Analyzing Missing Data

Introduction

Problems

Using Scripts

Missing data and data analysis

- Missing data is a problem in multivariate data because a case will be excluded from the analysis if it is missing data for any variable included in the analysis.
- If our sample is large, we may be able to allow cases to be excluded.
- If our sample is small, we will try to use a substitution method so that we can retain enough cases to have sufficient power to detect effects.
- In either case, we need to make certain that we understand the potential impact that missing data may have on our analysis.

Tools for evaluating missing data

- SPSS has a specific package for evaluating missing data, but it is included under the UT license.
- In place of this package, we will first examine missing data using SPSS statistics and procedures.
- After studying the standard SPSS procedures that we can use to examine missing data, we will use an SPSS script that will produce the output needed for missing data analysis without requiring us to issue all of the SPSS commands individually.

Key issues in missing data analysis

- We will focus on three key issues for evaluating missing data:
  - The number of cases missing per variable
  - The number of variables missing per case
  - The pattern of correlations among variables created to represent missing and valid data.

- Further analysis may be required depending on the problems identified in these analyses.

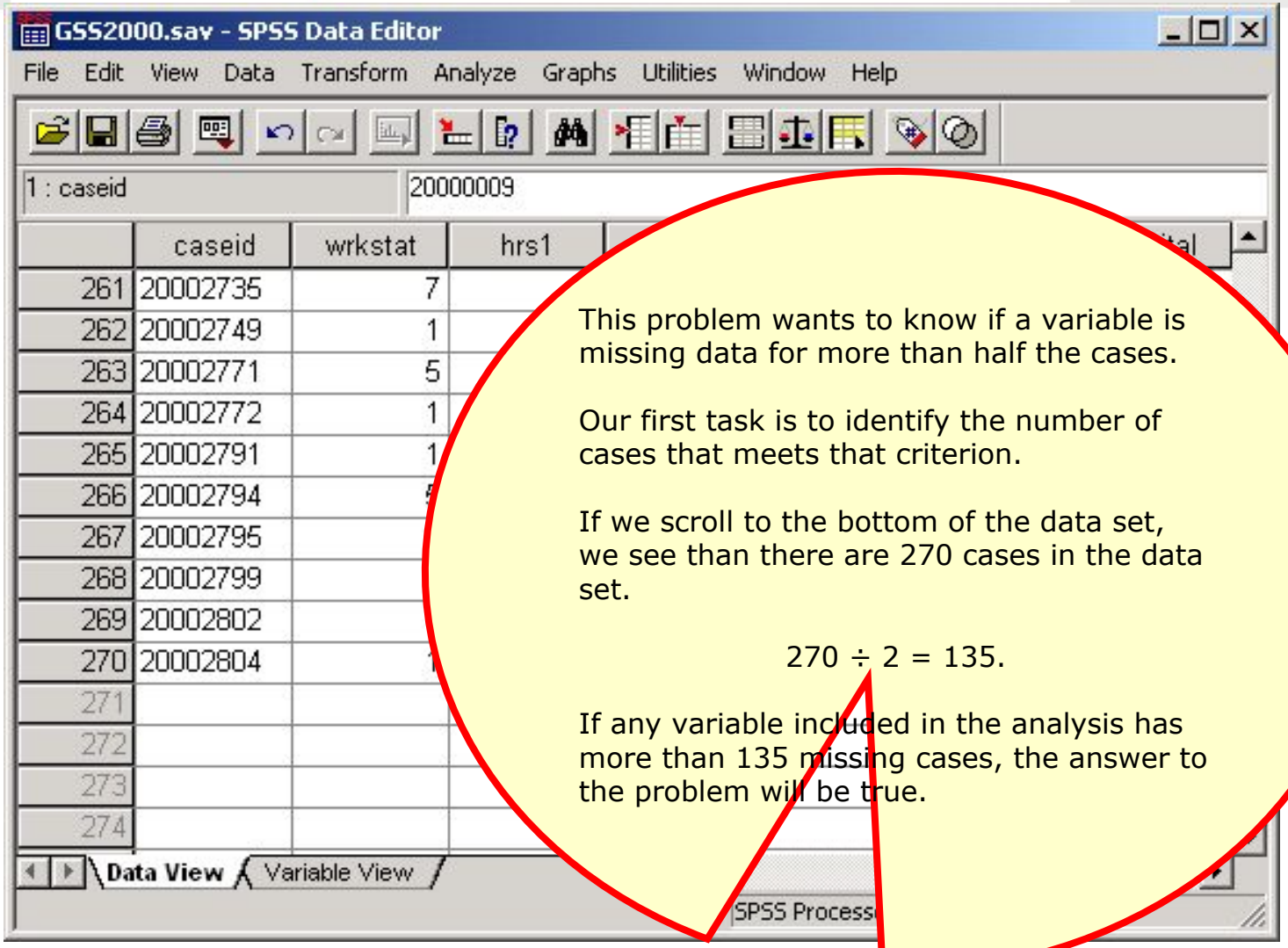
Problem 1

1. Based on a missing data analysis for the variables "employment status," "number of hours worked in the past week," "self employment," "governmental employment," and "occupational prestige score" in the dataset GSS2000.sav, is the following statement true, false, or an incorrect application of a statistic?

The variables "number of hours worked in the past week" and "employment status" are missing data for more than half of the cases in the data set and should be examined carefully before deciding how to handle missing data.

1. True
2. True with caution
3. False
4. Incorrect application of a statistic

Identifying the number of cases in the data set



This problem wants to know if a variable is missing data for more than half the cases.

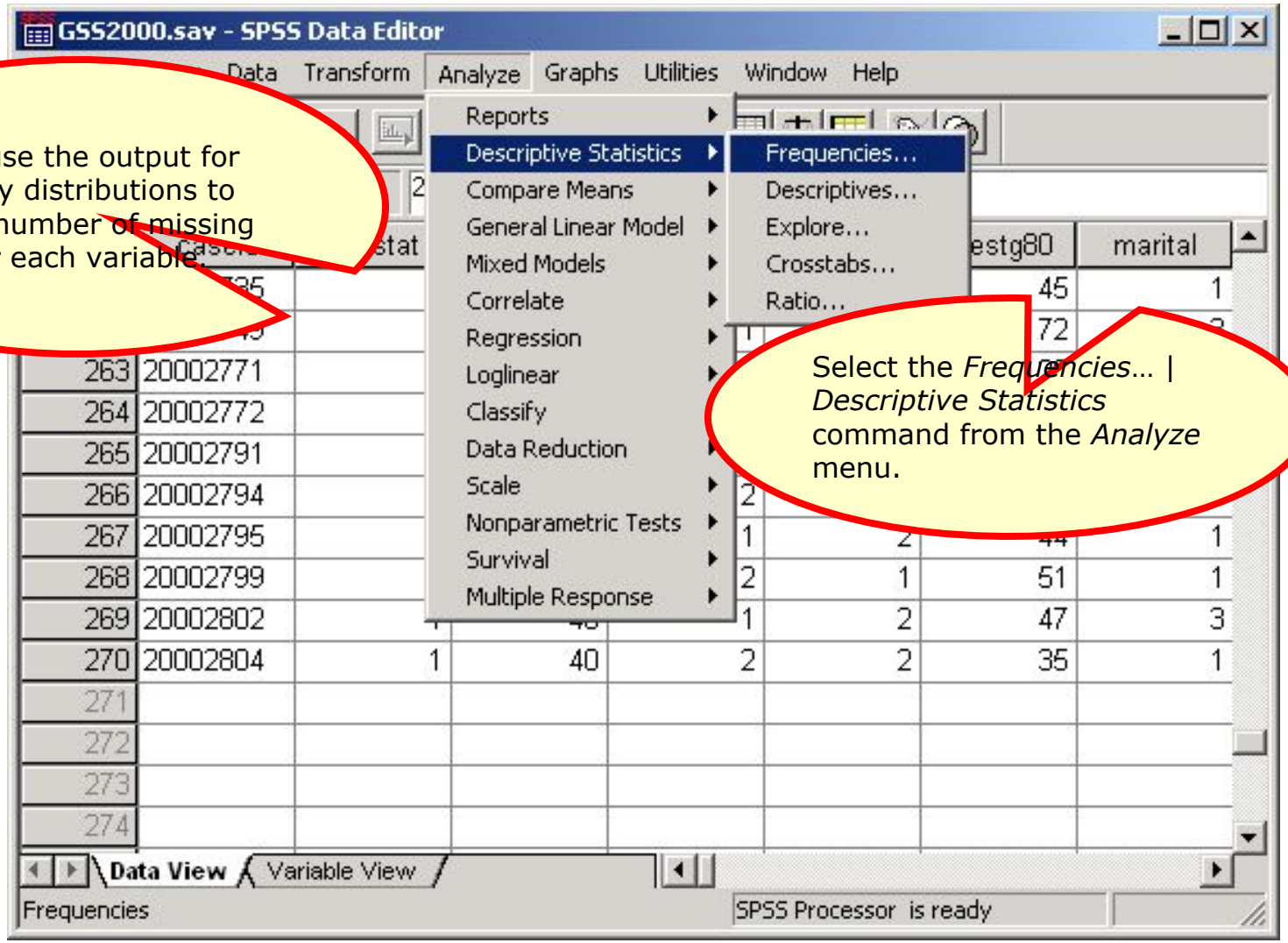
Our first task is to identify the number of cases that meets that criterion.

If we scroll to the bottom of the data set, we see than there are 270 cases in the data set.

$$270 \div 2 = 135.$$

If any variable included in the analysis has more than 135 missing cases, the answer to the problem will be true.

Request frequency distributions

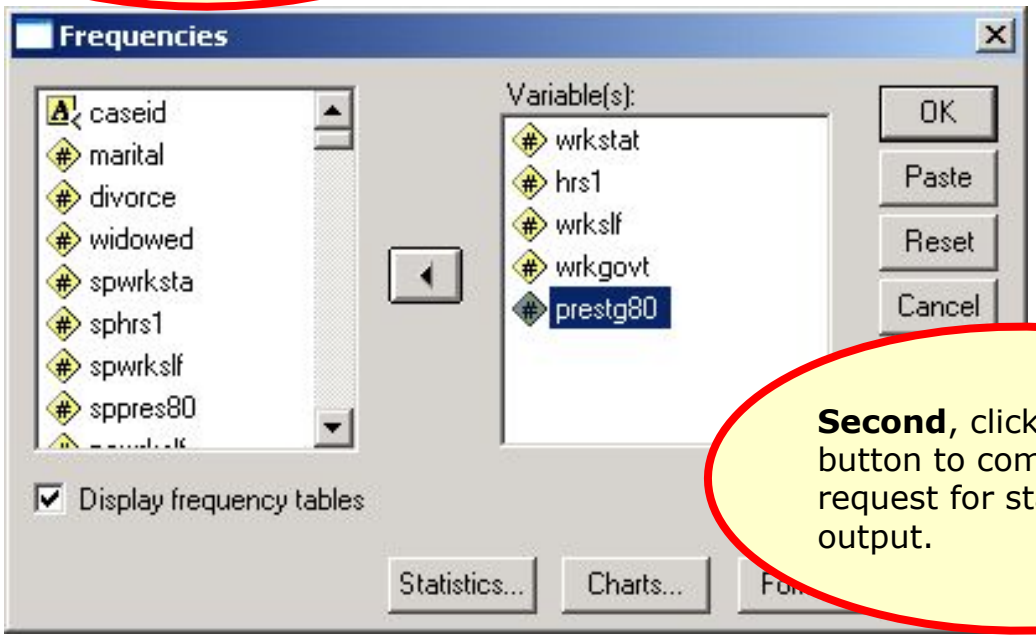


We will use the output for frequency distributions to find the number of missing cases for each variable.

Select the *Frequencies…* | *Descriptive Statistics* command from the *Analyze* menu.

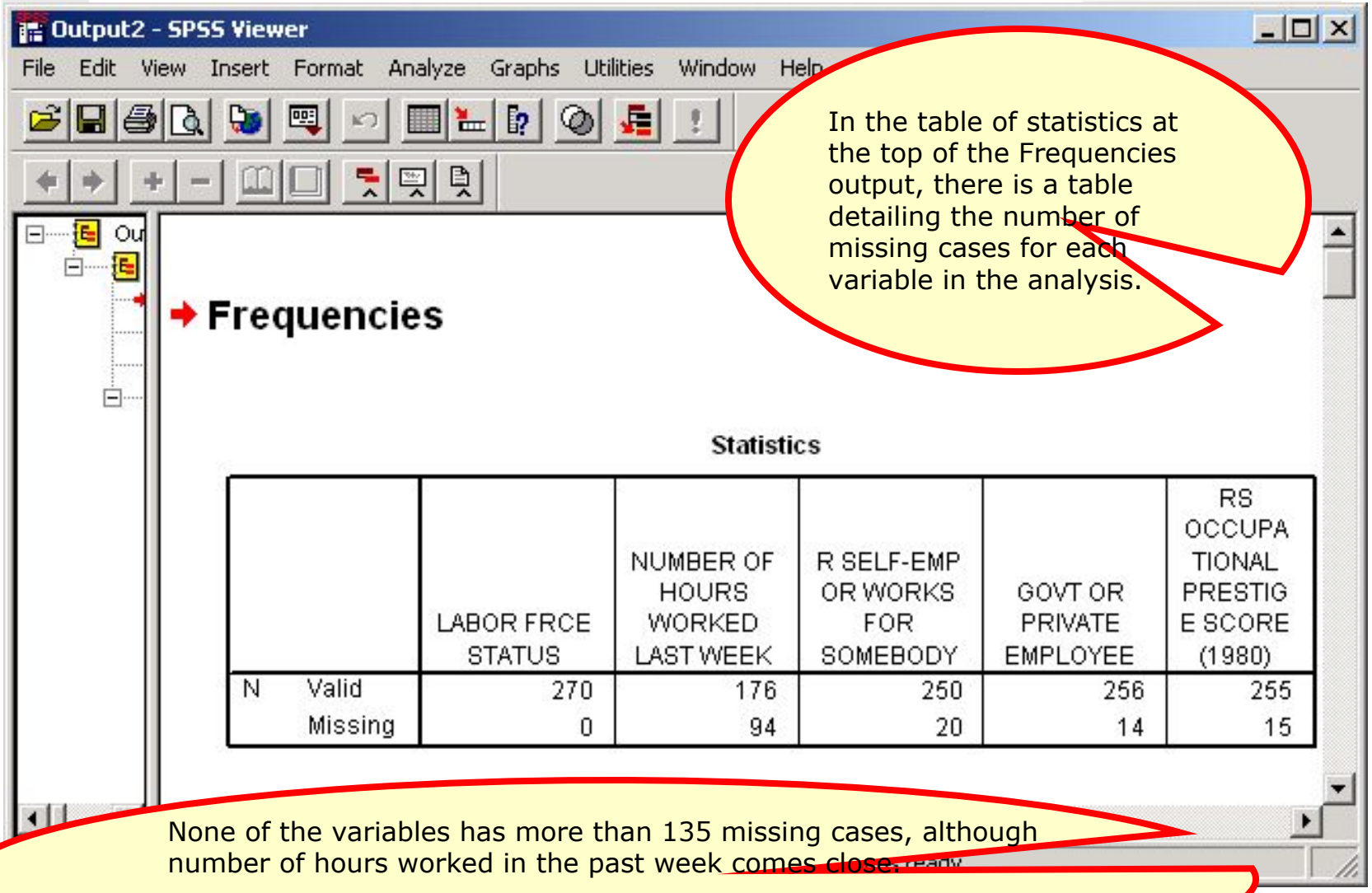
Completing the specification for frequencies

First, move the five variables included in the problem statement to the list box for variables.



Second, click on the OK button to complete the request for statistical output.

Number of missing cases for each variable



Statistics

		LABOR FRCE STATUS	NUMBER OF HOURS WORKED LAST WEEK	R SELF-EMP OR WORKS FOR SOMEBODY	GOVT OR PRIVATE EMPLOYEE	RS OCCUPATIONAL PRESTIGE SCORE (1980)
N	Valid	270	176	250	256	255
	Missing	0	94	20	14	15

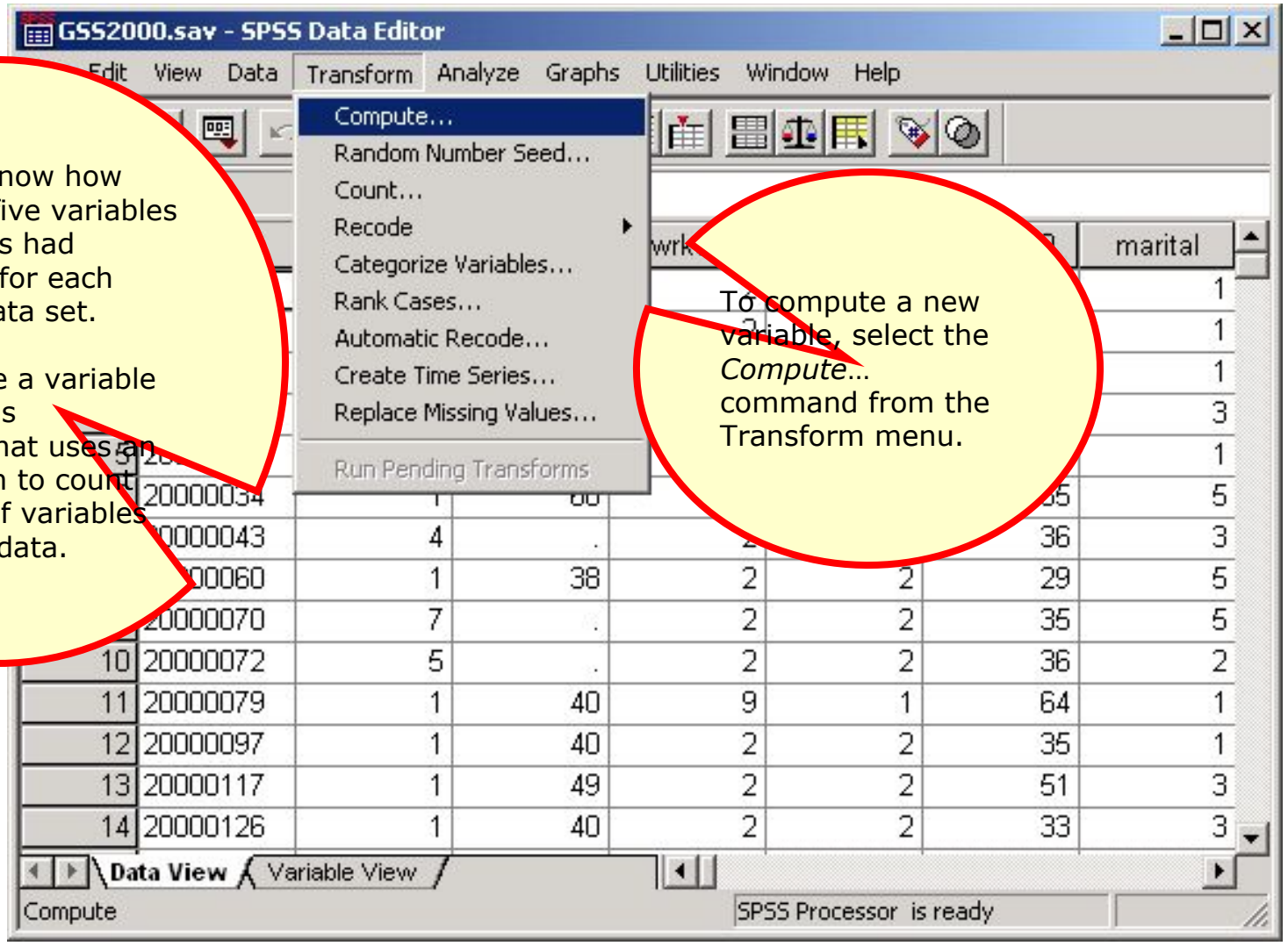
Problem 2

2. Based on a missing data analysis for the variables "employment status," "number of hours worked in the past week," "self employment," "governmental employment," and "occupational prestige score" in the dataset GSS2000.sav, is the following statement true, false, or an incorrect application of a statistic?

14 cases are missing data for more than half of the variables in the analysis and should be examined carefully before deciding how to handle missing data.

1. True
2. True with caution
3. False
4. Incorrect application of a statistic

Create a variable that counts missing data



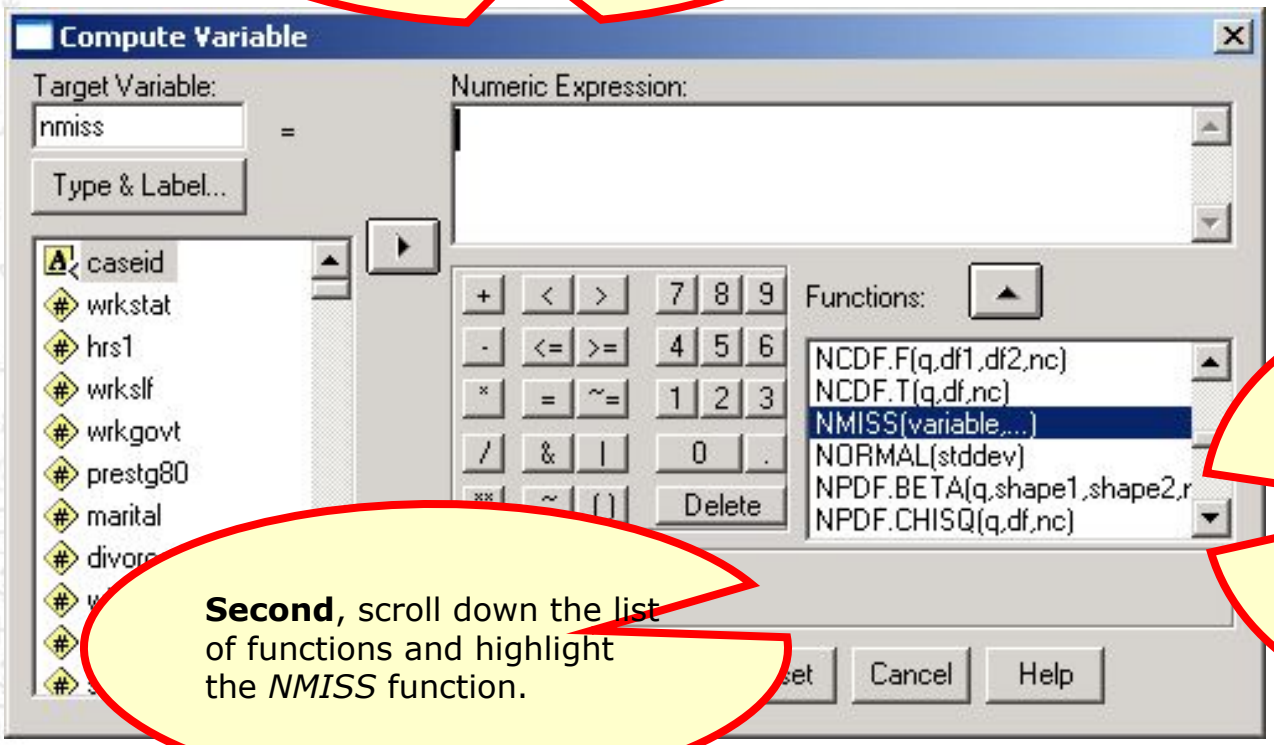
We want to know how many of the five variables in the analysis had missing data for each case in the data set.

We will create a variable containing this information that uses an SPSS function to count the number of variables with missing data.

To compute a new variable, select the *Compute…* command from the Transform menu.

Enter specifications for new variable

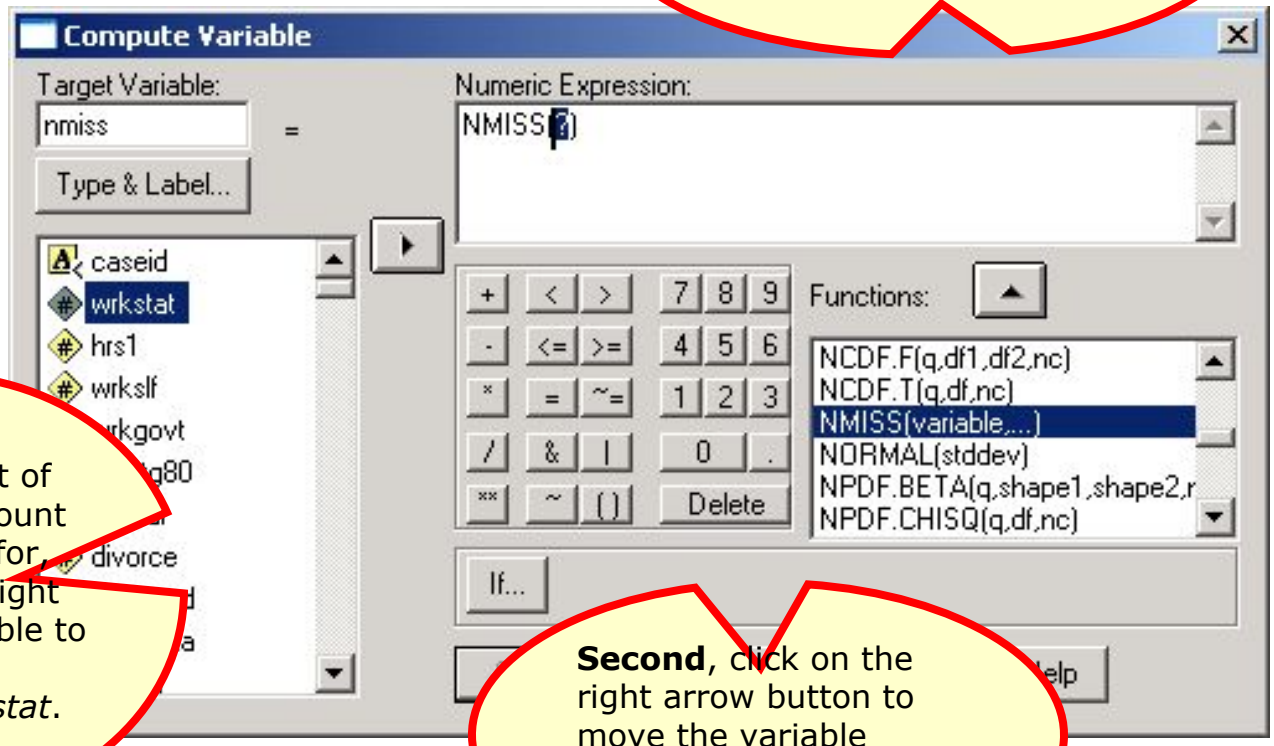
First, type in the name for the new variable *nmiss* in the Target variable text box.



Second, scroll down the list of functions and highlight the *NMISS* function.

Third, click on the up arrow button to move the *NMISS* function into the Numeric Expression text box.

Enter specifications for new variable

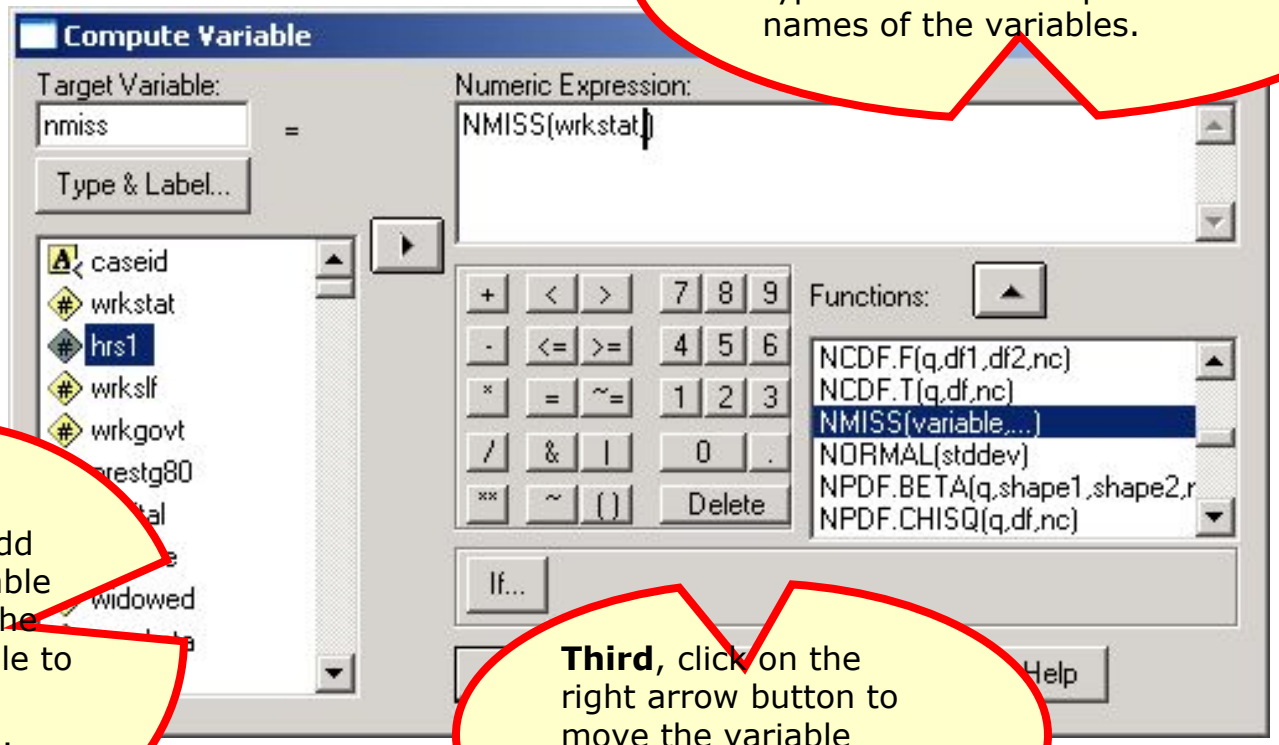


The *NMISS* function is moved into the Numeric Expression text box.

To add the list of variables to count missing data for, we **first** highlight the first variable to include in the function, *wrkstat*.

Second, click on the right arrow button to move the variable name into the function arguments.

Enter specifications for new variable

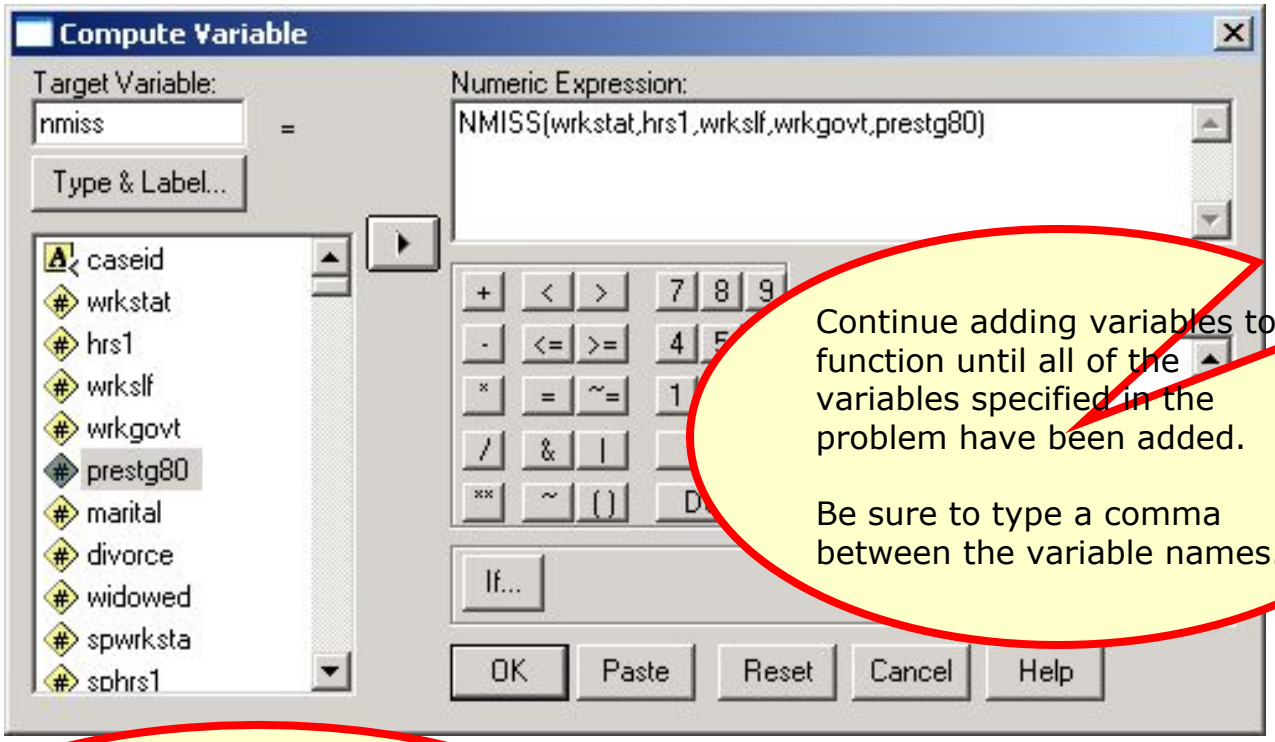


First, before we add another variable to the function, we type a comma to separate the names of the variables.

Second, to add the next variable we highlight the second variable to include in the function, *hrs1*.

Third, click on the right arrow button to move the variable name into the function arguments.

Complete specifications for new variable

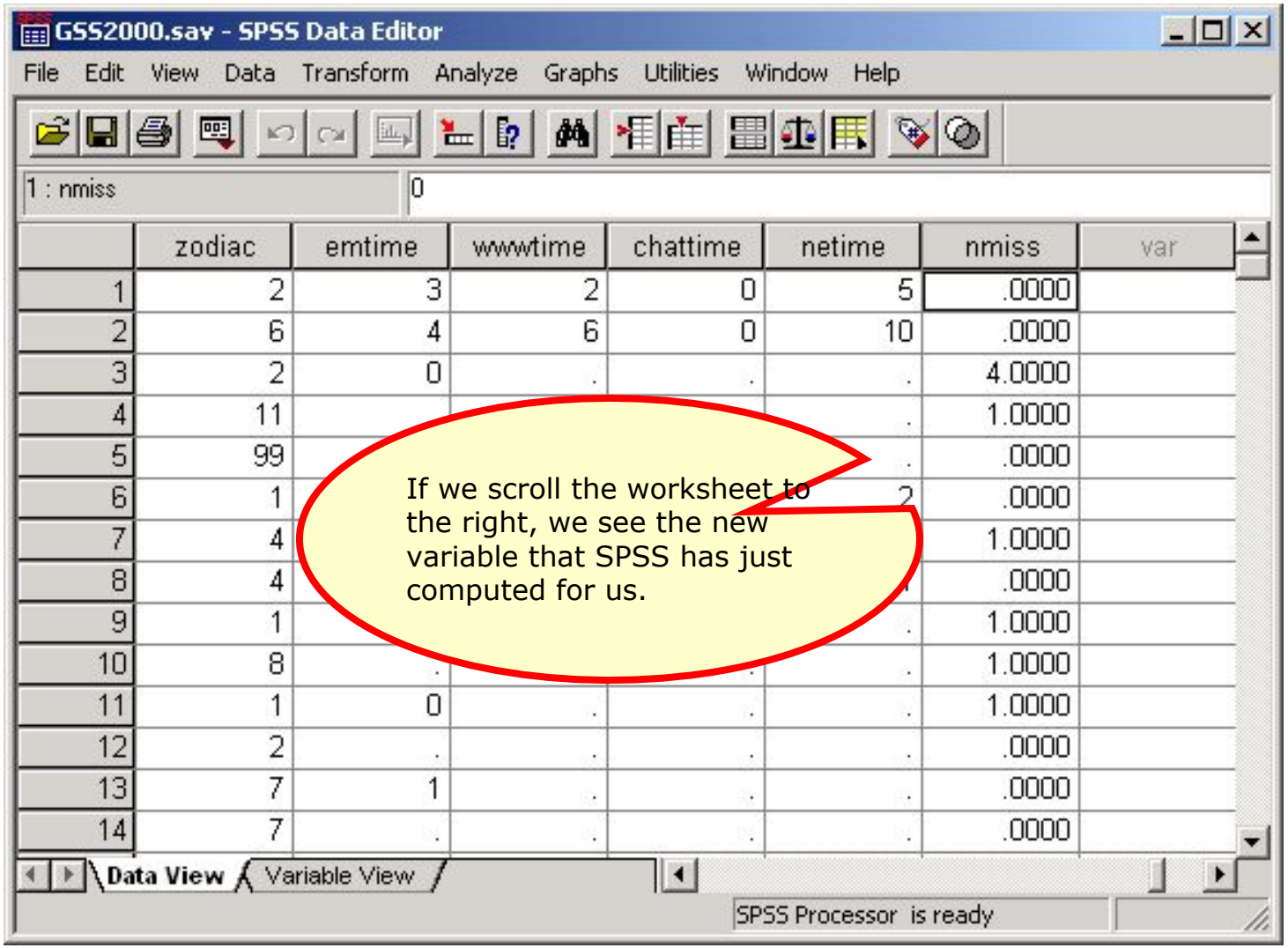


Continue adding variables to function until all of the variables specified in the problem have been added.

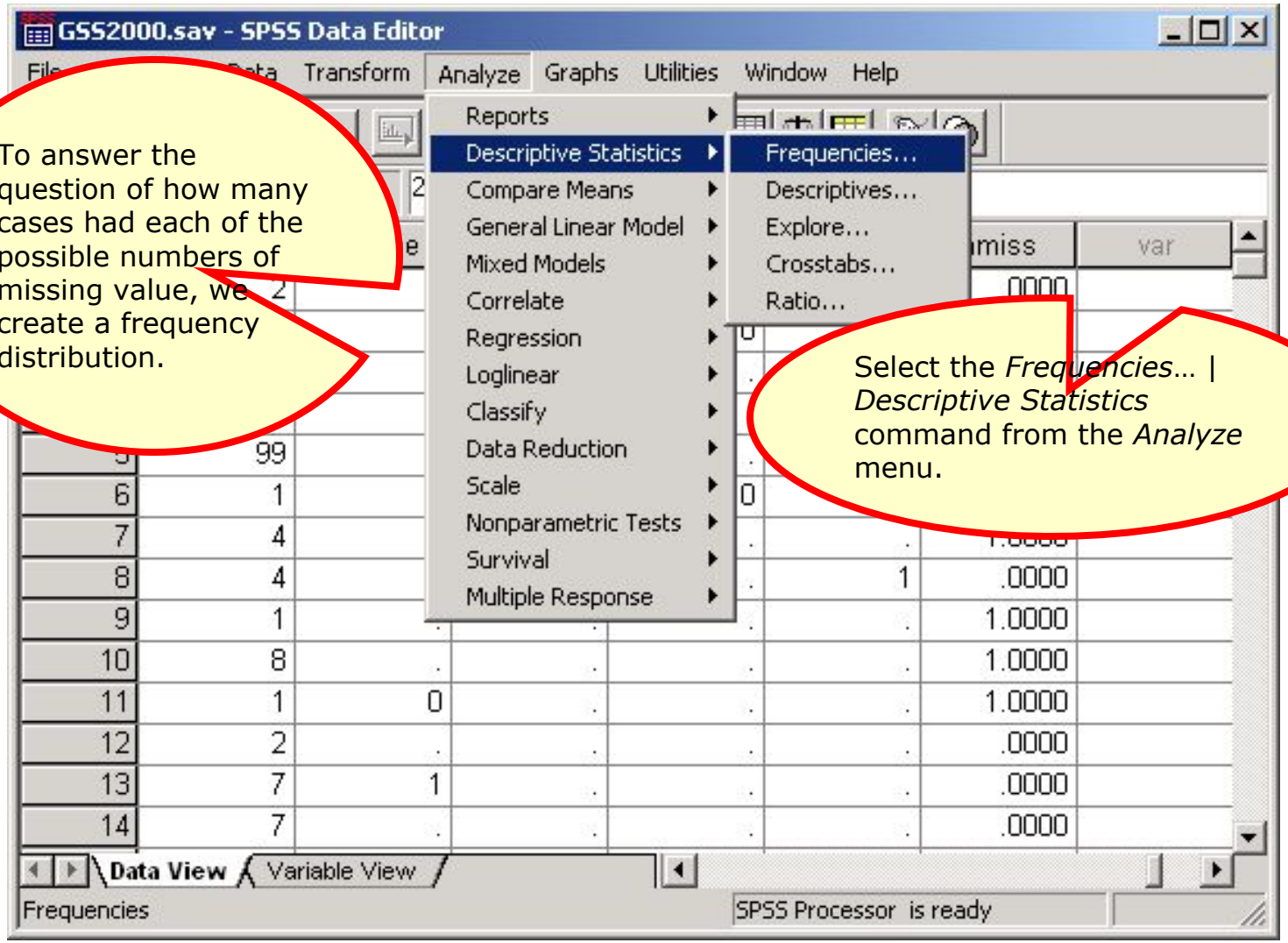
Be sure to type a comma between the variable names.

When all of the variables have been added to the function, click on the *OK* button to complete the specifications.

The *nmiss* variable in the data editor

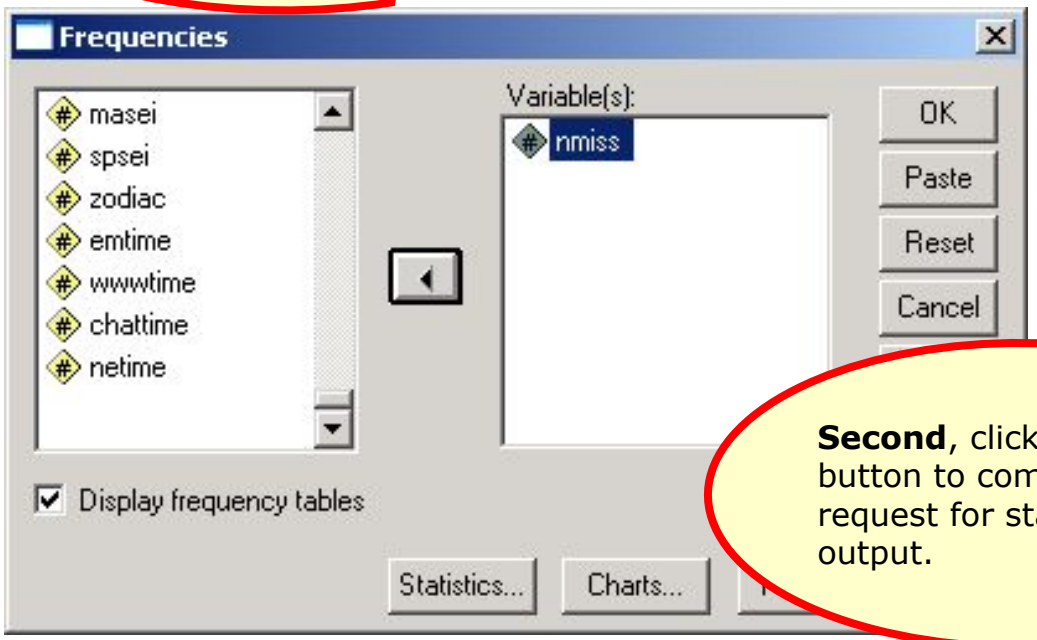


A frequency distribution for *nmiss*



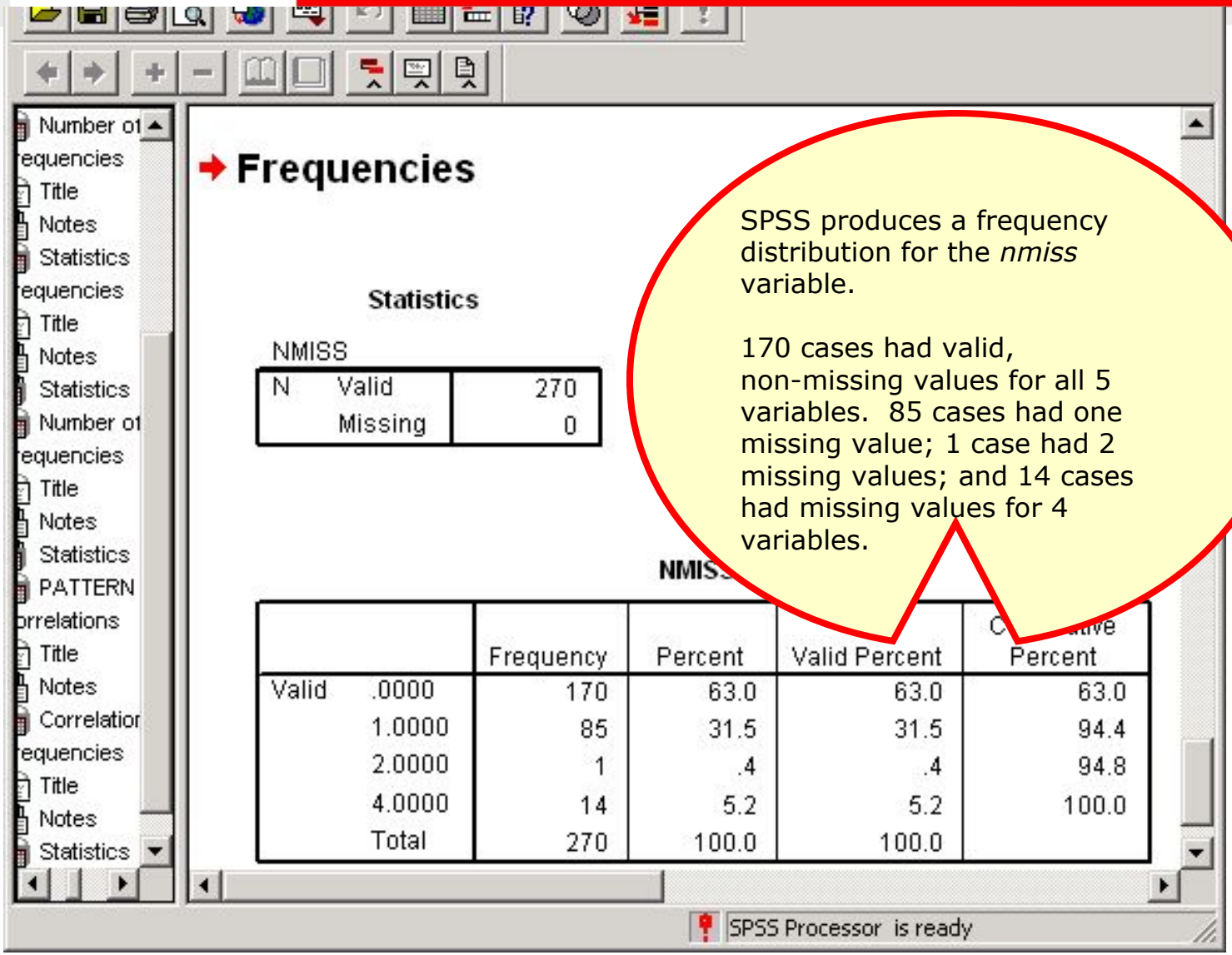
Completing the specification for frequencies

First, move the *nmiss* variable to the list of variables.



Second, click on the OK button to complete the request for statistical output.

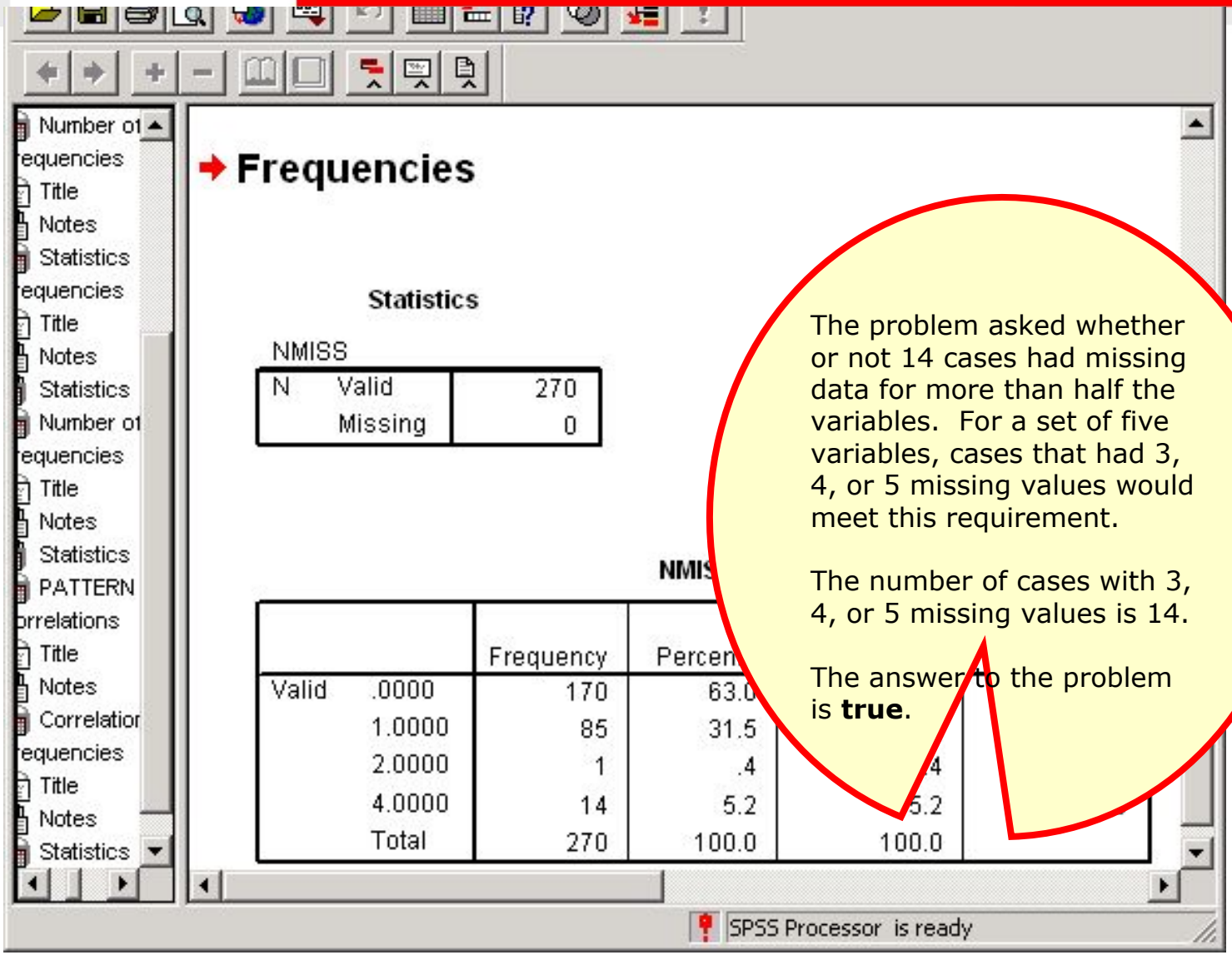
The frequency distribution



SPSS produces a frequency distribution for the *nmiss* variable.

170 cases had valid, non-missing values for all 5 variables. 85 cases had one missing value; 1 case had 2 missing values; and 14 cases had missing values for 4 variables.

Answering the problem



The problem asked whether or not 14 cases had missing data for more than half the variables. For a set of five variables, cases that had 3, 4, or 5 missing values would meet this requirement.

The number of cases with 3, 4, or 5 missing values is 14.

The answer to the problem is **true**.

Problem 3

3. Based on a missing data analysis for the variables "employment status," "number of hours worked in the past week," "self employment," "governmental employment," and "occupational prestige score" in the dataset GSS2000.sav, is the following statement true, false, or an incorrect application of a statistic? Use 0.01 as the level of significance.

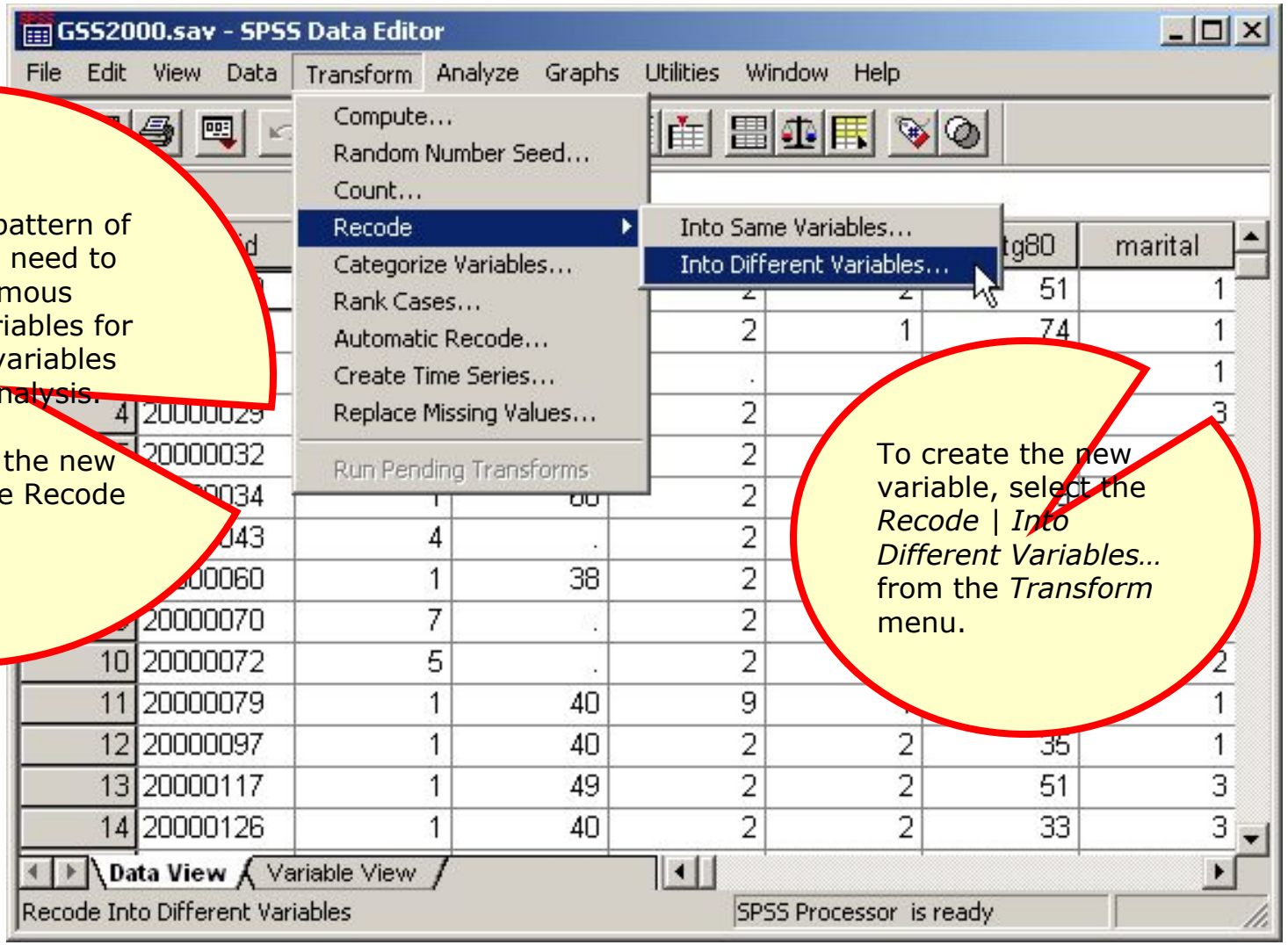
After excluding cases with missing data for more than half of the variables from the analysis if necessary, the presence of statistically significant correlations in the matrix of dichotomous missing/valid variables suggests that the missing data pattern may not be random.

1. True
2. True with caution
3. False
4. Incorrect application of a statistic

Compute valid/missing dichotomous variables

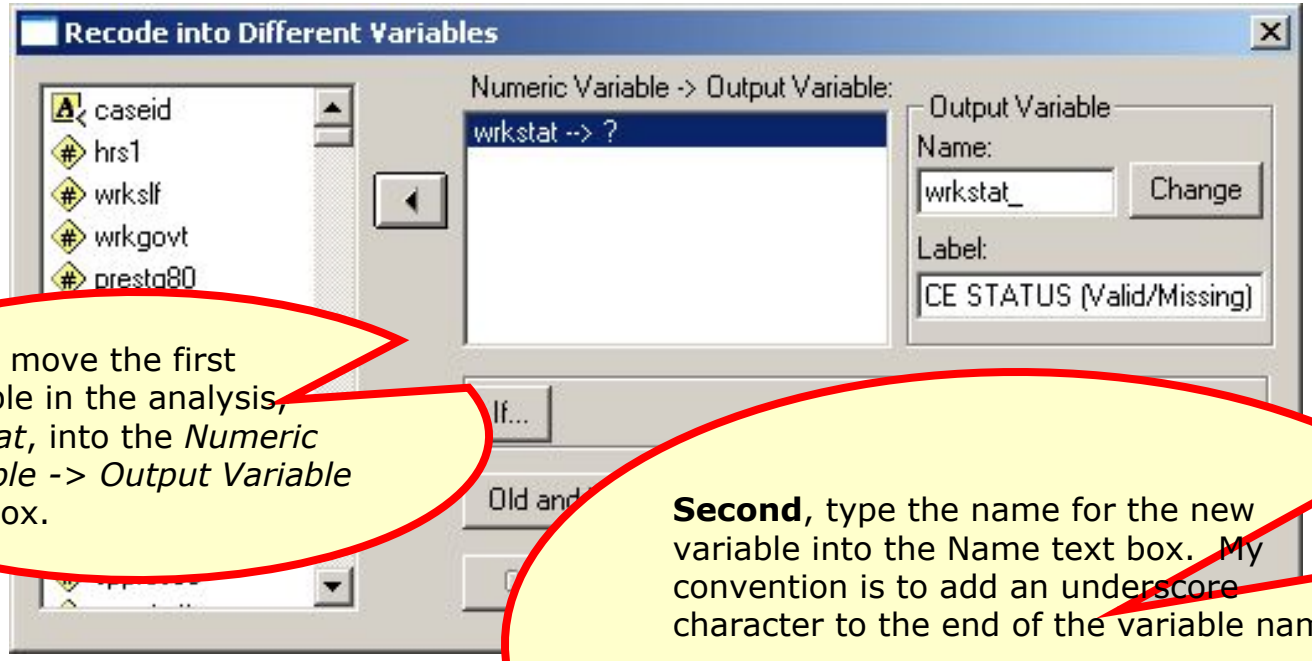
To evaluate the pattern of missing data, we need to compute dichotomous valid/missing variables for each of the five variables included in the analysis.

We will compute the new variable using the Recode command.



To create the new variable, select the *Recode* | *Into Different Variables…* from the *Transform* menu.

Enter specifications for new variable

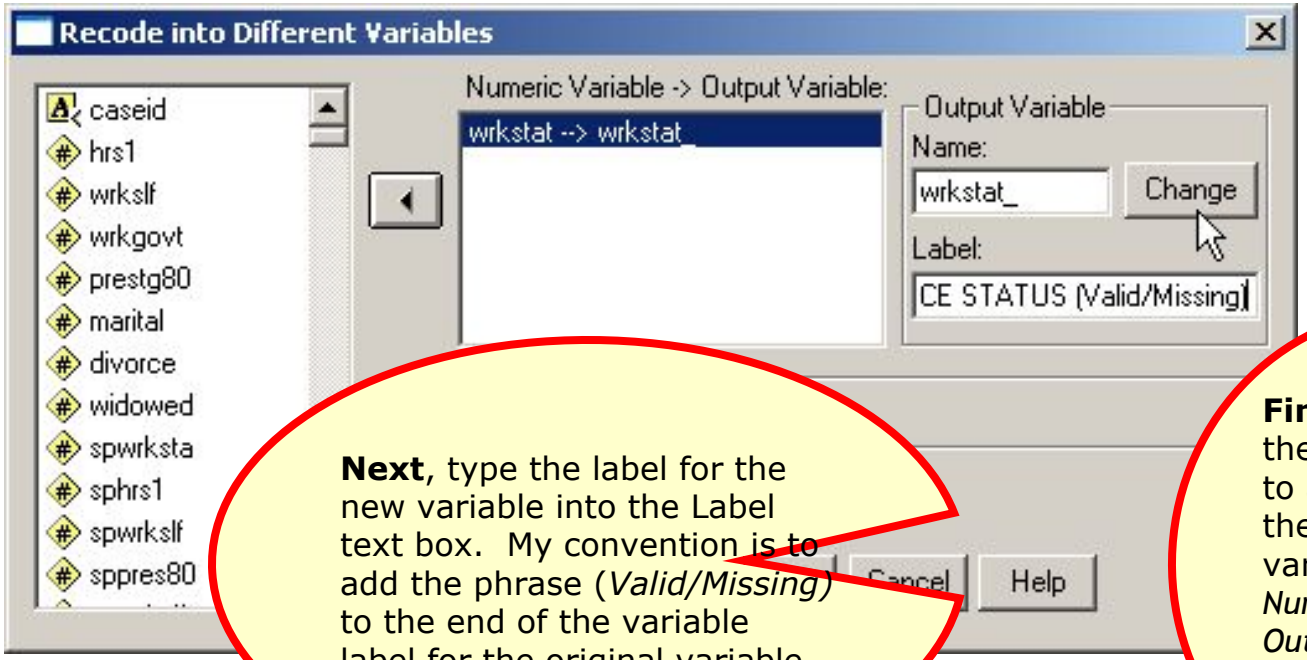


First, move the first variable in the analysis, *wrkstat*, into the *Numeric Variable -> Output Variable* text box.

Second, type the name for the new variable into the Name text box. My convention is to add an underscore character to the end of the variable name.

If this would make the variable more than 8 characters long, delete characters from the end of the original variable name.

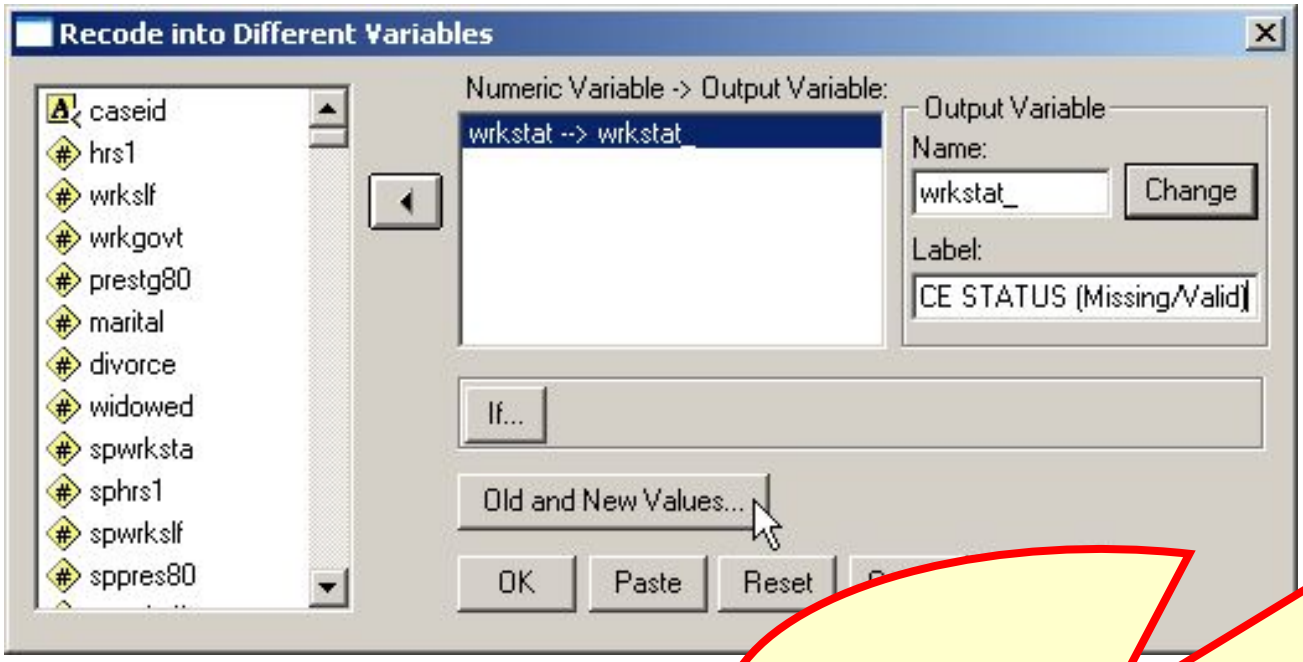
Enter specifications for new variable



Next, type the label for the new variable into the Label text box. My convention is to add the phrase (*Valid/Missing*) to the end of the variable label for the original variable.

Finally, click on the Change button to add the name of the dichotomous variable to the *Numeric Variable -> Output Variable* text box.

Enter specifications for new variable



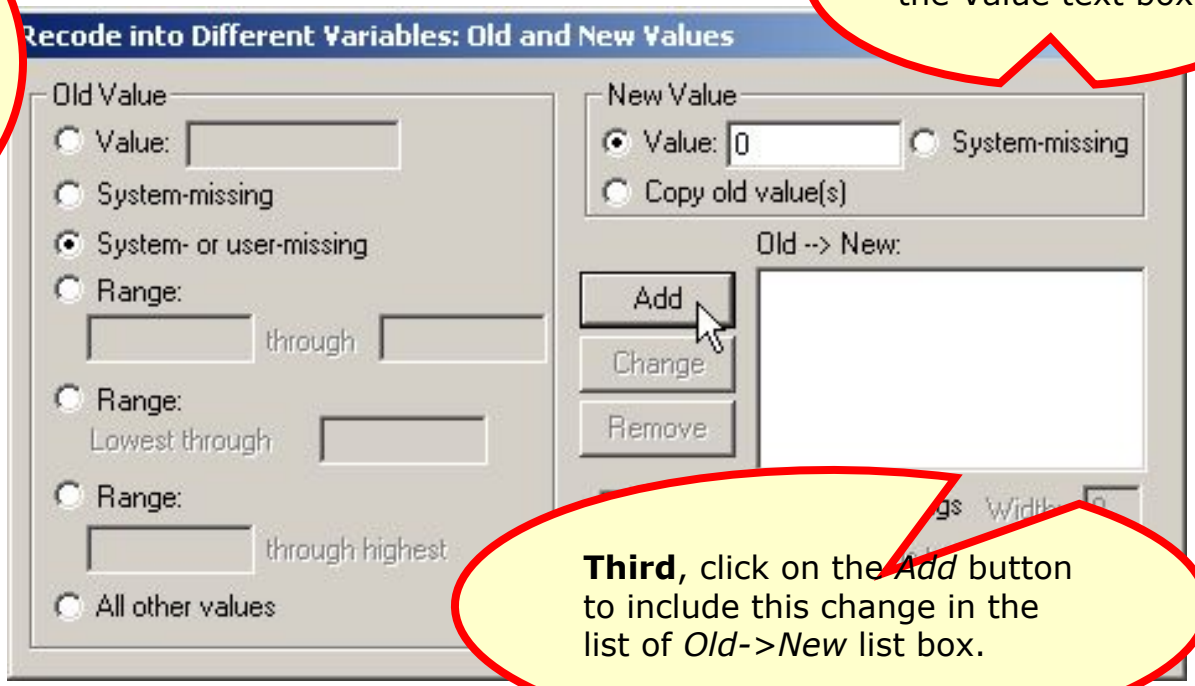
To specify the values for the new variable, click on the *Old and New Values…* button.

Change the value for missing data

The dichotomous variable should be coded 1 if the variable has a valid value, 0 if the variable has a missing value.

First, mark the *System- or user-missing* option button.

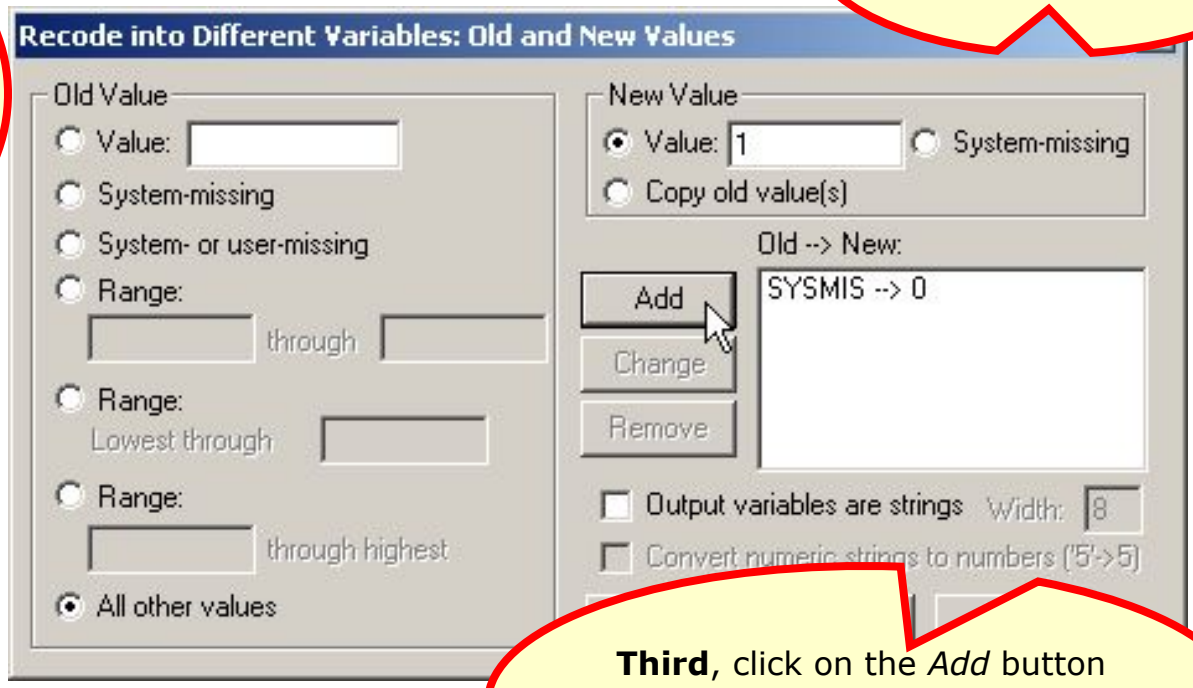
Second, type 0 in the Value text box.



Third, click on the *Add* button to include this change in the list of *Old->New* list box.

Change the value for valid data

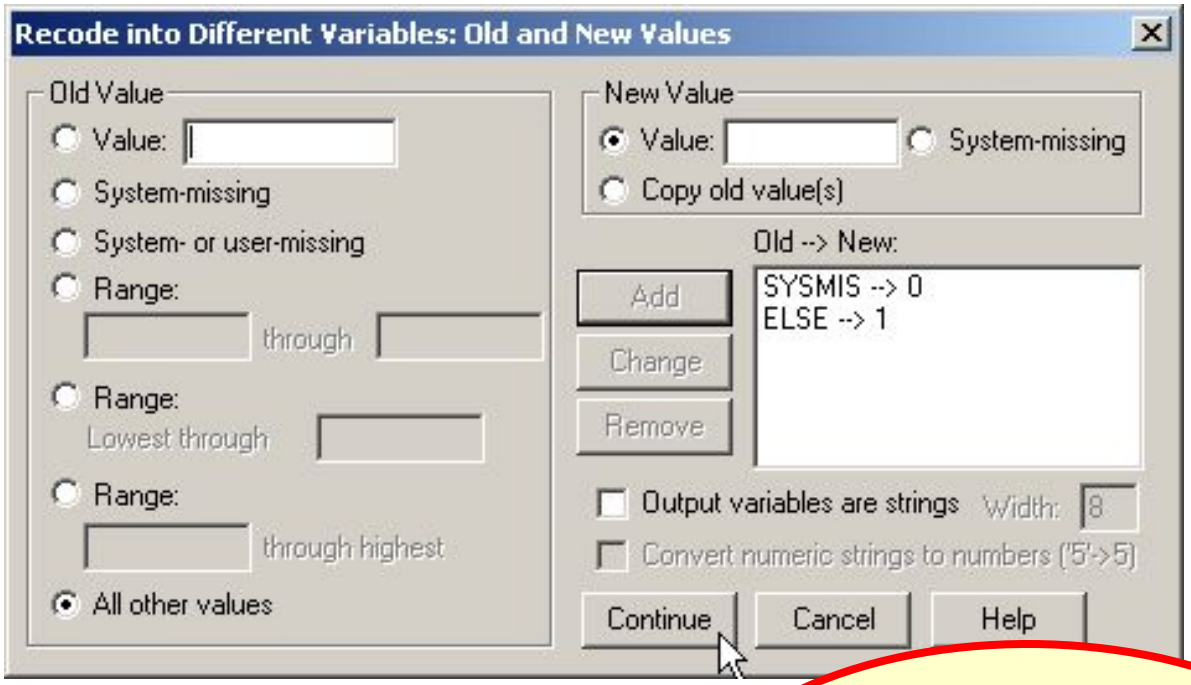
First, mark the *All other values* option button.



Second, type 1 in the Value text box.

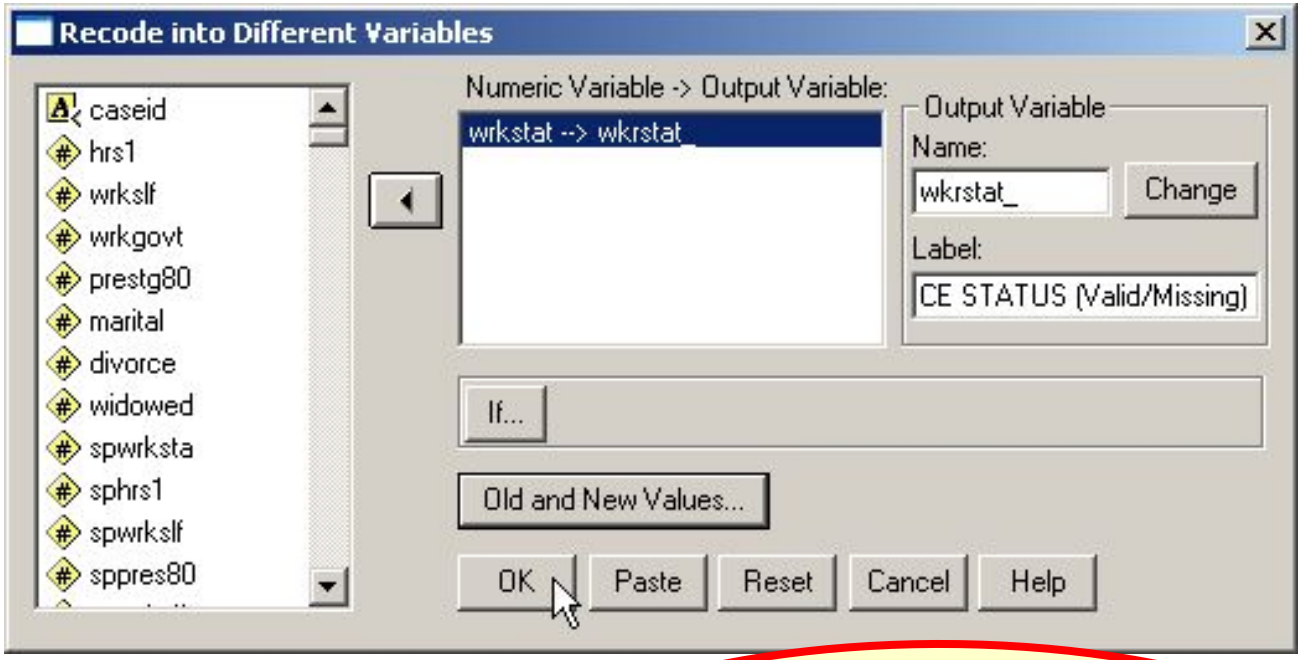
Third, click on the *Add* button to include this change in the list of *Old->New* list box.

Complete the value specifications



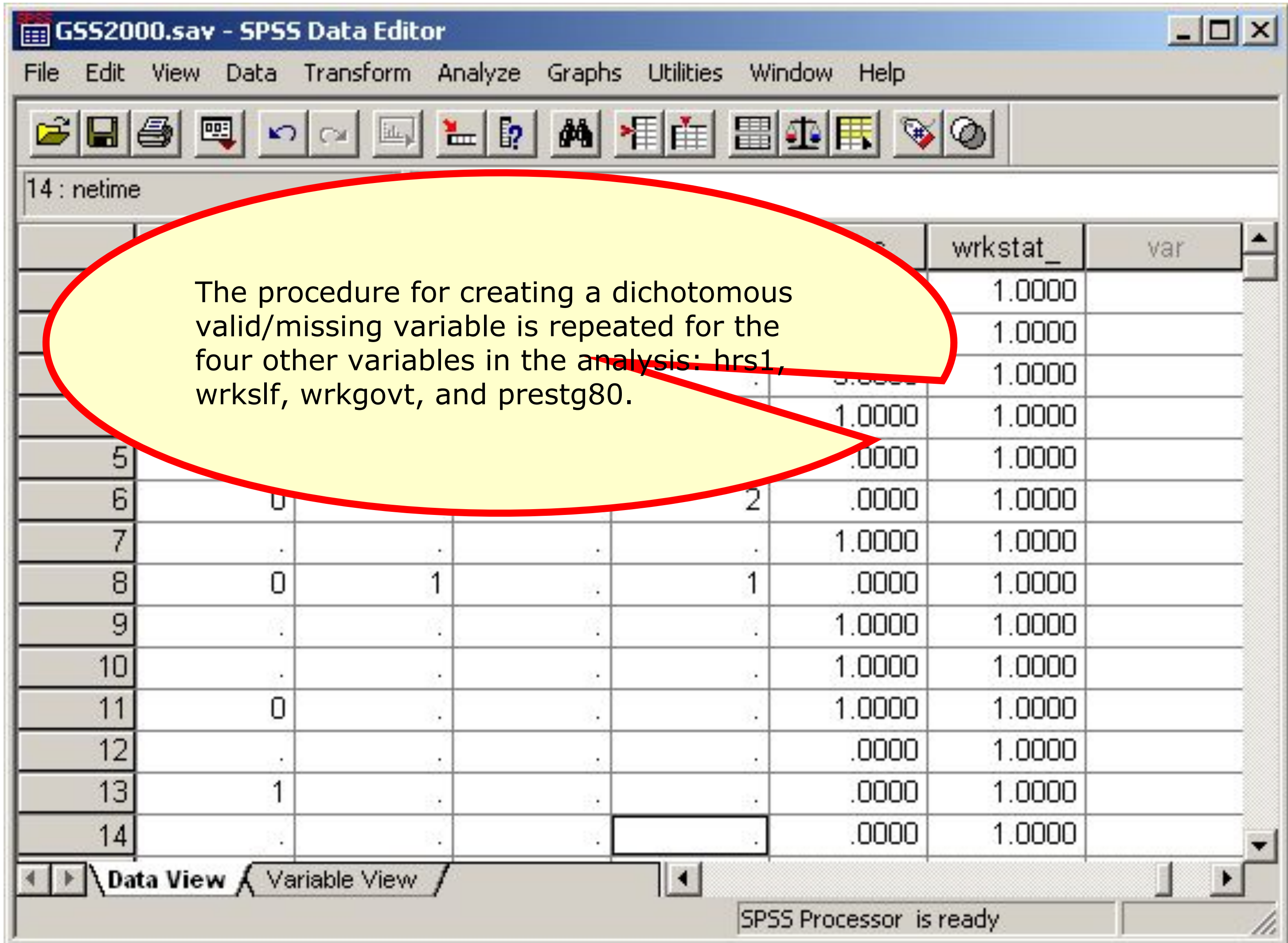
Having entered the values for recoding the variable into dichotomous values, we click on the *Continue* button to complete this dialog box.

Complete the recode specifications



Having entered specifications for the new variable and the values for recoding the variable into dichotomous values, we click on the *OK* button to produce the new variable.

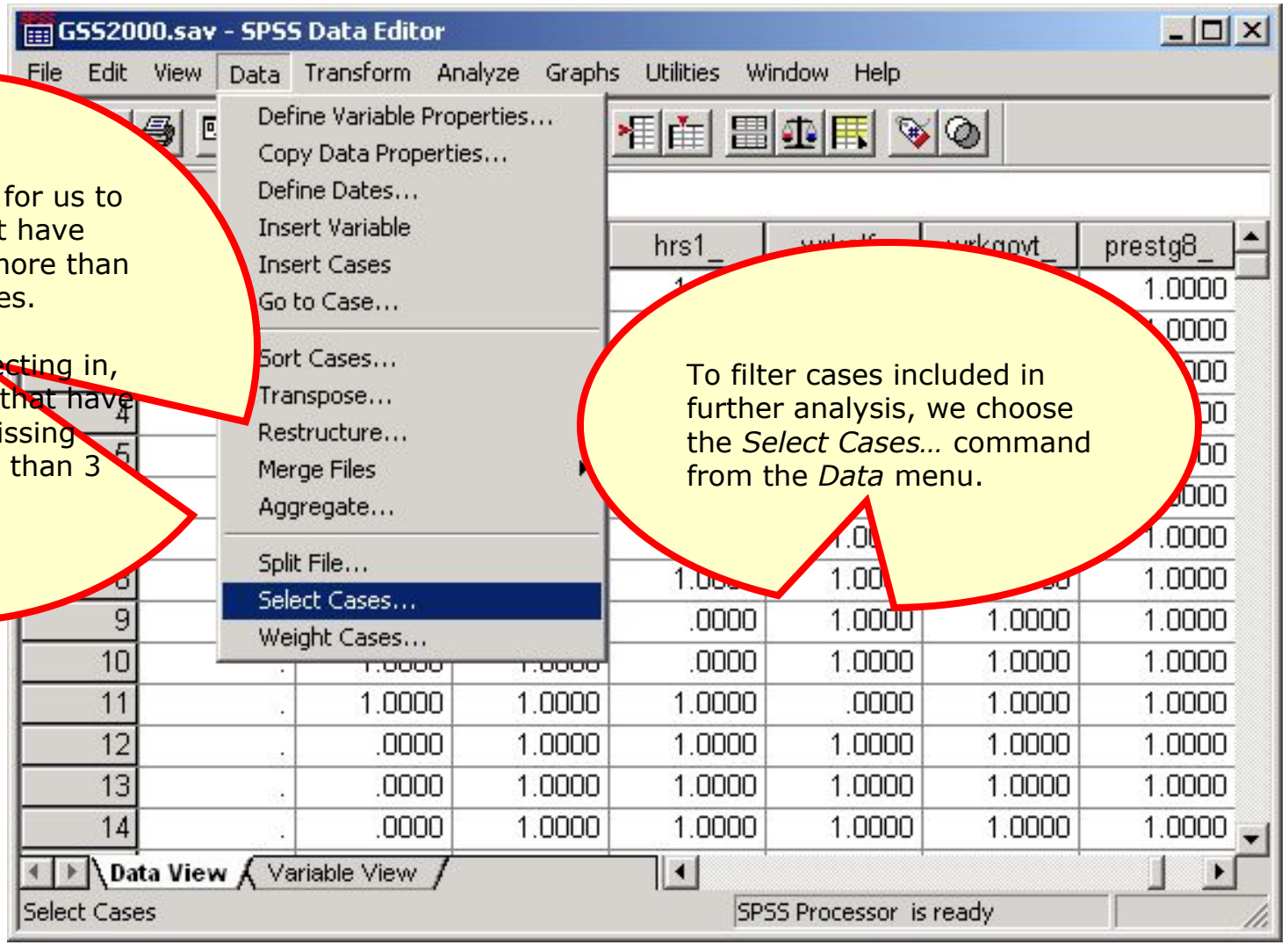
The dichotomous variable



Filtering cases with excessive missing variables

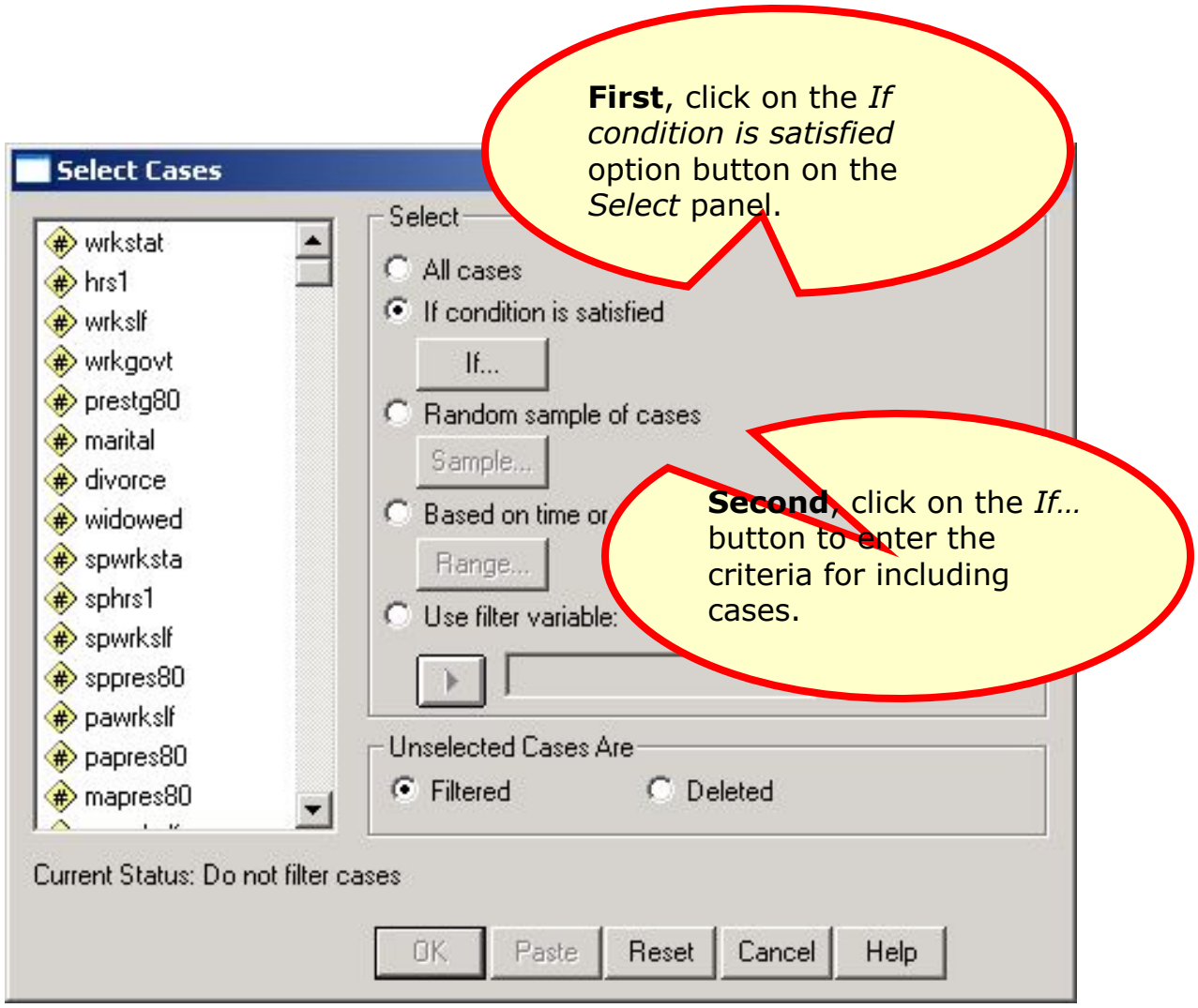
The problem calls for us to exclude cases that have missing data for more than half of the variables.

We do this by selecting in, or filtering, cases that have fewer than half missing variables, i.e. less than 3 missing variables.

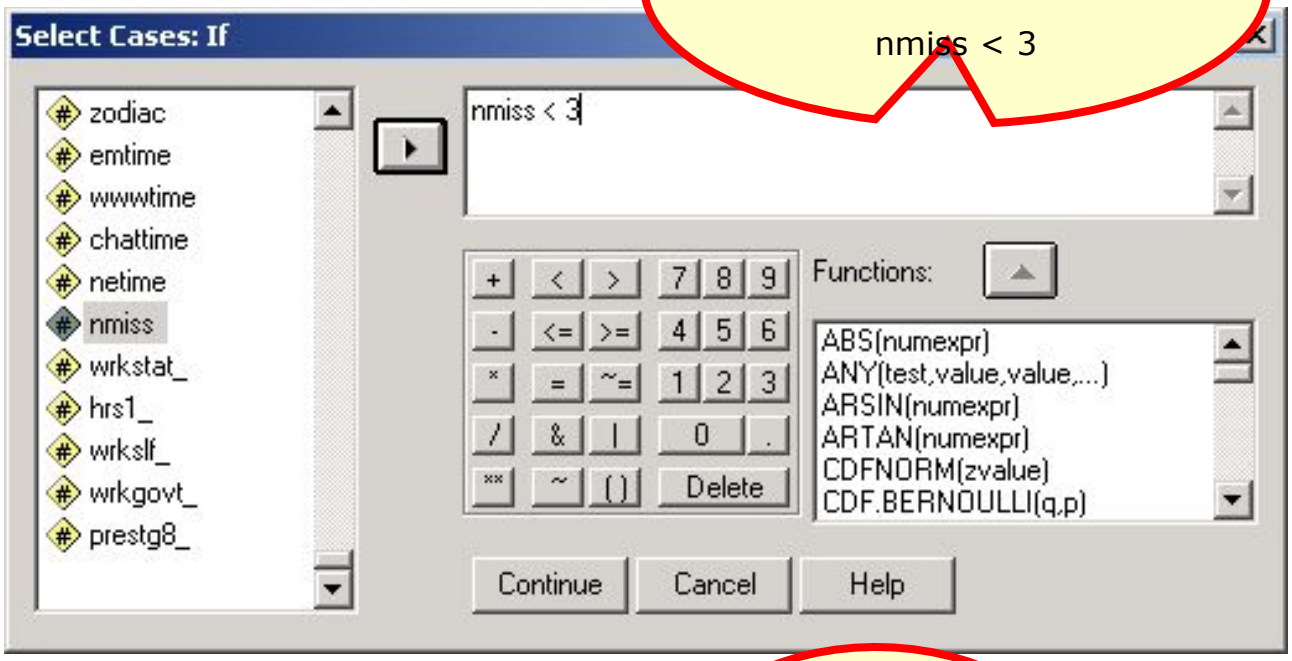


To filter cases included in further analysis, we choose the *Select Cases…* command from the *Data* menu.

Enter specifications for selecting cases



Enter specifications for selecting cases

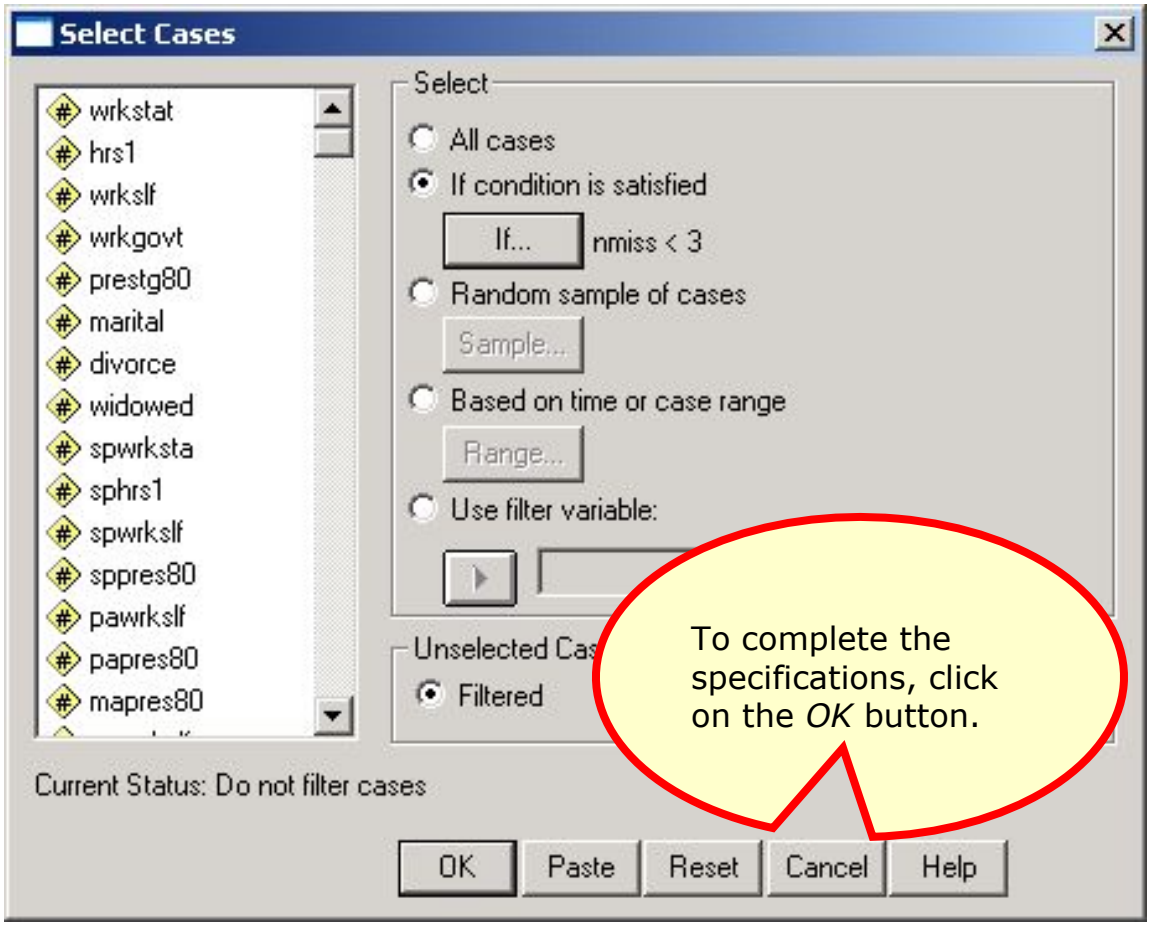


First, enter the criteria for including cases:

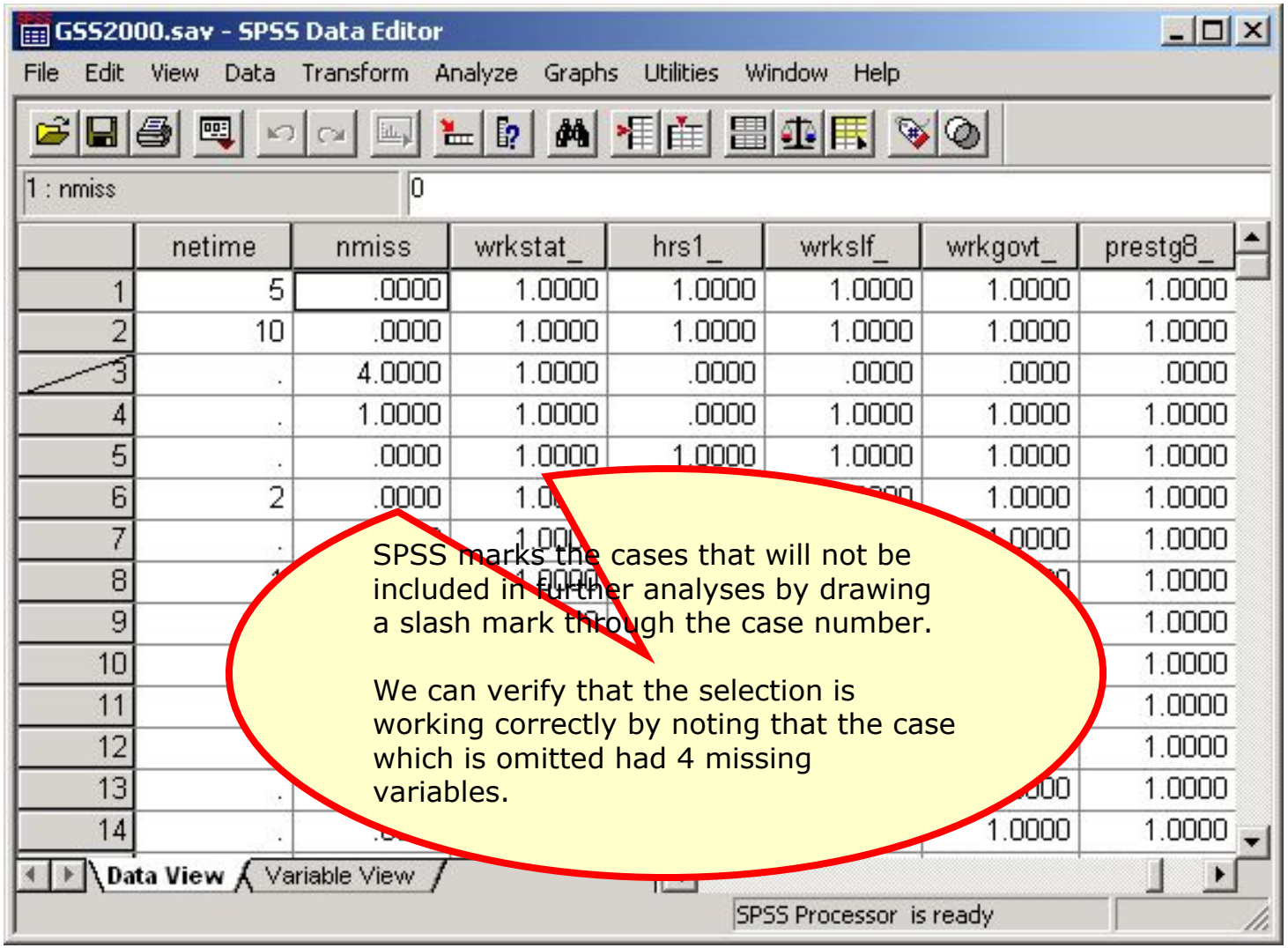
nmiss < 3

Second, click on the *Continue* button to complete the If specification.

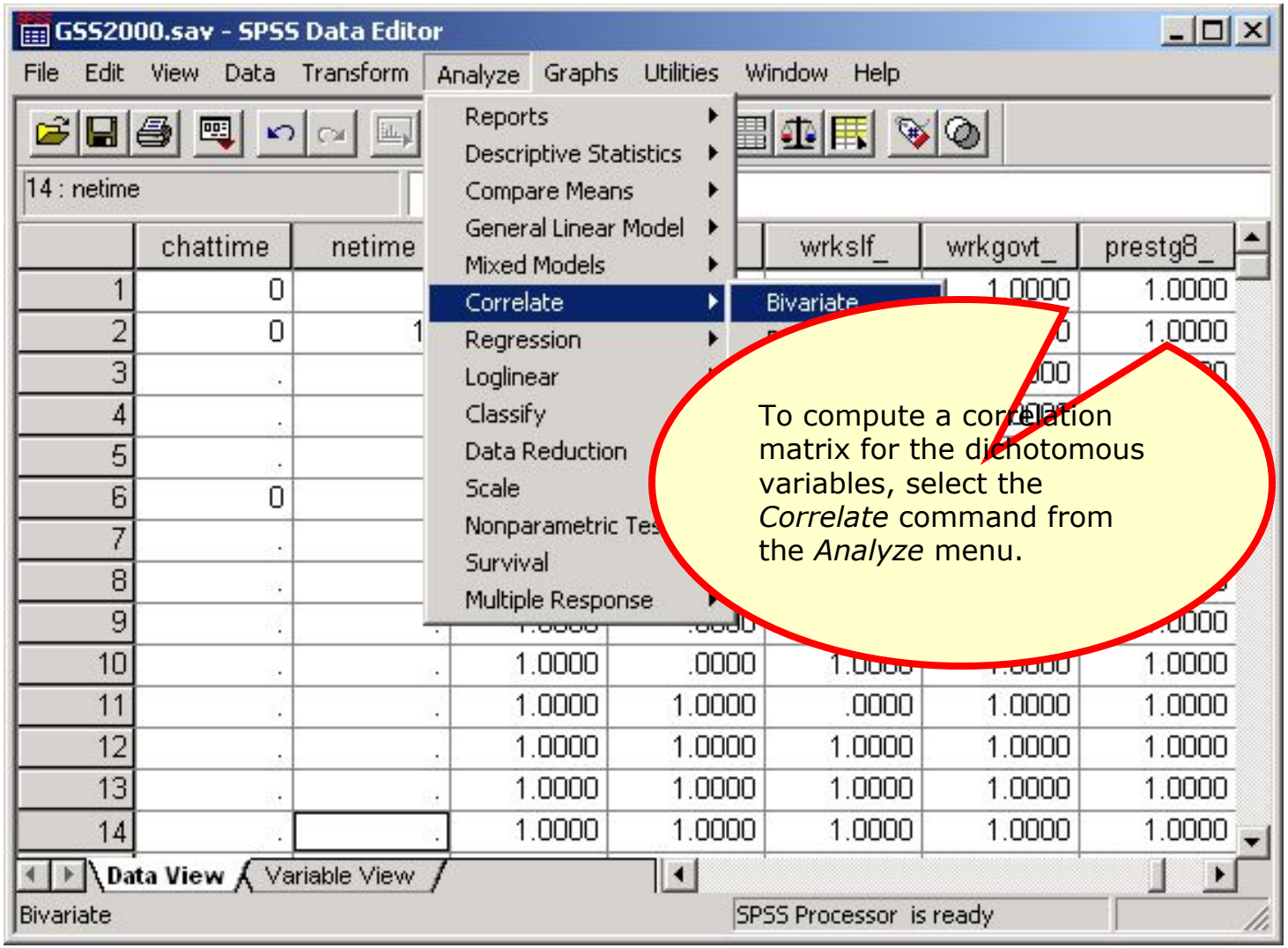
Complete the specifications for selecting cases



Cases excluded from further analyses

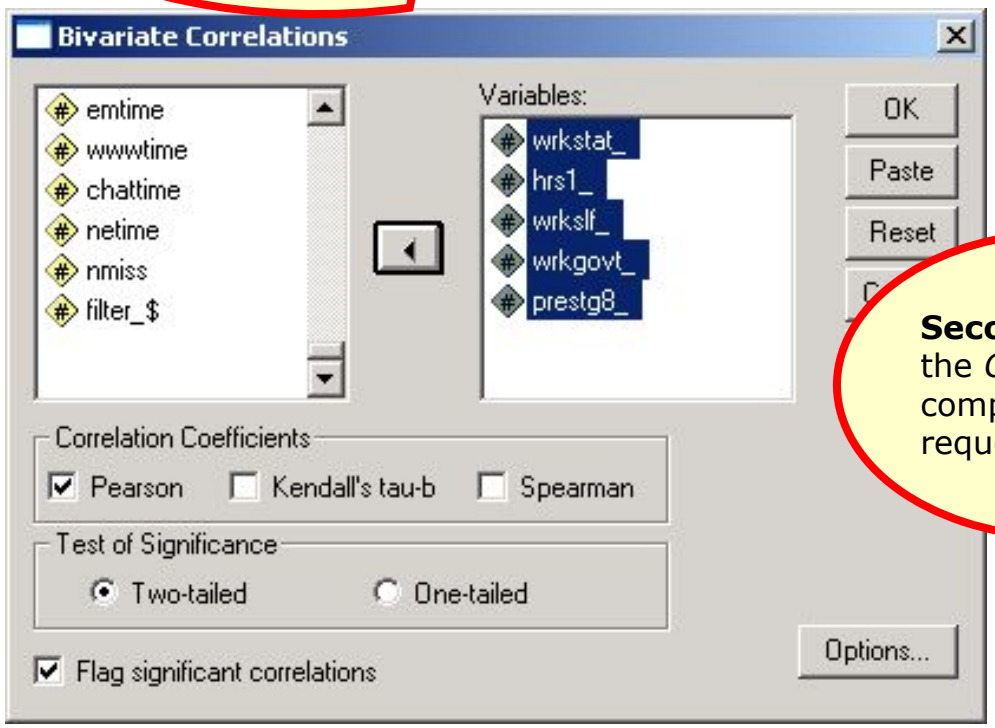


Correlating the dichotomous variables



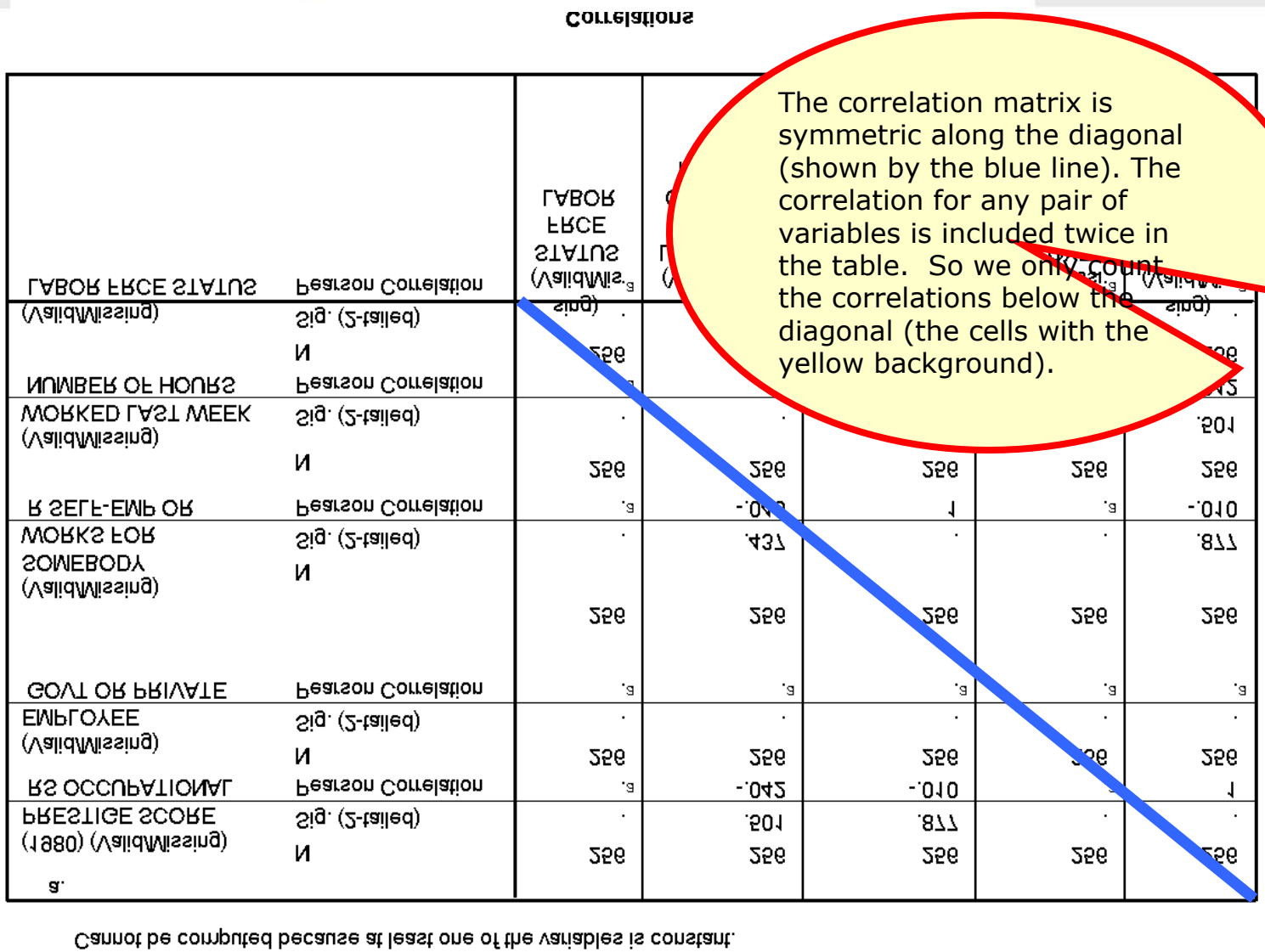
Specifications for correlations

First, move the dichotomous variables to the variables list box.



Second, click on the *OK* button to complete the request.

The correlation matrix



Correlations

		PRESTIGE SCORE (1980) (WalkWise)	R'S OCCUPATIONAL	GOVT OR PRIVATE	WORKS FOR SOMEBODY (WalkWise)	R SELF-EMP OR	WORKED LAST WEEK (WalkWise)	NUMBER OF HOURS	LABOR FRCE STATUS (WalkWise)
PRESTIGE SCORE (1980) (WalkWise)	Pearson Correlation	1	-.070	.	.118	-.070	.102		.
	Sig. (2-tailed)	.	.042	.	.	.	.		.
	N	520	520	520	520	520	520	520	520
R'S OCCUPATIONAL	Pearson Correlation	-.070	1	.	.174	.	.		.
	Sig. (2-tailed)	.	.	.	.	.	.		.
	N	520	520	520	520	520	520	520	520
EMPLOYEE (WalkWise)	Pearson Correlation	.	.	.	.	.	.		.
	Sig. (2-tailed)	.	.	.	.	.	.		.
	N	520	520	520	520	520	520	520	520
GOVT OR PRIVATE	Pearson Correlation	.	.	.	.	.	.		.
	Sig. (2-tailed)	.	.	.	.	.	.		.
	N	520	520	520	520	520	520	520	520
WORKS FOR SOMEBODY (WalkWise)	Pearson Correlation	.118	.174	.	1	.	.		.
	Sig. (2-tailed)	.	.	.	.	.	.		.
	N	520	520	520	520	520	520	520	520
R SELF-EMP OR	Pearson Correlation	-.070	.	.	.	1	.		.
	Sig. (2-tailed)	.	.	.	.	.	.		.
	N	520	520	520	520	520	520	520	520
WORKED LAST WEEK (WalkWise)	Pearson Correlation	.102	.	.	.	.	1		.
	Sig. (2-tailed)	.	.	.	.	.	.		.
	N	520	520	520	520	520	520	520	520
NUMBER OF HOURS	Pearson Correlation								
	Sig. (2-tailed)								
	N	520							
LABOR FRCE STATUS (WalkWise)	Pearson Correlation	.	.	.	.	.	.		1
	Sig. (2-tailed)	.	.	.	.	.	.		.
	N	520	520	520	520	520	520	520	520

a. Cannot be computed because at least one of the variables is constant.

The correlation matrix is symmetric along the diagonal (shown by the blue line). The correlation for any pair of variables is included twice in the table. So we only count the correlations below the diagonal (the cells with the yellow background).

The correlation matrix

Correlations

		LABOR FORCE STATUS (ValidMiss.)					
LABOR FORCE STATUS	Pearson Correlation						
(ValidMissing)	Sig. (2-tailed)	.					
	N	526					
NUMBER OF HOURS	Pearson Correlation	.a					
WORKED LAST WEEK (ValidMissing)	Sig. (2-tailed)	.					
	N	526	526				
R SELF-EMP OR	Pearson Correlation	.a	-.040				
WORKS FOR SOMEBODY (ValidMissing)	Sig. (2-tailed)	.	.431	.		.718	
	N	526	526	526	526	526	
GOVT OR PRIVATE	Pearson Correlation	.a	.a	.a	.a	.a	
EMPLOYEE (ValidMissing)	Sig. (2-tailed)	.	.	.	.	.	
	N	526	526	526	526	526	
R'S OCCUPATIONAL	Pearson Correlation	.a	-.045	-.070	.a	1	
PRESTIGE SCORE (1980) (ValidMissing)	Sig. (2-tailed)	.	.301	.117	.	.	
	N	526	526	526	526	526	

a. Cannot be computed because at least one of the variables is constant.

The correlations marked with footnote a could not be computed because one of the variables was a constant, i.e. the dichotomous variable has the same value for all cases.

This happens when one of the valid/missing variables has no missing cases, so that all of the cases have a value of 1 and none have a value of 0.

The correlation matrix

In the cells for which the correlation could be computed, the probabilities indicating significance are 0.437, 0.501, and 0.877.

None of the correlations are statistically significant. The answer to the question is **false**. We do not need to be concerned about a missing data problem for this set of variables.

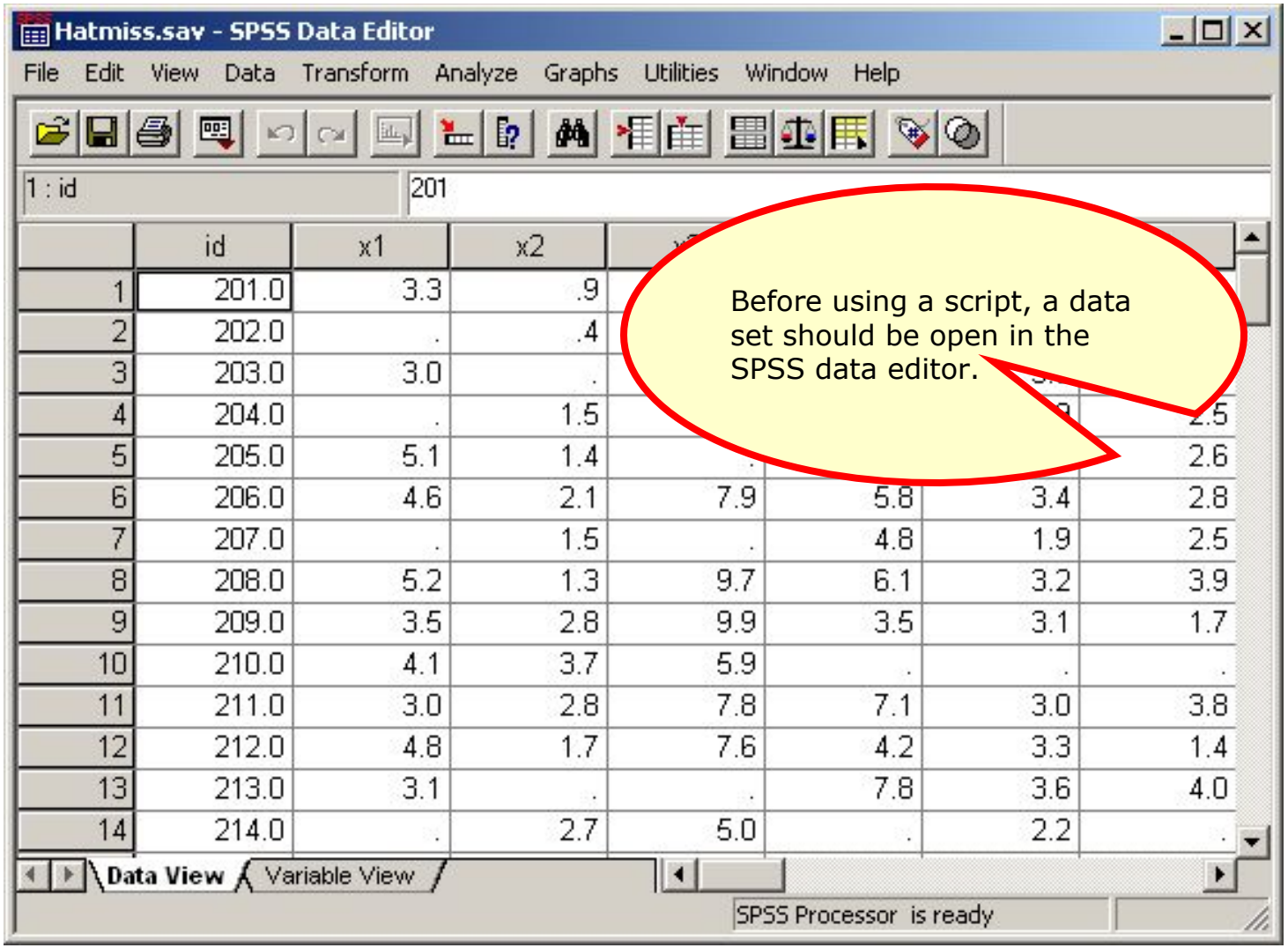
Using scripts

- The process of evaluating missing data requires numerous SPSS procedures and outputs that are time consuming to produce.
- These procedures can be automated by creating an SPSS script. A script is a program that executes a sequence of SPSS commands.
- Thought writing scripts is not part of this course, we can take advantage of scripts that I use to reduce the burdensome tasks of evaluating missing data.

Using a script for missing data

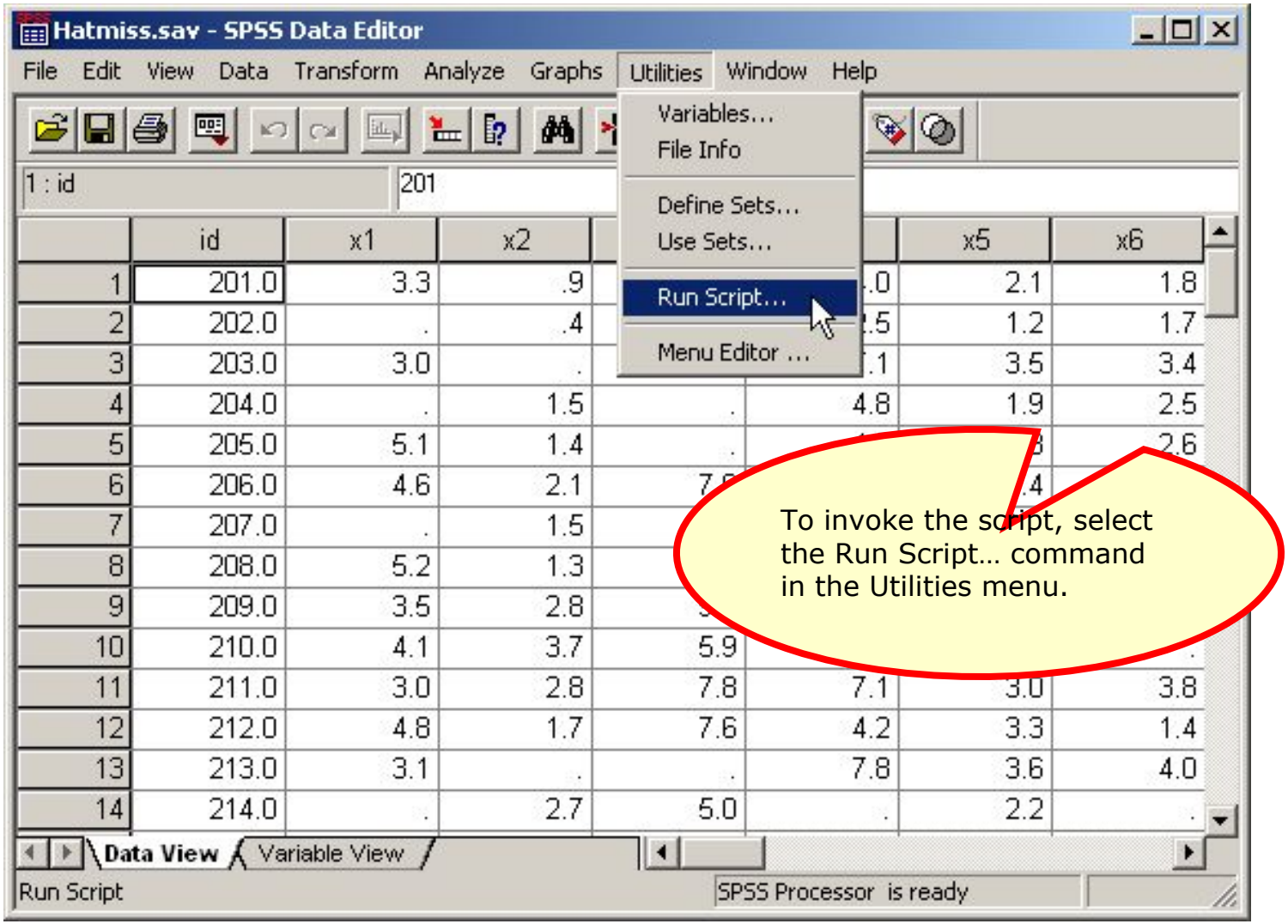
- The script “MissingDataCheck.sbs” will produce all of the output we have used for evaluating missing data, as well as other outputs described in the textbook.
- Navigate to the link “SPSS Scripts and Syntax” on the course web page.
- Download the script file “MissingDataCheck.exe” to your computer and install it, following the directions on the web page.

Open the data set in SPSS



	id	x1	x2				
1	201.0	3.3	.9				
2	202.0	.	.4				
3	203.0	3.0	.				
4	204.0	.	1.5				2.5
5	205.0	5.1	1.4				2.6
6	206.0	4.6	2.1	7.9	5.8	3.4	2.8
7	207.0	.	1.5	.	4.8	1.9	2.5
8	208.0	5.2	1.3	9.7	6.1	3.2	3.9
9	209.0	3.5	2.8	9.9	3.5	3.1	1.7
10	210.0	4.1	3.7	5.9	.	.	.
11	211.0	3.0	2.8	7.8	7.1	3.0	3.8
12	212.0	4.8	1.7	7.6	4.2	3.3	1.4
13	213.0	3.1	.	.	7.8	3.6	4.0
14	214.0	.	2.7	5.0	.	2.2	.

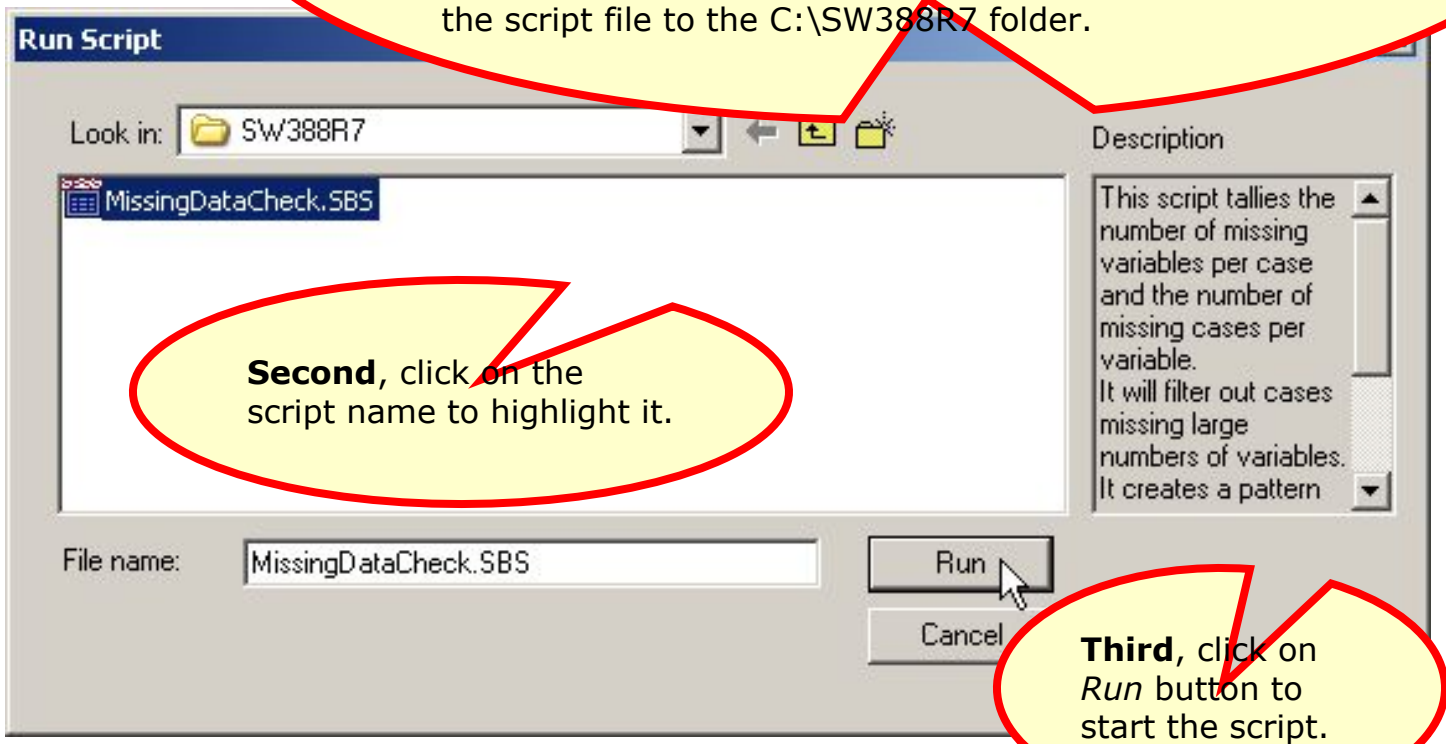
Invoke the script



Select the missing data script

First, navigate to the folder where you put the script. If you followed the directions, you will have a file with an ".SBS" extension in the C:\SW388R7 folder.

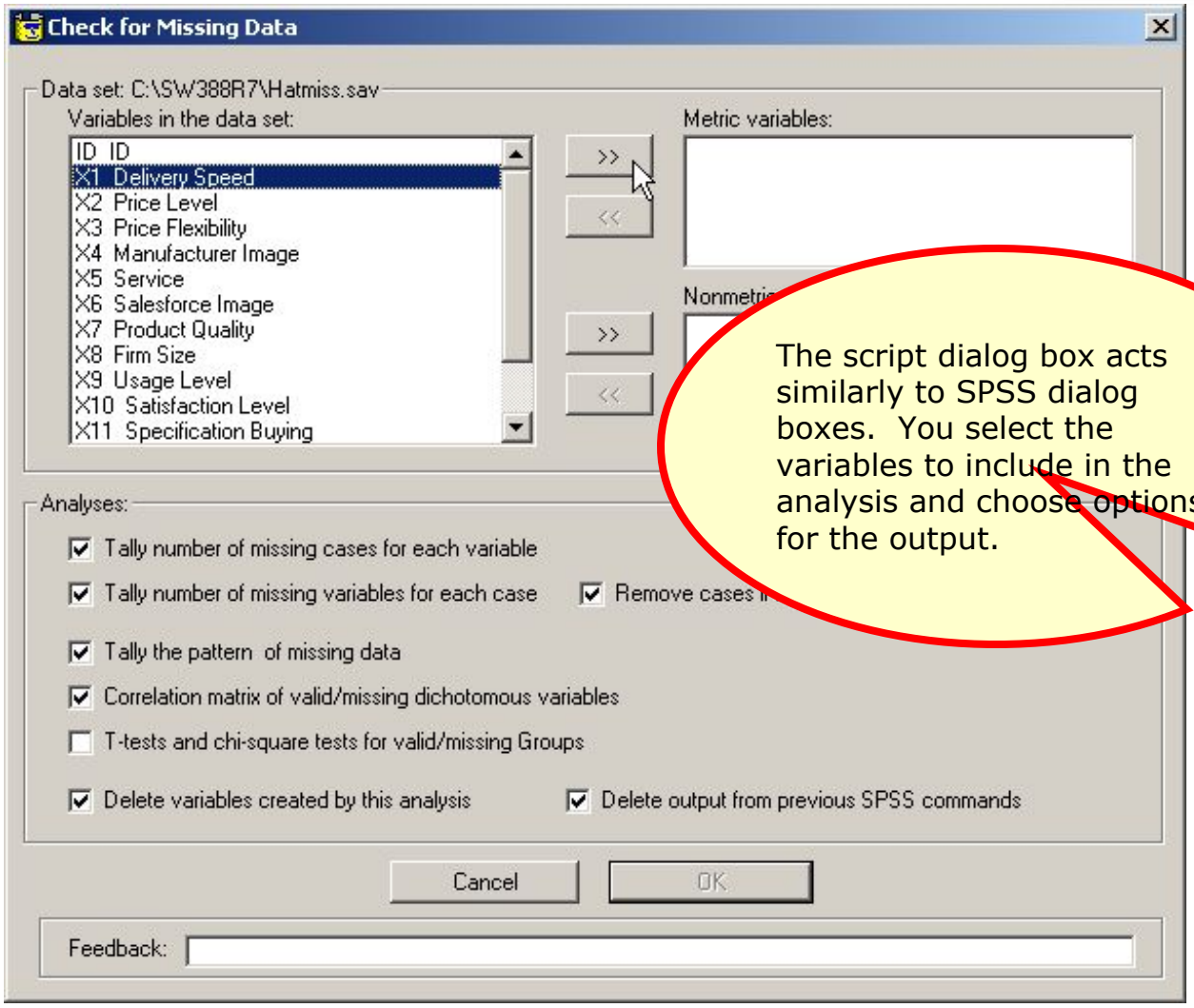
If you only see a file with an ".EXE" extension in the folder, you should double click on that file to extract the script file to the C:\SW388R7 folder.



Second, click on the script name to highlight it.

Third, click on *Run* button to start the script.

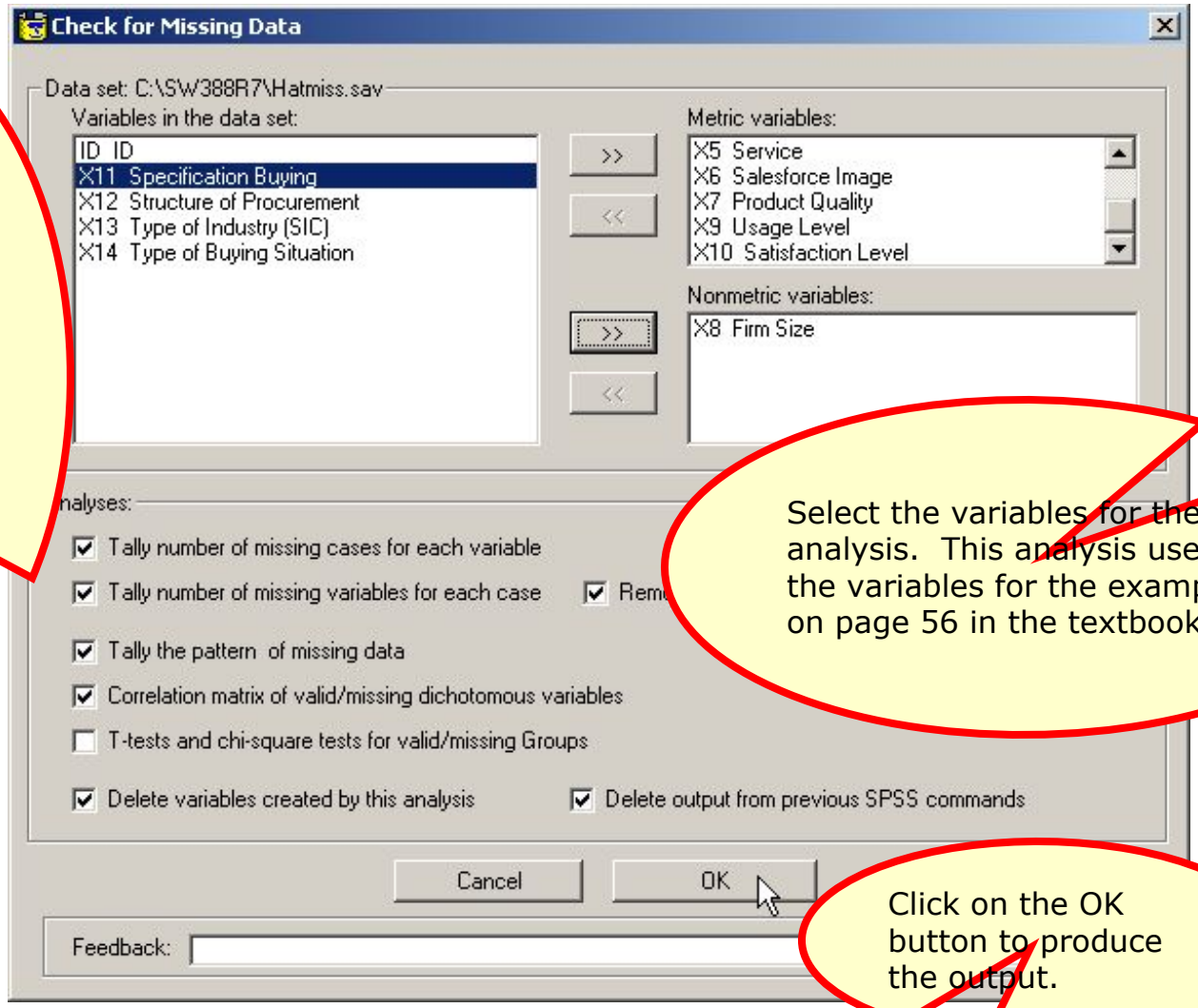
The script dialog



The script dialog box acts similarly to SPSS dialog boxes. You select the variables to include in the analysis and choose options for the output.

Complete the specifications

The checkboxes are marked to produce the output we need for our problems. The only additional option is to compute the t-tests and chi-square tests for all of the variables.



Select the variables for the analysis. This analysis uses the variables for the example on page 56 in the textbook.

Click on the OK button to produce the output.

The script finishes

If you SPSS output viewer is open, you will see the output produced in that window.

Since it may take a while to produce the output, and since there are times when it appears that nothing is happening, there is an alert to tell you when the script is finished.

Unless you are absolutely sure something has gone wrong, let the script run until you see this alert.

When you see this alert, click on the OK button.

Output from the script