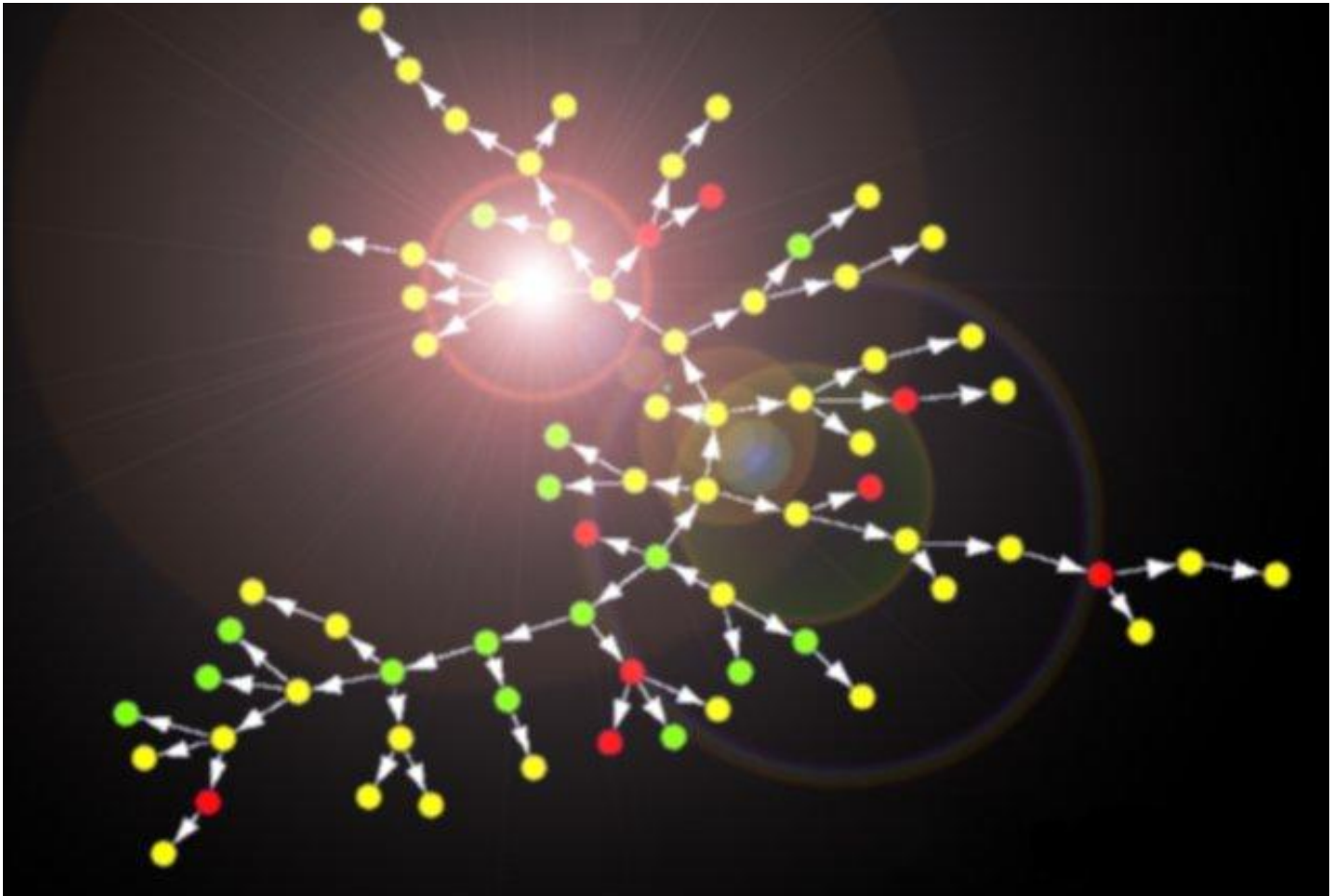


RDSAT 7.1 USER MANUAL



RDS Analysis Tool 7.1

User Manual

RDSAT 7.1 User Manual

By Michael W. Spiller, Chris Cameron, and Douglas D. Heckathorn

Last revised 25 November 2012

RDSAT 7.1 was jointly developed by Dr. Douglas Heckathorn, Michael W. Spiller, Vladimir Barash, Chris Cameron and Erik Volz of Cornell University and Ismail Degani of Degani Software with support from Cornell University. Neither Degani Software nor Cornell University make any guarantees that this software is appropriate or useful for addressing the needs of potential users. Cornell University, Degani Software and the program developers are not responsible or liable in any way for any consequences resulting from the use or misuse of the software or its documentation.

Copyright(c) 2012 Cornell University

This program may be freely used and distributed for non-commercial use. This copyright notice must appear in all copies and derivatives. The authors make no representations or warranties about the suitability of the software. The authors shall not be liable for any damages suffered by users as a result of using, modifying, or distributing this software or its derivatives.

Table of Contents

1 RDSAT 7.1 Basics	1
Installing RDSAT 7.1.....	1
Basic Layout Information	2
RDSAT Data Preparation Basics	3
Preparing Data from SAS.....	5
Preparing Data using the RDS Import Wizard.....	6
2 Loading, Viewing, and Editing Data in	
RDSAT 7.1	11
Loading Data	12
Viewing Data	13
3 Analyzing a Dataset.....	15
Analysis Overview	15
Setting Options for Analysis.....	15
Partition Analysis	19
Data Parsing Options.....	21
Breakpoint Analysis	23
4 Interpreting Analysis Results.....	26
Interpreting a Partition Analysis	26
Recruitment Tab.....	27
Estimation Tab	29
Network Sizes and Homophily Tab	33
Adjusted Average Network Sizes	33
Graphics and Histograms Tab.....	35
Interpreting a Breakpoint Analysis	42
5 Handling Missing Data in the Dataset	45
Replace Missing Data.....	45
Impute Median Values	46
Impute Degree	47
Add Field Sample Weights.....	48
6 The RDSAT 7.1 File Menu	49
RDSAT 7.1 File Menu Features	49
7 The RDSAT 7.1 Analyze Menu	55
Estimate Number of Waves Required.....	55
Estimate Prevalence	58

8 Batch Mode: Convert Files	61
9 Batch Mode: Calculate Estimates	71
Jobs and Subgroup Partitions	71
Creating a Batch in RDSAT	72
Running a Batch in RDSAT 7.1	81
Advanced Subgroup Analysis Features.....	83
10 Batch Mode: Table Builder Tool.....	89
Using the Table Builder Tool.....	90
Excluding and Combining Variable Values with the Table Builder Tool.....	94
Interacting Variables with the Table Builder Tool.....	98
Table Options in the Table Builder Tool.....	99
Table Builder Tool Output	101
Aggregating estimates across data files with the Table Builder Tool.....	104
RDS Glossary of Terms.....	105
References.....	108
Appendix 1: Frequently Asked Questions	110
Appendix 2: Graphing Recruitment	
Chains with NETDraw	112
Appendix 3: RDSAT 7.1 Performance Tuning	114

1 RDSAT 7.1 Basics

This chapter will introduce the basics of the RDS Analysis Tool 7.1. Topics covered include installing RDSAT 7.1, preparing data for RDS import, and importing the data using the RDSAT Import Wizard. SAS is a standard software package for managing data and will be described here.

Installing RDSAT 7.1

The RDS Analysis Tool is installed using a standard Windows or OS X installer application. First, download the installer to a temporary folder or your Desktop. Macintosh OS X 10.8 users may need to temporarily disable Gatekeeper by selecting “Allow applications downloaded from anywhere” in the Security and Privacy System Preferences. Re-enable Gatekeeper after RDSAT 7.1 is installed and has been opened one time.

Once the download is finished, double-click the newly downloaded installer application; the installer will guide you through the installation process. Default installation options are recommended and assumed throughout this manual.

To open the program, double click the “RDSAT” icon or (for Windows) select it from the Programs listing in the Start Menu.

Multicore Options

The RDS Analysis Tool installer will automatically configure RDSAT 7.1 to use multiple cores as long as the computer has sufficient installed RAM. If the computer RAM is upgraded after RDSAT 7.1 is installed, reinstall RDSAT 7.1 for optimal performance. See Appendix 3 for details about performance tuning options.

Basic Layout Information

RDSAT 7.1 has two modes of operation: interactive and batch modes. Interactive mode allows users to analyze one file at a time using interactive (point-and-click) menus; batch mode allows users to specify savable “jobs” and perform multiple analyses on one or more files. The tabs at the top left of the RDSAT 7.1 screen (see Figure 1.1) allow one to select which mode to use. Chapters 1-7 of this manual describe the interactive mode; chapters 8-10 describe the batch mode.

In interactive mode, all RDSAT 7.1 features are located in the right-hand side of the main screen as buttons, or in the menu bar (see Figure 1.1). The current dataset being analyzed is displayed in the selection menu beneath “RDS Data File:” When a dataset has been analyzed, all graphs and figures (output) can be found in the set of tabbed windows at the bottom of the main screen.

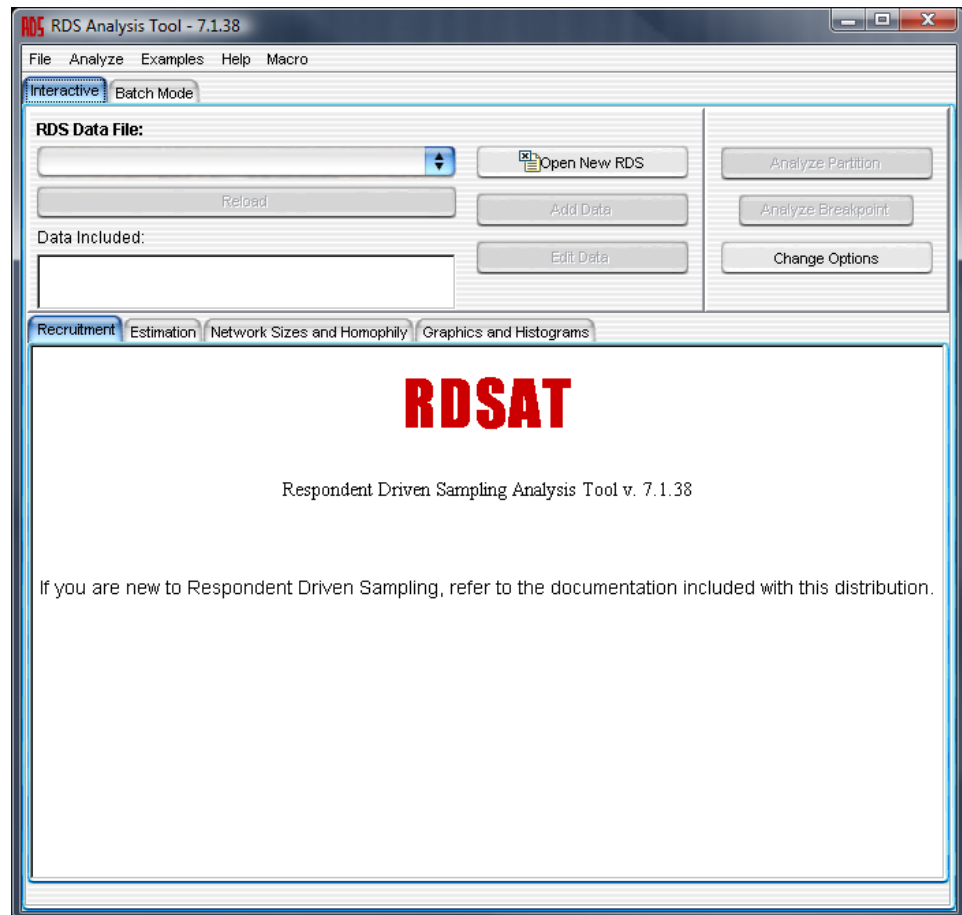


FIGURE 1.1 RDSAT 7.1 Main Window.

RDSAT Data Preparation Basics

RDSAT 7.1 uses a custom data format, so data must be imported using the import wizard, converted with the batch tool or prepared manually. This section describes manual data formatting and the following sections describe how to prepare data in SAS and import it using the import wizard. The batch file conversion tool is discussed in Chapter 8. Using the batch conversion tool is the recommended method to prepare data for use with RDSAT 7.1.

Note

Use the batch conversion tool (Discussed in Chapter 8) to prepare data for use with RDSAT 7.1. Manually formatting data is usually not necessary.

RDSAT 7.1 only analyzes RDS data files. RDS data files have three required properties. First, data must be in a tab or space-delimited text file with either the “.txt” or “.rds” suffix. Second, it must have a properly formatted header above the data, also known as the “RDS header”. Third, the RDS header must contain (at minimum) four pieces of information as detailed below. If the file conforms to these specifications it is an RDS data file.

The first two lines of the RDS data file contain the RDS header (see Figure 1.2). For the RDS header, the first line must have only the letters “RDS” (followed by an end-of-line character). This alerts RDSAT 7.1 that a file in RDS format is present. The second line must include three things: the number of respondents in the dataset, the maximum number of coupons given to a respondent to recruit others, and the value that represents missing data. After these items have been entered left to right (with a file delimiter after each), the second line continues with any additional variable names in the dataset in the same order as the data columns. Finally, the third and all subsequent rows of the RDS data file contain data (with one respondent per row).

RDS row	RDS												
Number of Respondents	3	33	-1	5	6	7	8	-1	-1	sex	agecat	race	
Number of Coupons	4	25	2	-1	-1	-1	-1	-1	-1	2	2	2	
Missing Value Code	5	50	3	17	608	607	609	18	-1	1	2	2	
	6	10	4	20	21	414	416	415	622	1	2	1	
	7	40	17	25	23	24	-1	-1	-1	1	2	2	

FIGURE 1.2 RDS HEADER

For example, the data fragment shown below says (from line 2) that the data file it is part of has 530 respondents, respondents are given a maximum of 6 coupons, the value “-1” represents missing data, and that it contains three other variables (sex, agecat, and race). Beginning on the third line, each row contains data on 1 respondent. Each column contains data on 1 variable. The first respondent in the data fragment below has the following characteristics (from line 3): his Survey ID is 3, he has a personal network size of 33, he was a seed (seeds have the missing data value for Coupon Received), he was given coupons 5, 6, 7, and 8 to recruit others, and he has values of 2 for sex, agecat, and race. Note that since the participant was given only 4 out of 6 possible coupons, the remaining two Coupon Given columns contain a missing data value.

The diagram illustrates the required variables for an RDS Data File. A table of variables is shown, with red boxes and arrows indicating the required variables for specific data types.

Respondent ID	Network Size	Coupon Received	Coupons Given
3	33	-1	5
4	25	2	-1
5	50	3	17
6	10	4	20
7	40	17	25

FIGURE 1.3 RDS Data File Required Variables

4

Note

The population size variable should be named “popsize”, and it must be constant (an identical value for every case). See the RDS file fragment below for a properly formatted RDS file with a population size of 10000.

RDS										Population Size					
										popsize	sex	agecat			
530	6	-1								10000	2	2			
3	33	0	5	6	7	8	0	0		10000	2	2			
4	25	2	0	0	0	0	0	0		10000	2	2			
5	50	3	17	608	607	609	18	0		10000	2	2			
6	10	4	20	21	414	416	415	622		10000	2	2			
7	40	17	25	23	24	0	0	0		10000	2	2			

FIGURE 1.4 RDSAT-readable data file with “popsize” variable for aggregate estimates

The following section will explain how to prepare data for RDSAT import using SAS.

Preparing Data from SAS

If the data to be analyzed are in a SAS data file, then the following steps will prepare the data to be converted into RDS format using the import wizard (see below) or the batch conversion tool (see Chapter 8).

Export the SAS data file to a flat text file using the following code fragment. The portions highlighted in blue are specific to the dataset and must be altered.

```
PROC EXPORT DATA=<libname.dataname>
  OUTFILE=<'Target Directory/RDSATdata.txt'>
  DBMS=TAB;
RUN;
```

There are three features of note in the above code. First, the output file must be a text file (suffix “.txt”). Second, the text-file delimiter is set to be tab with the “DBMS” option. Finally, the output file will contain all the variables present in the SAS data file (with variable names at the top of each column), so any variables that you do not want in the RDS data file should be removed from the SAS file before running the “proc export” code shown above. Any variable whose name or values contain spaces cannot be included in the RDS data file.

Note

RDSAT only recognizes one missing value code. Therefore, all the data values that should be treated as missing need to be converted to the same numeric value before the text file is exported.

Once the data has been exported, convert the text file to the RDS format using either the Import Wizard (see below) or the Batch File Conversion Tool (see Chapter 8).

Preparing Data using the RDS Import Wizard

The RDS Import Wizard is an interactive feature that converts delimited files (suffix “.txt”, “.csv” or “.dat”) or SAS XPORT files (suffix “.xpt”) into properly formatted .RDS files. To access the Import Wizard, open the File menu then click “Import data file...” (see Figure 1.5).

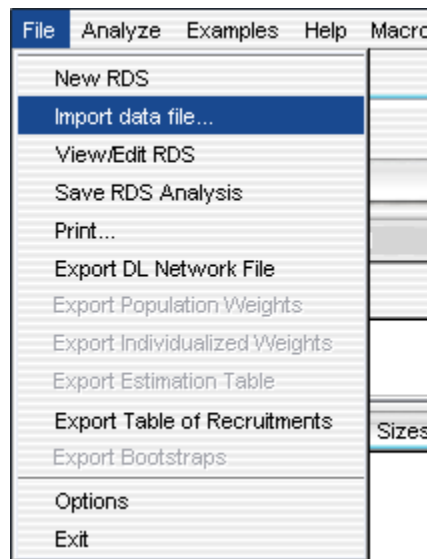


FIGURE 1.5 RDSAT 7.1 File Menu and Import data file... menu item.

When the Import Wizard has started, the front screen will appear (Figure 1.6). Locate the source data file by clicking the [Browse] button to open a standard file browser dialog. Select the source file, then click the [Next] button to continue.

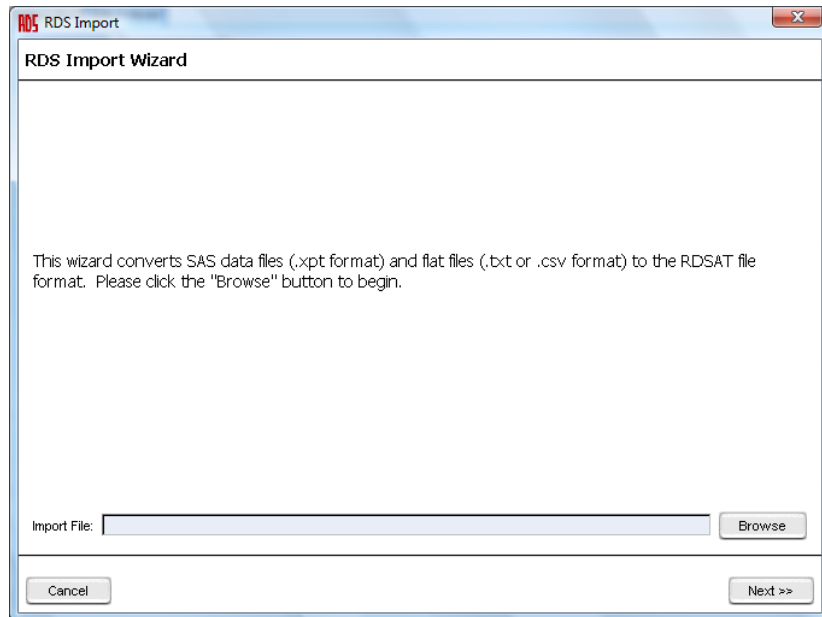


FIGURE 1.6 RDS Import Wizard front screen.

If the source file is not “.xpt” format, the Import Wizard will ask the user to specify whether the delimiter is:

- Tab
- Comma
- Space
- A custom/user-specified delimiter

After the delimiter has been specified, the Import Wizard will ask the user what the missing value code is. RDSAT 7.1 will treat the specified value as missing data, so the chosen value should not represent a valid data value for any variable in the file.

After the missing value has been specified (or immediately after the file has been chosen if it is “.xpt” format), the Import Wizard will ask the user to confirm the number of cases in the file (see Figure 1.7). This confirmation allows the user to be sure that the settings have been properly specified.

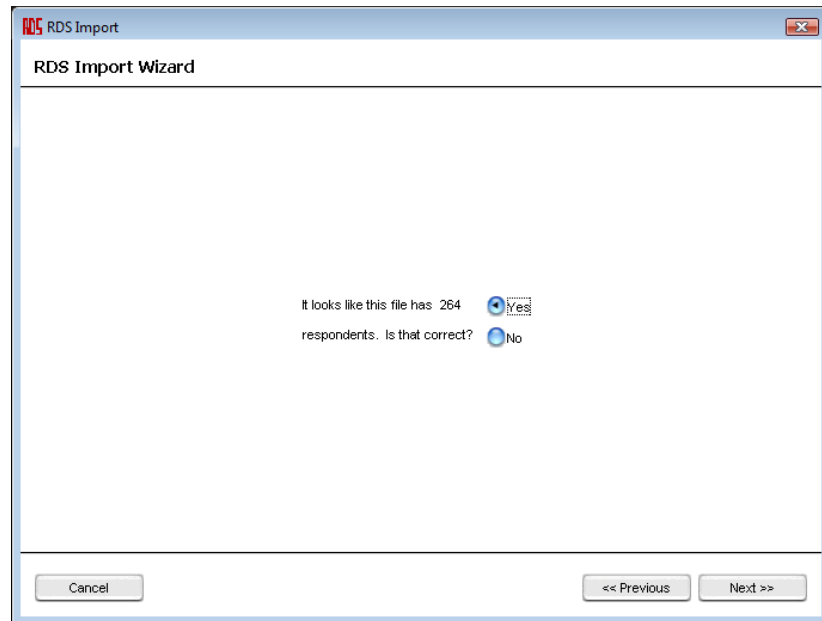


FIGURE 1.7 RDS Import Wizard number of respondents confirmation screen.

After reading the data file, the Wizard allows users to specify which variables should be included in the converted .RDS file. There is no hard limit on the number of variables that may be included in the converted file, although a file would be too large for RDSAT 7.1 to open if its size was greater than the user's computer's RAM (this case is extremely unlikely). Figure 1.8 displays the variable selection interface.

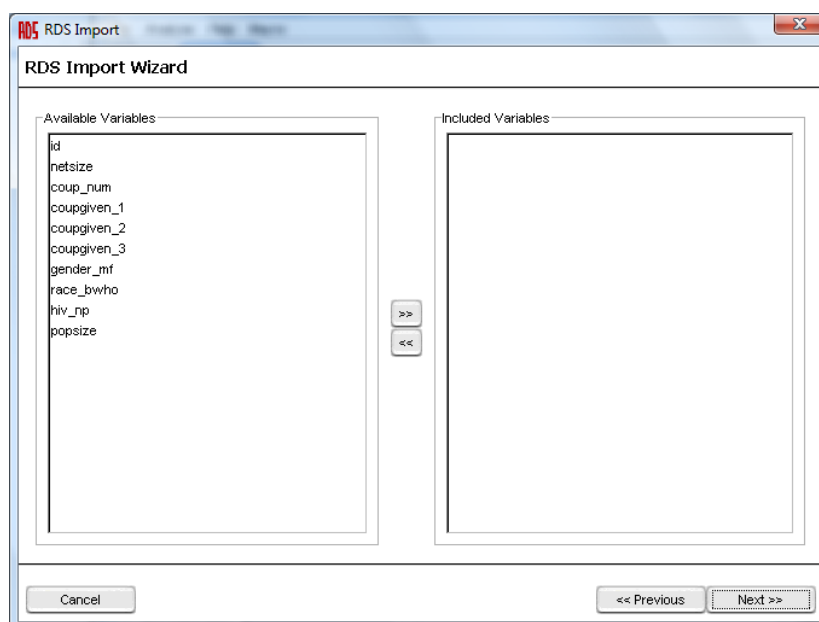


FIGURE 1.8 RDS Import Wizard variable selection screen.

Select the desired variables in the left-hand “Available Variables” pane and move them to the right-hand “Included Variables” pane by clicking the [>>] button. After the desired variables are included, click [Next] to continue.

At this stage, the user must tell the Import Wizard which variables should be assigned to the RDS header. Variables indicating the Respondent ID, respondent Network Size, respondent Coupon Received (the one with which he was recruited into the study), and the Coupons Given to a respondent to recruit others must be specified to create a RDS data file (see Figure 1.9). A final, optional, “Population Size” assignment is used to specify the variable indicating the population size associated with the file’s data; this field only needs to be specified if a user plans to aggregate estimates across files (see Chapter 9).

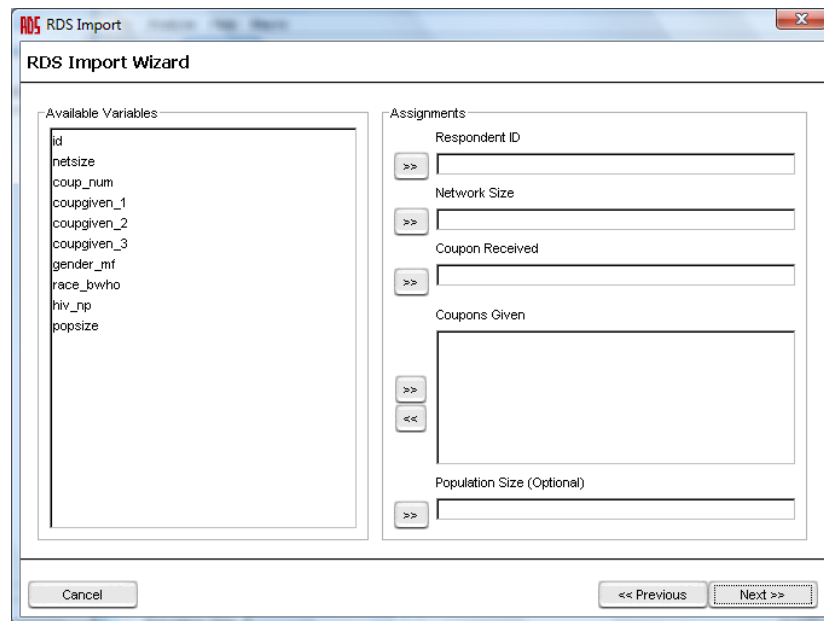


FIGURE 1.9 RDS Import Wizard header variable assignment.

After the RDS header variables have been assigned, the user specifies the output file name and save location for the RDS file that will be created.

Finally, the user clicks the [Convert] button on the final Import Wizard screen and a properly formatted RDS file is created. The new file is automatically loaded into the RDSAT 7.1 interactive mode, and users may begin analyzing the data immediately.

2 Loading, Viewing, and Editing Data in RDSAT 7.1

This chapter covers how to load, view, and edit data within RDSAT 7.1, using the Interactive Mode.

First open the .RDS formatted data file, which contains information about the sample size, missing data values, and number of coupons per respondent as well your survey data. Start RDSAT 7.1 and choose "Open New RDS" (see Figure 2.1), or select the file menu and click on "New RDS" (see Figure 1.5). When a file chooser dialog window appears, select the RDS data file and choose Open. The "nyjazz.rds" file included in this distribution is a good sample file to work with if no real dataset is available. This sample file may also be accessed through the "Load nyjazz.rds" option in the "Examples" menu.

Note

The sample RDS data set of New York jazz musicians was collected by Douglas Heckathorn and Joan Jeffri. See Heckathorn and Jeffri (2001) in references.

Loading Data

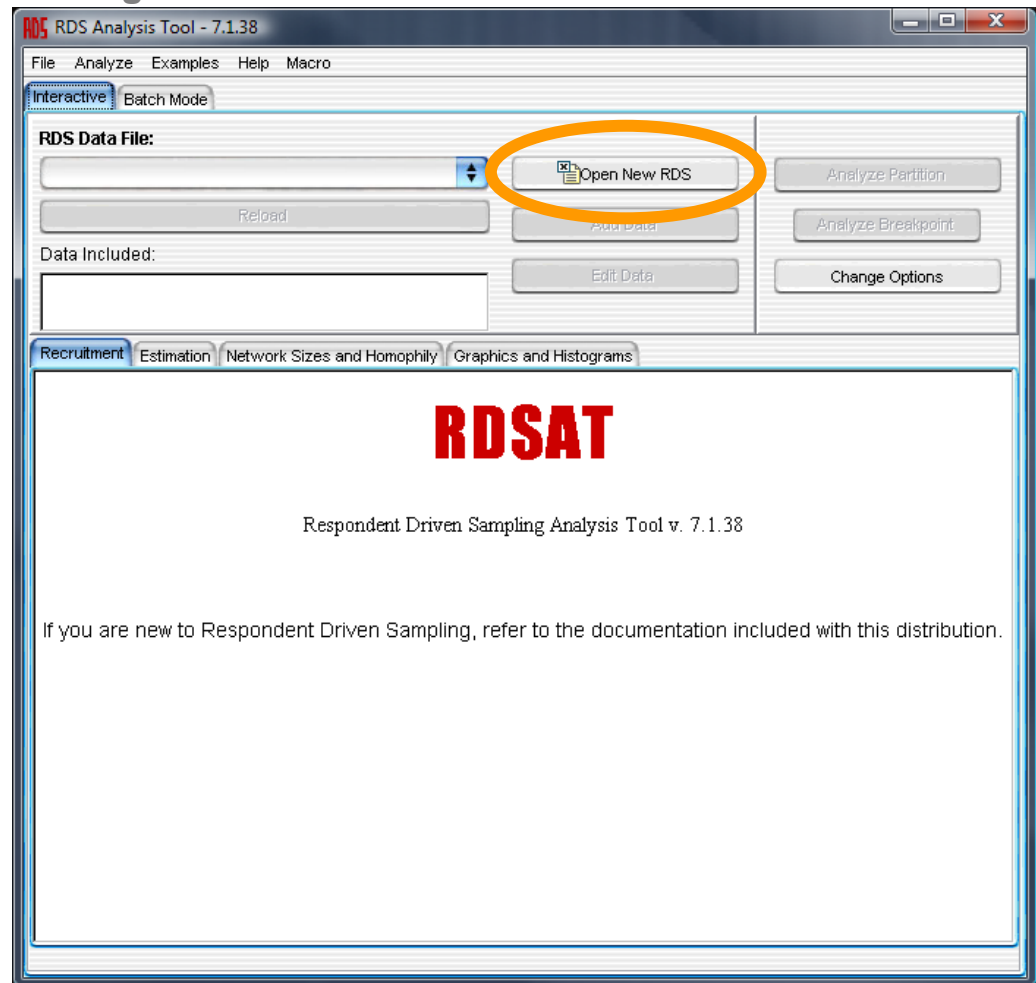


FIGURE 2.1 RDSAT 7.1 “Open New RDS” Button

Viewing Data

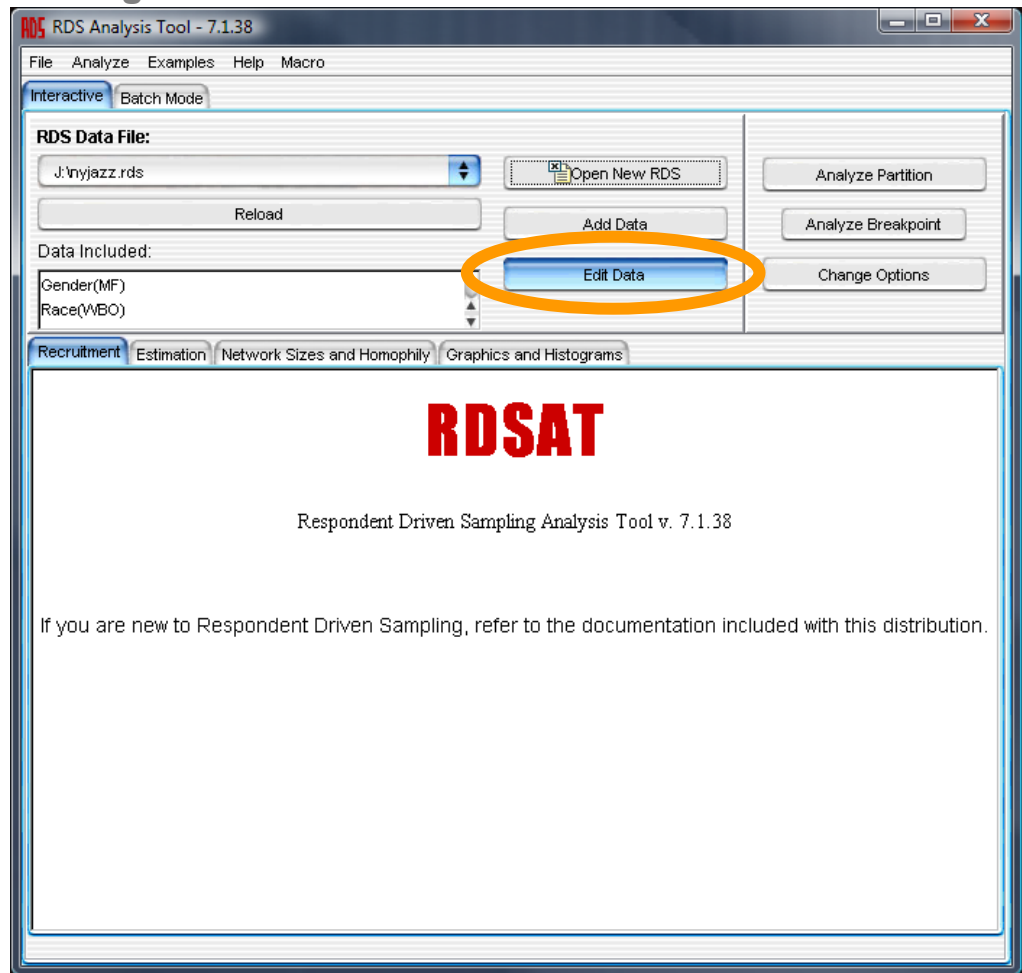


FIGURE 2.2 RDSAT 7.1 “Edit Data” Button

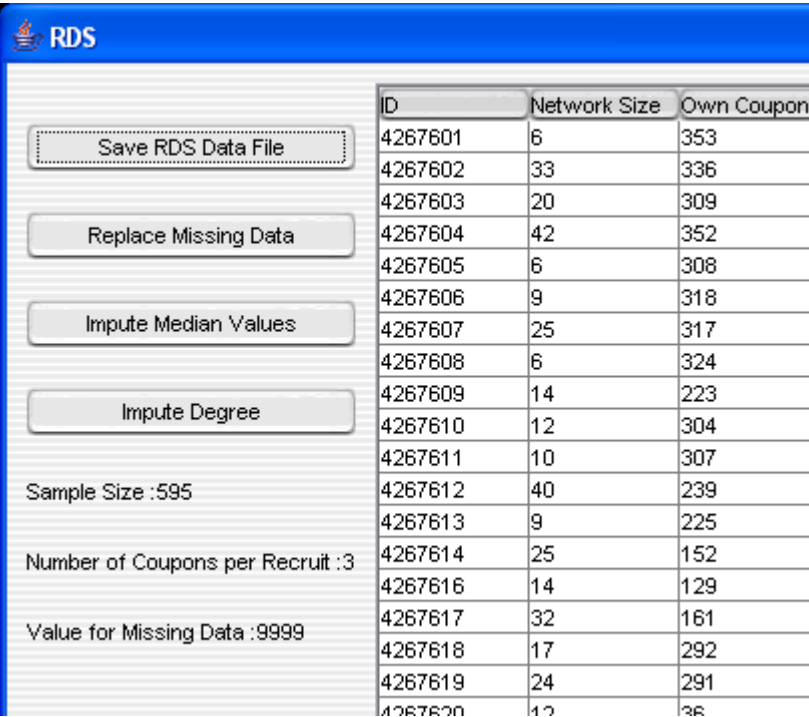
View the loaded data by clicking on the "Edit Data" Button, or select "View/Edit RDS" from the file menu. A new window will pop-up, displaying the contents of the data file you have loaded (see Figure 2.3). Sample size (595), the number of coupons per respondent (3), and the value for missing data (9999) are displayed on the left. Click and drag table columns to rearrange the column order.

Note

When a cell in the table is clicked on, its contents may be changed. The changes will be saved to any data file created with the [Save RDS Data] button. This option is NOT recommended because files created with “Save RDS Data” can only be used by RDSAT 7.1. All changes to data should be made in a statistical analysis program before opening the data with RDSAT 7.1.

Tip

Be careful not to delete or change data unintentionally when viewing data. If you mistakenly alter the data, close the Editor without saving and reload the dataset.



ID	Network Size	Own Coupon
4267601	6	353
4267602	33	336
4267603	20	309
4267604	42	352
4267605	6	308
4267606	9	318
4267607	25	317
4267608	6	324
4267609	14	223
4267610	12	304
4267611	10	307
4267612	40	239
4267613	9	225
4267614	25	152
4267616	14	129
4267617	32	161
4267618	17	292
4267619	24	291
4267620	12	336

FIGURE 2.3 RDSAT 7.1 Spreadsheet View

3 Analyzing a Dataset

This chapter introduces the analysis features of RDSAT 7.1. This is the heart of the software's functionality. This chapter provides an overview of partition and breakpoint analyses followed by detailed RDSAT 7.1 procedures for each.

Analysis Overview

Partition and breakpoint analyses were developed to handle different data types. Partition analysis was originally developed to handle categorical data and breakpoint analysis to handle continuous data. Presently, more sophisticated partition analysis techniques have extended partition analysis to both categorical and continuous variables.

A partition analysis divides the data into non-overlapping groups, or partitions, and provides estimates on those groups. A breakpoint analysis creates groups by cutting a continuous variable in two pieces at a specific variable value, or breakpoint. The value of the breakpoint changes in specified increments providing estimates for groups defined by each breakpoint. This allows the researcher to observe network structure based on a continuous variable.

Setting Options for Analysis

Before conducting an analysis, check the options that will be used. Click the [Options] button in the main window, and the window of Figure 3.1 will appear.

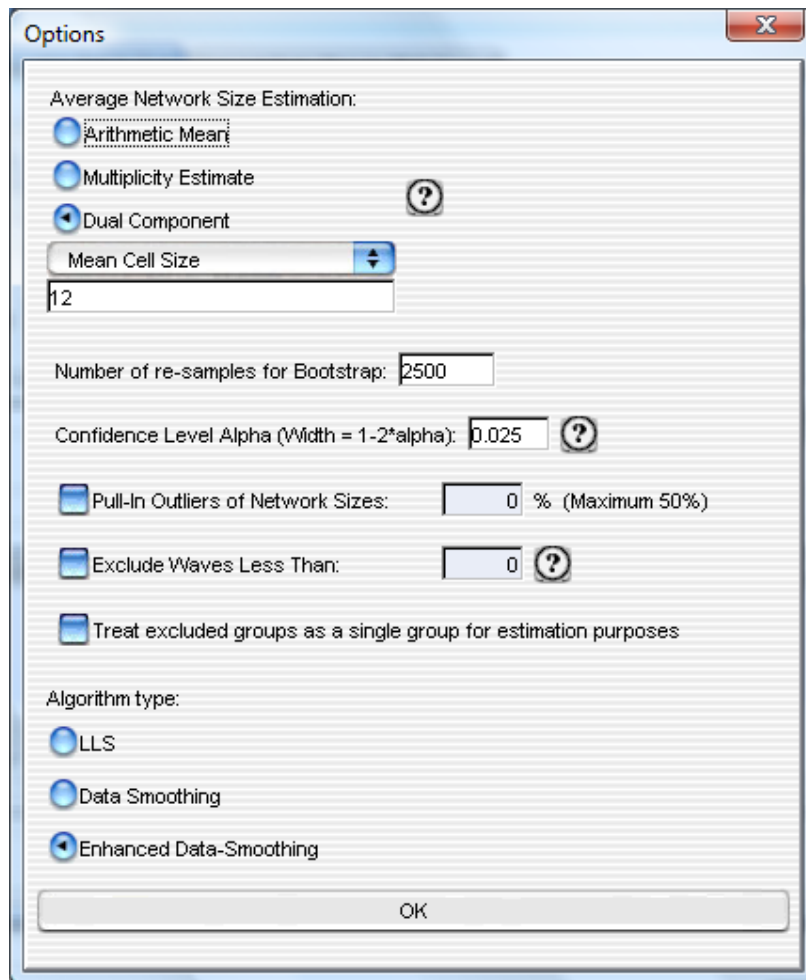


FIGURE 3.1 RDSAT 7.1 Options Window

Average Network Size Estimation

In a chain referral sample, those with more connections and larger personal network sizes tend to be over-represented in the sample. This can potentially bias sample estimates. The phenomenon can be corrected, however, using the recommended “Dual Component” estimate of average network size. To learn more about the methods used refer to Heckathorn (2007) (see “References” at the end of this manual).

Note

It is recommended to choose the “Dual Component” estimate with a mean cell size of 12. Current research indicates that this value produces the most stable estimates (see Heckathorn 2007 for details).

Number of Re-samples

This is the number of times the data is re-sampled to derive the bootstrap confidence intervals. For accurate confidence intervals, this option should be at least the default value of 2500. **For optimal accuracy (especially when estimates will be published), a number over 15,000 is recommended. Be aware, however, that the bootstrap resamples are demanding of CPU time. There may be a delay of several minutes if this value is set to a number high enough for publication-quality variance estimates.** In the RDSAT 7.1 Batch Mode (see Chapter 9), users may employ a quick estimation feature that calculates estimates without bootstrap resamples. This feature allows the analysis specifications to be examined for errors without users having to wait for many bootstrap resamples to be calculated.

Confidence Interval

The value of this parameter determines the level of confidence for the confidence intervals reported in the analysis. The default, .025, specifies a 95% confidence for the intervals reported in the analysis. The entered value is the proportion of bootstraps that are excluded from each tail of the bootstrap estimate distribution: for example, .025 indicates that 2.5% of bootstrap estimates are being excluded from each tail to create a $1 - (.025 * 2) = .95$ confidence interval.

Pull-In Outliers of Network Sizes

With this option you may eliminate extremely small and large outliers in network sizes. Check the box, and input the desired percentages of each end of the network distribution you would like to be pulled-in (for example, a value of 5% would pull-in the top 5% and bottom 5% of the network size values). If this option is selected, when the program encounters an individual whose network size is outside of the specified bounds, their network size will be set to the value of the nearest lower or upper bound (the 5th or 95th percentiles in the above example). If this feature is used, a modest value (less than 10%) is recommended.

Exclude Waves

The chain-referral process used in RDS studies allows respondents to be classified by their “recruitment wave”. A respondent’s wave is the number of recruitment links between him and the seed with which his recruitment chain began. For example, a seed would be wave 0 and a seed’s recruit would be wave 1. The “Exclude Waves Less Than” feature allows one to exclude the data collected in early recruitment waves from the RDS estimates. This feature was designed to assist methodological research and is not recommended for general use.

Note

For most estimates, the “Exclude Waves Less Than” option should be left unchecked.

Treatment of Excluded Groups

When using the Prevalence Tool (see Chapter 9 for details) or the Table Builder Tool (see Chapter 10 for details), users may specify variable values to be “excluded” from the estimates. Although they do not appear in some parts of the output, these excluded values still contribute to the estimation (this can be verified in the output).

For example, estimation of an HIV variable with “Positive”, “Negative”, and “Don’t Know” response categories would proceed as follows:

1. RDSAT 7.1 estimates a complete partition on the HIV variable including all three variable values.
2. RDSAT 7.1 calculates prevalence estimates using the HIV variable, leaving the user-specified “Excluded Values” out of the prevalence denominator.

The prevalence estimates in Step 2 are calculated automatically by the Table Builder tool when a Variable Value is excluded.

If users want a Variable Value to be treated as missing (ignored) by RDSAT 7.1, they should recode the variable value to the file’s missing value code in SAS (or their data preparation program) before analysis with the RDSAT 7.1 software.

The “Treatment of Excluded Groups” option determines how RDSAT 7.1 estimation proceeds when multiple groups are excluded. If the option box is ticked, RDSAT 7.1 will automatically recode the excluded variable values into a single group prior to the first step of estimation described above. If the option box is not ticked, RDSAT 7.1 will treat each excluded variable value as a distinct group during the first step of estimation.

This option would be desired if some of the excluded variable values have a small number of respondents (e.g., if very few respondents replied “Refuse” to the example HIV status question above and one wanted to include them in the estimation sample), in which case estimation would fail due to the small excluded groups.

Note

It is recommended that the “Treat excluded groups as a single group for estimation purposes” option be left unchecked unless some excluded groups are so small that estimation fails.

Algorithm Type

Three different algorithms are available for analyzing an RDS dataset: Linear Least Squares (LLS), Data Smoothing, and Enhanced Data Smoothing.

Note

The recommended algorithm is **Enhanced Data Smoothing**, which precludes divide by zero errors by adding a tiny, non-zero number (0.0001) to all cells in the recruitment matrix.

Partition Analysis

When an RDS dataset has been successfully loaded and options for analysis have been set, click "Analyze Partition" in the upper right of the main window (see Figure 3.2) to make the window in Figure 3.3 appear.

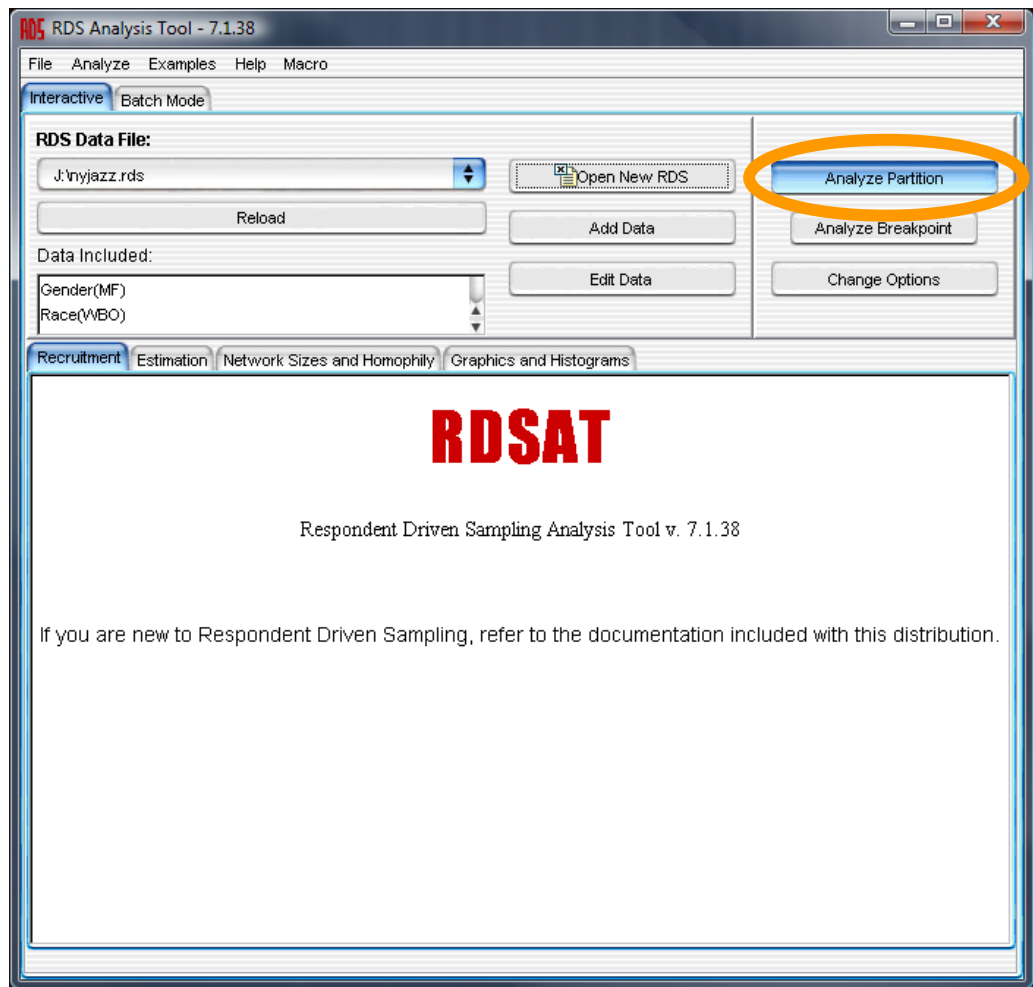


FIGURE 3.2 RDSAT 7.1 “Analyze Partition” Button

A "partition" is a user-defined set of groups. Everyone in the population belongs to a group in a partition. The groups are defined by common characteristic. For instance, a simple partition would consist of just one variable, such as gender. Those with a gender of 1 would form one group, those with gender of 2, another. A multi-trait partition of race and gender can also be created. A group would then be defined by both a gender and race value. For example, $(\text{race}, \text{gender}) = (1, 1)$ would be a separate group from $(\text{race}, \text{gender}) = (2, 1)$ although both groups have the same gender.

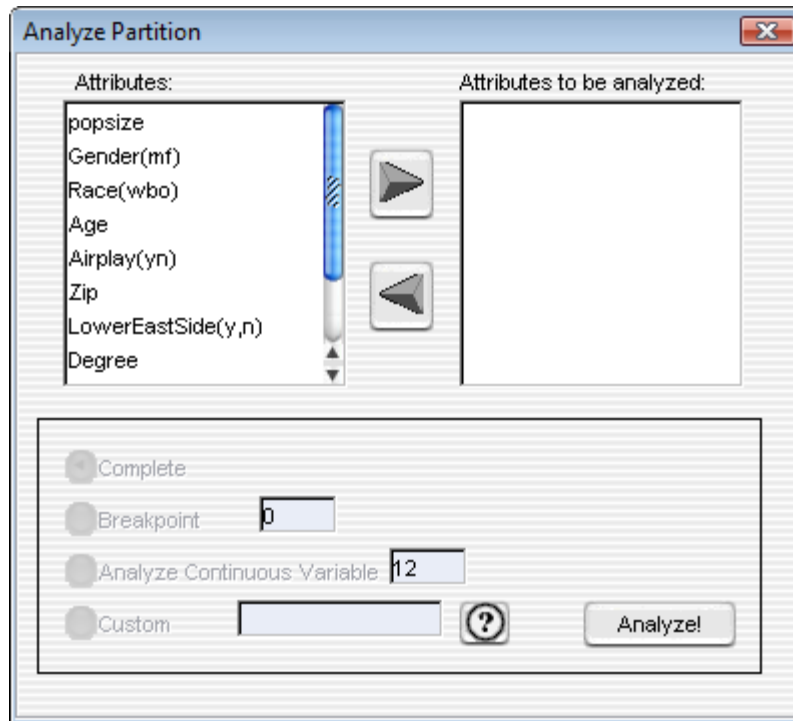


FIGURE 3.3 RDSAT 7.1 “Analyze Partition” Window

The partition panel is divided into three parts (see Figure 3.3). The top left (Attributes) contains a list of all variables that may be used for analysis. The top right (Attributes to be Analyzed) contains a list of all variables that will be used to make the partition. The bottom contains options for dividing, or parsing, the variable data.

To include a variable in the partition, select it in the left window and press the right-arrow. To remove it from the partition, select it in the right window and press the left-arrow. Data parsing options can be specified separately for each variable included in the analysis.

Data Parsing Options

Complete

This option will find every distinct value in the data file associated with that trait, and create partitions based on that value. For example, if “gender” has two values in the data file, (1, 2), the complete option will make a partition for each gender. If “race” has three values (10, 11, 12) then the complete option will create 3 partitions corresponding each race value. If both gender and race are included in the partition, there will be $2 \times 3 = 6$ partitions in all: $(\text{race}, \text{gender}) = \{(10, 1), (11, 1), (12, 1), (10, 2), (11, 2), (12, 2)\}$.

Breakpoint

For ordinal and continuous variables, this option will divide the sample into 2 groups: those respondents with a value less than the breakpoint, and those respondents with a

value greater than or equal to the breakpoint. This is different from a “breakpoint analysis” (discussed in the next section) in that only one breakpoint is chosen for the dataset, rather than a range of breakpoints. The analysis is identical to a complete partition analysis with the exception of creating exactly 2 groups from a partition in the dataset, rather than one for every possible variable value.


For example, the trait "age" has a range of values associated with it. It would be impractical to create a group for every distinct age, but by choosing breakpoint with a value of 40, the population can be divided into a group less than 40 years old and a group 40 years old and over.

Analyze Continuous Variable

This feature divides the sample into discrete groups based on the values of a continuous variable. The groups are automatically created so that the mean recruitment of the groups is approximately equal to the user-specified number (see Figure 3.3). The default is 12 because current research indicates that this value produces the most stable estimates (see Heckathorn 2007 for details). The results are interpreted in the same way as a “complete” RDS analysis of a categorical variable, except each group is defined by a range of values on the continuous variable.

Custom

This allows partitions to be specified as non-overlapping ranges of values. For instance, selecting a trait such as age and using a custom partition with parameters. “inf 20/21 30/31 40/41 inf” would create 5 groups based on 5 intervals of ages: the lowest age in the data to 20, 21 to 30, 31 to 40, and 41 to the highest age in the data (“inf” stands for the infinitely low or high value on the variable). Each range must be divided by a

forward slash, and intervals should not overlap. For more information click the  icon on the window pictured in Figure 3.3.

Breakpoint Analysis

A breakpoint analysis allows one variable to be analyzed over a range of possible values that divide the data in two groups. This is useful for analyzing the cumulative distribution of continuous variables such as age.

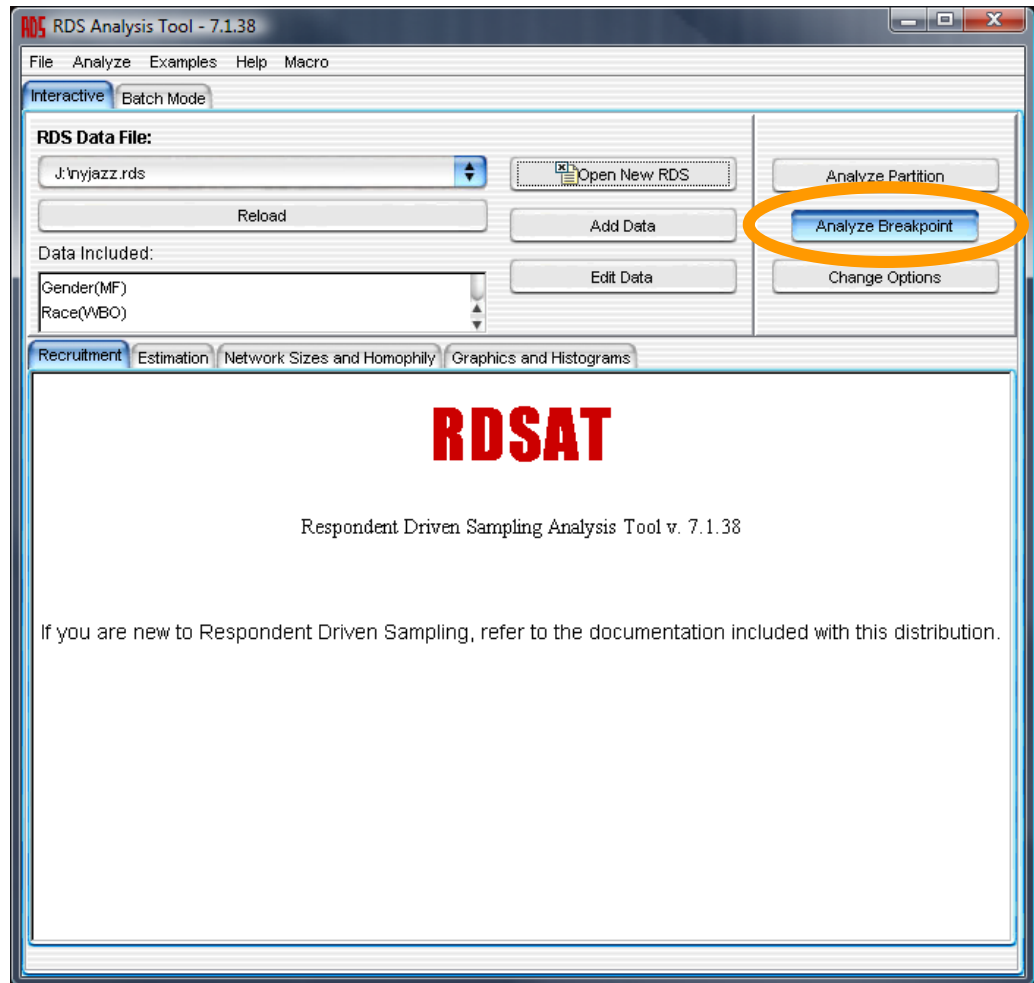


FIGURE 3.4 RDSAT 7.1 “Analyze Breakpoint” Button

To analyze a breakpoint, click on "Analyze Breakpoint" in the main window (see Figure 3.4). A Breakpoint analysis can be done on any variable, but it is more effective to use variables with many values, such as “age.”

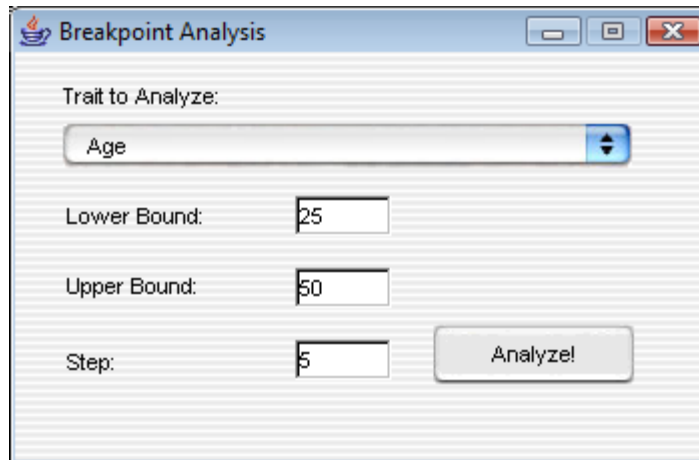


FIGURE 3.5 RDSAT 7.1 Breakpoint Analysis Window

In Figure 3.5, we are selecting “Age” as the variable to be analyzed and setting the location of the breakpoints. The “bound” fields define the range of values over which the breakpoints will be set. A “Step” of 5 with lower and upper bounds of 25 and 50 will break the dataset into the following 6 categories:

- Recruits younger than 25 versus 25 and older
- Recruits younger than 30 versus 30 and older
- Recruits younger than 35 versus 35 and older
- Recruits younger than 40 versus 40 and older
- Recruits younger than 45 versus 45 and older
- Recruits younger than 50 versus 50 and older

Likewise, a Step of 1 would produce 26 different categories, based on a breakpoint for every integer age between 25 and 50.

Note

The breakpoint analysis is performed as a series of estimates where each one divides the continuous variable into exactly two categories at a different variable value.

To run a single estimate with multiple, mutually exclusive categories for a continuous variable, users may specify the “Analyze Continuous Variable” data parsing option to have RDSAT 7.1 automatically divide the variable into categories or specify the “Custom” data parsing option to define the categories manually.

4 Interpreting Analysis Results

This chapter explains how to interpret the results of an RDSAT 7.1 analysis. The various proportion estimates are explained along with their corresponding graphs and diagrams.

Interpreting a Partition Analysis

First create a simple partition with one variable, and the “Complete” option, as shown in Figure 4.1. Click “Analyze!”.

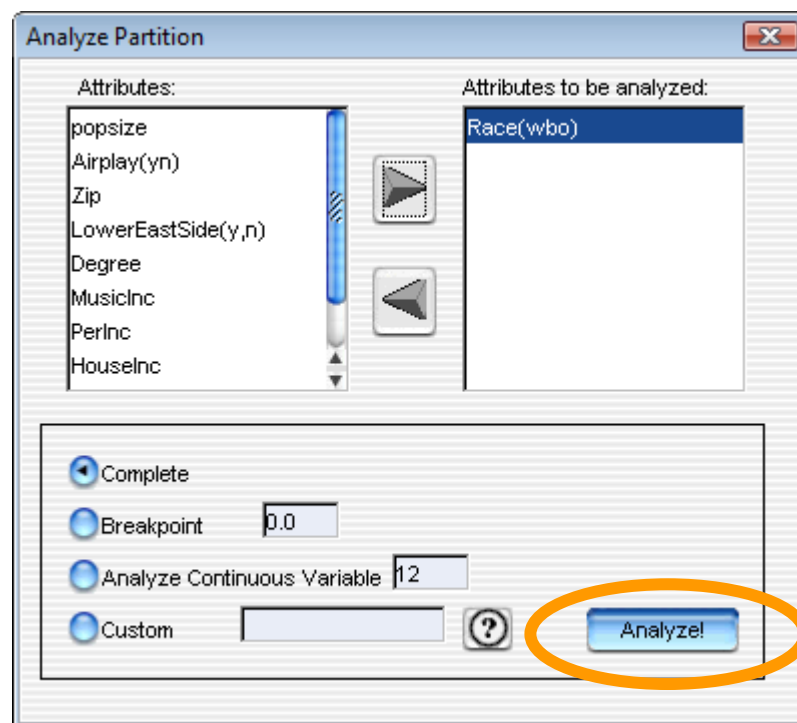


FIGURE 4.1 RDSAT 7.1 Single Variable Partition Analysis

After a moment, the results of the analysis will be output to the pages in the main window. To move between pages of the analysis, click on their corresponding tabs.

Recruitment Tab

The “Recruitment” tab displays general statistics regarding the recruitment (Figure 4.2).

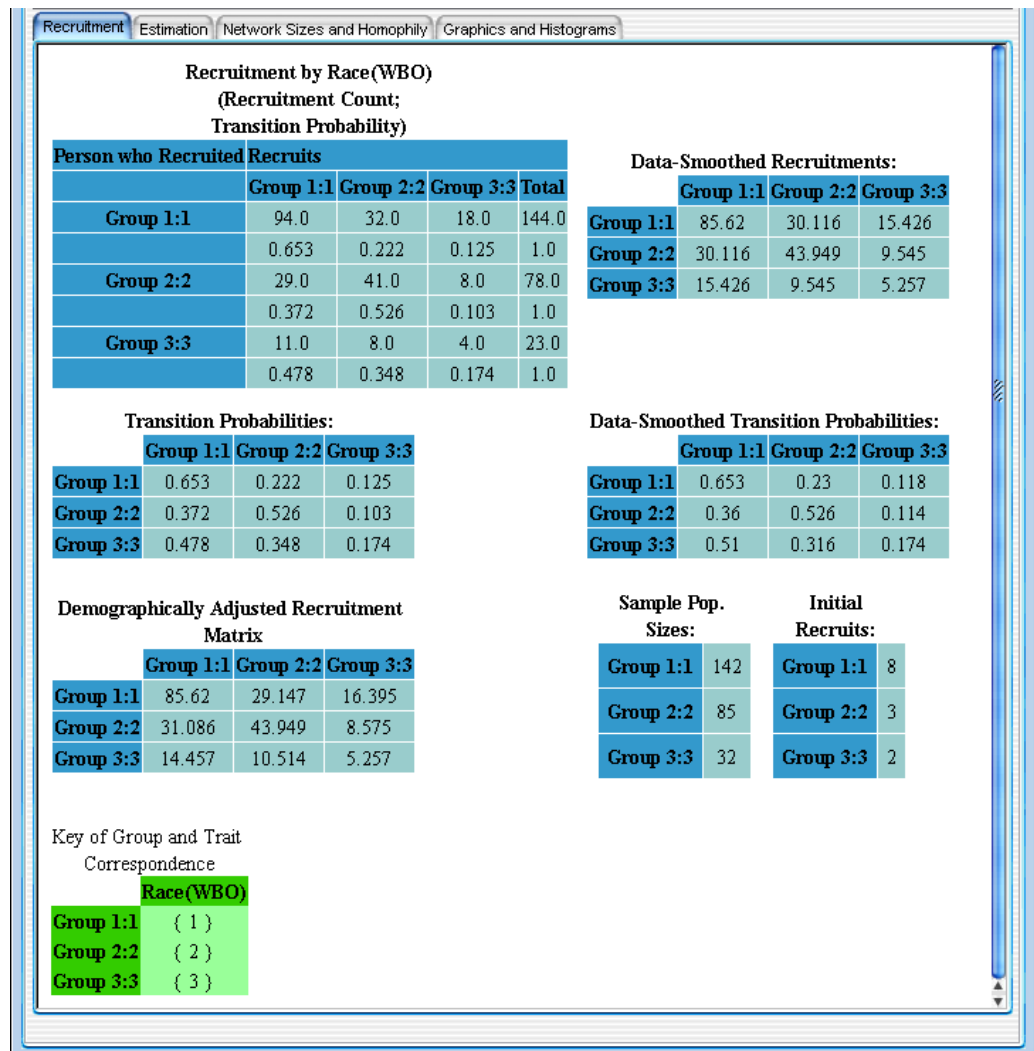


FIGURE 4.2 RDSAT 7.1 Single Variable Partition Analysis Recruitment Tab

Key of Group and Trait Correspondence

The green “Key of Group and Trait Correspondence” (at the bottom) is used to interpret the data related to recruitment in the analysis. It lists all of the various groups that were analyzed and assigns them numbers.

Recruitments

The top left “recruitment matrix” shows recruitments to and from each group. The horizontal axis (rows) depicts the recruiters and the vertical axis (columns) show recruits. For example, this matrix in Figure 4.2 tells us that Group 1 recruited 94 other people in Group 1.

Transition probabilities

The “Transition Probabilities” tab displays the probability of one group recruiting another. For example, Group 1 recruited 94 other members of Group 1 out of the total 144 recruitments made by Group 1, so the transition probability is $94 / (94 + 32 + 18) = .653$, where the denominator is the total number of recruits Group 1 made. Transition probabilities are reported in the recruitment matrix and as a separate table.

Note

Much of the data reported above also have corresponding data-smoothed estimates. Data-Smoothing is a method for eliminating deviations in cross-group recruitments that occur due to chance. For more information about data-smoothing, refer to Heckathorn (2002) in the “References” section of this manual.

Demographically-adjusted Recruitment Matrix

This option gives hypothetical recruitments if each group recruited with equal effectiveness. This is accomplished by adjusting recruitments until the number of recruitments by Group A (row sum in recruitment matrix) equals the number of times Group A was recruited (column sum in recruitment matrix). Similar to data-smoothing (see note above), demographic adjustment of recruitment is a way of eliminating deviations in recruitments that occur due to differential recruitment efficiency across groups.

Note

All RDS estimates that use the “Data Smoothing” or “Enhanced Data Smoothing” algorithms automatically incorporate Demographic Adjustment of the recruitment matrix.

Sample population sizes

Reports the total number of sample members in each group.

Initial Recruits

Reports the number of "seeds" from each group (i.e. people recruited by the researcher in each group).

Estimation Tab

The Estimation tab displays estimates of population proportions and their confidence intervals, which are the target estimates for most users (Figure 4.3). Along with these estimates, users should report adjusted average network sizes and the options associated with an estimate. See the “References” section at the end of this manual for examples of how these analyses are reported in published journal articles.

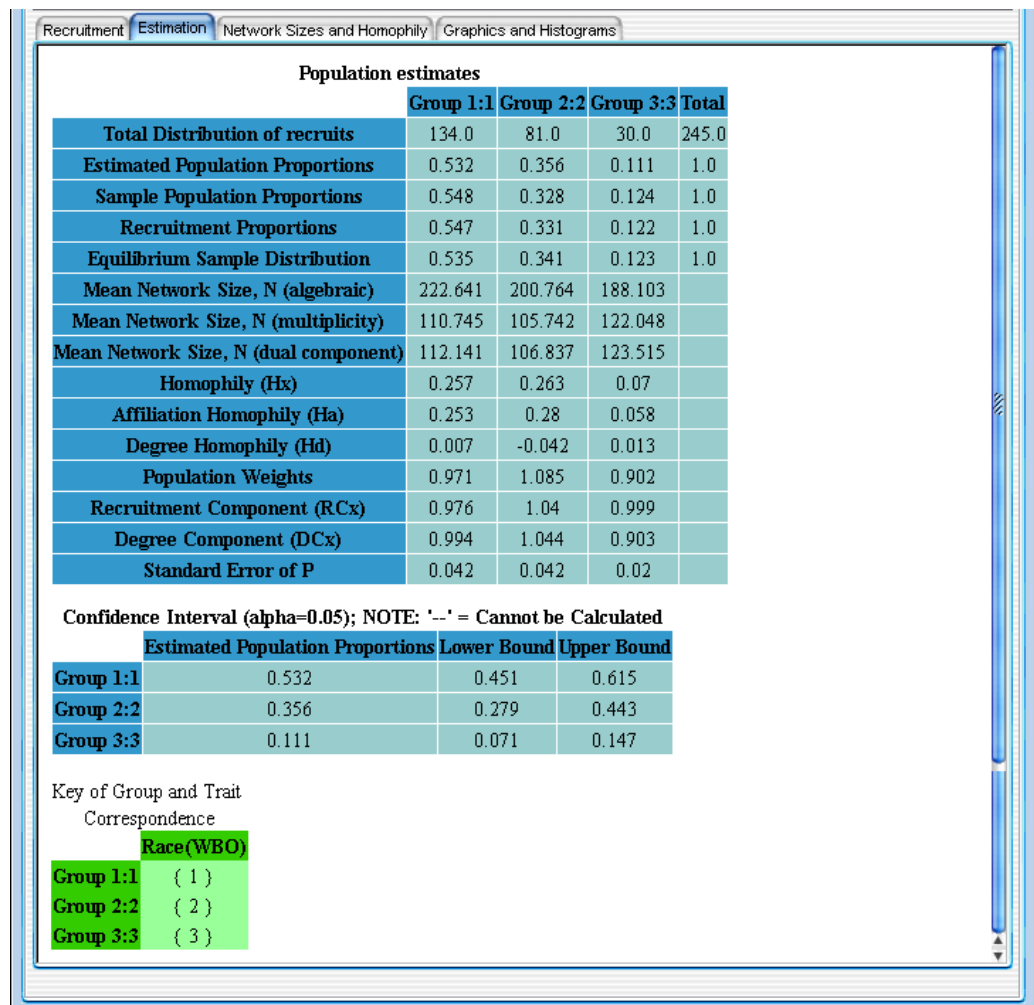


FIGURE 4.3 RDSAT 7.1 Single Variable Partition Analysis Estimation Tab

Total Distribution of Recruits

Displays the raw count of recruits in the data set for each group, which correspond to the column sums of the raw recruitment matrix. The Total is the sample size minus the number of seeds.

Estimated Population Proportions

Estimated Population Proportions are the RDS estimates of the population proportion of each group. This is the RDS estimator of primary interest for most users.

The estimated population proportion can either be calculated using the linear least squares algorithm, or the data-smoothing/enhanced data-smoothing algorithm, depending on how the options are set for the RDS analysis. In the above diagram, the enhanced data smoothing algorithm was used. See the “Algorithm Type” section of Chapter 3 for more information on the difference between various estimation algorithms in RDSAT 7.1 .

Sample Population Proportions

The sample population proportions are also called the "naïve" estimates of population proportions. The term naïve is used because the proportion is a simple ratio of how many members of a particular group were recruited to the total number of recruits. It is not adjusted for any statistical biases. (To learn more about the methods used refer to Salganik and Heckathorn 2004 and Heckathorn 2007).

Recruitment Proportions

The unadjusted recruitment proportions for the sample are the number of times members of group A were recruited divided by the total number of recruitments.

Equilibrium Sample Distribution

The equilibrium sample distribution indicates each group’s population proportion based only on the equilibrium distribution of that variable. These values are reported for diagnostic purposes; please see discussion of equilibrium and related concepts in the papers cited in the “References” section at the end of this manual.

Mean Network Size, N (algebraic)

This is the arithmetic mean of the sample’s network sizes.

Mean Network Size, N (multiplicity)

Network sizes are adjusted for over-sampling of high network respondents. In a chain referral sample, those with more connections and larger personal network sizes tend to be over-represented in the sample. (To learn more about the methods used refer to Salganik and Heckathorn 2004).

Mean Network Size, N (dual component)

Network sizes are adjusted for over-sampling of high network respondents and differential recruitment by network size. This is the recommended average network size estimator. (To learn more about the methods used refer to Heckathorn 2007).

Note

The “Dual-Component” mean network size estimator is preferred both for estimation and reporting.

Homophily (Hx)

Homophily is a measure of preference for connections to one's own group. Varies between -1 (completely heterophilous) and +1 (completely homophilous). For example, if males recruited exclusively other males, they would exhibit complete homophily.

Affiliation Homophily (Ha)

Affiliation homophily is a homophily measure based on the equilibrium proportions. It provides a measure of homophily which is not affected by differential network sizes across groups.

Degree Homophily (Hd)

Degree homophily is a measure of the level of homophily that is attributable to differential network size across groups.

Population Weights

The population weight is the multiplier that produces the RDS estimator. It provides a measure of bias accounted for with the RDS estimator. The weights are calculated as follows:

$$\text{population weight} = \frac{\text{estimated population proportion}}{\text{sample population proportion}}$$

Population weights can either be calculated using the linear least squares algorithm, or the data-smoothing/enhanced data-smoothing algorithm, depending on how the options are set for the RDS analysis. In Figure 4.3, the enhanced data smoothing algorithm was used. See the “Algorithm Type” section of Chapter 2 for more information on the difference between various estimation algorithms in RDSAT 7.1.

Recruitment Component (RCx)

The recruitment component of the population weight (refer to Heckathorn 2007 for a discussion).

$$\text{population weight} = (RCx) * (DCx)$$

Degree Component (DC_x)

The degree component of the population weight (refer to Heckathorn 2007 for a discussion).

$$population\ weight = (RC_x) * (DC_x)$$

Standard Error of P

The estimated standard error of the estimated population proportion, P_x, based on the results of the RDS bootstrapping algorithm.

Confidence Intervals

Confidence intervals are obtained by bootstrapping the original sample. The confidence intervals correspond to population proportion estimates calculated by the chosen estimation algorithm.

Network Sizes and Homophily Tab

This tab displays Homophily, Affiliation, and Average Network Sizes (Figure 4.4).

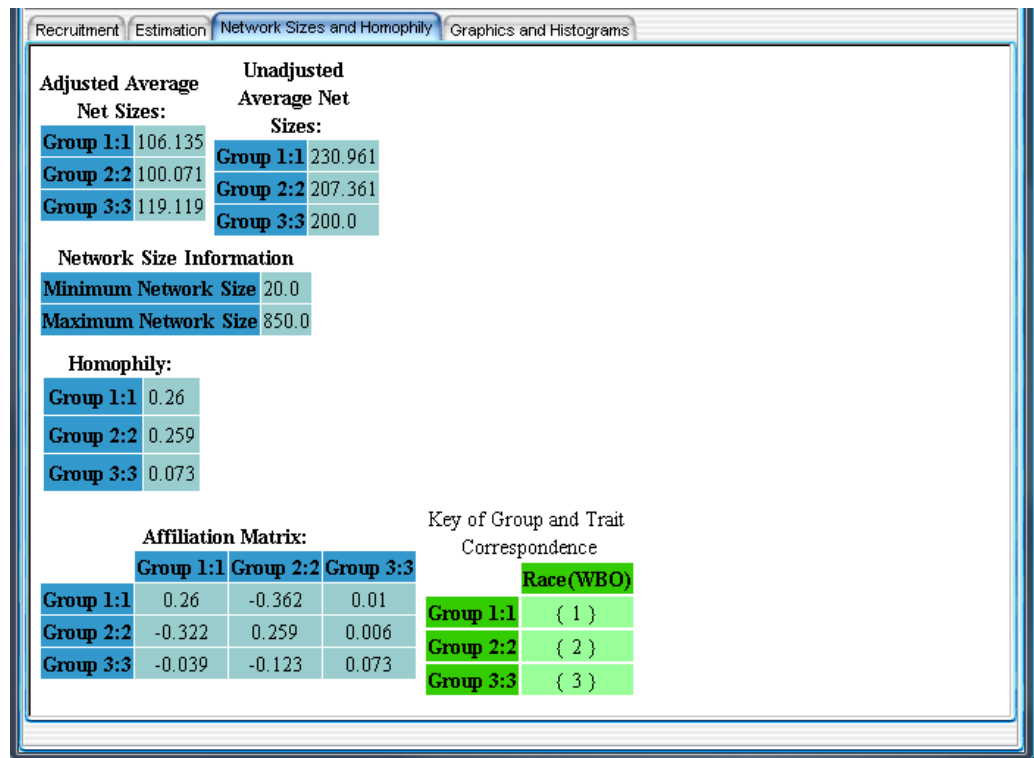


FIGURE 4.4 RDSAT 7.1 Single Variable Partition Analysis Network Sizes Tab

Adjusted Average Network Sizes

These are the same as the mean network size estimates in the previous tab for the chosen mean network size estimator (i.e, if you chose the “Dual Component” mean network size estimator, these values are the dual component estimates). These are the network size estimates used for the estimator, so they are the ones that should be reported (they are also displayed in the Estimation Tab).

Unadjusted Network Sizes

These are the same as “Mean network size, N (algebraic)” above. They are straightforward arithmetic means of the sample’s network sizes.

Network Size Information

Displays the minimum and maximum network sizes for the sample.

Homophily

Homophily is a measure of preference for connections to one's own group. Varies between -1 (completely heterophilous) and +1 (completely homophilous). For example, if HIV-positive respondents recruited no other HIV-positive respondents they would exhibit complete heterophily.

Affiliation Matrix

The affiliation matrix contains a measure of preference for connections to any group in the network. Varies between -1 (complete avoidance) and +1 (complete preference). Affiliation is a more general version of homophily. For example, if black respondents recruited exclusively white respondents they would exhibit complete preference (+1) for white respondents.

Graphics and Histograms Tab

This tab displays visual illustrations of data presented in the previous sections of this chapter.

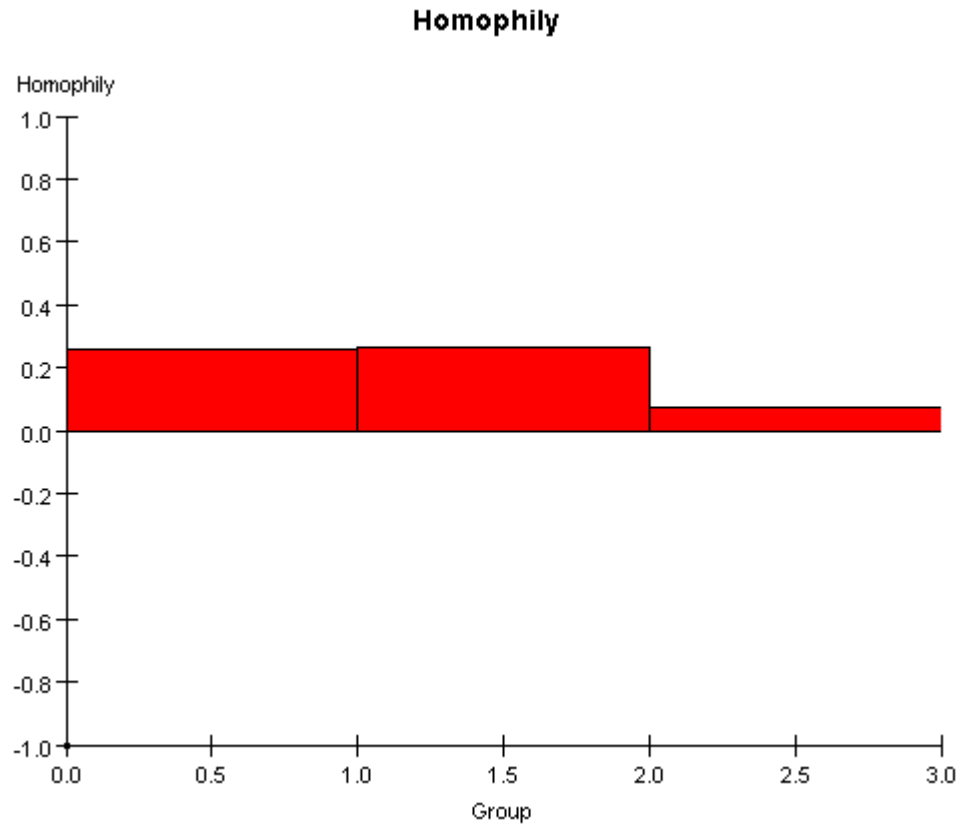


FIGURE 4.5 Homophily Chart on Graphics Tab

This graph displays homophily within 3 different analysis groups. Each group is shown as a separate bar. This graph illustrates that Group 2 (the middle bar) has the highest homophily (roughly .3), followed by Group 1 (the leftmost bar) and Group 3 (rightmost).

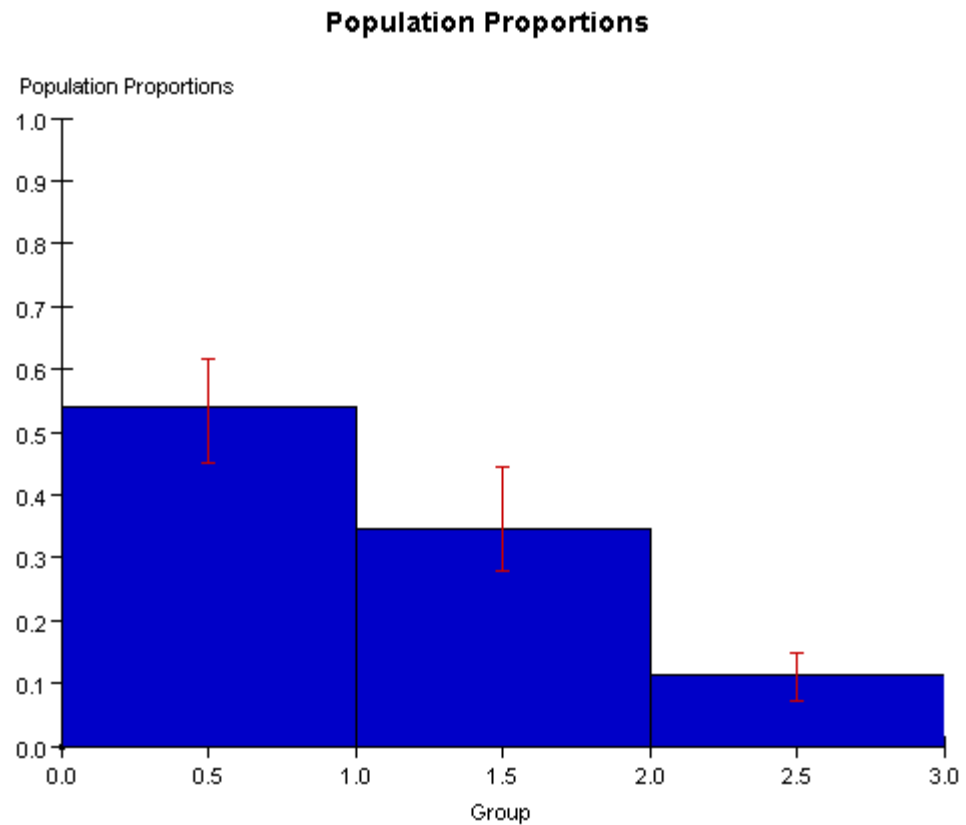


FIGURE 4.6 Population Proportions Chart on Graphics Tab

This graph displays the population proportions of each group. The y-axis is the population proportion, and should be read as a percentage. We see that Group 1, (the leftmost bar) comprises more than half the total population, followed by group 2 and 3.

The red “whisker” bars represent the values of the estimate’s confidence interval.

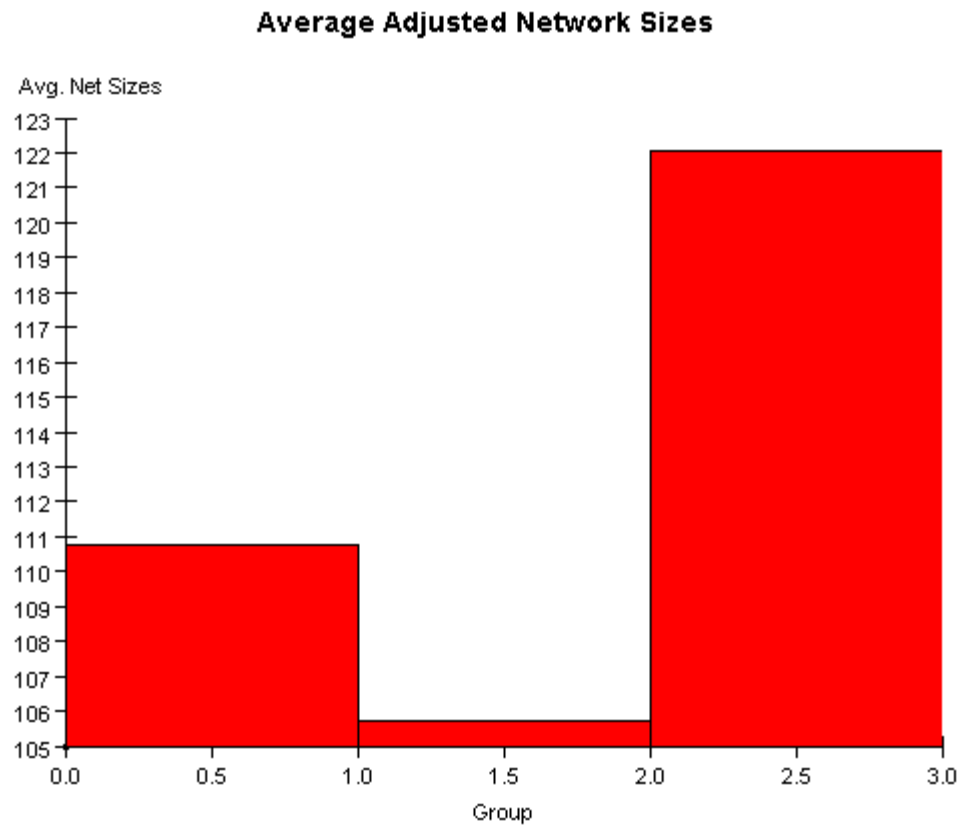


FIGURE 4.7 Adjusted Network Sizes Chart on Graphics Tab

This graph displays the adjusted network sizes of each group. Observe that group 3 (the rightmost bar) has the highest average network size.

Transition Probabilities

This is a 2 dimensional histogram of the transition probabilities. A brighter color corresponds to a higher value. It is a method of visualizing the corresponding transition matrix.

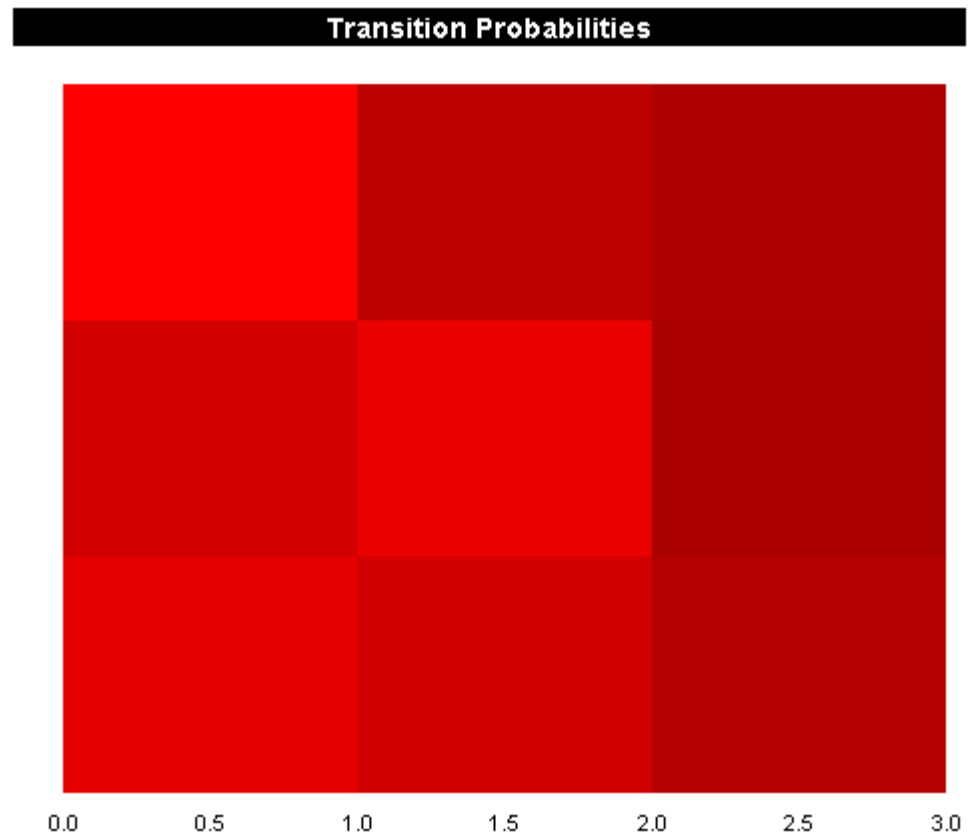


FIGURE 4.8 Transition Probabilities Matrix Visualization on Graphics Tab

Degree List

List of all network sizes (degrees) reported in the sample. The list is sorted from least to greatest for easy view of the distribution.

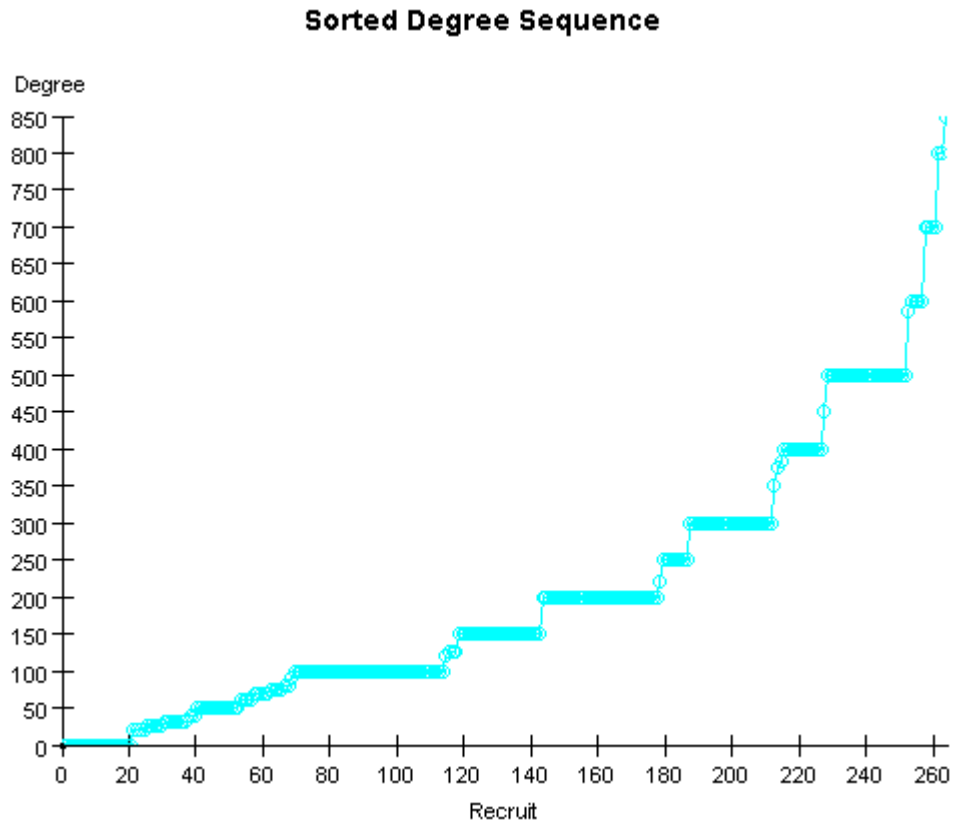


FIGURE 4.9 Degree Sequence Plot on Graphics Tab

In the graph above we see that there are a few respondents with networks as large as 800, but most respondents fall within a degree of 100-300.

Bootstrap Simulation Results

Shows the histogram of Bootstrap estimates of population proportions. The horizontal axis depicts population estimates for the specified group. The vertical axis shows the frequency of bootstrap estimates for the corresponding proportion.

Frequency of Population Proportions from Bootstrap Procedure

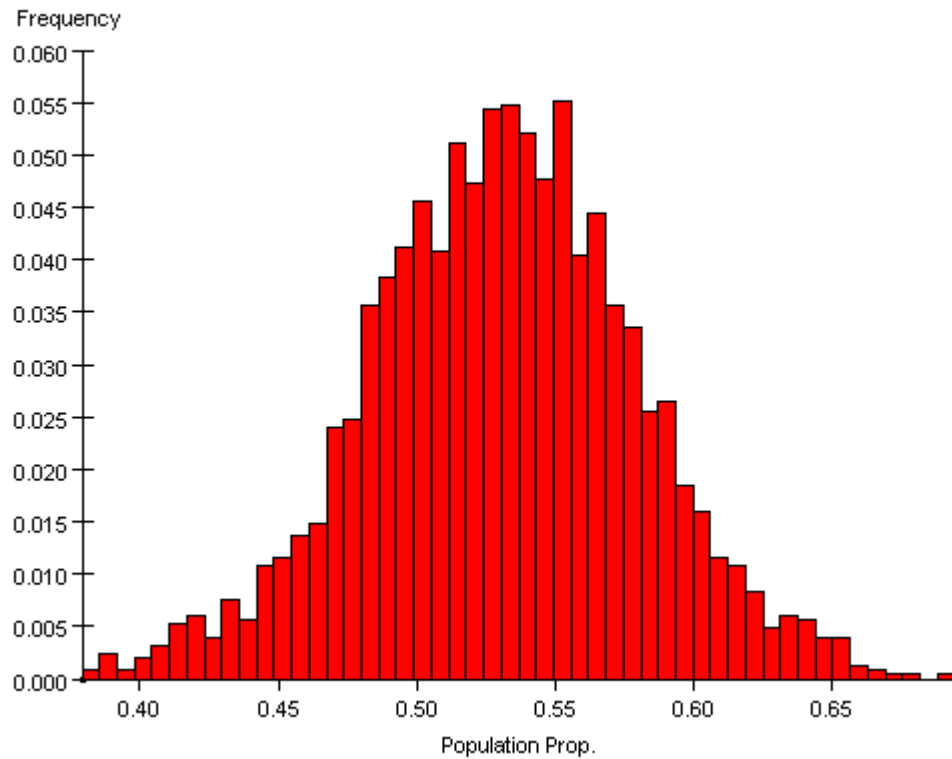


FIGURE 4.10 Bootstrap Results Histogram on Graphics Tab

Degree Distributions

Distribution of network sizes for each group and for the population as a whole. The diagram below is of the entire population. We see that most members of the population have network sizes close to 100 or 200, and the frequency of higher network sizes decreases with the exception of an anomaly at 500.

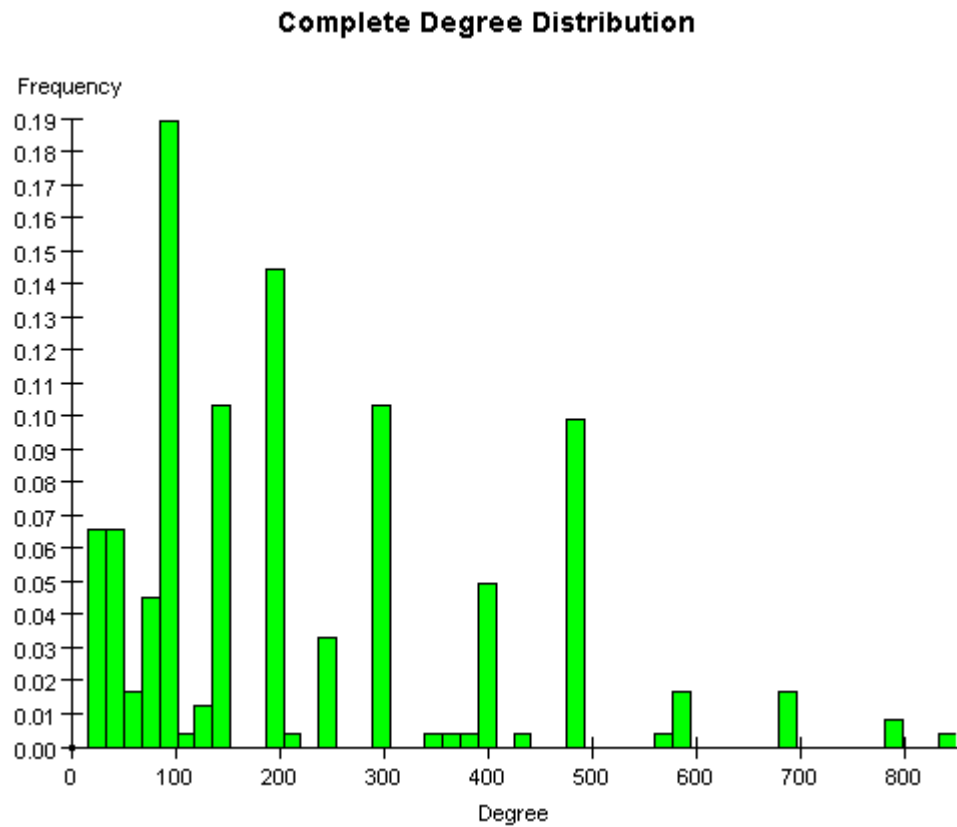


FIGURE 4.11 Degree Histogram on Graphics Tab

Interpreting a Breakpoint Analysis

A breakpoint analysis divides a dataset into groups based on a single continuous or ordinal variable. A variable of interest might be “Age,” where one wouldn’t examine each individual age as a separate group, but rather a range of ages. There is no recruitment data for breakpoint analyses. Rather there are interesting trends to notice in homophily and population proportions as the breakpoint is shifted and respondents are moved from the upper group of the lower group. The **Estimation** tab shows a table of Least Squares population estimates corresponding to each breakpoint value. Similarly, the **Network Sizes** and **Homophily** tables are arranged by breakpoint value (see Figure 4.12).



	30.0	35.0	40.0	45.0
Group Below Breakpoint	0.244	0.337	0.458	0.509
Group Above Breakpoint	0.755	0.662	0.541	0.490

FIGURE 4.12 RDSAT 7.1 Breakpoint Analysis Estimation Tab

Viewing the data in the graphics tab will often make patterns very clear. For example, in the example breakpoint analysis at the end of Chapter 3, New York Jazz musicians were analyzed based on their age; the 26-group analysis is shown in Figure 4.13.

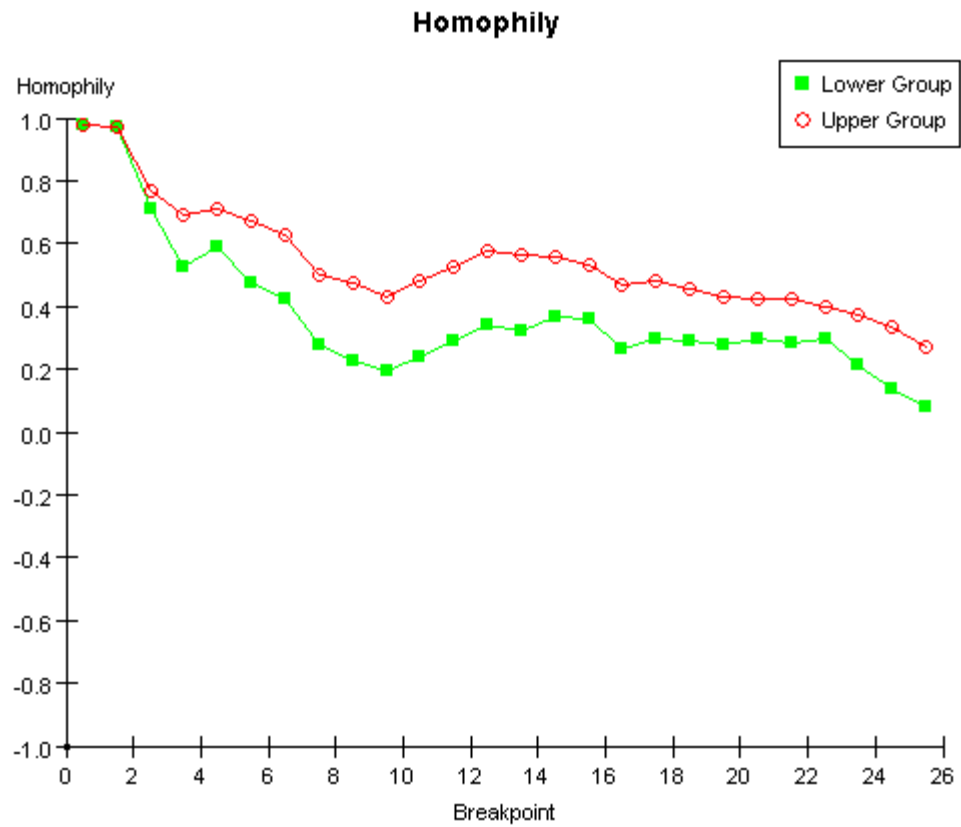


FIGURE 4.13 Homophily at different breakpoints among Jazz musicians

There are several visible patterns: Homophily tends to zero as the age variable increases. This implies that differences in age become less important for choosing relationships the older the recruits are. It is also notable at all breakpoints that the older group is more homophilous than the younger group. Finally, it is possible to see that homophily is strongest where age is the lowest (25). This implies that young jazz musicians show strong preference for relationships with other young jazz musicians.

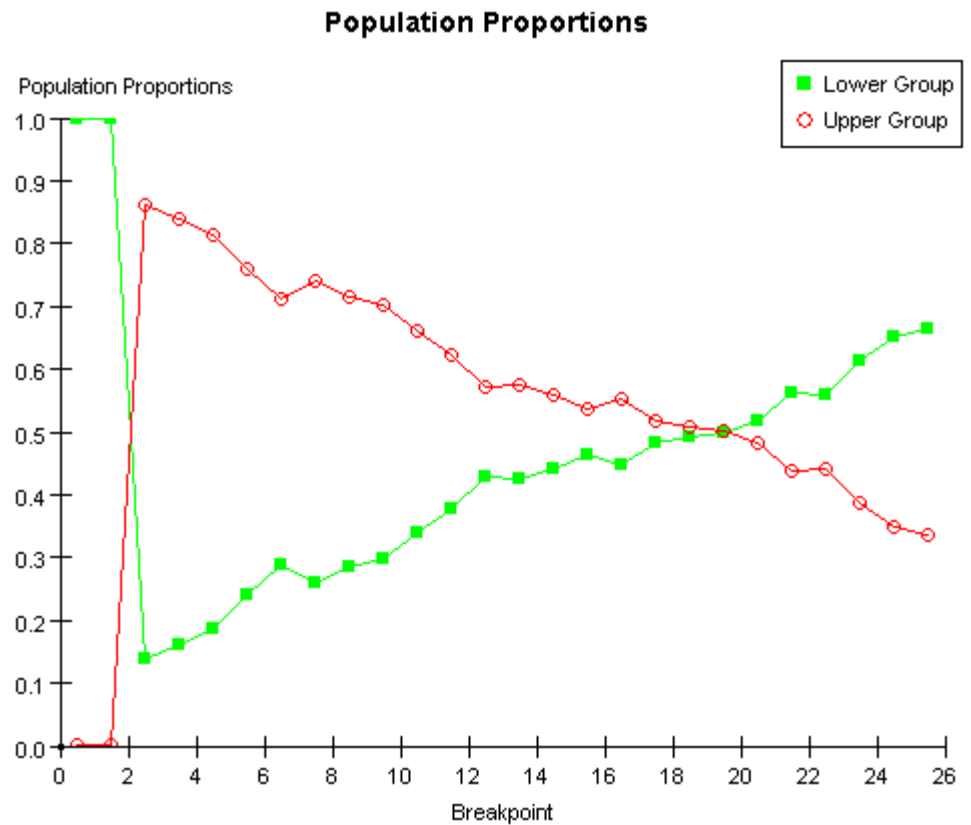


FIGURE 4.14 Population proportions at different breakpoints

Figure 4.14 shows the breakpoint where the population of the upper group equals that of the lower group. From this it can be inferred that half of the musicians are less than ~44 years old. Note that although the graph's x-axis ranges from 0 to 26, we are conducting a breakpoint analysis on groups age 25 to 50. Therefore the above intersection corresponds to an age of 44 ($19+25$), not 19.

5 Handling Missing Data in the Dataset

Most datasets contain missing data. RDSAT 7.1 offers two ways of handling missing data. Both of these options will be covered in this chapter.

RDSAT 7.1 employs two features to handle missing data. The first makes it possible to reassign another value to missing data. In this way, respondents for whom data is missing can be included in the analysis as a separate category. The other procedure imputes missing values at the median of the variable. These features are located in the Edit Data screen.

Note

Replacing and imputing data is not recommended. The proper coding of missing data should be handled in statistical analysis software prior to analysis using RDSAT 7.1.

Replace Missing Data

This feature replaces all missing data cells with a user-specified value. First, click “Replace Missing Data” on the left side of the Edit Data screen. Select the variable you want to replace values in, enter the new value for missing data, and click “Commit Changes.” To make the changes permanent, click “Save RDS Data File.” (see Figure 5.1).

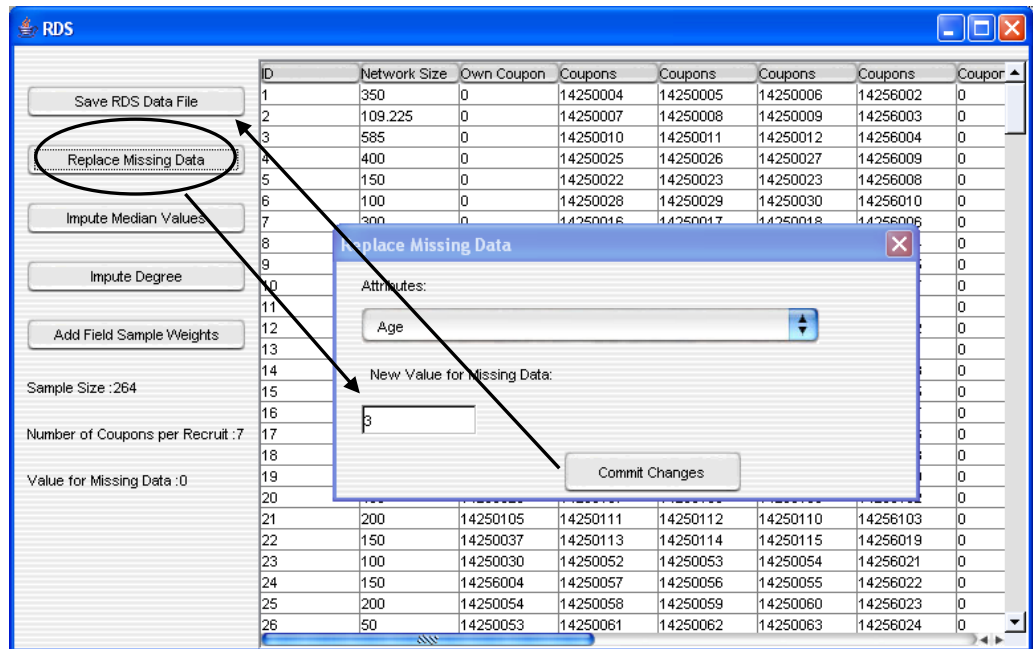


FIGURE 5.1 RDSAT 7.1 Replace Missing Data

Impute Median Values

This feature calculates the median value of the variable being analyzed and replaces all missing data cells with this median value. First, click “Impute Median Values” on the left side of the Edit Data screen. Select the variable you want to replace values in and click “Commit Changes.” To make the changes permanent, click “Save RDS Data File.” See Figure 5.2.

Note

Make sure the median value of a variable is reasonable before using Median value imputation. Median value imputation is only useful for continuous variables and ordinal/sequential categorical variables. For example, median value imputation is valid for variables such as “age” or “level of education.” For a categorical variable, such as gender, imputation would produce a nonsensical value that is half way between “male” and “female”.

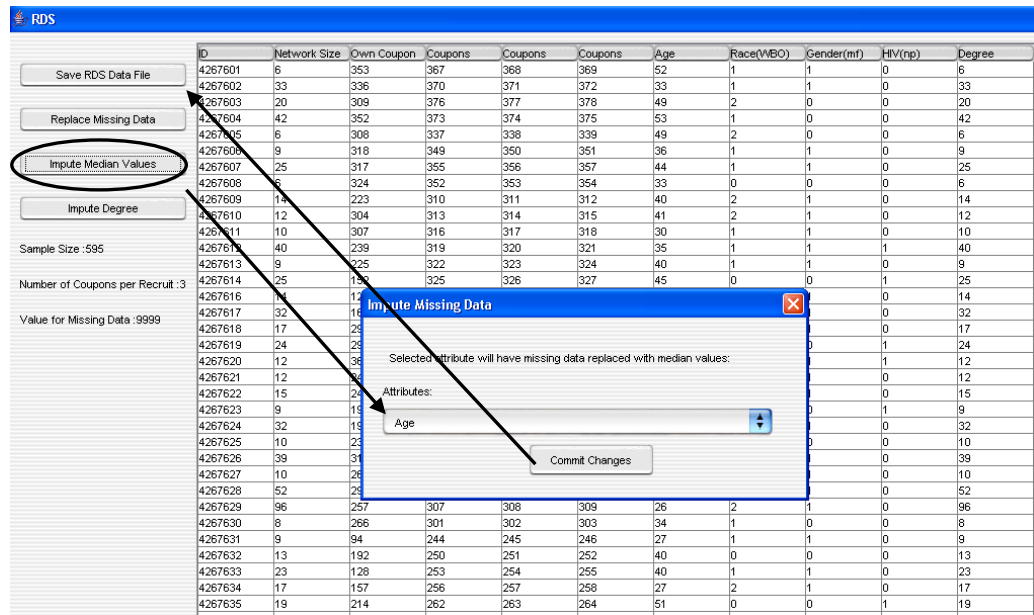


FIGURE 5.2 RDSAT 7.1Impute Median Values

Impute Degree

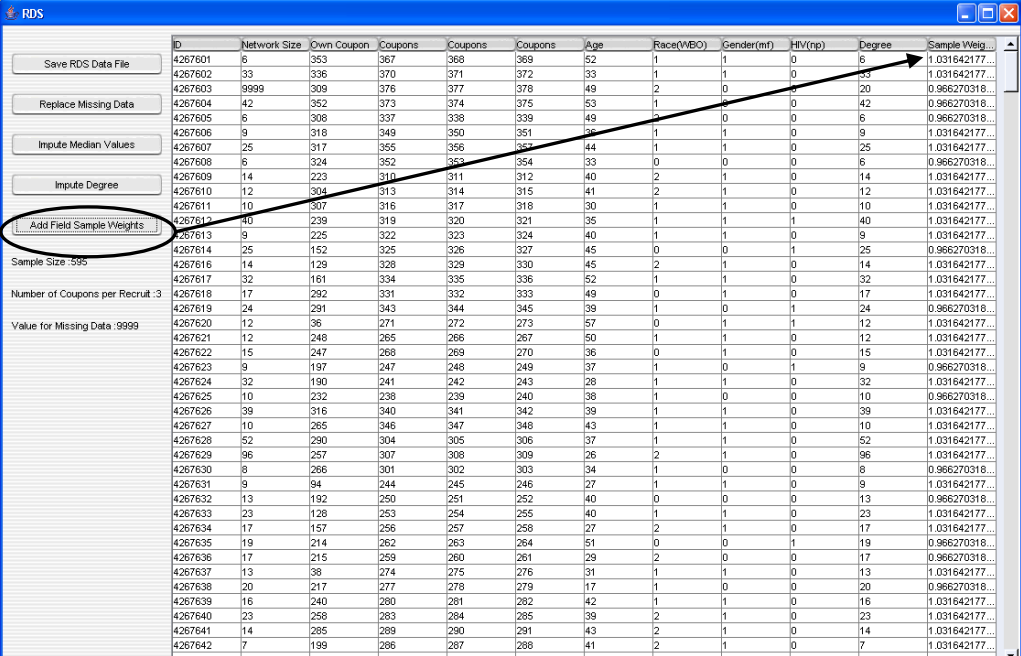
This feature imputes missing values on Network Size. To use this feature, first run a partition analysis on the Network Size variable. This analysis defines the groups that will be used to impute the Network Size. Next, click “Impute Degree.” To make changes permanent, click “Save RDS Data File.”

Note

The “Impute Degree” feature only functions after a partition has been analyzed because it uses the adjusted mean network size for the group (defined by the partition) in which each respondent is a member to impute the degree. To learn more about partition analysis, see Chapters 3 and 4 of this manual. Additionally, pulling in network size outliers will not affect degree imputation.

Add Field Sample Weights

This feature adds the Field Sample Weights to the RDS data file. It only appears in the Edit Data screen when a partition has been analyzed. In the Edit Data screen, click “Add Field Sample Weights.” A new column of data will appear that contains the Field Sample Weights. Click “Save RDS Data File” to make this change permanent. A field sample weight for a respondent is the population weight (see Estimation section in chapter 4) corresponding to the respondent’s variable value for the last partition. For example, if the most recent partition analysis was on “gender” and the respondent is male the population weight for males is that respondent’s field sample weight (see Figure 5.3).



ID	Network Size	Own Coupon	Coupons	Coupons	Coupons	Age	Race(WBO)	Gender(mf)	HIV(np)	Degree	Sample Weight
4267601	6	353	367	368	369	52	1	1	0	6	1.031642177...
4267602	33	336	370	371	372	33	1	1	0	23	1.031642177...
4267603	9999	309	376	377	378	49	2	0	0	20	0.966270318...
4267604	42	352	373	374	375	53	1	0	0	42	0.966270318...
4267605	6	308	337	338	339	49	0	0	0	6	0.966270318...
4267606	9	318	349	350	351	36	1	1	0	9	1.031642177...
4267607	25	317	355	356	357	44	1	1	0	25	1.031642177...
4267608	6	324	352	353	354	33	0	0	0	6	0.966270318...
4267609	14	223	310	311	312	40	2	1	0	14	1.031642177...
4267610	12	304	313	314	315	41	2	1	0	12	1.031642177...
4267611	10	307	316	317	318	30	1	1	0	10	1.031642177...
4267612	40	239	319	320	321	35	1	1	1	40	1.031642177...
4267613	9	225	322	323	324	40	1	1	0	9	1.031642177...
4267614	25	152	325	326	327	45	0	0	1	25	0.966270318...
4267616	14	129	328	329	330	45	2	1	0	14	1.031642177...
4267617	32	161	334	335	336	52	1	1	0	32	1.031642177...
4267618	17	292	331	332	333	49	0	1	0	17	1.031642177...
4267619	24	291	343	344	345	39	1	0	1	24	0.966270318...
4267620	12	36	271	272	273	57	0	1	1	12	1.031642177...
4267621	12	248	265	266	267	50	1	1	0	12	1.031642177...
4267622	15	247	268	269	270	36	0	1	0	15	1.031642177...
4267623	9	197	247	248	249	37	1	0	1	9	0.966270318...
4267624	32	190	241	242	243	28	1	1	0	32	1.031642177...
4267625	10	232	238	239	240	38	1	0	0	10	0.966270318...
4267626	39	316	340	341	342	39	1	1	0	39	1.031642177...
4267627	10	265	346	347	348	43	1	1	0	10	1.031642177...
4267628	52	290	304	305	306	37	1	1	0	52	1.031642177...
4267629	96	257	307	308	309	26	2	1	0	96	1.031642177...
4267630	8	266	301	302	303	34	1	0	0	8	0.966270318...
4267631	9	94	244	245	246	27	1	1	0	9	1.031642177...
4267632	13	192	250	251	252	40	0	0	0	13	0.966270318...
4267633	23	128	253	254	255	40	1	1	0	23	1.031642177...
4267634	17	157	256	257	258	27	2	1	0	17	1.031642177...
4267635	19	214	262	263	264	51	0	0	1	19	0.966270318...
4267636	17	215	259	260	261	29	2	0	0	17	0.966270318...
4267637	13	38	274	275	276	31	1	1	0	13	1.031642177...
4267638	20	217	277	278	279	17	1	0	0	20	0.966270318...
4267639	16	240	280	281	282	42	1	1	0	16	1.031642177...
4267640	23	258	283	284	285	39	2	1	0	23	1.031642177...
4267641	14	285	289	290	291	43	2	1	0	14	1.031642177...
4267642	7	199	286	287	288	41	2	1	0	7	1.031642177...
4267643	6	284	295	296	297	37	0	1	1	6	1.031642177...

FIGURE 5.3 RDSAT 7.1 Add Field Sample Weights

6 The RDSAT 7.1 File Menu

The RDSAT 7.1 File Menu contains several features. This chapter describes how to use them.

RDSAT 7.1 File Menu Features

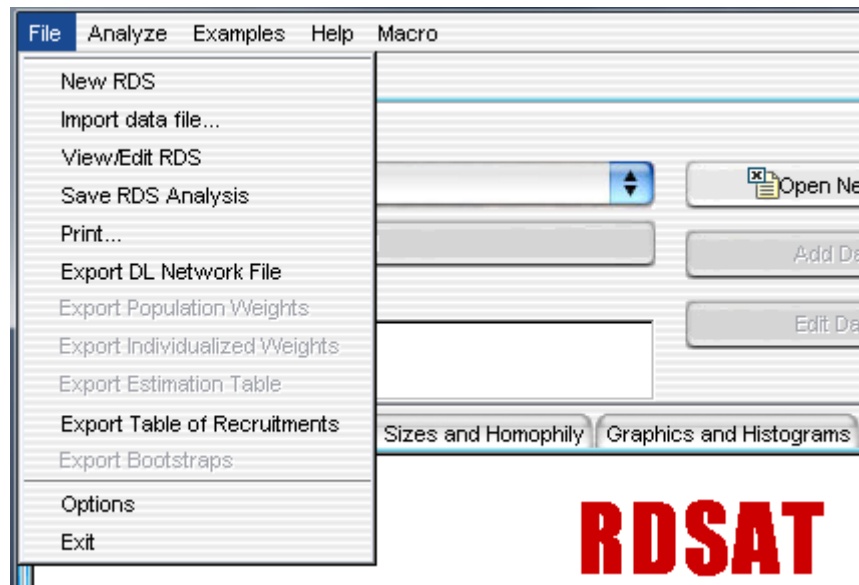


Figure 6.1 RDSAT 7.1 File Menu

New RDS

This feature allows the user to open a new RDS data set. The [Open New RDS] button (on the main screen) serves the same function.

Import Data File

This feature opens the import wizard, which can be used to properly format a data file for use by RDSAT 7.1. See Chapter 1 for more information on this feature.

View/Edit RDS

This feature opens the Edit Data screen. The [Edit Data] button (on the main screen) serves the same function.

Save RDS Analysis

This feature saves an RDS partition analysis in the form of a text file. It can be imported to Excel as a tab-delimited file.

Print...

This feature prints an RDS analysis.

Export DL Network File

Allows a DL network file to be exported to the recruitment chain data. DL format is recognized by numerous network analysis packages, including [UCI-net](#), [NetDraw](#), and [Pajek](#). NetDraw in particular, can be used to create attractive social network visualizations as seen in Figure 6.2.

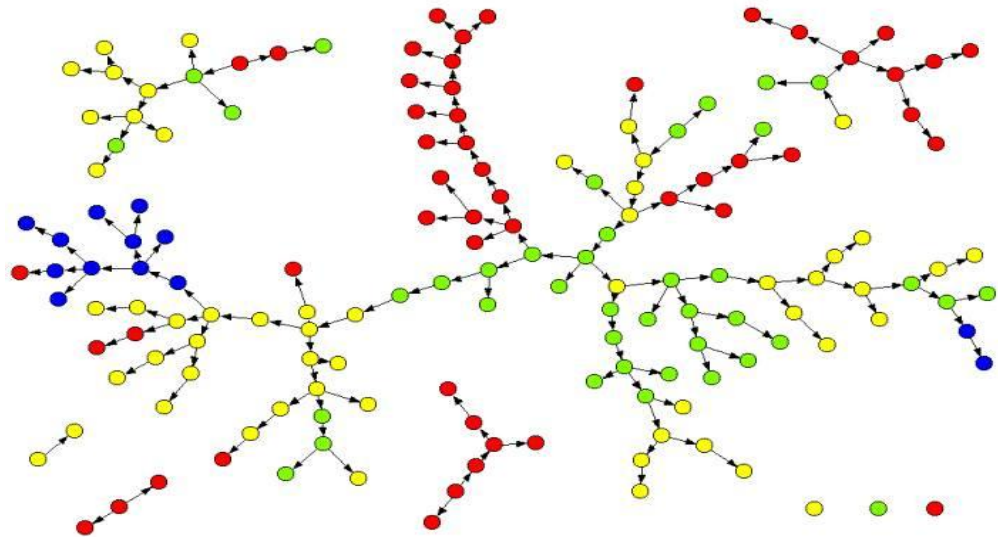


FIGURE 6.2 NetDraw-Generated Social Network Visualization

Export Population Weights

This function exports a text file of Population Weights (from “Population Estimates” table under “Estimation” tab, see Chapter 4), for each respondent based on the most recent partition analysis. Weights are linked to respondents by the Respondent ID. There will be a different weight for each group in the partition, and every individual in the group will be assigned the same group weight.

Export Individualized Weights

This function exports a text file of individualized RDS weights for each respondent. The weights are calculated based on respondents’ individual network sizes and the latest partition analysis performed. When generated for a dependent variable, these weights can be used to weight an entire data set for multivariate

analysis in a statistics program. These differ from Population Weights because they take each individual's network size into account. Therefore, each respondent will have a different weight (whereas all members of a given group have the same Population Weight). Both weights are used by statistics programs (e.g. SAS, SPSS) to adjust for an individual's probability of being sampled. Individualized weights are recommended for multivariate analysis.

Export Estimation Table

This function exports a text file of output and weights, corresponding to the most recent partition analysis performed, for each respondent in the data. In essence, this reproduces the "Population Estimates" table from the "Estimation" tab in RDSAT 7.1, so a partition analysis **MUST** be performed in order for this function to be available (see Chapter 4 in this manual for more detailed explanation of the "Population Estimates" table). The exported fields are:

RID: The Respondent ID

Group: Group number to which the respondent belongs

PopEst: The RDS population proportion estimate of the respondent's group.

Sample: The sample proportion of the respondent's group.

RecruitProp: The recruitment proportion of the respondent's group.

Equilibrium: The equilibrium proportion of the respondent's group.

Hx: The RDS homophily measure for the respondent's group.

Ha: The affiliation homphily measure for the respondent's group.

Hd: The degree homphily measure for the respondent's group.

Weight: The population weight for the respondent's group.

RecComponent: The recruitment component for the respondent's group (RCx).

DegComponent: The degree component for the respondent's group (DCx).

IndDegreeComp: The degree component based on the respondent's individual degree. This value is unique to the respondent.

IndweightComp: The individualized RDS estimator weight based on respondent's degree and the partition variable. When calculated for a

dependent variable, the data set can be weighted by this value for multivariate analysis.

Degree: The respondent's degree or personal network size.

The exported text file will look like this in Notepad:

estimationtable - Notepad

File Edit Format View Help

Key of Group and Trait Correspondence
Gender(MF) Language(ESC)

RID	Seed	Group	PopEst	Sample	RecruitProp	Equilibrium	Hx	Ha	Hd	weight	RecComponent	BegComponent	Indegree	Inc
1	1	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.702	0.1
2	1	1	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
3	1	1	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.76781	0.1
4	1	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.13556	0.1
5	1	1	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.51187	0.1
6	1	1	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.63984	0.1
7	1	1	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.07678	0.1
8	1	1	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	1.27969	1.4
9	1	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.25594	0.1
10	1	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.07678	0.1
101	0	3	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	3.83906	4.1
102	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
103	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
104	0	2	0	0.06098	0.05797	0	0	-1	1	0	0	-2016300516729058000000000001	0	0.1
105	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
106	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
107	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
108	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
109	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
110	0	4	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.10652	0.1
111	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
112	0	2	0	0.06098	0.05797	0	0	-1	1	0	0	-2016300516729058000000000001	0	0.1
113	0	2	0	0.06098	0.05797	0	0	-1	1	0	0	-2016300516729058000000000001	0	0.1
114	0	2	0	0.06098	0.05797	0	0	-1	1	0	0	-2016300516729058000000000001	0	0.1
115	0	4	0	0	0	0	0	0	0	0	0	0	0	0.1
116	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.85312	1.1
117	0	3	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.30712	0.1
118	0	3	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.38931	0.4
119	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.07678	0.1
120	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
121	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
122	0	3	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.63984	0.1
123	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.09598	0.1
124	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.07678	0.1
125	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.10669	0.1
126	0	3	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	1.53562	1.7
127	0	3	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.12797	0.1
128	0	3	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	1.27969	1.4
129	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
130	0	4	0	0.19512	0.2029	0	0.73887	0.73887	1	0	0	-3321969174965271400000000000	0	0.1
131	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.30712	0.4
132	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.30712	0.1
133	1	0	0	0	0	0	0	0	0	0	0	0	0	0
134	1	0	0	0	0	0	0	0	0	0	0	0	0	0
135	1	0	0	0	0	0	0	0	0	0	0	0	0	0
136	1	0	0	0	0	0	0	0	0	0	0	0	0	0
137	1	0	0	0	0	0	0	0	0	0	0	0	0	0
138	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.25594	0.1
139	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.76781	1.1
140	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.07678	0.1
141	0	1	0.47104	0.45122	0.47826	0.671	0.51213	0.21562	0.21456	1.04393	1.48707	0.702	0.76781	1.1
142	0	3	0.52896	0.29268	0.26087	0.329	-0.1045	0.21562	-0.37802	1.80728	1.1241	1.60776	0.95976	1.1

FIGURE 6.3 RDSAT 7.1 Exported Estimation Table (text file)

Export Table of Recruitments

This feature exports a text file containing a list of every recruitment in the dataset. When this feature is clicked in the File menu, the following menu appears:

Export Table of Recruitments

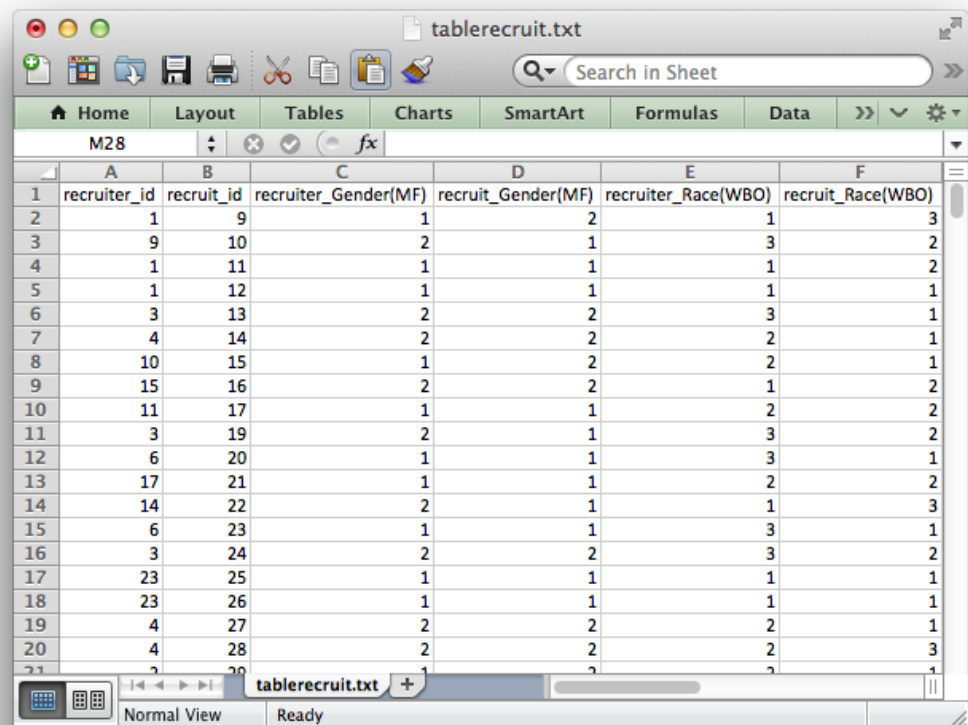
Available Variables	Variables to be Exported
Language(ESC)	
Age	
Gender(MF)	
Race(WBCHO)	

Add Remove

OK Cancel

FIGURE 6.4 Export Table of Recruitments pop-up

Similar to the way one would define a partition analysis, variables are moved to the right-hand “Variables to be Exported” pane by clicking a variable’s name in the “Available Variables” pane and clicking the [Add] button below (Figure 6.4). When the desired variables have been moved, click the [OK] button at the bottom of the window, and a standard Windows Save menu will appear. Enter the filename and specify the save location, then click “Save.” The file will contain a list of all recruitments based on the defined groups. The output is shown in Figure 6.5.



	A	B	C	D	E	F
	recruiter_id	recruit_id	recruiter_Gender(MF)	recruit_Gender(MF)	recruiter_Race(WBO)	recruit_Race(WBO)
1						
2	1	9	1	2	1	3
3	9	10	2	1	3	2
4	1	11	1	1	1	2
5	1	12	1	1	1	1
6	3	13	2	2	3	1
7	4	14	2	2	2	1
8	10	15	1	2	2	1
9	15	16	2	2	1	2
10	11	17	1	1	2	2
11	3	19	2	1	3	2
12	6	20	1	1	3	1
13	17	21	1	1	2	2
14	14	22	2	1	1	3
15	6	23	1	1	3	1
16	3	24	2	2	3	2
17	23	25	1	1	1	1
18	23	26	1	1	1	1
19	4	27	2	2	2	1
20	4	28	2	2	2	3
21	2	29	1	2	2	1

FIGURE 6.5 RDSAT 7.1 Exported Table of Recruitments (text file)

The first column is recruiter’s Recruiter ID, and the second column is recruit’s Recruiter ID. The third column is recruiter’s value on the selected variable, and the fourth column is recruit’s value on the selected variable. If more than one variable is specified, the columns continue with the recruiter’s variable value then the recruit’s variable value. This file can be used by some network analysis computer programs.

Export Bootstraps

This feature exports a text file containing the data used to generate the bootstrap results histogram.

- a) The tab delineated text output contains the count of bootstraps results that fall within bins .001 wide.
- b) The file will always contain 1002 rows, where the first row contains variable names, and the subsequent rows contain histogram bins 0 through 1 by .001.
- c) The first column in the file, named “bootstrap_value” contains the bin labels for 0 through 1 by .001.
- d) After the first column, there will be one column for each value/group in the most recent partition. These columns contain the frequency of that row’s bin value in that variable value’s bootstrap list.

Options

This feature opens the options menu. The [Change Options] button (on the main screen) serves the same function.

Exit

This feature exits the RDSAT 7.1 program.

7 The RDSAT 7.1 Analyze Menu

The RDS Analysis Tool offers several features not directly associated with partition and breakpoint analyses. They will be discussed in this chapter.

Estimate Number of Waves Required

The Estimate Number of Waves Required feature allows hypothetical recruitment scenarios to be examined through simulation. A group is selected to be the initial recruiters (seeds), and they are allowed to recruit based on the estimated transition probabilities until the sample proportion stabilize. This helps in determining how many waves of recruitment are necessary before the sample reaches equilibrium.

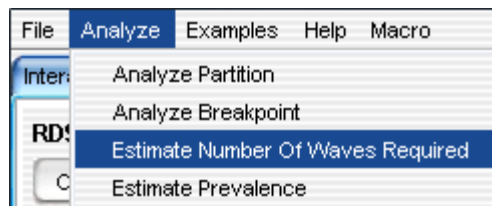


FIGURE 7.1 RDSAT 7.1 Estimate Number of Waves Required Menu Item

To use this feature, first analyze a partition on the variable for which you want to estimate number of waves required (see Chapters 3 and 4 for information on analyzing a partition).

After you have analyzed a partition, click on “Estimate Number of Waves Required” in RDSAT 7.1’s Analyze menu (Figure 7.1). This will cause the window of Figure 7.2 to appear. Then select a starting group (from the variable you analyzed a partition on) for a hypothetical sample. Next, choose a convergence radius.

The waves estimation feature estimates how many sample recruitment waves would be required for a given subgroup partition to reach "equilibrium." It estimates this by determining the point at which the sample proportions for the subgroup partition change very little as new recruitment waves are added to the sample. The convergence radius is the maximum allowed change in sample composition values between waves when a sample has reached equilibrium. For a given subgroup partition, a smaller convergence radius will always take at least as many waves to reach equilibrium as a larger convergence radius (and will often

increase the computing time required). Because a smaller convergence radius means the estimates must be more stable across recruitment waves to be considered in equilibrium, the estimated total number of waves required will be more conservative than it would be for a larger convergence radius.

Note

The default Convergence Radius in the “Estimate Waves” feature is .02, which serves as a good starting point for a waves analysis.

A radius of .02 means that the sample population proportions are considered converged (at equilibrium) when the change in population proportions in between waves is less than the convergence radius. Click OK, and this utility will use the Markov process implicit in the calculated transition probabilities to check how many waves are required for the sample proportions of your variable to reach equilibrium. The results of the analysis will be output to a new report page (see Figure 7.3).

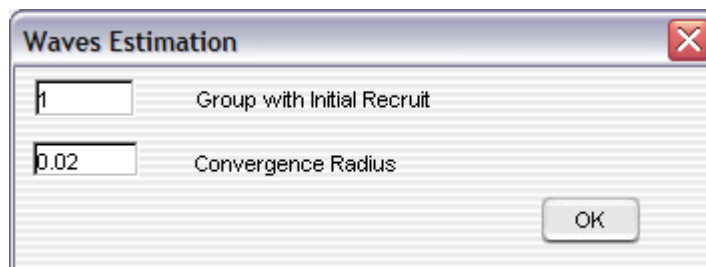


FIGURE 7.2 RDSAT 7.1 Waves Estimation Window

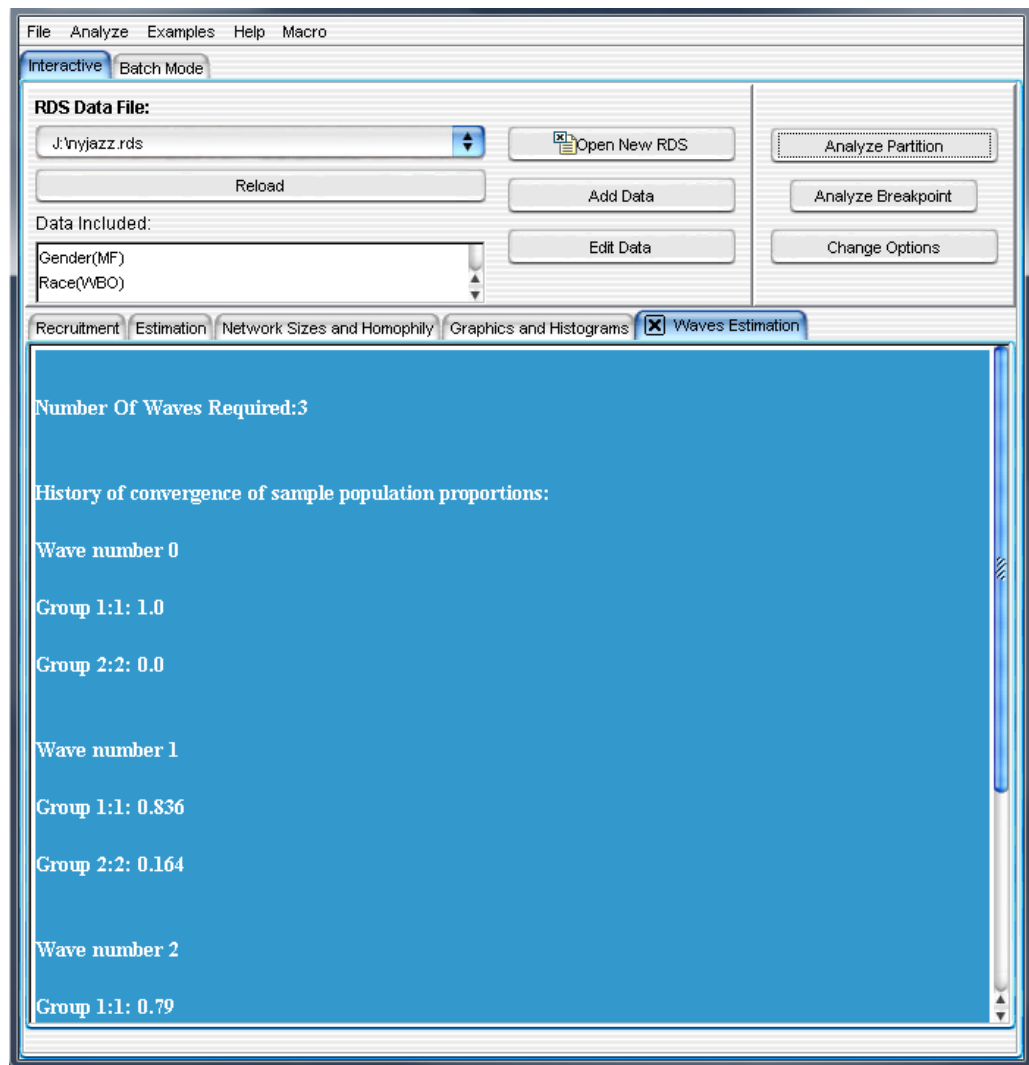


FIGURE 7.3a RDSAT 7.1 Waves Estimation

Figure 7.3a is a screenshot of the waves estimation output for a partition analysis of the New York Jazz dataset. The reformatted output is listed below (Figure 7.3b)

Number Of Waves Required: 3

History of convergence of sample population proportions:

Wave number 0

Group 1:1: 1.0

Group 2:2: 0.0

Wave number 1

Group 1:1: 0.836

Group 2:2: 0.164

Wave number 2

Group 1:1: 0.79

Group 2:2: 0.21

Wave number 3

Group 1:1: 0.778

Group 2:2: 0.22

FIGURE 7.3b RDSAT 7.1 Waves Estimation – Formatted Results

What this information means is that it took a total of 3 recruitment waves before the sample proportions changed by less than .02 (with a convergence radius of .02). As we can see, the change in sample proportion of Group 1 from wave 2 to 3 is $.79 - .778 = .012$, which is less than .02. The same is true of Group 2.

Estimate Prevalence

Prevalence estimation is similar to partition analysis, only more complicated ratio estimates can be produced. As an example, we will determine the HIV prevalence and confidence interval among males in an RDS sample (Figure 7.4).

First, a partition analysis of the relevant variables must be run (see Chapters 3 and 4 for more information on executing a partition analysis). Once you have done a partition analysis, identify the groups of interest for prevalence estimation using the “Key”. In our example, HIV positive males are Group 1.1 and non-HIV positive males are Group 1.2.

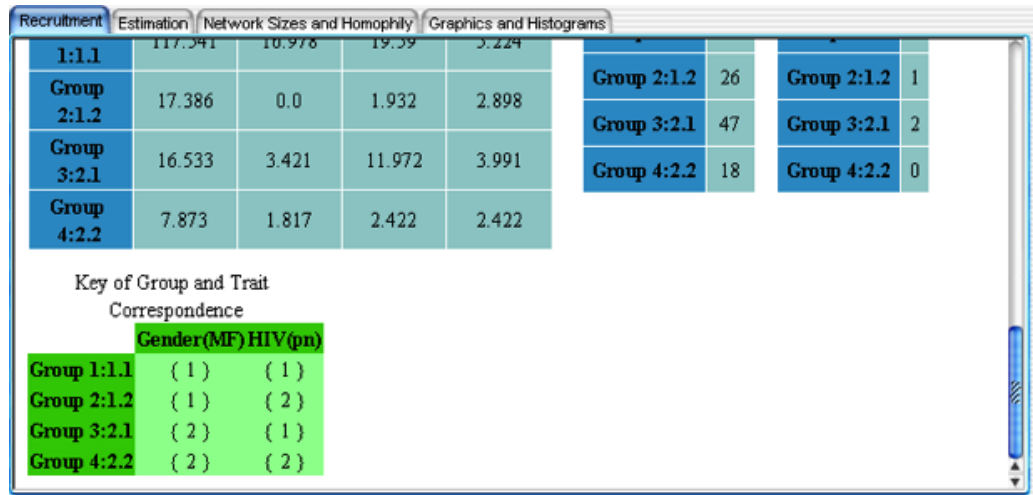


FIGURE 7.4 Key of Group and Trait Correspondence in Recruitment Tab

We are now ready to perform prevalence estimation. From the menu items select: **Analyze → Estimate Prevalence**, as shown below:

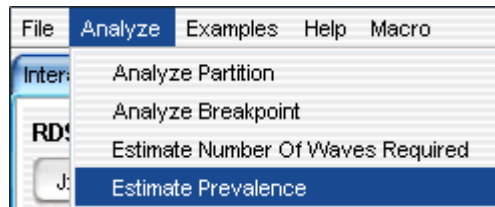


FIGURE 7.5 Analyze → Estimate Prevalence

The prevalence function requires you to enter the denominator and numerator used for estimation. Use the “Select Group” buttons to enter these fields. The groups appearing in the pull down menu correspond to groups from *the most recent partition analysis performed*. Then click “OK”.

In our case, we want the prevalence of HIV among males within the population. Thus, the numerator is Group 1.1 (HIV positive males) and the denominator is BOTH Group 1:1.1 (HIV positive males) and Group 2:1.2 (non-HIV positive males) (Figure 7.6).

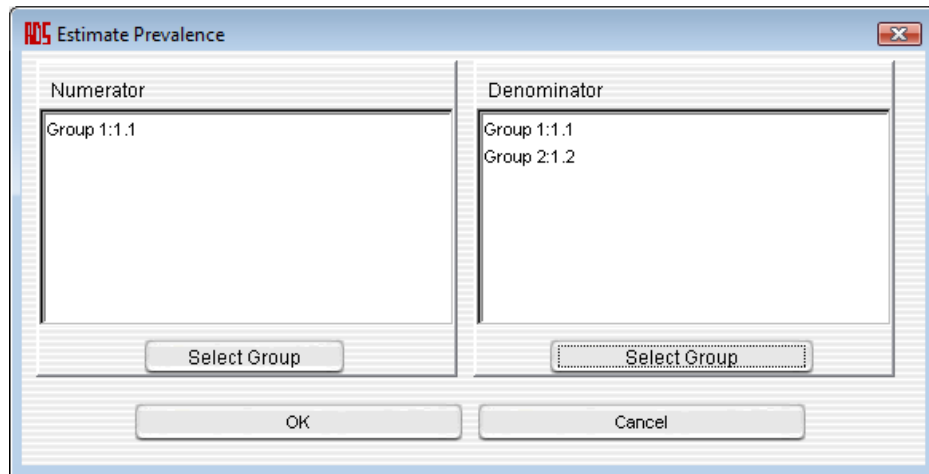


FIGURE 7.6 Estimate Prevalence Window

Once the analysis is performed, the output will appear in a new tab called “Ratio”. The output contains a prevalence estimate and confidence interval for that estimate as well as those groups used by the function and Key of Group and Trait Correspondence.

In our example, 87.6% of males are estimated to be HIV positive. The confidence interval for this estimate is 81.9% to 92.1%.

Recruitment

Estimation

Network Sizes and Homophily

Graphics and Histograms

Ratio

Prevalence Data

	Prevalence	Lower Bound	Upper Bound
Ratio Estimator	0.876	0.819	0.921

Ratio Composition

	Ratio Composition
Numerator	Group 1:1.1
Denominator	Group 1:1.1, Group 2:1.2

Key of Group and Trait Correspondence

	Gender(MF)	HIV(pn)
Group 1:1.1	(1)	(1)
Group 2:1.2	(1)	(2)
Group 3:2.1	(2)	(1)

FIGURE 7.7 Estimate Prevalence Output Screen - Ratio Tab

8 Batch Mode: Convert Files

RDSAT 7.1 has two modes of operation: interactive and batch modes. Interactive mode allows users to analyze one file at a time using interactive (point-and-click) menus; batch mode allows users to specify savable “jobs” that can perform multiple analyses on one or more files.

Accessing Batch Mode Tools

The row of tabs below the menu bar labeled “Interactive” and “Batch Mode” are used to switch between operating modes. (The features available in prior versions of RDSAT 7.1 are available through the “interactive” tab and the new batch tools can be accessed by clicking on the “batch mode” tab (see Figure 7.1).)

When to use Batch Mode

Use the RDSAT 7.1 batch processing tool to facilitate multi-variable and multi-site analysis and to simplify analysis replication. Batch processing is a type of automation where a set of user-specified functions are applied to a collection, or batch, of files. Users should be familiar with interactive mode of RDSAT 7.1 before attempting to use batch mode.

Batch Mode is useful for the following types of tasks:

- Convert a large number of data files to the .RDS format.
- Analyze multiple data sets using the same RDSAT 7.1 settings.
- Estimate the prevalence of a variable within multiple partitions on one or more data sets.
- Aggregate RDS estimates across multiple files.
- Create a record of the analysis settings and estimates produced for archival purposes.

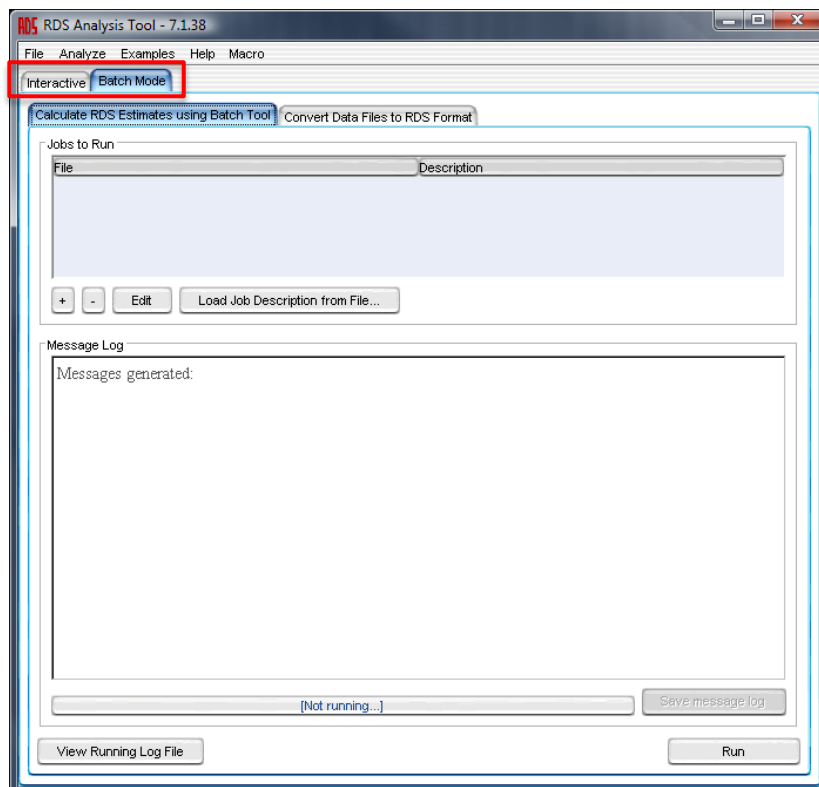
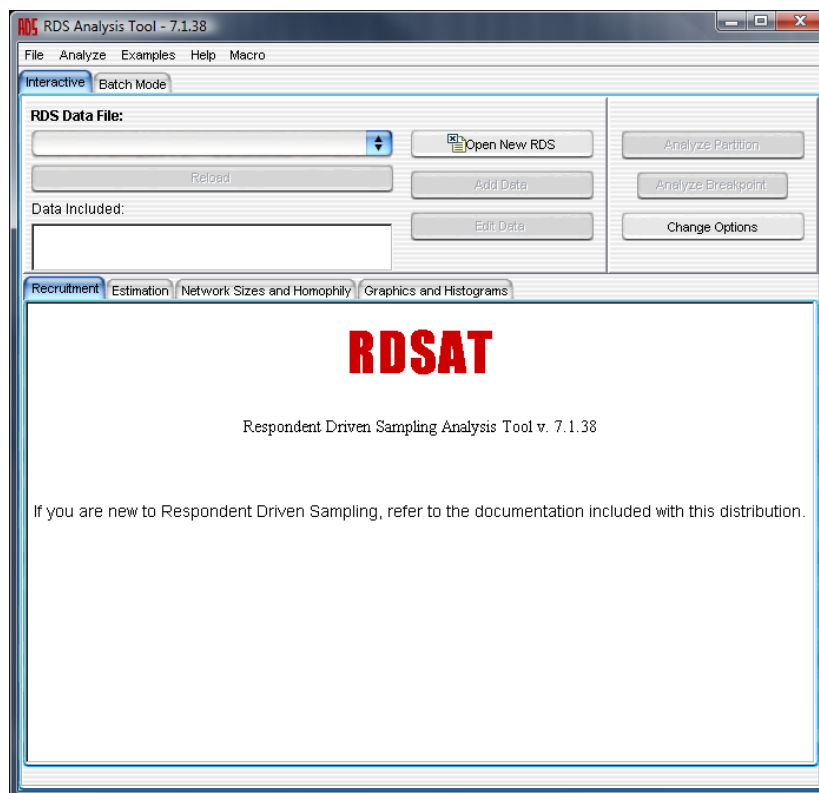


FIGURE 7.1 The RDSAT 7.1 operating mode is selected using the tabs below the menu bar.

Batch File Conversion Tool

Batch file conversion tool is useful when multiple data sets need to be converted to the RDS format. RDSAT 7.1 can convert SAS export files (.xpt) and character delimited text formats such as the comma-separated-value format (.csv).

Access the Batch File Conversion Tool by clicking the “Convert Data Files to RDS Format” tab in the “Batch Mode” interface (see Figure 8.2).

Batch File Conversion Settings can be saved to file and reloaded using ‘Save Batch Convert Setting to File’ and ‘Load Batch Convert Settings From File’ buttons. The saved “file conversion settings” include both the actual settings and the list of files to which these settings apply. This is particularly useful for ongoing studies where new data can be added to the file, but variables names are static. By reloading settings from a previous import, the updated files can be easily converted to RDS format.

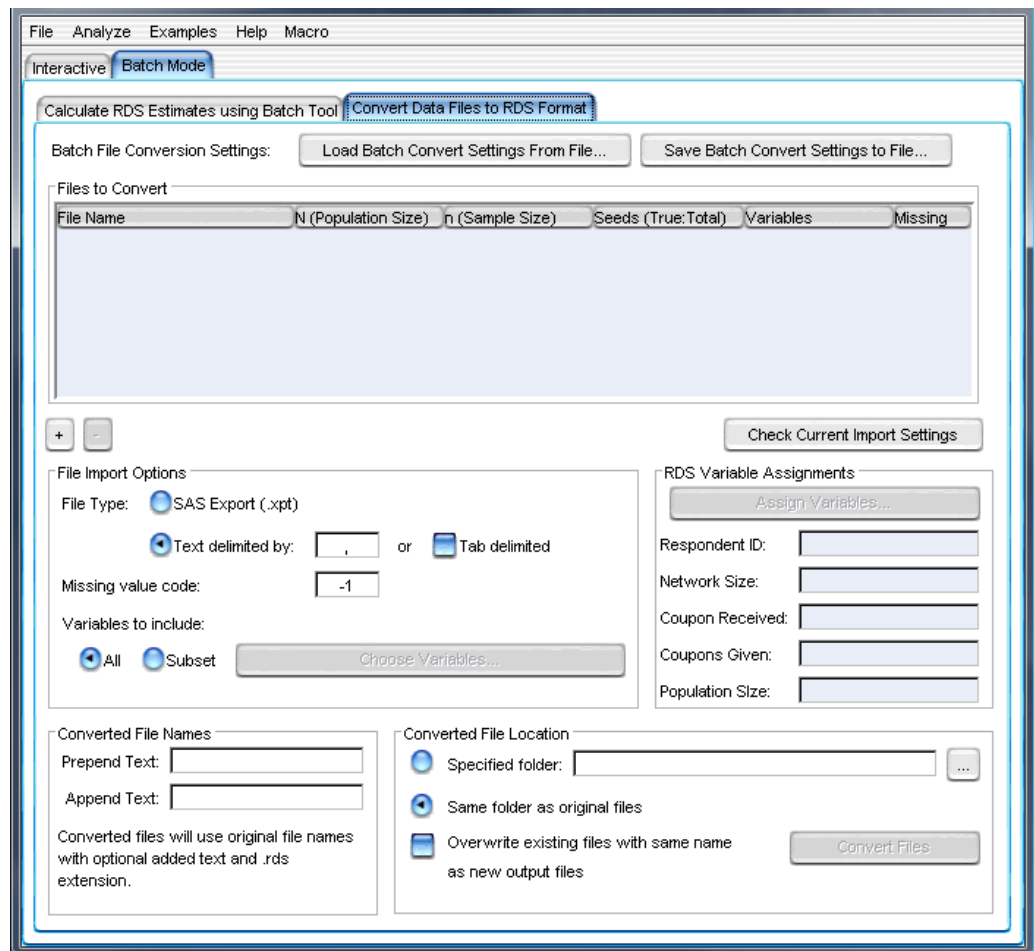


FIGURE 8.2 Convert Data Files to RDS Format Window

Converting files with RDSAT

The batch conversion tool allows a single set of import settings to be applied to multiple files.

The Convert Data Files to RDS Format window is divided into five sub-sections where import settings are specified (see Figure 8.2):

Files to Convert. The data files to import.

File Import Options. Specify the file format, missing data and variables to include in the converted file.

RDS Variable Assignments. Specify which variable names are associated with each of the RDS-specific variables required for RDS estimates.

Converted File Names. Text to add to converted file names.

Converted File Locations. Select where converted files should be saved.

1. Files to Convert. Files are added to the file list with the [+] button and removed with the [-] button. In order to successfully convert multiple files, each file must be in the same format and must use the same names for the variables that correspond to the RDS header variables: Respondent ID, Network Size, Coupon Received (from recruiter), Coupons Given (to recruit others) and, optionally, Population Size ('popsize').

2. File Import Options. Select the file type and the delimiter if necessary. Delimiters can be entered by typing the appropriate characters or using standard escaped character notation for non-printing characters. The CSV format is indicated by typing “,” in the delimiter field while a tab-delimited file would be indicated by typing “\t” in the delimiter field. As a convenience to users, there is a check box to indicate “tab delimited”. Note that checking the tab-delimited box overrides any text entered in the delimiter field.

The missing value code is specified by typing the missing character or character sequence that represents missing values in the data set. The missing value code may not contain spaces.

RDSAT 7.1 can either include all variables in each file or a subset of variables common to all files. The “All” option will include every variable in each file in the converted version of that file, including variables unique to that particular data file. When a subset is defined, those variables must be present in every file or an error will be generated for that file during conversion.

If desired, use the [Choose Variables] dialog to specify the subset of variables to be included in the converted files. The Choose Variables dialog shows two lists of variables (Figure 8.3). The Available Variables are all the variables in the first file on the Files to Convert list. Move the desired variables to the Included Variables list by highlighting the variable name and clicking the [>>] (move right) button.

Variables can be removed from Included Variables list by highlighting the variable name and clicking [<<] .

Tip

When importing multiple data files, the RDS header variable names and missing value must be the same across all the files. If RDSAT 7.1 should include all the variables found in the source file in the converted files, choose “All” for “Variables to include” in the import settings. This feature works on a file-by-file basis, so a single batch convert can import similar files even if each contains some unique variables.

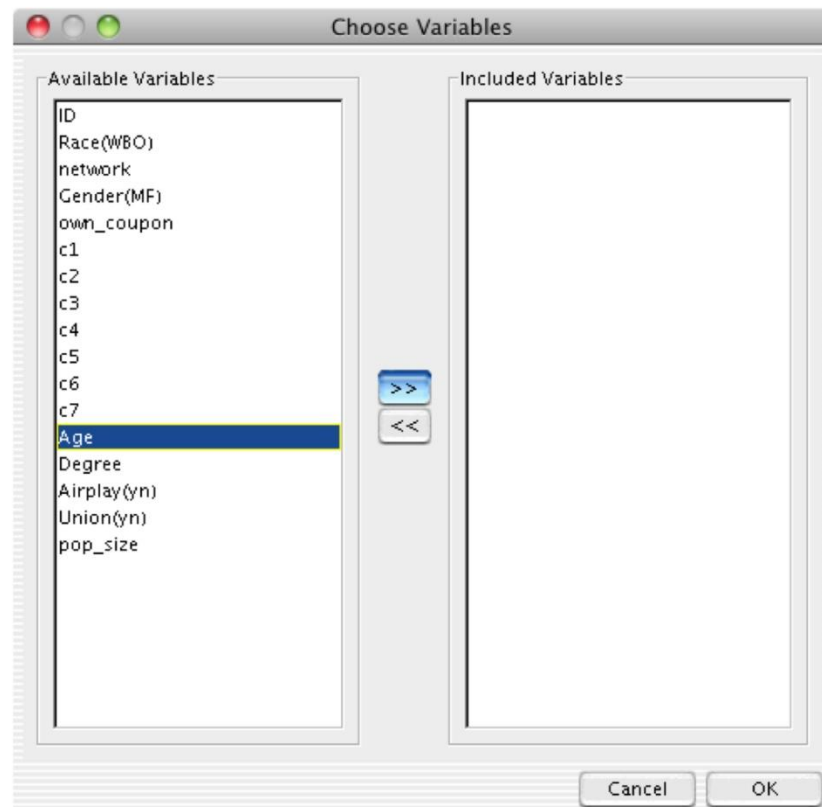


FIGURE 8.3 Optional dialog to select a subset of variables to include in the converted file(s).

File Conversion Notes:

- (1) If variables have different names across data files, the variables should be renamed using a different program prior to converting these files in RDSAT 7.1.
- (2) The RDS header variables must be present in every file and the variable names must be the same.

(3) The “Choose Variables...” and “Assign Variables” dialogs are populated using the contents of the first data file in the list. It is possible to select a subset of variables to include that is not valid across all files. This situation will generate an error when file conversion is attempted.

(4) Variables that correspond to “Coupons Given” must be the present in every data file, so it may be necessary to add extra columns of missing values to pad files with fewer “Coupons Given” variables. This would be most common when different sites issue different numbers of coupons per respondent.

3. RDS Header Variable Assignments. RDS requires certain data columns to calculate weighted estimates; these data columns must be matched to corresponding RDS header variables. The required variables for the RDS header are Respondent ID, Network Size, Coupon Received (from recruiter) and Coupons Given (to recruits). If cross-file aggregation will be used, a Population Size variable is necessary as well. Use the “Assign Variables” dialog to match the data column names with their corresponding RDS header fields (see Figure 8.4). Highlight the desired variable in the Available Variables list on the left and click the [>] button next to the corresponding field on the right to assign that variable to one of the RDS header variable roles.

These assignments must be valid across all files or the file conversion will fail. The assigned variables must be present and identically named in each file included in the file list.

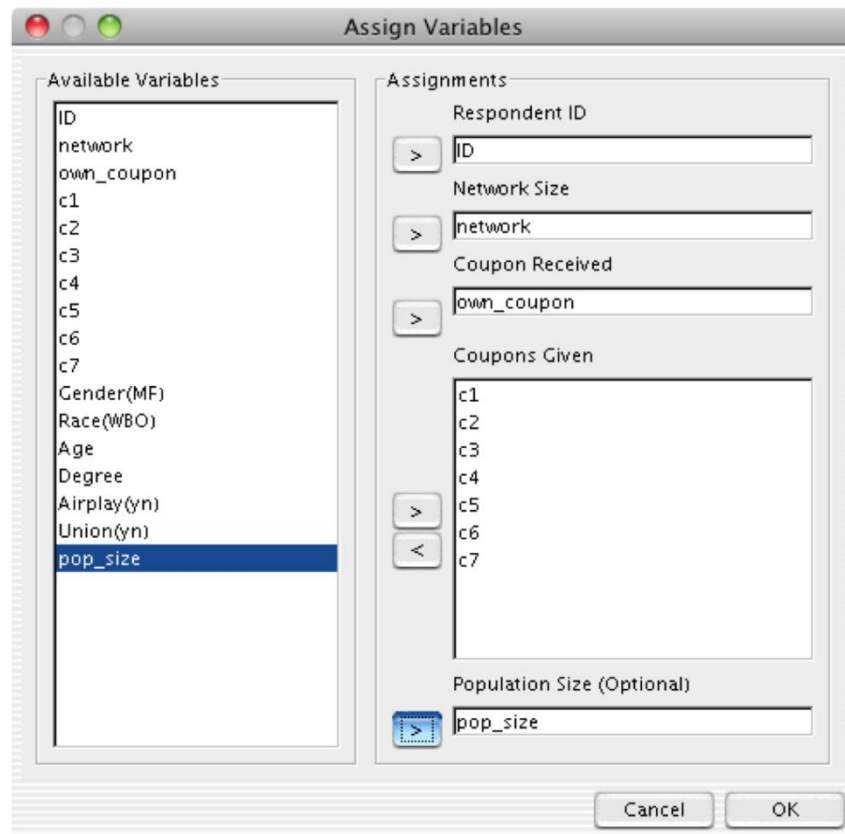


FIGURE 8.4 Variable assignment dialog.

4. *Converted File Names.* By default, converted files have the same name as the input file except that the file extension is changed to “.rds”. Additional text can be prepended or appended to the filename by typing in the appropriate boxes. It is helpful to use underscores or hyphens to separate the added text from the original file name.

5. *Converted File Location.* Converted files can be saved to the same directory as the original files or all converted files can be saved to a single directory.

RDSAT 7.1 offers the option to overwrite existing files to improve file management and accommodate workflows that require generating updated estimates as data files are updated.

Verify the Conversion Settings

After the five sections of the convert files dialog are set, the Convert Data Files dialog will resemble Figure 8.5. The Files to Convert list includes 4 columns with demographic and diagnostic information intended to help the user identify errors in the import settings. *Population Size* indicates the value set for population size for each file. If population size is not present, aggregated estimates cannot be

produced with the converted files. *Sample Size* indicates the number of data rows RDSAT 7.1 recognizes and *Seeds* shows how many respondents have missing or invalid recruiter coupons. The true seed count is the number of seeds with missing recruiter coupons. The total seeds are the number of true seeds plus the number of respondents with coupon numbers that were not issued to any recruiter. Comparing the true and total seed count can help identify data entry errors. **The seeds not indicated as true seeds will be weighted as if they were validly recruited. To override this behavior, use an external data editor to set the *coupon_recieved* value for these respondents to missing.** *Variables* indicates the number of columns identified and *Missing* indicates the proportion of cells with the *Missing value code* specified in the File Import Options. Users should verify that these values match expected values. A small amount of missing data is expected because seeds are indicated by the missing value in the Coupon Received variable.

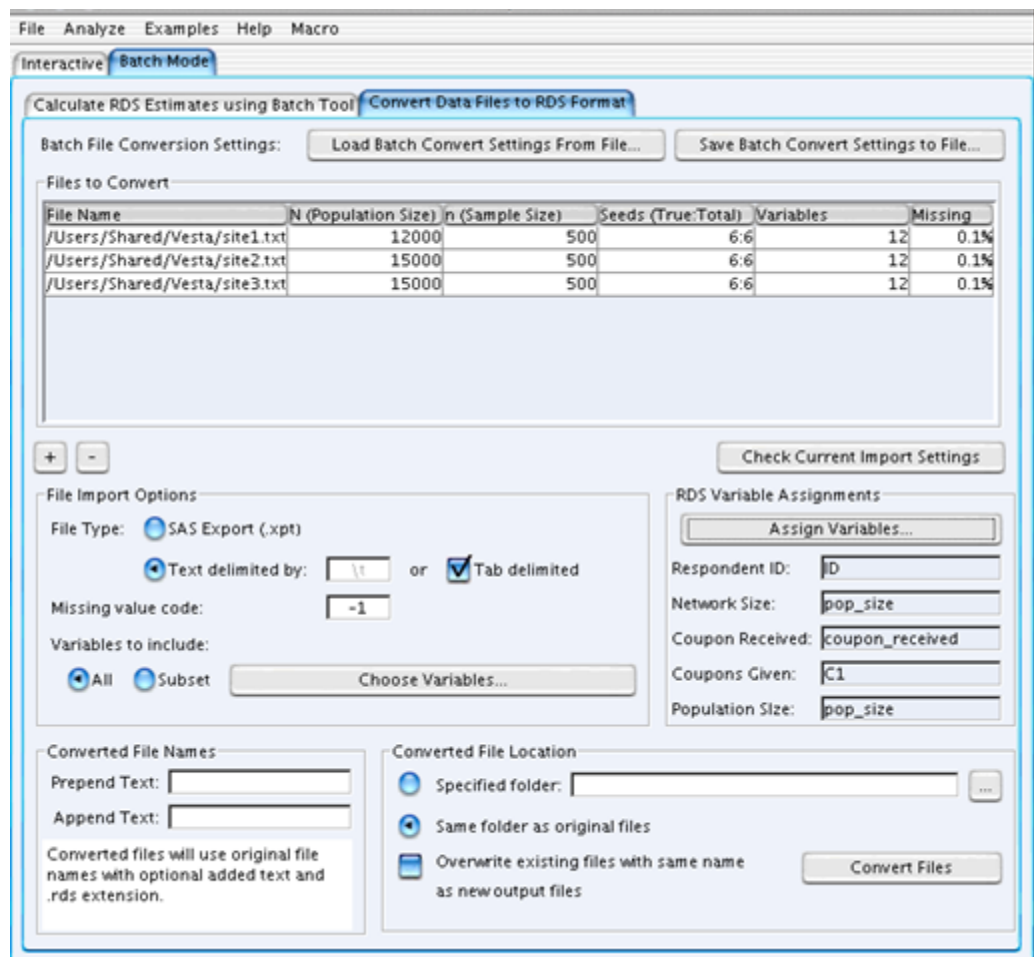


FIGURE 8.5 Convert Data Files with set of files to import and import options specified.

Correcting errors in the Conversion Settings

If the current conversion settings cannot be applied to a file, that file will be shown in red text in the “Files to Convert” list. If none of the listed files can be converted, it is likely that the file type settings are incorrect. Verify the file format and delimiter.

If only one of several similar files cannot be imported, it is likely that the file is missing either a variable required for the RDS Header from the “Assign Variables” list or that the file is missing a variable specified in the “Subset” option of “Variables to include.” Try changing “Variables to include” to “All” and verify that the file contains the expected variables.

File Analyze Examples Help Macro

Interactive **Batch Mode**

Calculate RDS Estimates using Batch Tool **Convert Data Files to RDS Format**

Batch File Conversion Settings: Load Batch Convert Settings From File... Save Batch Convert Settings to File...

Files to Convert

File Name	N (Population Size)	n (Sample Size)	Seeds (True:Total)	Variables	Missing
/Users/Shared/Vesta/site1.txt	--	500	0.0	1	none
/Users/Shared/Vesta/site2.txt	--	500	0.0	1	none
/Users/Shared/Vesta/site3.txt	--	500	0.0	1	none

NOTE: Files in red indicate parsing errors. Please check your import settings match the files being added.

+ -

Check Current Import Settings

File Import Options

File Type: ☒ SAS Export (.xpt)

☒ Text delimited by: or ☐ Tab delimited

Missing value code:

Variables to include: ☒ All ☐ Subset

RDS Variable Assignments

Respondent ID:

Network Size:

Coupon Received:

Coupons Given:

Population Size:

Converted File Names

Prepend Text:

Append Text:

Converted files will use original file names with optional added text and .rds extension.

Converted File Location

☒ Specified folder:

☒ Same folder as original files

☐ Overwrite existing files with same name as new output files

FIGURE 8.6 Convert Data Files Panel will show files in red text to indicate that the import settings cannot be applied to a file. All files will show red until the import options are specified and match the file format.

Convert Files

Clicking the [Convert Files] button begins the conversion process and opens a log window that shows the conversion process. When complete, the user can see any errors encountered and then save or dismiss the log (see Figure 8.7).

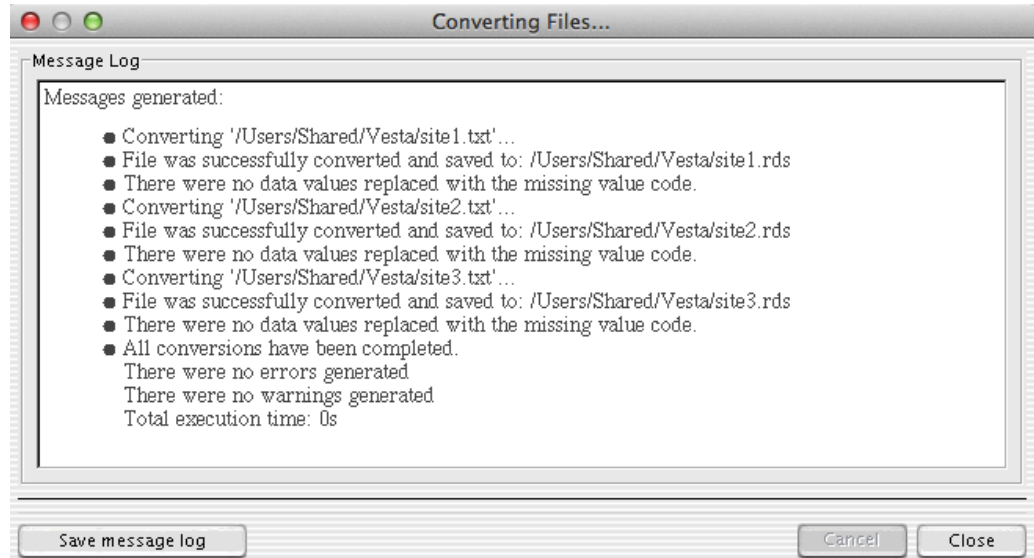


FIGURE 8.7 Conversion Log Window

If no errors are reported, the conversion was successful and the converted files are ready to be loaded into RDSAT 7.1. If desired, the import settings can be saved for later re-use by using the [Save Batch Conversion Settings to File...] button at the top of the main batch conversion dialog (see Figure 8.5).

9 Batch Mode: Calculate Estimates

RDSAT 7.1 has two modes of operation: interactive and batch modes. Interactive mode allows users to analyze one file at a time using interactive (point-and-click) menus; batch mode allows users to specify savable “jobs” that can perform multiple analyses on one or more files.

Accessing Batch Mode Tools

The row of tabs below the menu bar labeled “Interactive” and “Batch Mode” are used to switch between operating modes (see Figure 7.1). The features available in prior versions of RDSAT 7.1 are available through the “interactive” tab and the new batch tools can be accessed by clicking on the “batch mode” tab (see Figure 7.1). Batch calculation allows user-defined *jobs* to be executed sequentially with no user interaction.

Jobs and Subgroup Partitions

RDS analysis relies on cross-recruitment among two or more subsets of the population in order to generate weighted estimates. These population subsets are created by partitioning the population into distinct groups based on a set of attributes called a *subgroup partition*. A subgroup partition may be defined on a single variable (race) or a set of variables (race by gender). Each RDSAT 7.1 analysis is based on a user-defined subgroup partition (see Figure 9.1).

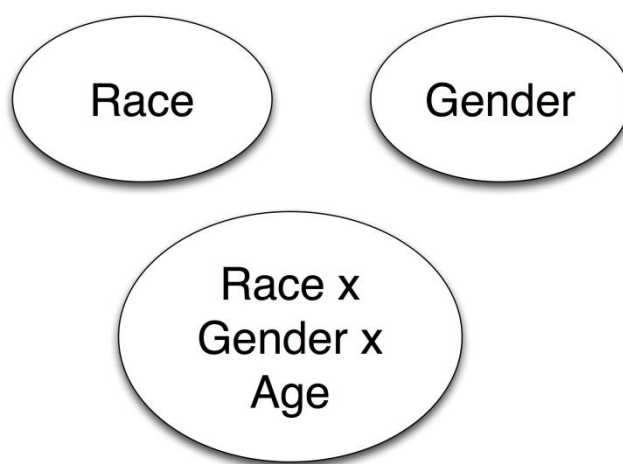


FIGURE 9.1 Conceptual Diagram: Examples of Subgroup Partitions

A *job* specifies a set of analyses to perform on a set of files. The job contains all the estimation options and the location where the output will be saved. Once a job is created, it can be saved as a file and reloaded into RDSAT 7.1 in the future. The job can be executed multiple times, so analyses can be easily repeated in the future (see Figure 9.2).

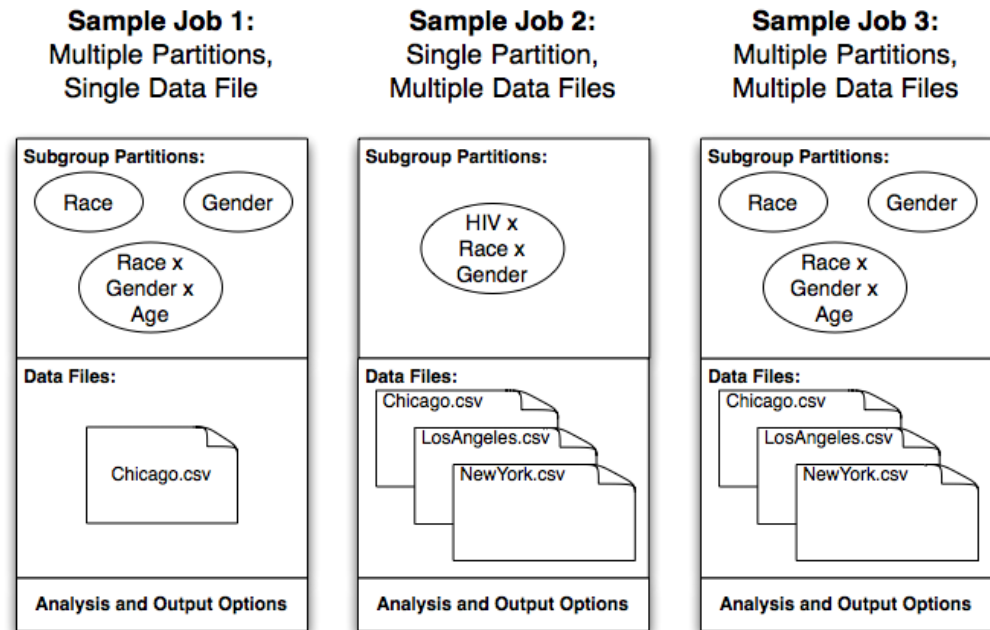


FIGURE 9.2 Conceptual Diagram: Sample Jobs

Creating a Batch in RDSAT

The Calculate Estimates window consists of two parts: a list of jobs which can be executed by clicking the [Run] button and a message log that reports the status of the job execution. In order to run jobs in batch mode, the user must first create or load a job.

The row of buttons below the job queue is used to create, load and edit jobs. A new job is created by clicking the add [+] button. The subtract [-] button removes the selected job from the list. A previously saved job can be loaded into the job queue by clicking the [Load Job Description from File...] button. A selected job can be edited by clicking the [Edit] button (see Figure 9.3).

Use the [Run] button to execute the jobs listed in the jobs list.

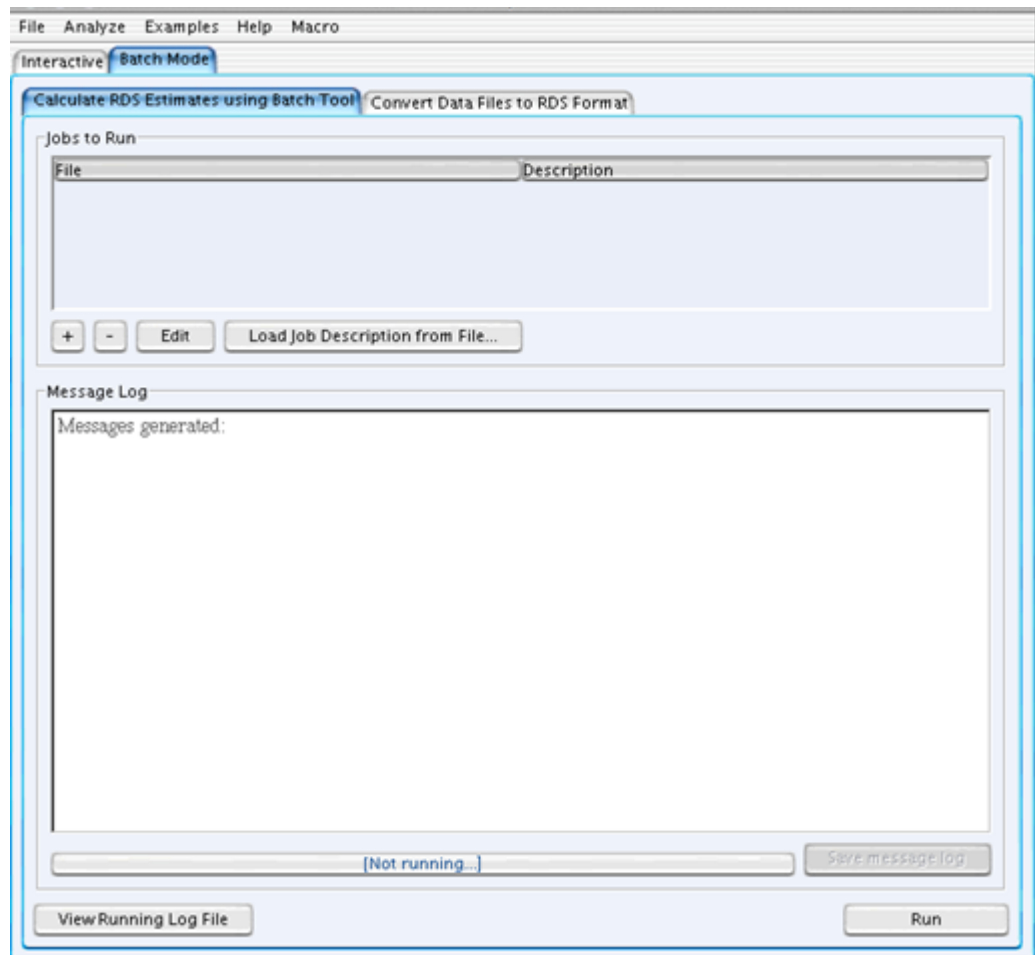


FIGURE 9.3 Estimates from jobs are generated via the “Calculate Estimates” tab in the “Batch Mode” interface.

Specifying a RDSAT 7.1 Job

A RDSAT 7.1 job is a file that specifies how RDSAT 7.1 should perform the analyses. A job is made of three parts:

Files. The data files to use.

Analysis. The subgroups to estimate and options specifying how the estimates are calculated.

Output. The report contents and where to save the output file.

Create a new job by clicking the add [+] button to open the Job Creation Wizard. The wizard has four screens—the first three screens correspond to one of the major parts of the job specification and the last screen provides a way to error check and save the job to a file (see Figure 9.4).

1. Files. Type a brief description about the analysis in the “Job Description” field. Use the add [+] and subtract [-] buttons to specify a list of data files that RDSAT 7.1 will use for this job.

Note

Each data file included in a single job must use the same variable names. Files may contain unique variables but analysis is only possible on variables present in all files.

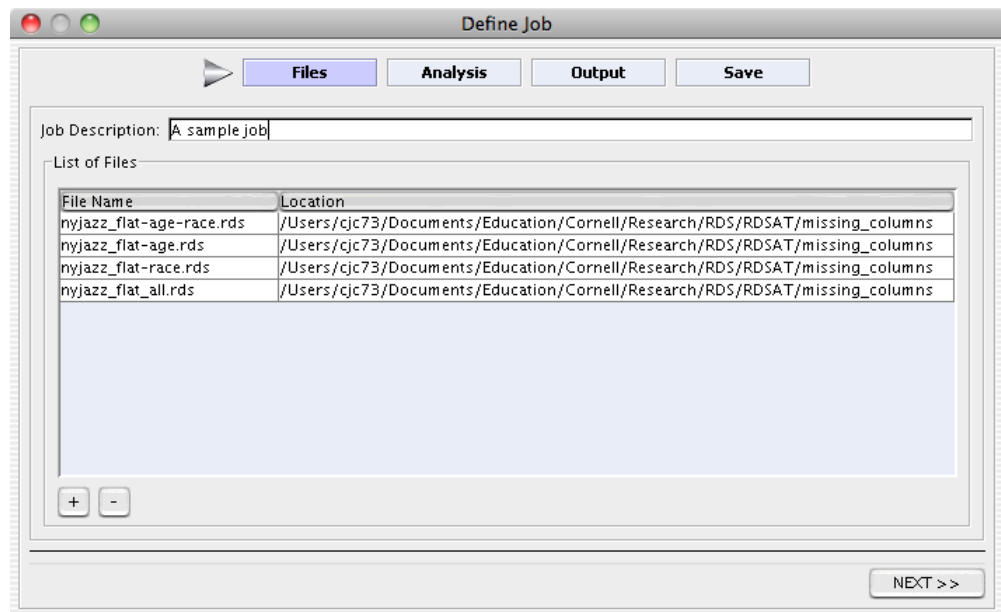


FIGURE 9.4 Job Creation Wizard - File Specification

2. Analysis. Set the default options for the analysis by clicking the [Set Default Options] button (shown in Figure 9.6). Setting the default options is an efficient way to apply the same estimation options to many subgroup partitions. Once defaults are specified, they can be saved and loaded again for use in other jobs. The options include the method used to calculate network size, a choice of RDS estimators, variance estimation options and data manipulation options (see Figure 9.5). These correspond to the options discussed at length in the “Setting Options For Analysis” section of chapter 3. The options shown below are recommended for generating an initial analysis and verification. To generate the most reliable confidence intervals, the number of bootstrap re-samples should be at least 15000.

Options

Average Network Size Estimation:

☐ Arithmetic Mean

☐ Multiplicity Estimate

☒ Dual Component

Mean Cell Size: 12

Number of re-samples for Bootstrap: 15000

Confidence Level Alpha (Width = $1-2*\alpha$): 0.025

☐ Pull-In Outliers of Network Sizes: 0 % (Maximum 50%)

☐ Exclude Waves Less Than: 0

☐ Treat excluded groups as a single group for estimation purposes

Algorithm type:

☐ LLS

☐ Data Smoothing

☒ Enhanced Data-Smoothing

OK

FIGURE 9.5 Job Creation Wizard - Subgroup Specification – Set Default Options. See Chapter 3 “Setting Options for Analysis” for detailed information about these options.

Use the text field labeled “Prevalence Options: Levels to Exclude from all Variables” to specify any variable levels that should be excluded from the prevalence estimates (see Figure 9.6). The most common use of this field is to exclude codes for technically missing data categories like “Don’t Know” and “Refused.” Enter the codes as they appear in the data file, separating the codes with a comma. See Chapter 3 for a detailed description of the Excluded Values estimation procedures. Chapter 10 describes the use of the table builder and table builder options buttons in the “Analysis” dialog box.

Once default options are set, use the add [+] and subtract [-] buttons to specify a list of subgroup partitions RDSAT 7.1 should estimate (see Figure 9.6).

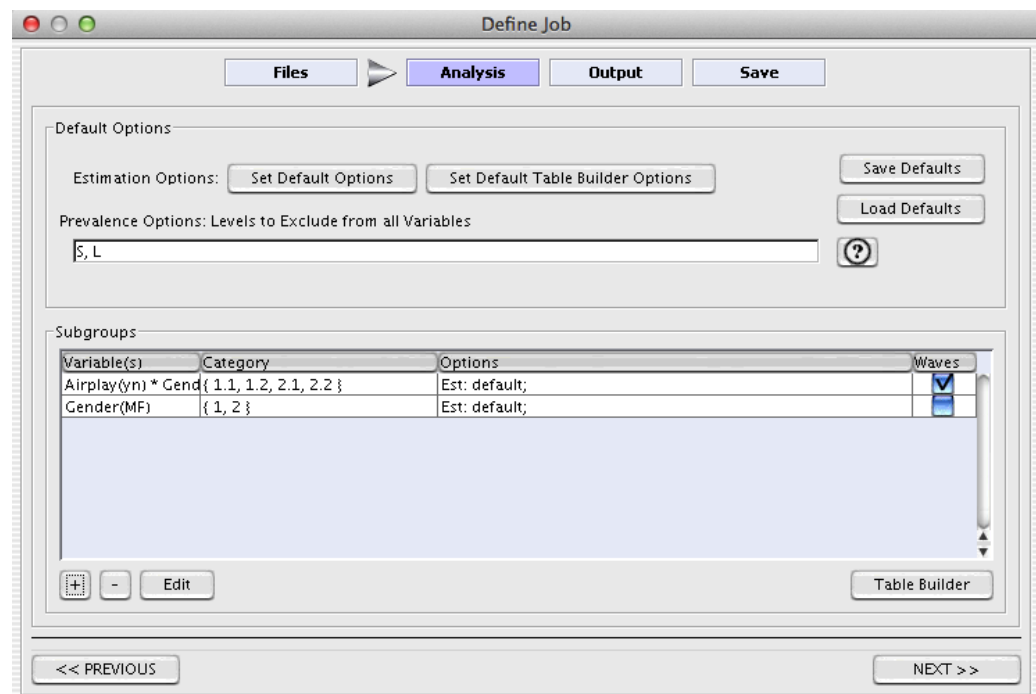


FIGURE 9.6 Job Creation Wizard – Subgroup List Screen

The add [+] button will open the “Define Subgroup” window, which consists of variable selection and options. Select a variable in the list on the left and include it in the analysis by clicking the “move right” button: [>>]. Move a variable out of the “Included” list by using the “move left” button: [<<]. The subgroup is defined by the unique combinations of the levels of the included variable(s). Most of the options available in batch mode are the same as those available when using RDSAT 7.1 in interactive mode, but the layout is different (see Figure 9.7).

Per Variable Options apply to the highlighted variable only and specify how RDSAT calculates the levels of that variable. These settings correspond to the options discussed in the “Partition Analysis” section of chapter 3.

Calculate Equilibrium Waves will include a computation of the number of recruiting waves in the sample and estimate the number of waves required to reach equilibrium. This is a diagnostic tool used to understand how the particular seeds that generated a sample might have biased the sample. Advanced use of this feature is discussed in the “Advanced Subgroup Analysis Features” section at the end of this chapter.

Prevalence Reports instruct RDSAT 7.1 to calculate the prevalence of one variable in the subgroup among the subgroups defined by the rest of the variable in the subgroup. It is possible to define multiple prevalence reports per subgroup. Click the [+] button to create a default prevalence report. The default report is the prevalence of the first variable in the list for the subgroups defined by the remaining variables. In the example shown in figure 9.7, the prevalence variable is Airplay because it is first on the list, and the subgroups for which prevalence will be reported are the combination of the factor levels for Gender and Race. The excluded column indicates if any variable levels will be excluded from the prevalence report. By default, only those variable levels defined as excluded in the Default Prevalence Options will appear in the excluded column. Note that default excluded values will only appear if present in the files, so verifying that the default excluded values appear as expected will help catch input errors. Customized prevalence reports are discussed in the Advanced Subgroup Analysis Features section at the end of this chapter.

Estimation Options Specify the *Custom* estimation option if the defined subgroup requires options different from the default options.

Calculate Aggregate Estimates When data files contain a valid population size variable, RDSAT 7.1 can generate weighted aggregated estimates across multiple data files. The default settings also generate estimates for each site individually, but this can be suppressed in the output if only the aggregate estimates are desired. The weight used to aggregate each group value is the estimated population proportion for the group multiplied by the overall population size assigned to the file. This weighting strategy accounts for differences in population sizes and population composition.

The [Add this Subgroup] button adds the specified subgroup with selected options to the subgroup partition list and resets the “Define Subgroup” window so a new subgroup can be defined. Once all desired subgroups are defined, use the [Done adding Subgroups] to dismiss the window.

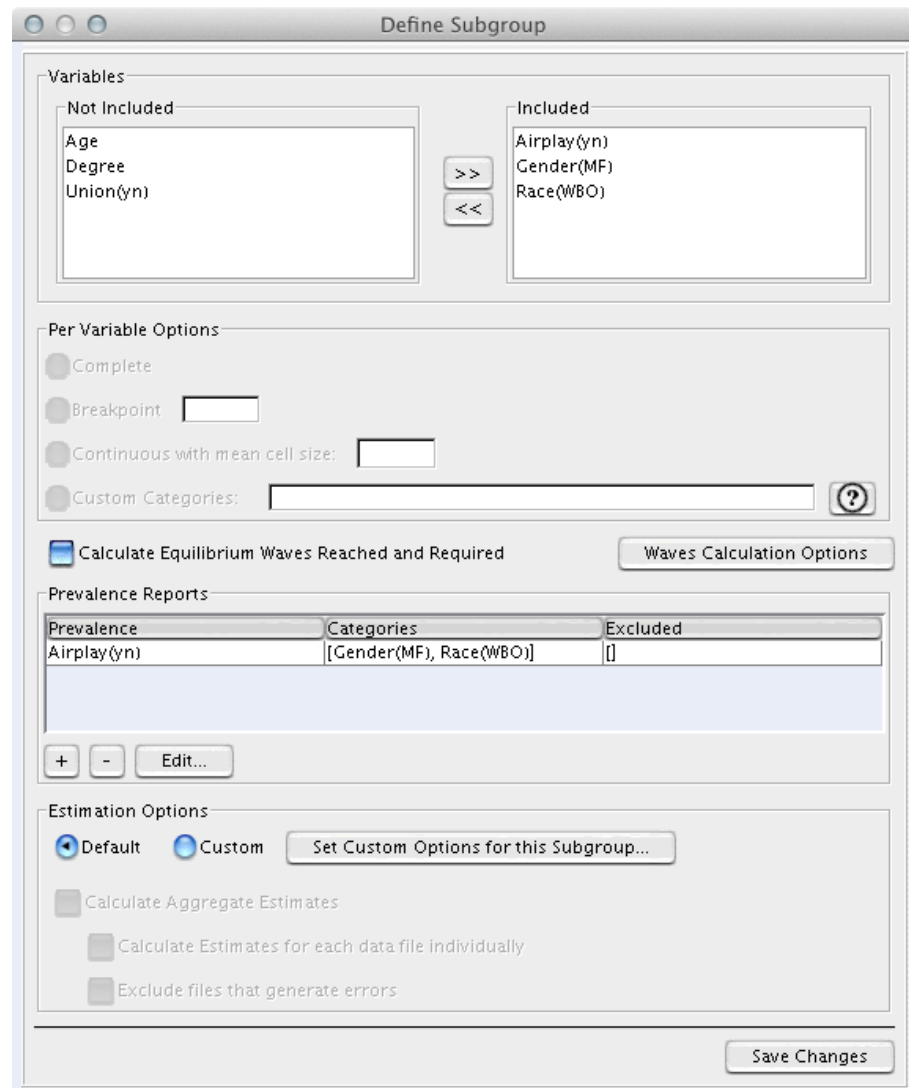


FIGURE 9.7 The Define Subgroup Window showing a Subgroup Partition of Airplay x Gender x Race

When all desired subgroup partitions have been specified and show in the Subgroups list (Figure 9.6) click the [Next >>] button to proceed.

3. Output. The *Output File Contents* lists the subgroups and the information that will be included in the output from the job. The *Output File Format* specifies how and where the output will be saved (see Figure 9.8).

Specify a location to save the results by clicking the [...] button next to the “Save As:” field. Note that the results from a single job can be saved to one or many files and in Excel compatible (.xls or .xlsx) or Comma-separated-value (.csv) formats. When multiple file output is selected, each subgroup partition will be reported in a separate file. These files will have the name specified in the “Save As” field with the partition name appended to the file name. When “.xls” or “.xlsx” formats are

specified, the results for each input data file will appear on a single worksheet within the Excel workbook. If the Multiple File Output option is not selected, all output will appear in the same file.

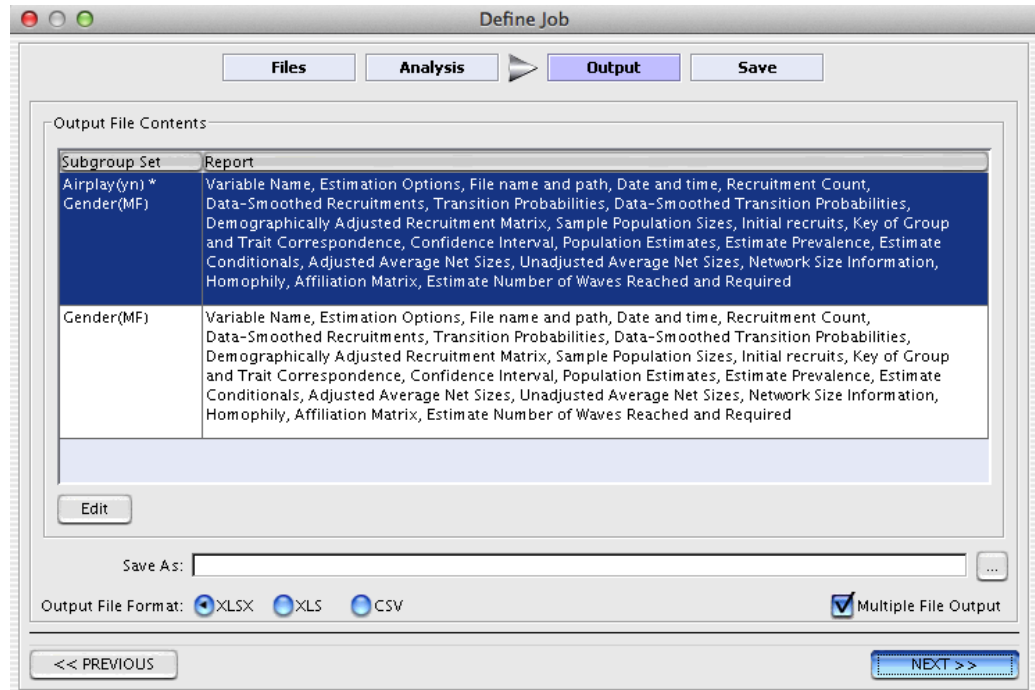


FIGURE 9.8 Job Creation Wizard - Output Specification

Use the [Edit] button to open the “Define Report Profile” window (see Figure 9.9). The report content for each subgroup is specified in the “Define Report Profile” window. Any of the standard RDSAT 7.1 results for each partition can be included or suppressed in the output files. If population or individualized weights are desired, these can be generated as separate files by checking the appropriate box. Report profiles can be saved and reused in other RDSAT 7.1 job specifications, so a standard format can be used for multiple batches or at multiple study sites.

Note:

Individualized and population weight files are automatically named by including the text “_ind-weights” or “_pop-weights” after the file name specified in the Save As field.

Once the Output File Contents are set, click [Next >>] to proceed to the final step.

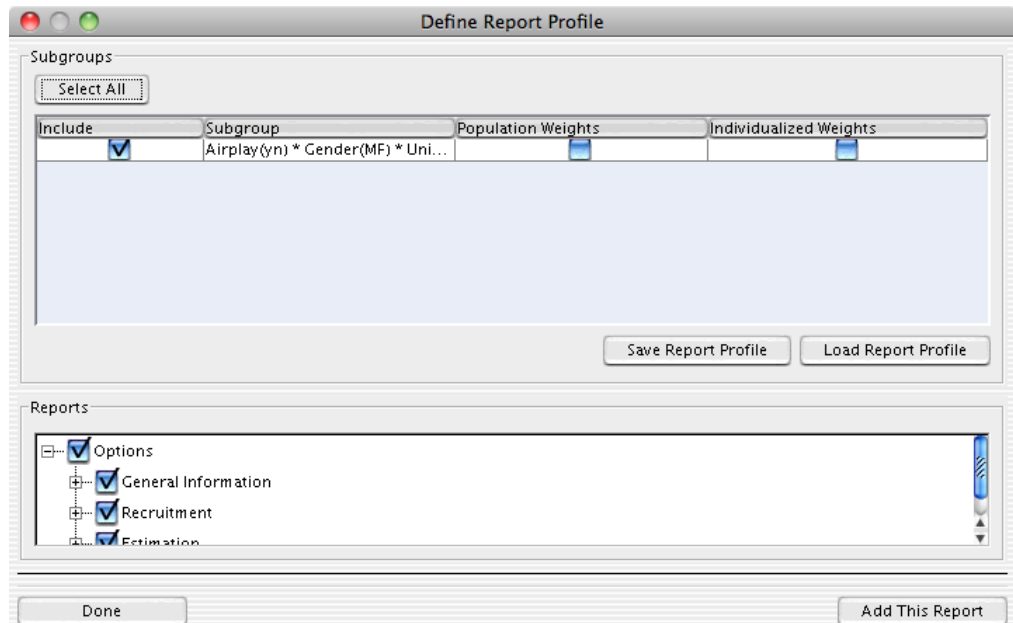


FIGURE 9.9 Job Creation Wizard - Output Specification - Define Report dialog. Output can be specified on a partition-by-partition basis.

4. Save. The final step when defining a job is to save the job definition file. RDSAT 7.1 also offers a chance to generate preliminary analysis without calculating confidence intervals. Since generating confidence intervals is the most time consuming aspect of generating RDS estimates, this quick check feature can be used to confirm adequate cross recruitment and verify report formats before allowing RDSAT 7.1 to begin a longer batch run.

Note

The [Generate Preliminary Analysis without Confidence Intervals] feature attempts to identify potential errors in the job specification or data that would keep the full analysis from finishing. If this feature is used, any errors are reported in the Verification Log (Figure 9.10).

During the Save step of the Define Job dialog (Figure 9.10), specify the job file's save location by clicking the [...] button next to the "Save Job As:" field (see Figure 9.10). Click the [Save] button to save the job to a file without adding the job to the current batch. The Save feature is appropriate when creating a job to be run at another time. To run the job, use [Save and Add to Batch] to save the specified job and add it to the current batch.

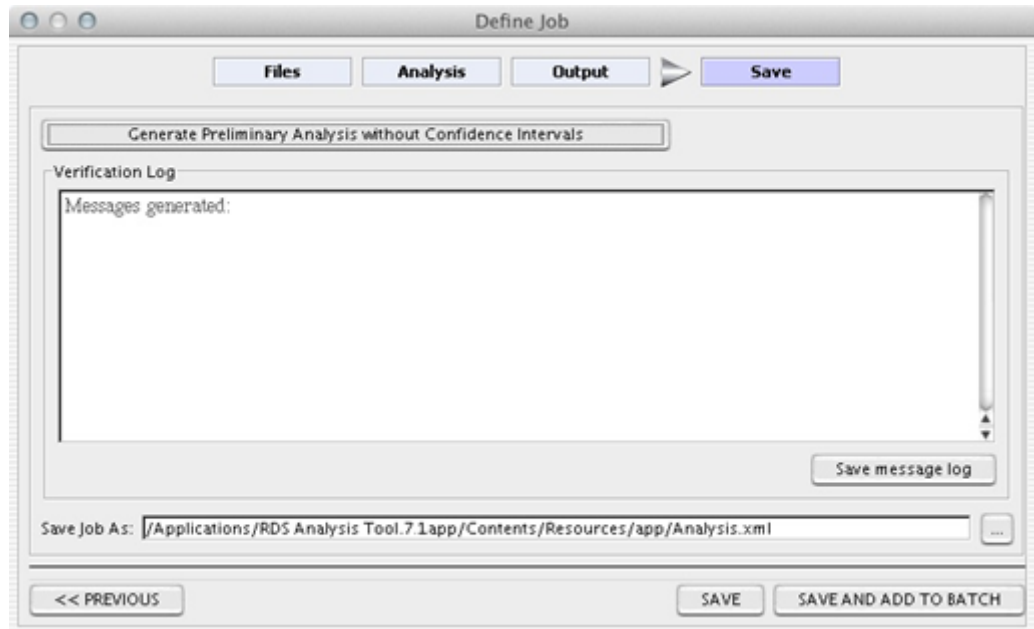


FIGURE 9.10 Job Creation Wizard – Save Job. Jobs may be tested using the “Preliminary Analysis” feature, and jobs may be saved to a batch.

Running a Batch in RDSAT 7.1

When at least one job is specified or loaded from file, use the [Run] button to run the listed jobs (see Figure 9.11).

RDSAT 7.1 will report activity in the Message Log and show elapsed time on the progress bar below the Message Log. A running batch can be canceled by using the [Cancel] button, which replaces the [Run] button when the batch is in progress.

If RDSAT 7.1 reports errors during the batch, see the message log for details about the errors. If RDSAT 7.1 is interrupted during batch processing, use the [View Running Log File] upon re-launching RDSAT 7.1 to see the last messages posted to the message log.

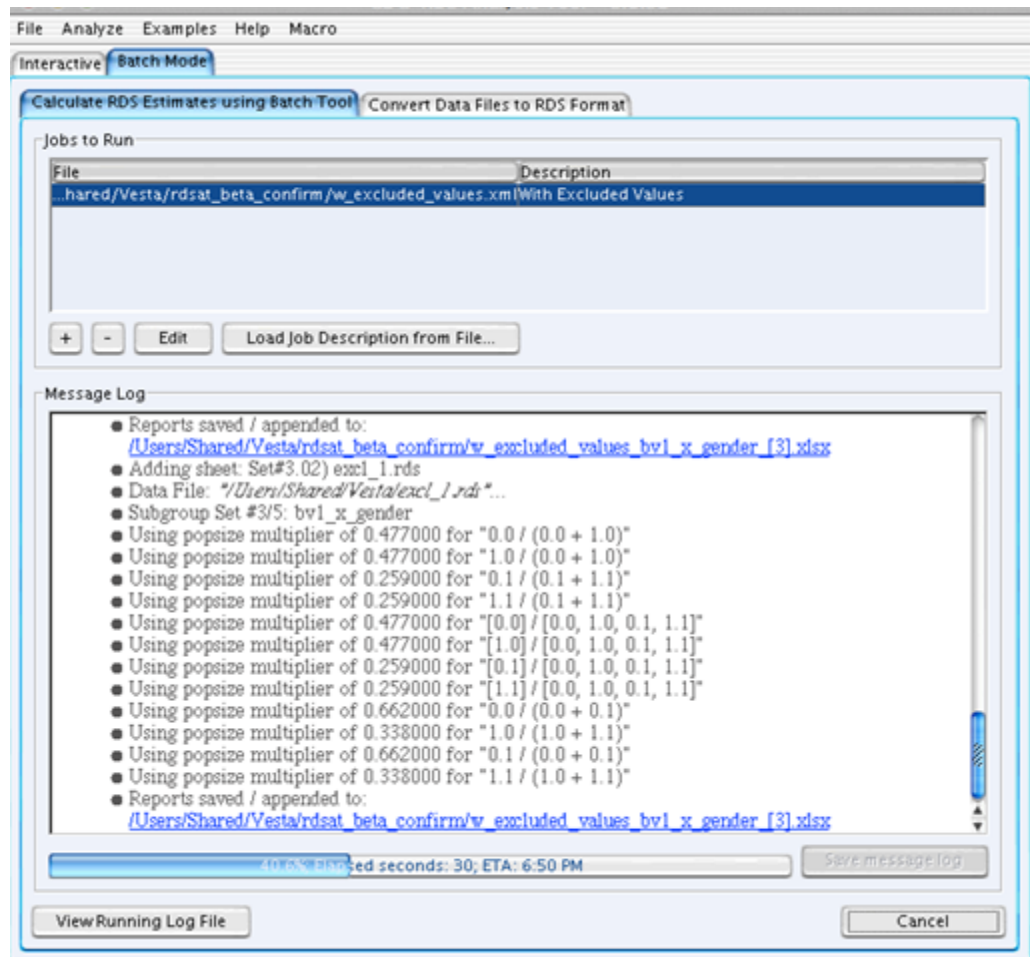


FIGURE 9.11 Job Execution. Specified jobs are listed and executed here; the message log from an executed batch is displayed.

Advanced Subgroup Analysis Features

RDSAT 7.1 includes specialized functionality for advanced users in the Calculate Equilibrium Waves and Prevalence Report sections of the Define Subgroup dialog (Figure 9.12). This section covers the use of these specialized options.

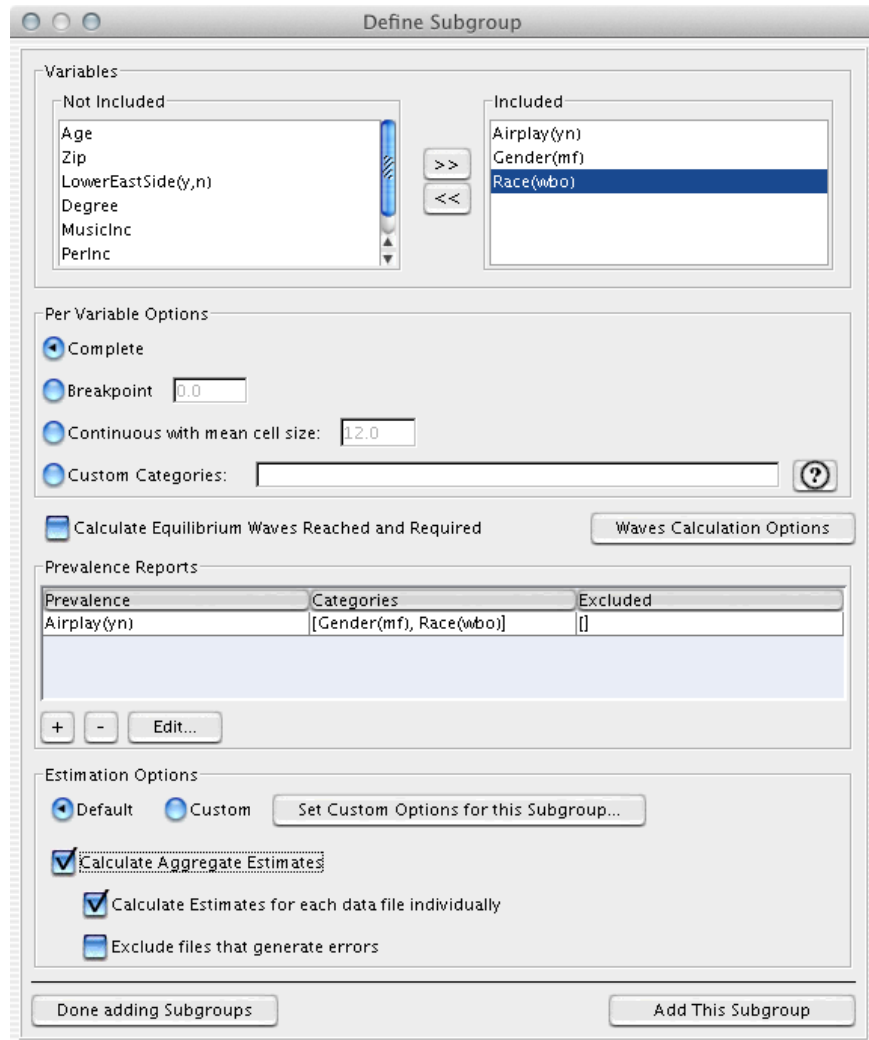


FIGURE 9.12 The Define Subgroup Dialog.

Calculate Equilibrium Waves involves two different calculations. The number of waves reached is a computation of the number of recruiting waves in the sample. The number of waves reached is compared to the number of simulated waves required to reach equilibrium. This is a diagnostic tool used to understand how the particular seeds that generated a sample might have biased the sample. The [Seed Composition Options] dialog can be used to change the seed composition and the algorithm used to calculate the equilibrium waves required (Figure 9.13).

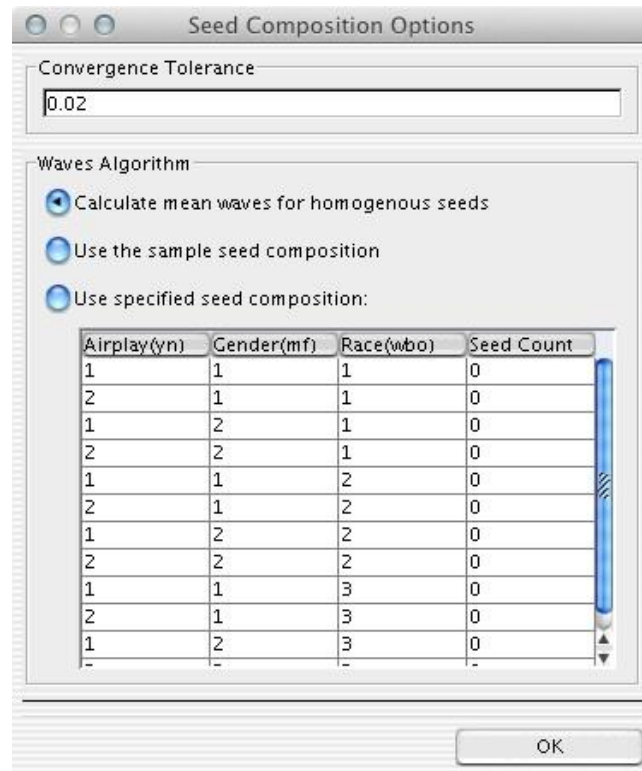


FIGURE 9.13 Seed Composition Options

As successive waves are added to the sample, the sample composition will stabilize. The number of waves required to reach equilibrium is related to the density of cross-cutting ties among the subgroups in the population. The Convergence Tolerance defines the minimum required change in sample proportions between successive waves (see discussion in Chapter 7). When the difference between waves is less than or equal to the convergence tolerance times the population proportion, the simulated sample has reached equilibrium. The Waves Algorithm determines how the equilibrium is generated.

The default option, calculate mean waves for homogenous seeds, calculates several equilibrium scenarios starting with homogenous seeds from each subgroup defined by the partition. The final output includes the minimum, mean and maximum number of waves required to reach equilibrium across all these scenarios and represents a worst, best and average case.

The sample seed composition algorithm uses the distribution of seeds in the data file to calculate the equilibrium. This can be helpful as a diagnostic when seeds are drawn heavily from a few subgroups. When the seed composition is close to the sample equilibrium, the number of waves required may be quite small. In this case the small number of waves is not a good indicator of the subgroup mixing. Analysts should interpret results generated with this algorithm with care.

Finally, it is possible to specify a custom combination of seeds from different subgroups. This feature might be used early in the second year of a study to determine if, given the previously observed recruiting behavior, the proposed seed diversity is adequate to reach sample equilibrium after a few waves.

Prevalence Reports instruct RDSAT 7.1 to calculate the prevalence of the levels of one variable in the partition among the subgroups defined by the rest of the variables in the subgroup. Prevalence Reports are introduced at the end of Chapter 7. Using the Prevalence Report tool allows RDSAT 7.1 to generate sets of prevalence estimates automatically. It is also possible to define multiple sets of prevalence reports using different options within a single subgroup definition. Most uses of the Prevalence Report tool are now better addressed with the Table Builder tool discussed in Chapter 10. The Prevalence report tool is more flexible, so it may be helpful for less standard analyses.

Click the [+] button to create a default prevalence report. The default report is the prevalence of the first variable in the list for the subgroups defined by the remaining variables. The [-] button will delete the selected report. In the example shown in figure 9.7, the prevalence variable is Airplay because it is first on the list, and the subgroups for which prevalence will be reported are the combination of the factor levels for Gender and Race. The excluded column indicates if any variable levels will be excluded from the prevalence report. By default, only those variable levels defined as excluded in the Default Prevalence Options will appear in the excluded column. Note that default excluded values will only appear if present in the files, so verifying that the default excluded values appear as expected will help catch input errors.

Create a customized prevalence report by generating the default report as defined above. Either double click the newly added prevalence report or select the report and click the [Edit...] button to open the Prevalence report dialog (Figure 9.14). The Prevalence window has two top tabs: Assisted and Custom. The Assisted tab will generate the most common types of prevalence reports. The Custom tab can be used to define any possible prevalence report, but the process is quite laborious and not recommended.

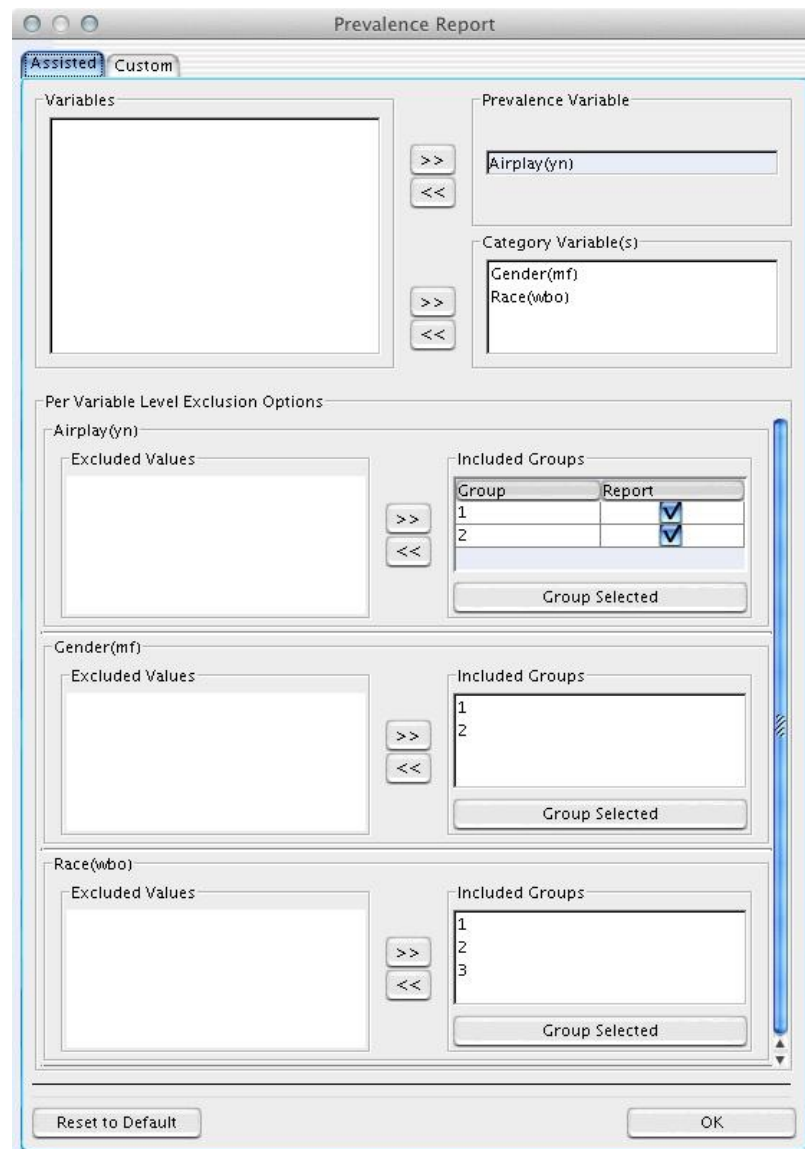


FIGURE 9.14 Prevalence Report tool.

The default report is automatically populated with the variables from the subgroup definition. The population proportions of the levels of the Prevalence Variable levels are estimated within the subgroups defined by the interaction of the Category Variables. In the example shown, the report will include the prevalence of each level of Airplay—yes and no—for each of the six subgroups defined by the cross of Race and Gender.

The lower section of the Prevalence Report window is arranged in rows corresponding to the variables defining the partition. Each row has an Excluded Values list and an Included Groups list. Only the variable values listed in the Included Groups list will be used to generate the prevalence report. The top row

in this section represents the prevalence variable. For a binary variable like airplay, it is sufficient to calculate the prevalence of Airplay=yes, so an analyst might prefer to edit the default settings to report only the Airplay=yes prevalence. In the Airplay: Included Groups list, uncheck group 2 to suppress the prevalence output for this variable level. Variables levels suppressed in this way are still part of the denominator for the prevalence calculation. Suppressing a value reduces the amount of output but does not change the calculated values. The remaining rows reflect the Category Variables.

The prevalence report tool supports excluding values from the calculations. Subgroups composed of at least one excluded value are not included in the denominator of the prevalence so this feature can be used to exclude technically missing data like “Don’t Know,” “Lost” or “Refused.” Any variables entered in the prevalence field of the Default Options on the Define Job: Subgroups dialog (figure 9.6) will be automatically entered in the Excluded Values list. Excluded values can be included in the analysis by selecting the value and clicking the appropriate [>>] button. Some variables may have additional levels that should be excluded for some analyses. For instance transgendered individuals might be excluded when males and females are the only genders of interest. Any variable level can be excluded by clicking on the group and using [<<] to move it to the excluded column.

Numerically small groups can be joined into a single category to improve the estimation outcome. For example, if the proportion of black and “other” races in the sample was small, these two groups could be joined into a single category corresponding to non-white. To join one or more groups, highlight multiple groups by holding the command (Macintosh) or control (Windows) key while clicking in the groups. Click the appropriate [Group Selected] button to join the groups into a single category. RDSAT 7.1 will temporarily internally recode the data when producing the prevalence report. To separate a composite group, use the [<<] to Exclude the composite group. This will restore the original variable values. Move the values back to the Included Groups list by selecting the values and clicking the [>>] button.

The prevalence variable can be changed by selecting the variable in the Prevalence Variable field and clicking the corresponding [<<] button. Repeat the process with the Category Variables. Select a new prevalence variable by clicking a variable name in the “Variables” list and clicking the [>>] button to the left of the Prevalence Variable field. Leave the Prevalence Variable field empty to produce demographic estimates for the groups defined by the Category Variable(s). The prevalence calculations produced are demographic estimates normalized with respect to any excluded variable levels.

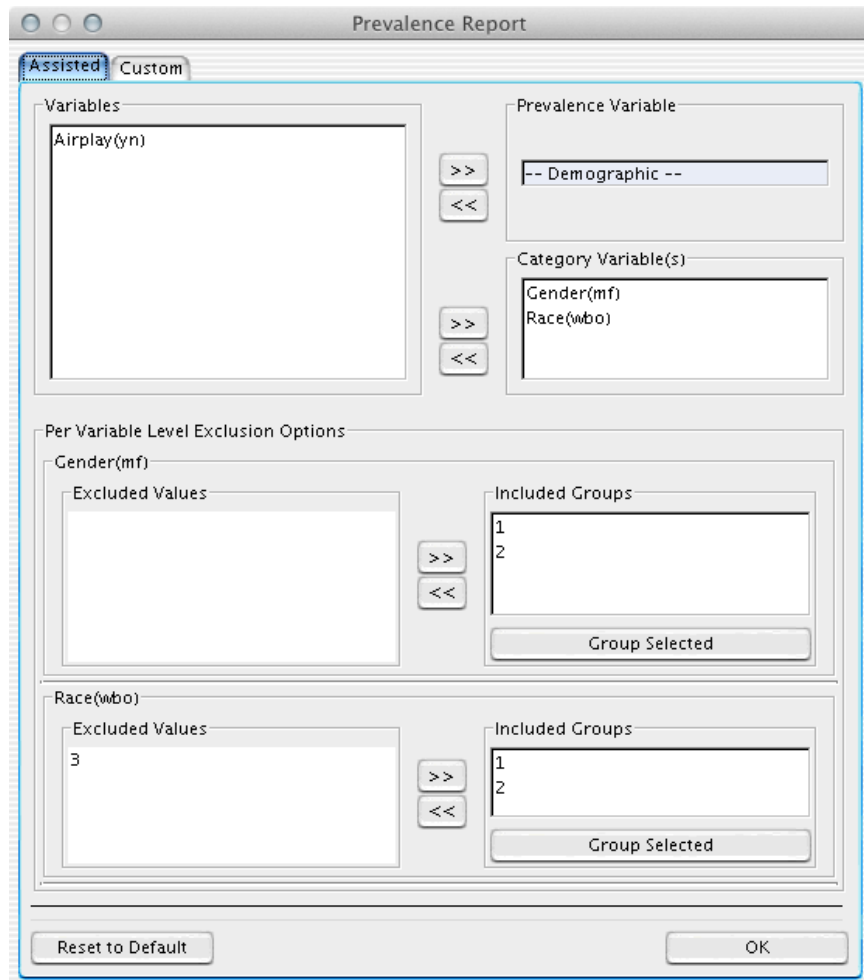


FIGURE 9.15 Prevalence Report tool set to produce demographic estimates.

Click [OK] to accept the report definition and return to the Define Subgroup dialog. The [Reset to Default] button can be used to undo any changes made to the prevalence report. Multiple prevalence reports can be defined for a single subgroup partition by clicking the [+] button multiple times.

10 Batch Mode: Table Builder Tool

RDSAT 7.1 introduces a “Table Builder tool” designed to assist users with specifying the subgroup partitions and prevalence estimates they desire to estimate (see Chapter 7 for a definition and discussion of “Prevalence estimates”; see Chapter 9 for more information on basic estimation with RDSAT Batch Mode).

The Table Builder tool does not add any estimation procedures beyond those available in the Batch Mode. Its purpose is to allow users to more easily create pre-formatted tables containing the sets of estimates most commonly desired for RDS publications.

Note

Because users can estimate multiple related subgroup partitions and prevalence estimates with a single table specification, the Table Builder tool both saves time and decreases the likelihood of errors in the specification process. For these reasons, **we recommend that users employ the Table Builder tool for almost all of their estimation procedures.**

Preparing to Use the Table Builder Tool

The Table Builder tool is part of RDSAT 7.1’s batch mode. The row of tabs below the menu bar labeled “Interactive” and “Batch Mode” are used to switch between operating modes (see Figure 8.2). Batch calculation allows user-defined *jobs* to be executed sequentially with no user interaction. (See Chapter 9 for a discussion of the terms *jobs* and *subgroup partitions*.)

After accessing the “batch mode” tab, users should proceed as with standard batch mode estimation discussed in Chapter 9.

1. Clicking on the “Calculate RDS Estimates using Batch Tool” tab (see Figure 9.3)
2. Adding RDS formatted data files to the job (see Figure 9.4)
3. Specifying the recommended default analysis options (see Figure 9.5)

After these steps have been completed, users are ready to use the Table Builder tool.

Using the Table Builder Tool

After completing the steps listed above, users may choose to change the default table options by clicking the button to [Set Default Table Builder Options] or access the Table Builder tool by clicking the [Table Builder] button below the “Subgroups” field (see Figure 10.1).

Setting the default table options will change the settings shown in Figure 10.8 and determine the columns included in the table and some aspects of table formatting. If changes are made, new tables will use the new default settings.

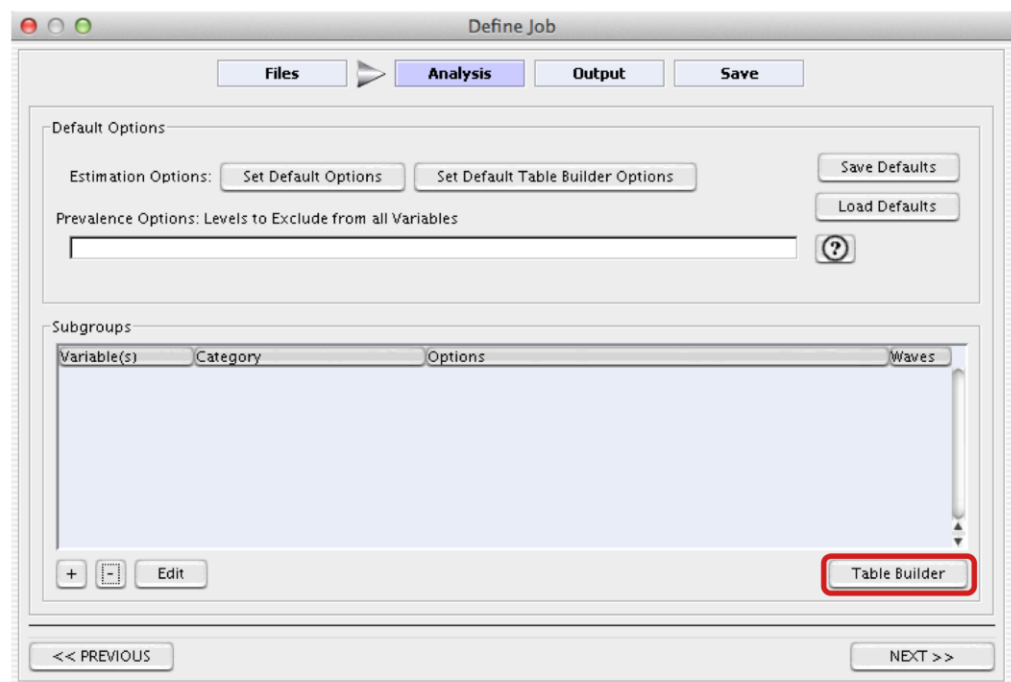


FIGURE 10.1 RDSAT 7.1 Job Creation Wizard – [Table Builder] button on the Subgroup List Screen

After clicking the [Table Builder] button, the Table Builder interface will appear (Figure 10.2). The interface contains five sections of note:

1. “Table Title” field
2. “Variables” list of variables in the jobs data files
3. “Rows: Categorical Variables” list
4. “Columns: Prevalence Variables” list
5. “Options” menu
 - a. “Row Variables” tab
 - b. “Column Variables” tab
 - c. “Table Options” tab
6. Buttons bar

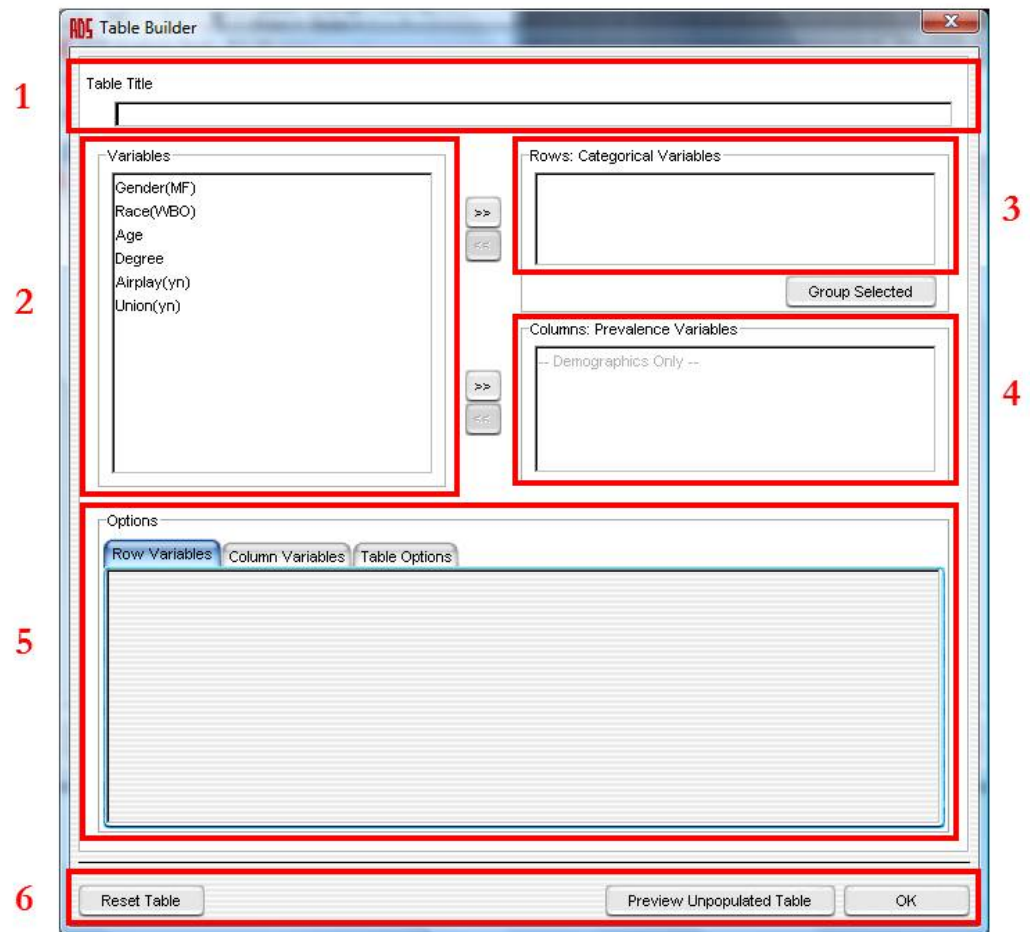


FIGURE 10.2 Table Builder interface – section numbers correspond to list above

The “Table Title” field allows users to specify a title for the table they are making. Generally, this title will contain the names of the variables being estimated and any other information that might be useful to have in the table output.

Note

The Table Builder tool requires that every table have a title. Users may click the [Preview Unpopulated Table] button in the bottom button bar to view the table layout and automatically generate a table title.

The “Variables” list contains a list of all variables available for estimation in the job’s data files.

The “Rows: Categorical Variables” field will contain the categorical row variables a user desires for the table. For example, if a user were estimating airplay prevalence within race groups, the variable representing race groups would belong in this field (see Figure 10.3). Users may move variables from the “Variables” list to the “Rows: Categorical Variables” field by clicking the variable name in the “Variables” list and clicking the [>>] button to the left of the “Rows: Categorical Variables” field.

The “Columns: Prevalence Variables” field will contain the prevalence column variables a user desires for the table. For example, if a user were estimating airplay prevalence within race groups, the variable representing airplay would belong in this field (see Figure 10.3). Users may move variables from the “Variables” list to the “Columns: Prevalence Variables” field by clicking the variable name in the “Variables” list and clicking the [>>] button to the left of the “Columns: Prevalence Variables” field.

As variables are added to the “Rows: Categorical Variables” and “Column: Prevalence Variables” fields, they will appear in the corresponding tab of the “Per Variable Level Exclusion Options” menu (compare Figures 10.2 and 10.3).

The “Button bar” at the bottom of the screen contains three buttons. The [Reset to Default] button resets the Table Builder interface to its empty status upon opening (i.e., it undoes all additions to a table). The [Preview Unpopulated Table] button displays a mockup of the specified table and also automatically generates a Table Title if one has not been specified. The [OK] button closes the Table Builder interface and adds the specified table to the Subgroup List.

Tip

Use the [Preview Unpopulated Table] feature to ensure that the table has been specified correctly without having to wait for the estimation to run (which may be time-consuming and would need to be repeated in case of errors).

The screenshot shows the 'ADS Table Builder' window. At the top, the 'Table Title' is 'Prevalence of Airplay within Race groups'. Below this, there are three main sections: 'Variables', 'Rows: Categorical Variables', and 'Columns: Prevalence Variables'. The 'Variables' list includes 'Gender(MF)', 'Age', 'Degree', and 'Union(yn)'. The 'Rows' section contains 'Race(WBO)'. The 'Columns' section contains 'Airplay(yn)'. There are arrows between these sections to move variables. Below these is the 'Options' section with three tabs: 'Row Variables', 'Column Variables', and 'Table Options'. The 'Row Variables' tab is selected, showing 'Race(WBO)' as the variable. Under 'Analysis Type', 'Complete' is selected. There are 'Excluded Values' and 'Included Groups' lists. The 'Included Groups' list contains '1', '2', and '3'. At the bottom, there are buttons for 'Reset Table', 'Preview Unpopulated Table', and 'OK'.

FIGURE 10.3 Table Builder Tool – Prevalence of airplay within race categories table specification

Users may optionally add more than one variable to the “Rows: Categorical Variables” and “Columns: Prevalence Variables” fields to add additional analyses to the table. Adding the “Gender(MF)” variable to the “Rows: Categorical Variables” field in Figure 10.3 would generate a table containing the prevalence of airplay within race groups and prevalence of airplay within gender groups.

Similarly, adding the “Union(yn)” variable to the “Columns: Prevalence Variables” field in Figure 10.3 would generate a table containing the prevalence of airplay within race groups and the prevalence of union membership within race groups. Users may optionally add multiple variables to each of the “Rows” and “Columns” variable fields.

Excluding and Combining Variable Values with the Table Builder Tool

The “Per Variable Level Exclusion Options” menu in the Table Builder interface (see Figure 10.2) allows users to customize the estimation for each individual variable in either the “Rows: Categorical Variables” or “Columns: Prevalence Variables” fields. Users may customize the estimation in three ways:

- a. Change the “Analysis Type” from the default “Complete” to “Custom”
- b. Exclude “Variable Values” from the table
- c. “Group” selected Variable Values in the table

The analysis types that RDSAT 7.1 supports are described extensively in Chapter 3. Of the four types discussed there, only two are available for variables in the Table Builder: “Complete” and “Custom.” A “Complete” analysis type will treat every variable value as a category for estimation; a “Custom” analysis type allows users to specify how RDSAT 7.1 should convert the raw variable values to categories for estimation.

Sometimes variables contain valid values that users desire to exclude from RDS estimation. For example, users might want to exclude the “Don’t Know” value from an HIV variable containing “Positive”, “Negative”, and “Don’t Know” answers (note that these response categories would be coded as numbers in the actual raw data).

To exclude a Variable Value from the table for a Row variable, first click the “Row Variables” tab in the “Per Variable Level Exclusion Options” menu. Next, select the desired value in the “Included Groups” field, then click the [<<] button to move the value into the “Excluded Values” field (see Figure 10.4). See Chapter 3 for a detailed description of estimation when variable values have been excluded.

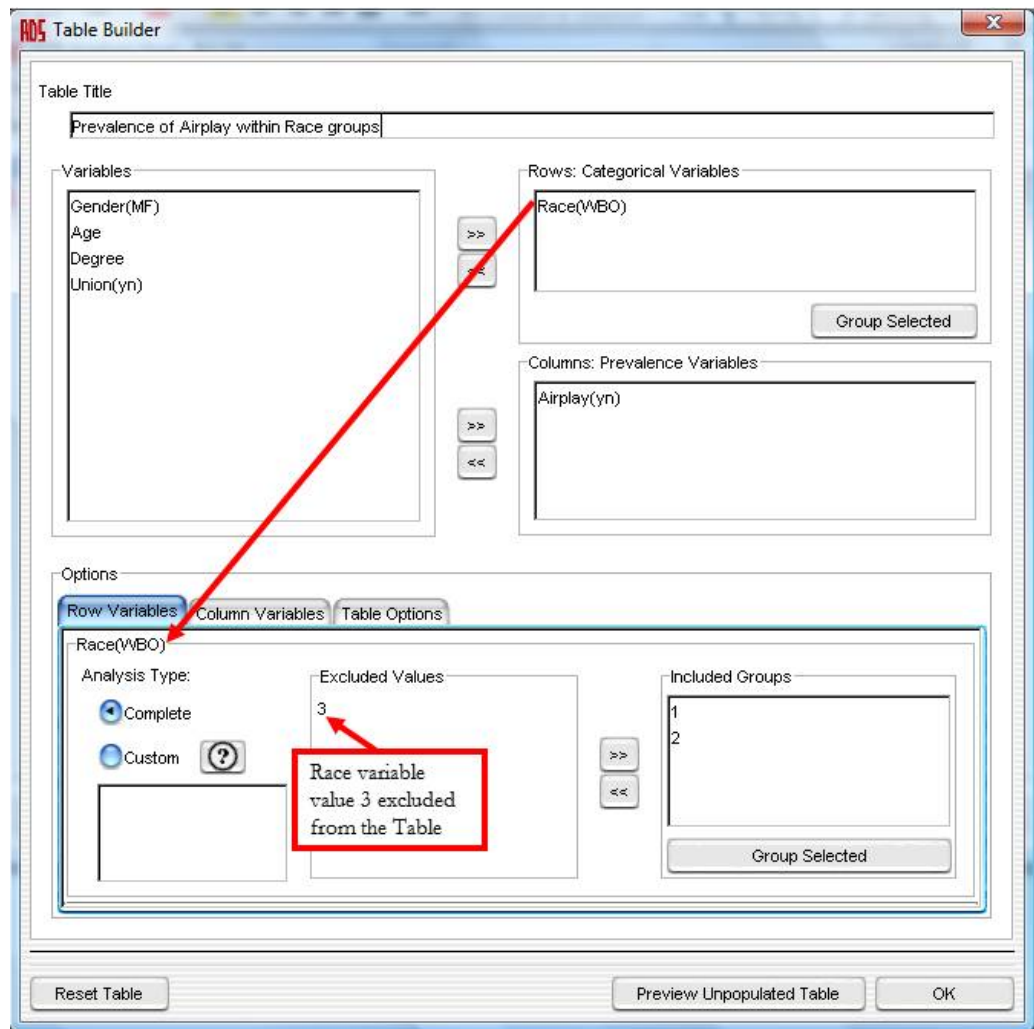


FIGURE 10.4 Table Builder – Prevalence of airplay within race groups table specification with Race variable value 3 excluded from the table.

To exclude a Variable Value from the table for a Column variable, perform the same steps but in the “Column Variables” tab in the “Per Variable Level Exclusion Options” menu.

Notice that the variables in the “Column Variables” tab have tick boxes next the variable values (see Figure 10.5). All values are ticked by default; un-ticking one does not alter the estimation but omits that value’s columns from the table output.

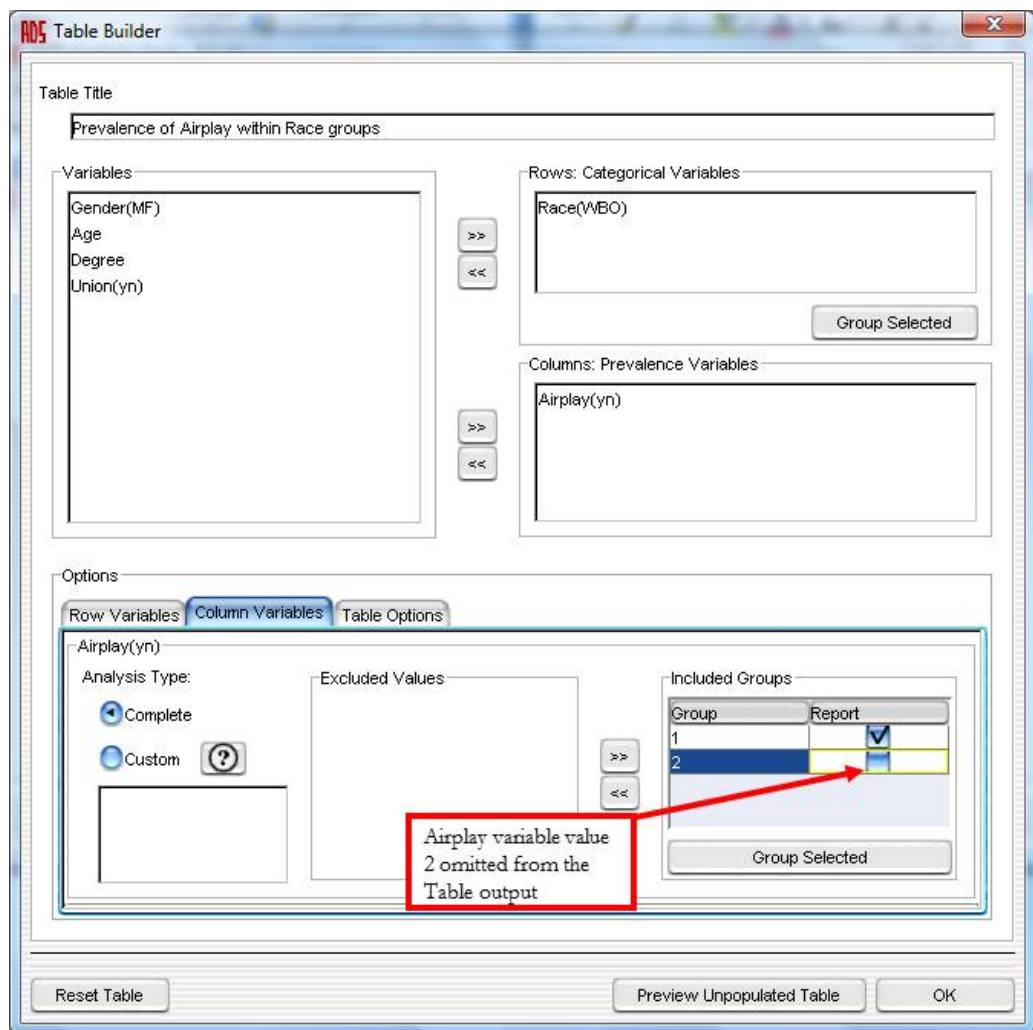


FIGURE 10.5 Table Builder - Prevalence of airplay within race groups table specification with Airplay variable value 2 omitted from the output.

Note that this variable-specific Variable Value exclusion procedure will be performed automatically if users have specified a “Level to Exclude from all Variables” in the Subgroup List screen (see Figure 9.6 for more information).

Sometimes users wish to combine, or “group,” variable values, either for theoretical reasons or because of small sample sizes for the values. This action can be performed by using the [Group Selected] button for the variable in the relevant tab of the “Per Variable Level Exclusion Options” menu. Users should highlight the values they desire to group together in the “Included Groups” menu by holding the “Ctrl” keyboard button (or the command key on a Macintosh) and clicking on the values. Next, click the [Group Selected] button to combine the values into a new estimation group (see Figure 10.6).

Tip

The Table Builder will correctly group variable values for estimation and will document the grouping in its job code and output. However, we recommend that all recoding procedures (including grouping variable values) occur in SAS or another data preparation program before the data is brought into RDSAT 7.1 for estimation.

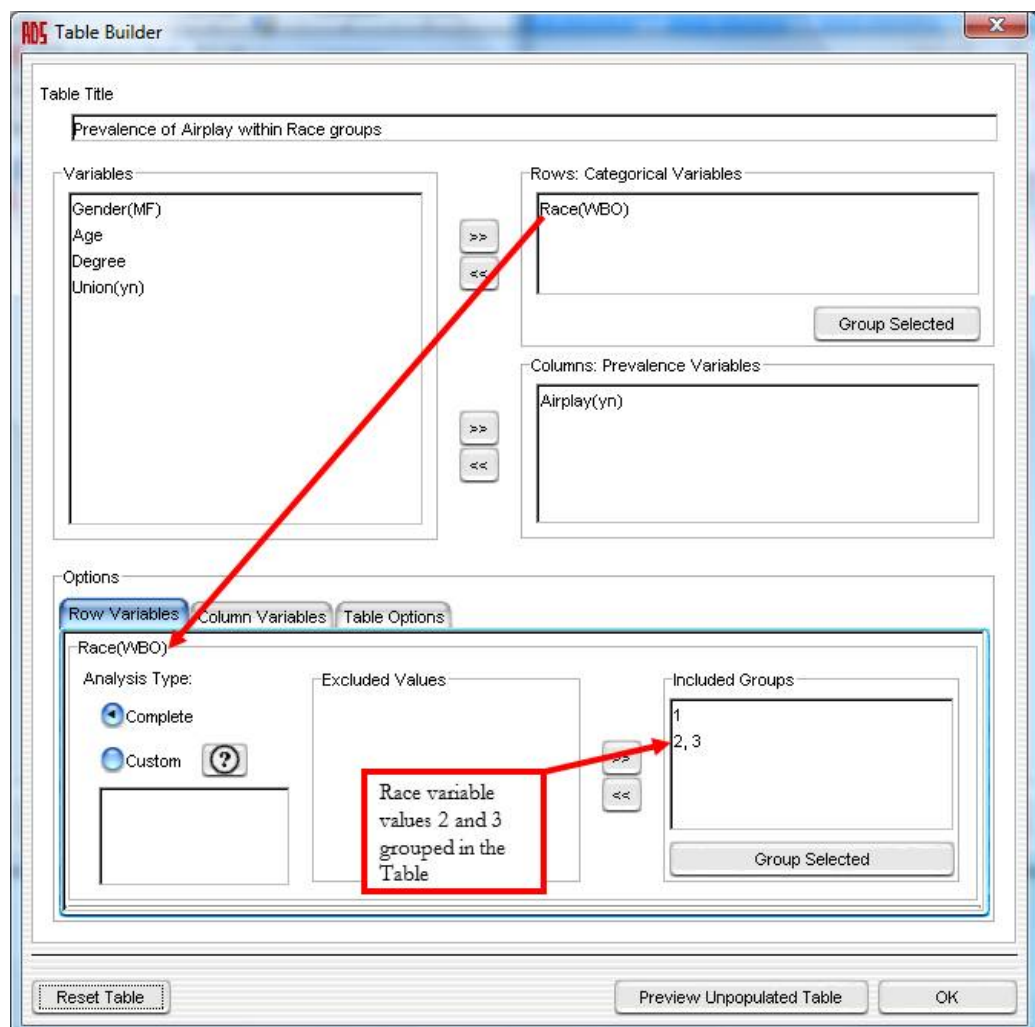


FIGURE 10.6 Table Builder – Prevalence of airplay within race groups table specification with Race variable values 2 and 3 grouped.

Interacting Variables with the Table Builder Tool

Unlike the standard “Subgroup Partition” interface (see Figure 9.7 and discussion), the “Rows: Categorical Variables” and “Columns: Prevalence Variables” fields do not automatically “interact”/“cross” the variables populating the field. To interact two “Row” variables (e.g., race and gender), users should:

- Move both variables to the “Rows: Categorical Variables fields”
- Highlight the variables to interact by holding the “Ctrl” keyboard button and clicking on the desired variables
- Click the [Group Selected Button]

Once these steps have been completed, an interacted variable will appear in the list (in addition to the base variables that were interacted). Note that Column prevalence variables cannot be interacted. See Figure 10.7 below.

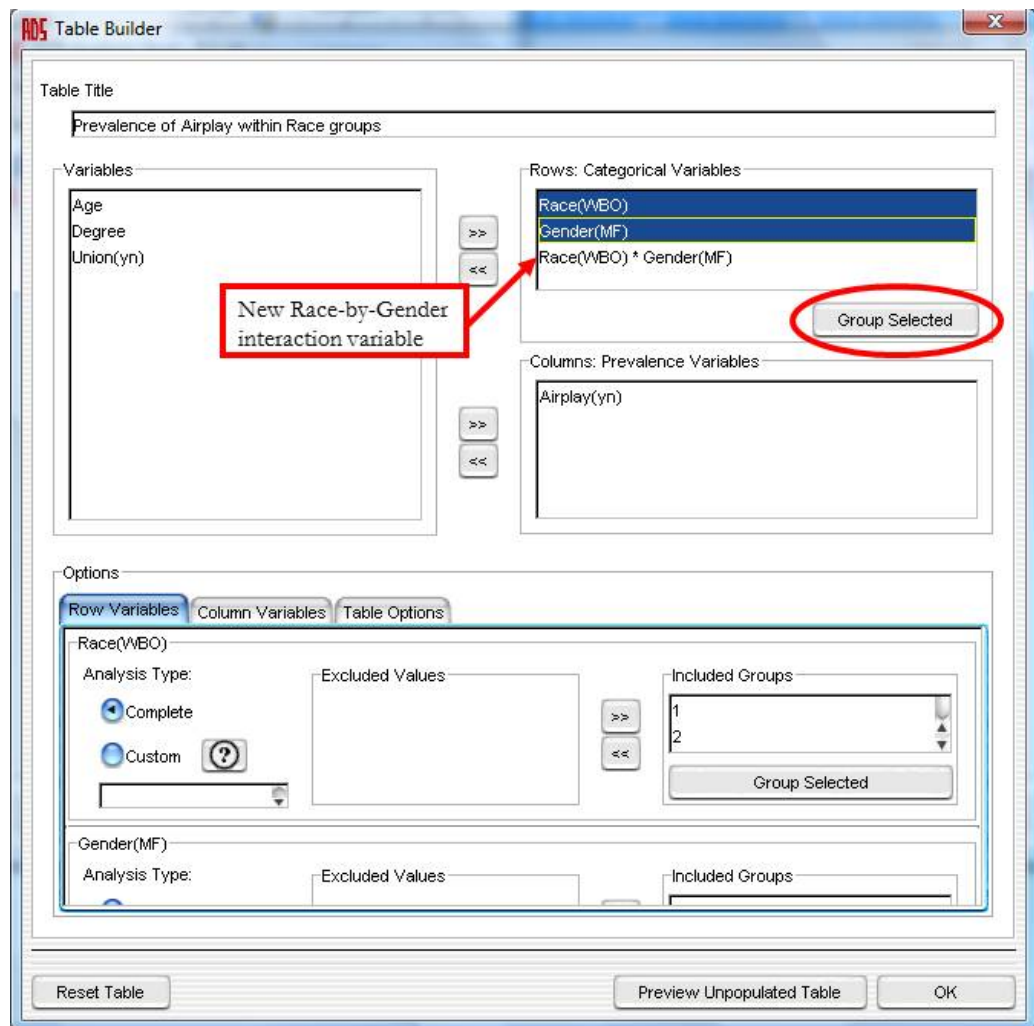


FIGURE 10.7 Table Builder – Prevalence of Airplay within Race, Gender, and Race-by-Gender interaction groups.

Table Options in the Table Builder Tool

The Table Builder tool options are located in the “Table Options” tab in the “Options” menu. The Table Options tab contains three sections (see Figure 10.8):

- a. Output to include in Table
- b. Equilibrium Waves Required
- c. Estimation Options

The “Output Options” section contains tick boxes for each type of output available for tables; users may tick each output item they desire. Each of these output items is described in detail in Chapter 4. When some groups are excluded from row variables, some cells can be normalized by the sizes of the non-excluded groups. Checking the boxes in the normalized column will produce additional columns of output.

The “Confidence Intervals in Separate Columns” tick box determines whether the estimates will appear with both point estimate and confidence interval in a single Excel cell [e.g., “0.502 (.401, .603)”] or each of the point, CI lower bound, and CI upper bound will be placed in their own cells without punctuation (such that calculations can be performed on them directly in Excel).

The “Calculate Equilibrium Waves Reached and Required” section contains a tick box to activate estimation of the equilibrium waves reached and required for homogenous seeds (see the Advanced Subgroup Analysis Section of Chapter 9 for a detailed discussion of this feature). A convergence tolerance of .02 is recommended as the starting point for a waves analysis.

The “Estimation Options” section is identical to the “Estimation Options” section in the “Define Subgroup” menu (Figure 9.5); see Chapter 3 “Setting Options for Analysis” for detailed information about these options. ***Calculate Aggregate Estimates*** is a new feature in RDSAT 7.1. When data files contain a valid population size variable, RDSAT can generate weighted aggregated estimates across multiple data files. The default settings also generate estimates for each site individually, but this can be suppressed in the output if only the aggregate estimates are desired.

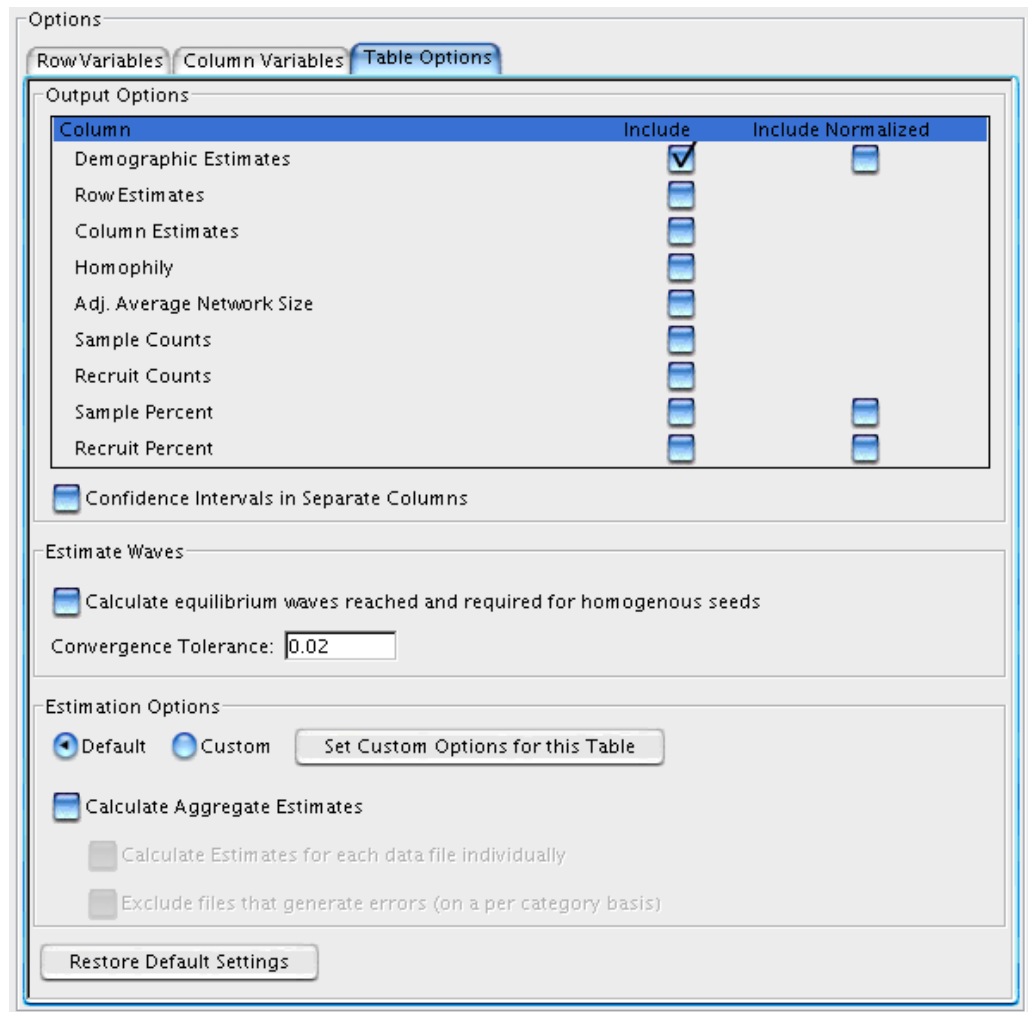


FIGURE 10.8 Table Builder – Table Options menu

After users have specified the table variables, specified the variable-specific options in the “Row Variables” and “Column Variables” tabs, and specified the table options in the “Table Options” tab, they may click the [OK] button on the bottom Button Bar to add the specified table to the job. After clicking [OK], the specified table will appear in the list in the Subgroup List screen (see Figure 10.9).

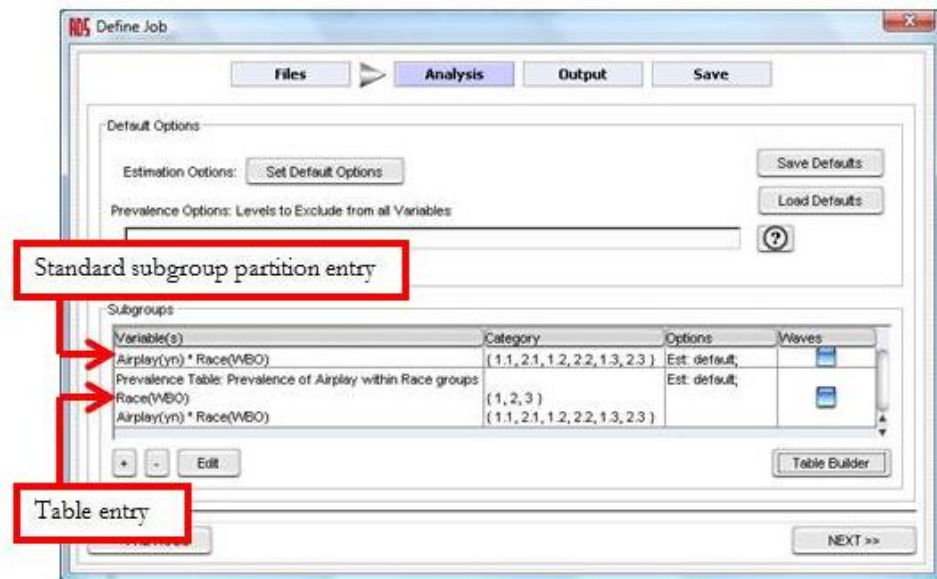


FIGURE 10.9 Define Job Screen – Job with standard subgroup partition and table specified

Once a table has been added to the Subgroup list in the “Define Job” screen, users may proceed with job specification as described in Chapter 9. The standard RDSAT 7.1 batch output will be produced for every subgroup partition included in the job and in the table along with a separate Excel file containing the table and an estimation error log (see below).

Table Builder Tool Output

Each table produced using the Table Builder tool will be contained in its own Excel file containing two tabs: the “Table” tab and the “Errors” tab (see Figure 10.10).

Table output will always contain the “Overall” RDS proportion estimates for every table variable. The overall estimates for the row categorical variable(s) are in the leftmost “Overall” and second-to-left “Normalized Overall” columns (the estimates in these two columns will match unless a row Variable Value has been excluded, in which case the Normalized Totals will not contain an estimate for the Excluded Value). The “Overall” column in Figure 10.11 tells us that 53.1% of the population is estimated to be Race=1, with a confidence interval of (42.3%, 63.4%).

The overall estimates for the column prevalence variable(s) are in the “Overall” row at the bottom of the table. The “Overall” row in Figure 10.11 tells us that 75.1% of the population is estimated to be Airplay(yr)=1, with a confidence interval of (66.1%, 85.3%).

The most commonly used (and included by default) Table Builder estimates are “Demographic”, “Row”, and “Column” estimates. “Demographic” proportion estimates sum to 1 over all of the Column Variable Value columns. The “Demographic / Airplay(yr) / 1” column in Figure 10.11 tells us that 31.0% of the population is estimated to be [Race(WBO)=2 and Airplay(yr)=1], with a confidence interval of (21.8%, 42.5%).

Table Builder “Row” estimates sum to 1 within row over all of the Column Variable Value columns. The “Row / Airplay(yr) / 2” column in Figure 10.11 tells us that 16.2% of [Race(WBO)=3] population members are estimated to be [Airplay(yr)=2], with a confidence interval of (2.8%, 35.2%).

Table Builder “Column” estimates sum to 1 within column over all of the Row Variable Value rows. The “Column / Airplay(yr) / 2” column in Figure 10.11 tells us that 41.3% of [Airplay(yr)=1] population members are estimated to be [Race(WBO)=2], with a confidence interval of (29.3%, 53.7%).

		Overall Row Variable Estimates		Demographic Estimates		Row Estimates		Column Estimates	
Table 1: Prevalence of Airplay within Race groups		Overall	Normalized Overall	Demographic Airplay (yr)	Demographic Airplay (yr)	Row Airplay (yr)	Row Airplay (yr)	Column Airplay (yr)	Column Airplay (yr)
				1	2	1	2	1	2
Race(WBO)									
1		0.531 (0.431, 0.637)	0.531 (0.43, 0.638)	0.346 (0.266, 0.46)	0.171 (0.089, 0.257)	0.67 (0.549, 0.814)	0.33 (0.188, 0.454)	0.462 (0.35, 0.588)	0.684 (0.515, 0.858)
2		0.36 (0.259, 0.463)	0.36 (0.264, 0.469)	0.31 (0.218, 0.425)	0.061 (0.019, 0.1)	0.836 (0.722, 0.946)	0.164 (0.051, 0.269)	0.413 (0.293, 0.537)	0.243 (0.094, 0.395)
3		0.109 (0.061, 0.154)	0.109 (0.063, 0.156)	0.094 (0.049, 0.139)	0.018 (0.003, 0.039)	0.838 (0.652, 0.979)	0.162 (0.028, 0.352)	0.125 (0.066, 0.177)	0.072 (0.011, 0.167)
Excluded		--	--						
Overall		--	--	0.751 (0.66, 0.852)	0.249 (0.148, 0.34)	0.751 (0.662, 0.849)	0.249 (0.151, 0.339)	--	--

FIGURE 10.12 Table Builder Output – Tables tab – Table estimates with no Excluded Values

Table Builder Tool Output – Errors Tab

The “Errors” tab contains the exact same table “shell”, or formatting layout, as the “Tables tab” but does not contain any RDS estimates. Instead, it reports whether there was an estimation error in any cell of the table by placing the word “ERROR” in that table cell (see Figure 10.13).

Estimation error reported for the Race=1, Airplay=1 Column estimate.

	Overall	Normalized Overall	Demographic Airplay(yn)	Demographic Airplay(yn)	Row Airplay(yn)	Row Airplay(yn)	Column Airplay(yn)	Column Airplay(yn)
18 Race(WBO)								
19 1								ERROR
20 2								
21 3								
22 Excluded								
23 Overall								

FIGURE 10.13 Table Builder Output – “Errors” tab

Aggregating estimates across data files with the Table Builder Tool

If the user has added more than one RDS data file to the job and at least 2 data files have a valid “popsize” variable, the “Calculate Aggregate Estimates” tick box in the “Calculation Options” tab of the Table Builder interface will be clickable (i.e., will not be greyed out as in Figure 10.7). See the “Subgroup Partition Options” section in Chapter 9 above for a discussion of cross-file aggregation in RDSAT 7.1.

If the user has selected the “Calculate Aggregate Estimates” option, the “Tables” tab in the Table Builder output will contain one estimated table for every data file along with an additional “Aggregated Table” containing the aggregation of the file-specific table estimates.

Each of the file-specific and aggregated tables will have a counterpart in the “Errors” tab of the Table Builder output. The file-specific tables will display errors as described above, and the Aggregated Table will display either “EXCLUDED” or “ERROR” in every cell that had a file-specific error, depending on whether the “Exclude Files that Generate Errors” option was selected during table specification.

RDS Glossary of Terms

Adjust Average Network Size Option

In a chain referral sample, those with more connections and larger personal network sizes tend to be over-represented in the sample. This option corrects this bias.

Adjusted Average Network Sizes

Network sizes that are adjusted for sampling bias.

Affiliation Matrix

Displays preference measures for connections between all group pairs. The diagonal of this matrix is Homophily within a group.

Bootstrap Simulation Results

Shows the histogram of Bootstrap estimates of Least Squares population proportions. The horizontal axis depicts population estimates for the specified group. The vertical axis shows the frequency of the Bootstrap estimate.

Breakpoint Analysis

A Breakpoint analysis allows one trait to be analyzed over a range of possible breakpoints. This is very useful for continuous variables, such as age.

Complete Variable Analysis

This option will find every distinct value in the data file associated with a variable trait, and create new groups based on that value.

Confidence Interval

The value of this parameter determines the level of confidence for the confidence intervals reported in the analysis. The default, .05, measures the normalized length of a tail of the distribution of population proportions. In short, it determines 90% confidence for the intervals reported in the analysis.

Draw in Outliers

An analysis option that recodes extremely small and large outliers in network sizes from the dataset.

Data-Smoothed Population Proportions

Reports estimated population proportions for the Data-Smoothed population equations.

Data-Smoothed Population Weights

Multiplicative factors by which the Data-Smoothed Estimates are different from the naive estimates.

Degree Distributions

Distribution of network sizes for each group and for the population as a whole.

Degree List

List of all network sizes reported in the sample. The list is sorted from least to greatest for easy view of the distribution.

Demographically-adjusted Recruitment Matrix

Gives hypothetical recruitments if each group recruited with equal effectiveness. Transition probabilities implied by this matrix are identical to those of the original Recruitment Matrix.

DL Network File

DL format is recognized by numerous network analysis packages, including *UCInet* and *NetDraw*. *NetDraw* in particular can be used to create attractive social network visualizations (Appendix 2).

Enhanced Data Smoothing

An option that allows analysis to take place even in a dataset with no recruitment data for a particular group.

Homophily

A measure of preference for connections to one's own group. Varies between -1 (completely heterophilous) and +1 (completely homophilous).

Impute Missing Data and Re-Analyze

Sets missing data to their most probable value, given the transition probabilities.

Initial Recruits

Reports the number of "seeds", i.e. people recruited by the researcher in each group.

Least-Squares Population Proportions

Reports the estimated population proportions of each group using linear least squares to solve the population equations.

LLS Population Weights

Multiplicative factors by which the Least Squares Estimates are different from the naive estimates.

Partition

A user-defined set of groups. Everyone in the population belongs to a group in a partition. The groups are defined by common traits.

Re-Analyze with Specified Missing Data

This feature allows each trait to be chosen and to specify which value the missing data within that trait to have. It can also be used to give missing data a unique value to allow groups to form on the basis of whether they have missing data.

Recruitment Matrix

Matrix of recruitments by and of each group. The vertical axis (rows) depicts the recruiter groups and the horizontal axis (columns) show recruit groups.

Re-samples

This is the number of times random subsets of the data are sampled to derive the bootstrap confidence intervals. More re-sampling will result in better confidence intervals, but will be more CPU intensive.

Respondent

A participant in an RDS sampling study.

Respondent ID

A unique integer representing a respondent in a given RDS dataset.

Sample Population Proportions

The "naive" estimates of population proportions, without correction of over-sampling and other biases.

Sample Population Sizes

The total number of recruits in each group.

Self-Reported Network Size

The number of individuals a respondent reports he or she has in his/her network.

Transition Probabilities

Normalizes recruitments by dividing by the total number of recruitments and gives the probability of one group recruiting another.

Unadjusted Network Sizes

A straight-forward arithmetic mean of the sample's network sizes.

Waves Estimation

This feature allows hypothetical recruitment scenarios to be examined. The sample population proportions are considered converged when the change in population proportions in between waves is less than the convergence radius.

References

Note

Many of these references are available for download online at:

www.RespondentDrivenSampling.org.

- 1) Heckathorn, D. D. (1997) "Respondent driven sampling: A new approach to the study of hidden populations." *Social Problems* 44:174-199.
- 2) Heckathorn, D. D. (2002) "Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations." *Social Problems* 49:11-34.
- 3) Heckathorn, D.D. (2002b) "Development of a Theory of Collective Action: From the emergence of norms to AIDS prevention and the analysis of social structures." In Joseph Berger and Morris Zelditch Jr. (Eds.), *New Directions in Sociological Theory* (pp. 79-108). Oxford: Rowman and LittleField.
- 4) Heckathorn, D. D. and J. Jeffri. (2001) "Finding the beat: Using respondent-driven sampling to study jazz musicians." *Poetics* 28:307-329.
- 5) Heckathorn, D. D. and J. Jeffri. (2003) "Jazz networks: Using respondent-driven sampling to study stratification in two jazz communities." Presented at the *Annual Meeting of the American Sociological Association*. Atlanta, GA. August 2003.
- 6) Heckathorn, D. D. and J. E. Rosenstein (2002) "Group Solidarity as the Product of Collective Action: Creation of Solidarity in a Population of Injection Drug Users." *Advances in Group Processes* 19: 37-66.
- 7) Heckathorn, D. D. Salaam Semaan, Robert S. Broadhead, and James j. Hughes (2002) "Extensions of Respondent-Driven Sampling: A new Approach to the Study of Injection Drug Users." *AIDS and Behavior* 6: 55-67
- 8) Heckathorn, D. D. (2007) "Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment." *Sociological Methodology* 37(1): 151-207.

- 9) Magnani, Robert, Keith Sabin, Tobi Saidel, and Douglas Heckathorn (2005) "Review of sampling hard-to-reach and hidden populations for HIV surveillance" *AIDS Review* 19:67-72.
- 10) Salganik, M. J. and D. D. Heckathorn (2004) "Sampling and estimation in hidden populations using respondent-driven sampling." *Sociological Methodology* 34:193-239.
- 11) Semaan, Saleem, Jennifer Lauby, and Jon Liebman (2002) "Street and Network Sampling in Evaluation Studies of HIV Risk-Reduction

Appendix 1: Frequently Asked Questions

Are there any other essential variables we should be analyzing in RDSAT 7.1 other than gender, race and age?

The variables to be analyzed depend on the research questions being addressed. Recording and analyzing socially salient variables can be helpful for diagnostic reasons, but the selection of these variables requires an understanding of the group under investigation. RDS is a method for drawing statistically valid samples, so its role is to help ensure that the answers are statistically valid.

Are seeds included in the RDSAT 7.1 analyses calculations?

Seeds are not included in RDS estimation of average group network sizes because a member of the study population did not recruit them into the sample. However, respondents recruited *by* seeds do count and therefore the recruitments *by* seeds are included.

One of the respondents in my study said that he has a network size of 0 – how does RDSAT 7.1 handle this?

Because respondents must know at least one person (their recruiter), it is not possible for a respondent to have a valid network size of 0. Additionally, it is highly likely that any respondent who gave a network size of 0 did not understand the network size question. Therefore, respondents with a network size of 0 are assigned the average network size of their group in any given partition analysis.

Note: We do not recommend imputing a network size of 1 for these respondents due to the deleterious impact on standard errors/design effects (and because it is highly likely the respondent did not understand the network size question). See the data appendix of Heckathorn 2007 for more details.

If a participant reports that the person who gave them a coupon is a stranger, what are the implications for the recruitment chains that follow?

In RDS studies, recruitment rights are both scarce and valuable, so respondents tend not to waste them on strangers; recruitment by strangers tends to be rare,

generally 1% to 3%. A reasonable research strategy is to check to see if the respondents recruited by strangers differ significantly from other respondents, and if not, then to treat these as valid recruitments.

How does restricting recruitment to specific races affect the legitimacy of the survey and or RDSAT 7.1 analysis?

This restriction of the sampling frame narrows the scope of the study, e.g., limiting recruitment to Latino IDU would mean that the study would yield no information about non-Latino IDU or Latina IDU. How to best choose the sampling frame depends on the aims of the study.

How does RDSAT 7.1 account for missing data? For example, one of our sites lost 2 interviews (handheld computer malfunction) - one from a seed and the other from a non-seed respondent.

RDSAT 7.1 excludes cases with missing data on a variable-by-variable basis. For variables for which a respondent's data is missing, neither the respondent's network size, the recruitment of the respondent with missing data, nor the recruitments by the respondent with missing data are included in **RDSAT 7.1** calculations. If the respondent only has missing data on some variables, his recruitments will be included for the variables with valid data.

RDSAT 7.1 Interactive mode won't load my data file. Why?

The most common reason **RDSAT 7.1** interactive mode won't load a file is that there is an empty cell or space somewhere in the file. In general, we recommend that all users export their data to a flat file (".txt", ".csv") then use the Import Wizard or Batch Conversion Tool to create properly formatted RDS data files.



Appendix 2: Graphing Recruitment Chains with NETDraw

Graphing recruitment chains can be done using NetDraw, a freely available network graphing program. Graphing an RDS recruitment chain requires 2 different data files:

- 1) The **DL File**, created with RDSAT, contains information on the structure of the chains (who recruited whom).
- 2) The **Attribute File** contains information of the respondents and is created from the RDSAT data file.

The DL File:

- 1) To create the file, load your data into RDSAT 7.1.
Select File → Export DL Network File
Save the file.
- 2) Open NetDraw
- 3) Once you have opened NetDraw (It should say “NetDraw – Visualization Software” at the top, open the DL File you saved by selecting:

File → Open → Ucinet DL text file → Network (1-mode)

Open the DL file you created. You should see a few red dots on the screen.

- 4) To view the recruitment chain select:
Layout → Graph-Theoretic layout → Spring Embedding

Select the following criteria in the popup box:

Layout Criteria: Distances + N.R. + Equal Edge Lengths

Starting Positions: Current positions

No. of iterations: 1000 (If you get overlapping chains, increase this #)

Distance Between Components: 10 (This may need to be adjusted to as high as 20)

Proximities: geodesic distances

Click “OK” and you should see your recruitment chains.

The Attribute File:

The attribute file is VERY similar to the RDS data file. To make it:

- 1) Open the RDS data file with Excel.
- 2) Replace “RDS” with “*node data” in the first line (all lower case, no space between “*” and “node”, 1 space between “*node” and “data”)
- 3) Replace the *sample size* (row 2, column 1) with “ID”
- 4) Delete the columns of Coupon #s (since they are not needed)
- 5) Save the file as a “Tab delimited text file”, **do not overwrite your RDS file.**
- 6) Go back to NetDraw and select
File→Open →VNA Text File→Attributes

In the popup, select the file you just saved and **Select the “Node Attribute(s)” bullet under “Type of Data”**. Click “OK.”

- 7) Your attributes are now loaded.
- 8) NetDraw is almost completely interactive and fairly straight forward to use.
You can control individual nodes by clicking on them or groups of nodes by using the popup menus on the side.

For **example**: select: Properties Nodes Color Attribute based. This will bring up a popup box with a pull down menu with all your attributes in it. Selecting an attribute will color code the node for that attribute.

A detailed discussion of the various features of NetDraw is beyond the scope of this document.



Appendix 3: RDSAT 7.1 Performance Tuning

The RDSAT 7.1 Installer configures RDSAT 7.1 to make optimal use of available RAM and processing power for most jobs. Analyses with many complex partitions may benefit from adjustments to the default settings, particularly if an analysis fails from lack of available ram or insufficient heap space. Performance tuning involves changing the number of threads and the maximum ram allocated to each thread to best accommodate the job. More threads will complete an analysis faster but the RAM available for each thread will be reduced.

These settings are controlled by the virtual machine options. Changing these settings requires administrator access and is not generally recommended. Be aware that 32-bit systems are limited in the amount of RAM that can be allocated, so values in excess of 1G may not work on these systems.

Editing VM Options on Mac OS X:

- 1) Navigate to the RDSAT 7.1.x.app. The default location is /Applications/RDSAT 7.1.x.app. Control-click on the RDSAT application icon and select “Show Package Contents” from the contextual menu.
- 2) Find the file .../ RDSAT 7.1.x.app/Contents/Info.plist in a text editor or plist editor and save a backup copy.
- 3) Open the file .../ RDSAT 7.1.x.app/Contents/Info.plist in a text editor or plist editor.
- 4) Look for the lines (numeric values may differ):


```
<key>VMOptions</key>
<string>-Xmx2g </string> <!-- I4J_INSERT_VMOPTIONS -->
```

and the lines:

```
<key>rds.max.threads</key>
<string>4</string>
```

This is located within the Information Property List/Java level in the XML

- 5) -Xmx controls the amount of RAM available to each thread. In the example above, this value is set to 2 GB, indicated by the '2g' following -Xmx.
- 6) The rds.max.threads is the number of cores RDSAT will attempt to use, in this case, the rds.max.threads is set to 4
- 7) The computer in this example has four cores and 8 gigabytes of RAM, so RDSAT is configured to make full use of these resources (4*2 GB = 8 GB).
- 8) If a job was running out of memory, these settings could be modified to increase the amount of ram to 4GB. This requires a reduction in the number of threads to 2, to keep the total resource use less than or equal to that available (2 * 4GB = 8GB). These settings are reflected in the sample text below:

```
<key>VMOptions</key>
<string>-Xmx4g </string> <!-- I4J_INSERT_VMOPTIONS -->
```

and the lines:

```
<key>rds.max.threads</key>
<string>2</string>
```

This is located within the Information Property List/Java/Properties level in the XML

- 9) Save Info.plist and quit the editor. Relaunch RDSAT for the new settings to take effect.

Editing VM Options on Windows:

- 1) *Note: Users must have administrator privileges for the computer to change the VM Options.*
- 2) Close the RDSAT program.
- 3) Open a text editor program (such as Notepad) with elevated privileges by right-clicking on the icon and selecting "Run as administrator" and clicking "Continue" in the pop-up window.
- 4) In the text editor program, click File -> Open...
- 5) In the Open dialog, navigate to the RDSAT 7.1.x installation folder. The default location is "C:\Program Files\ RDSAT 7.1.x" for 64-bit installations and "C:\Program Files (x86)\ RDSAT 7.1.x" for 32-bit installations.
- 6) In the bottom right of the Open dialog, above the [Open] button, click the drop-down menu and select "All Files (*.*)".
- 7) Select the "rdsat.vmoptions" file and click [Open].
- 8) Each line of text in this file controls a different Java specification.
- 9) "-Xmx" controls the amount of RAM available to each thread. For example, this value might be set to 1 GB, indicated by the '1g' following "-

Xmx". *Note: the "-Xmx" specification does not accept decimals. For example, to allocate 1.5 GB of RAM per thread, one would specify the equivalent 1500 MB of RAM instead as "1500m".*

- 10) "-Drds.max.threads" specifies the number of processor cores RDSAT will attempt to use. For a machine with 2 cores, this might be set to 2 with the "=2" following "-Drds.max.threads".
- 11) By default, RDSAT will use all available cores with the maximum amount of memory per thread such that the number of threads times the amount of RAM per thread is less than or equal to the amount of RAM available on the computer. If a job is running out of memory, the amount of memory per thread can be increased by changing the "-Xmx" specification. However, the number of threads times the amount of RAM per thread must not be greater than the amount of RAM available on the computer.
- 12) Save rdsat.vmoptions and quit the editor. Relaunch RDSAT for the new settings to take effect.