

Analyzing Missing Data

Introduction

Problems

Using Scripts

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

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.

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

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 two key issues for evaluating missing data:
 - The number or proportion of cases missing for each variable
 - Whether or not cases with missing data had statistically significant differences from cases with valid data for the other variables included in the analysis.
- Further analysis may be required depending on the problems identified in these analyses.

Benchmark for evaluating missing data

- The text suggests that, in general, if no more than 5% of the cases in the sample were missing data for a variable and if the pattern of missing data is random, missing data is not especially problematic for the analysis.

Our strategy for evaluating missing data

- The criteria lead us to a two stage strategy for evaluating the pattern of missing data.
- First, we will identify variables that are missing data for more than 5% of the cases in the sample.
 - If no variables are missing more than 5% of the cases, we will assume that there is not a problematic pattern.
- Second, for each variable that is missing data for more than 5% of the cases, we create a dichotomous missing/valid variable that is coded 0 for cases missing data and 1 for cases with valid data and test for statistically significant differences between the valid and missing groups for all other variables in the analysis.
 - If significant differences are found, we will attach a caution to our analysis with a recommendation for further study of the problems.

Testing for differences in missing/valid groups

- If the variable to be tested is metric, we use a t-test to compare the missing and valid groups.
- If the variable is nonmetric, we use a chi-square test of independence to compare the missing and valid groups.
- In all tests, we will use the level of significance stated in the problem for evaluating missing data and assumptions.

7

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

Example

- For example, suppose we are testing the relationship between the independent variables sex and age, and the dependent variable respondent's income. A frequency distribution on income indicates that 37.8% of the cases did not answer the question, so we create a dichotomous variable that is coded 0 for missing income and 1 for valid income.
- Since sex is a nonmetric variable, we do a chi-square test of independence with the missing/valid income as the independent variable and sex as the dependent variable to see if there is a relationship.
- Since age is a metric variable, we do a t-test to see if the average age for subjects who answered the question is different than the average age for subjects who skipped the question.

$$\begin{aligned} H_1: \mu < 0 \\ H_0: \mu = 0 \\ \bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \\ s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ s = \sqrt{s^2} \\ t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \end{aligned}$$

Problem 1

In the dataset GSS2000R, is the following statement true, false, or an incorrect application of a statistic? Use a level of significance of 0.01 for evaluating missing data and assumptions.

In pre-screening the data for use in a multiple regression of the dependent variable "total hours spent on the Internet" [netime] with the independent variables "age" [age], "highest year of school completed" [educ], and "sex" [sex], the missing data analysis did not indicate any need for caution or further analysis for a problematic pattern of missing data.

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

Checking level of measurement

10

Since we are pre-screening for a multiple regression problem, we should make sure we satisfy the level of measurement before proceeding.

For GSS2000R, is the following statement true, level of measurement for

"Total hours spent on the Internet" [netime] is interval, satisfying the metric level of measurement requirement for the dependent variable.

In pre-screening the data for use in a multiple regression of the dependent variable "total hours spent on the Internet" [netime] with the independent variables "age" [age], "highest year of school completed" [educ], and "sex" [sex], the missing data analysis did not indicate any need for caution or further analysis for a problematic pattern of missing data.

1. "Age" [age] and "highest year of school completed" [educ] are interval, satisfying the metric or dichotomous level of measurement requirement for independent variables.

2. "Sex" [sex] is dichotomous, satisfying the metric or dichotomous level of measurement requirement for independent variables.

3. False

4. Inappropriate application of a statistic

Request frequency distributions

11

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.

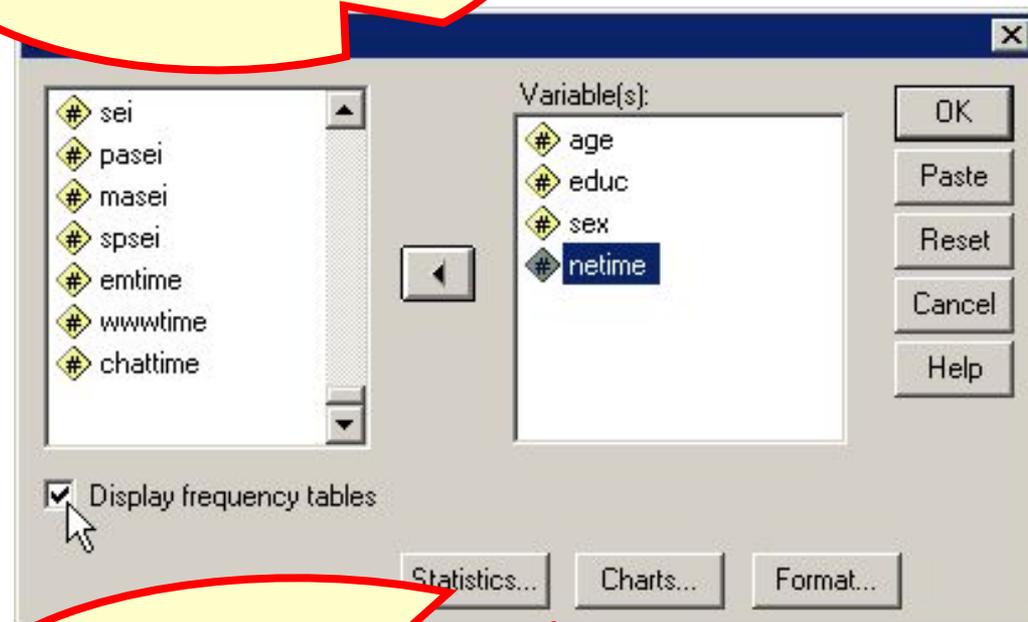
The screenshot shows the SPSS Data Editor window for 'GSS2000R.sav'. The 'Analyze' menu is open, and the 'Descriptive Statistics' sub-menu is selected, with 'Frequencies...' highlighted. The background shows a data table with columns 'estg80' and 'marital'. The status bar at the bottom indicates 'SPSS Processor is ready'.

	estg80	marital
1	51	1
2	74	1
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		

Completing specifications for frequencies - 1

12

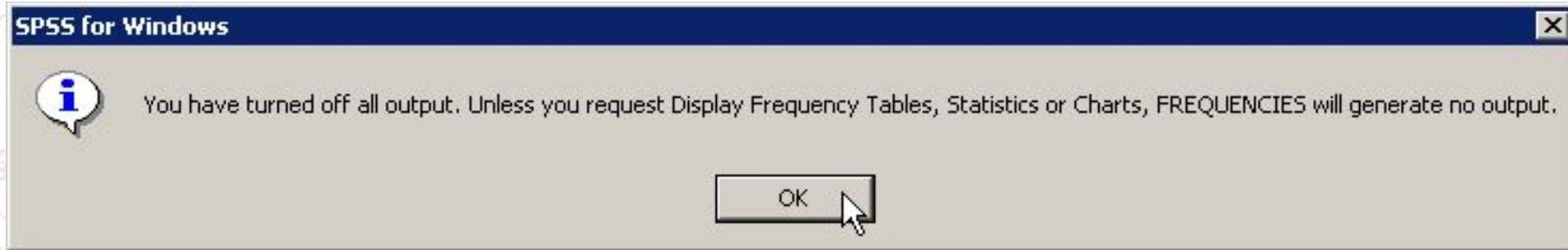
First, move the four variables included in the problem statement to the list box for variables.



Second, click on the *Display frequency tables* check box to clear it, since all we want is the statistics for missing and valid cases.

Completing specifications for frequencies - 2

13

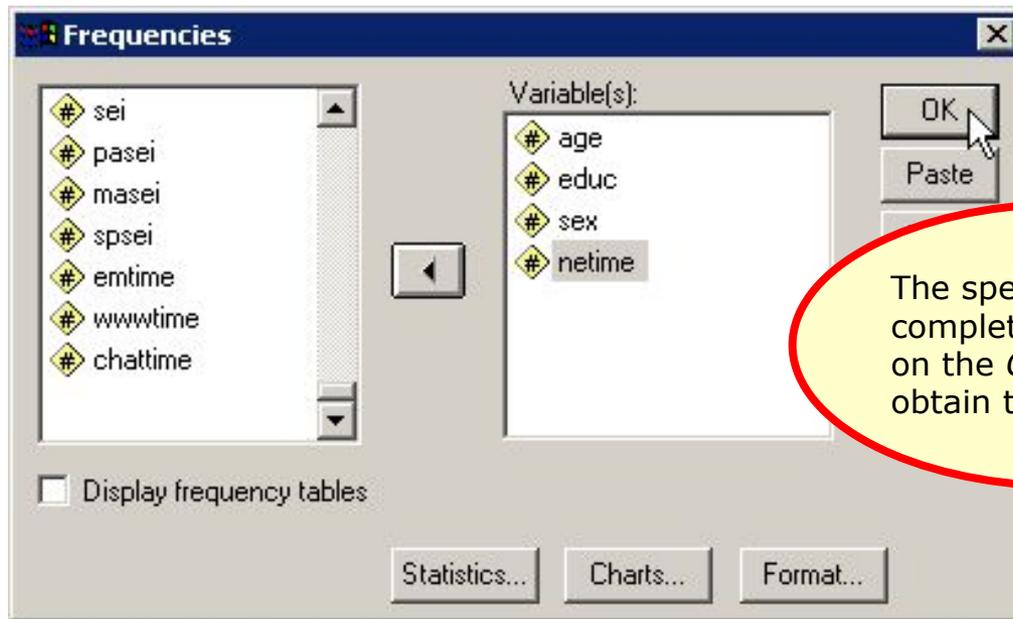


SPSS give us a warning message that we will not generate any output. However, it will produce the statistics for valid an missing data which is want we want.

Click on the *OK* button to close the warning.

Completing specifications for frequencies - 3

14



The specifications are complete, so we click on the OK button to obtain the output.

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

Number of missing cases for each variable - 1

15

Output1 - SPSS Viewer

File Edit View Insert Format Analyze Graphs Utilities

With 270 cases in the data set, a variable is missing more than 5% of the cases if it had 14 or more cases with missing value.

→ **Frequencies**

Statistics

		AGE	EDUC	SEX	NETIME
N	Valid	270	269	270	93
	Missing	0	1	0	177

The variables "age" [age], "highest year of school completed" [educ], and "sex" [sex] were missing data for less than 5% of the cases in the data set. T-tests and chi-square tests to compare cases with missing data to cases with valid data for the other variables included in the analysis were not conducted.

Number of missing cases for each variable - 2

16

With 270 cases in the data set, a variable is missing more than 5% of the cases if it had 14 or more cases with missing value.

Statistics

		AGE	EDUC	SEX	NETIME
N	Valid	270	269	270	93
	Missing	0	1	0	177

One variable was missing data for more than 5% of the cases in the data set: "total hours spent on the Internet" [netime] was missing data for 65.6% of the cases in the data set (177 of 270 cases). A missing/valid dichotomous variables was created for this variable to test whether the group of cases with missing data differed significantly from the group of cases with valid data on the other variables included in the analysis.

Creating the missing/valid variable - 1

17

We will create a new variable whose values represent cases with missing or valid data.

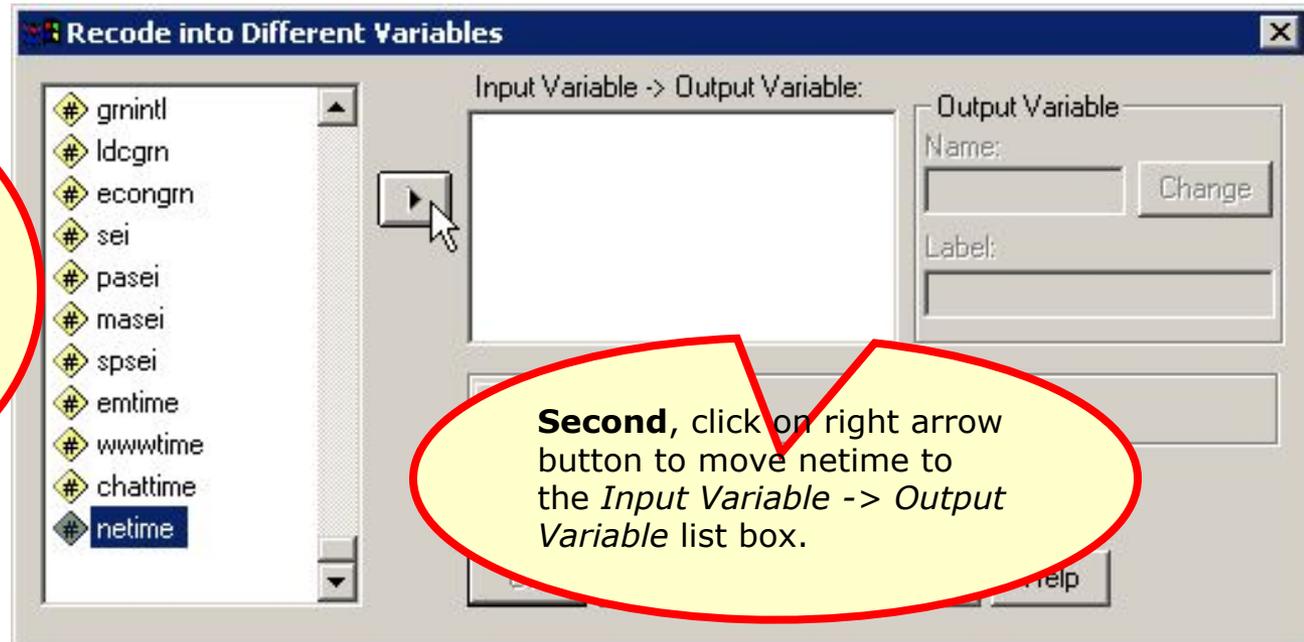
First, select the *Recode* | *Into Different Variables...* command from the *Transform* menu.

caseid	tg80	marital
1 20000009	51	1
2 20000012	74	1
3 20000020		1
4 20000029		
5 20000032		
6 20000034		
7 20000043		
8 20000060	38	5
9 20000070	35	5
10 20000072	36	2
11 20000079	64	1
12 20000097	35	1
13 20000117	51	3
14 20000126	33	3

Creating the missing/valid variable - 2

18

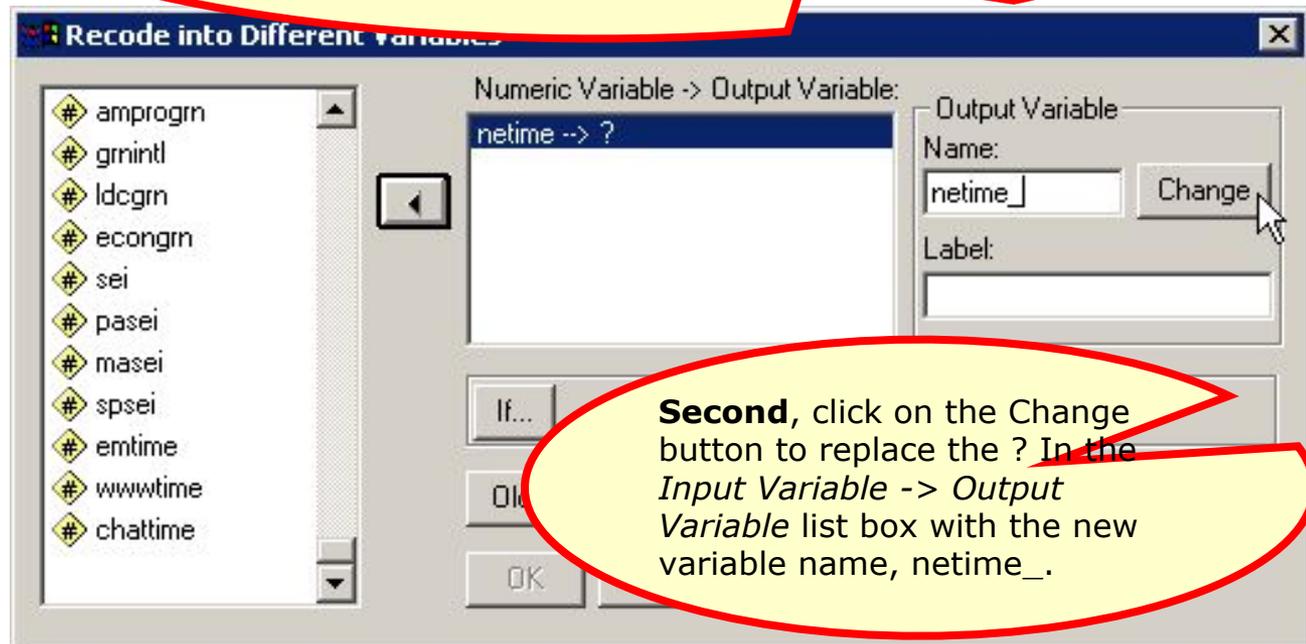
First, highlight the variable netime, which is the variable which had more than 5% missing data, for which we want to create the missing/valid variable.



Creating the missing/valid variable - 3

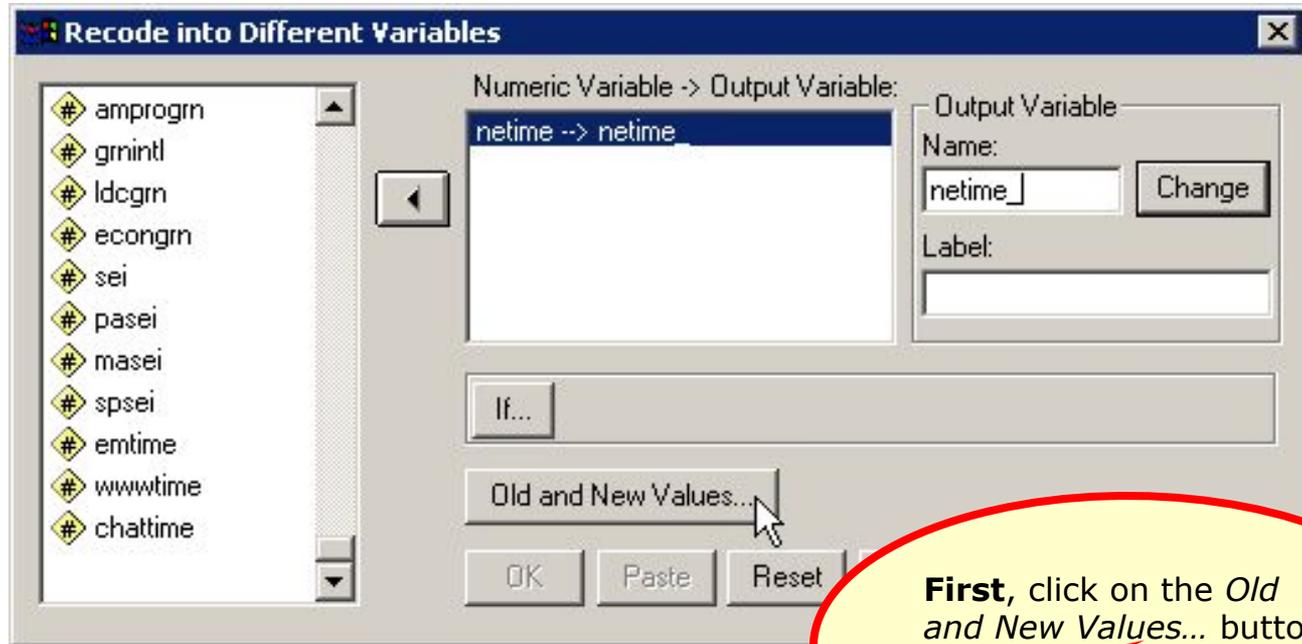
19

First, type a name for the new variable into the Name: text box. I usually just add an underscore to the variable name if the original variable name is 7 letters or less. If the variable is 8 letters, I delete the last letter so that I do not exceed the SPSS requirement that a variable name be 8 characters or less.



Creating the missing/valid variable - 4

20



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

Creating the missing/valid variable - 5

21

First, to create the code 0 for missing data, we mark the *System- or user-missing* option button on the *Old Value* panel.

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:

Second, in the Value: text box in the New Value panel, we type a zero.

Third, click on the *Add* button to add the change from missing to zero to the list *Old* \square *New*.

Creating the missing/valid variable - 6

22

First, to create the code 1 for valid data, we mark the *All other values* option button on the *Old Value* panel.

Second, in the Value: text box in the New Value panel, we type a one.

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:

MISSING -> 0

Buttons: Add, Change, Remove, Continue, Cancel, Help

Third, click on the *Add* button to add the change from other values to one to the list *Old* \square *New*.

Creating the missing/valid variable - 7

23

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

MISSING -> 0
ELSE -> 1

Output variables are strings Width:

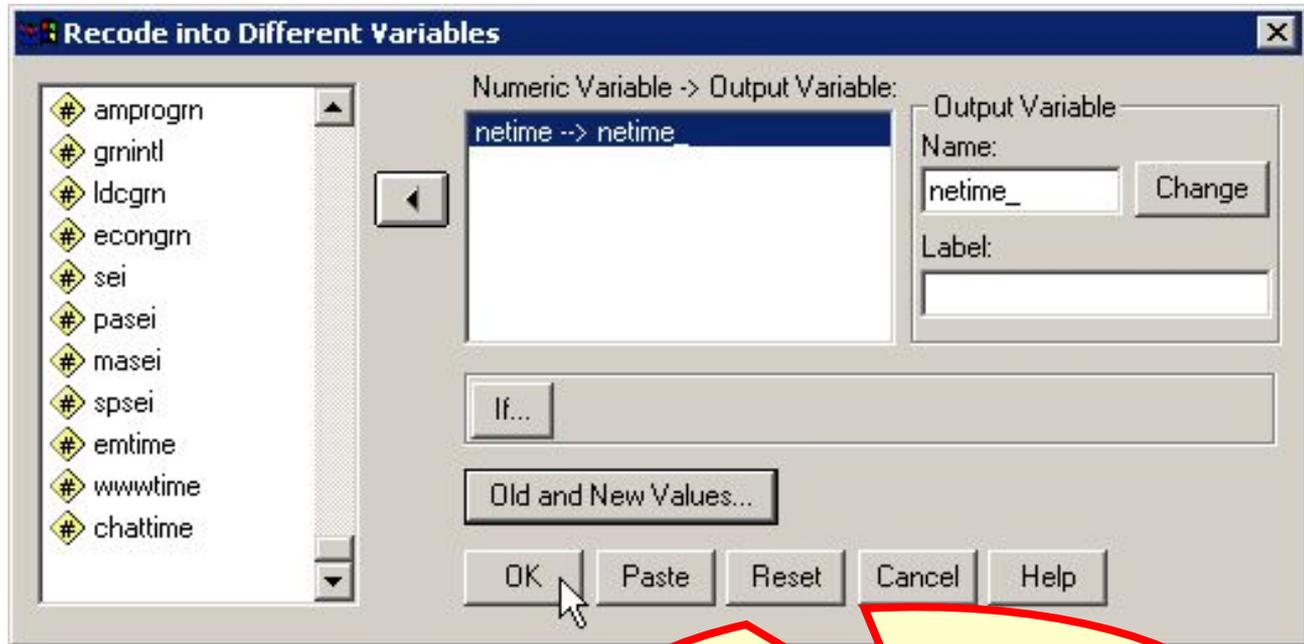
Convert numeric strings to numbers ('5'>5)

Continue Cancel Help

Having completed the changes, we click on the *Continue* button to close the dialog box.

Creating the missing/valid variable - 8

24



Click on the OK button to indicate the completion of the specifications for the new variable.

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

The missing/valid variable in the data editor

25

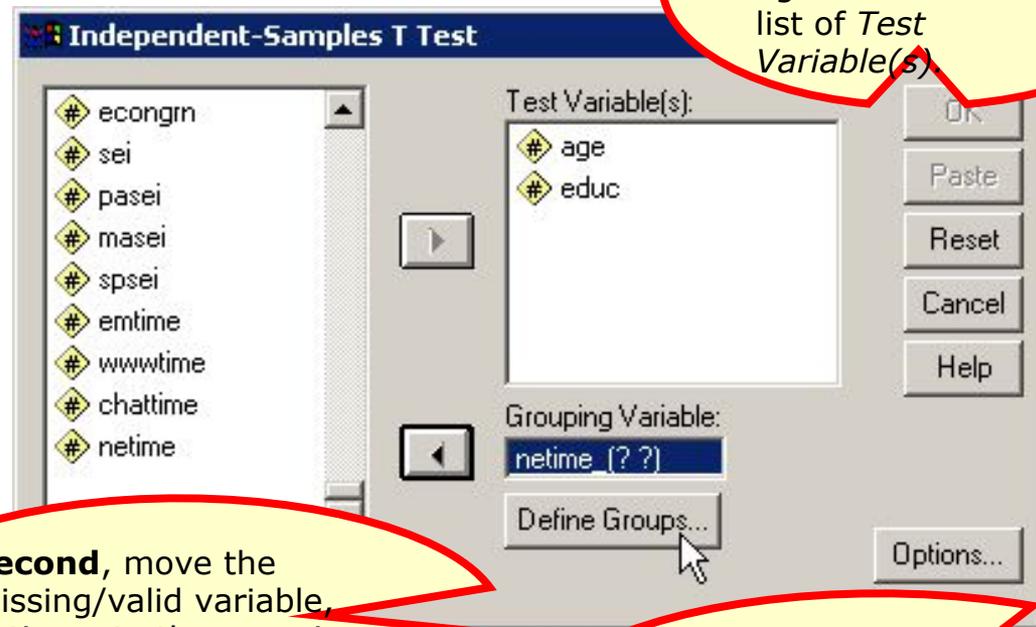
The screenshot shows the SPSS Data Editor window for 'GSS2000R.sav'. The data is displayed in 'Data View' with columns: spsei, emtime, wwwtime, chattime, netime, netime_, and var. The 'netime_' column contains values 1 for non-missing 'netime' and 0 for missing values. A yellow callout box with a red border contains the following text:

If we look at the newly created netime_ variable in the data editor, we see that valid data for netime (4.50, 10.0, etc) correspond to a 1 for netime_, while missing data indicators, ".", correspond to 0.

	spsei	emtime	wwwtime	chattime	netime	netime_	var
1	92.3	3.00	1.50	.00	4.50	1	
2	63.5	4.00	6.00	.00	10.00	1	
3	53.3	.00	.	.	.	0	
4	0	
5	38.2	0	
6	.	.30	.	.	.	1	
7	0	
8	1	
9	0	
10	0	
11	78.5	0	
12	29.2	0	
13	.	.50	.	.	.	0	
14	0	

T-tests comparing missing and valid cases - 2

27



First, move the metric variables age and educ to the list of Test Variable(s).

Second, move the missing/valid variable, netime_ to the grouping variable text box.

Third, click on the *Define Groups...* button to specify the codes for the groups to compare in the analysis.

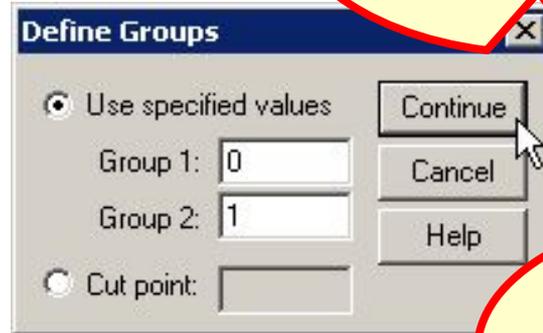
$H_1: \mu < 0$
 $\sum_{i=1}^n w_i x_i = \bar{x}$
 $\sum_{i=1}^n w_i (x_i - \bar{x}) = 0$
 $H_0: \mu = 0$
 $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
 $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$
 $\sigma^2 = E(x - \mu)^2$
 $\bar{y} = \frac{1}{2}(x_j + x_{j+1})$

T-tests comparing missing and valid cases - 3

28

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

First, type the number 0 for the missing group into the *Group 1* text box.

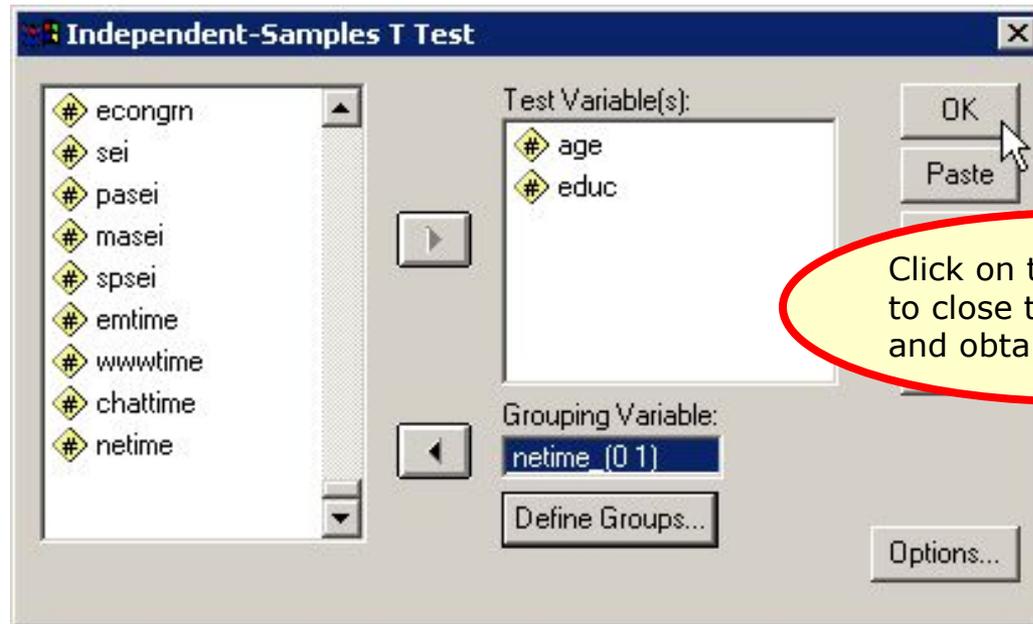


Second, type the number 1 for the valid group into the *Group 2* text box.

Third, click on the *Continue* button complete the definition of the groups for the independent variable.

T-tests comparing missing and valid cases - 4

29



Click on the OK button to close the dialog box and obtain the output.

$H_1: \mu < 0$
 $\sum_{i=1}^n w_i (x_i - \mu) = 0$
 $\sum_{i=1}^n w_i x_i = \sum_{i=1}^n w_i \mu$
 $\sum_{i=1}^n w_i x_i = \mu \sum_{i=1}^n w_i$
 $\mu = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$
 $\sigma^2 = E(x - \mu)^2 = E(x^2 - 2x\mu + \mu^2) = E(x^2) - 2\mu E(x) + \mu^2$
 $\sigma^2 = E(x^2) - 2\mu \sum_{i=1}^n w_i x_i + \mu^2 \sum_{i=1}^n w_i$
 $\sigma^2 = E(x^2) - 2\mu \sum_{i=1}^n w_i x_i + \mu^2 \sum_{i=1}^n w_i$
 $\sigma^2 = E(x^2) - 2\mu \sum_{i=1}^n w_i x_i + \mu^2 \sum_{i=1}^n w_i$

Output for the t-tests - 1

30

Group Statistics

	NETIME_	N	Mean	Std. Deviation	Std. Error Mean
AGE	0	177	48.32	18.414	1.384
	1	93	41.55	12.117	1.256
EDUC	0	177	12.34	2.812	.212
	1	92	14.62	2.554	.255

There were significant differences in the statistical tests comparing cases with missing data to cases with valid data.

Cases who had missing data for the variable "total hours spent on the Internet" [netime] had an average score on the variable "age" [age] that was 6.77 units higher than the average for cases who had valid data (t=3.624, p<0.001).

Indepe

		Levene's Test for Equality of Variances		Independent-Samples T-Test			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
AGE	Equal variances assumed	22.640	.000	3.201	268	.002	6.77
	Equal variances not assumed			3.624	254.703	.000	6.77
EDUC	Equal variances assumed	.050	.823	-6.507	267	.000	-2.28
	Equal variances not assumed			-6.708	200.606	.000	-2.28

Chi-square tests comparing missing and valid cases - 1

32

We use chi-square tests of independence to test for differences in the breakdown between the missing and valid groups for the nonmetric variables in the analysis.

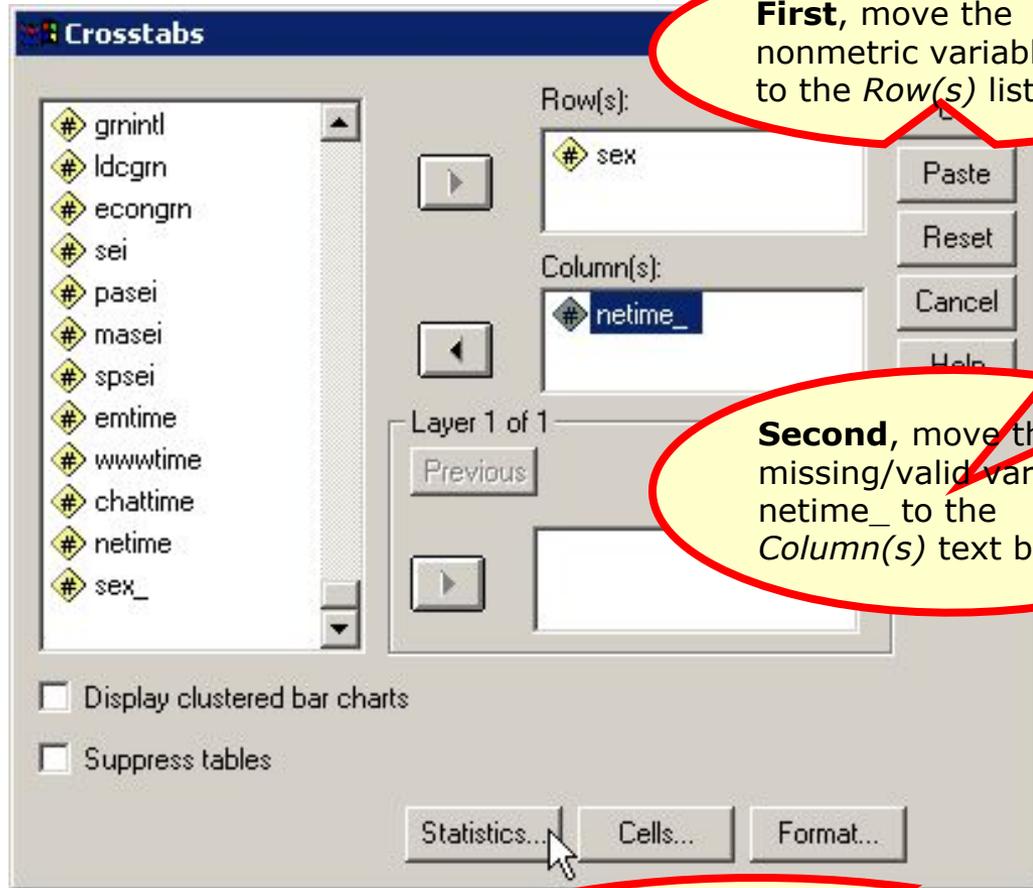
The screenshot shows the SPSS Data Editor window. The 'Analyze' menu is open, and the 'Descriptive Statistics' sub-menu is also open, with 'Crosstabs...' selected. The background shows a data grid with columns 'estg80' and 'marital'.

	estg80	marital
3	20000020	
4	20000029	
5	20000032	
6	20000034	
7	20000043	
8	20000060	
9	20000070	
10	20000072	5
11	20000079	1
12	20000097	1
13	20000117	1
14	20000126	1

First, select the *Descriptive Statistics | Crosstabs...* command from the *Analyze* menu.

Chi-square tests comparing missing and valid cases - 2

33



First, move the nonmetric variable sex to the Row(s) list box.

Second, move the missing/valid variable, netime_ to the Column(s) text box.

Third, click on the Statistics... button to specify the chi-square test.

$$H_1: \mu < 0$$

$$W = \sum_{i=1}^n w_i x_i (g-1) = \sum_{i=1}^n w_i x_i$$

$$H_0: \mu = 0$$

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

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

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

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

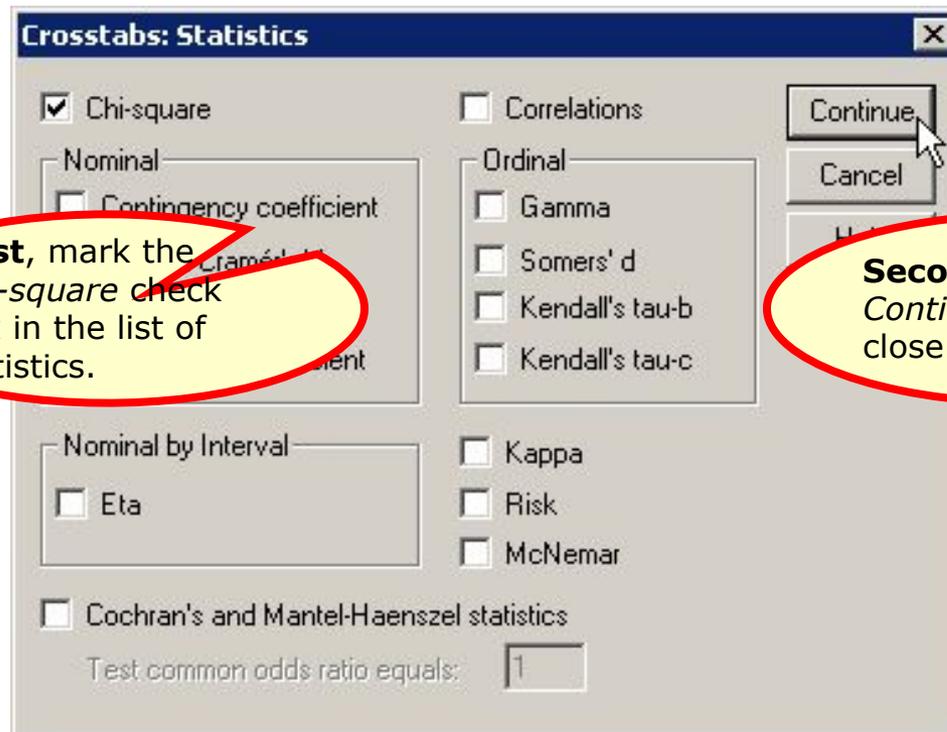
$$y = \frac{1}{2} (x_j + x_{j+1})$$

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

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

Chi-square tests comparing missing and valid cases - 3

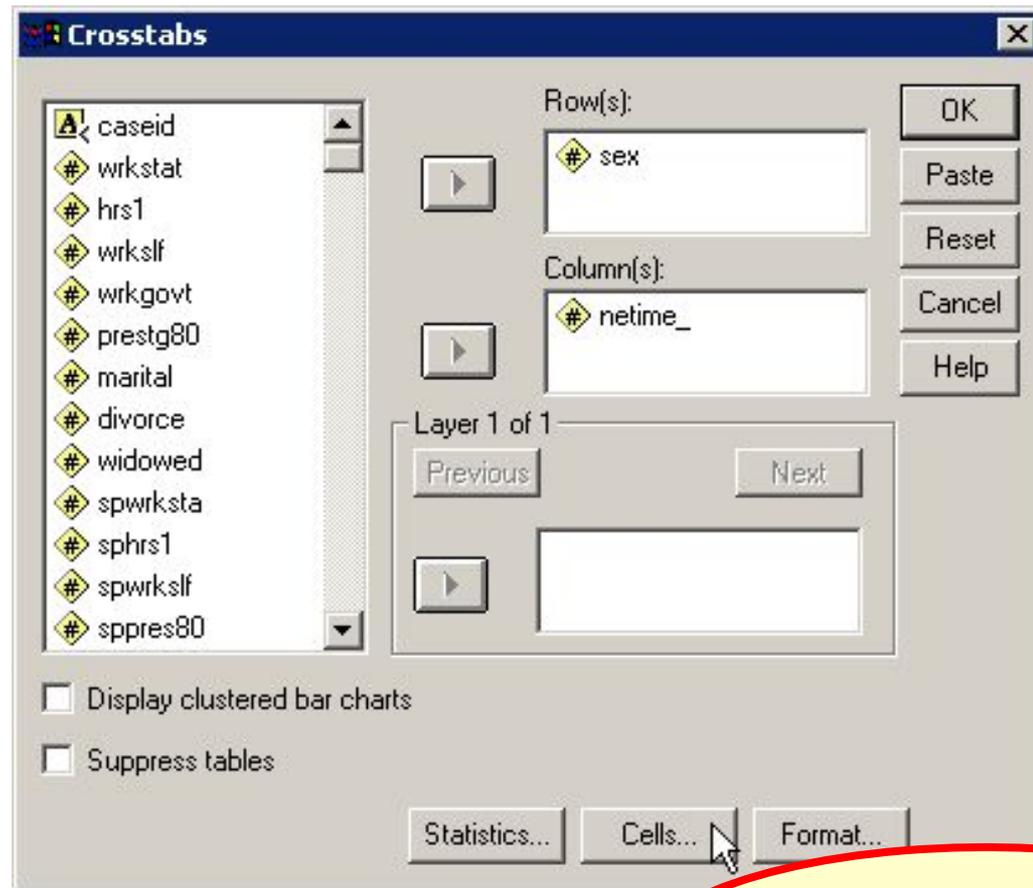
34



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

Chi-square tests comparing missing and valid cases - 4

35



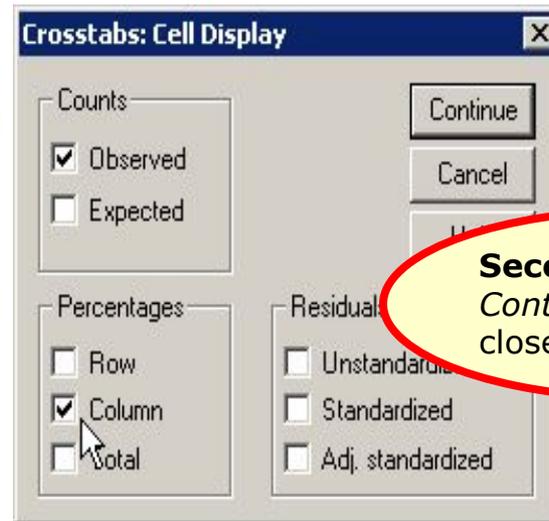
Click on the *Cells..* button to request that column percentages be included in the cross tabulated table.

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

Chi-square tests comparing missing and valid cases - 5

36

First, mark the *Column* check box in the *Percentages* panel.

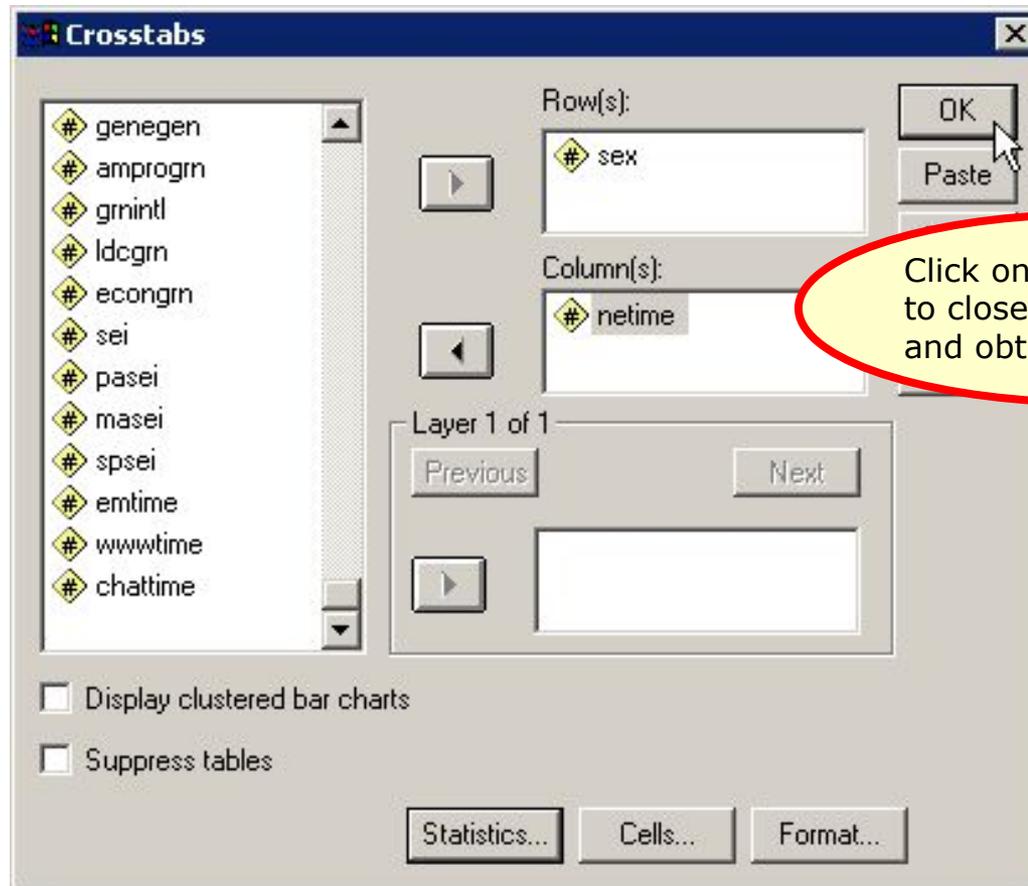


Second, click on the *Continue* button to close the dialog box.

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

Chi-square tests comparing missing and valid cases - 6

37



Click on the OK button to close the dialog box and obtain the output.

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

Output for the chi-square test

RESPONDENTS SEX * Valid/Missing TOTAL TIME SPENT ON THE INTERNET
Crosstabulation

			Valid/Missing TOTAL TIME SPENT ON THE INTERNET		Total
			0	1	
RESPONDENTS SEX	1	Count	73	38	111
		% within Valid/Missing TOTAL TIME SPENT ON THE INTERNET	41.2%	40.9%	41.1%
	2	Count	104	55	159
		% within Valid/Missing TOTAL TIME SPENT ON THE INTERNET	58.8%	40.9%	
Total		Count			
		% within Valid/Missing TOTAL TIME SPENT ON THE INTERNET			

On the chi-square test, the difference in the breakdown for the missing cases is not statistically different from the breakdown for the valid cases.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.004 ^b	1	.952		
Continuity Correction ^a	.000	1	1.000		
Likelihood Ratio	.004	1	.952		
Fisher's Exact Test				1.000	.529
Linear-by-Linear Association	.004	1	.952		
N of Valid Cases	270				

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

Answer 1

39

In the dataset GSS2000R, is the following statement true, false, or an incorrect application of a statistic? Use a level of significance of 0.01 for evaluating missing data and assumptions.

In pre-screening the data for use in a multiple regression of the dependent variable "total hours spent on the Internet" [netime] with the independent variables "age" [age], "highest year of school completed" [educ], and "sex" [sex], the missing data analysis did not indicate any need for caution or further analysis for a problematic pattern of missing data.

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

Since there were significant differences in the statistical tests comparing cases with missing data to cases with valid data, a caution was added to the interpretation of any findings, pending further analysis of the missing data pattern.

The answer to the question is false.

Using scripts

40

- 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

41

- The script “EvaluatingAssumptionsAndMissingData.exe” will produce all of the output we have used for evaluating missing data (as well as output for testing assumptions).
- Navigate to the link “SPSS Scripts and Syntax” on the course web page.
- Download the script file “EvaluatingAssumptionsAndMissingData.exe” to your computer and install it, following the directions on the web page.

Open the data set in SPSS

42

The screenshot shows the SPSS Data Editor window titled "GSS2000R.sav - SPSS Data Editor". The window contains a menu bar (File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Window, Help) and a toolbar with various icons. The main area displays a data table with columns: caseid, wrkstat, hrs1, wrk, and others. The first row is selected, showing caseid 20000009. A yellow callout box with a red border is overlaid on the table, containing the text: "Before using a script, a data set should be open in the SPSS data editor." The status bar at the bottom indicates "SPSS Processor is ready".

	caseid	wrkstat	hrs1	wrk	...		
1	20000009	1	50				
2	20000012	1	40				
3	20000020	6	.				
4	20000029	5	.		3		
5	20000032	1	40	2	1		
6	20000034	1	60	2	2	55	5
7	20000043	4	.	2	2	36	3
8	20000060	1	38	2	2	29	5
9	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

Invoke the script

43

The screenshot shows the SPSS Data Editor window for 'GSS2000R.sav'. The Utilities menu is open, and the 'Run Script...' option is highlighted. A yellow callout box with a red border points to this option, containing the text: 'To invoke the script, select the Run Script... command in the Utilities menu.'

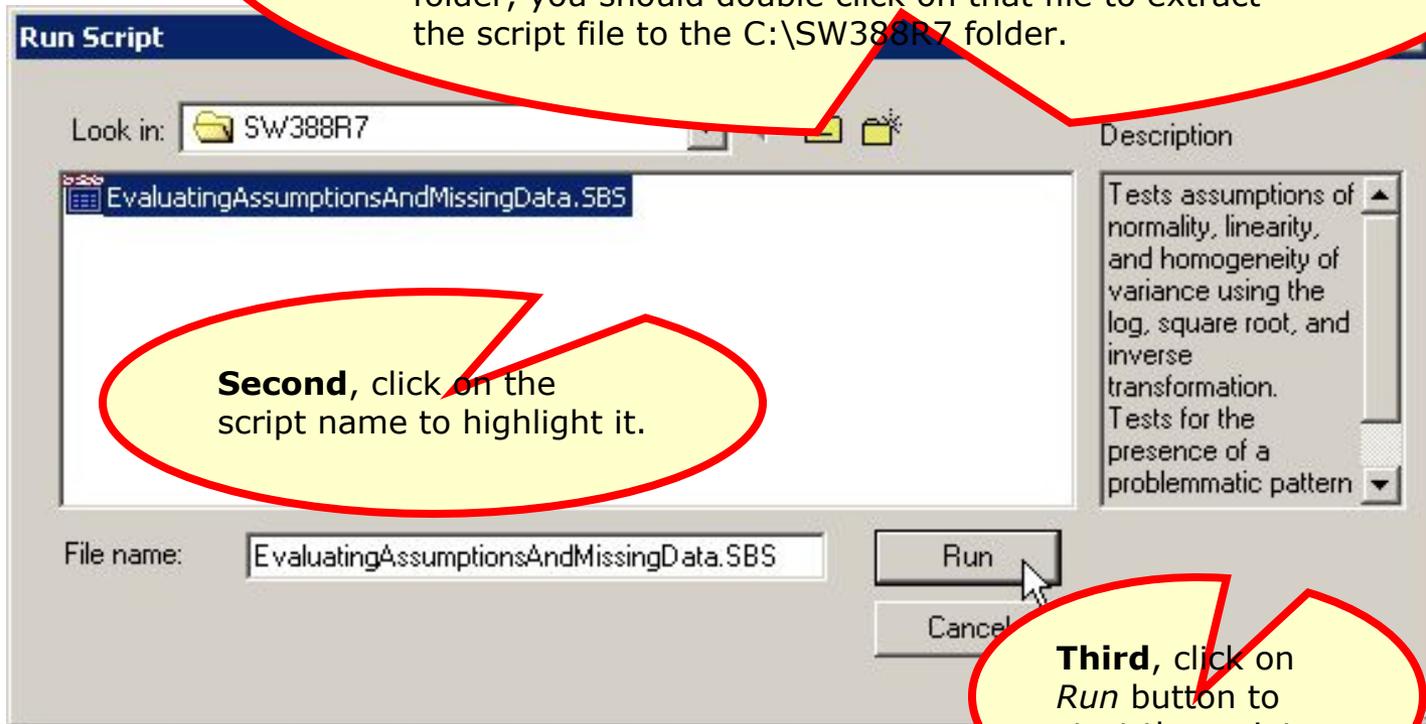
	caseid	wrkstat	hrs1		prestg80	marital
1	20000009	1	50			
2	20000012	1	40	2	51	1
3	20000020	6	.	1	74	1
4	20000029	5	.	2	40	3
5	20000032	1	40	2	.	1
6	20000034	1	60	2	55	.
7	20000043	4	.	2	.	.
8	20000060	1	38	2	.	.
9	20000070	7	.	2	.	.
10	20000072	5	.	2	.	.
11	20000079	1	40	9	64	1
12	20000097	1	40	2	35	1
13	20000117	1	49	2	51	3
14	20000126	1	40	2	33	3

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

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.

$$H_1: \mu < 0$$

$$W = \sum_{i=1}^n w_i \frac{x_i(\theta-1)}{1+x_i\theta} = \beta$$

$$H_0: \mu = 0$$

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

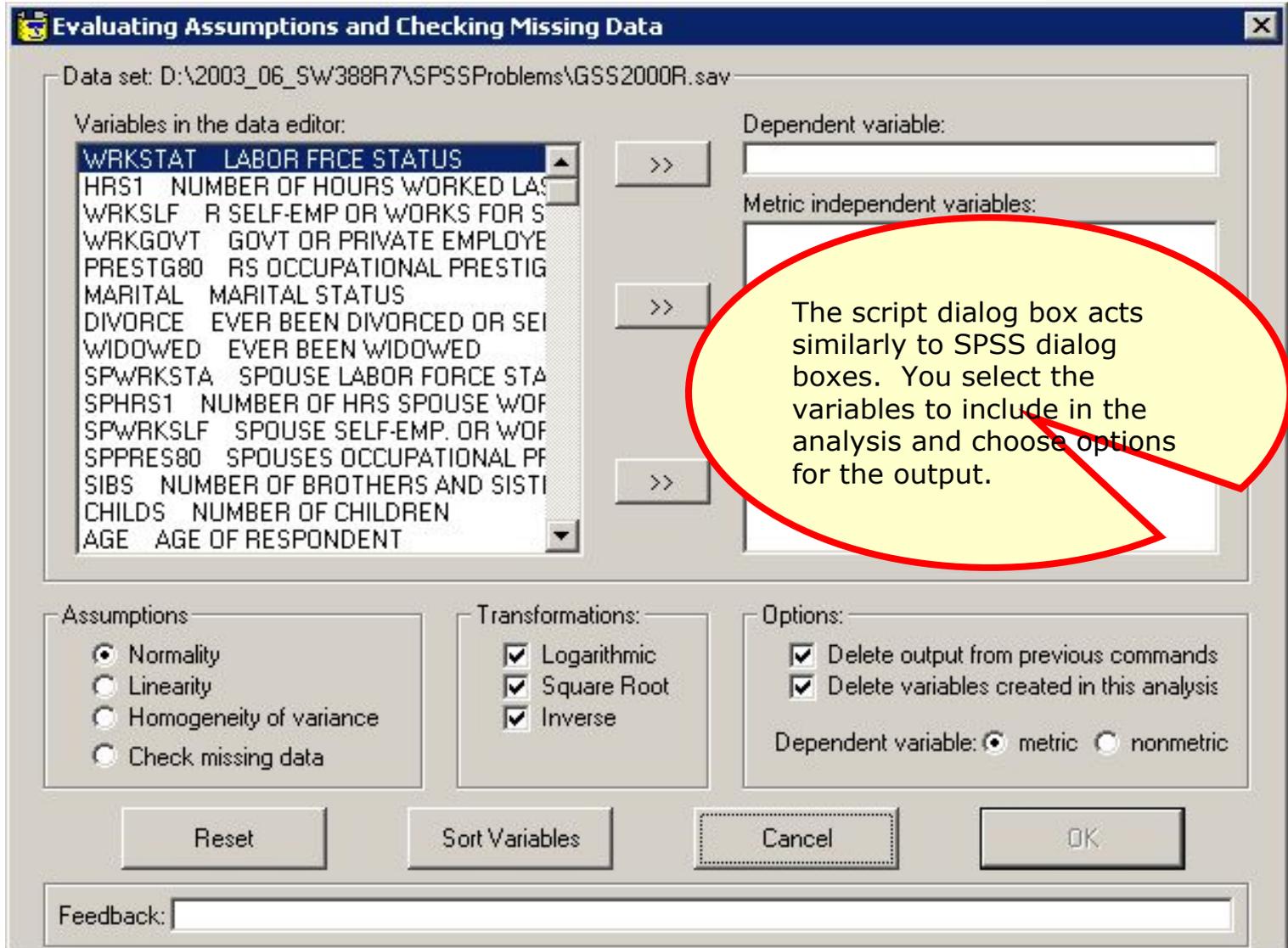
$$\mu = E(x) = \int_0^1 x f(x) dx$$

$$\sigma^2 = E(x - \mu)^2 = \int_0^1 (x - \mu)^2 f(x) dx$$

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

The script dialog

45



Evaluating Assumptions and Checking Missing Data

Data set: D:\2003_06_SW388R7\SPSSProblems\GSS2000R.sav

Variables in the data editor:

- WRKSTAT LABOR FRCE STATUS
- HRS1 NUMBER OF HOURS WORKED LAST WEEK
- WRKSLF R SELF-EMP OR WORKS FOR SELF
- WRKGOVT GOVT OR PRIVATE EMPLOYER
- PRESTG80 RS OCCUPATIONAL PRESTIGE
- MARITAL MARITAL STATUS
- DIVORCE EVER BEEN DIVORCED OR SEPARATED
- WIDOWED EVER BEEN WIDOWED
- SPWRKSTA SPOUSE LABOR FORCE STATUS
- SPHRS1 NUMBER OF HRS SPOUSE WORKS LAST WEEK
- SPWRKSLF SPOUSE SELF-EMP. OR WORKS FOR SELF
- SPPRES80 SPOUSES OCCUPATIONAL PRESTIGE
- SIBS NUMBER OF BROTHERS AND SISTERS
- CHILDS NUMBER OF CHILDREN
- AGE AGE OF RESPONDENT

Dependent variable:

Metric independent variables:

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.

Assumptions:

- Normality
- Linearity
- Homogeneity of variance
- Check missing data

Transformations:

- Logarithmic
- Square Root
- Inverse

Options:

- Delete output from previous commands
- Delete variables created in this analysis

Dependent variable: metric nonmetric

Reset Sort Variables Cancel OK

Feedback:

Complete the specifications - 1

46

Evaluating Assumptions and Checking Missing Data

Data set: D:\2003_06_SW388R7\SPSSProblems\GSS2000R.sav

Variables in the data editor:

DOWNBLUE FELT DOWN AND BLUE IN F
SDCACTS PHYSICAL AND EMOTION SOC
GRNEXAGG ENVIRONMENTAL THREATE
GENEGEN HOW DANGEROUS MODIFYIN
AMPROGRM AMERICAN DOING ENOUGH
ENVIF

Dependent variable (DV):
NETIME TOTAL TIME SPENT ON THE INT

Metric independent variables (IV):
AGE AGE OF RESPONDENT
EDUC HIGHEST YEAR OF SCHOOL COMF

Nonmetric independent variables (IV):
SEX RESPONDENTS SEX

Assumptions:

- Normality
- Linearity
- Homogeneity of variance
- Check missing data

Transformations:

- Logarithmic
- Square root
- Inverse

Dependent variable: metric nonmetric

Delete variables created in this analysis

Reset Sort Variables Cancel OK

Feedback: _____

Move the the dependent and independent variables from the list of variables to the list boxes. Metric and nonmetric variables are moved to separate lists so the computer knows how you want them treated.

You must also indicate the level of measurement for the dependent variable. In this case, the metric option button is marked.

Complete the specifications - 2

47

Evaluating Assumptions and Checking Missing Data

Data set: D:\2003_06_SW388R7\SPSSProblems\GSS2000R.sav

Variables in the data editor:

- DOWNBLUE FELT DOWN AND BLUE IN F
- SOCFACTS PHYSICAL AND EMOTION SOC
- GRNEXAGG ENVIRONMENTAL THREATE
- GENEGEN HOW DANGEROUS MODIFYN
- AMPROGRN AMERICAN DOING ENOUGH
- GRNINTL INTL AGREEMENTS FOR ENVIF
- POOR COUNTRIES LESS THAN
- ECONOMIC PROGRESS DEPE
- IDENT'S SOCIOECONOMIC IN
- ER'S SOCIOECONOMIC INDE:
- HER'S SOCIOECONOMIC INDI
- OUSE'S SOCIOECONOMIC INDE
- TIME SPENT USING E-MAIL
- WWWTIME TIME SPENT ON THE WWW
- CHATTIME TIME SPENT ON CHAT

Dependent variable (DV):

- NETIME TOTAL TIME SPENT ON THE INT

Metric independent variables (IV):

- AGE AGE OF RESPONDENT
- EDUC HIGHEST YEAR OF SCHOOL COMF

Nonmetric independent variables (IV):

- SEX RESPONDENTS SEX

Assumptions:

- Normality
- Linearity
- Homogeneity of variance
- Check missing data

Transformations:

Options:

- Delete output from previous commands
- Tables created in this analysis
- metric nonmetric

Reset Sort Variables Cancel **OK**

Feedback: _____

Mark the option button for the type of output you want the script to compute.

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.

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

Output from the script - 1

49

Number of Valid and Missing Cases per Variable

		Statistics			
		TOTAL TIME SPENT ON THE INTERNET	AGE OF RESPONDENT	HIGHEST YEAR OF SCHOOL COMPLETED	RESPONDENTS SEX
N	Valid	93	270	269	270
	Missing	177	0	1	0

T-Test for AGE OF RESPONDENT HI by Valid/Missing TOTAL TIME SP

	Valid/Missing	TOTAL TIME SPENT ON	
		0	1
AGE OF RESPONDENT	0		
	1		
HIGHEST YEAR OF SCHOOL COMPLETED	0		
	1		

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

Scroll through the script to locate the outputs needed to answer the question.

Complete the specifications - 2

50

The script dialog box does not close automatically because we often want to run another test right away. There are two methods for closing the dialog box.

Click on the X close box to close the script.

Click on the *Cancel* button to close the script.

Click on the *OK* button to close the script.

Assumptions:
 Normality
 Linearity
 Homogeneity of variance
 Check missing data

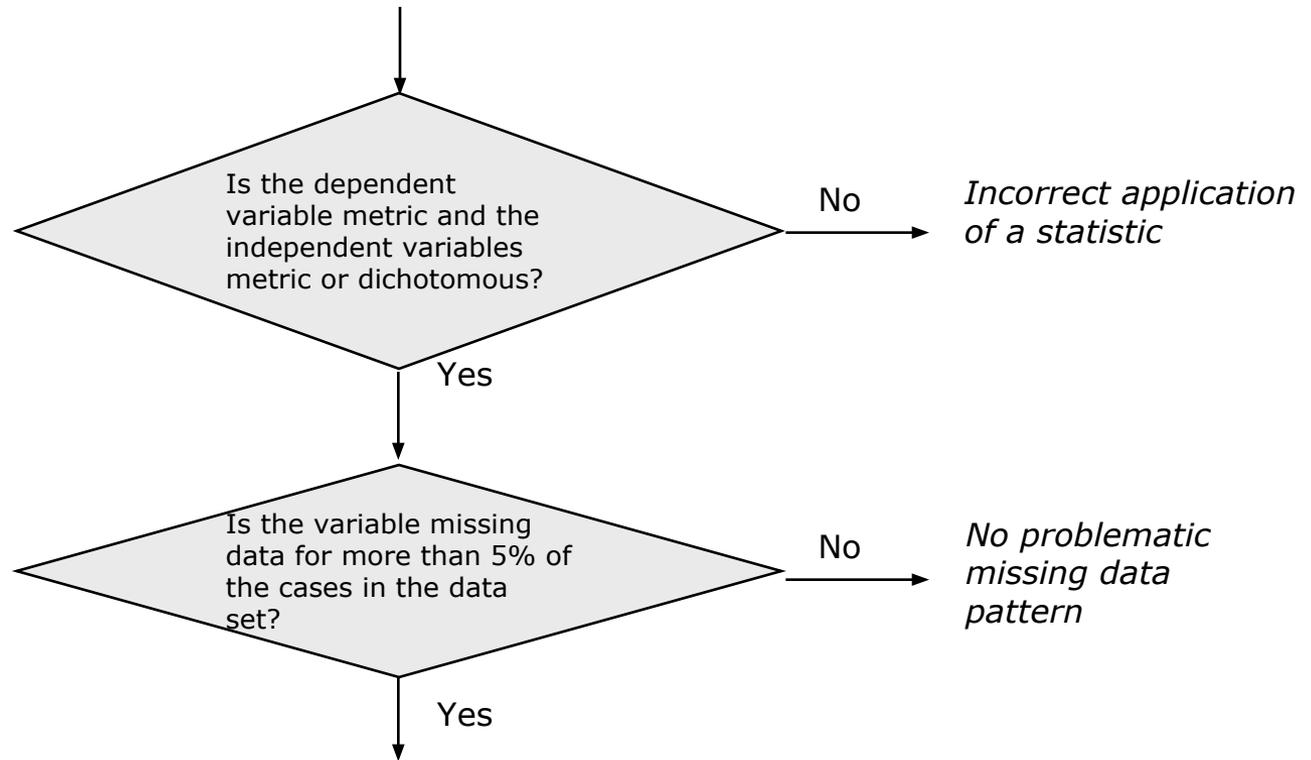
Transformations:
 None
 Logarithmic

Options:
 Delete output from previous commands
 Delete variables created in this analysis
Dependent variable: metric nonmetric

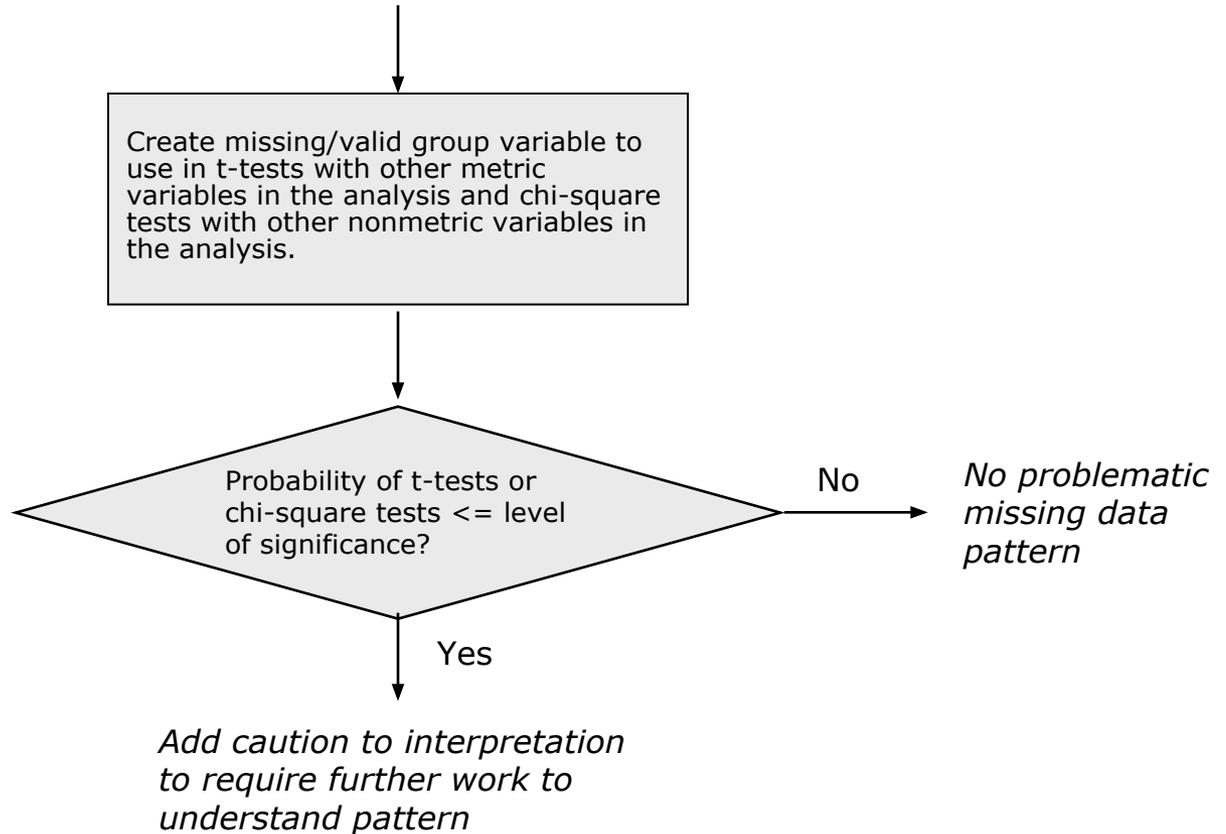
Steps in analyzing missing data

51

The following is a guide to the decision process for answering problems about problematic patterns of missing data:



Steps in analyzing missing data



52

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