InitialRisk User's Guide

Version 3.2

January 2012

Prepared for:

National Center for
Education Statistics

Prepared by:

WESTAT
Rockville, Maryland

TABLE OF CONTENTS

LIST OF APPENDIXES

LIST OF EXHIBITS

LIST OF FIGURES

1. INTRODUCTION

1.1 Purpose of *InitialRisk*

The question of whether the release of statistical data for public use may lead to the disclosure of the identity of individual units is a long-standing concern. Federal and state agencies are struggling with the need to release study data while protecting the confidentiality of the individuals or institutions included in these data. There are several laws to ensure that information provided by individuals is kept private. These include the Privacy Act of 1974,[1] the Education Sciences Reform Act of 2002, the USA Patriot Act of 2001, and the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA).[2] Failure to protect the confidentiality of individuals in accordance with these laws can result in a fine and/or a prison term, as well as doing irreparable damage to the study and reputation of the Federal agency.

The goal for the Federal agency is to ensure confidentiality while limiting data perturbations in a manner that would minimize the impact on data quality. Some Federal agencies provide guidelines and standardized procedures for ensuring data confidentiality and quality. The National Center for Education Statistics (NCES) has implemented confidentiality standards for all NCES data releases.[3]

The IES confidentiality standards generally require two separate procedures for public-use microdata files: (1) identify any sensitive variables and records through external matching and mask the variables and records that match using directed data swapping, and (2) introduce an additional measure of uncertainty into the data using random swapping. The *DataSwap* software was developed and enhanced in an effort to ease the implementation of both procedures. The *InitialRisk* software is designed to help the user analyze the disclosure risk elements in the data set. The results from the initial risk analysis can be used to guide the disclosure control treatments such as data suppression, recoding, swapping (directed or random), or other data perturbation techniques.

1.2 Software

The *InitialRisk* software is imbedded as an application in the *DataSwap* software so that users may perform an initial risk analysis prior to finalizing data swapping parameters. The process

---

[1] Section 552a protects records maintained on individuals: http://www.justice.gov/opcl/privstat.htm.

[2] Title V of the E-Government Act of 2002, Public Law 107-347 (December 17, 2002). These standards affect all Executive Branch agencies: http://www.eia.doe.gov/oss/CIPSEA.pdf

[3] For the detailed standards, see http://nces.ed.gov/statprog/2002/std4_2.asp.

provides a risk stratum that classifies each record from low to high risk using the counts of violations. The risk stratum can be used in the swapping approach by assigning a larger measure of size for target record selection for higher risk records, or differential swapping rates can be assigned across the risk strata. The standardized software allows for consistency across studies and provides systematic swapping outputs. The software interface and parameter input sheet option makes the software easy to use and share with other organizations.

The *InitialRisk* software is also a stand-alone software application containing the same algorithm as is imbedded in the *DataSwap* software, without any data swapping capabilities. Since the confidential data swapping algorithm is not included in the *InitialRisk*, the software can be more widely distributed than *DataSwap*.

1.3 Method

The methodology for the *InitialRisk* software is simple and extensive. It is simple because it is runs crosstabulations as the basic approach. It is extensive since it processes an exhaustive number of tabulations and captures key information along the way. The basic idea of *InitialRisk* is to identify the attributes or combinations of attributes that make a record different from the others in survey data. The unique or rare cases are associated with high disclosure risk. If a sample case can be uniquely identified by a small number of less detailed attributes, it is even more risky because it is highly likely to be a population unique. Disclosure risk may arise if an intruder intends to identify individuals and disclose their identities or attributes through the matching of known information to external sources.

For forming the tabulations, it is recommended to use factual identifiers such as demographic and geographical variables, etc. Variables with subjective responses, such as cognitive items, are not visible or identifiable by data intruders and would typically be excluded from the extensive multi-way analysis.

Within *InitialRisk*, a limited recoding utility is offered, if requested by the user. The user can recode some non-missing values of a variable to be missing so that these values will not be involved in the subsequent risk analysis. That is, the values will not be included in any combination of attributes to determine sample uniqueness or rareness. After initial recodes are implemented, the next phase of the risk analysis is to process all possible tables of certain dimensions for a specified number of variables. Violations are flagged when table cell counts are less than a given threshold rule – three for example (i.e., Rule of 3). For each category of each variable, the proportion of table cells with violations is computed

among all cells in which a variable is involved. The algorithm counts the number of violations in which a record is involved for the set of tables generated.

This process provides a summary that informs the decisions about the need for further recoding or suppression. Recoding (or collapsing) non-missing values to non-missing values needs to be done outside of *InitialRisk*. The process also provides a risk stratum that classifies each record from low to high risk using the counts of violations. The risk stratum can be used in the swapping approach by assigning a larger measure of size for target record selection for higher risk records, or differential swapping rates can be assigned across the risk strata.

Section 2 describes the parameters and syntax and provides some user guidelines, while Section 3 presents a technical description of the algorithm. Section 4 describes a test dataset that accompanies the software and Section 5 includes an example using these data. Finally, Section 6 describes the Windows-based interface as an alternative way to use *InitialRisk*.

2. INITIAL RISK ANALYSIS

This section describes the user specifications, provides some notes on processing, and gives an overview of the output reports from *InitialRisk*. A description of the *InitialRisk* user specification is provided in Section 2.1. Some processing notes are given and a list of error messages is provided in Section 2.2. Section 2.3 provides an overview of the output reports.

2.1 Specifications

User specifications for the SAS macro can be submitted on a parameter sheet (see Exhibit 2-1).[4] Once the parameters are entered into the program, then the program is invoked through a SAS macro call. The call is detailed in Section 2.1.1. Each parameter, including controllers for input, algorithm, and output, is described in Section 2.1.2. An example of a SAS macro call for *InitialRisk* is provided in Section 2.1.3.

2.1.1 Syntax

InitialRisk is invoked by using a SAS macro call, as follows:

% *InitialRisk* ([macro parameter = value],[macro parameter = value], . . .)

For PC/VAX/LINUX users, the following statements must precede the macro call:

%INCLUDE ‘file name containing the macro statements’;

[4] Alternatively, the Windows-based interface can be used as described in section 6.

Initial Risk Analysis Macro
Parameter specification form

3.2

Study Name:
Charge Number:
Date:
Project Director:
Statistician:

Parameter		*	Entry	Default	Description
Input Controllers	**DATA=**	**R**			input file
	WEIGHT=	**O**			case survey weight - a single variable
	ID=	**R**			case identification - a single variable
Algorithm Controllers	**VARPOOL=**	**R**			initial risk analysis variables – a list of variables delimited by # (maximum number of variables = 20)
	MISSINGDEF=	**O**		**' ' or .**	values to be defined as missing – a list of values separated by #
	MINDIM=	**O**		**1**	minimum dimension of a table
	MAXDIM=	**O**		**2**	maximum dimension of a table
	THRESHOLD=	**O**		**3**	unweighted threshold value to determine violation – a violation occurs if cell count < THRESHOLD
	WGTTHRESHOLD=	**O**		**0**	weighted threshold value to determine violation – a violation occurs if cell sum of weights < WGTTHRESHOLD
	NUMGROUPS=	**O**		**5**	number of risk strata to form
Output Controllers	**OUT=**	**R**			output file
	RISKSTRT=	**O**		**__RISKSTRT**	name of risk stratum variable
	CUTOFF=	**O**		**50**	Shows the top CUTOFF categories of variables that contribute to the highest violation rates

* O: Optional R: Required

Version 3.2, December 2011

Exhibit 2-1. - *InitialRisk* Macro Parameter Specification Form

2.1.2 Parameter Description

The *InitialRisk* macro uses the following parameters.

Input Controllers

DATA = *SAS dataset* specifies the SAS dataset name of the file used in *InitialRisk*. This parameter is required. The user may specify a one- or two-level name (i.e., a temporary or permanent SAS dataset). A one-level name may be specified only if running the macro outside the Windows-based interface. However, if a completed parameter sheet is to be used with the Windows-based interface, the full directory path must be designated (see Section 6.1 for more details.) The dataset must be SAS version 7 or higher (i.e., have a sasb7dat extension) and may contain up to 255 variables.

WEIGHT = *variable* specifies the weight variable to be used to calculate the weighted counts. This parameter is only required if WGTTHRESHOLD is specified.

ID = *variable* specifies one variable to uniquely identify each record. This parameter is required.

Algorithm Controllers

VARPOOL = *list of variables*[5] specifies the categorical variables used to form tables. This parameter is required. The user may specify up to 20 numeric or character variables. The list is delimited by space.

MISSINGDEF = *xvalue1 xvalue2 … xvaluen # yvalue1 yvalue2 … yvaluen # … # zvalue1 zvalue2 … zvaluen* indicates values of VARPOOL to be recoded as SAS missing value, which is **.** for numeric variables and " " for character variables. Character values must be surrounded by quotation marks. This parameter is optional. The "#" entry provides the separator for each VARPOOL variable. If MISSINGDEF is left blank, then no recoding will occur. The recoding is temporary and only lasts until the macro run is finished. In the output file all the variables will keep their original values as in the input file.

[5] Note that each variable must be specifically listed and delimited by spaces. Shortcuts such as VAR1-VAR5 or VAR* may not be used.

MINDIM = *integer* specifies the minimum number of variables that can be used to form tables. This parameter is optional. The default is set equal to 1.

MAXDIM = *integer* specifies the maximum number of variables that can be used to form tables. This parameter is optional. The default is set equal to 2.

THRESHOLD = *integer* specifies the threshold rule that is used to identify violation cells in terms of unweighted counts. If the number of records in a cell is less than THRESHOLD, the records in the cell are flagged as violations, and their violation counts increase by 1. This parameter is optional. The default is set equal to 3.

WGTTHRESHOLD = *constant* specifies the threshold rule that is used to identify violation cells in terms of weighted counts. If the sum of weights in a cell is less than WGTTHRESHOLD, the records in the cell are flagged as violations, and their violation counts increase by 1. This parameter is optional. The default is set equal to 0, which indicates WGTTHRESHOLD will not be applied in the initial risk analysis.

NUMGROUPS = *integer* specifies the number of risk strata to form in terms of the violation counts computed for each record. This parameter is optional. The default is set equal to 5. By default, the records are ranked by descending order of violation counts and grouped into five risk strata, with stratum 0 containing the records associated with the lowest risk and stratum 4 containing the records associated with highest risk. Usually strata 1 through 4 contain similar numbers of records. Unbalanced strata or collapsed strata may result if there are many tied values in violation counts.

Output Controllers

OUT = *SAS dataset* specifies the SAS dataset file name containing the initial risk analysis results. This parameter is required. The file has the same content as the file specified in DATA parameter, except that it contains an additional variable specified by RISKSTRT. The file can have a one- or two-level name (i.e., may refer to a temporary or permanent SAS dataset). A one-level name may only be specified if running the macro outside the Windows-based interface.

RISKSTRT = *variable* specifies the name of the risk stratum variable that is created during the initial risk analysis. This parameter is optional. By default, the risk stratum variable is named as _RISKSTRT. If the user specifies a name that is the same as an existing variable in DATA, that variable in DATA will be overwritten, and no warning will be provided.

CUTOFF = *integer* specifies that the output report will display, for each table dimension, the top CUTOFF variable categories with the highest violation rates. The violation rate for a variable/category is computed as the percent of violation cells in all possible cells involving this variable/category at a given table dimension. This parameter is optional. The default is set equal to 50.

2.1.3 An Example Macro Call

The following is an example of an *InitialRisk* program with example values for each parameter:

```
%InitialRisk(
   DATA=_IN.exampledata,
   ID=CASEID,
   WEIGHT=WEIGHT,
   VARPOOL=BIB1201 BIC0501 BID0101 BIE0601 BORNUSA CENREG DAGE3
           DRACE3 EDUC3 GENDER,
   OUT=_OUT.ira_out,
   THRESHOLD=3,
   WGTTHRESHOLD=0,
   NUMGROUPS=5,
   MINDIM=2,
   MAXDIM=3,
   CUTOFF=15,
   RISKSTRT=RiskStratum,
   MISSINGDEF=NULL#NULL#NULL#5#NULL#NULL#NULL#NULL#NULL#NULL
)
```

The parameter values in this program are interpreted as follows. First the program selects 10 categorical variables to form all possible two-way and three-way tables. If a record has a value of 5 for BIE0601, it will be recoded as SAS missing and not used to form any tables involving BIE0601. The table cells containing fewer than three records are identified as violation cells. The violation count, i.e., the number of times that a record is in any violation cells, is computed for each record. All the records are divided into five risk strata according to their violation counts. The risk stratum variable is named as RiskStratum, and will be included in the output file IRA_OUT. It has five values as 0 through 4. The output report will show the top 15 variable categories with the highest violation rates for table dimensions two and three, respectively.

2.2 Processing

Some system requirements and notes for using *InitialRisk* are described in Section 2.2.1. Section 2.2.2 includes a list of errors that may occur during processing.

2.2.1 System Requirements and Processing Notes

InitialRisk is a SAS macro with a Windows-based interface. SAS version 8 or 9 is required to run the macro. Windows XP and SAS Data Provider 9.12 or higher are required to run the interface (described in Section 6). *InitialRisk* has also been tested and adapted for use on other Windows environments and is available for implementation on VISTA and Windows 7 OS machines. However, the software has been tested in limited VISTA and Windows 7 environments. It should be noted that Vista 32-bit Business version or higher, and Windows 7 Professional version or higher are required to run SAS as well as *InitialRisk*. It is possible that *InitialRisk* may not work in certain customized VISTA and Windows 7 configurations where components may be dropped or missing. Since Windows 7 will become the de facto operating system over the next few years, further testing of *InitialRisk* in various Windows 7 versions and configurations will continue.

The following are some helpful programming notes for using *DataSwap*.

1. Parameters are separated by commas. For example, in % INITIALRISK (VARPOOL=A, THRESHOLD=B,…).

2. Parameters do not have to be in order. They can be specified as %INITIALRISK(VARPOOL=A, THRESHOLD=B) or %INITIALRISK(THRESHOLD=B, VARPOOL=A).

3. SAS shortcut techniques for variable lists do not work. For example, VAR1-VAR3 should be written as VAR1 VAR2 VAR3.

4. All intermediate SAS names created by the program start with a __ (two underscores).

5. The messages generated by *InitialRisk* start with "*InitialRisk* Error" so they can be distinguished from SAS messages.

6. TITLE3, TITLE4, TITLE5, FOOTNOTE1, FOOTNOTE2, and FOOTNOTE3 are reserved for use by *InitialRisk*. Users should avoid these titles/footnotes since they will be overwritten.

2.2.2 Errors

The program aborts with an error message under the following conditions:

1) If DATA, OUT, ID, WEIGHT, or VARPOOL is left blank, for example:

   InitialRisk Error: Parameter DATA is required.

2) If MINDIM, MAXDIM, THRESHOLD, WGTTHRESHOLD, NUMGROUPS, or CUTOFF is specified as a non-numeric value, for example:

   InitialRisk Error: Parameter MINDIM must be numeric.

2.3 Output Reports

The output of *InitialRisk* contains five sections.

First, an information page is provided as a check to ensure parameters have been specified correctly. The information sheet is, in general, a reflection of the parameter sheet. The user should check the specifications of the parameters. The information sheet also provides the macro version # indicating the current version of the software.

Second, two-way frequencies of original VARPOOL variables against their temporary recodes are provided, if recoding is specified in MISSINGDEF for any VARPOOL variables. The frequencies serve as a check to ensure the values specified in MISSINGDEF have been correctly recoded to SAS missing values during the initial risk analysis. A footnote at the bottom of each page in this section indicates that the recodes are temporary and will not be included in the output file. This output summarizes the risk at the record level, which helps the user target records or values with high disclosure risk when SDC treatments are applied.

Third, a printout shows statistics of violation counts by risk stratum. In each stratum, it presents the number and percentage of records in the stratum, as well as minimum, median, maximum, mean, and sum of violation counts. A footnote at the bottom of each page in this section indicates that unbalanced or collapsed strata may be resulted from tied values in violation counts.

Fourth, a printout shows percentage of violations by table dimension, variable, and category of variable for each table dimension. The printout is sorted by ascending table dimension and descending percentage of violations. The percentage of violations for a category of a variable indicates the percentage

of violation cells among all possible table cells which are of a certain dimension and involve this specific variable/category. This output summarizes the risk at the variable level, which helps the user to make decisions on SDC treatments such as suppression or recoding.

Fifth, a printout is provided to show the contents of the output file. The user may use it to check if the output file contains the same information as the input file except for an additional risk stratum variable.

3. TECHNICAL DESCRIPTION

This section provides a technical description of the *InitialRisk* algorithm including the definition of missing values (Section 3.1), formation of tabulations (Section 3.2), identification of violations (Section 3.3) and the interpretation of the risk measures (Section 3.4).

3.1 Definition of Missing Values

In the first phase of *InitialRisk*, the user may recode some values of VARPOOL variables from non-missing to missing through the option MISSINGDEF. *InitialRisk* only examines the tables based on complete cases. In other words, cases with missing values in any of the variables in a table will be excluded. Sometimes answers such as "Don't Know", "Refused", "Inapplicable", etc, are coded into special non-missing values such as 7, 8, 9, -1, 999, … in the survey data. If the user determines that these values do not carry useful information to the intruders and should not be involved in the risk analysis, he may choose to recode these values to be missing in the *InitialRisk* macro. These recodes are temporary and only last until the end of the macro run. The output file still contains the original variables and values.

3.2 Formation of Tabulations

The macro exhaustively forms *m*-way tables using the variables in VARPOOL or the temporarily recoded variables if MISSINGDEF is specified. The lower and upper bounds for the dimension parameter *m* are specified in MINDIM and MAXDIM. The tables contain complete cases only. In total, the macro forms and scans *n* tables, where

$$n = \sum_{m=MINDIM}^{MAXDIM} \binom{p}{m} = \sum_{m=MINDIM}^{MAXDIM} \frac{p(p-1)\cdots(p-m+1)}{m(m-1)\cdots 1},$$

and *p* is the number of variables in VARPOOL.

3.3 Identification of Violations

If the number of cases in a cell is less than THRESHOLD or the weighted cell count is less than WGTTHRESHOLD, this cell is flagged as a violation cell, and the variables/categories (e.g. SEX = 2 and REGION = 3) used to define this cell are identified as "contributing to cell violations." Meanwhile, violation counts, which keep track of the number of times a case has been in any violation cells, of the cases in this cell increase by one.

3.4 Interpretation of the Risk Measures

Two risk measures are summarized in the output report: (1) violation counts by risk stratum; (2) percent of cell violations by variable/category. The first measure indicates the risk level of each data record. The cases in the data set are divided into different risk strata according to the rankings of their violation counts. Stratum 0 contains the riskless cases with zero violation counts. The other strata contain similar number of cases for each risk level containing records with at least one violation. Sometimes risk strata can be combined if the violation counts of their cases are tied. The output report shows some summary statistics, such as minimum, median, maximum, mean, and sum, of the violation counts within each risk stratum. The results can be helpful for applying data perturbation techniques such as data swapping.

The second measure indicates which variables/categories contribute to cell violations more than others. The percent of cell violations for a variable/category is computed as the number of violation cells involving this variable/category divided by the total number of cells formed by this variable/category. In the report, the percent of cell violations is displayed by table dimension. The CUTOFF variables/categories with the highest percentages are shown in descending order of percentages for each table dimension. The results can be useful for determining variable suppression or recoding.

4. TEST DATA

A subset of the 1992 National Adult Literacy Study (NALS) prison study public-use microdata file is used in Section 5 as a test dataset. The values of the variance unit and variance stratum are not those created for that release. The SAS dataset is called EXAMPLEDATA, and it has 20 variables and 182 records. The name, description, values, and value labels for each of the 20 variables are given below.

Variable Information, by Description and Possible Values

Variables	Description	Possible values
BIB1201	Ever been in program to improve basic skills?	1 (Yes), 2 (No)
BIC0501	Ever been placed on probation?	1 (Yes), 2 (No)
BID0101	Do you have work assignments inside or outside?	1 (Yes), 2 (No)
BIE0601	How often write letters/memos in English?	1 (Yes), 2 (No)
BORNUSA	Born in USA?	1 (Yes), 2 (No)
CASEID	Identification No.	ID from "90110104" to "93210309"
CENREG	Census region	1 (Northeast), 2 (Midwest), 3 (South), 4 (West)
DAGE	Age derived from date of birth	values ranging from 17 to 63
DAGE3	Derived age with three categories	1 (DAGE<30), 2 (30<=DAGE<50), 3 (DAGE>=50)
DIC0401	Derived years since admission	values ranging from 0.08 to 17.6
DRACE3	Derived Race/ethnicity with three categories	1 (Hispanic); 2 (NH Black); 3 (Other)
EDUC3	Recoded highest education level with three categories	1:less than high school, 2: high school , 3: >high school
EDUC_DET	Detailed highest level of education received	values ranging from 2 to 11
GENDER	Gender	1 (male), 2 (female)
RATE	Sampling rate	0.05; 0.02
RiskStratum[6]	Risk stratum calculated from the example	0-4
SCORE	Average literacy score	values ranging from 13 to 400
VARSTRAT	Variance stratum	values ranging from 1 to 91
VARUNIT	Variance unit	1 or 2
WEIGHT	Full sample weight	continuous with values ranging from 136 to 1788

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Adult Literacy Study, 1992 Public-use File.

[6] Provided for reference. Running the example as written will overwrite this variable. Alternatively, a name other than "RiskStratum" may be specified by the user to create a new risk stratum variable.

5. EXAMPLE

An example is presented here to illustrate how to use *InitialRisk* to analyze disclosure risks in microdata. The example illustrates the basic features of *InitialRisk* including specifying variables, table dimensions, and threshold rules involved in the analysis, recoding nonmissing values to missing, defining risk stratum variable, and displaying the standard output. The input dataset is the test data (EXAMPLEDATA), which is described in the previous Section 4. The output from the example is provided in appendix A. Appendix B contains the corresponding screen shots for the example from the Windows-based interface.

The example is set up to scan through all possible two-way (MINDIM = 2) and three-way (MAXDIM = 3) tables formed by a list of 10 variables (VARPOOL= BIB1201 BIC0501 BID0101 BIE0601 BORNUSA CENREG DAGE3 DRACE3 EDUC3 GENDER). Table cells with less than three units (THRESHOLD = 3) are identified as sparse cells or violation cells. Value 5 of the fourth variable in the VARPOOL, BIE0601, is recoded to missing so that this value is not involved in the risk analysis. Or in other words, units with BIE0601 = 5 are excluded from the tables formed by BIE0601 and other variables; therefore, BIE0601 = 5 does not contribute to any violations of threshold rules. The units in the input data are divided into five risk strata (NUMGROUPS = 5) in terms of their violation counts. The risk stratum variable is named as RiskStratum and is included in the OUT file.

The filled-in parameter sheet and entire standard output for the example are provided in appendix A. In the output for *InitialRisk* "Percent violations by table dimension, variable, and category of variable" on page A-6, 15 variables/categories (CUTOFF = 15) with the highest percentages of violations are displayed for each table dimension.

6. WINDOWS-BASED INTERFACE

The Windows-based interface provides a means to view the input data file, run frequencies and/or access summaries of the variables, and automatically create the *InitialRisk* parameter sheet and SAS program containing the *InitialRisk* macro call, run the macro, and view output all in the same interface. The sections below describe navigating the various screens in the interface, selecting parameters, running the *InitialRisk* macro, and viewing the output. Please refer to Section 2 for definitions of each of the parameters and their syntax.

6.1 Getting Started

Upon opening the *InitialRisk* Windows-based interface, the user is immediately prompted to select a source file, either a SAS dataset, a prefilled parameter sheet, a SAS program containing the *InitialRisk* macro call, or an MS Access database as shown in Figure 1.

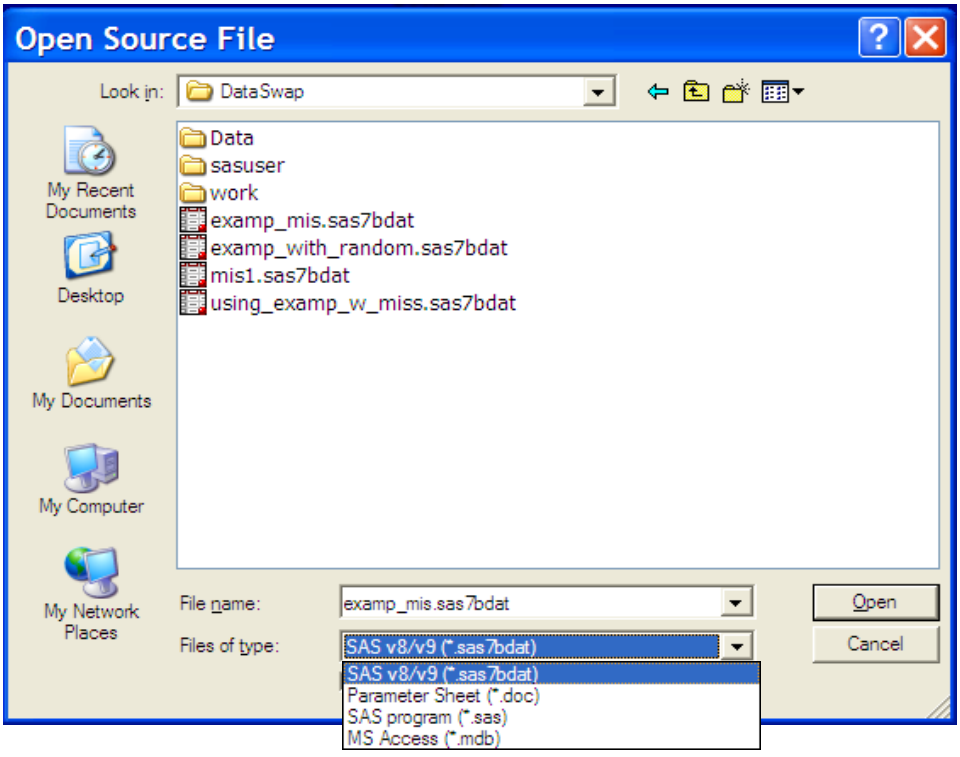


Figure 1 – Source File Selection

If a dataset or database is selected, a screen (see Figure 2) is displayed with the variables on the dataset filling in the *Variable Pool*. If a parameter sheet or SAS program is selected, the same screen appears with all parameters prefilled.

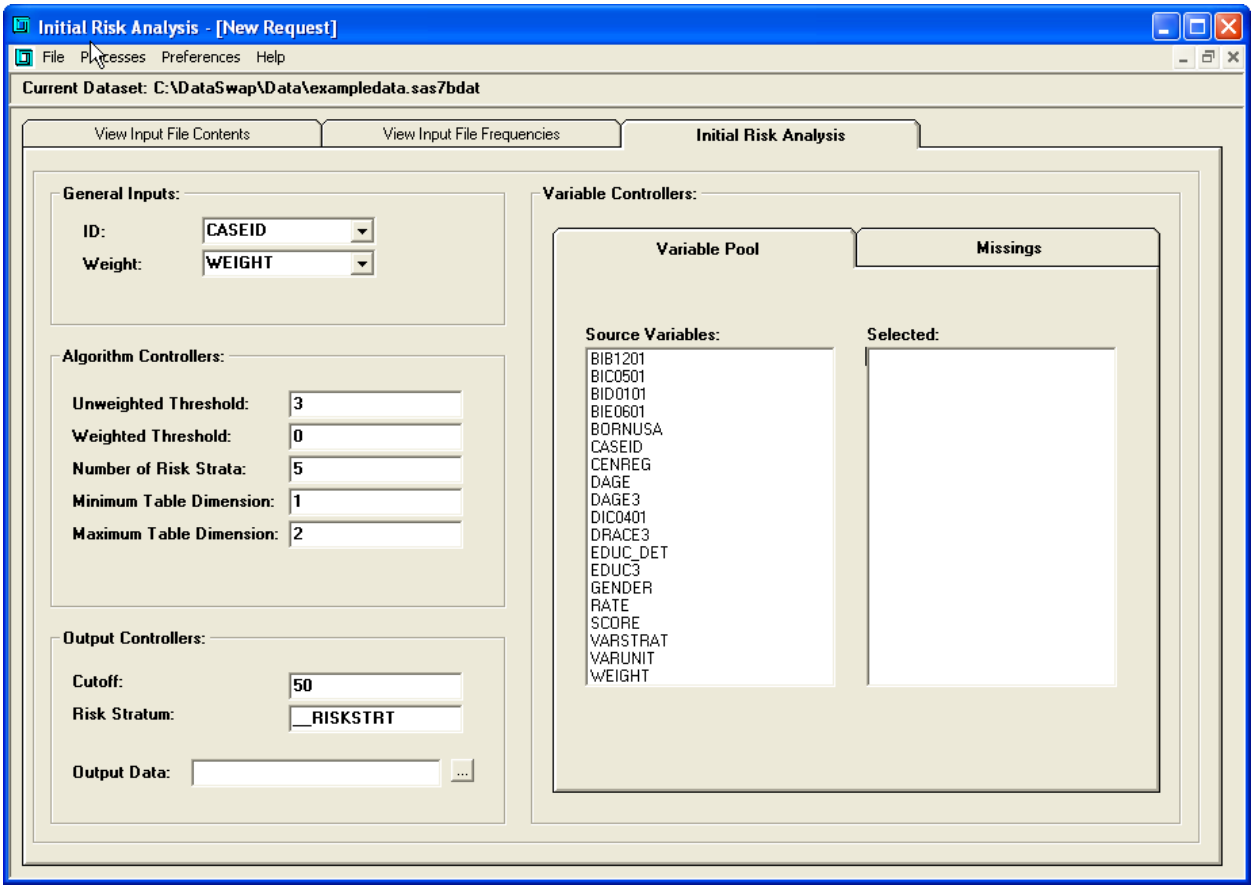
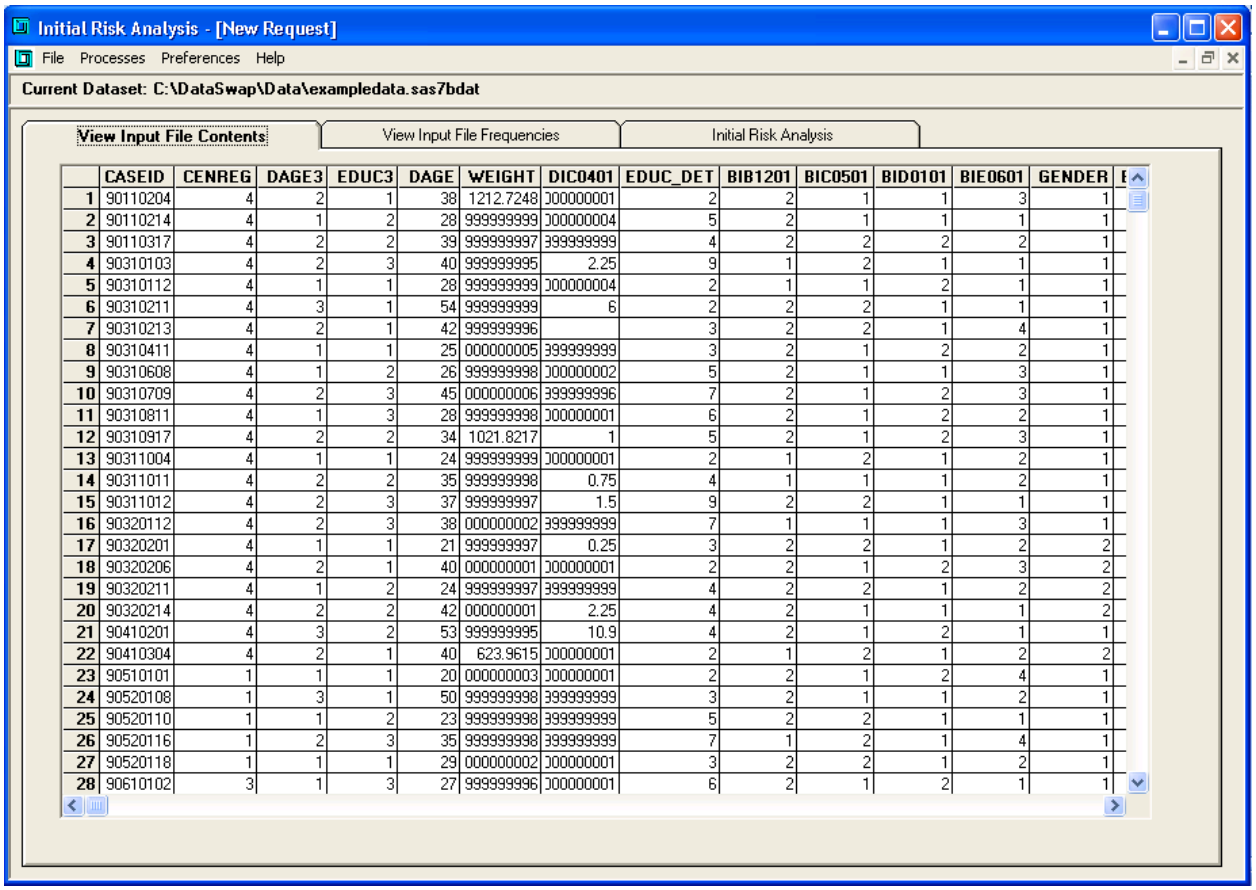


Figure 2 - Specifications Screen

There are three tabs in the interface: the *Initial Risk Analysis* tab, *View Input File Contents* tab, and *View Input File Frequencies* tab.

The *Initial Risk Analysis* screen (see Figure 2) is where each parameter may be specified using the drop-down boxes and the lists provided. Selecting parameters and completing the parameter sheet or SAS program are described in Section 6.2.

The *View Input File Contents* screen (see Figure 3) allows the user to view the values for all variables for each case (observation) in the selected data file. This screen allows the user to verify values of potential parameter variables without having to open a separate SAS session.



	CASEID	CENREG	DAGE3	EDUC3	DAGE	WEIGHT	DIC0401	EDUC_DET	BIB1201	BIC0501	BID0101	BIE0601	GENDER
1	90110204	4	2	1	38	1212.7248	000000001	2	2	1	1	3	1
2	90110214	4	1	2	28	999999999	000000004	5	2	1	1	1	1
3	90110317	4	2	2	39	999999997	999999999	4	2	2	2	2	1
4	90310103	4	2	3	40	999999995	2.25	9	1	2	1	1	1
5	90310112	4	1	1	28	999999999	000000004	2	1	1	2	1	1
6	90310211	4	3	1	54	999999999	6	2	2	2	1	1	1
7	90310213	4	2	1	42	999999996		3	2	2	1	4	1
8	90310411	4	1	1	25	000000005	999999999	3	2	1	2	2	1
9	90310608	4	1	2	26	999999998	000000002	5	2	1	1	3	1
10	90310709	4	2	3	45	000000006	999999996	7	2	1	2	3	1
11	90310811	4	1	3	28	999999998	000000001	6	2	1	2	2	1
12	90310917	4	2	2	34	1021.8217	1	5	2	1	2	3	1
13	90311004	4	1	1	24	999999999	000000001	2	1	2	1	2	1
14	90311011	4	2	2	35	999999998	0.75	4	1	1	1	2	1
15	90311012	4	2	3	37	999999997	1.5	9	2	2	1	1	1
16	90320112	4	2	3	38	000000002	999999999	7	1	1	1	3	1
17	90320201	4	1	1	21	999999997	0.25	3	2	2	1	2	2
18	90320206	4	2	1	40	000000001	000000001	2	2	1	2	3	2
19	90320211	4	1	2	24	999999997	999999999	4	2	2	1	2	2
20	90320214	4	2	2	42	000000001	2.25	4	2	1	1	1	2
21	90410201	4	3	2	53	999999995	10.9	4	2	1	2	1	1
22	90410304	4	2	1	40	623.9615	000000001	2	1	2	1	2	2
23	90510101	1	1	1	20	000000003	000000001	2	2	1	2	4	1
24	90520108	1	3	1	50	999999998	999999999	3	2	1	1	2	1
25	90520110	1	1	2	23	999999998	999999999	5	2	2	1	1	1
26	90520116	1	2	3	35	999999998	999999999	7	1	2	1	4	1
27	90520118	1	1	1	29	000000002	000000001	3	2	2	1	2	1
28	90610102	3	1	3	27	999999996	000000001	6	2	1	2	1	1

Figure 3 - *View Input File* Contents Screen

The *View Input File Frequencies* screen (Figure 4) allows the user to run frequencies or access summary statistics on any variables in the selected file. The user may run frequencies by selecting the *Show Frequencies* radio button and then selecting one or more variables for which frequencies are desired (by either dragging and dropping or double-clicking the variable name). Selecting more than one variable results in a crosstabulation of the variables. It is suggested that users examine the frequencies of VARPOOL variables before running *InitialRisk*. Selecting the *Show Summary* radio button allows the user to view the average, minimum, maximum, and total sum of variable values for one or more variables. Only one frequency may be shown on the screen, but summaries for more than one variable may be viewed simultaneously.

To change the dataset or import a new parameter sheet file, the user must first close the current source information by selecting *File* ➔ *Close*. If parameters have changed while using the source file, the user is prompted to save the changes. Selecting *File* ➔ *Open* opens a new source file using the screen shown in Figure 1.

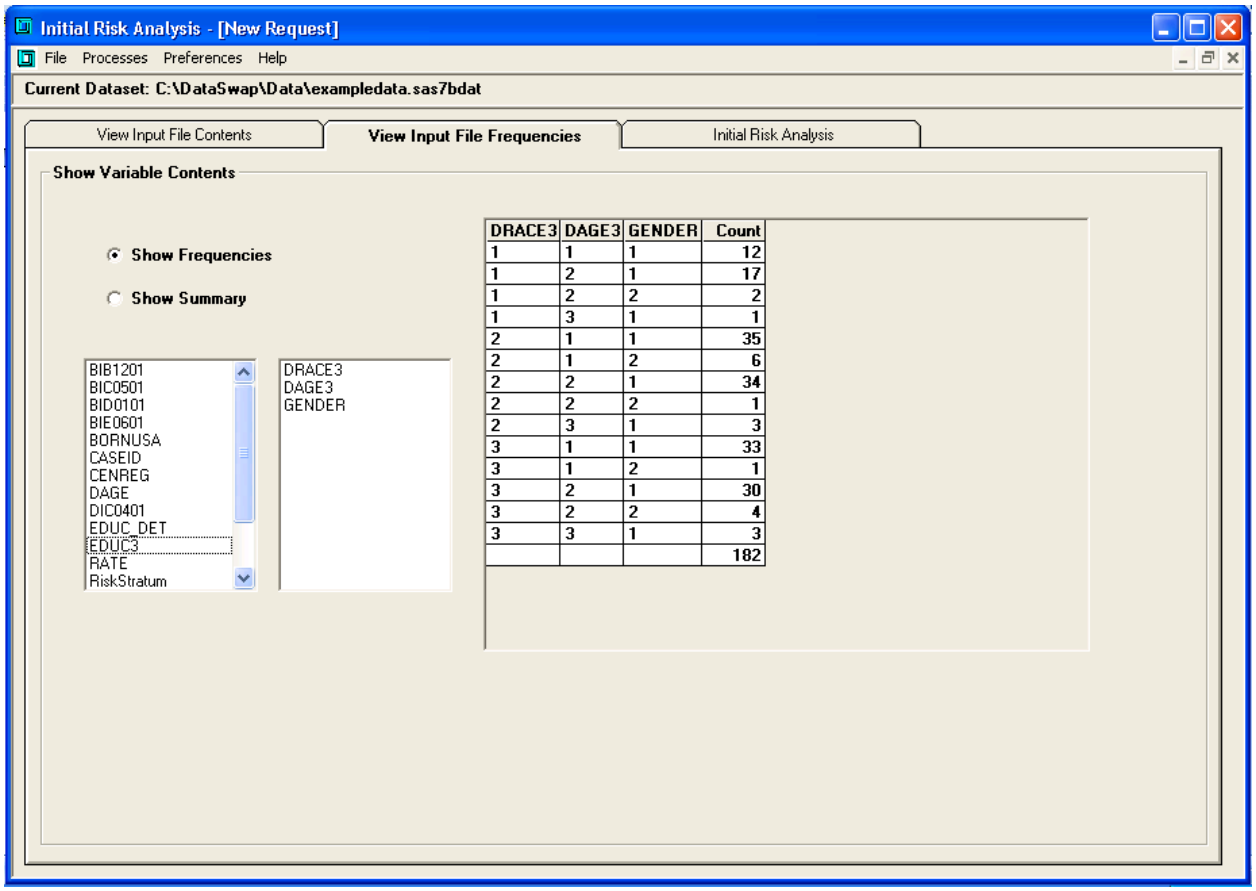


Figure 4 - *View Input File* Frequencies Screen

6.2 Initial Risk Analysis

6.2.1 Specifying Parameters

If a parameter sheet or SAS program is not imported upon opening *InitialRisk*, the parameters for processing can be specified on the *Initial Risk Analysis* screen shown in Figure 2. With the exception of VARPOOL and MISSINGDEF, all parameters may be entered by either selecting variables from the drop-down boxes or by typing variable names and values directly.

If a parameter sheet resides in a different directory from the dataset, the input and output datasets must be specified with the appropriate directory path such as *C*:\<*data path*>\<*dataset name*>; it is not necessary to include the file extension. If the dataset and the parameter sheet are in the same directory, only the dataset name is required.

The VARPOOL parameters may be specified by double-clicking on the desired variable name(s) in the displayed list of source variables. The selected variables will be shown in the *Selected* window (see Figure 5).

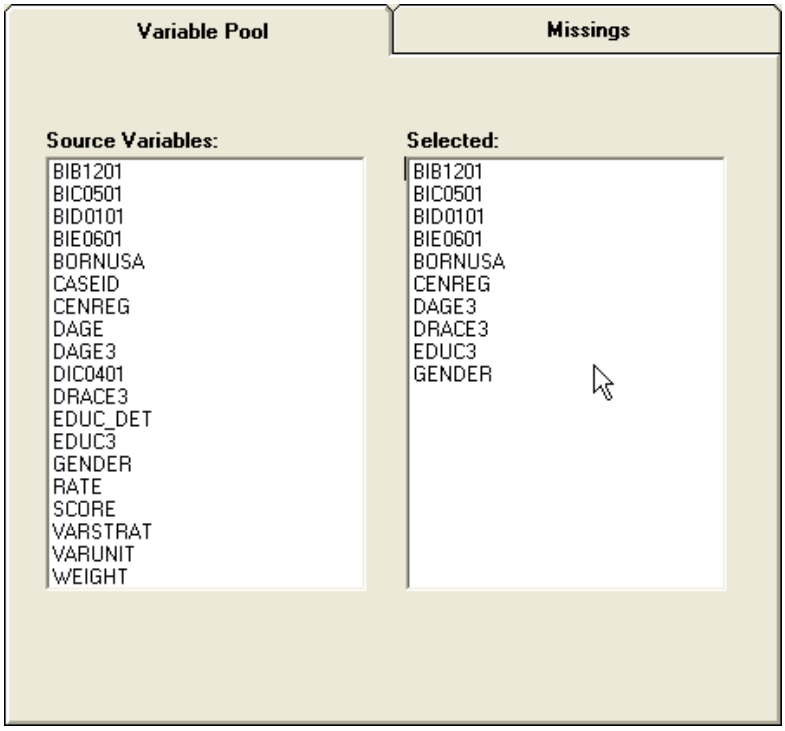


Figure 5 – *Variable Pool* Tab

The user may change the MISSINGDEF parameters on the *Missings* tab in the *Variable Controllers* section (see Figure 6). The user may select the variable to be recoded by clicking it in the *Selected Variables* window. The possible values of the selected variable are then shown in the *Values* window in the middle. Next the user may double click the value that needs to be recoded as missing. Then this value will be moved from the *Values* window to the *To be recoded* window on the right. If the user would like to recode another variable, he may repeat this process. He can also check the recodes at any time by clicking the variable name in the *Selected Variables* window.

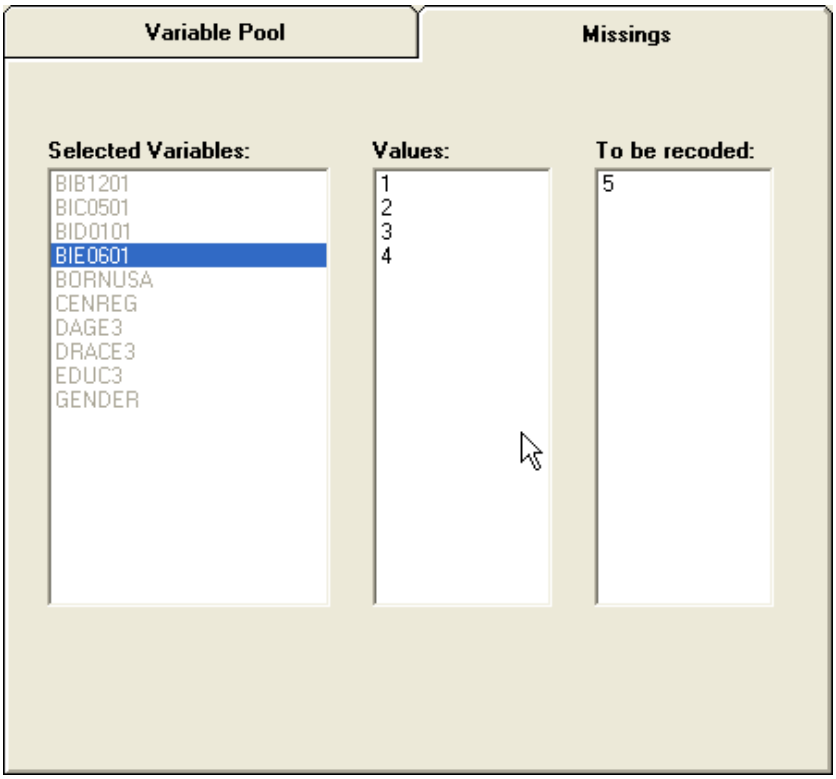


Figure 6 – *Missings* Tab

Once all the parameters have been specified, a parameter sheet can be created by selecting *File* ➔ *Save As* ➔ *Initial Risk Parameter Sheet* from the menu. Alternatively, the user can have the interface generate SAS code immediately before running the program as described in the next section.

6.2.2 Running the Program

InitialRisk can be run one of three ways once the parameters have been specified in the *Initial Risk Analysis* screen. One way is to use the parameter sheet generated from the Windows-based interface and then use the SAS macro as shown in Section 2.

Another way to run *InitialRisk* is to generate SAS code from the interface by selecting *File* ➔ *Save As* ➔ *Initial Risk SAS Code* from the menu and save the code as a SAS System program. (The default location for this program is the same location as the input dataset.) The SAS program can then be run outside the interface in SAS.

A third way to run *InitialRisk* is directly from the Windows interface by selecting *Processes* ➔ *Initial Risk*. After running the software from the interface, one of two messages will be displayed.

If the program was successful, a message such as that shown in Figure 7 will be displayed indicating the location of the .log and .lst file (a SAS program file with a .sas extension and the same prefix is also created). TextPad (text editing software) is the default program the macro uses to display the output; if TextPad is not installed, Notepad is used. The output contained in the .lst file is identical to that created when running the macro outside the interface.

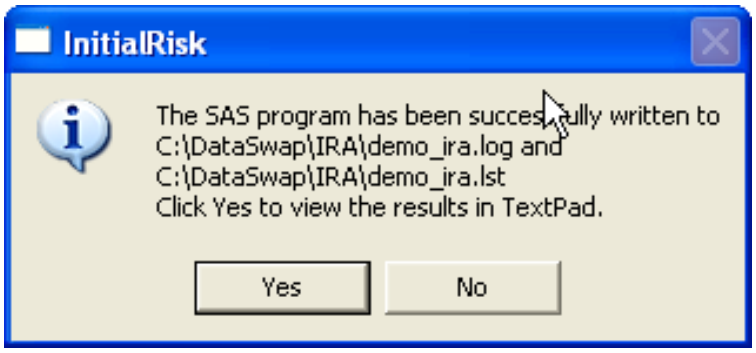


Figure 7 – Successful Completion Message

If the program does not run successfully, the message in Figure 8 is displayed. If the user selects *Yes*, the .log file is displayed so that the user can view any messages that may clarify why the program did not run successfully.

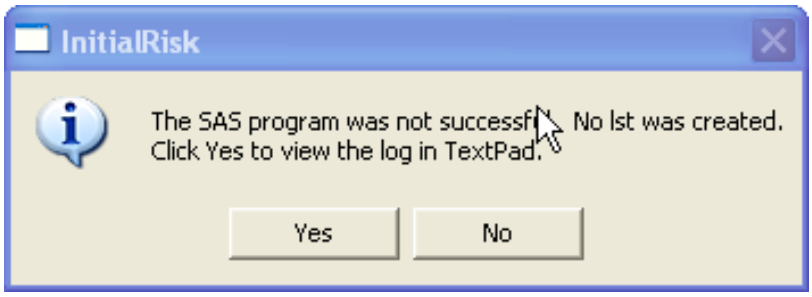


Figure 8 – Unsuccessful Completion Message

Clicking *Yes* on either of the screens shown in Figure 11 or 12 does not close the interface or change any of the parameters on the *Specifications* screen.

It is recommended that the user rename any .sas, .log, and .lst files that were created using the default naming convention if they expect to refer back to these runs at a later time. The macro eventually reuses these default names after several runs, so these default-named files may be overwritten by newer files.

APPENDIX A

Example Parameter Sheet and Standard Output

Initial Risk Analysis Macro
Parameter specification form

3.2

Study Name:

Charge Number:

Date:

Project Director:

Statistician:

Parameter		*	Entry	Default	Description
Input Controllers	**DATA=**	**R**	\\westat.com\DFS\DATASWAP\Y2011\Demo\Data\exampledata		input file
	WEIGHT=	**R**	WEIGHT	**1**	case survey weight - a single variable
	ID=	**R**	CASEID		case identification - a single variable
	VARPOOL=	**R**	BIB1201 BIC0501 BID0101 BIE0601 BORNUSA CENREG DAGE3 DRACE3 EDUC3 GENDER		initial risk analysis variables – a list of variables delimited by # (maximum number of variables = 20)
	MISSINGDEF=	**O**	NULL#NULL#NULL#5#NULL#NULL#NULL#NULL#NULL#NULL	**' ' or .**	values to be defined as missing – a list of values separated by #
Algorithm Controllers	**MINDIM=**	**O**	2	**1**	minimum dimension of a table
	MAXDIM=	**O**	3	**2**	maximum dimension of a table
	THRESHOLD=	**O**	3	**3**	unweighted threshold value to determine violation – a violation occurs if cell count < THRESHOLD
	WGTTHRESHOLD=	**O**	0	**0**	weighted threshold value to determine violation – a violation occurs if cell sum of weights < WGTTHRESHOLD
	NUMGROUPS=	**O**	5	**5**	number of risk strata to form
Output Controllers	**OUT=**	**R**	\\westat.com\DFS\DATASWAP\Y2011\Demo\IRA\ira_out		output file
	RISKSTRT=	**O**	RiskStratum	**__RISKSTRT**	name of risk stratum variable
	CUTOFF=	**O**	15	**50**	Shows the top CUTOFF categories of variables that contribute to the highest violation rates

* O: Optional R: Required

Version 3.2, December 2011

INITIAL RISK REPORT
THE INFORMATION PAGE

INITIALRISK MACRO VERSION #:	1.0
INPUT DATA SET:	_IN.exampledata
OUTPUT DATA SET:	_OUT.ira_out
CASE IDENTIFICATION:	CASEID
CASE WEIGHT:	WEIGHT
ANALYSIS VARIABLES (VARPOOL):	BIB1201 BIC0501 BID0101 BIE0601 BORNUSA CENREG DAGE3 DRACE3 EDUC3 GENDER
MISSINGDEF:	NULL#NULL#NULL#5#NULL#NULL#NULL#NULL#NULL#NULL
MINDIM:	2
MAXDIM:	3
THRESHOLD:	3
WGTTHRESHOLD:	0
NUMGROUPS:	5
RISKSTRT:	RiskStratum
CUTOFF:	15

INITIAL RISK REPORT
Check temporary recodes

The FREQ Procedure

Original BIE0601	Recoded BIE0601	Frequency
1	1	53
2	2	66
3	3	28
4	4	20
5	.	15

Note: The recodes are temporary and will not be included in the output data set

INITIAL RISK REPORT
Statistics of violation counts by risk stratum: threshold=3

InitialRisk: Risk Stratum	N	Percent	Minimum	Median	Maximum	Mean	Sum
0	81	44.51	0	0.00	0	0.00	0
1	25	13.74	1	1.00	1	1.00	25
2	29	15.93	2	3.00	4	3.00	87
3	22	12.09	5	6.00	8	6.32	139
4	25	13.74	9	15.00	42	18.60	465

The risk strata are assigned using PROC RANK. Unbalanced groups or collapsed groups may be resulted from many tied values in counts or violations.

INITIAL RISK REPORT

_OUT.ira_out: Percent violations by TabDim, Var, Level: threshold = 3, weighted threshold = 0

Number of variables involved in tables	Variable	Category of variable	Proportion of cells with violations of Rule of 3
2	DAGE3	3	0.55
2	GENDER	2	0.26
2	BORNUSA	2	0.21
2	BIE0601	4	0.14
2	CENREG	2	0.13
2	DRACE3	1	0.13
2	BIE0601	3	0.09
2	CENREG	1	0.09
2	EDUC3	2	0.08
2	EDUC3	3	0.08
2	BIB1201	1	0.08
2	BID0101	2	0.08
2	BIE0601	1	0.04
2	BIE0601	2	0.04
3	DAGE3	3	0.77
3	GENDER	2	0.57
3	BORNUSA	2	0.53
3	BIE0601	4	0.51
3	BIE0601	3	0.35
3	BIB1201	1	0.35
3	CENREG	1	0.33
3	EDUC3	3	0.32
3	DRACE3	1	0.31
3	CENREG	2	0.26
3	CENREG	4	0.26
3	EDUC3	2	0.25
3	BID0101	2	0.25
3	CENREG	3	0.22

INITIAL RISK REPORT
Contents of the output dataset - _OUT.ira_out

The CONTENTS Procedure

Data Set Name	_OUT.IRA_OUT	Observations	182
Member Type	DATA	Variables	20
Engine	V9	Indexes	0
Created	Monday, December 12, 2011 03:57:59 PM	Observation Length	176
Last Modified	Monday, December 12, 2011 03:57:59 PM	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	16384
Number of Data Set Pages	3
First Data Page	1
Max Obs per Page	92
Obs in First Data Page	72
Number of Data Set Repairs	0
Filename	C:\DataSwap\IRA\ira_out.sas7bdat
Release Created	9.0202M3
Host Created	XP_PRO

Alphabetic List of Variables and Attributes

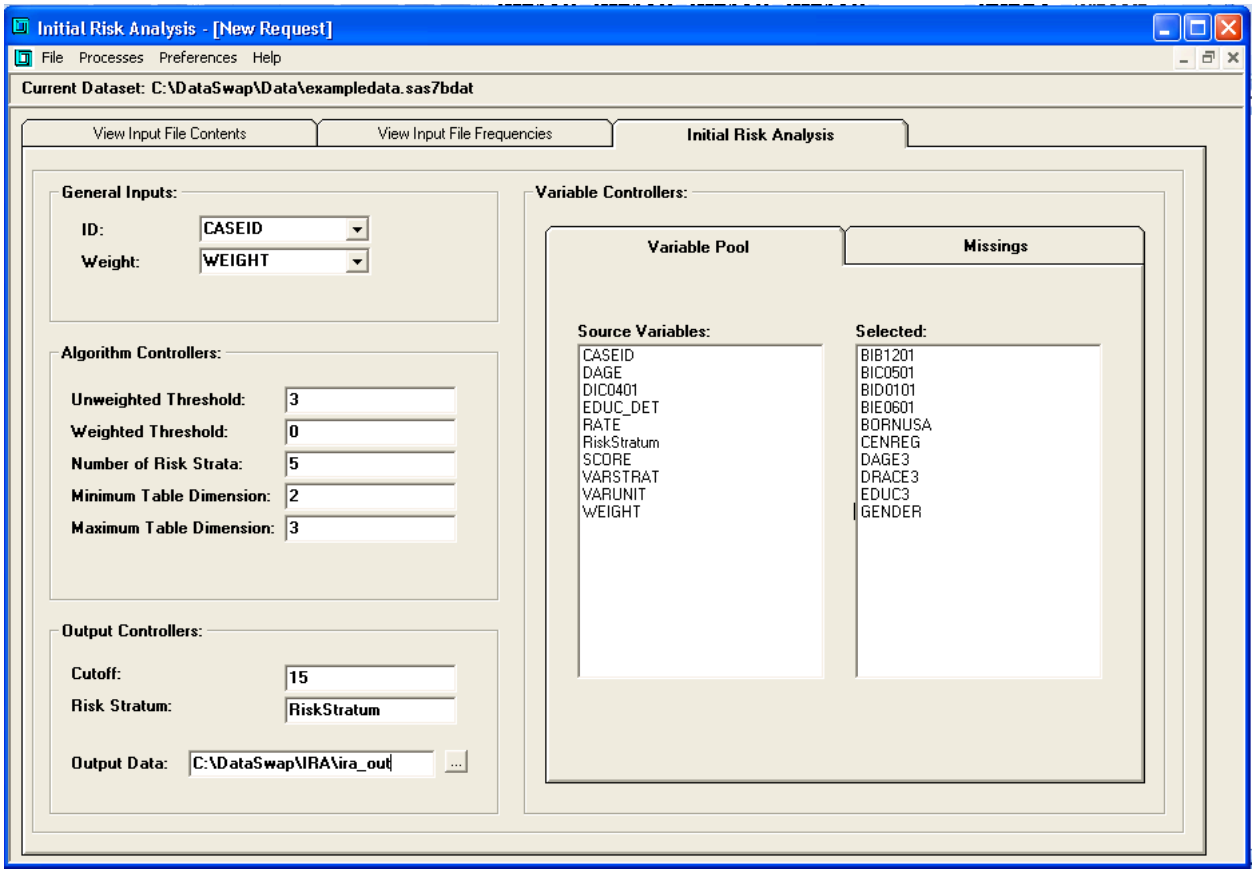
#	Variable	Type	Len	Label
9	BIB1201	Num	8	Ever been in pgm to improve basic skills
10	BIC0501	Num	8	Ever been placed on probation?
11	BID0101	Num	8	Nay work assignments inside or outside?
12	BIE0601	Num	8	How often write letters/memos in Engl?
14	BORNUSA	Num	8	if born in USA 1=Yes, 2=No
1	CASEID	Char	18	Identification no.
2	CENREG	Num	8	Census region
5	DAGE	Num	8	Age derived fr date of birth or Screener
3	DAGE3	Num	8	Derived age category - 1:<30, 2:<50, 3: >=50
7	DIC0401	Num	8	Derived years since admission
15	DRACE3	Num	8	Derived Race/ethnicity - 1:Hispanic, 2:NH Black, 3:Other
4	EDUC3	Num	8	Recoded highest education level - 1:less HS, 2:=HS, 3: >HS
8	EDUC_DET	Num	8	Highest level of education
13	GENDER	Num	8	Gender (sex)
17	RATE	Num	8	sampling rate
20	RiskStratum	Num	8	InitialRisk: Risk Stratum
16	SCORE	Num	8	Literature score
19	VARSTRAT	Num	8	Variance Stratum
18	VARUNIT	Num	8	Variance Unit
6	WEIGHT	Num	8	Final weight

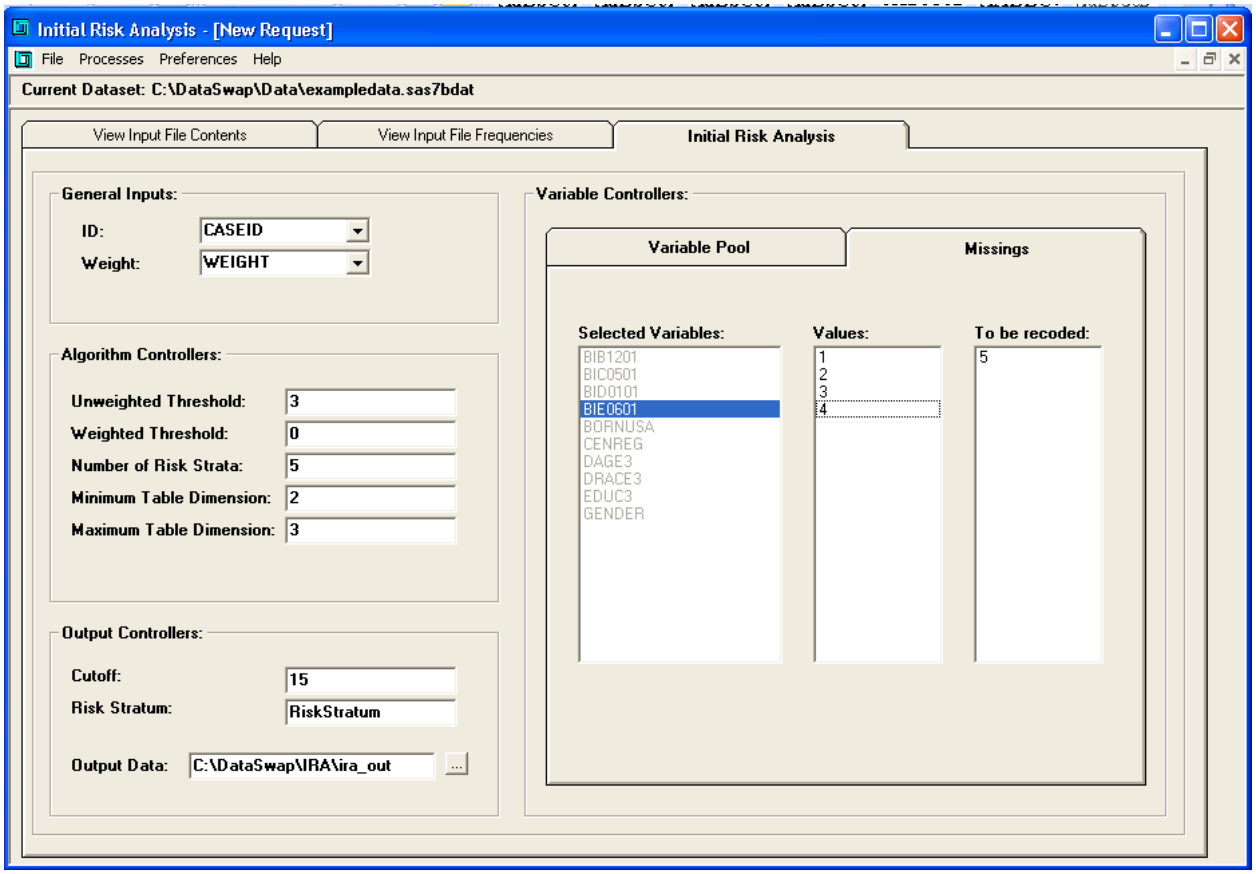
APPENDIX B

Screen Shots from the Windows-Based Interface for *InitialRisk* Example

Appendix B. Screen shots from the Windows-based interface for *InitialRisk* Example

Below are screen shots demonstrating the parameter specifications for the *InitialRisk* example shown in Appendix A.




Initial Risk Analysis - [New Request]
File Processes Preferences Help
Current Dataset: C:\DataSwap\Data\exampledata.sas7bdat
View Input File Contents
View Input File Frequencies
Initial Risk Analysis
General Inputs:
ID:
CASEID
Weight:
WEIGHT
Algorithm Controllers:
Unweighted Threshold:
3
Weighted Threshold:
0
Number of Risk Strata:
5
Minimum Table Dimension:
2
Maximum Table Dimension:
3
Output Controllers:
Cutoff:
15
Risk Stratum:
RiskStratum
Output Data:
C:\DataSwap\IRA\ira_out
Variable Controllers:
Variable Pool
Missings
Selected Variables:
BIB1201
BIC0501
BID0101
BIE0601
BORNUSA
CENREG
DAGE3
DRACE3
EDUC3
GENDER
Values:
1
2
3
4
To be recoded:
5