# 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.

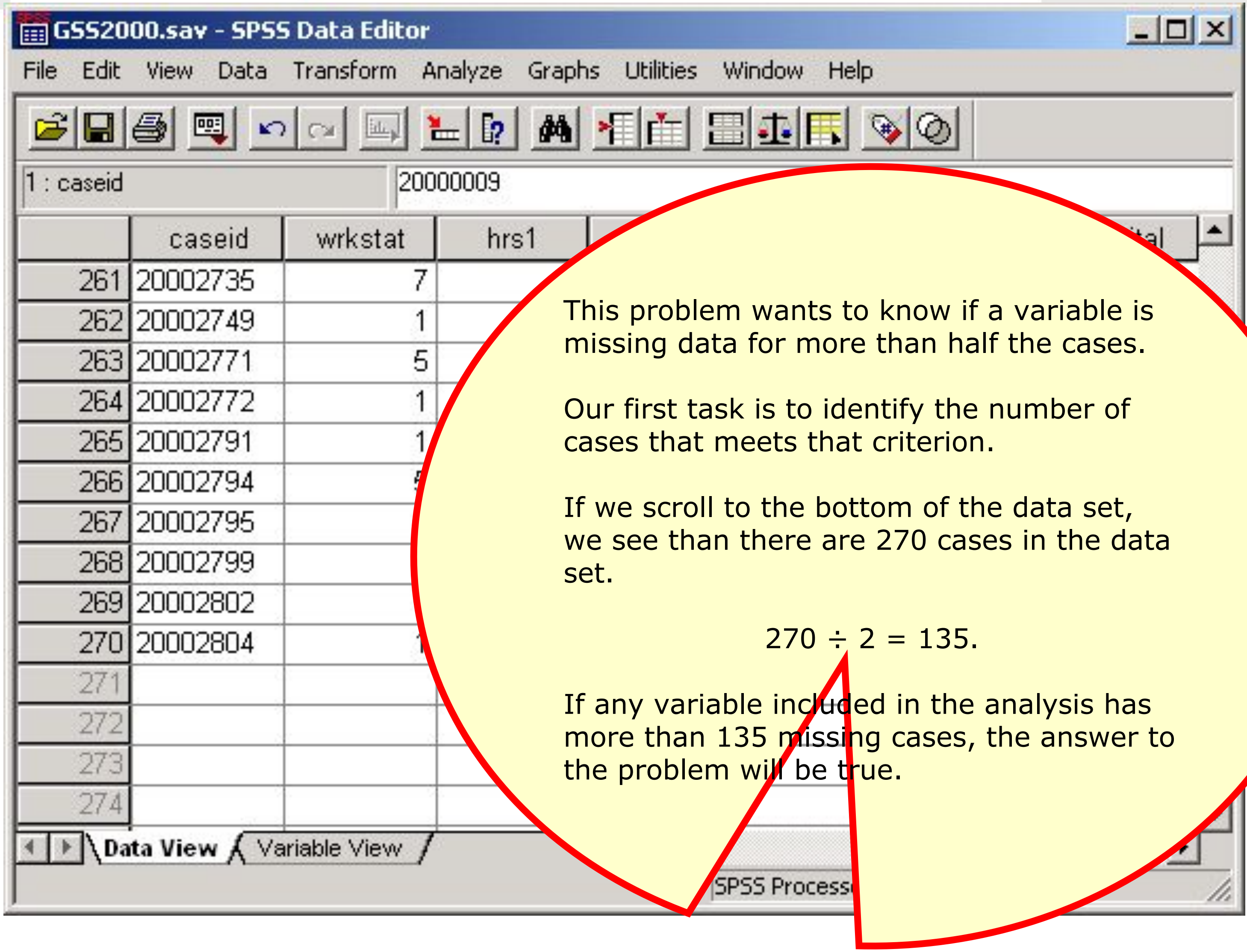
# Problem 1

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


GSS2000.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Process

1 : caseid 20000009

|  | caseid | wrkstat | hrs1 |
|---|---|---|---|
| 261 | 20002735 | 7 |  |
| 262 | 20002749 | 1 |  |
| 263 | 20002771 | 5 |  |
| 264 | 20002772 | 1 |  |
| 265 | 20002791 | 1 |  |
| 266 | 20002794 | 5 |  |
| 267 | 20002795 |  |  |
| 268 | 20002799 |  |  |
| 269 | 20002802 |  |  |
| 270 | 20002804 | 1 |  |
| 271 |  |  |  |
| 272 |  |  |  |
| 273 |  |  |  |
| 274 |  |  |  |

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

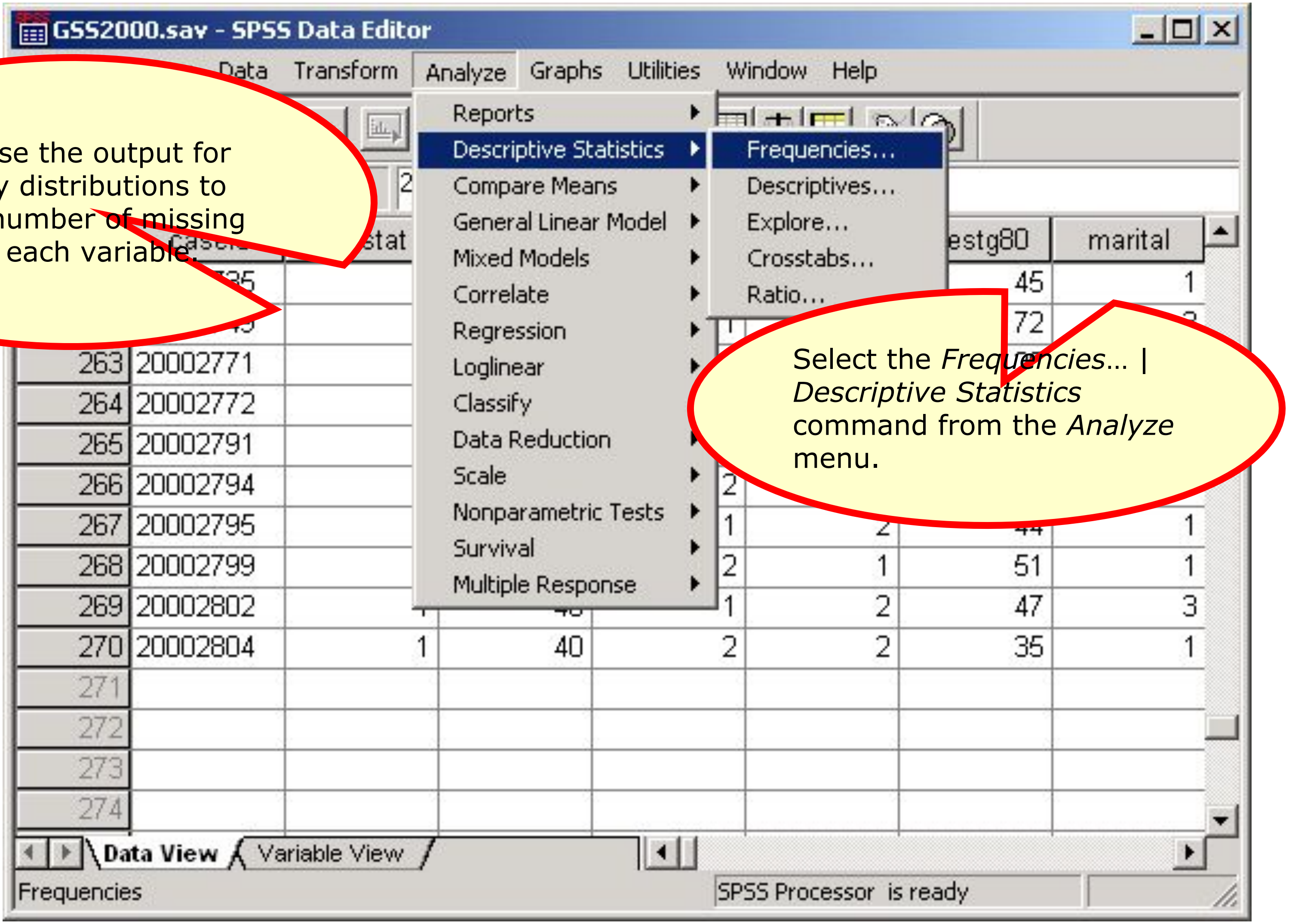
Our first task is to identify the number of cases that meets that criterion.

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

270 ÷ 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

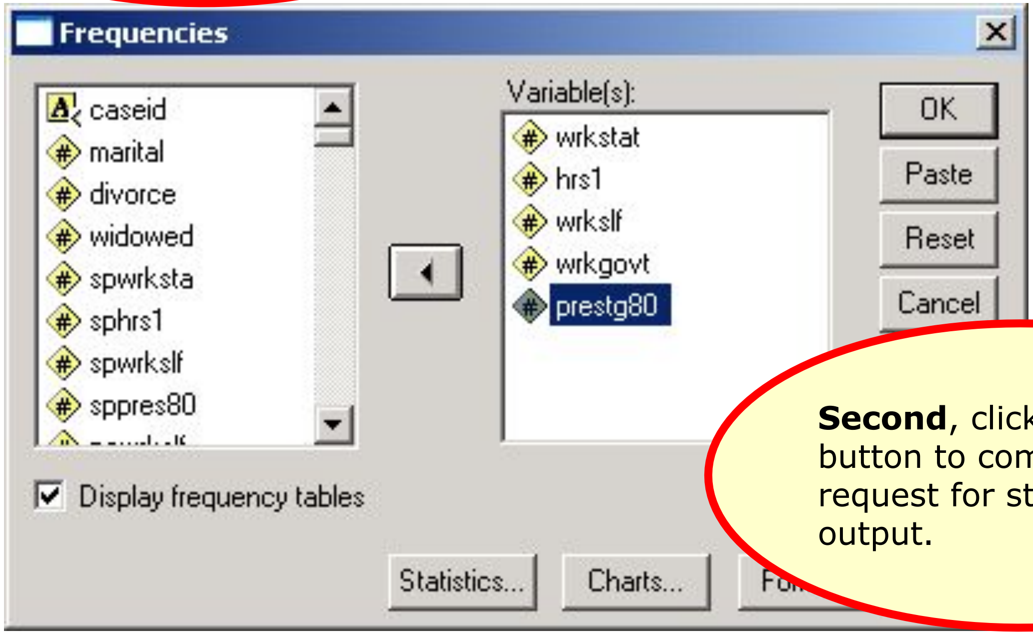


We will use the output for frequency distributions to find the number of missing cases for each variable.

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.



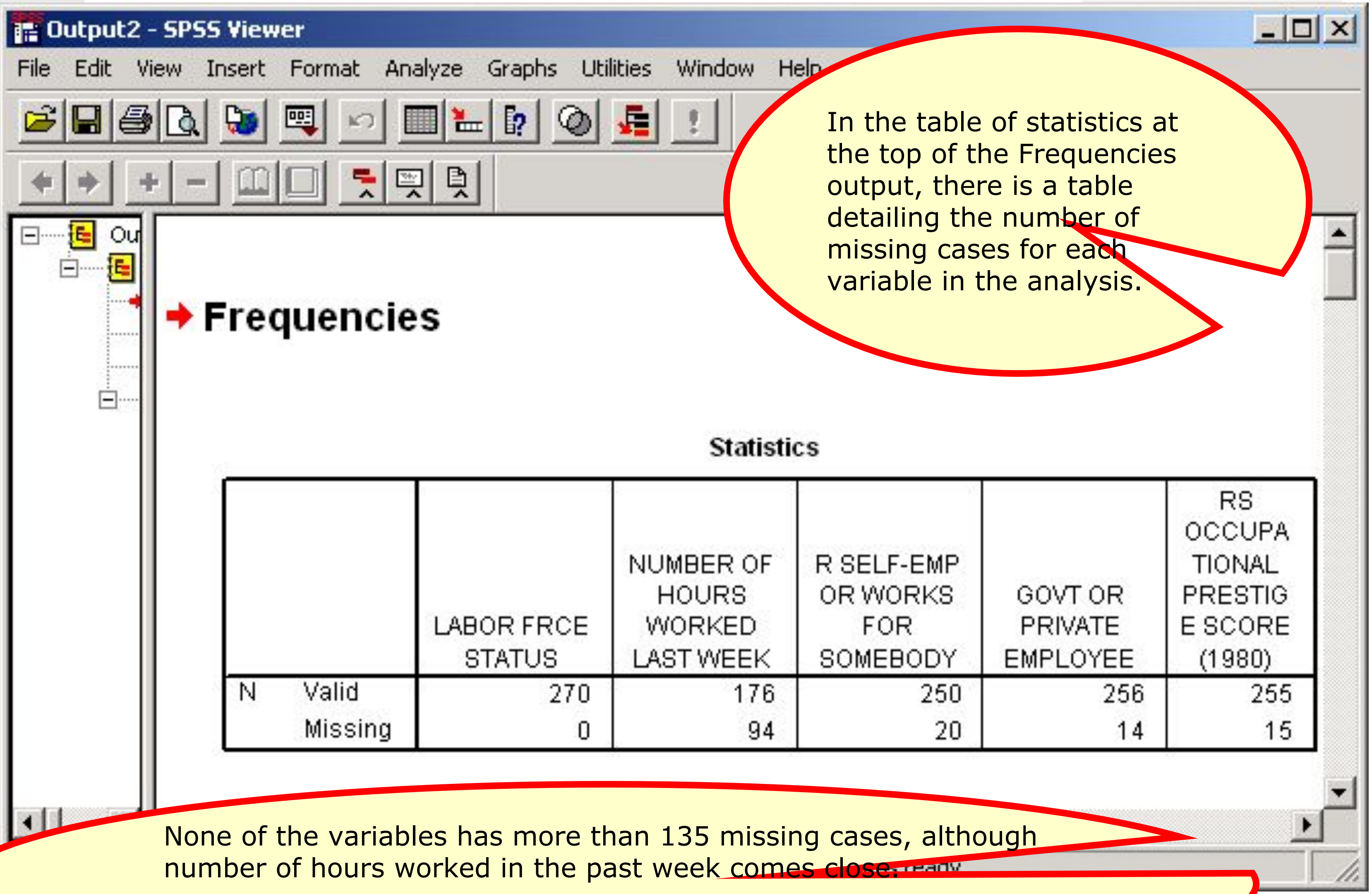
caseid
marital
divorce
widowed
spwrksta
sphrs1
spwrkslf
sppres80

Variable(s):
wrkstat
hrs1
wrkslf
wrkgovt
prestg80

☑ Display frequency tables

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

# Number of missing cases for each variable


Output2 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Ou

## Frequencies

**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 |

In the table of statistics at the top of the Frequencies output, there is a table detailing the number of missing cases for each variable in the analysis.

None of the variables has more than 135 missing cases, although number of hours worked in the past week comes close.

The answer to the question is **false**.

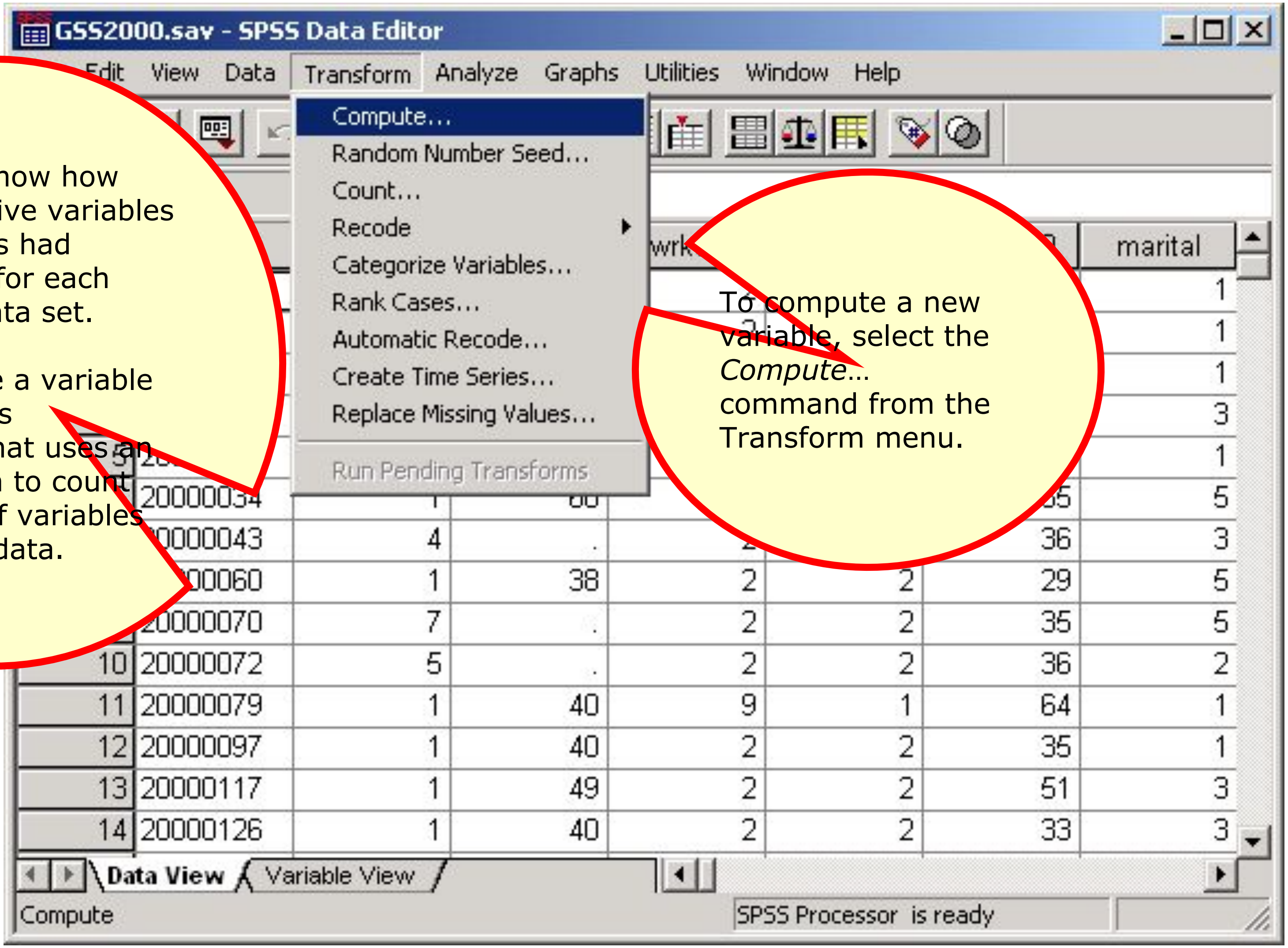
# Problem 2

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



| | | | | wrk | | | marital |
|---|---|---|---|---|---|---|---|
| | | | | | | | 1 |
| | | | | | | | 1 |
| | | | | | | | 1 |
| | | | | | | | 3 |
| 5 | 20 | | | | | | 1 |
| | 20000034 | 1 | 60 | | | 35 | 5 |
| | 0000043 | 4 | . | 2 | | 36 | 3 |
| | 00060 | 1 | 38 | 2 | 2 | 29 | 5 |
| | 20000070 | 7 | . | 2 | 2 | 35 | 5 |
| 10 | 20000072 | 5 | . | 2 | 2 | 36 | 2 |
| 11 | 20000079 | 1 | 40 | 9 | 1 | 64 | 1 |
| 12 | 20000097 | 1 | 40 | 2 | 2 | 35 | 1 |
| 13 | 20000117 | 1 | 49 | 2 | 2 | 51 | 3 |
| 14 | 20000126 | 1 | 40 | 2 | 2 | 33 | 3 |

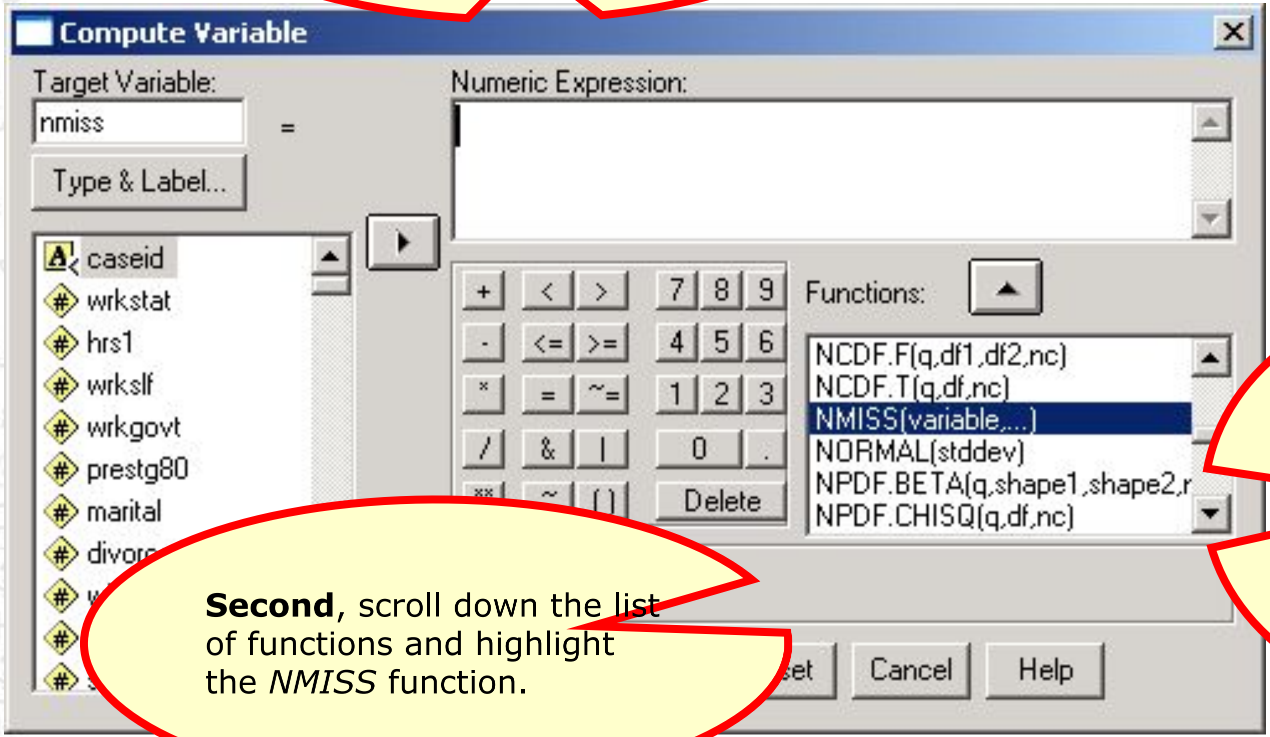
We want to know how many of the five variables in the analysis had missing data for each case in the data set.

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.



Compute Variable

Target Variable: nmiss

Numeric Expression:

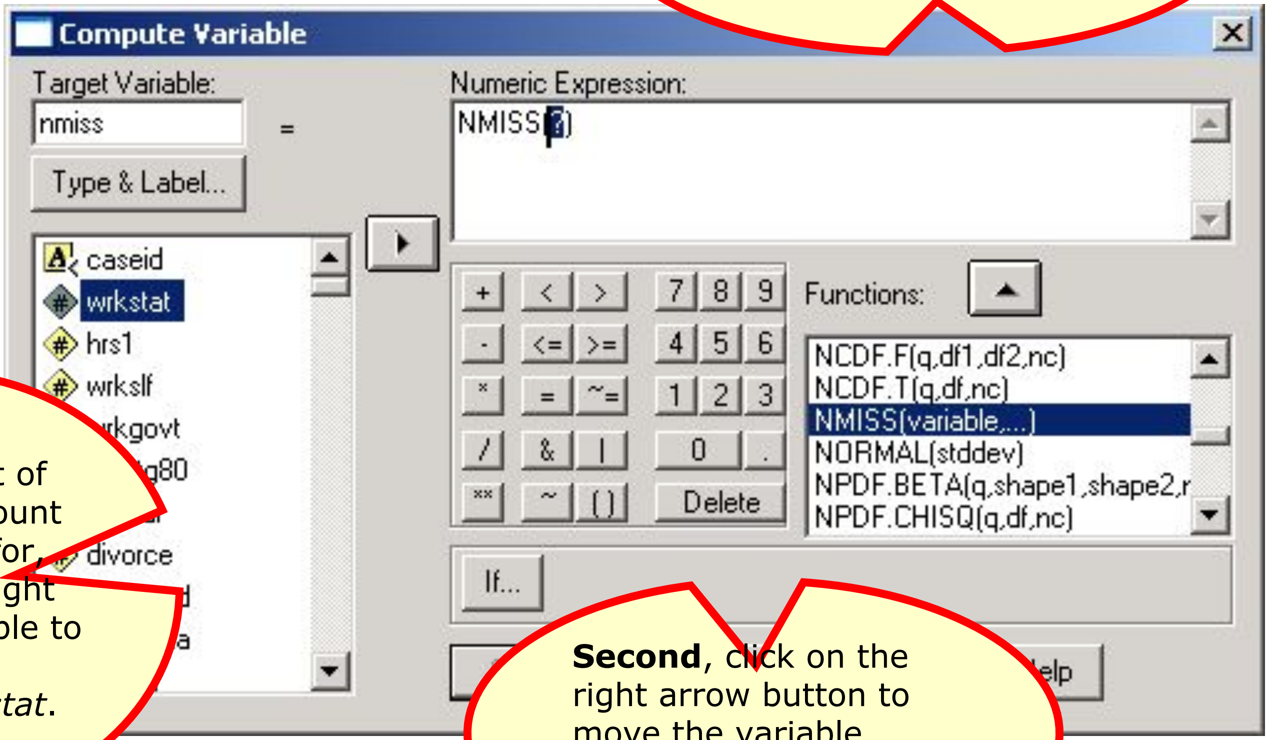
Functions:

NCDF.F(q,df1,df2,nc)
NCDF.T(q,df,nc)
NMISS(variable,...)
NORMAL(stddev)
NPDF.BETA(q,shape1,shape2,r
NPDF.CHISQ(q,df,nc)

**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



Compute Variable

Target Variable: nmiss

Numeric Expression: NMISS(?)

caseid
wrkstat
hrs1
wrkslf
wrkgovt

divorce

Functions:

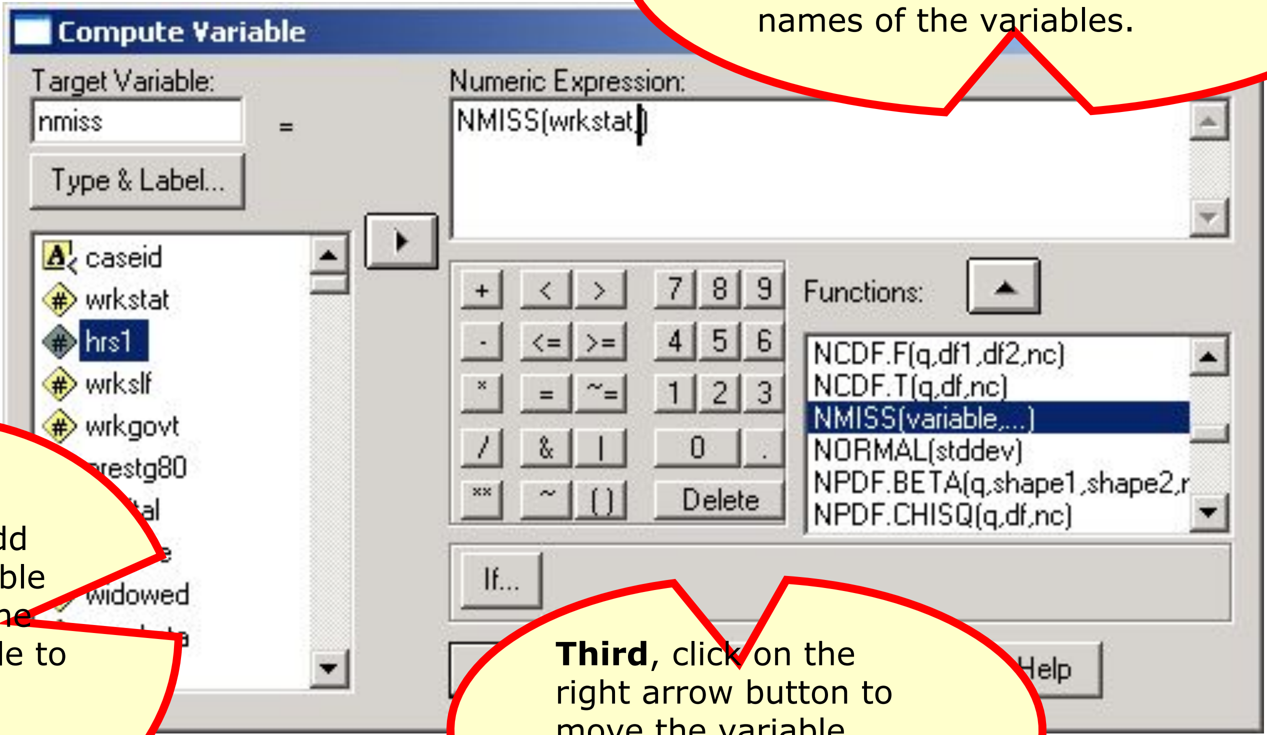
NCDF.F(q,df1,df2,nc)
NCDF.T(q,df,nc)
NMISS(variable,...)
NORMAL(stddev)
NPDF.BETA(q,shape1,shape2,r
NPDF.CHISQ(q,df,nc)

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



Target Variable: nmiss =

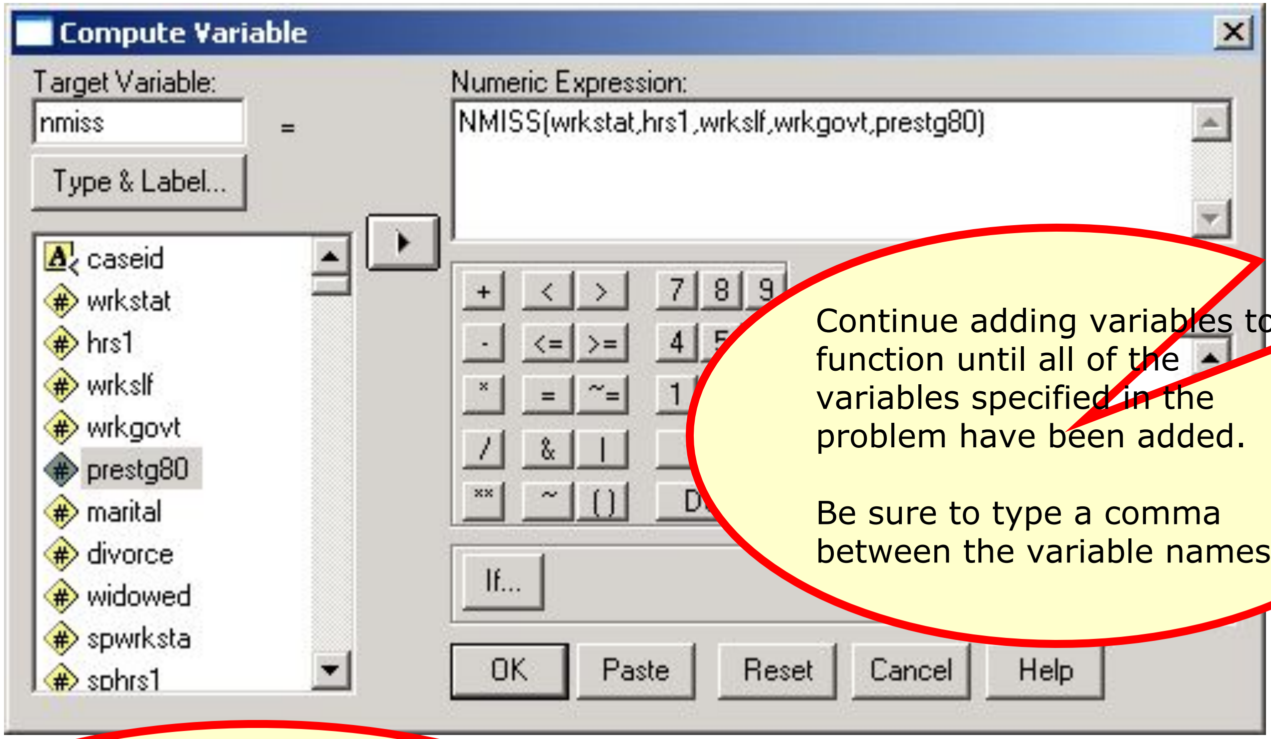
Numeric Expression: NMISS(wrkstat,)

**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



Compute Variable

Target Variable: nmiss

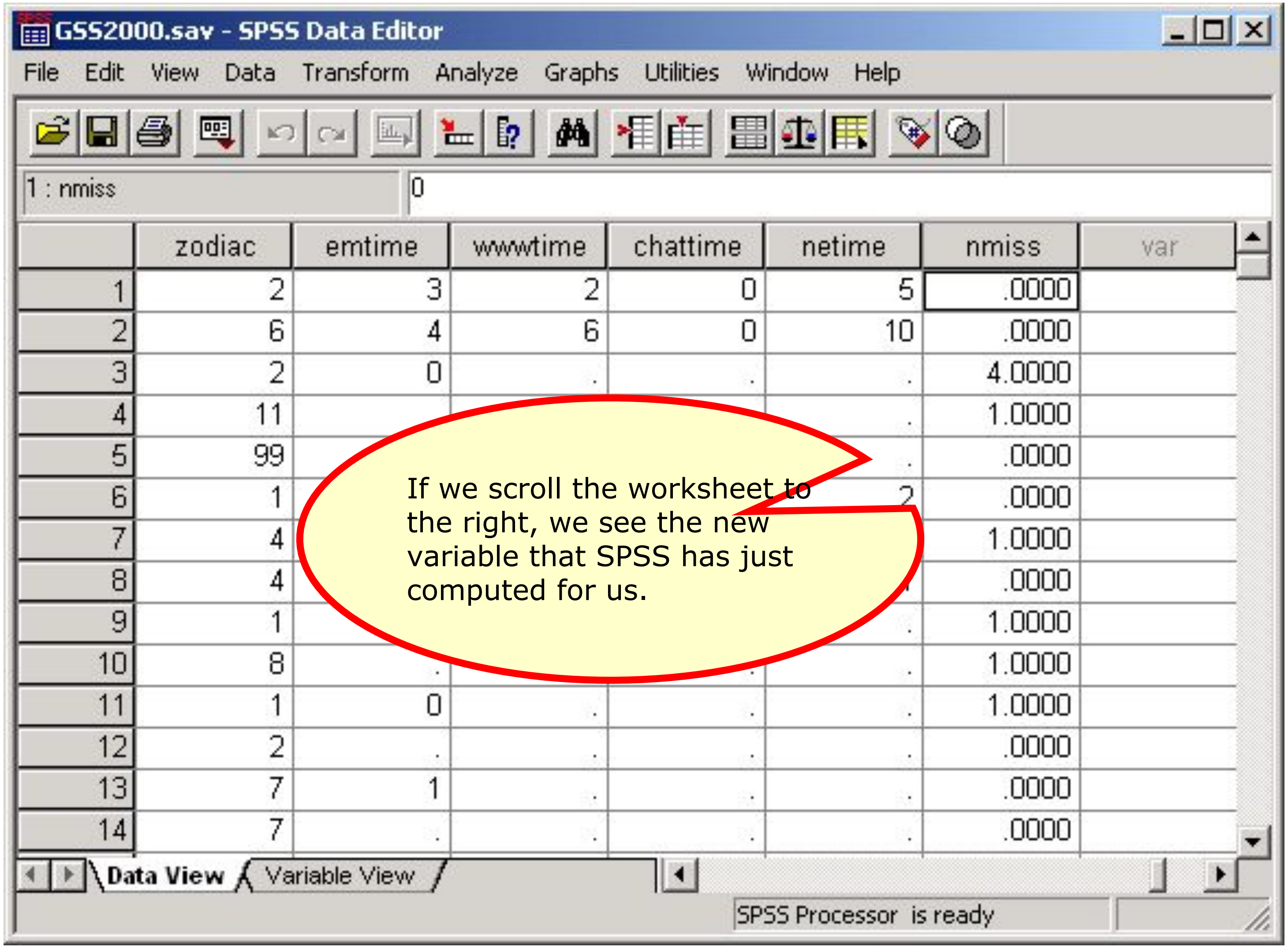
Numeric Expression: NMISS(wrkstat,hrs1,wrkslf,wrkgovt,prestg80)

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

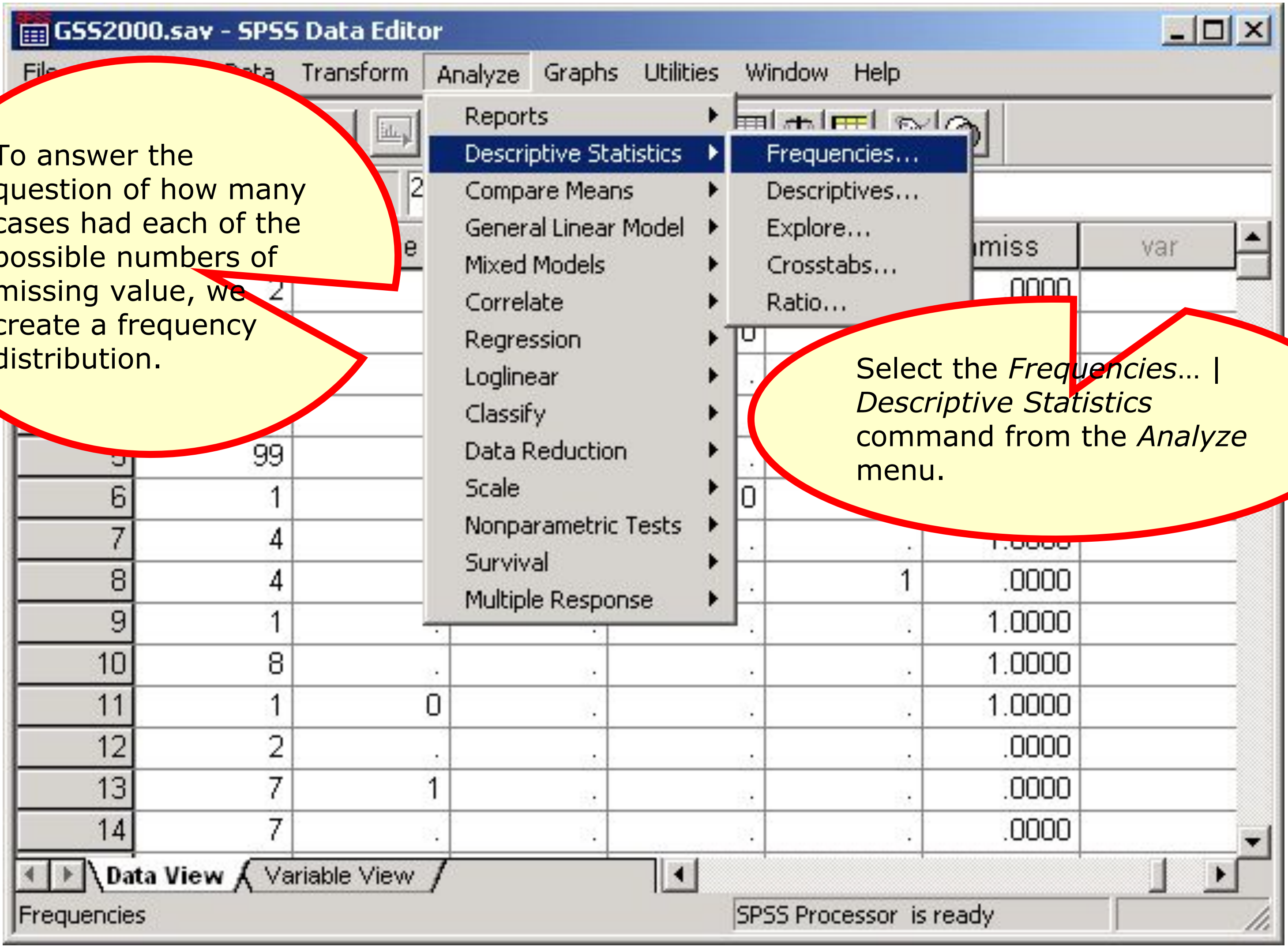

GSS2000.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

1 : nmiss 0

|  | zodiac | emtime | wwwtime | chattime | netime | nmiss | var |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 2 | 0 | 5 | .0000 |  |
| 2 | 6 | 4 | 6 | 0 | 10 | .0000 |  |
| 3 | 2 | 0 | . | . | . | 4.0000 |  |
| 4 | 11 |  |  |  | . | 1.0000 |  |
| 5 | 99 |  |  |  | . | .0000 |  |
| 6 | 1 |  |  |  | 2 | .0000 |  |
| 7 | 4 |  |  |  |  | 1.0000 |  |
| 8 | 4 |  |  |  |  | .0000 |  |
| 9 | 1 |  |  |  | . | 1.0000 |  |
| 10 | 8 | . |  | . | . | 1.0000 |  |
| 11 | 1 | 0 | . | . | . | 1.0000 |  |
| 12 | 2 | . | . | . | . | .0000 |  |
| 13 | 7 | 1 | . | . | . | .0000 |  |
| 14 | 7 | . | . | . | . | .0000 |  |

If we scroll the worksheet to the right, we see the new variable that SPSS has just computed for us.

# A frequency distribution for *nmiss*



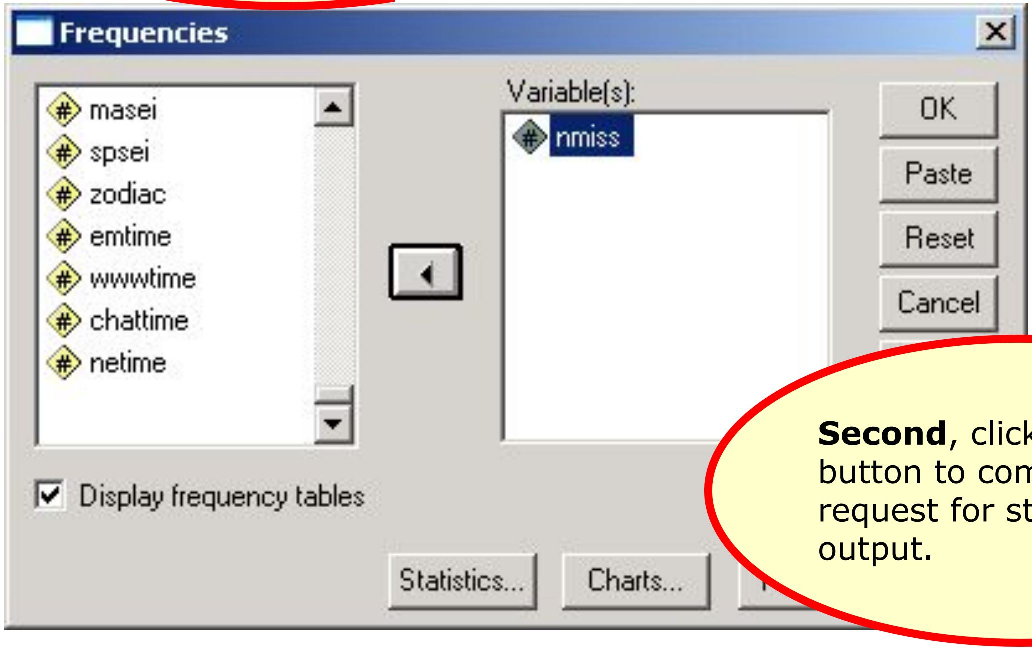
|  |  |  |  |  |  | miss | var |
|---|---|---|---|---|---|---|---|
| 5 | 99 |  |  |  |  |  |  |
| 6 | 1 |  |  | 0 |  |  |  |
| 7 | 4 |  |  | . | . |  |  |
| 8 | 4 |  |  | . | 1 | .0000 |  |
| 9 | 1 | . | . | . | . | 1.0000 |  |
| 10 | 8 | . | . | . | . | 1.0000 |  |
| 11 | 1 | 0 | . | . | . | 1.0000 |  |
| 12 | 2 | . | . | . | . | .0000 |  |
| 13 | 7 | 1 | . | . | . | .0000 |  |
| 14 | 7 | . | . | . | . | .0000 |  |

To answer the question of how many cases had each of the possible numbers of missing value, we create a frequency distribution.

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

# Completing the specification for frequencies

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

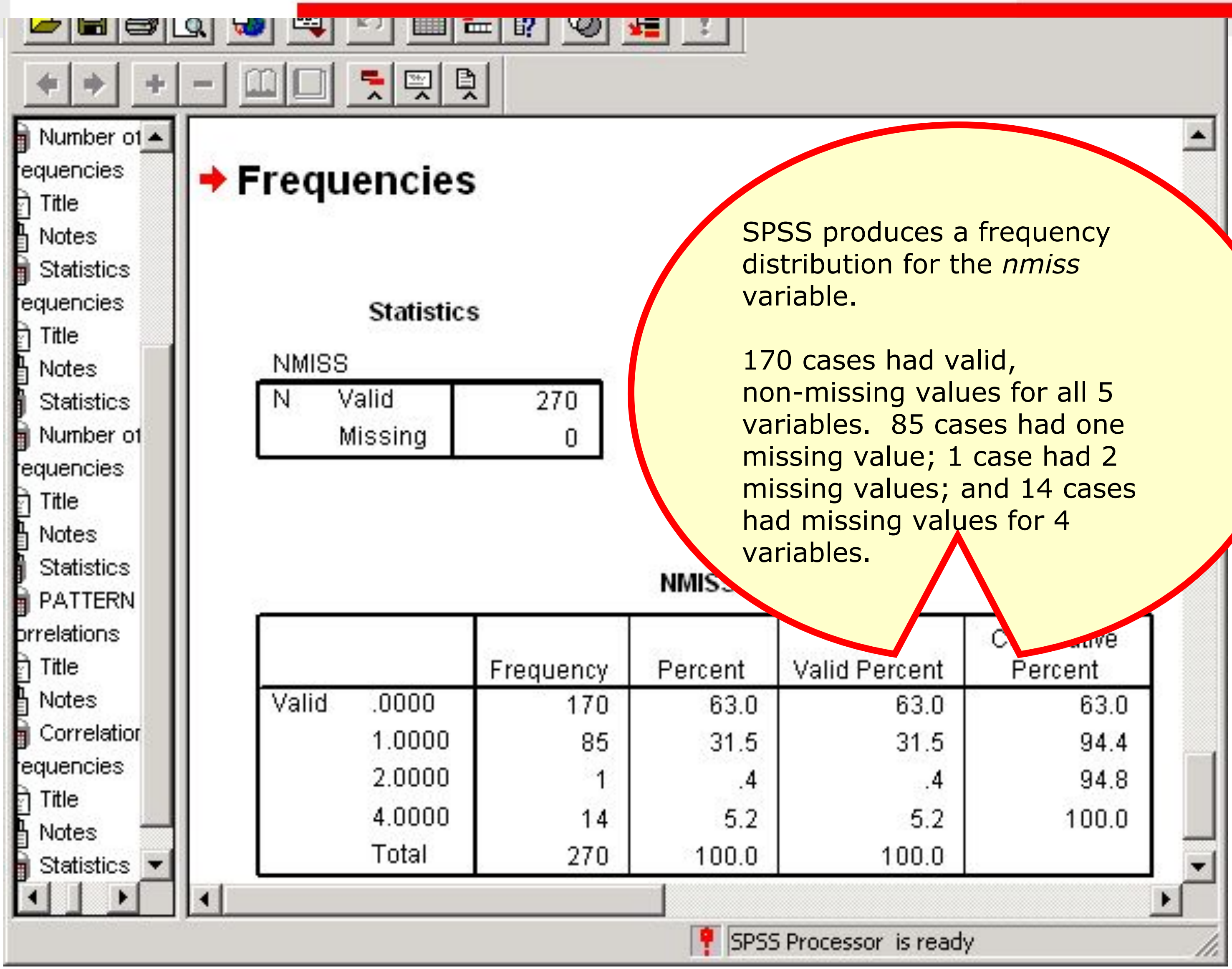


Variable(s): nmiss

☑ Display frequency tables

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

# The frequency distribution



## Frequencies

**Statistics**

NMISS

| | | |
|---|---|---|
| N | Valid | 270 |
| | Missing | 0 |

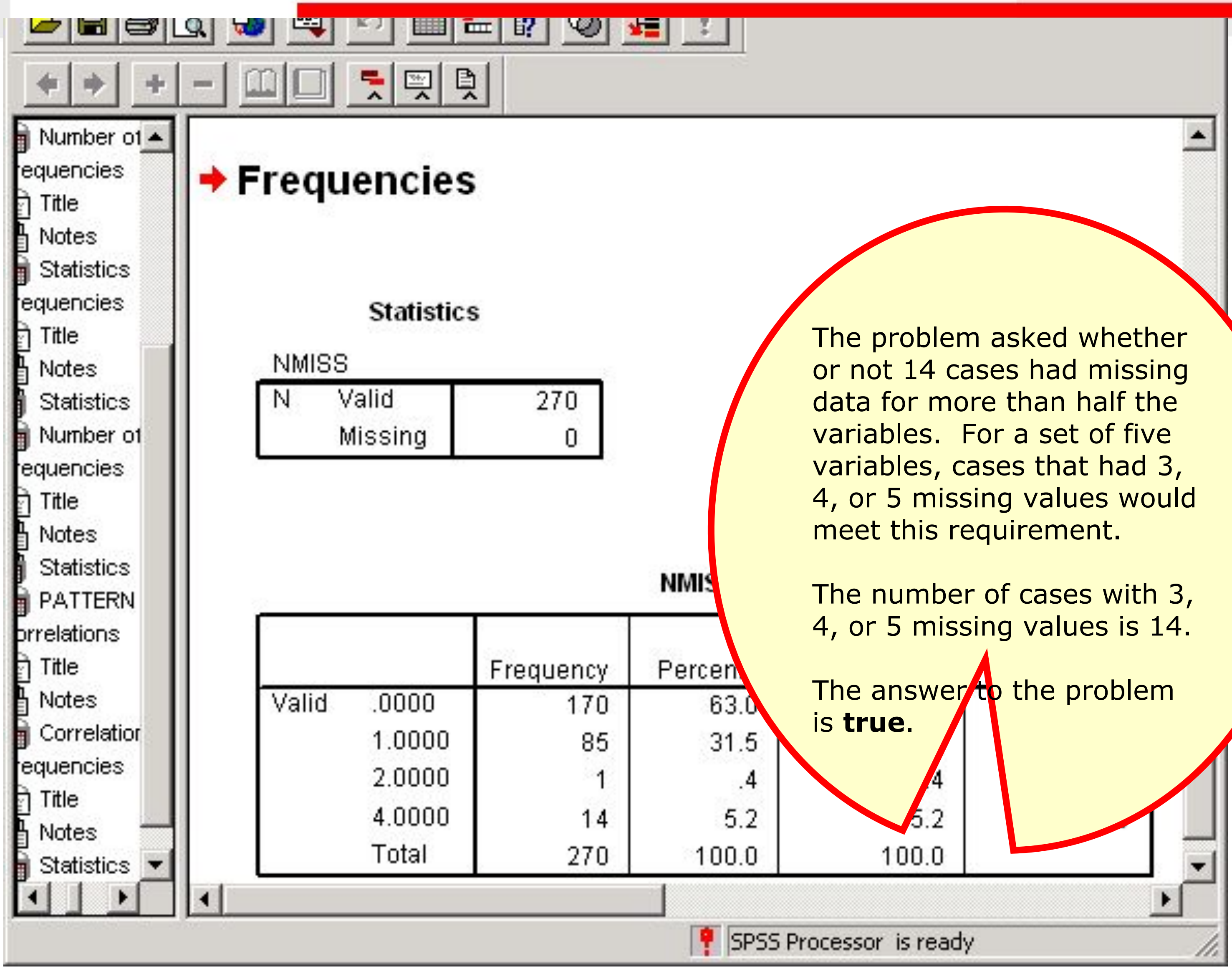
**NMISS**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | .0000 | 170 | 63.0 | 63.0 | 63.0 |
| | 1.0000 | 85 | 31.5 | 31.5 | 94.4 |
| | 2.0000 | 1 | .4 | .4 | 94.8 |
| | 4.0000 | 14 | 5.2 | 5.2 | 100.0 |
| | Total | 270 | 100.0 | 100.0 | |

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



## Frequencies

**Statistics**

NMISS

| | | |
|---|---|---|
| N | Valid | 270 |
| | Missing | 0 |

**NMIS**

| | | Frequency | Percen | |
|---|---|---|---|---|
| Valid | .0000 | 170 | 63.0 | |
| | 1.0000 | 85 | 31.5 | |
| | 2.0000 | 1 | .4 | .4 |
| | 4.0000 | 14 | 5.2 | 5.2 |
| | Total | 270 | 100.0 | 100.0 |

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**.

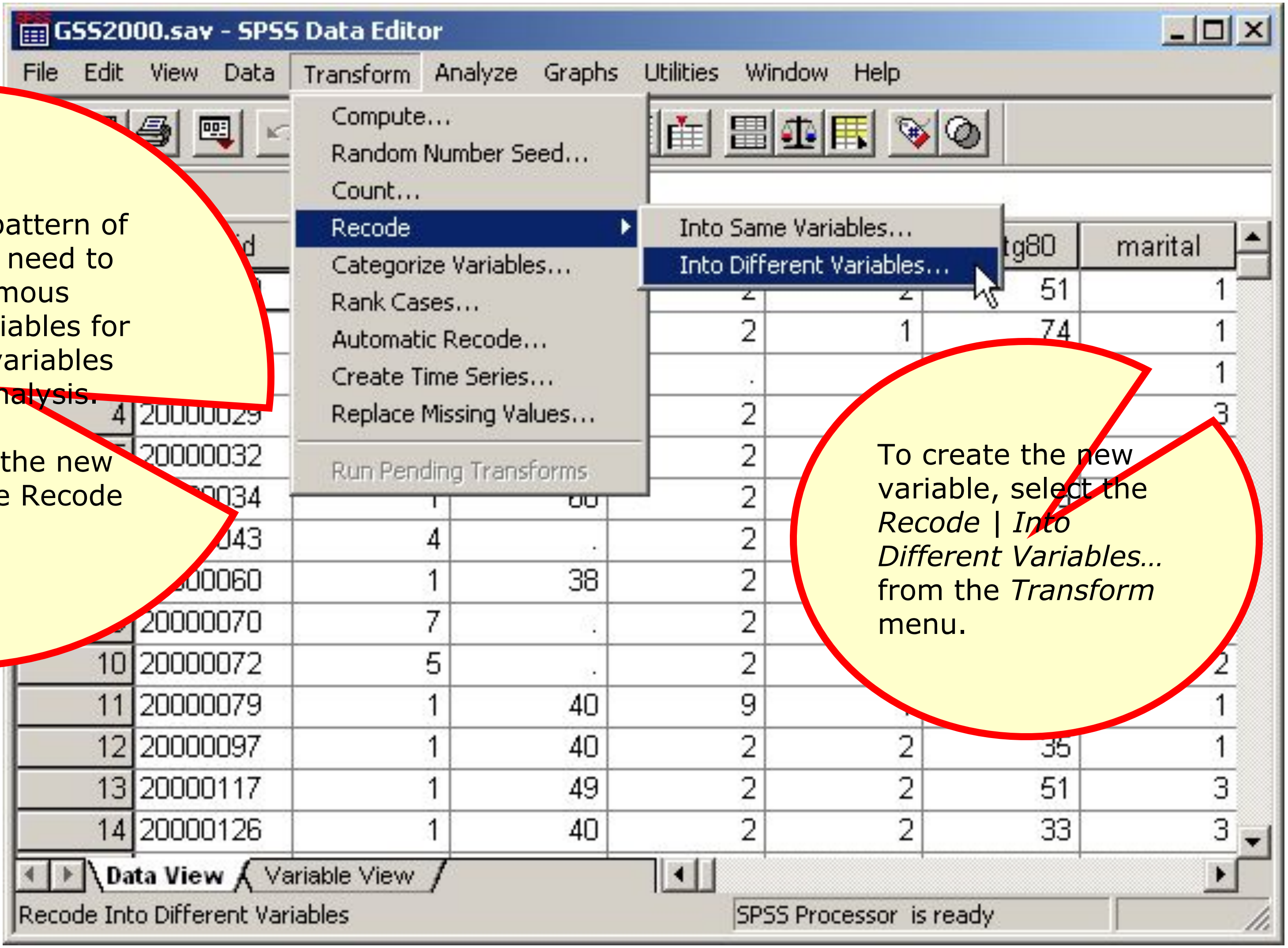
# Problem 3

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

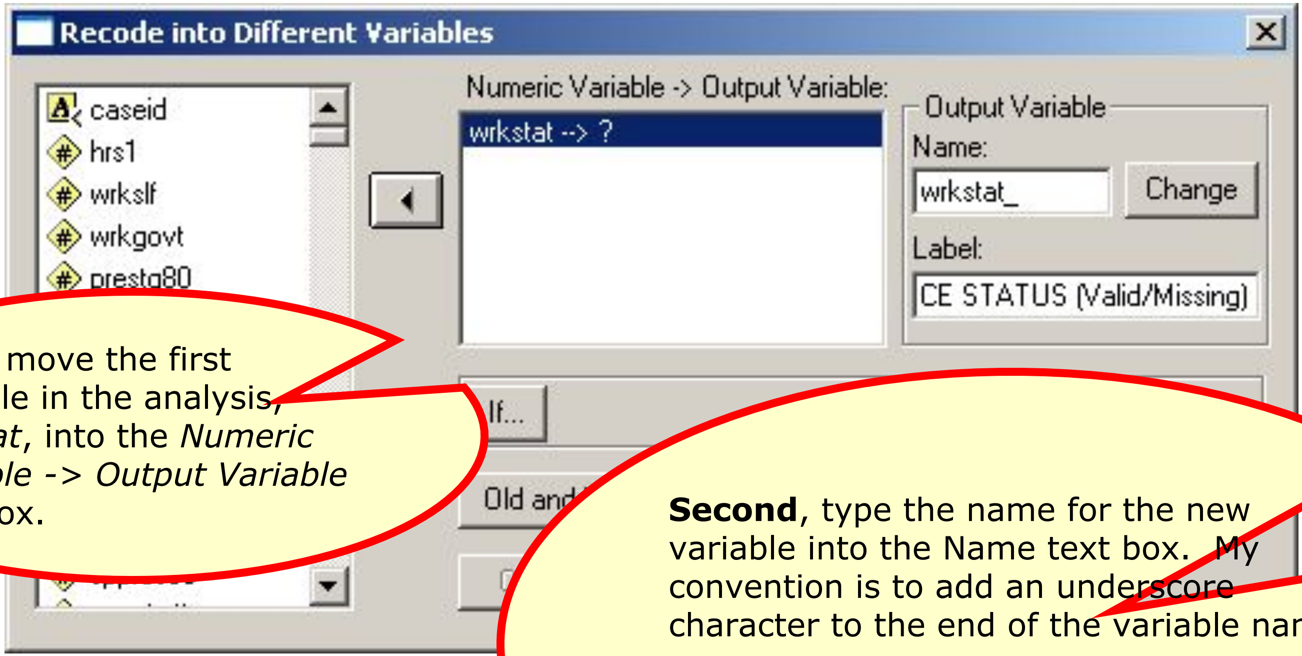


To evaluate the pattern of missing data, we need to compute dichotomous valid/missing variables for each of the five variables included in the analysis.

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

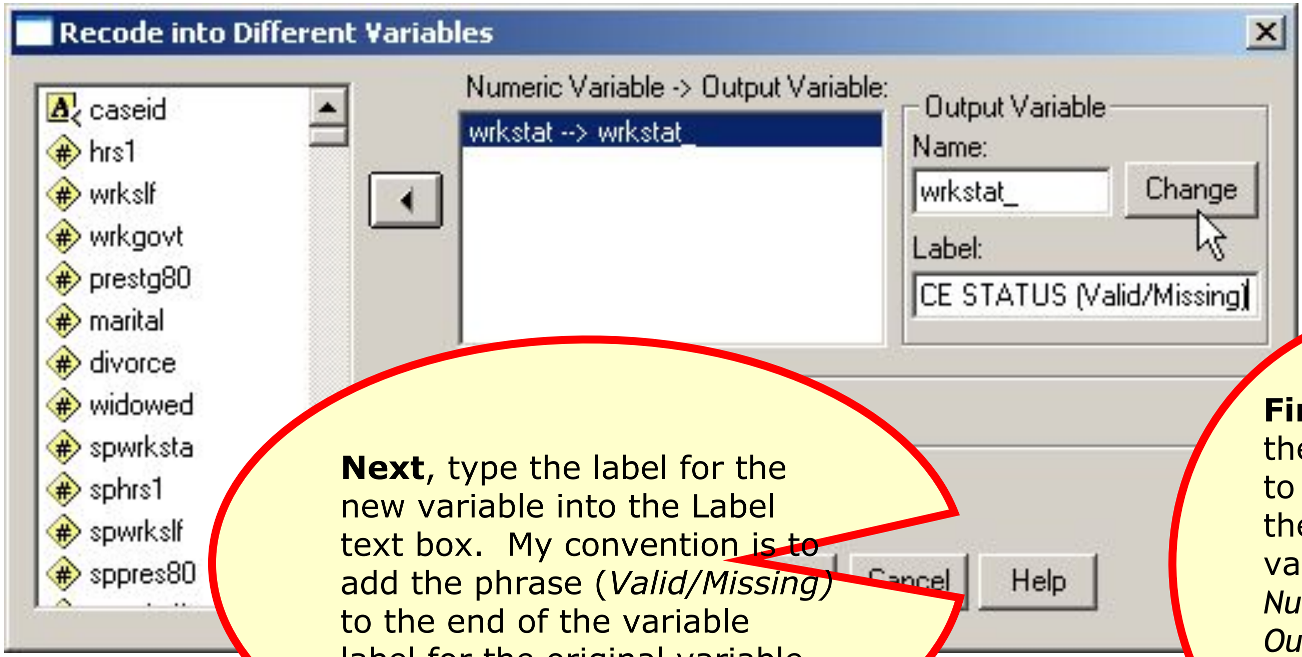


**First**, move the first variable in the analysis, *wrkstat*, into the *Numeric Variable -> Output Variable* text box.

**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



Numeric Variable -> Output Variable:
wrkstat --> wrkstat_

Output Variable

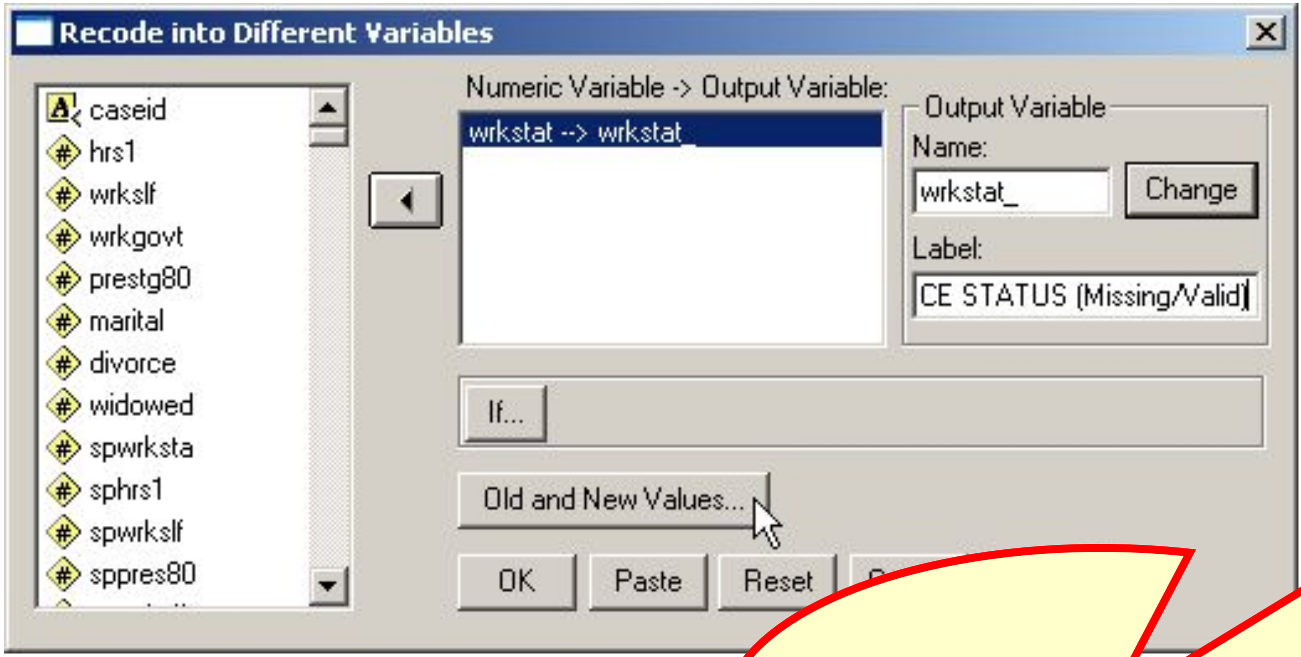
Name: wrkstat_

Label: CE STATUS (Valid/Missing)

**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



Numeric Variable -> Output Variable: wrkstat --> wrkstat_

Name: wrkstat_

Label: CE STATUS (Missing/Valid)

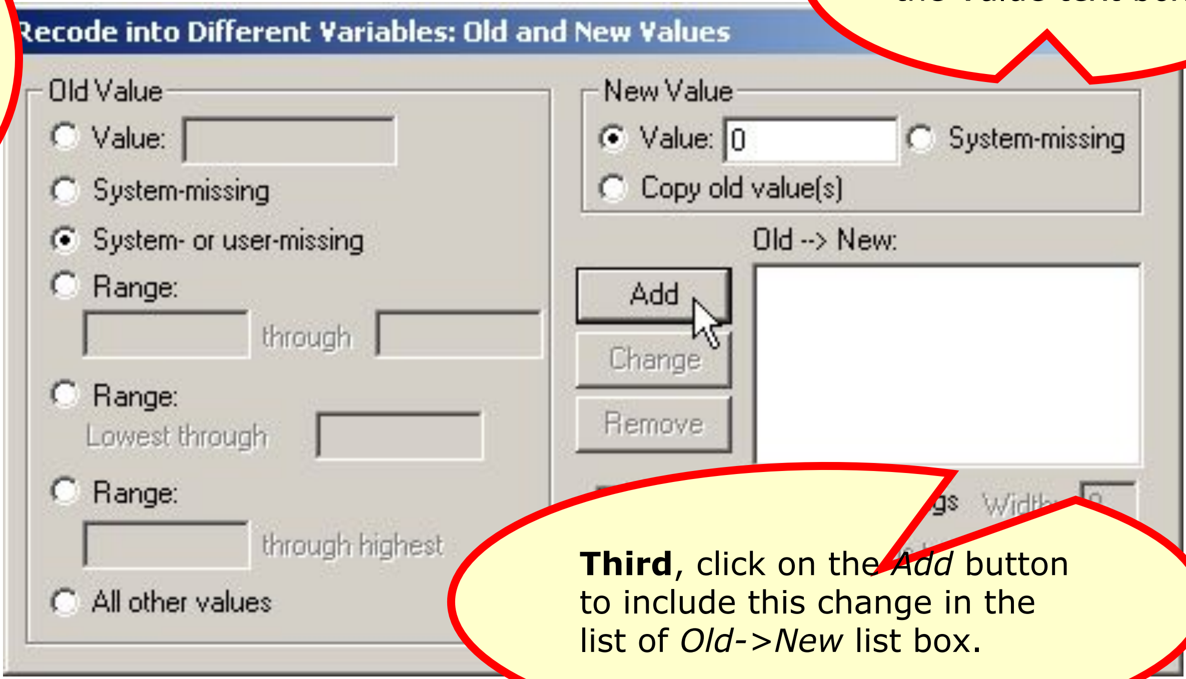
To specify the values for the new variable, click on the *Old and New Values…* button.

# 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.



Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

● System- or user-missing

○ Range: through

○ Range: Lowest through

○ Range: through highest

○ All other values

New Value

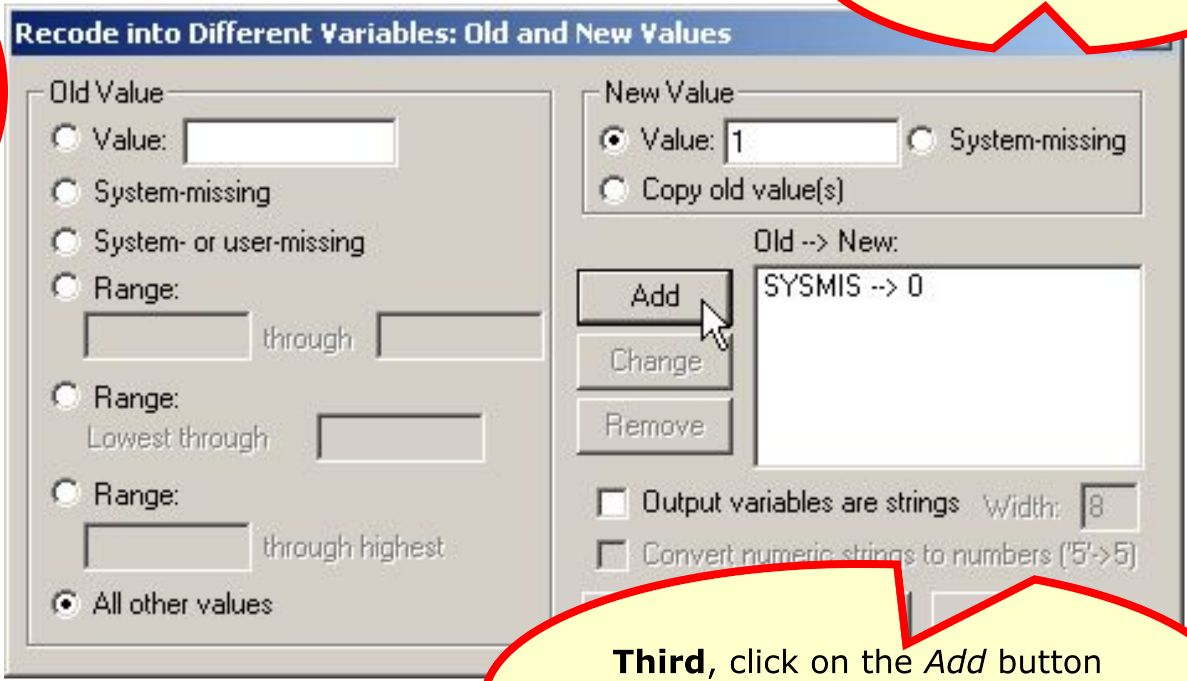
● Value: 0

○ System-missing

○ Copy old value(s)

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

# Change the value for valid data



Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

○ System- or user-missing

○ Range:

through

○ Range:

Lowest through

○ Range:

through highest

● All other values

New Value

● Value: 1 ○ System-missing

○ Copy old value(s)

Old --> New:

SYSMIS --> 0

□ Output variables are strings Width: 8

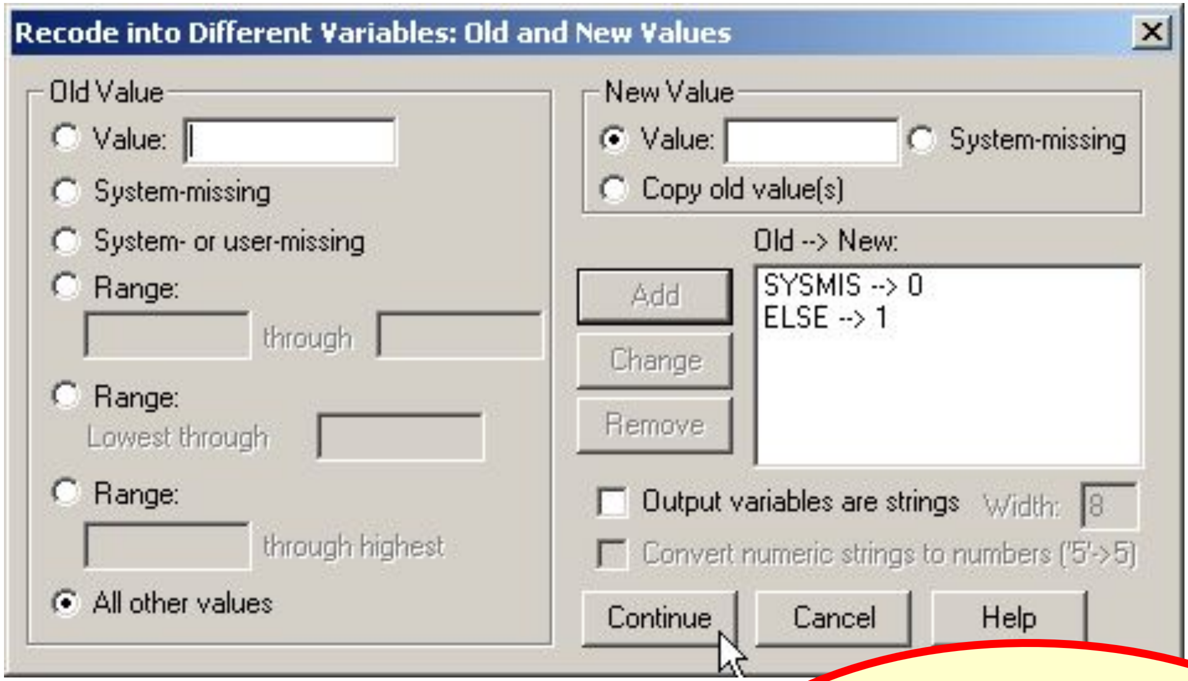
□ Convert numeric strings to numbers ('5'->5)

**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



Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

○ System- or user-missing

○ Range:

through

○ Range:

Lowest through

○ Range:

through highest

● All other values

New Value

● Value:

○ System-missing

○ Copy old value(s)

Old --> New:

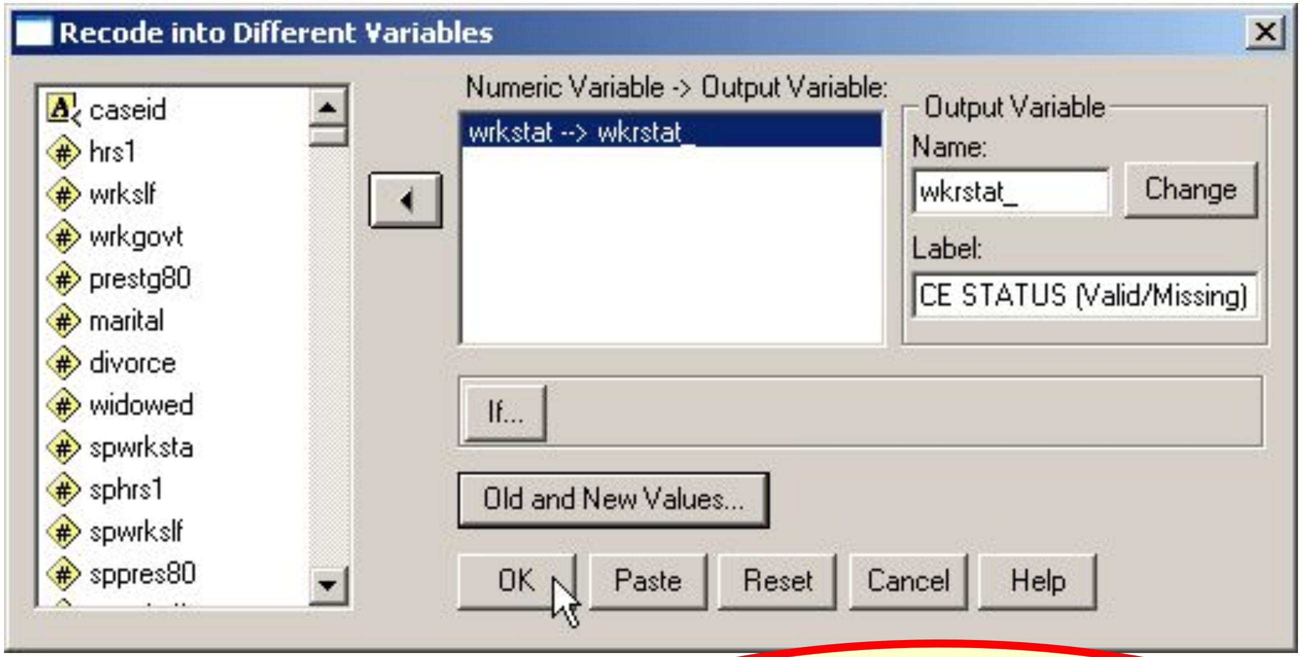
SYSMIS --> 0
ELSE --> 1

□ Output variables are strings Width: 8

□ Convert numeric strings to numbers ('5'->5)

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



Recode into Different Variables

Numeric Variable -> Output Variable: wrkstat --> wkrstat_

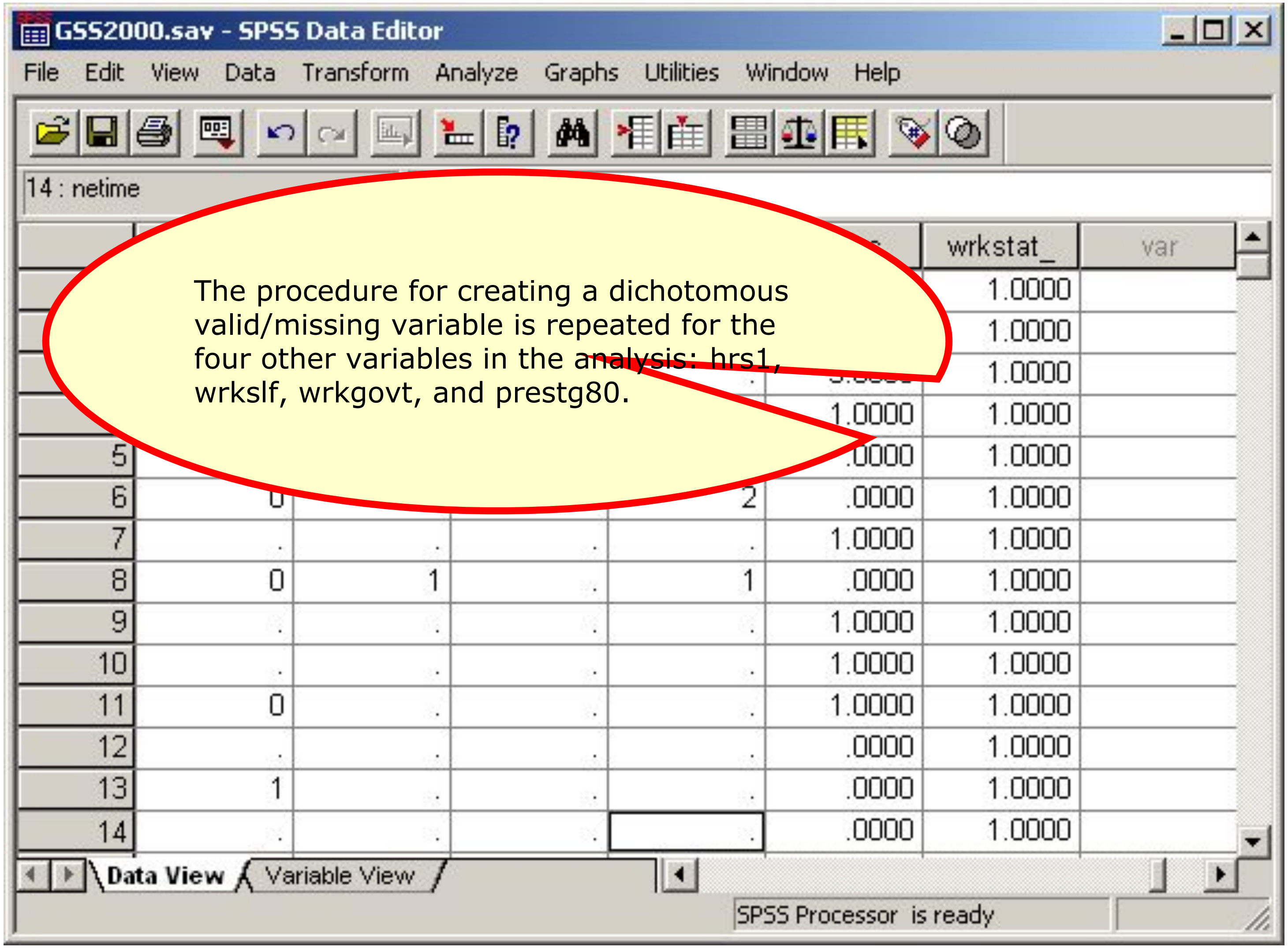
Output Variable

Name: wkrstat_

Label: CE STATUS (Valid/Missing)

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


GSS2000.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

14 : netime

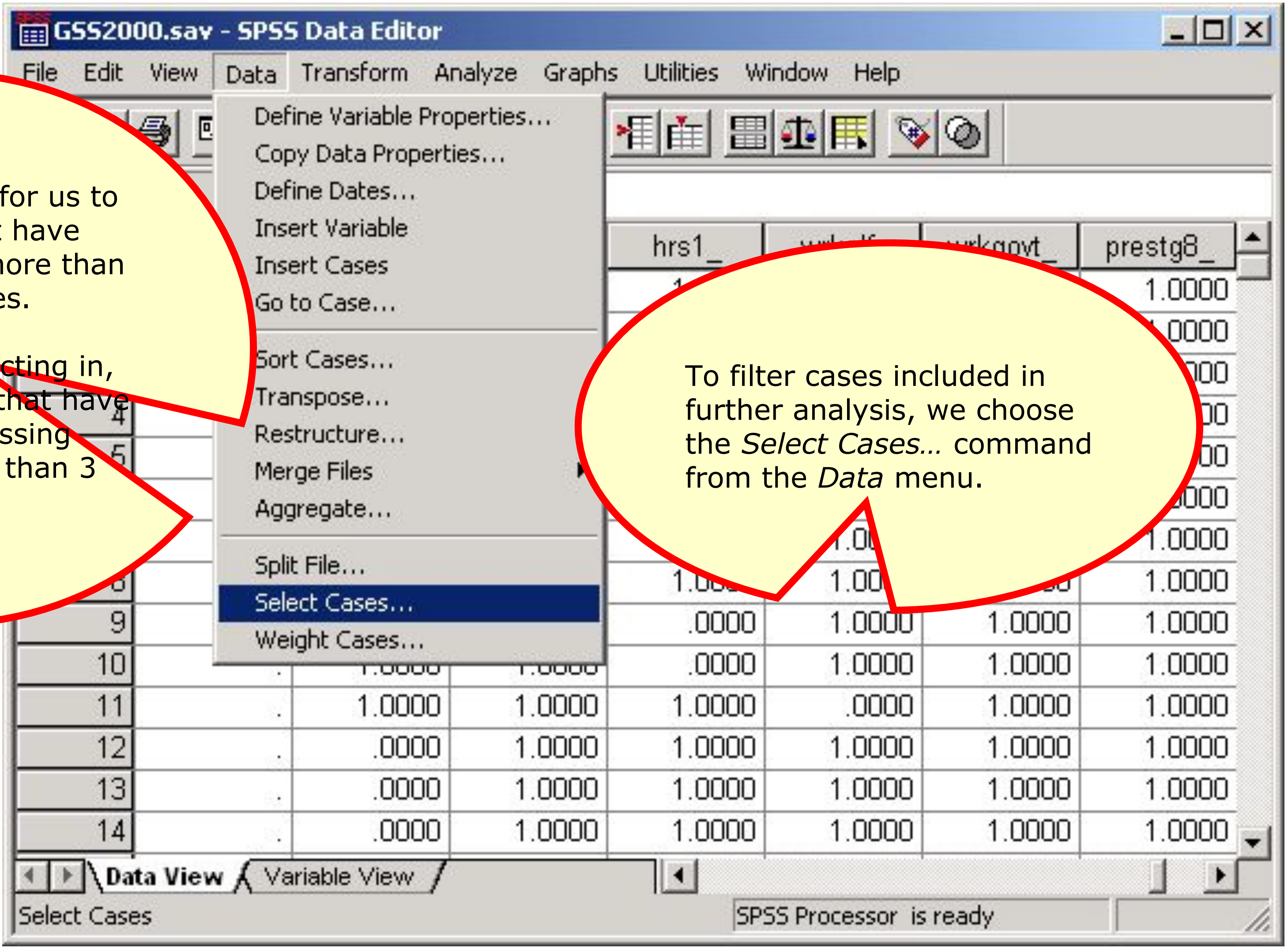
|  |  |  |  |  |  | wrkstat_ | var |
|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  | 1.0000 |  |
|  |  |  |  |  |  | 1.0000 |  |
|  |  |  |  | . |  | 1.0000 |  |
|  |  |  |  |  | 1.0000 | 1.0000 |  |
| 5 |  |  |  |  | .0000 | 1.0000 |  |
| 6 | 0 |  |  | 2 | .0000 | 1.0000 |  |
| 7 | . | . | . | . | 1.0000 | 1.0000 |  |
| 8 | 0 | 1 | . | 1 | .0000 | 1.0000 |  |
| 9 | . | . | . | . | 1.0000 | 1.0000 |  |
| 10 | . | . | . | . | 1.0000 | 1.0000 |  |
| 11 | 0 | . | . | . | 1.0000 | 1.0000 |  |
| 12 | . | . | . | . | .0000 | 1.0000 |  |
| 13 | 1 | . | . | . | .0000 | 1.0000 |  |
| 14 | . | . | . | . | .0000 | 1.0000 |  |

The procedure for creating a dichotomous valid/missing variable is repeated for the four other variables in the analysis: hrs1, wrkslf, wrkgovt, and prestg80.

# 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.



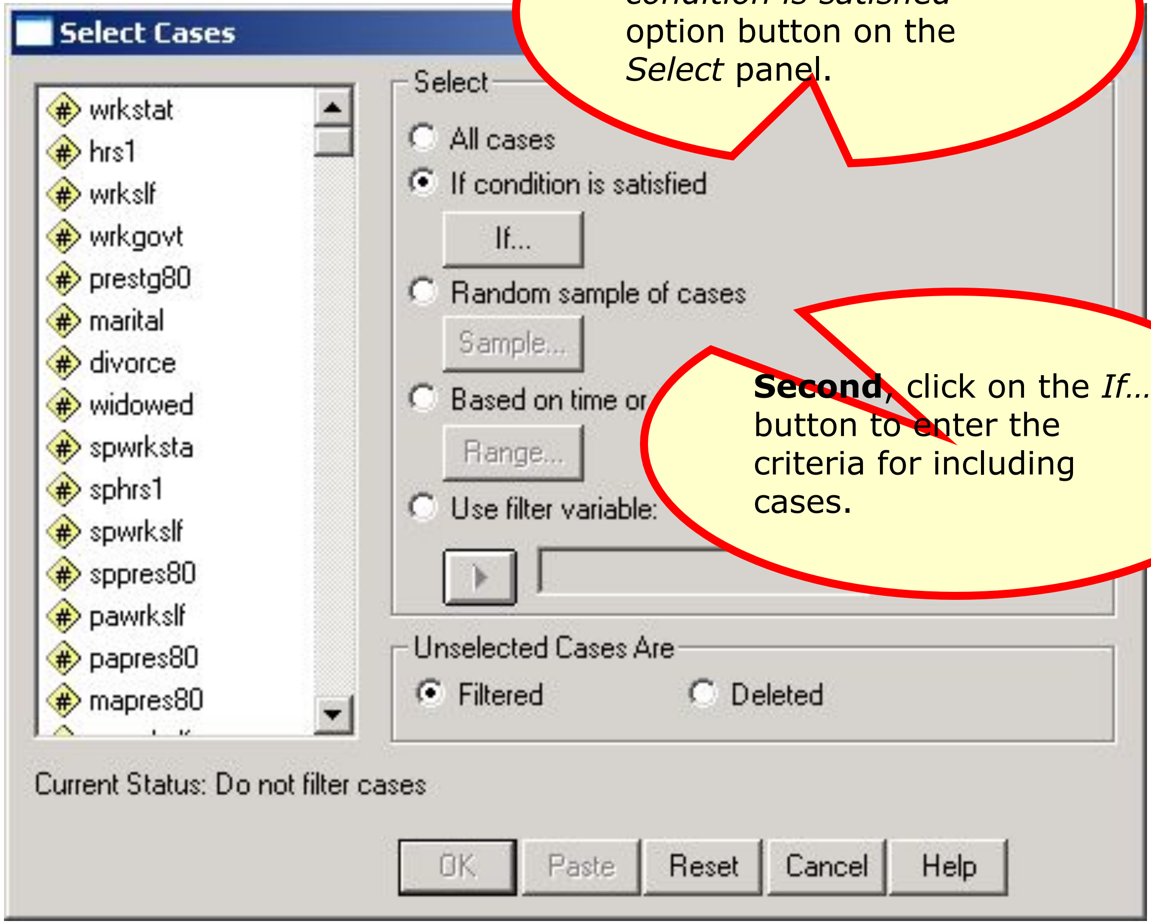
| | | | | hrs1_ | wrkslf_ | wrkgovt_ | prestg8_ |
|---|---|---|---|---|---|---|---|
| 9 | | | | .0000 | 1.0000 | 1.0000 | 1.0000 |
| 10 | . | 1.0000 | 1.0000 | .0000 | 1.0000 | 1.0000 | 1.0000 |
| 11 | . | 1.0000 | 1.0000 | 1.0000 | .0000 | 1.0000 | 1.0000 |
| 12 | . | .0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 13 | . | .0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 14 | . | .0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

Select Cases

SPSS Processor is ready

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

# Enter specifications for selecting cases



Select Cases

wrkstat
hrs1
wrkslf
wrkgovt
prestg80
marital
divorce
widowed
spwrksta
sphrs1
spwrkslf
sppres80
pawrkslf
papres80
mapres80

Select

○ All cases

● If condition is satisfied

If...

○ Random sample of cases

Sample...

○ Based on time or

Range...

○ Use filter variable:

Unselected Cases Are

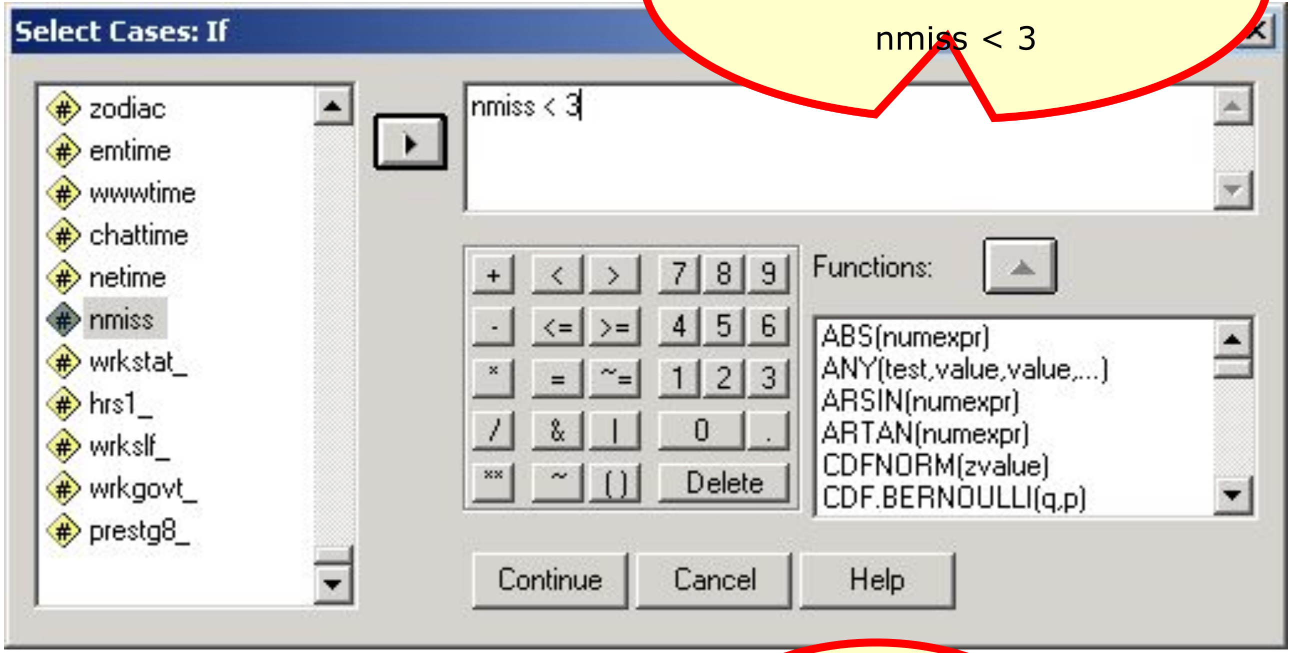
● Filtered ○ Deleted

Current Status: Do not filter cases

**First**, click on the *If condition is satisfied* option button on the *Select* panel.

**Second**, click on the *If…* button to enter the criteria for including cases.

# Enter specifications for selecting cases



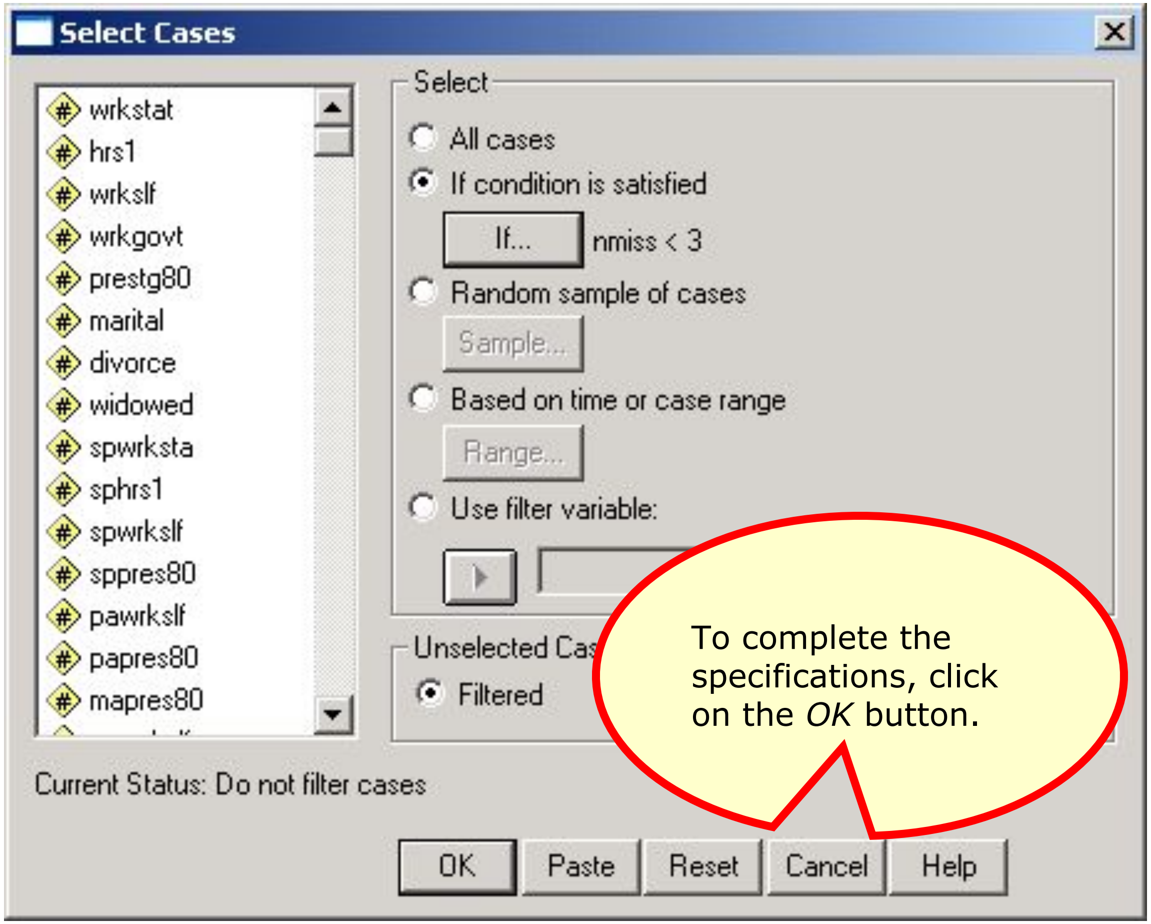
nmiss < 3

**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



Select

○ All cases

● If condition is satisfied

If... nmiss < 3

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

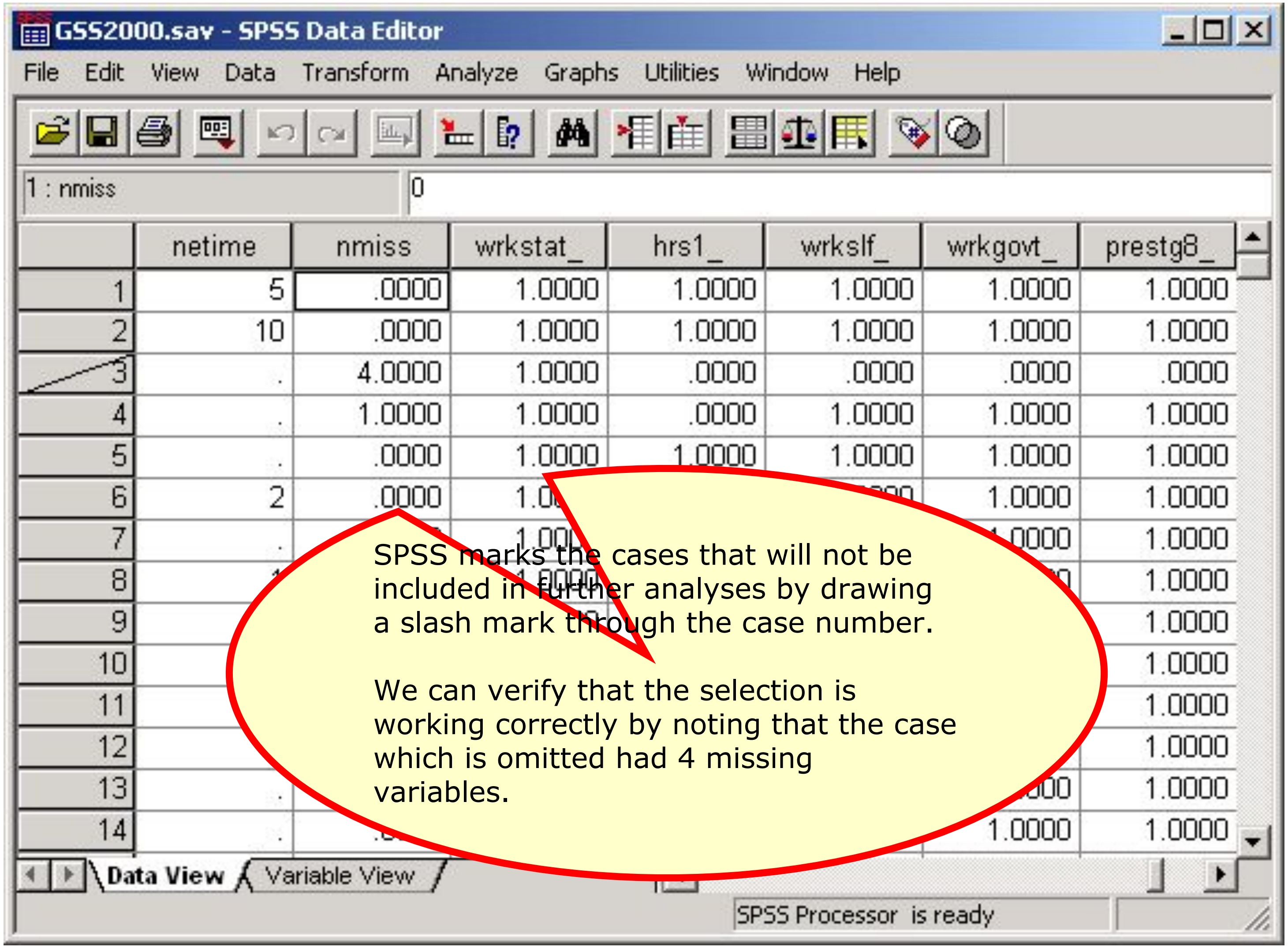
Unselected Cas

● Filtered

Current Status: Do not filter cases

To complete the specifications, click on the *OK* button.

# Cases excluded from further analyses


GSS2000.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

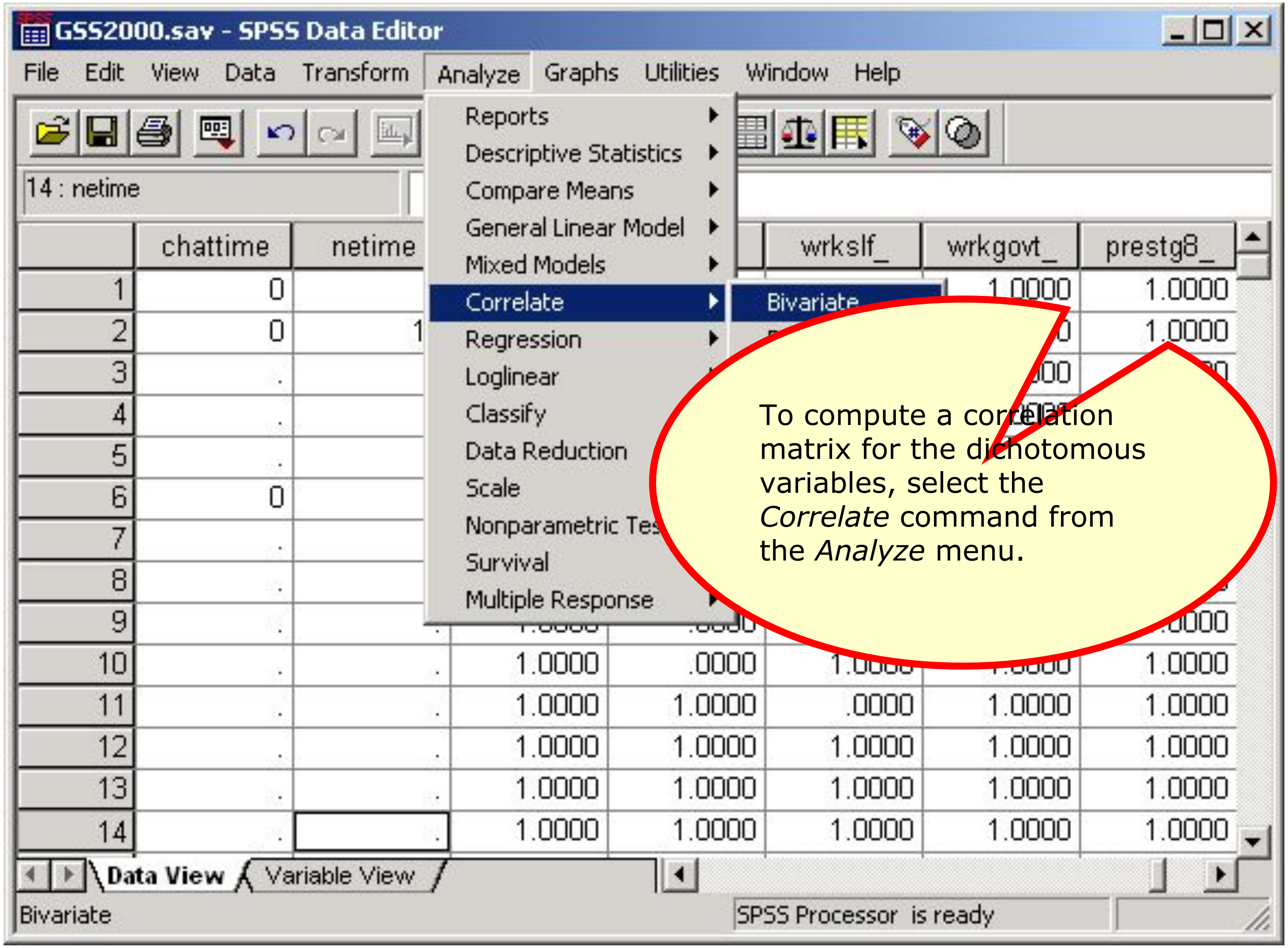
1 : nmiss 0

|  | netime | nmiss | wrkstat_ | hrs1_ | wrkslf_ | wrkgovt_ | prestg8_ |
|---|---|---|---|---|---|---|---|
| 1 | 5 | .0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 2 | 10 | .0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 3 | . | 4.0000 | 1.0000 | .0000 | .0000 | .0000 | .0000 |
| 4 | . | 1.0000 | 1.0000 | .0000 | 1.0000 | 1.0000 | 1.0000 |
| 5 | . | .0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 6 | 2 | .0000 | 1.0 |  |  | 1.0000 | 1.0000 |
| 7 | . |  | 1.00 |  |  | 1.0000 | 1.0000 |
| 8 | 1 |  | 1.000 |  |  |  | 1.0000 |
| 9 |  |  |  |  |  |  | 1.0000 |
| 10 |  |  |  |  |  |  | 1.0000 |
| 11 |  |  |  |  |  |  | 1.0000 |
| 12 |  |  |  |  |  |  | 1.0000 |
| 13 | . |  |  |  |  | 00 | 1.0000 |
| 14 | . | .0 |  |  |  | 1.0000 | 1.0000 |

SPSS marks the cases that will not be included in further analyses by drawing a slash mark through the case number.

We can verify that the selection is working correctly by noting that the case which is omitted had 4 missing variables.

# Correlating the dichotomous variables



14 : netime

| | chattime | netime | | | wrkslf_ | wrkgovt_ | prestg8_ |
|---|---|---|---|---|---|---|---|
| 1 | 0 | | | | | 1.0000 | 1.0000 |
| 2 | 0 | 1 | | | | | 1.0000 |
| 3 | . | | | | | | |
| 4 | . | | | | | | |
| 5 | . | | | | | | |
| 6 | 0 | | | | | | |
| 7 | . | | | | | | |
| 8 | . | | | | | | |
| 9 | . | | | | | | |
| 10 | . | . | 1.0000 | .0000 | | | 1.0000 |
| 11 | . | . | 1.0000 | 1.0000 | .0000 | 1.0000 | 1.0000 |
| 12 | . | . | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 13 | . | . | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 14 | . | . | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

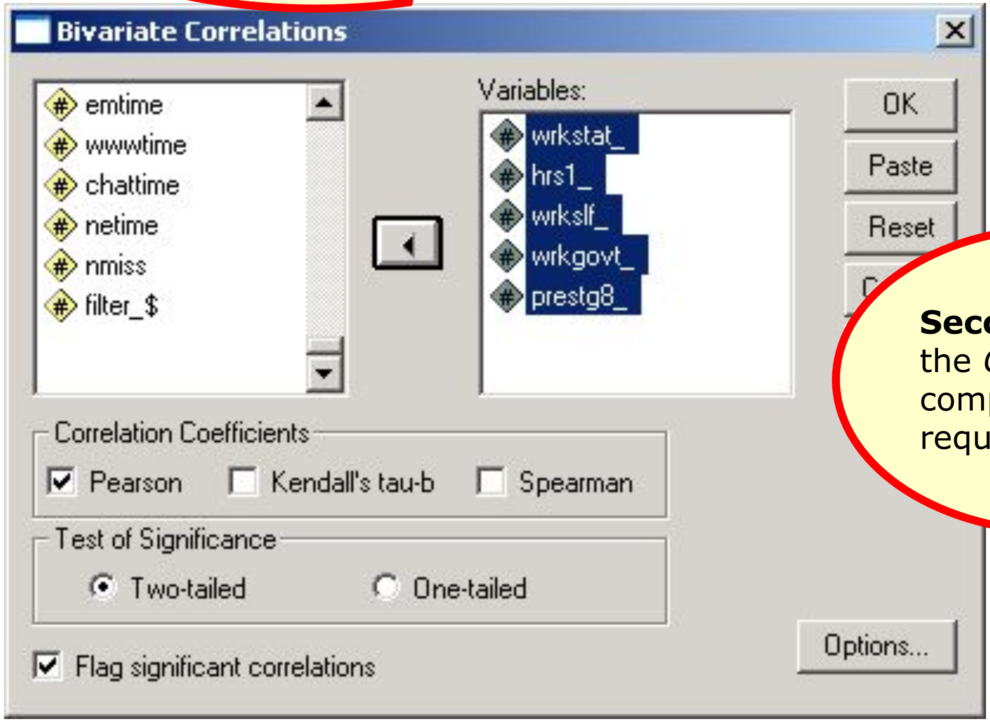
Bivariate

SPSS Processor is ready

To compute a correlation matrix for the dichotomous variables, select the *Correlate* command from the *Analyze* menu.

# Specifications for correlations

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



Bivariate Correlations

Variables: wrkstat_, hrs1_, wrkslf_, wrkgovt_, prestg8_

Correlation Coefficients

☑ Pearson □ Kendall's tau-b □ Spearman

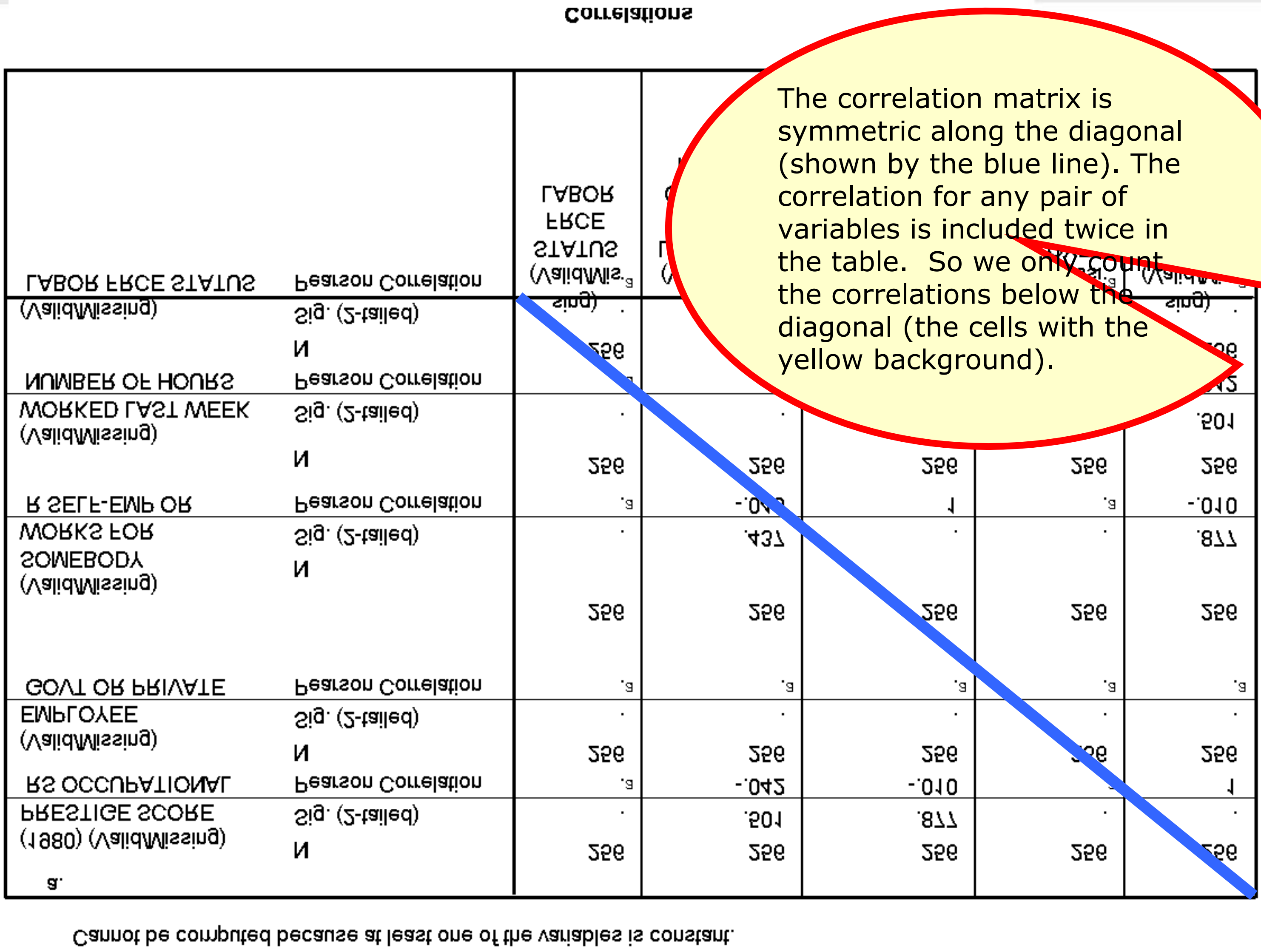
Test of Significance

● Two-tailed ○ One-tailed

☑ Flag significant correlations

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

# The correlation matrix



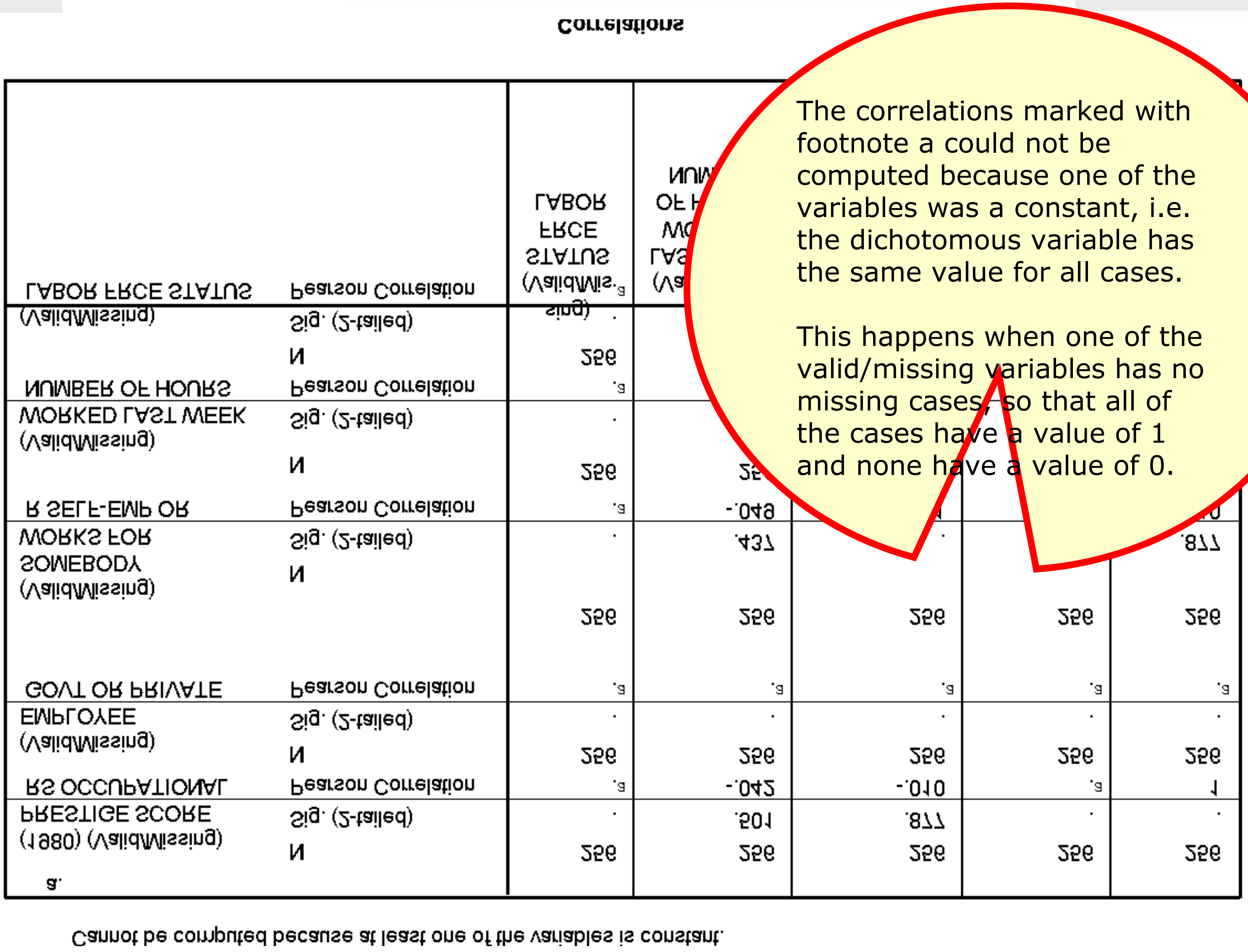
Correlations

| | | LABOR FRCE STATUS (ValidMis-sing) | | | (ValidMis.a | (ValidMis-sing) |
|---|---|---|---|---|---|---|
| LABOR FRCE STATUS (ValidMissing) | Pearson Correlation | .a | | | | .a |
| | Sig. (2-tailed) | . | | | | . |
| | N | 256 | | | | 256 |
| NUMBER OF HOURS WORKED LAST WEEK (ValidMissing) | Pearson Correlation | .a | | | | -.042 |
| | Sig. (2-tailed) | . | . | | | .501 |
| | N | 256 | 256 | 256 | 256 | 256 |
| R SELF-EMP OR WORKS FOR SOMEBODY (ValidMissing) | Pearson Correlation | .a | -.039 | 1 | .a | -.010 |
| | Sig. (2-tailed) | . | .437 | . | . | .877 |
| | N | 256 | 256 | 256 | 256 | 256 |
| GOVT OR PRIVATE EMPLOYEE (ValidMissing) | Pearson Correlation | .a | .a | .a | .a | .a |
| | Sig. (2-tailed) | . | . | . | . | . |
| | N | 256 | 256 | 256 | 256 | 256 |
| RS OCCUPATIONAL PRESTIGE SCORE (1980) (ValidMissing) | Pearson Correlation | .a | -.042 | -.010 | .a | 1 |
| | Sig. (2-tailed) | . | .501 | .877 | . | . |
| | N | 256 | 256 | 256 | 256 | 256 |

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

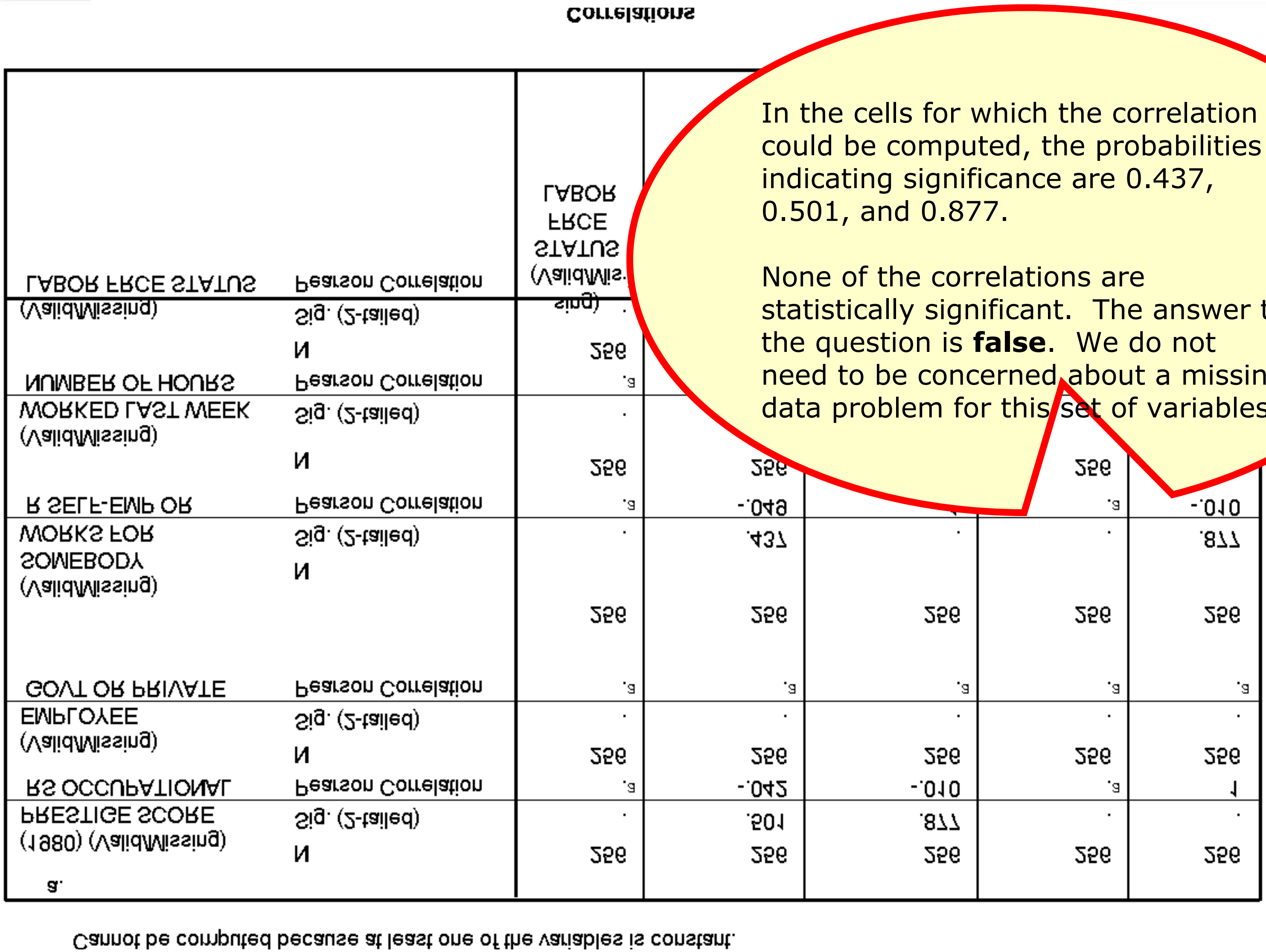
| | | LABOR FRCE STATUS (ValidMissing)[a] | NUM... OF H... WO... LAS... (Va... | [illegible] | [illegible] | [illegible] |
|---|---|---|---|---|---|---|
| LABOR FRCE STATUS (ValidMissing) | Pearson Correlation | .a | [illegible] | [illegible] | [illegible] | [illegible] |
| | Sig. (2-tailed) | . | [illegible] | [illegible] | [illegible] | [illegible] |
| | N | 256 | [illegible] | [illegible] | [illegible] | [illegible] |
| NUMBER OF HOURS WORKED LAST WEEK (ValidMissing) | Pearson Correlation | .a | [illegible] | [illegible] | [illegible] | [illegible] |
| | Sig. (2-tailed) | . | [illegible] | [illegible] | [illegible] | [illegible] |
| | N | 256 | 25[illegible] | [illegible] | [illegible] | [illegible] |
| R SELF-EMP OR WORKS FOR SOMEBODY (ValidMissing) | Pearson Correlation | .a | -.049 | [illegible] | [illegible] | [illegible] |
| | Sig. (2-tailed) | . | .437 | . | . | .877 |
| | N | 256 | 256 | 256 | 256 | 256 |
| GOVT OR PRIVATE EMPLOYEE (ValidMissing) | Pearson Correlation | .a | .a | .a | .a | .a |
| | Sig. (2-tailed) | . | . | . | . | . |
| | N | 256 | 256 | 256 | 256 | 256 |
| RS OCCUPATIONAL PRESTIGE SCORE (1980) (ValidMissing) | Pearson Correlation | .a | -.042 | -.010 | .a | 1 |
| | Sig. (2-tailed) | . | .201 | .877 | . | . |
| | N | 256 | 256 | 256 | 256 | 256 |

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



Correlations

| | | LABOR FRCE STATUS (ValidMissing) | | | | |
|---|---|---|---|---|---|---|
| LABOR FRCE STATUS (ValidMissing) | Pearson Correlation | | | | | |
| | Sig. (2-tailed) | . | | | | |
| | N | 256 | | | | |
| NUMBER OF HOURS WORKED LAST WEEK (ValidMissing) | Pearson Correlation | .a | | | | |
| | Sig. (2-tailed) | . | | | | |
| | N | 256 | 256 | | 256 | |
| R SELF-EMP OR WORKS FOR SOMEBODY (ValidMissing) | Pearson Correlation | .a | -.049 | | .a | -.010 |
| | Sig. (2-tailed) | . | .437 | . | . | .877 |
| | N | 256 | 256 | 256 | 256 | 256 |
| GOVT OR PRIVATE EMPLOYEE (ValidMissing) | Pearson Correlation | .a | .a | .a | .a | .a |
| | Sig. (2-tailed) | . | . | . | . | . |
| | N | 256 | 256 | 256 | 256 | 256 |
| RS OCCUPATIONAL PRESTIGE SCORE (1980) (ValidMissing) | Pearson Correlation | .a | -.042 | -.010 | .a | 1 |
| | Sig. (2-tailed) | . | .501 | .877 | . | . |
| | N | 256 | 256 | 256 | 256 | 256 |

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

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.

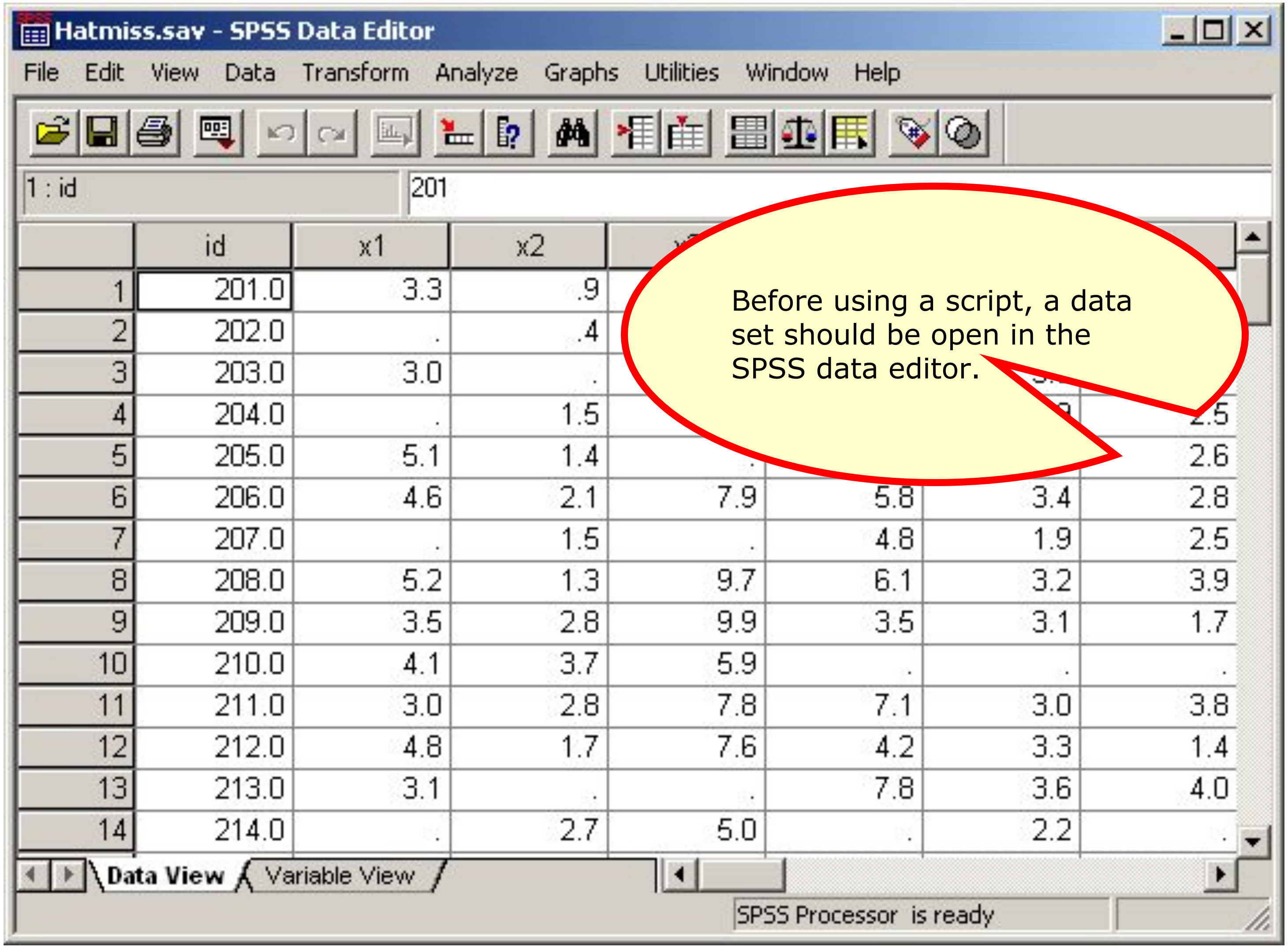
# Using a script for missing data

- 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



Hatmiss.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View

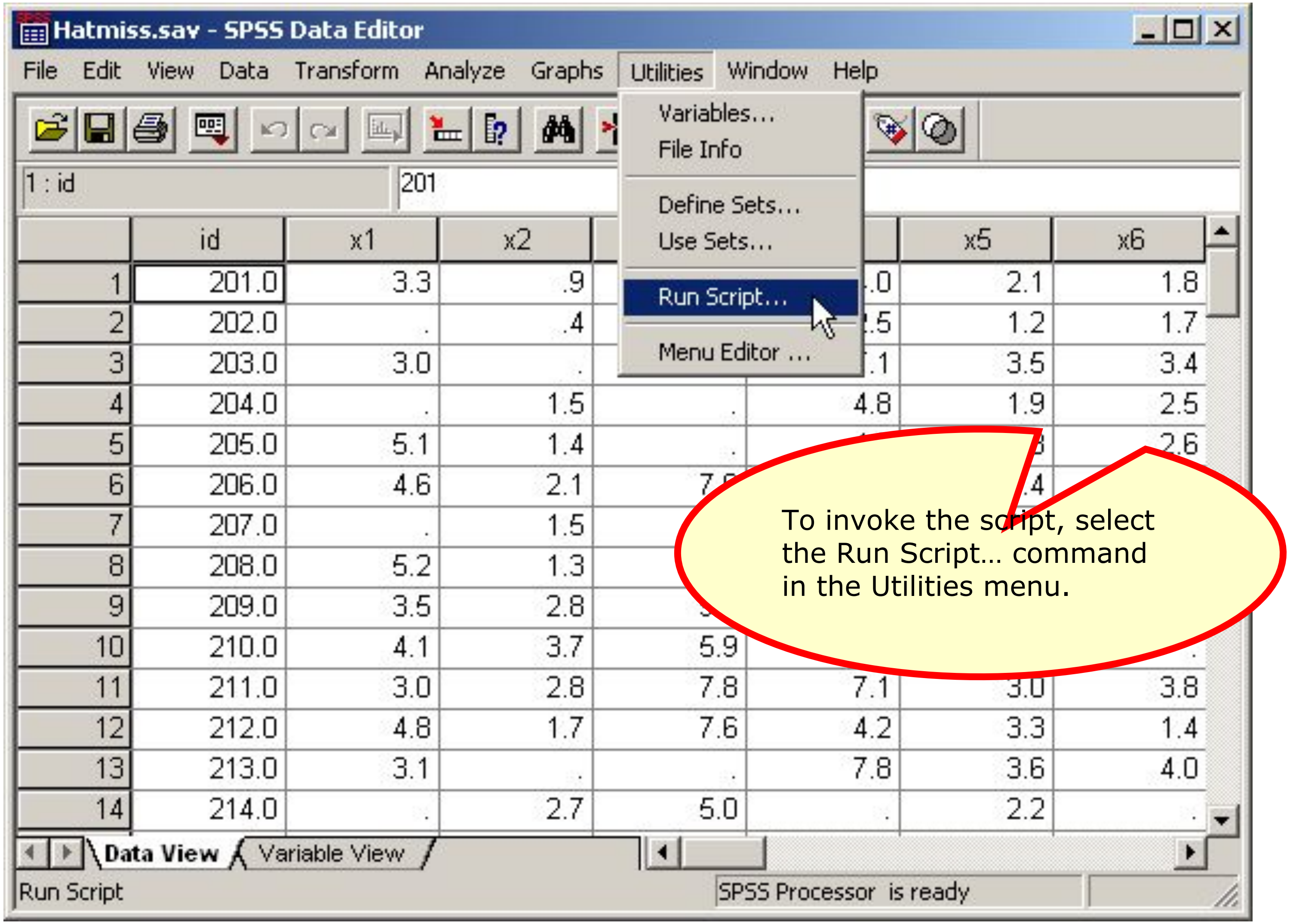
1 : id 201

| | id | x1 | x2 | [illegible] | [illegible] | [illegible] | [illegible] |
|---|---|---|---|---|---|---|---|
| 1 | 201.0 | 3.3 | .9 | [illegible] | [illegible] | [illegible] | [illegible] |
| 2 | 202.0 | . | .4 | [illegible] | [illegible] | [illegible] | [illegible] |
| 3 | 203.0 | 3.0 | . | [illegible] | [illegible] | [illegible] | [illegible] |
| 4 | 204.0 | . | 1.5 | [illegible] | [illegible] | [illegible] | 2.5 |
| 5 | 205.0 | 5.1 | 1.4 | . | [illegible] | [illegible] | 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 | . |

SPSS Processor is ready

Before using a script, a data set should be open in the SPSS data editor.

# Invoke the script

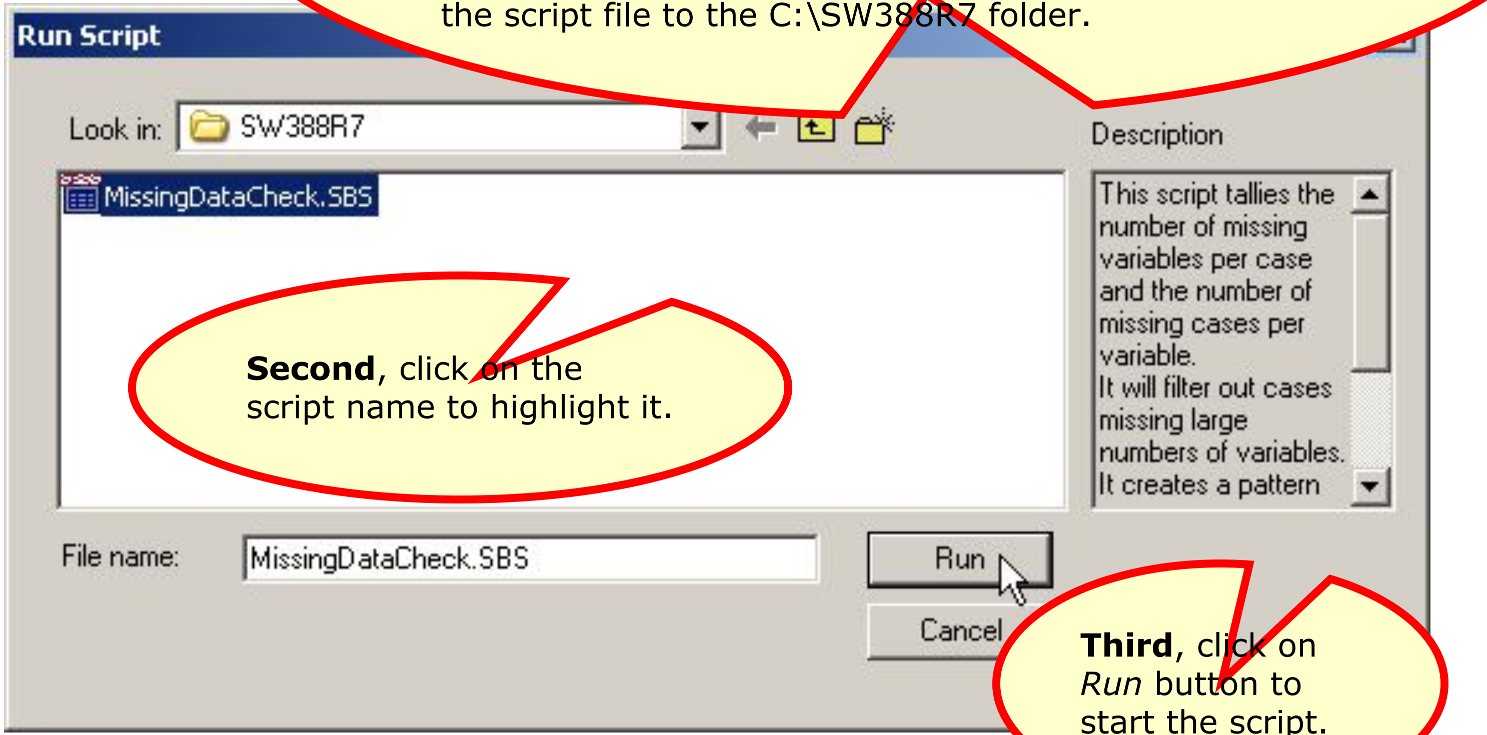


1 : id 201

|  | id | x1 | x2 |  |  | x5 | x6 |
|---|---|---|---|---|---|---|---|
| 1 | 201.0 | 3.3 | .9 |  | .0 | 2.1 | 1.8 |
| 2 | 202.0 | . | .4 |  | .5 | 1.2 | 1.7 |
| 3 | 203.0 | 3.0 | . |  | .1 | 3.5 | 3.4 |
| 4 | 204.0 | . | 1.5 | . | 4.8 | 1.9 | 2.5 |
| 5 | 205.0 | 5.1 | 1.4 | . |  |  | 2.6 |
| 6 | 206.0 | 4.6 | 2.1 |  |  | .4 |  |
| 7 | 207.0 | . | 1.5 |  |  |  |  |
| 8 | 208.0 | 5.2 | 1.3 |  |  |  |  |
| 9 | 209.0 | 3.5 | 2.8 |  |  |  |  |
| 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 | . |

To invoke the script, select the Run Script… command in the Utilities menu.

# Select the missing data script



Look in: SW388R7

MissingDataCheck.SBS

Description

This script tallies the number of missing variables per case and the number of missing cases per variable.
It will filter out cases missing large numbers of variables.
It creates a pattern

File name: MissingDataCheck.SBS

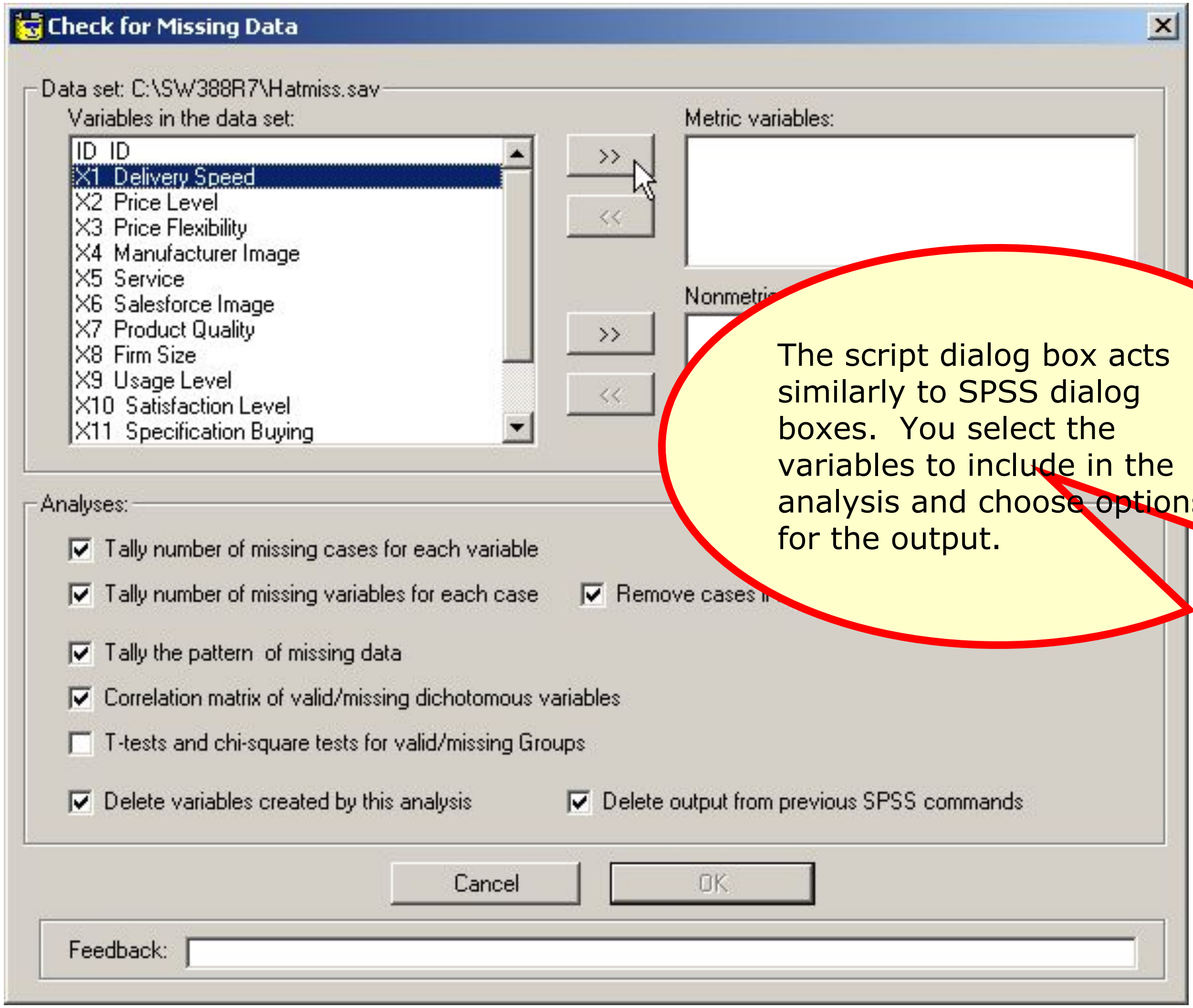
**First**, navigate to the folder where you put the script. If you followed the directions, you will have a file with an ".SBS" extension in the C:\SW388R7 folder.

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



Check for Missing Data

Data set: C:\SW388R7\Hatmiss.sav

Variables in the data set:

ID ID
X1 Delivery Speed
X2 Price Level
X3 Price Flexibility
X4 Manufacturer Image
X5 Service
X6 Salesforce Image
X7 Product Quality
X8 Firm Size
X9 Usage Level
X10 Satisfaction Level
X11 Specification Buying

Metric variables:

Nonmetri

Analyses:

☑ Tally number of missing cases for each variable

☑ Tally number of missing variables for each case ☑ Remove cases i

☑ Tally the pattern of missing data

☑ Correlation matrix of valid/missing dichotomous variables

□ T-tests and chi-square tests for valid/missing Groups

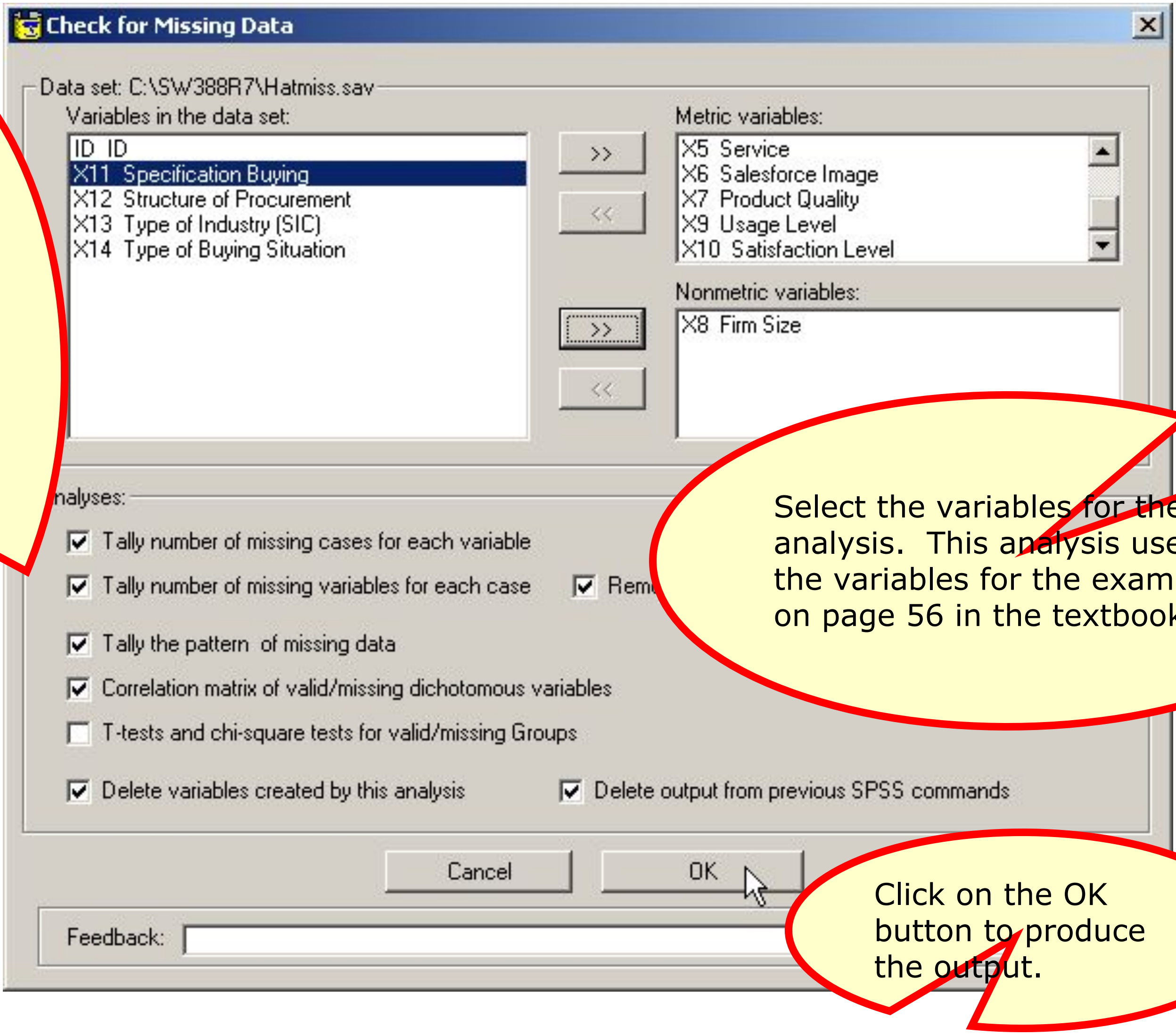
☑ Delete variables created by this analysis ☑ Delete output from previous SPSS commands

Feedback:

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.



Check for Missing Data

Data set: C:\SW388R7\Hatmiss.sav

Variables in the data set:

ID ID
X11 Specification Buying
X12 Structure of Procurement
X13 Type of Industry (SIC)
X14 Type of Buying Situation

Metric variables:

X5 Service
X6 Salesforce Image
X7 Product Quality
X9 Usage Level
X10 Satisfaction Level

Nonmetric variables:

X8 Firm Size

nalyses:

☑ Tally number of missing cases for each variable

☑ Tally number of missing variables for each case ☑ Rem

☑ Tally the pattern of missing data

☑ Correlation matrix of valid/missing dichotomous variables

□ T-tests and chi-square tests for valid/missing Groups

☑ Delete variables created by this analysis ☑ Delete output from previous SPSS commands

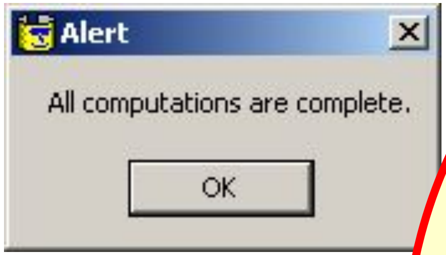
Feedback:

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.



Alert

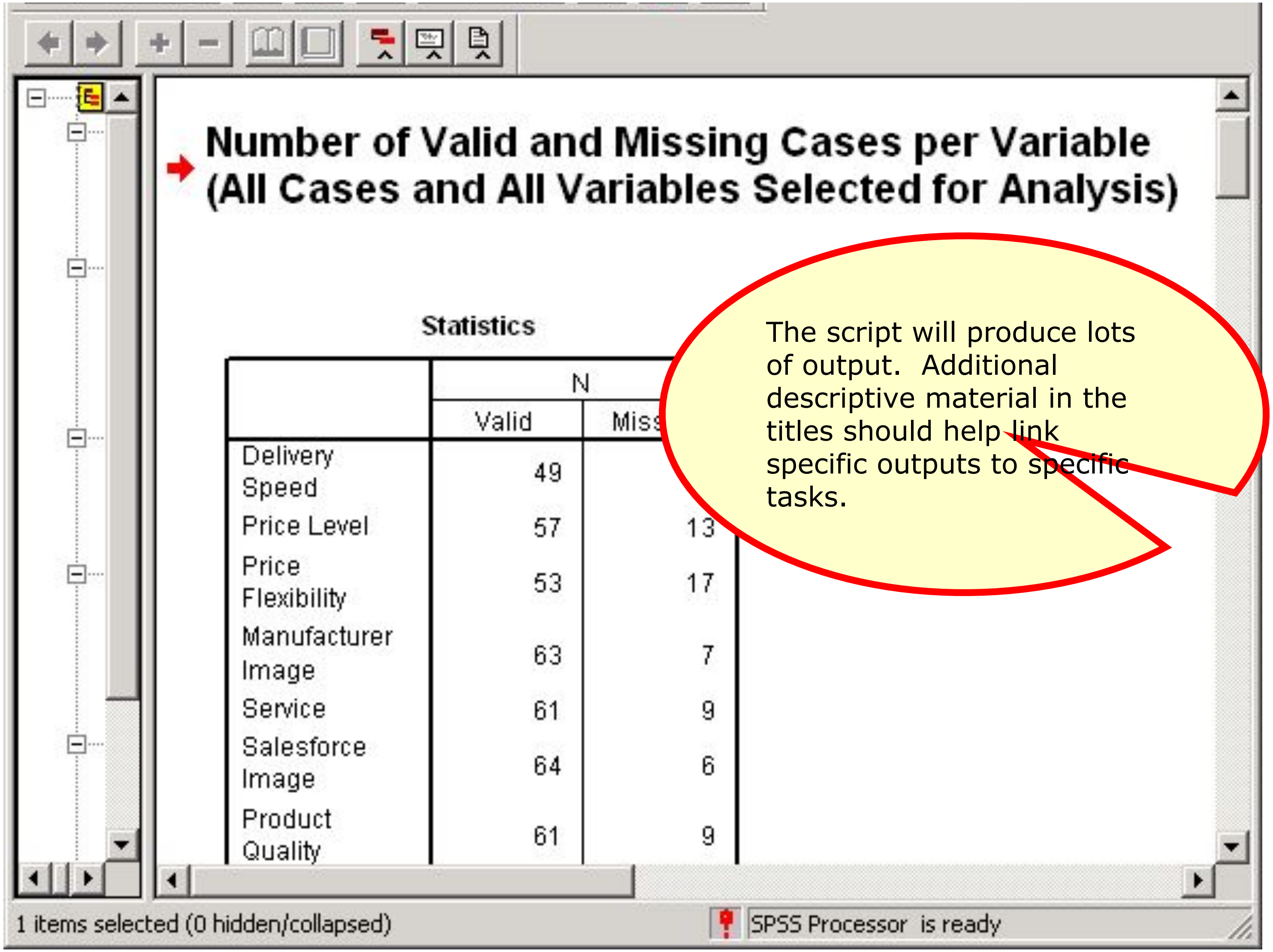
All computations are complete.

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



**Number of Valid and Missing Cases per Variable (All Cases and All Variables Selected for Analysis)**

**Statistics**

| | N | |
|---|---|---|
| | Valid | Miss |
| Delivery Speed | 49 | |
| Price Level | 57 | 13 |
| Price Flexibility | 53 | 17 |
| Manufacturer Image | 63 | 7 |
| Service | 61 | 9 |
| Salesforce Image | 64 | 6 |
| Product Quality | 61 | 9 |

1 items selected (0 hidden/collapsed)

SPSS Processor is ready

The script will produce lots of output. Additional descriptive material in the titles should help link specific outputs to specific tasks.