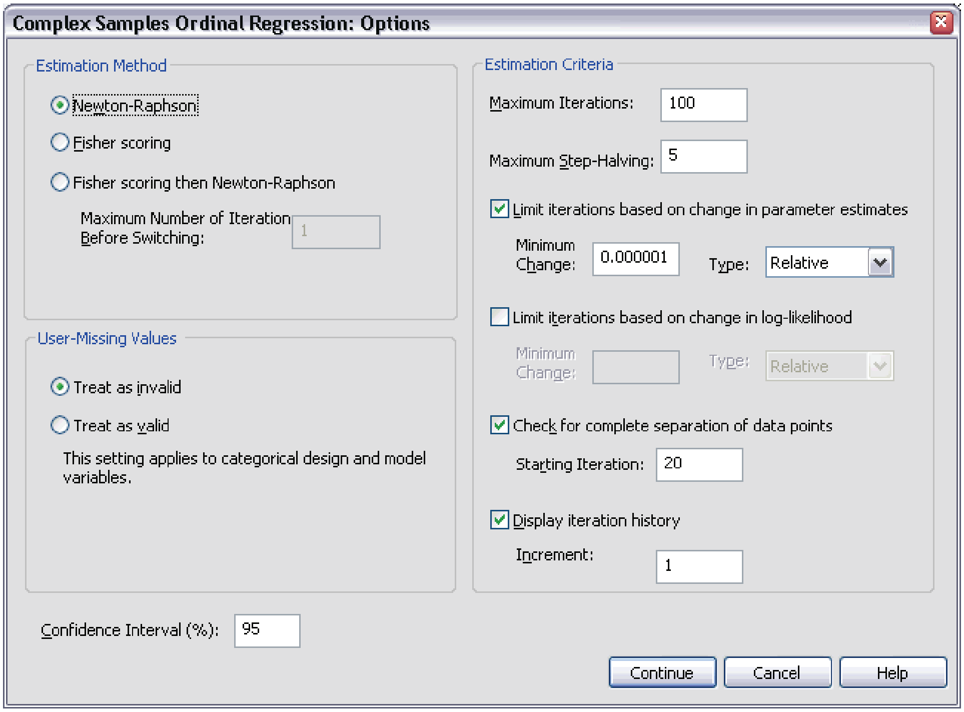
Figure 20-15
Complex Samples Ordinal Regression main dialog box



▶ Deselect *Gender* and *Voted in last election* as factors.

▶ Click Options.

Figure 20-16
Options dialog box



- Select Display iteration history.

  The iteration history is useful for diagnosing problems encountered by the estimation algorithm.

- Click Continue.
- Click OK in the Complex Samples Ordinal Regression dialog box.

Warnings

Figure 20-17
Warnings for reduced model

The log-likelihood value cannot be increased after the maximum number of steps in the step-halving method.
The CSORDINAL procedure continues despite the above warning(s). Subsequent results shown are based on the last iteration. Validity of the model fit is uncertain.
The following message applies to the generalized cumulative model.
The log-likelihood value cannot be increased after the maximum number of steps in the step-halving method.

The warnings note that estimation of the reduced model ended before the parameter estimates reached convergence because the log-likelihood could not be increased with any change, or "step," in the current values of the parameter estimates.

Figure 20-18
Warnings for reduced model

			Threshold			Regression							
Iteration Number[b]	N Step-halving	Pseudo -2 Log Likelihood	[opinion_gastax=1]	[opinion_gastax=2]	[opinion_gastax=3]	[agecat=1]	[agecat=2]	[agecat=3]	[drivefreq=1]	[drivefreq=2]	[drivefreq=3]	[drivefreq=4]	[drivefreq=5]
0	0	326640.3	-1.309	-.058	1.020	.000	.000	.000	.000	.000	.000	.000	.000
1	0	303567.5	-3.242	-1.881	-.704	-.323	-.137	-.094	-3.841	-2.970	-2.248	-1.563	-.835
2	0	303336.3	-3.327	-1.897	-.664	-.325	-.139	-.095	-3.740	-2.998	-2.291	-1.568	-.811
3	0	303335.9	-3.333	-1.900	-.664	-.326	-.139	-.096	-3.750	-3.003	-2.295	-1.570	-.812
4	0	303335.9	-3.333	-1.900	-.664	-.326	-.139	-.096	-3.750	-3.003	-2.295	-1.570	-.812
5[a]	5	303335.9	-3.333	-1.900	-.664	-.326	-.139	-.096	-3.750	-3.003	-2.295	-1.570	-.812

Redundant parameters are not displayed. Their values are always zero in all iterations.
Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, drivefreq
Link function: Logit

a. The log-likelihood value cannot be increased after the maximum number of steps in the step-halving method.

b. Newton-Raphson method was used to estimate the parameters.

Looking at the iteration history, the changes in the parameter estimates over the last few iterations are slight enough that you're not terribly concerned about the warning message.

Comparing Models

Figure 20-19
Pseudo R-Squares for reduced model

Cox and Snell	.179
Nagelkerke	.191
McFadden	.071

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, gender, votelast, drivefreq
Link function: Logit

The R^2 values for the reduced model are identical to those for the original model. This is evidence in favor of the reduced model.

Figure 20-20
Classification table for reduced model

Observed	Predicted: Strongly agree	Agree	Disagree	Strongly disagree	Percent Correct
Strongly agree	7067.567	12823.258	3183.380	2058.750	28.1%
Agree	4271.234	15684.090	6100.963	6205.137	48.6%
Disagree	2024.816	13157.809	5654.047	8640.746	19.2%
Strongly disagree	889.869	9226.578	5889.053	15308.703	48.9%
Overall Percent	12.1%	43.1%	17.6%	27.3%	37.0%

Dependent Variable: The legislature should enact a gas tax (Ascending)
Model: (Threshold), agecat, drivefreq
Link function: Logit

The classification table somewhat complicates matters. The overall classification rate of 37.0% for the reduced model is comparable to the original model, which is evidence in favor of the reduced model. However, the reduced model shifts the predicted response of 3.8% of the voters from *Disagree* to *Agree*, more than half of whom were observed to respond *Disagree* or *Strongly disagree*. This is a very important distinction that deserves careful consideration before choosing the reduced model.

Summary

Using the Complex Samples Ordinal Regression Procedure, you have constructed competing models for the level of support for the proposed bill based on voter demographics. The test of parallel lines shows that a generalized cumulative model is not necessary. The tests of model effects suggest that *Gender* and *Voted in last election* could be dropped from the model, and the reduced model performs well in terms of pseudo-R^2 and overall classification rate compared to the original model. However, the reduced model misclassifies more voters across the *Agree/Disagree* split, so the legislators prefer to keep the original model for now.

Related Procedures

The Complex Samples Ordinal Regression procedure is a useful tool for modeling an ordinal variable when the cases have been drawn according to a complex sampling scheme.

- The Complex Samples Sampling Wizard is used to specify complex sampling design specifications and obtain a sample. The sampling plan file created by the Sampling Wizard contains a default analysis plan and can be specified in the Plan dialog box when you are analyzing the sample obtained according to that plan.
- The Complex Samples Analysis Preparation Wizard is used to specify analysis specifications for an existing complex sample. The analysis plan file created by the Sampling Wizard can be specified in the Plan dialog box when you are analyzing the sample corresponding to that plan.
- The Complex Samples General Linear Model procedure allows you to model a scale response.
- The Complex Samples Logistic Regression procedure allows you to model a categorical response.

Bibliography

Cochran, W. G. 1977. *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons.

Cox, D. R., and E. J. Snell. 1989. *The Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley and Sons.

Kish, L. 1987. *Statistical Design for Research*. New York: John Wiley and Sons.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In: *Frontiers in Economics,* P. Zarembka, ed. New York: Academic Press.

Murthy, M. N. 1967. *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

Nagelkerke, N. J. D. 1991. A note on the general definition of the coefficient of determination. *Biometrika*, 78:3, 691–692.

Särndal, C., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Index

adjusted chi-square
in Complex Samples, 60, 73, 86
adjusted *F* statistic
in Complex Samples, 60, 73, 86
adjusted residuals
in Complex Samples Crosstabs, 46
analysis plan, 21

Bonferroni
in Complex Samples, 60, 73, 86
Brewer's sampling method
in Sampling Wizard, 8

chi-square
in Complex Samples, 60, 73, 86
classification table
in Complex Samples Logistic Regression, 205
in Complex Samples Ordinal Regression, 220
classification tables
in Complex Samples Logistic Regression, 71
in Complex Samples Ordinal Regression, 84
clusters
in Analysis Preparation Wizard, 23
in Sampling Wizard, 6
coefficient of variation (COV)
in Complex Samples Crosstabs, 46
in Complex Samples Descriptives, 41
in Complex Samples Frequencies, 36
in Complex Samples Ratios, 52
column percentages
in Complex Samples Crosstabs, 46
Complex Samples
hypothesis tests, 60, 73, 86
missing values, 37, 48
options, 37, 42, 48, 53
Complex Samples Analysis Preparation Wizard, 144
public data, 144
related procedures, 158
sampling weights not available, 147
summary, 147, 158
Complex Samples Crosstabs, 44, 171
crosstabulation table, 175
related procedures, 177
relative risk, 171, 175, 177
statistics, 46
Complex Samples Descriptives, 39, 165
missing values, 42
public data, 165
related procedures, 170
statistics, 41, 168
statistics by subpopulation, 169
Complex Samples Frequencies, 34, 159
frequency table, 162
frequency table by subpopulation, 163
related procedures, 164
statistics, 36
Complex Samples General Linear Model, 55, 186
command additional features, 65
estimated means, 62
marginal means, 194
model, 57
model summary, 191
options, 64
parameter estimates, 193
related procedures, 197
save variables, 63
statistics, 59
tests of model effects, 192
Complex Samples Logistic Regression, 66, 199
classification table, 205
command additional features, 78
model, 69
odds ratios, 74, 208
options, 77
parameter estimates, 207
pseudo R^2 statistics, 204
reference category, 68
related procedures, 210
save variables, 75
statistics, 71
tests of model effects, 206

Complex Samples Ordinal Regression, 79, 211
classification table, 220
generalized cumulative model, 223
model, 82
odds ratios, 87, 221
options, 90
parameter estimates, 218
pseudo R^2 statistics, 217, 228
related procedures, 229
response probabilities, 81
save variables, 89
statistics, 84
tests of model effects, 218
warnings, 227
Complex Samples Ratios, 50, 179
missing values, 53
ratios, 182
related procedures, 184
statistics, 52
Complex Samples Sampling Wizard, 94
PPS sampling, 126
related procedures, 143
sampling frame, full, 94
sampling frame, partial, 107
summary, 104–105, 138
complex sampling
analysis plan, 21
sample plan, 4
confidence intervals
in Complex Samples Crosstabs, 46
in Complex Samples Descriptives, 41, 168–169
in Complex Samples Frequencies, 36, 162–163
in Complex Samples General Linear Model, 59, 64
in Complex Samples Logistic Regression, 71
in Complex Samples Ordinal Regression, 84
in Complex Samples Ratios, 52
confidence level
in Complex Samples Logistic Regression, 77
in Complex Samples Ordinal Regression, 90
contrasts
in Complex Samples General Linear Model, 62
correlations of parameter estimates
in Complex Samples General Linear Model, 59
in Complex Samples Logistic Regression, 71
in Complex Samples Ordinal Regression, 84
covariances of parameter estimates
in Complex Samples General Linear Model, 59
in Complex Samples Logistic Regression, 71
in Complex Samples Ordinal Regression, 84
crosstabulation table
in Complex Samples Crosstabs, 175
cumulative probabilities
in Complex Samples Ordinal Regression, 89
cumulative values
in Complex Samples Frequencies, 36

degrees of freedom
in Complex Samples, 60, 73, 86
design effect
in Complex Samples Crosstabs, 46
in Complex Samples Descriptives, 41
in Complex Samples Frequencies, 36
in Complex Samples General Linear Model, 59
in Complex Samples Logistic Regression, 71
in Complex Samples Ordinal Regression, 84
in Complex Samples Ratios, 52
deviation contrasts
in Complex Samples General Linear Model, 62
difference contrasts
in Complex Samples General Linear Model, 62

estimated marginal means
in Complex Samples General Linear Model, 62
expected values
in Complex Samples Crosstabs, 46

F statistic
in Complex Samples, 60, 73, 86
Fisher scoring
in Complex Samples Ordinal Regression, 90

generalized cumulative model
in Complex Samples Ordinal Regression, 223

Helmert contrasts
in Complex Samples General Linear Model, 62

inclusion probabilities
in Sampling Wizard, 12

input sample weights
in Sampling Wizard, 6
iteration history
in Complex Samples Logistic Regression, 77
in Complex Samples Ordinal Regression, 90
iterations
in Complex Samples Logistic Regression, 77
in Complex Samples Ordinal Regression, 90

least significant difference
in Complex Samples, 60, 73, 86
likelihood convergence
in Complex Samples Logistic Regression, 77
in Complex Samples Ordinal Regression, 90

marginal means
in GLM Univariate, 194
mean
in Complex Samples Descriptives, 41, 168–169
measure of size
in Sampling Wizard, 8
missing values
in Complex Samples, 37, 48
in Complex Samples Descriptives, 42
in Complex Samples General Linear Model, 64
in Complex Samples Logistic Regression, 77
in Complex Samples Ordinal Regression, 90
in Complex Samples Ratios, 53
Murthy's sampling method
in Sampling Wizard, 8

Newton-Raphson method
in Complex Samples Ordinal Regression, 90

odds ratios
in Complex Samples Crosstabs, 46, 171
in Complex Samples Logistic Regression, 74, 208
in Complex Samples Ordinal Regression, 87, 221

parameter convergence
in Complex Samples Logistic Regression, 77
in Complex Samples Ordinal Regression, 90
parameter estimates
in Complex Samples General Linear Model, 59, 193
in Complex Samples Logistic Regression, 71, 207
in Complex Samples Ordinal Regression, 84, 218
plan file, 3
polynomial contrasts
in Complex Samples General Linear Model, 62
population size
in Complex Samples Crosstabs, 46
in Complex Samples Descriptives, 41
in Complex Samples Frequencies, 36, 162–163
in Complex Samples Ratios, 52
in Sampling Wizard, 12
PPS sampling
in Sampling Wizard, 8
predicted categories
in Complex Samples Logistic Regression, 75
in Complex Samples Ordinal Regression, 89
predicted probability
in Complex Samples Logistic Regression, 75
in Complex Samples Ordinal Regression, 89
predicted values
in Complex Samples General Linear Model, 63
pseudo R^2 statistics
in Complex Samples Logistic Regression, 71, 204
in Complex Samples Ordinal Regression, 84, 217, 228
public data
in Analysis Preparation Wizard, 144
in Complex Samples Descriptives, 165

R^2 statistic
in Complex Samples General Linear Model, 59, 191
ratios
in Complex Samples Ratios, 182
reference category
in Complex Samples General Linear Model, 62
in Complex Samples Logistic Regression, 68
relative risk
in Complex Samples Crosstabs, 46, 171, 175, 177
repeated contrasts
in Complex Samples General Linear Model, 62
residuals
in Complex Samples Crosstabs, 46

in Complex Samples General Linear Model, 63
response probabilities
in Complex Samples Ordinal Regression, 81
risk difference
in Complex Samples Crosstabs, 46
row percentages
in Complex Samples Crosstabs, 46

Sampford's sampling method
in Sampling Wizard, 8
sample plan, 4
sample proportion
in Sampling Wizard, 12
sample size
in Sampling Wizard, 10, 12
sample weights
in Analysis Preparation Wizard, 23
in Sampling Wizard, 12
sampling
complex design, 4
sampling estimation
in Analysis Preparation Wizard, 25
sampling frame, full
in Sampling Wizard, 94
sampling frame, partial
in Sampling Wizard, 107
sampling method
in Sampling Wizard, 8
separation
in Complex Samples Logistic Regression, 77
in Complex Samples Ordinal Regression, 90
sequential Bonferroni correction
in Complex Samples, 60, 73, 86
sequential sampling
in Sampling Wizard, 8
sequential Sidak correction
in Complex Samples, 60, 73, 86
Sidak correction
in Complex Samples, 60, 73, 86
simple contrasts
in Complex Samples General Linear Model, 62
simple random sampling
in Sampling Wizard, 8
square root of design effect
in Complex Samples Crosstabs, 46
in Complex Samples Descriptives, 41
in Complex Samples Frequencies, 36
in Complex Samples General Linear Model, 59
in Complex Samples Logistic Regression, 71
in Complex Samples Ordinal Regression, 84
in Complex Samples Ratios, 52
standard error
in Complex Samples Crosstabs, 46
in Complex Samples Descriptives, 41, 168–169
in Complex Samples Frequencies, 36, 162–163
in Complex Samples General Linear Model, 59
in Complex Samples Logistic Regression, 71
in Complex Samples Ordinal Regression, 84
in Complex Samples Ratios, 52
step-halving
in Complex Samples Logistic Regression, 77
in Complex Samples Ordinal Regression, 90
stratification
in Analysis Preparation Wizard, 23
in Sampling Wizard, 6
sum
in Complex Samples Descriptives, 41
summary
in Analysis Preparation Wizard, 147, 158
in Sampling Wizard, 104–105, 138
systematic sampling
in Sampling Wizard, 8

t test
in Complex Samples General Linear Model, 59
in Complex Samples Logistic Regression, 71
in Complex Samples Ordinal Regression, 84
table percentages
in Complex Samples Crosstabs, 46
in Complex Samples Frequencies, 36, 162–163
test of parallel lines
in Complex Samples Ordinal Regression, 84, 223
tests of model effects
in Complex Samples General Linear Model, 192
in Complex Samples Logistic Regression, 206
in Complex Samples Ordinal Regression, 218

unweighted count
in Complex Samples Crosstabs, 46
in Complex Samples Descriptives, 41
in Complex Samples Frequencies, 36

in Complex Samples Ratios, 52

warnings
in Complex Samples Ordinal Regression, 227