

Analyzing Missing Data

Introduction

Problems

Using Scripts

$$\begin{aligned} H_1: \mu < 0 \\ H_0: \mu = 0 \\ \sigma^2 &= E(x - \mu)^2 \\ \mu &= \frac{1}{2}(x_j + x_{j+1}) \\ t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \end{aligned}$$

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.

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i = \bar{x}$
 $H_0: \mu = 0$
 $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
 $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$
 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
 $\sigma^2 = E(x - \mu)^2$
 $\bar{y} = \frac{1}{2}(x_j + x_{j+1})$

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.

3

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i = \bar{x}$
 $\sum_{i=1}^n w_i x_i^2 = \bar{x}^2 + s^2$
 $H_0: \mu = 0$
 $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
 $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
 $s = \sqrt{s^2}$
 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
 $\sigma^2 = E(x - \mu)^2$
 $\bar{y} = \frac{1}{2}(x_j + x_{j+1})$

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.

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i = \bar{x}$
 $H_0: \mu = 0$
 $\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
 $\sigma^2 = E(x - \mu)^2$
 $\bar{y} = \frac{1}{2}(x_j + x_{j+1})$

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

1 : caseid 20000009

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

Data View Variable View

SPSS Process

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 that 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.

6

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i (b-1) = b^s$
 $v = x_j$
 $H_0: \mu = 0$
 $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
 $\sigma^2 = E(x - \mu)^2$
 $y = \frac{1}{2}(x_j + x_{j+1})$

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.

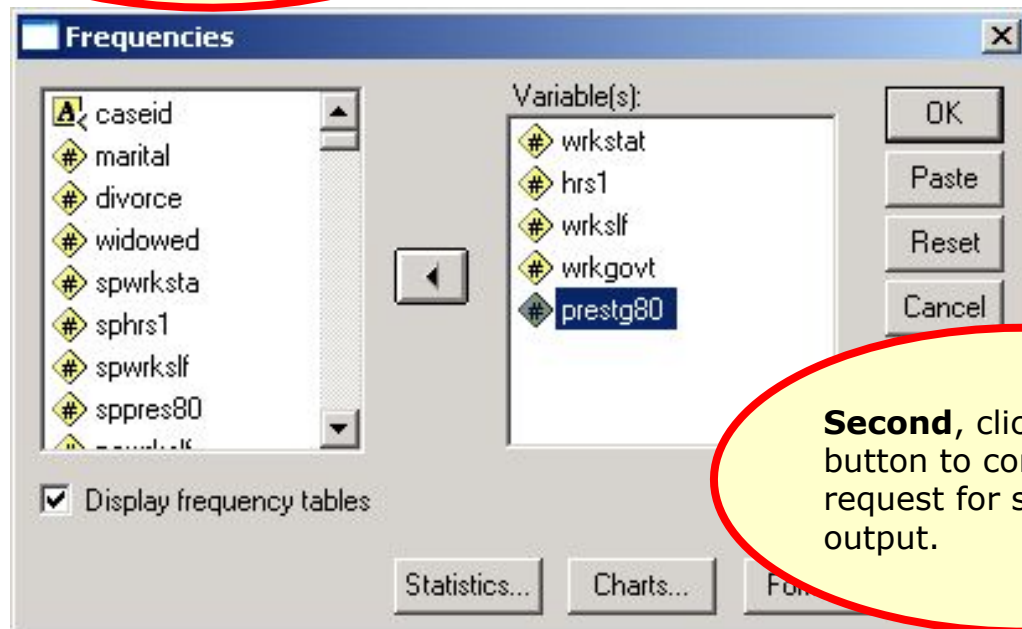
The screenshot shows the SPSS Data Editor window for 'GSS2000.sav'. The 'Analyze' menu is open, and the 'Descriptive Statistics' sub-menu is selected, with 'Frequencies...' highlighted. The background shows a data grid with columns 'estg80' and 'marital'. The status bar at the bottom indicates 'SPSS Processor is ready'.

Case #	estg80	marital					
263	20002771						
264	20002772						
265	20002791						
266	20002794						
267	20002795						
268	20002799						
269	20002802						
270	20002804	1	40	2	2	35	1
271							
272							
273							
274							

Completing the specification for frequencies

8

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

The screenshot shows the SPSS Output2 - SPSS Viewer window. The main content is a table titled "Statistics" under the heading "Frequencies". The table lists the number of valid and missing cases for several variables. A red oval highlights a text box that says: "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." Another red oval highlights a text box at the bottom that says: "None of the variables has more than 135 missing cases, although number of hours worked in the past week comes close."

		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

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

10

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

11

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.

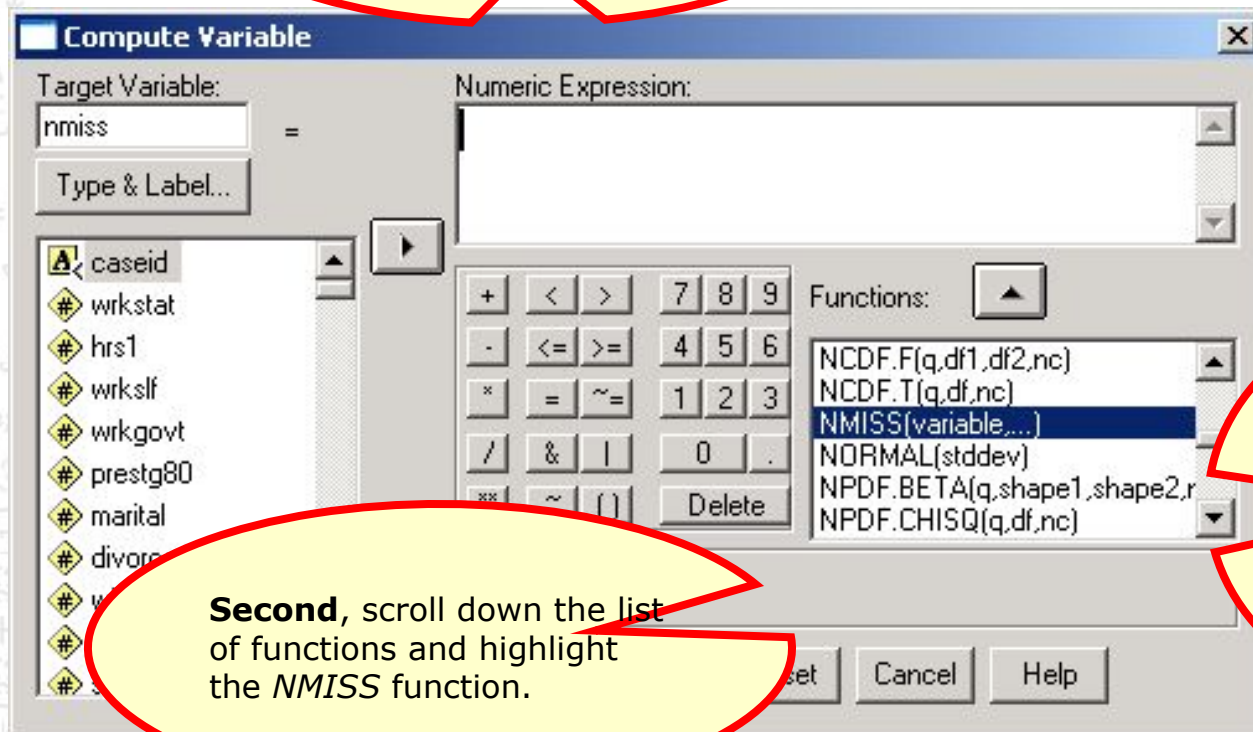
The screenshot shows the SPSS Data Editor window for 'GSS2000.sav'. The 'Transform' menu is open, and the 'Compute...' option is selected. The background shows a data table with columns 'wrk', 'marital', and others. The status bar at the bottom indicates 'SPSS Processor is ready'.

Case	wrk	marital
1	1	1
2	1	1
3	1	1
4	3	3
5	1	1
6	5	5
7	3	3
8	5	5
9	2	2
10	2	2
11	2	2
12	2	2
13	2	2
14	2	2

Enter specifications for new variable

12

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



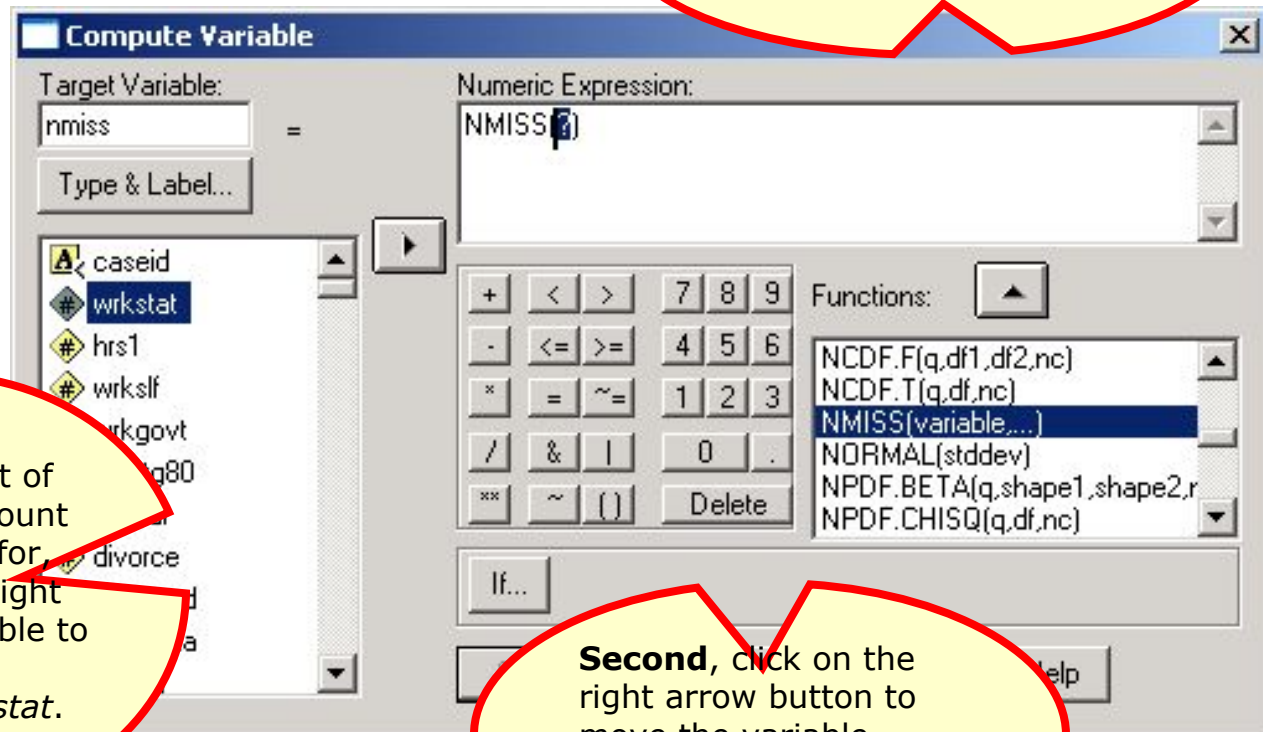
Second, scroll down the list of functions and highlight the *NMISS* function.

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

Enter specifications for new variable

13

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

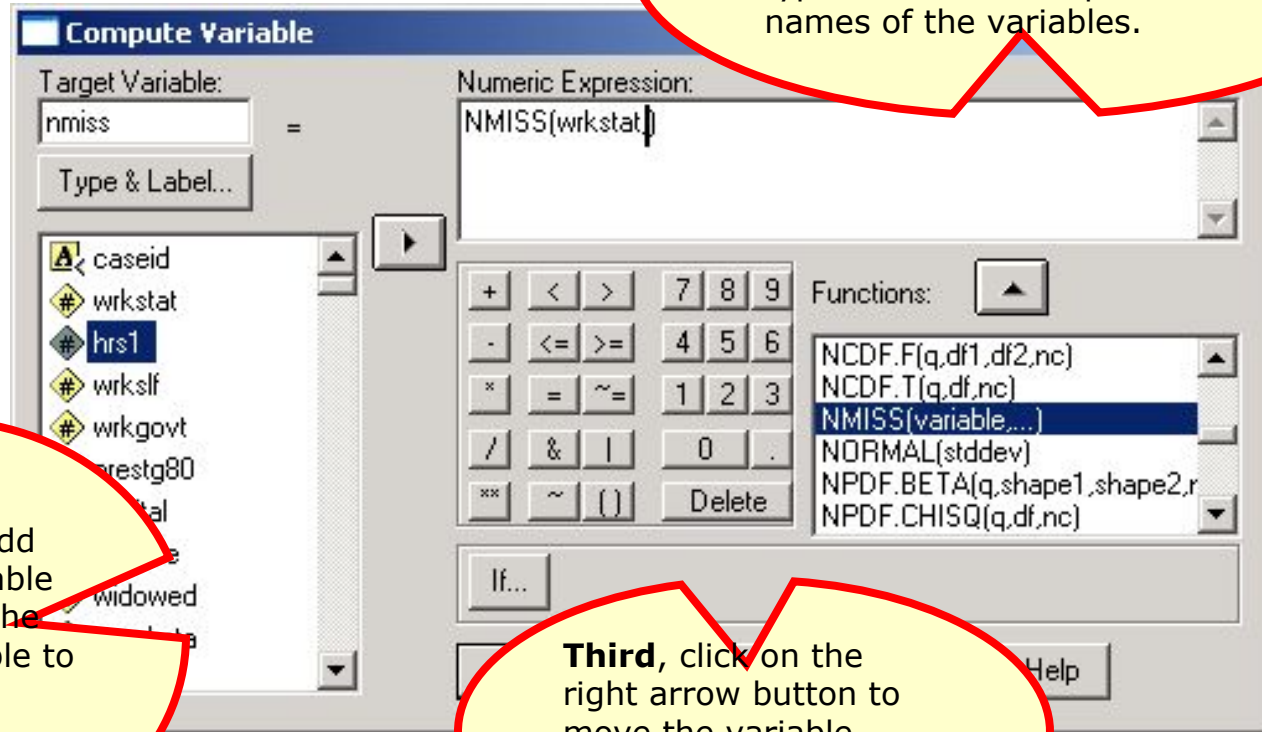


To add the list of variables to count missing data for, we **first** highlight the first variable to include in the function, *wrkstat*.

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

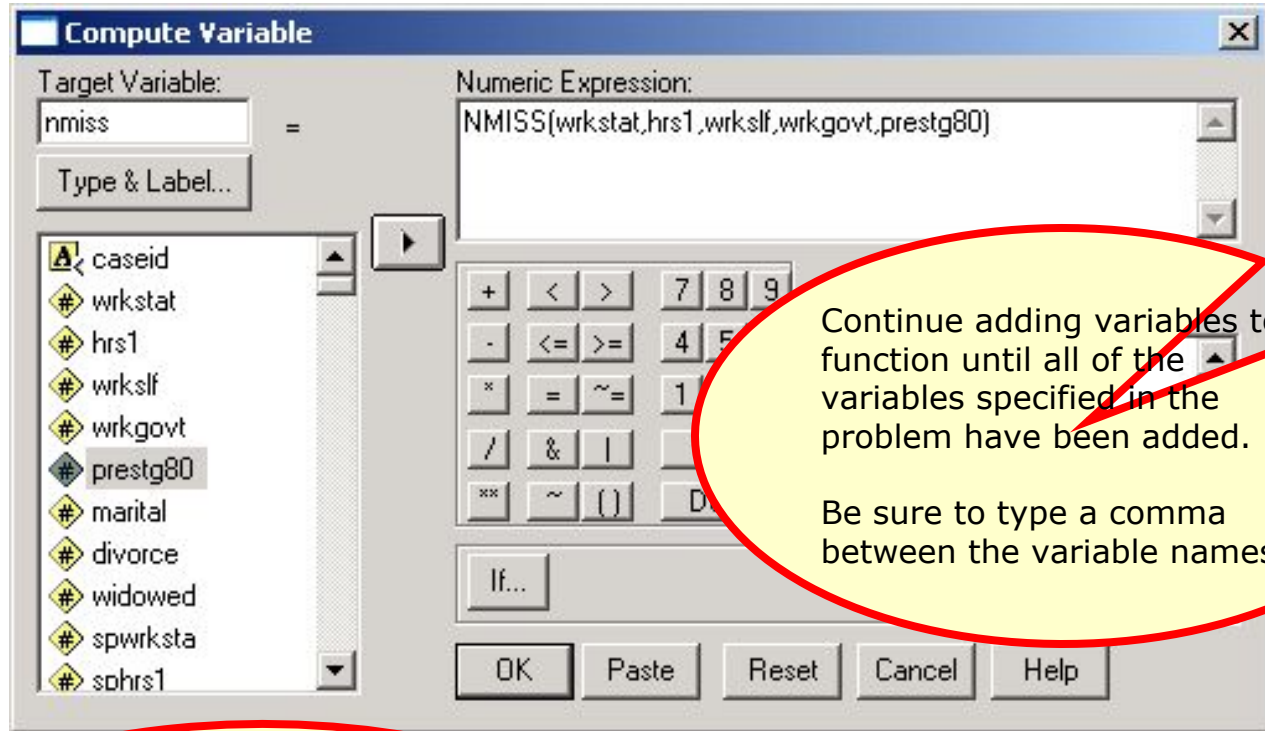
Enter specifications for new variable

14



Complete specifications for new variable

15



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

16

The screenshot shows the SPSS Data Editor window for 'GSS2000.sav'. The 'Data View' tab is active, displaying a dataset with 14 rows and 8 columns. The columns are 'zodiac', 'emtime', 'wwwtime', 'chattime', 'netime', 'nmiss', and 'var'. The 'nmiss' column contains values representing missing data for each row. A yellow callout box with a red border points to the 'nmiss' column, containing the text: 'If we scroll the worksheet to the right, we see the new variable that SPSS has just computed for us.'

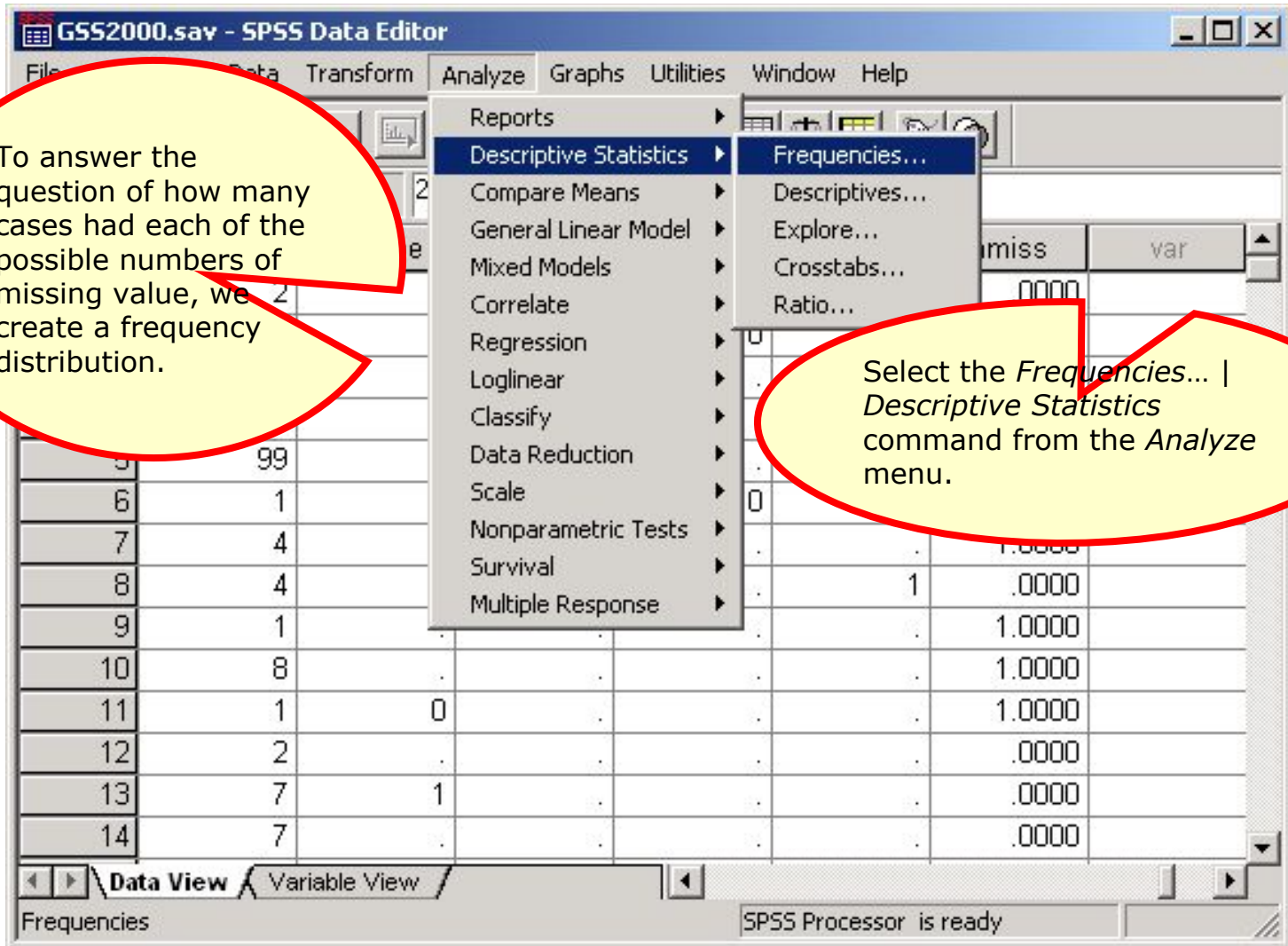
	zodiac	emtime	wwwtime	chattime	netime	nmiss	var
1	2	3	2	0	5	.0000	
2	6	4	6	0	10	.0000	
3	2	0	.	.	.	4.0000	
4	11	1.0000	
5	990000	
6	1	.	.	.	2	.0000	
7	4	1.0000	
8	40000	
9	1	1.0000	
10	8	1.0000	
11	1	0	.	.	.	1.0000	
12	20000	
13	7	10000	
14	70000	

A frequency distribution for *nmiss*

17

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

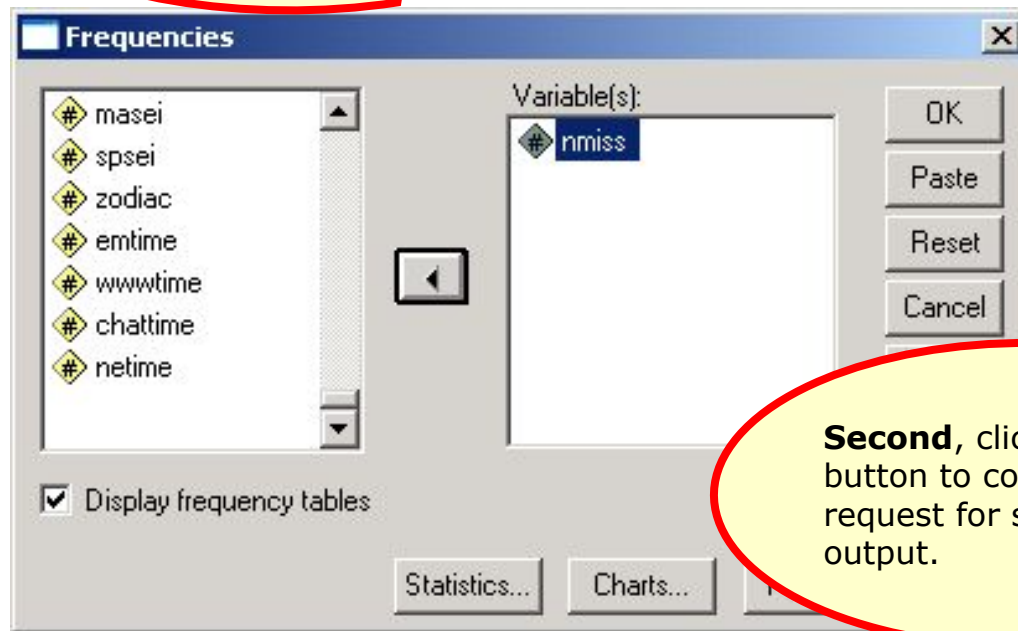
Select the *Frequencies...* | *Descriptive Statistics* command from the *Analyze* menu.



Completing the specification for frequencies

18

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

19

The screenshot shows the SPSS 'Frequencies' dialog box for the variable 'NMISS'. The 'Statistics' section is expanded, showing a summary table for NMISS. Below this, a larger table displays the frequency distribution for the variable, including counts, percentages, and cumulative percentages for valid and missing values.

Statistics

NMISS		
N	Valid	270
	Missing	0

NMISS

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

SPSS Processor is ready

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.

$H_1: \mu < 0$

$\sum_{i=1}^n w_i x_i = \beta$

$\sum_{i=1}^n w_i x_i + \beta = \beta$

$H_0: \mu = 0$

$\bar{x} = \frac{\sum x_i}{n}$

$s = \frac{\sum (x_i - \bar{x})^2}{n-1}$

$\sigma^2 = E(x - \mu)^2 = E(x_i + x_{i+1})$

$y = \frac{1}{2}(x_j + x_{j+1})$

Answering the problem

20

The screenshot shows the SPSS 'Frequencies' dialog box. The 'Statistics' section is expanded to show 'NMISS' (Number of Missing Values). Below it, a table shows that there are 270 valid cases and 0 missing cases. The 'PATTERN' section is also expanded, showing a table with columns for 'Frequency' and 'Percent' for various missing value counts (0.0000, 1.0000, 2.0000, 4.0000, and Total). A yellow callout bubble highlights the text: '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**.'

Statistics

NMISS		
N	Valid	270
	Missing	0

PATTERN

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

SPSS Processor is ready

Problem 3

21

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

22

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.

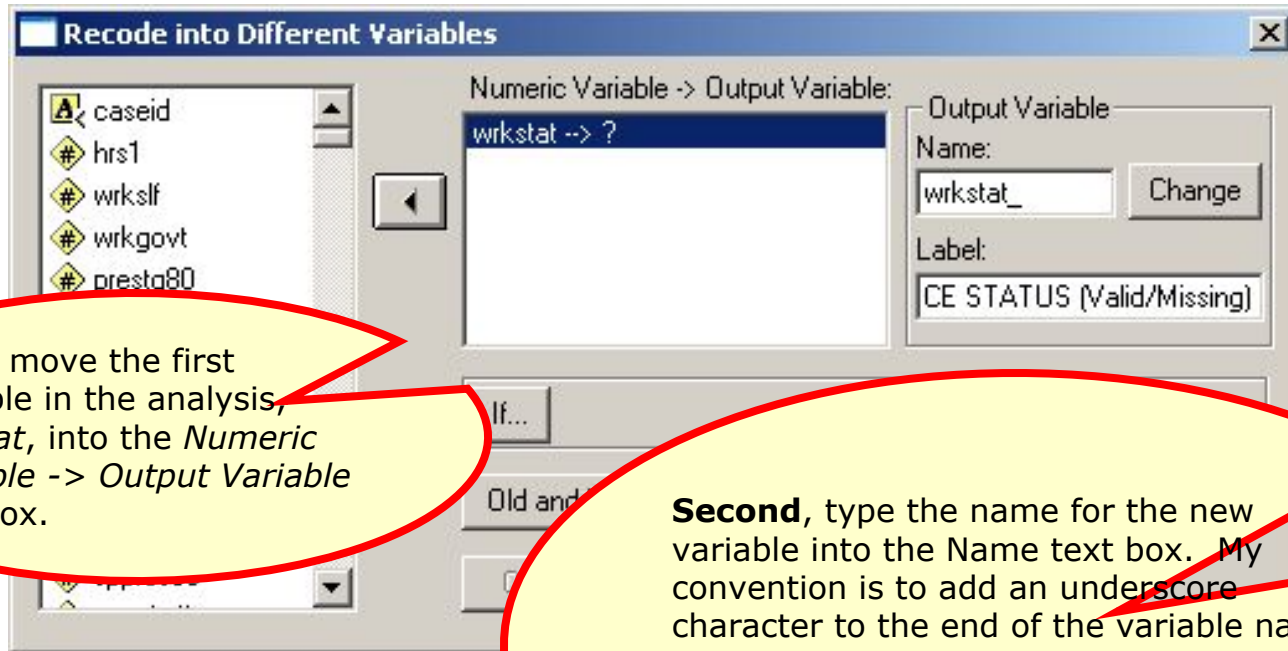
The screenshot shows the SPSS Data Editor window titled "GSS2000.sav - SPSS Data Editor". The "Transform" menu is open, and the "Recode" option is selected. A sub-menu is open, showing "Into Different Variables..." as the selected option. The background shows a data table with columns "marital" and "g80".

	marital	g80
4	1	51
5	1	74
7	1	38
10	2	40
11	9	40
12	2	35
13	2	51
14	2	33

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

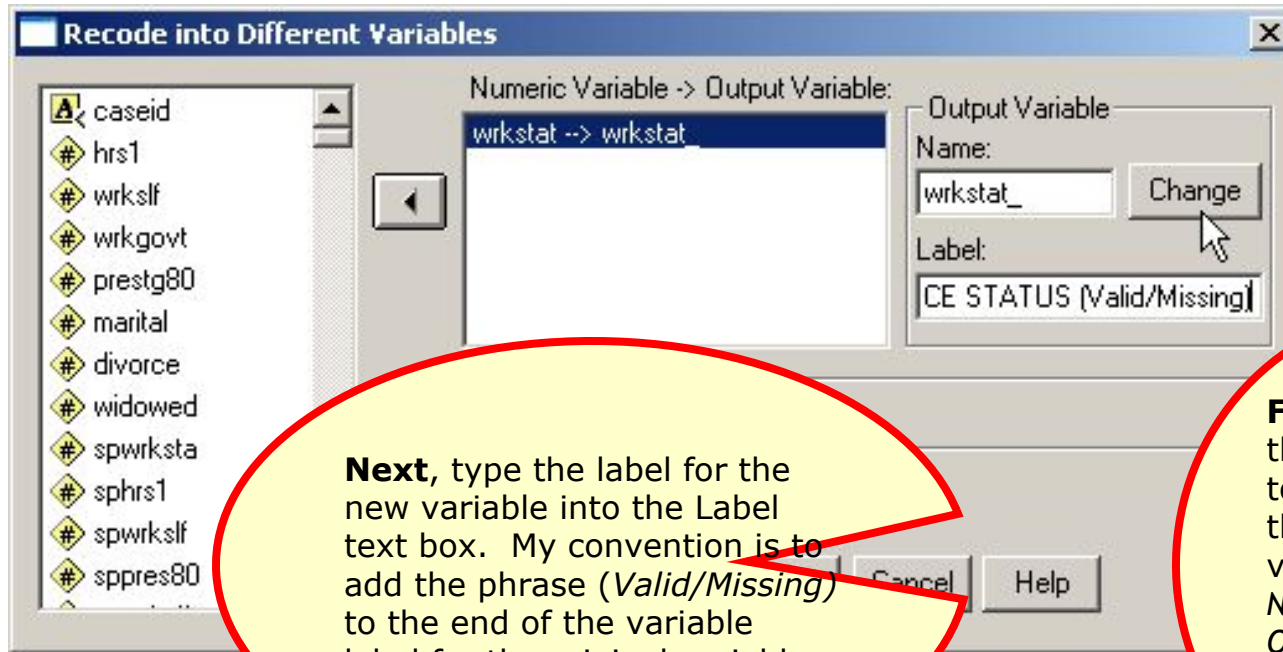
Enter specifications for new variable

23



Enter specifications for new variable

24

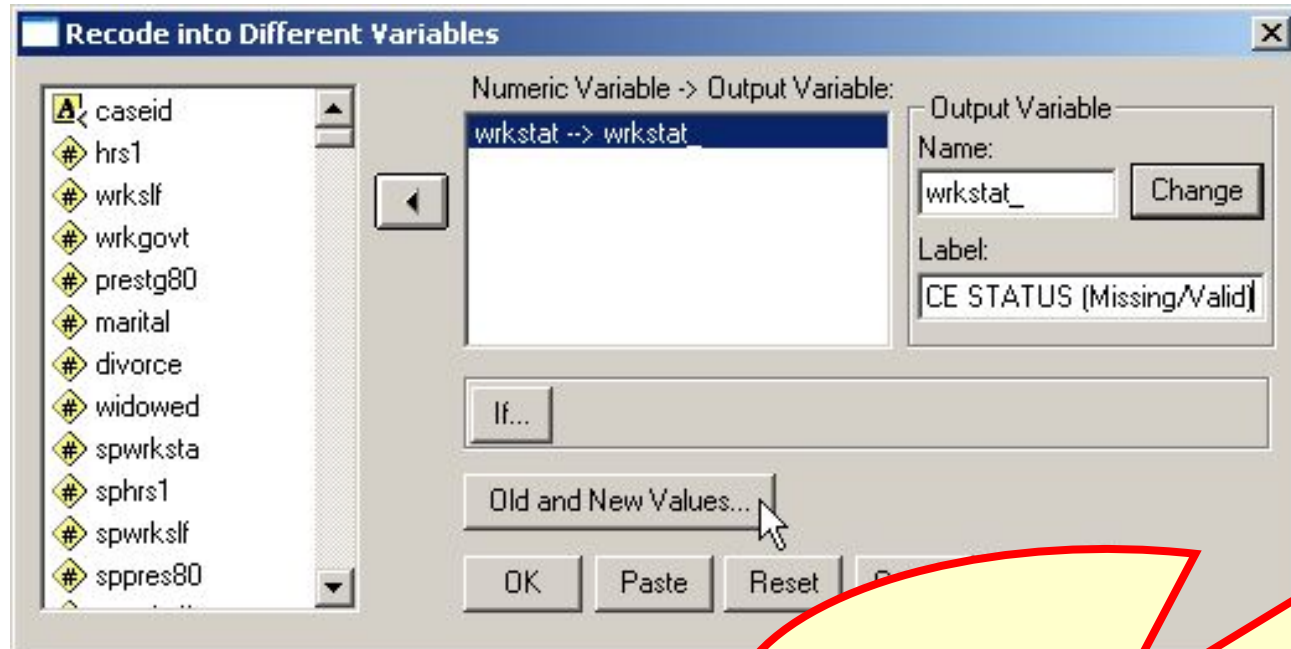


Next, type the label for the new variable into the Label text box. My convention is to add the phrase *(Valid/Missing)* to the end of the variable label for the original variable.

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

25



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

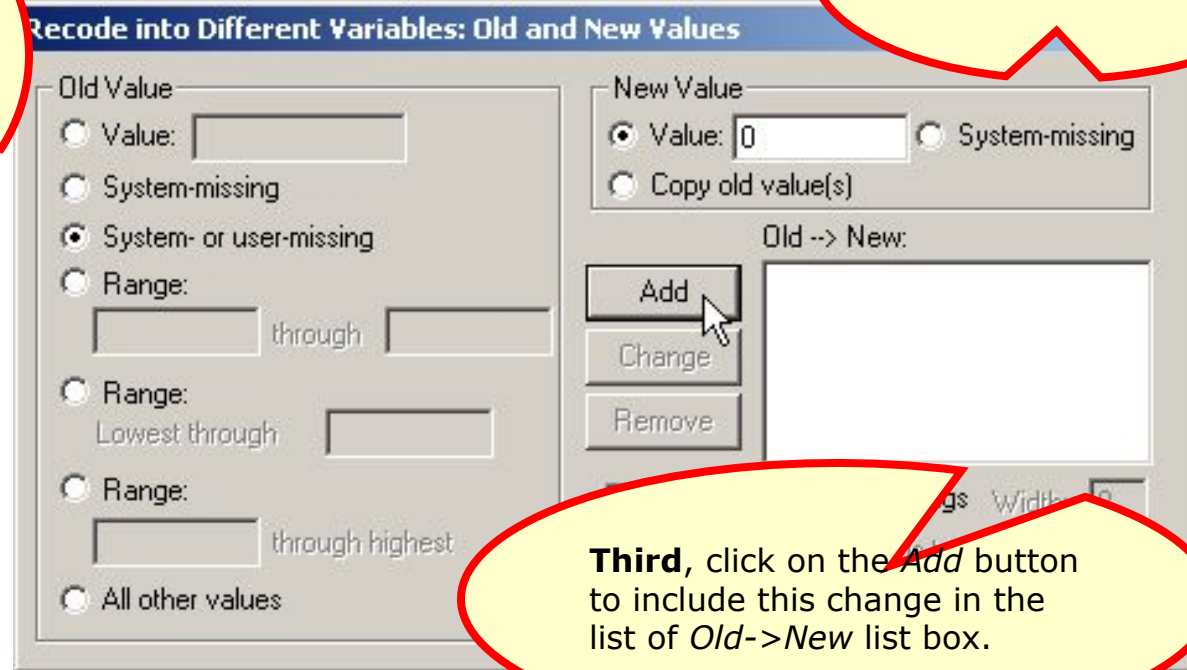
Change the value for missing data

26

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: System-missing
- Copy old value(s)

Old --> New:

Add
Change
Remove

Width:

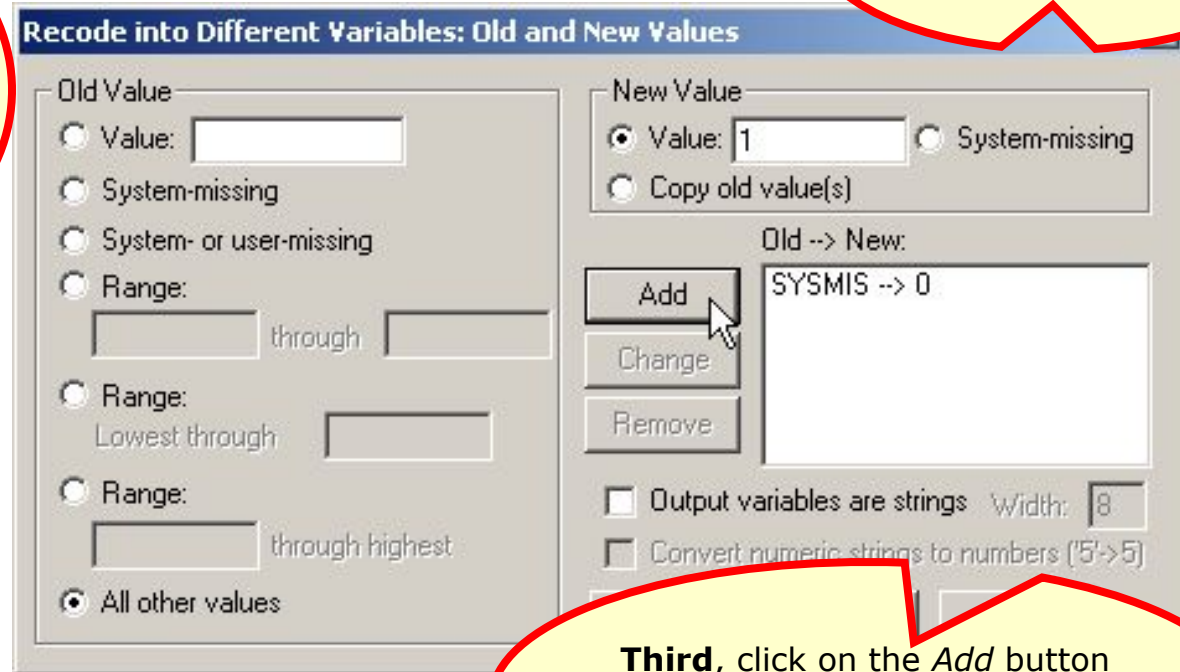
Third, click on the *Add* button to include this change in the list of *Old->New* list box.

Change the value for valid data

27

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

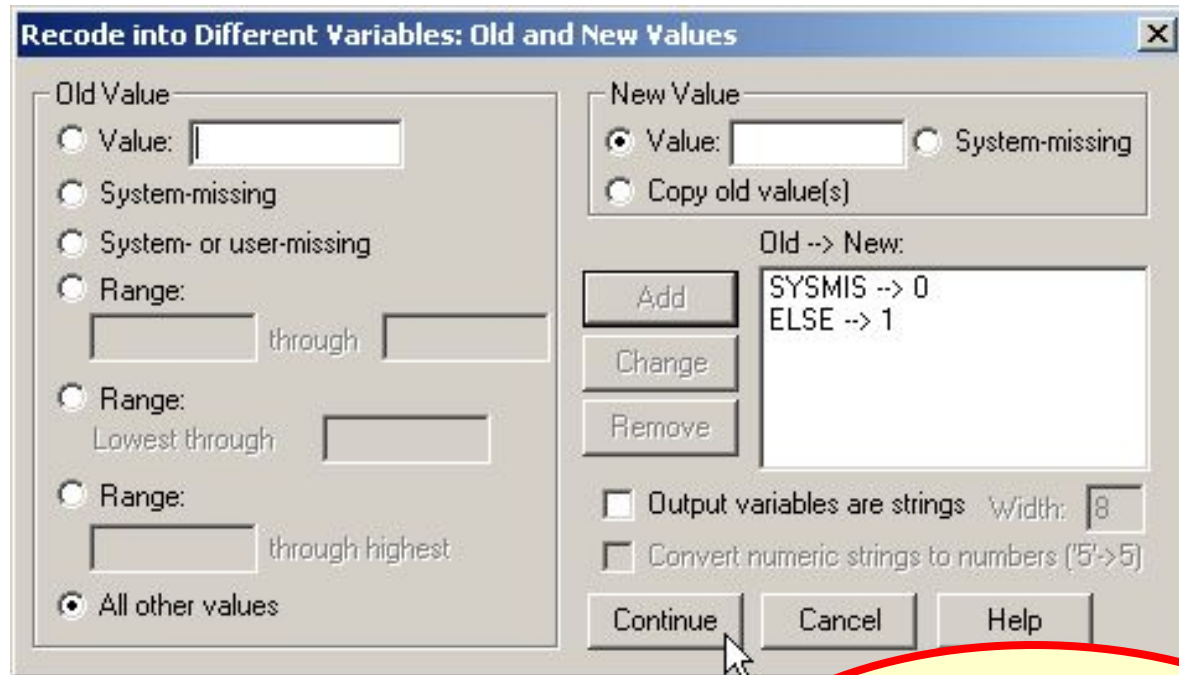
Second, type 1 in the Value text box.



The image shows the 'Recode into Different Variables: Old and New Values' dialog box in SPSS. The 'Old Value' section has the 'All other values' radio button selected. The 'New Value' section has the 'Value' radio button selected with '1' entered in the text box. The 'Old -> New:' list contains 'SYSMIS -> 0'. The 'Add' button is highlighted with a mouse cursor. The 'Output variables are strings' checkbox is checked with a width of 8. The 'Convert numeric strings to numbers' checkbox is unchecked.

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

Complete the value specifications



The image shows the 'Recode into Different Variables: Old and New Values' dialog box in SPSS. The 'Old Value' section has 'All other values' selected. The 'New Value' section has 'Value' selected. The 'Old --> New' list contains 'SYSMIS --> 0' and 'ELSE --> 1'. The 'Continue' button is highlighted with a red circle.

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:

- Add
- Change
- Remove
- SYSMIS --> 0
- ELSE --> 1

Output variables are strings Width:

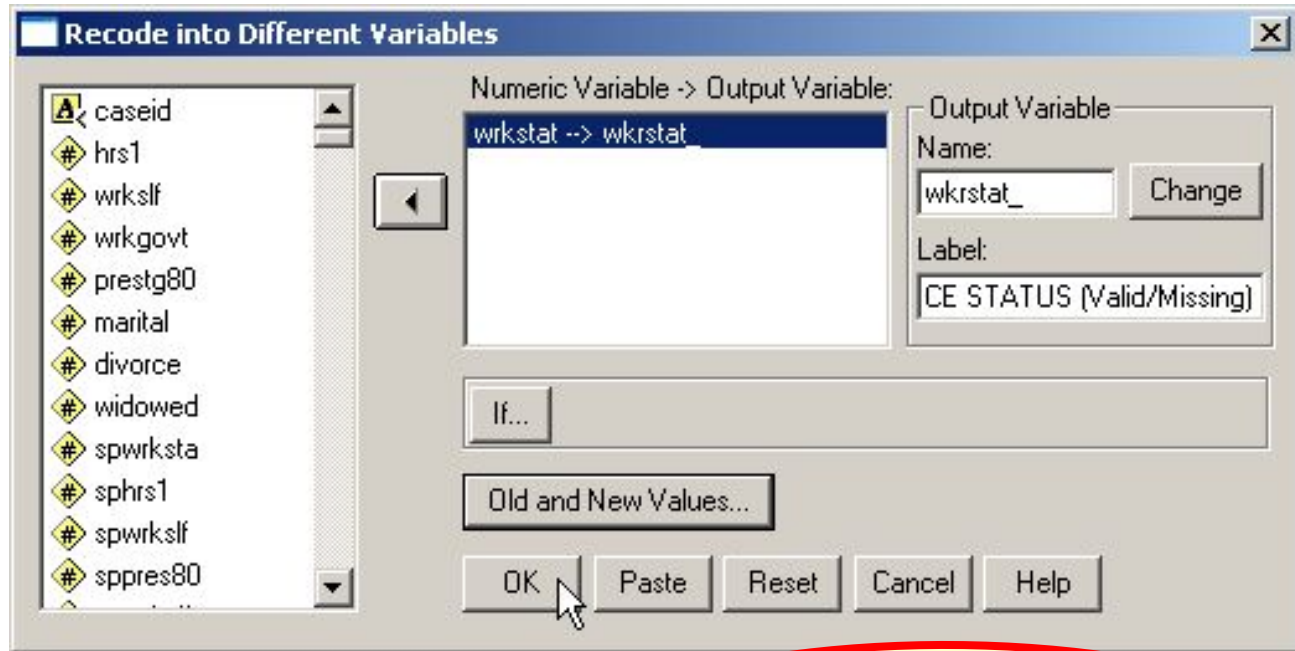
Convert numeric strings to numbers ('5'>5)

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

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i (b-1) = b^s$
 $\sum_{i=1}^n w_i x_i + \sum_{i=1}^n w_i y = x_j$
 $W = \sum_{i=1}^n w_i x_i (b-1) = b^s$
 $H_0: \mu = 0$
 $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = t$
 $\sigma^2 = E(x - \mu)^2 = t = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$
 $y = \frac{1}{2} (x_j + x_{j+1})$

Complete the recode specifications

29



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.

Filtering cases with excessive missing variables

31

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.

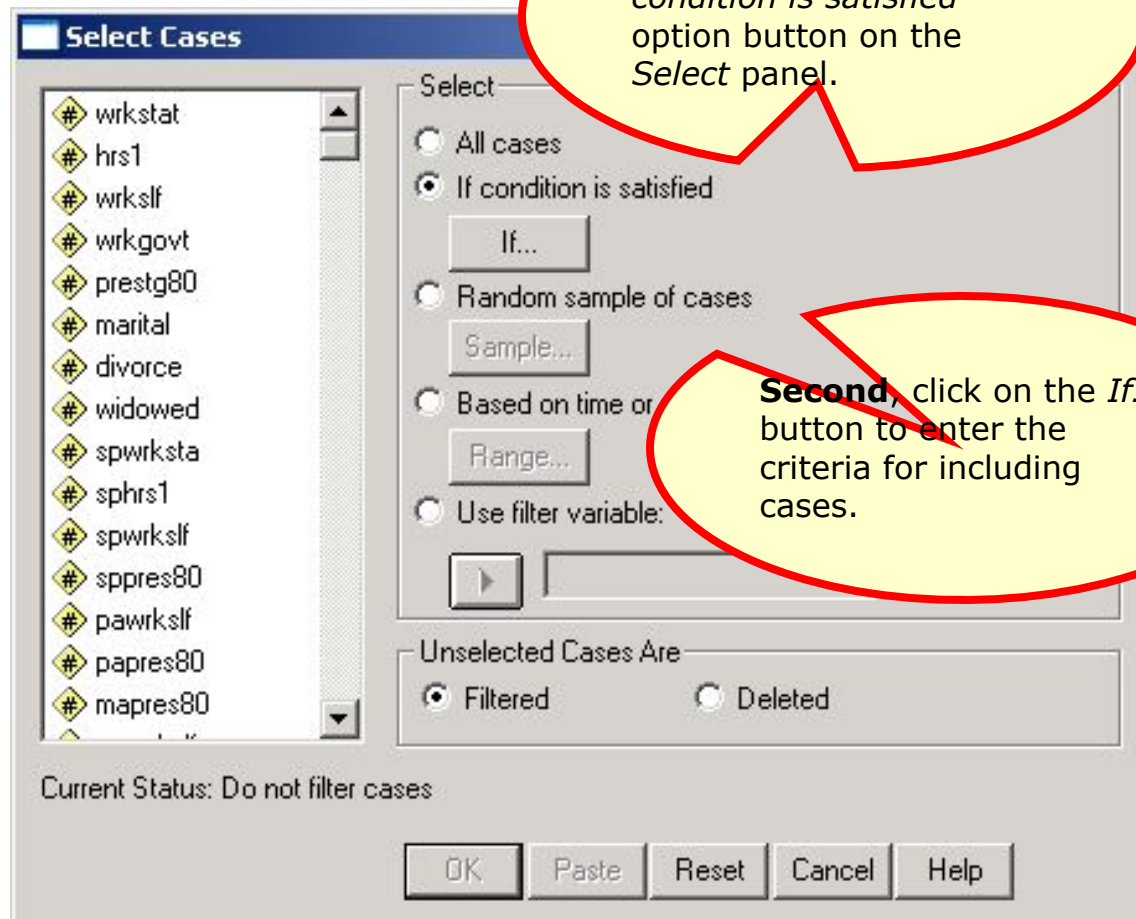
To filter cases included in further analysis, we choose the *Select Cases...* command from the *Data* menu.

The screenshot shows the SPSS Data Editor window for 'GSS2000.sav'. The 'Data' menu is open, and 'Select Cases...' is highlighted. The background shows a data table with columns 'hrs1_', 'work_16', 'worknovt_', and 'prestg8_'. The status bar at the bottom indicates 'Select Cases' and 'SPSS Processor is ready'.

	hrs1_	work_16	worknovt_	prestg8_
1	1	1	1	1.0000
2	1	1	1	1.0000
3	1	1	1	1.0000
4	1	1	1	1.0000
5	1	1	1	1.0000
6	1	1	1	1.0000
7	1	1	1	1.0000
8	1	1	1	1.0000
9	1	1	1	1.0000
10	1	1	1	1.0000
11	1	1	1	1.0000
12	1	1	1	1.0000
13	1	1	1	1.0000
14	1	1	1	1.0000

Enter specifications for selecting cases

32



$H_1: \mu < 0$

$\sum_{i=1}^n w_i x_i (\delta - 1) = 0$

$y = x_j$

$H_0: \mu = 0$

$\bar{x} = \frac{\sum x_j}{n}$

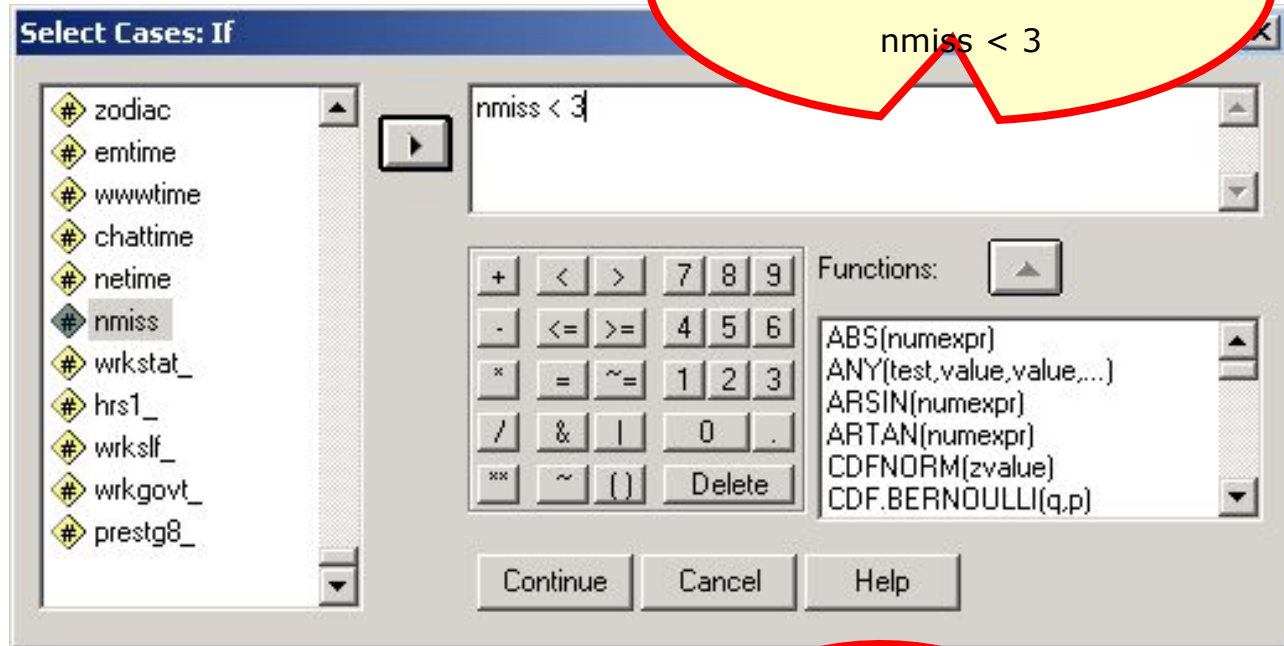
$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$\sigma^2 = E(x - \mu)^2$

$\bar{y} = \frac{1}{2} (x_j + x_{j+1})$

Enter specifications for selecting cases

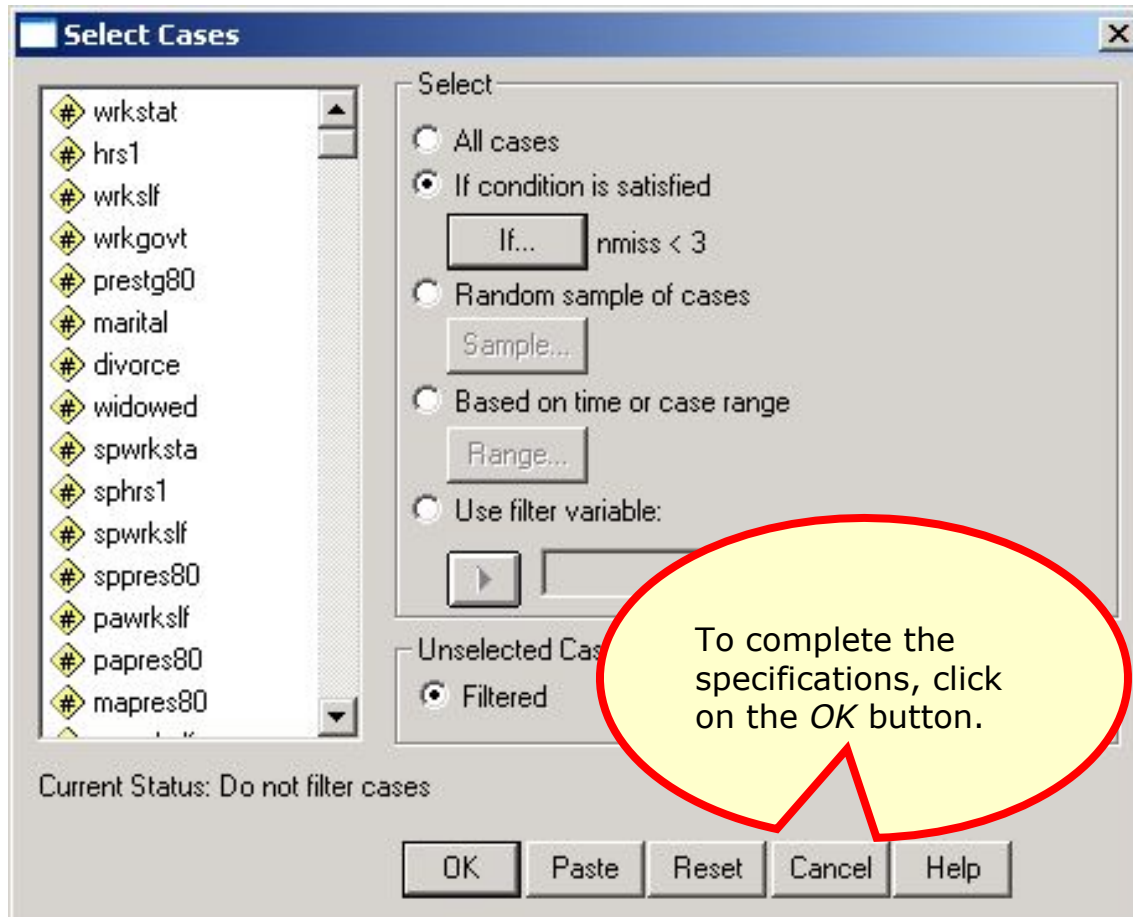
33



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

Complete the specifications for selecting cases

34



Current Status: Do not filter cases

OK Paste Reset Cancel Help

To complete the specifications, click on the OK button.

Cases excluded from further analyses

35

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.

	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	1.0000	1.0000	1.0000	1.0000	1.0000
7	.	.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	.	.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	.	.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	.	.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	.	.0000	1.0000	1.0000	1.0000	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

Correlating the dichotomous variables

36

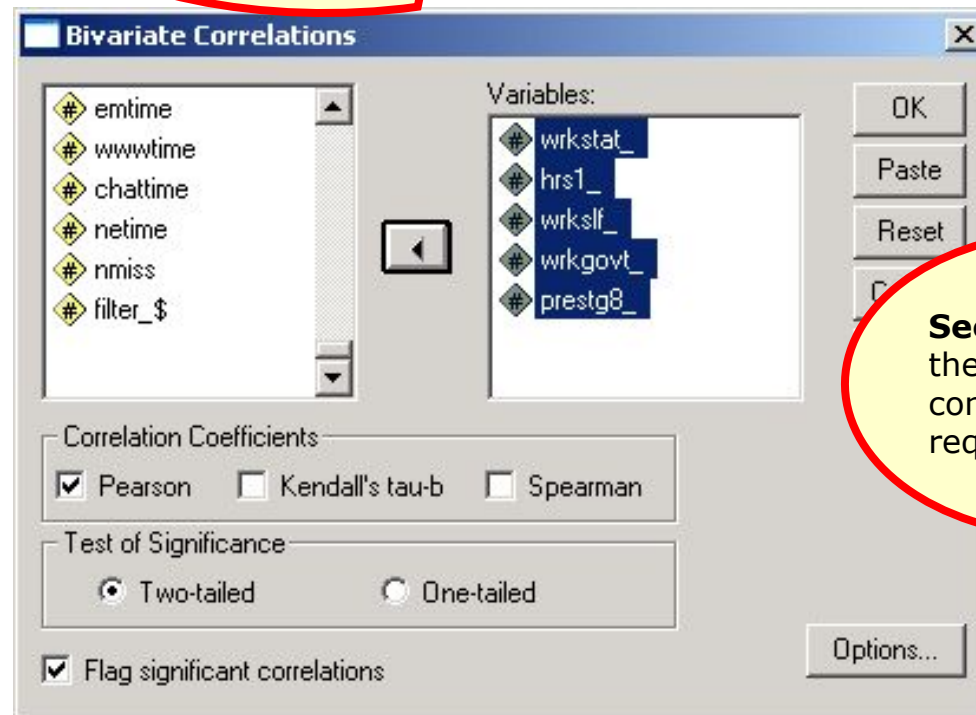
The screenshot shows the SPSS Data Editor window for a file named 'GSS2000.sav'. The 'Analyze' menu is open, and the 'Correlate' option is highlighted. A yellow callout box with a red border contains the following text: 'To compute a correlation matrix for the dichotomous variables, select the *Correlate* command from the *Analyze* menu.'

The data table in the background shows variables 'chattime' and 'netime' for cases 1 through 14. The 'netime' variable has values 0 or 1. Other variables visible in the table include 'wrkslf_', 'wrkgovt_', and 'prestg8_'. The status bar at the bottom indicates 'SPSS Processor is ready'.

Specifications for correlations

37

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



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

The correlation matrix

Correlations

		LABOR FORCE STATUS (Variable)	NUMBER OF HOURS WORKED LAST WEEK (Variable)	IS SELF-EMP OR WORKS FOR SOMEBODY (Variable)	GOVT OR PRIVATE EMPLOYEE (Variable)	IS OCCUPATIONAL PRESTIGE SCORE (Variable)
LABOR FORCE STATUS (Variable)	Pearson Correlation	1				
NUMBER OF HOURS WORKED LAST WEEK (Variable)	Pearson Correlation		1			
IS SELF-EMP OR WORKS FOR SOMEBODY (Variable)	Pearson Correlation			1		
GOVT OR PRIVATE EMPLOYEE (Variable)	Pearson Correlation				1	
IS OCCUPATIONAL PRESTIGE SCORE (Variable)	Pearson Correlation					1

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).

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

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i + g(x_i - 1) = 0$
 $y = x_j$
 $H_0: \mu = 0$
 $\bar{x} - \mu_0$
 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
 $\sigma^2 = E(x_i - \mu)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$
 $y = \frac{1}{2} (x_j + x_{j+1})$

The correlation matrix

Correlations

		LABOR FORCE STATUS (Variable)	NUMBER OF HOURS WORKED LAST WEEK (Variable)	IS SELF-EMP OR WORKS FOR SOMEBODY (Variable)	GOVT OR PRIVATE EMPLOYEE (Variable)	IS OCCUPATIONAL PRESTIGE SCORE (Variable)
LABOR FORCE STATUS (Variable)	Pearson Correlation	1				
NUMBER OF HOURS WORKED LAST WEEK (Variable)	Pearson Correlation	.228	1			
IS SELF-EMP OR WORKS FOR SOMEBODY (Variable)	Pearson Correlation	.228	.228	1		
GOVT OR PRIVATE EMPLOYEE (Variable)	Pearson Correlation	.228	.228	.228	1	
IS OCCUPATIONAL PRESTIGE SCORE (Variable)	Pearson Correlation	.228	.228	.228	.228	1

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.

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

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i + g x_{i+1} = R$
 $W = \sum_{i=1}^n w_i x_i + g x_{i+1} = R$
 $H_0: \mu = 0$
 $\bar{x} = \frac{\sum x_j}{n}$
 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
 $\sigma^2 = E(x - \mu)^2 = \frac{1}{n} \sum (x_j - \mu)^2$
 $\bar{y} = \frac{1}{2} (x_j + x_{j+1})$

The correlation matrix

Correlations

		LABOR FORCE STATUS (Variable)					
LABOR FORCE STATUS (Variable)	Pearson Correlation						
	Sig. (2-tailed)						
NUMBER OF HOURS WORKED LAST WEEK (Variable)	Pearson Correlation						
	Sig. (2-tailed)						
% SELF-EMP OR WORKS FOR SOMEBODY (Variable)	Pearson Correlation						
	Sig. (2-tailed)						
GOVT OR PRIVATE EMPLOYEE (Variable)	Pearson Correlation						
	Sig. (2-tailed)						
% OCCUPATIONAL PRESTIGE SCORE (Variable)	Pearson Correlation						
	Sig. (2-tailed)						

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.

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

$H_1: \mu < 0$
 $W = \sum_{i=1}^n w_i x_i + g(x_i - 1) = 0$
 $H_0: \mu = 0$
 $\bar{x} = \frac{\sum x_i}{n}$
 $s^2 = E(x_i - \mu)^2 = \frac{1}{n} \sum (x_i - \mu)^2$
 $s = \sqrt{s^2}$
 $y = \frac{1}{2} (x_i + x_{i+1})$
 $v = x_j$

Using scripts

41

- 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.
- Though 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

42

- 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

43

1 : id 201

	id	x1	x2	x3	x4	x5	x6
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	.

SPSS Processor is ready

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

Invoke the script

44

The screenshot shows the SPSS Data Editor window titled 'Hatmiss.sav - SPSS Data Editor'. The 'Utilities' menu is open, and the 'Run Script...' option is highlighted with a mouse cursor. A yellow callout bubble with a red border points to this option, containing the text: 'To invoke the script, select the Run Script... command in the Utilities menu.'

	id	x1	x2	x3	x4	x5	x6
1	201.0	3.3	.9	.0	.0	2.1	1.8
2	202.0	.	.4	.5	.0	1.2	1.7
3	203.0	3.0	.	.1	.0	3.5	3.4
4	204.0	.	1.5	.0	4.8	1.9	2.5
5	205.0	5.1	1.4	.0	.0	.0	2.6
6	206.0	4.6	2.1	7.0	.0	.4	.0
7	207.0	.	1.5	.0	.0	.0	.0
8	208.0	5.2	1.3	.0	.0	.0	.0
9	209.0	3.5	2.8	.0	.0	.0	.0
10	210.0	4.1	3.7	5.9	.0	.0	.0
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	.	.0	7.8	3.6	4.0
14	214.0	.	2.7	5.0	.0	2.2	.0

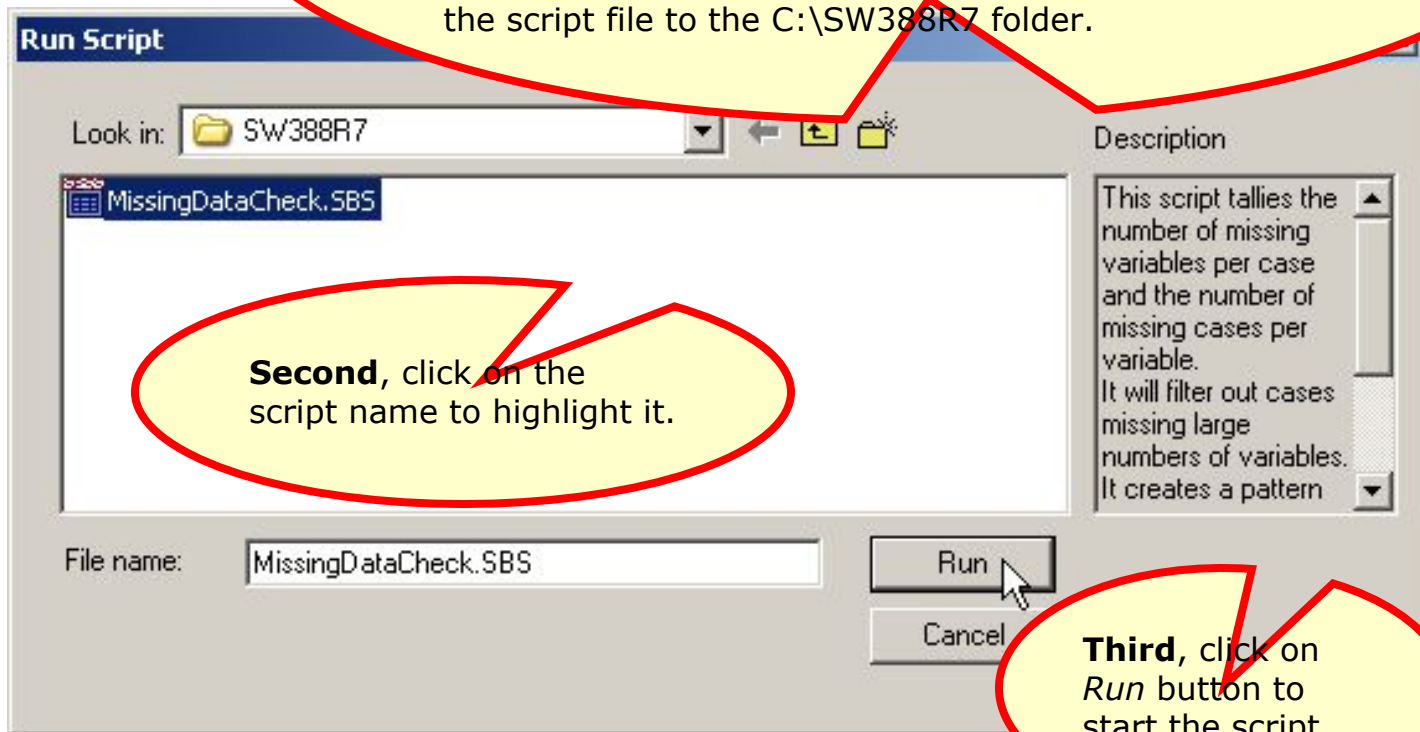
$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i (b-1) = b^0$
 $W = \sum_{i=1}^n w_i x_i + \sum_{i=1}^n x_i$
 $H_0: \mu = 0$
 $\bar{x} = \frac{\sum x_i}{n}$
 $s = \frac{\sum (x_i - \bar{x})^2}{n-1}$
 $\sigma^2 = E(x - \mu)^2 = \frac{1}{n} \sum (x_i - \mu)^2$
 $y = \frac{1}{2} (x_j + x_{j+1})$

Select the missing data script

45

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.

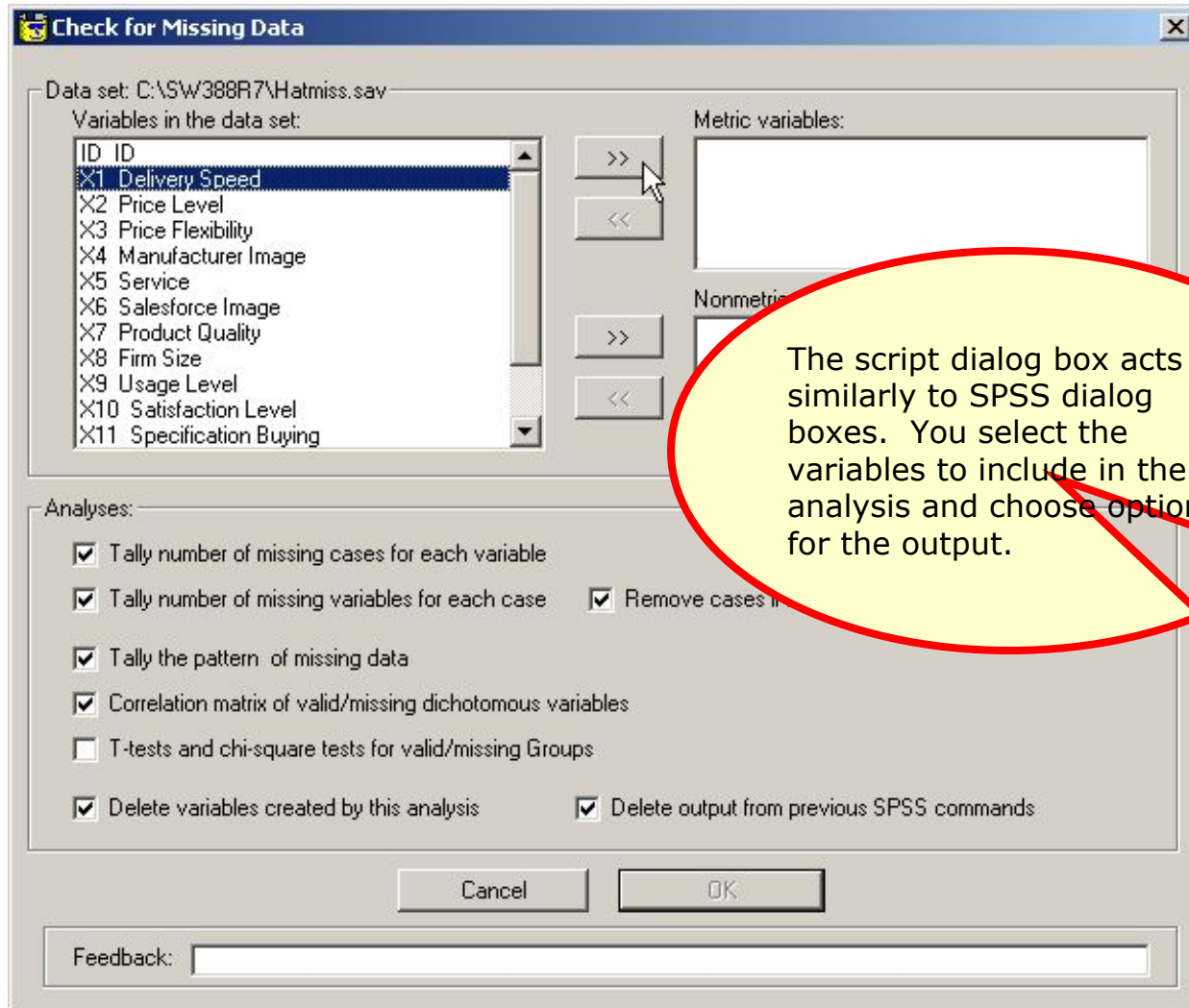


Second, click on the script name to highlight it.

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

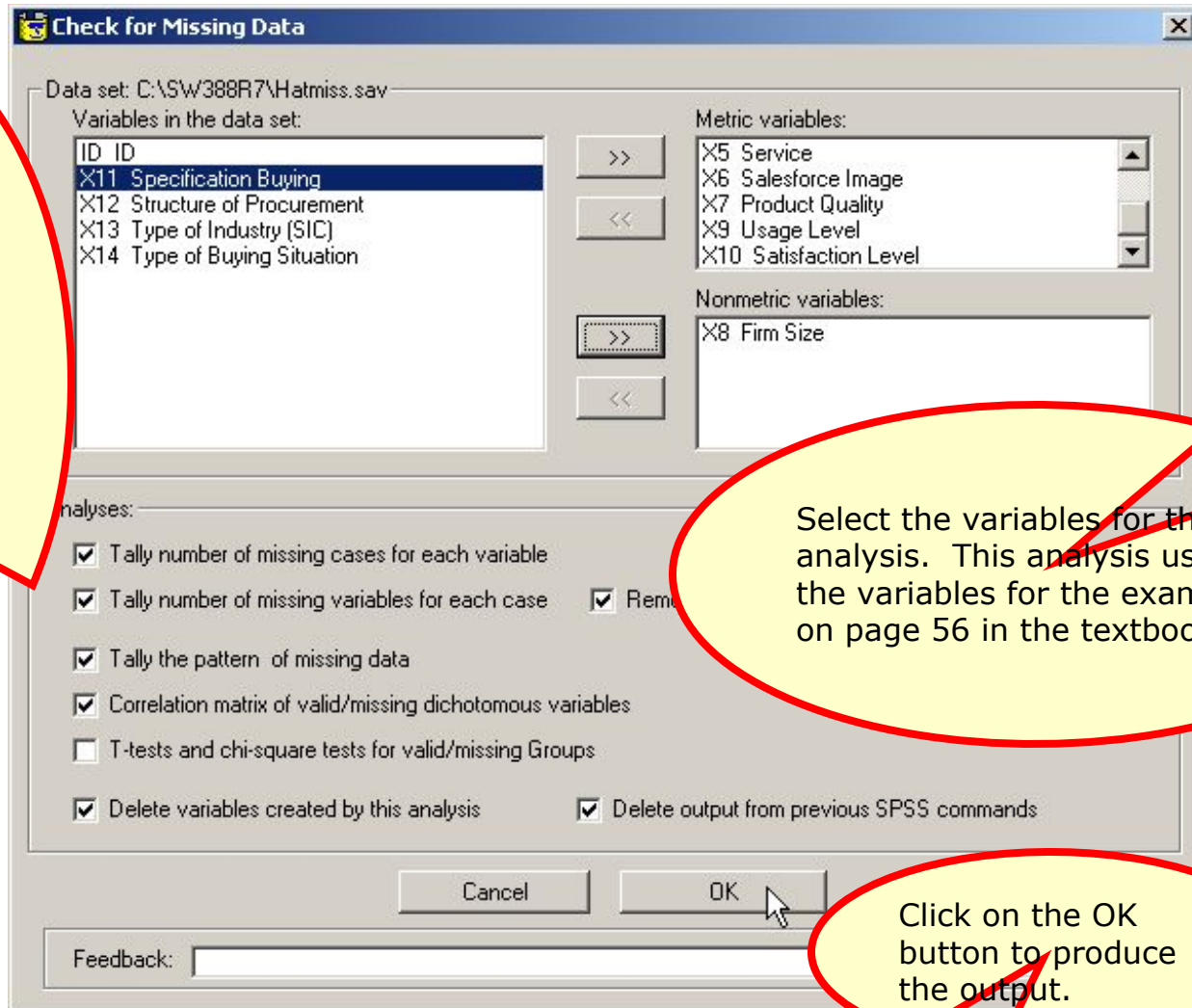
The script dialog

46



Complete the specifications

47



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

48

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.

$H_1: \mu < 0$
 $H_0: \mu = 0$
 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
 $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$
 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
 $\sigma^2 = E(x - \mu)^2$
 $\bar{y} = \frac{1}{2} (x_j + x_{j+1})$

Output from the script

49

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

Statistics

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

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

1 items selected (0 hidden/collapsed) SPSS Processor is ready

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i = \beta$
 $\sum_{i=1}^n w_i x_i + \beta = x_j$
 $H_0: \mu = 0$
 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
 $\sigma^2 = E(x - \mu)^2 = E(x_j + x_{j+1})$
 $y = \frac{1}{2}(x_j + x_{j+1})$