Introduction to SPSS

Mr. Kongmany Chaleunvong

GFMER - WHO - UNFPA - LAO PDR
Training Course in Reproductive Health Research
Vientiane, 22 October 2009

Object of the Course

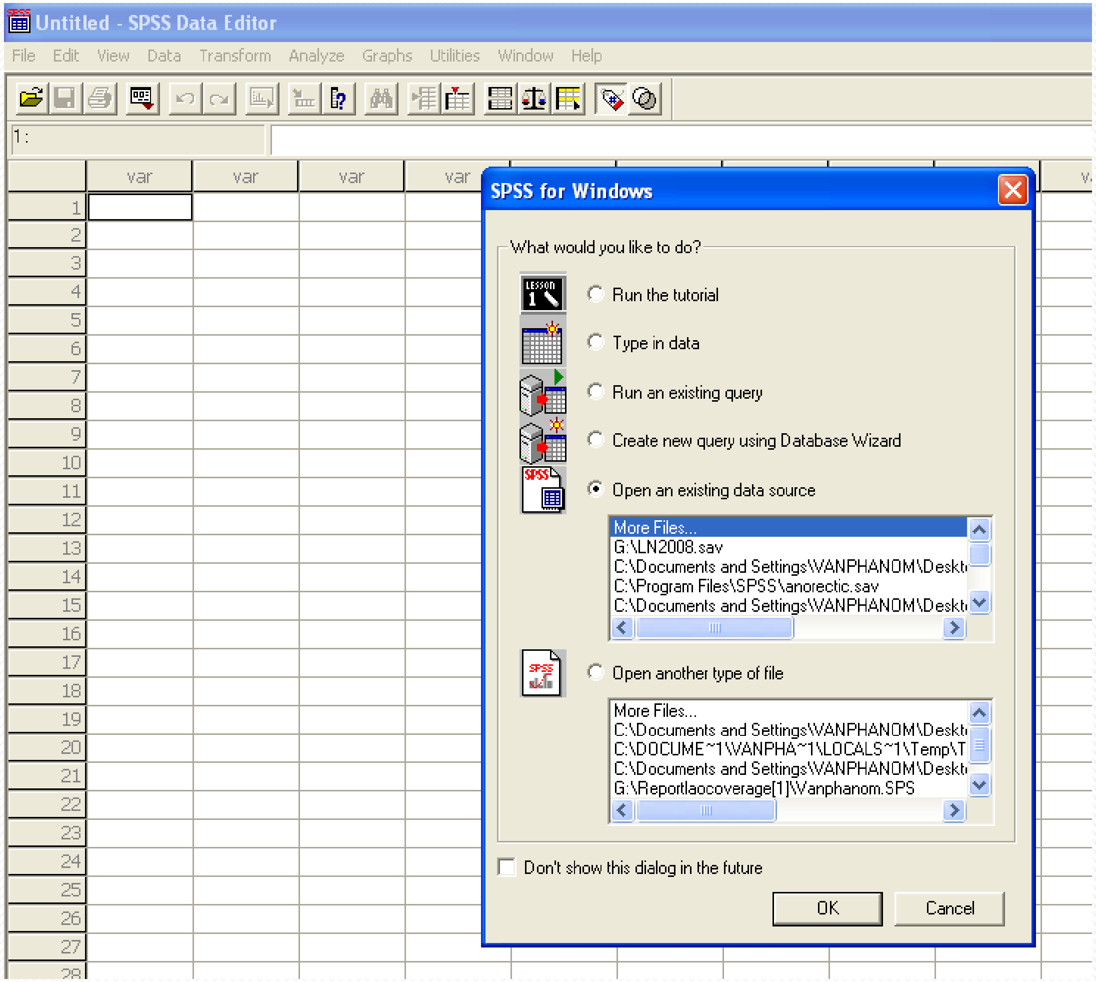
- Introduction to SPSS
- The basics of managing data files

Introduction: What is SPSS?

- SPSS is a statistical package for beginning, intermediate, and advanced data analysis
- Originally it is an acronym of Statistical Package for the Social Science but now it stands for Statistical Product and Service Solutions
- One of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions

Starting SPSS for Windows

Launch SPSS either by double-clicking the SPSS icon on the desktop, or from the Start menu –SPSS will have a group under programs. The opening screen should appear as



Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 :
var var var var
SPSS for Windows
OK
Cancel

What would you like to do?

○ Run the tutorial

○ Type in data

○ Run an existing query

○ Create new query using Database Wizard

● Open an existing data source

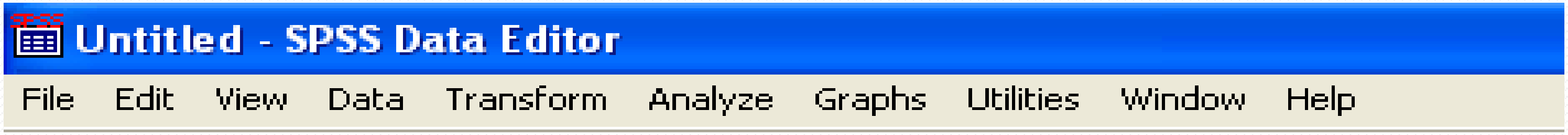
More Files...
G:\LN2008.sav
C:\Documents and Settings\VANPHANOM\Deskt
C:\Program Files\SPSS\anorectic.sav
C:\Documents and Settings\VANPHANOM\Deskt

○ Open another type of file

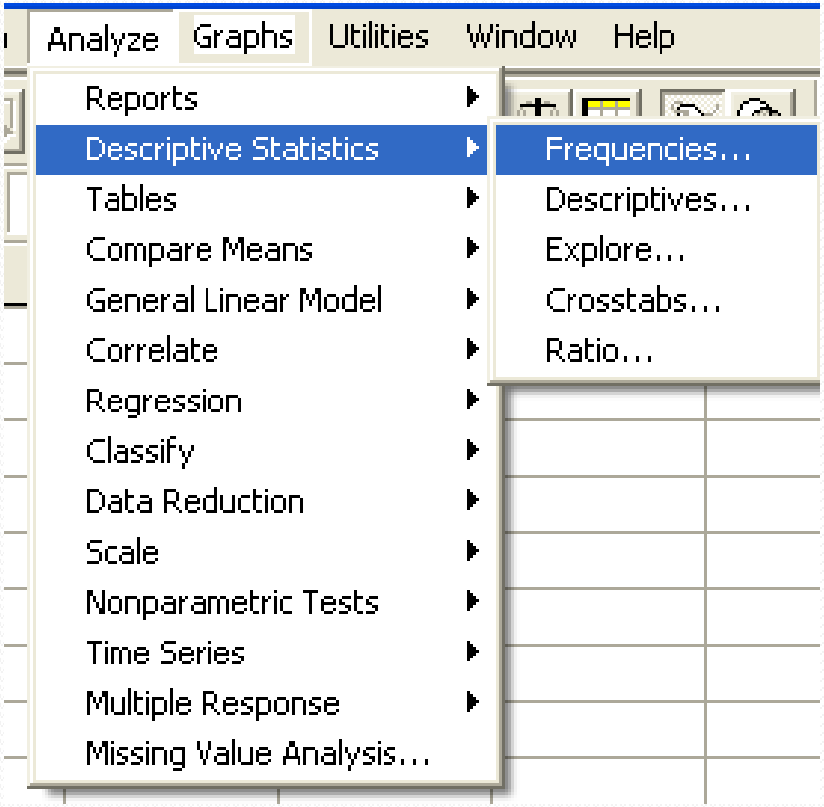
More Files...
C:\Documents and Settings\VANPHANOM\Deskt
C:\DOCUME~1\VANPHA~1\LOCALS~1\Temp\T
C:\Documents and Settings\VANPHANOM\Deskt
G:\Reportlaocoverage[1]\Vanphanom.SPS

□ Don't show this dialog in the future

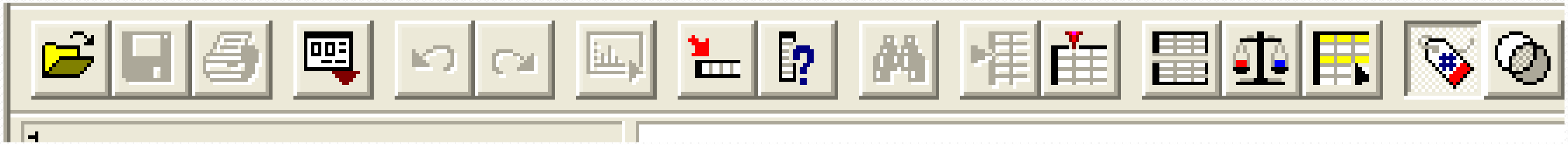
The Menu bar



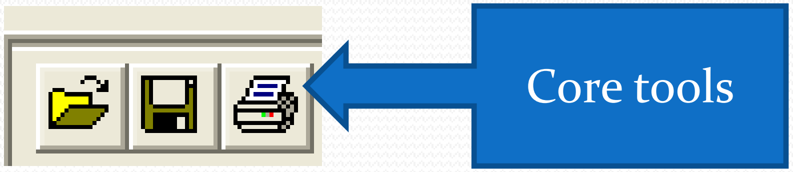
The Menu bar lists 10 pull down menu, grouping the available SPSS commands. Some of these have sub-menus, the Analyze menu is like this



The Toolbar

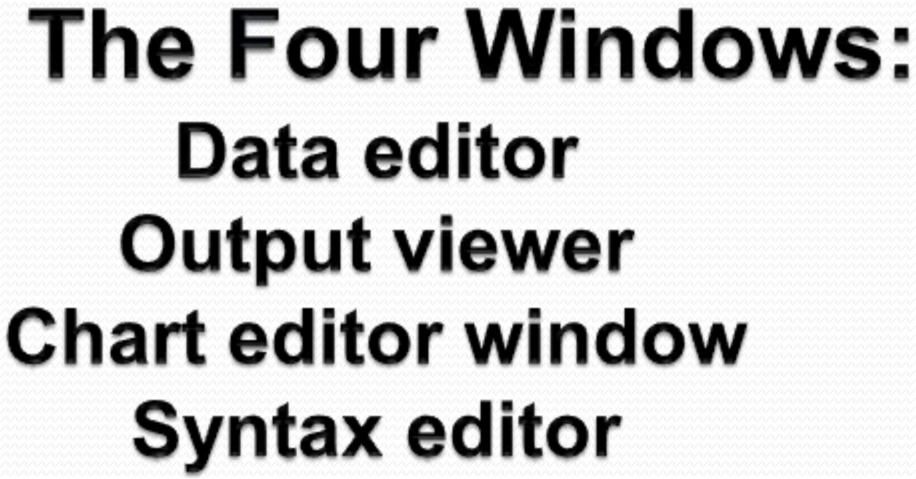


The toolbar, located just below the menu bar, provides quick and easy access to many frequently used facilities



- Open File: Displays the Open File dialog box for the type of windows that is active.
- Save File: Saves the working file, if the file has no name, it displays the Save File dialog box for the type of document that is active.
- Print : Displays the print dialog box.

About the four-windows in SPSS



The Four Windows:

Data editor

Output viewer

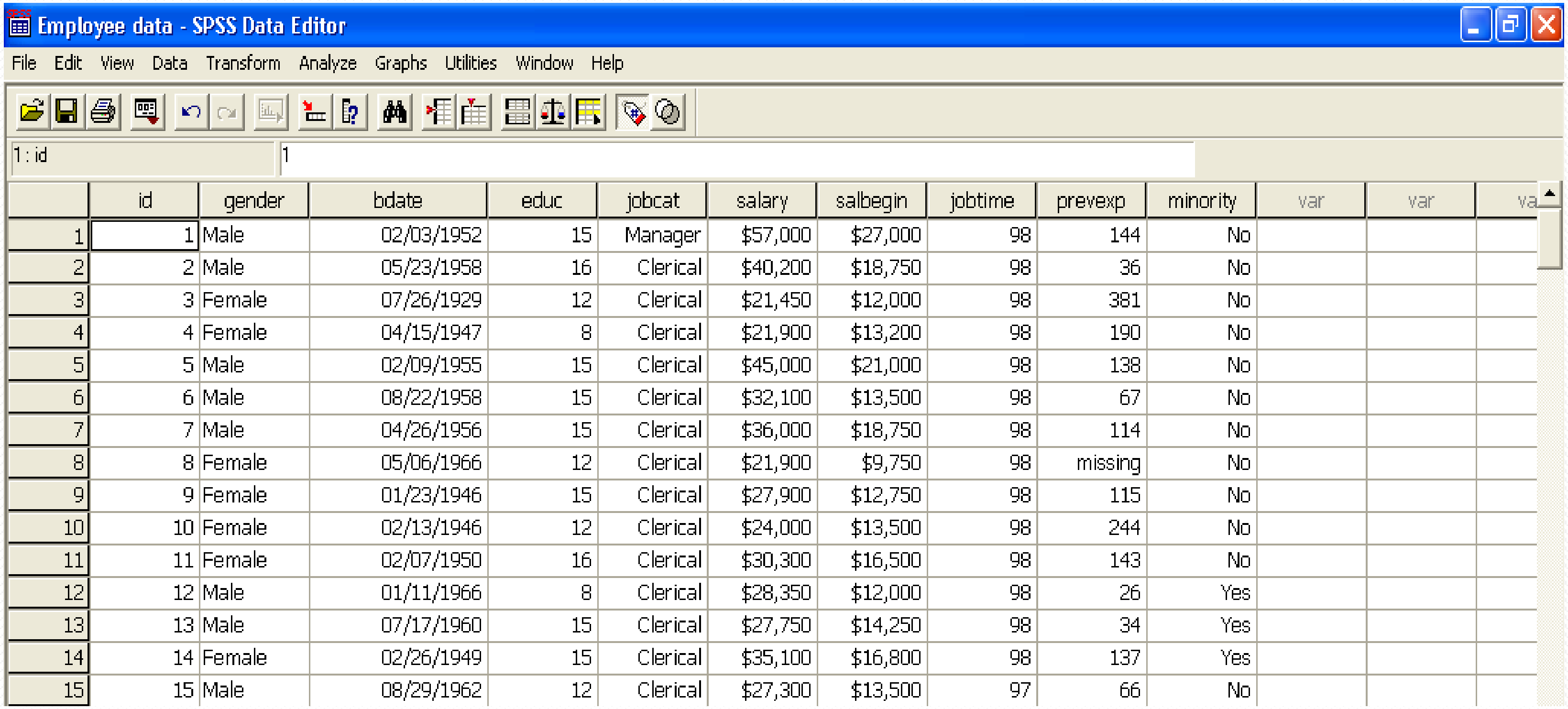
Chart editor window

Syntax editor

The Four Windows: Data Editor

- Data Editor

  Spreadsheet-like system for defining, entering, editing, and displaying data. Extension of the saved file will be “sav.”


Employee data - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help

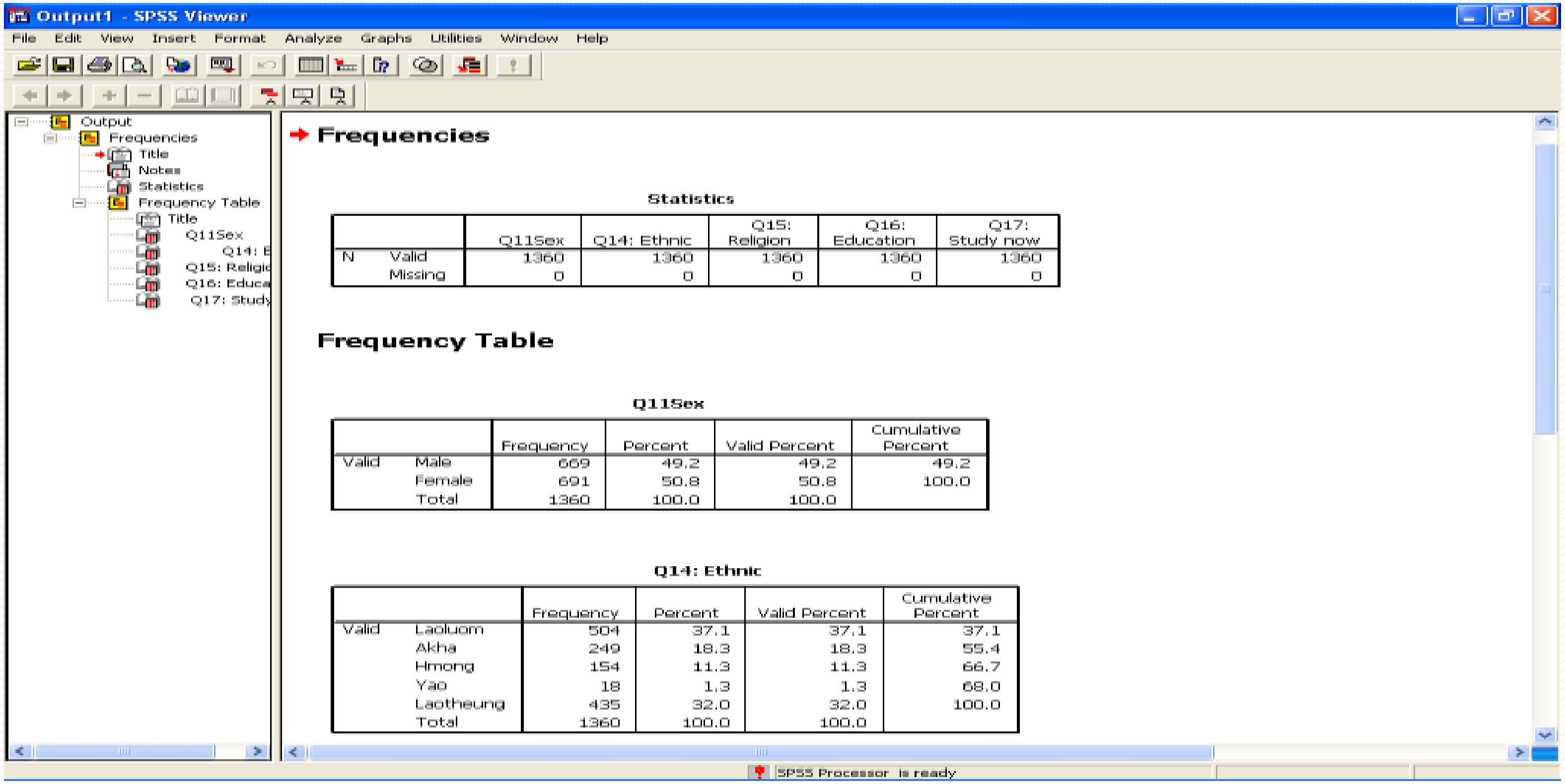
1 : id 1

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var	var	va
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98	144	No			
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98	36	No			
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98	381	No			
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98	190	No			
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98	138	No			
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98	67	No			
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98	114	No			
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98	missing	No			
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98	115	No			
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98	244	No			
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98	143	No			
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98	26	Yes			
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98	34	Yes			
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98	137	Yes			
15	15	Male	08/29/1962	12	Clerical	$27,300	$13,500	97	66	No			

The Four Windows: Output Viewer

- Output Viewer

  Displays output and errors. Extension of the saved file will be “spo.”



Frequencies

Statistics

		Q11Sex	Q14: Ethnic	Q15: Religion	Q16: Education	Q17: Study now
N	Valid	1360	1360	1360	1360	1360
	Missing	0	0	0	0	0

Frequency Table

Q11Sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	669	49.2	49.2	49.2
	Female	691	50.8	50.8	100.0
	Total	1360	100.0	100.0	

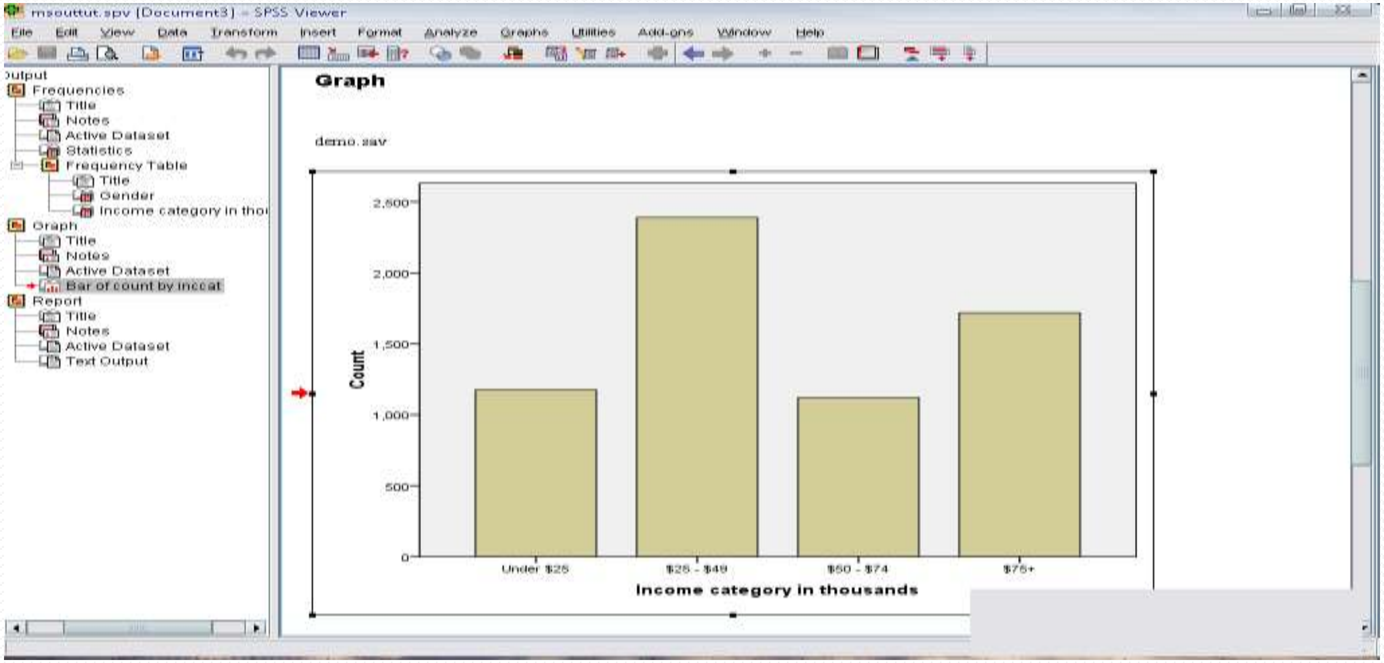
Q14: Ethnic

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Laoluom	504	37.1	37.1	37.1
	Akha	249	18.3	18.3	55.4
	Hmong	154	11.3	11.3	66.7
	Yao	18	1.3	1.3	68.0
	Laotheung	435	32.0	32.0	100.0
	Total	1360	100.0	100.0	

The Four Windows: Chart editor window

- Output Viewer

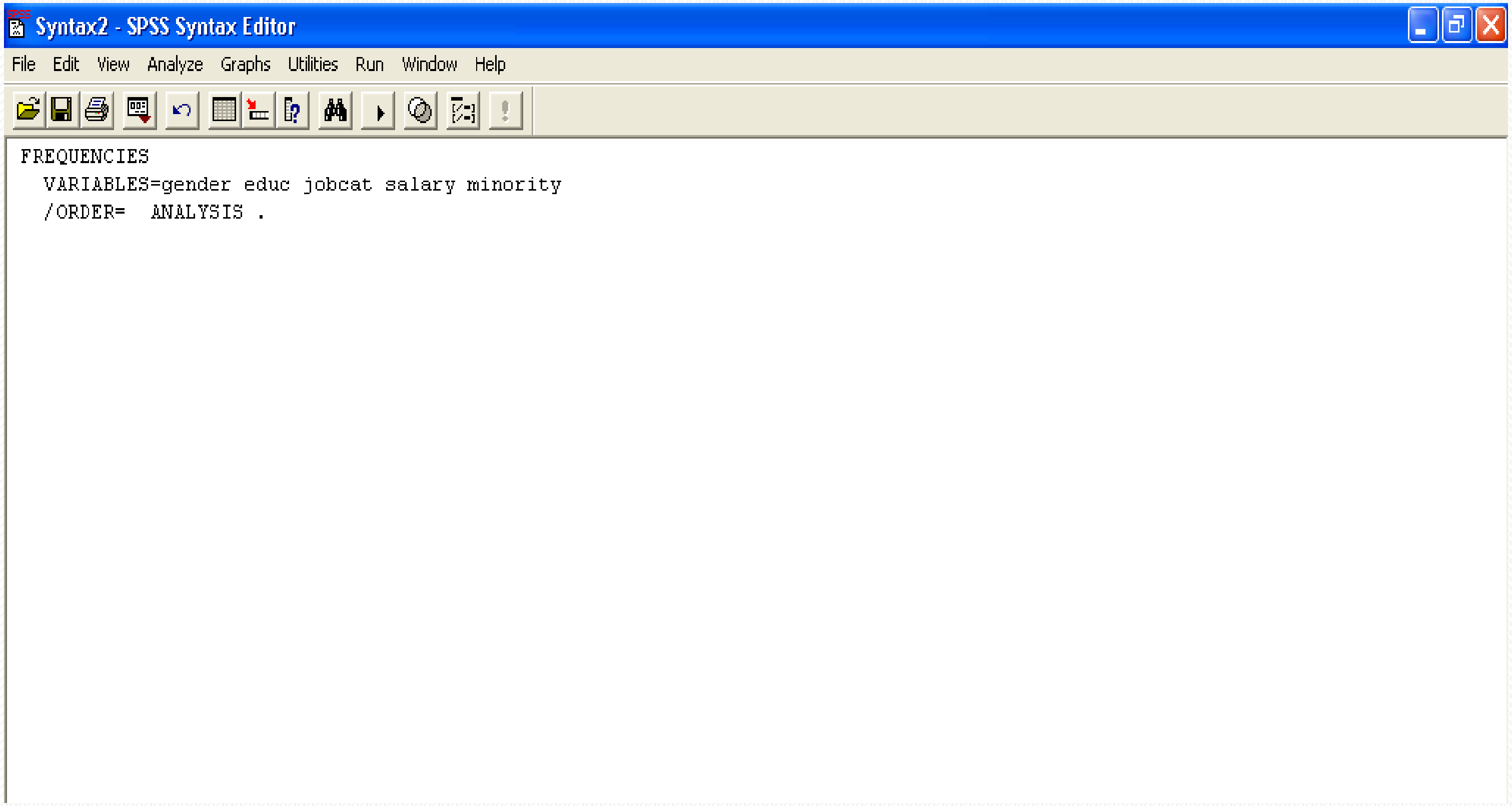
  Displays output and errors. Extension of the saved file will be “spo.”



The Four Windows: Syntax editor

- Syntax Editor

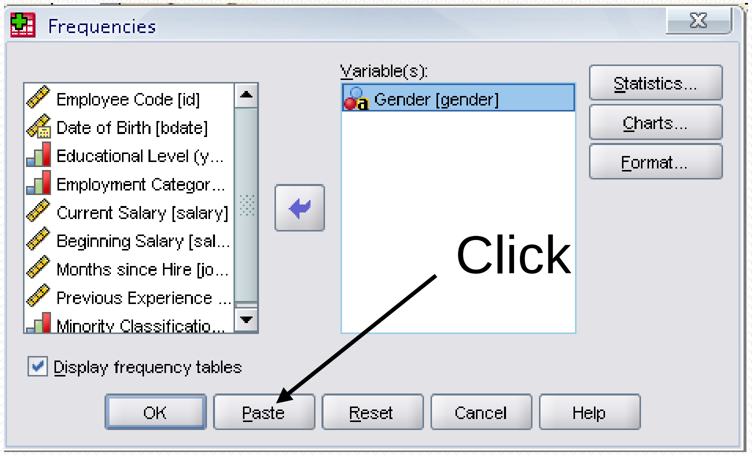
  Text editor for syntax composition. Extension of the saved file will be “sps.”


Syntax2 - SPSS Syntax Editor
File Edit View Analyze Graphs Utilities Run Window Help

```
FREQUENCIES
  VARIABLES=gender educ jobcat salary minority
  /ORDER=  ANALYSIS .
```

Using the Syntax editor

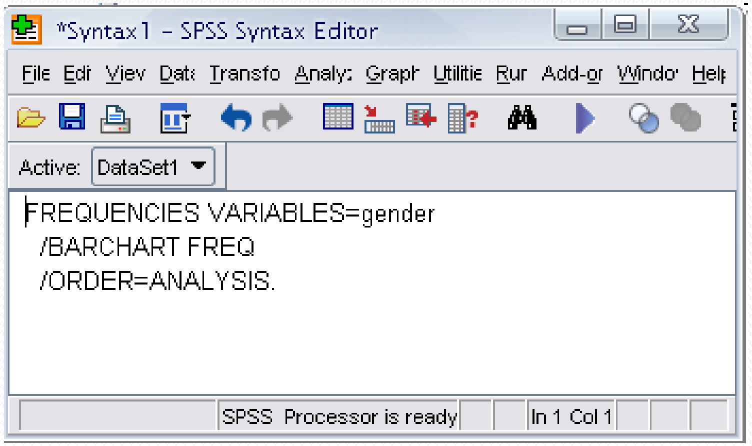
- Click ‘Analyze,’ ‘Descriptive statistics,’ then click ‘Frequencies.’
- Put ‘Gender’ in the Variable(s) box.
- Then click ‘Charts,’ ‘Bar charts,’ and click ‘Continue.’
- Click ‘Paste.’



Variable(s): Gender [gender]

☑ Display frequency tables

Click


*Syntax1 - SPSS Syntax Editor
File Edi Viev Dat: Transfo Analy: Grapt Utilitie Rur Add-or Windo Hel
SPSS Processor is ready
In 1 Col 1

Active: DataSet1

```
FREQUENCIES VARIABLES=gender
  /BARCHART FREQ
  /ORDER=ANALYSIS.
```

The basics of managing data files

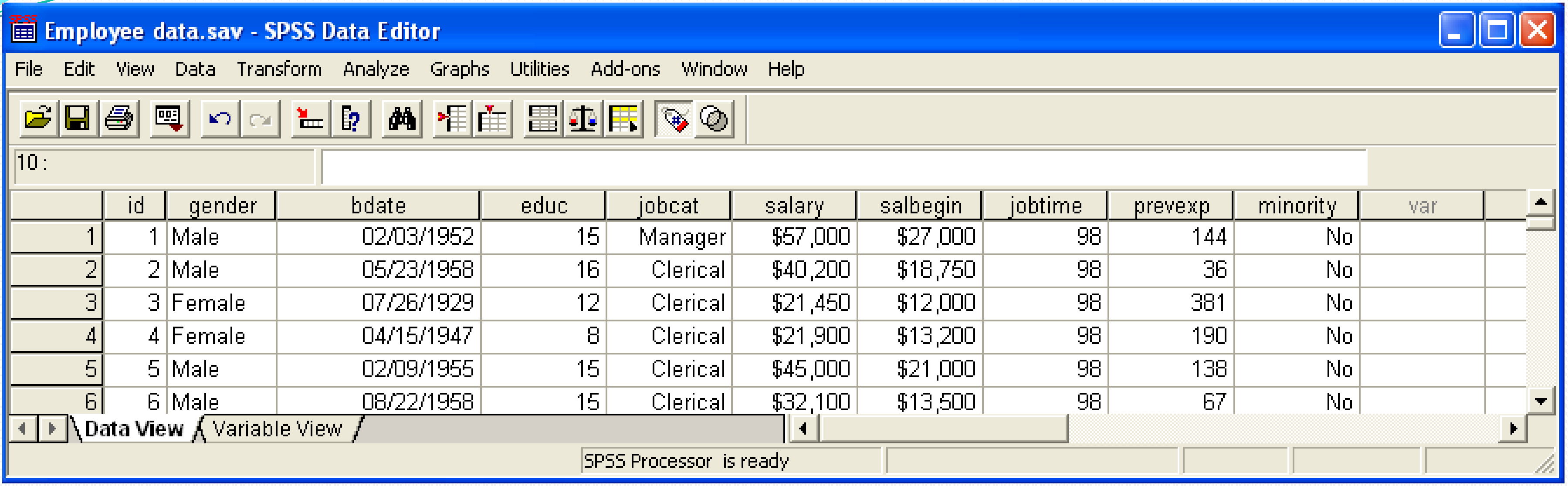
Data Entry & Coding

- Before describing the process for defining variables, an important distinction should be made between two terms that are often confused: *variable* and *value*
- A variable is a measure or classification scheme that can have several values
- Values are the numbers or categorical classification representing individual instances of the variable being measured

Data Entry

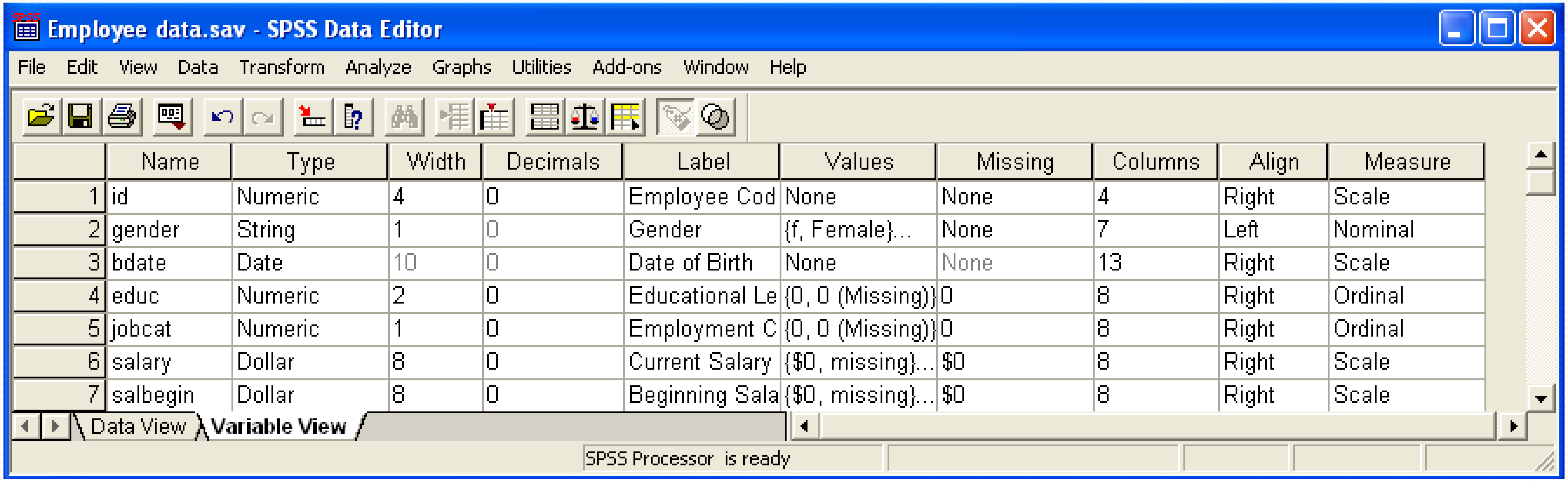
- You may create a data file using one of your favorite text editors, or word processing packages (e.g., Word Perfect, MS-Word). Files created using word processing software should be saved in text format before trying to read them into an SPSS session.
- You may enter your data into a spreadsheet (e.g., Lotus 123, Excel, dBASE) and read it directly into SPSS for Windows.
- Finally, you may enter the data directly into the spreadsheet-like Data Editor of SPSS for Windows.
  - In this document we are going to examine one data entry methods: using the Data Editor of SPSS for Windows.

The Data View


Employee data.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
10 :
Data View
Variable View
SPSS Processor is ready

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98	144	No	
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98	36	No	
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98	381	No	
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98	190	No	
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98	138	No	
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98	67	No	

The Variable View


Employee data.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Data View
Variable View
SPSS Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	4	0	Employee Cod	None	None	4	Right	Scale
2	gender	String	1	0	Gender	{f, Female}...	None	7	Left	Nominal
3	bdate	Date	10	0	Date of Birth	None	None	13	Right	Scale
4	educ	Numeric	2	0	Educational Le	{0, 0 (Missing)}	0	8	Right	Ordinal
5	jobcat	Numeric	1	0	Employment C	{0, 0 (Missing)}	0	8	Right	Ordinal
6	salary	Dollar	8	0	Current Salary	{$0, missing}...	$0	8	Right	Scale
7	salbegin	Dollar	8	0	Beginning Sala	{$0, missing}...	$0	8	Right	Scale

Define Information – The Variable View

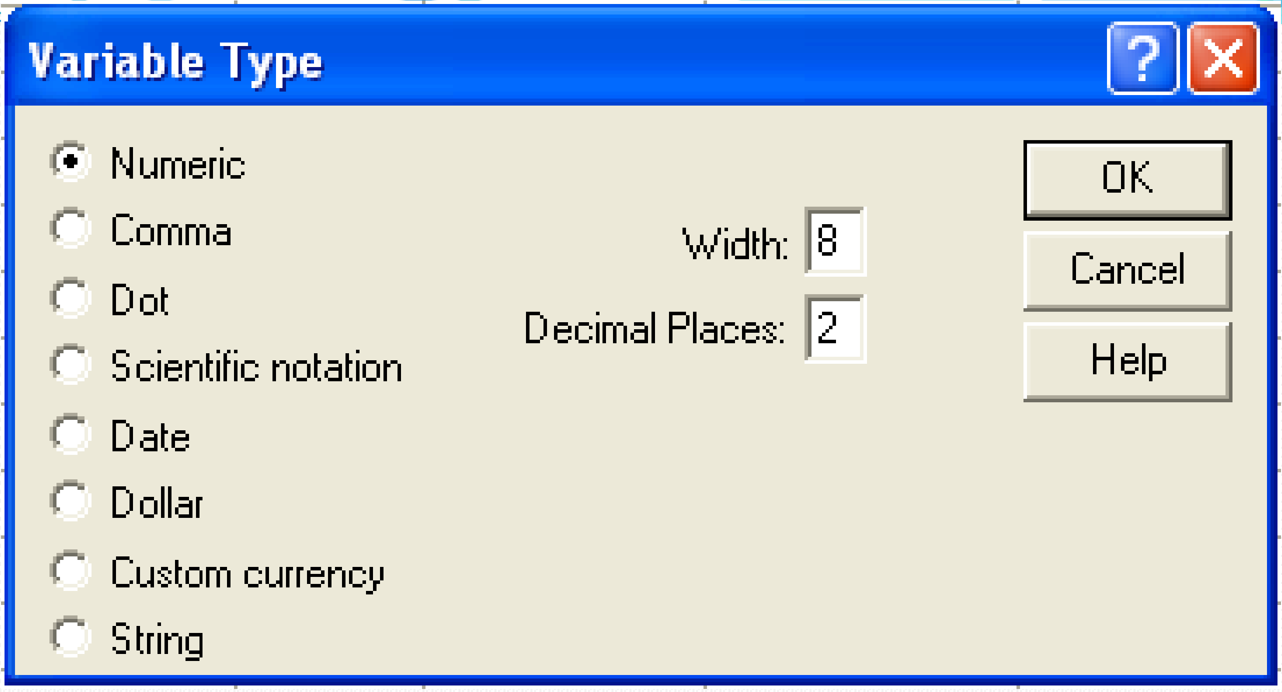
- Name
  - Each variable name must be unique; duplication is not allowed.
  - Start with a letter.
  - May have up to 8 characters, including letters, numbers, and the symbols (@, #, _, or $).
  - Variable names cannot end with a period.

The Variable View (con’t)

- Name (con’t)
  - Variable names that end with an underscore should be avoided.
  - The certain key words are reversed and may not be used as variable names, e.g. “compute”, “sum” and so forth.
  - Ex. Subject_ID, but not “subject-ID”, and not “Subject ID”.

The Variable View (con’t)

- Type
  - Basic type – numeric and string
  - Maximum width for numeric variables is 40 characters, the maximum number of decimal positions is 16.
  - String variables may contain letters or numbers. For string values a blank is considered a valid value.
  - Numeric operations on the string variables will NOT be allowed, e.g. finding the mean, variance, standard deviation, etc…



Variable Type

● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String

Width: 8

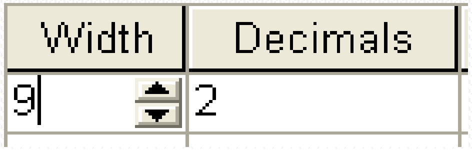
Decimal Places: 2

– If you select a string variable, you can tell SPSS how much “room” to leave in memory for each value, indicating the number of characters to b allowed for data entry in this string variable

The Variable View (con't)

- Width
  - The number of characters.

    SPSS will allow to be entered for the variable.
  - For a numerical value with decimals, this total width has to include a spot for each decimal, as well as one for the decimal point.



Width	Decimals
9	2

- Decimals
  - If more decimals have been entered or computed by SPSS, the additional information will be retained internally but not displayed on screen.

The Variable View (con’t)

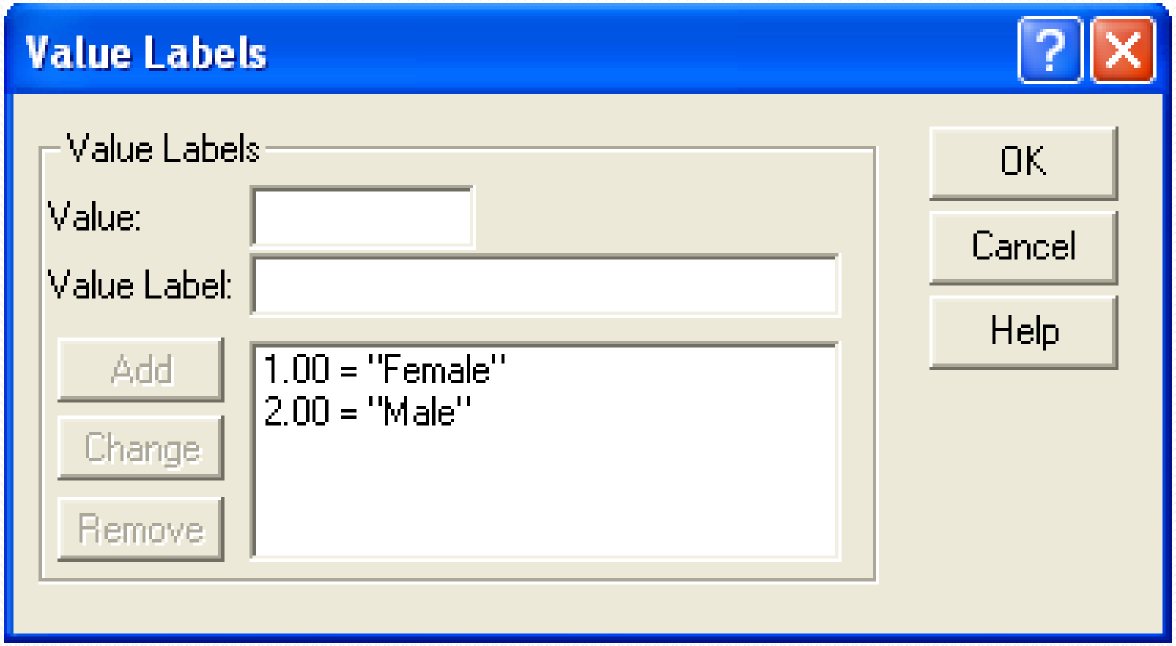
- Label
  - A string to identify in detail what a variable represents.
  - Is limited to 255 characters
  - May contain spaces and punctuation.

The Variable View (con’t)

- Values
  - Indicate how the numbers are assigned for categorical data.
  - Instead of typing into the computer the full answer to each question, codes are typed in (e.g., 1 if the respondent is female, 2 if male).
  - Codes are usually numerical, because this is what most statistical software expects, and using only numerical codes makes data entry faster.
  - These are easier to remember, and therefore tend to have lower error rates.

The Variable View (con’t)

- Values (con’t)
  - To code categorical variables in numeric format.
  - The Value Labels will be used.



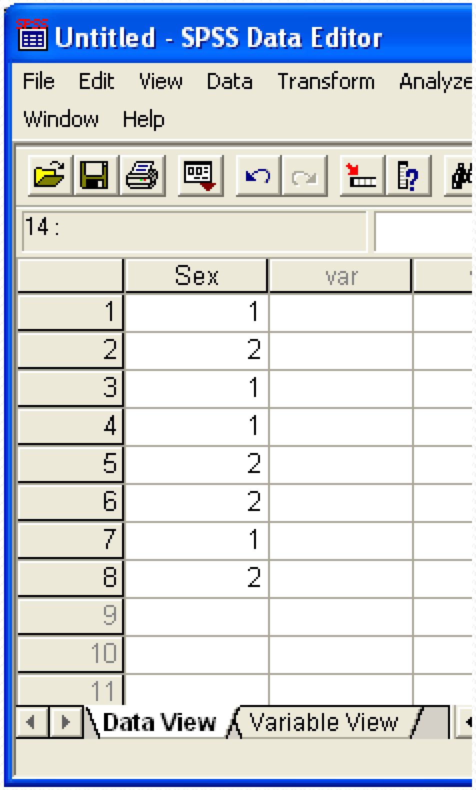
Value:

Value Label:

1.00 = "Female"
2.00 = "Male"

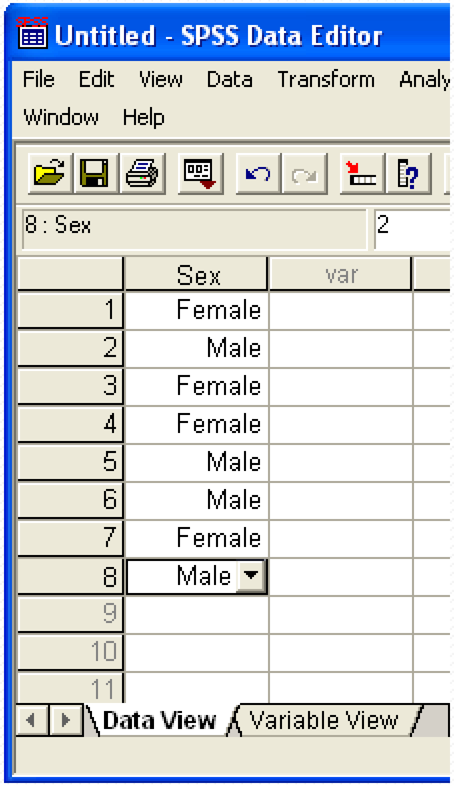
The Variable View (con't)

The labels can be seen in the Data View by clicking on the "toe tag" icon in the tool bar , which switches between the numeric values and their labels.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze
Window Help
Data View
Variable View

14 :

	Sex	var
1	1	
2	2	
3	1	
4	1	
5	2	
6	2	
7	1	
8	2	
9		
10		
11		

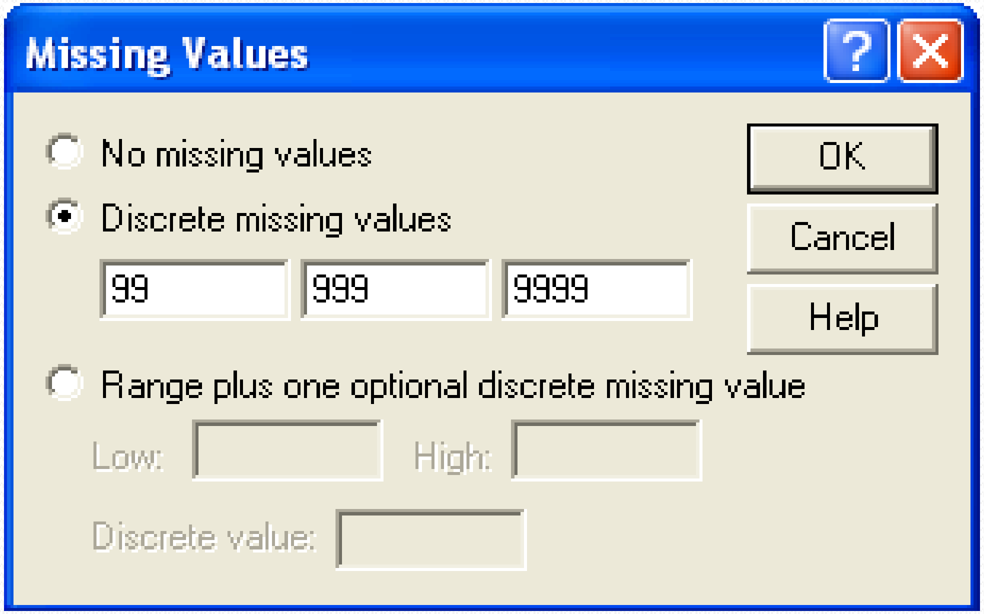

Untitled - SPSS Data Editor
File Edit View Data Transform Analy
Window Help
Data View
Variable View

8 : Sex 2

	Sex	var
1	Female	
2	Male	
3	Female	
4	Female	
5	Male	
6	Male	
7	Female	
8	Male	
9		
10		
11		

The Variable View (con’t)

- Missing
  - Signal to SPSS which data should be treated as missing.
  - System Missing data – SPSS display a single period.



○ No missing values

● Discrete missing values

99 | 999 | 9999

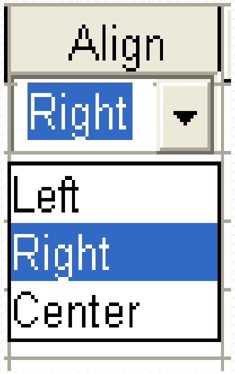
○ Range plus one optional discrete missing value

Low: High:

Discrete value:

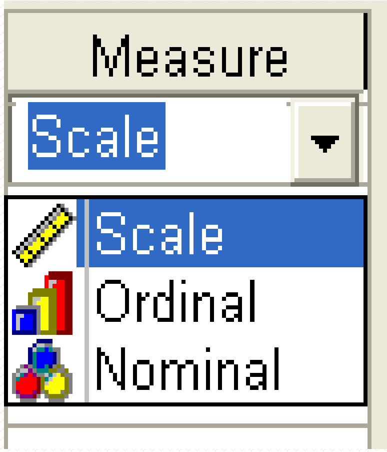
The Variable View (con't)

- Columns
  - How wide the column should be for each variable
  - Columns affect only the display of values in the Data Editor. Changing the column width does not change the defined width of a variable.
- Align



The Variable View (con't)

- Measure
  - Indicates the level of measurement.
  - Since SPSS does NOT differentiate between interval and ratio levels of measurement, both of these quantitative variable types are lumped together as "Scale".
  - Nominal and ordinal levels of measurement ARE differentiated.



Type of Measurement

The answers to the "numerical questions" are real numbers, not just arbitrary codes. There are four types of numerical scales that exist: nominal scales, ordinal scales, interval scales, and ratio scales.

- Scale
  - A ratio scale is one in which the answers are real numbers, and an answer of zero means what it says. "What age are you?" - "How tall are you?" - "How many children do you have?"
  - An interval scale (meaning equal-interval) - if there’s a zero point, it’s arbitrary, but the difference between two successive possible answers is the same. For example, the scale of temperature.

Type of Measurement (con’t)

- Ordinal
  - Frequently, categorical data responses represent more than two possible outcomes, and often these possible outcomes take on some inherent ordering.
  - No clue as to the relative distances between the levels.
  - For example, low – medium – high

    50% – 75% – 100% – 200%

    strong agree – agree – neutral – disagree – strongly disagree.

Type of Measurement (con’t)

- Nominal
  - A nominal scale isn’t really a scale at all, but an arbitrary code value to distinguish the different groups.
  - No inherent ordering to the categories.
  - For example, “Do you prefer the beach, mountains, or lake for a vacation?”

    “Which color is your favorite?”

Data Cleaning

- What most data entry programs will not do is warn the user when unlikely (but possible) codes occur. For example, if a respondent's age is shown as 99, this may be true, but it may also be a mistake.
- Therefore it's not only wild values that need to be checked. The first frequencies check from a program needs to be looked at very carefully to detect this kind of mistake.

Data Cleaning (con't)

- Check missing values - If the question was "Which sex are you, male or female?" and the possible answers are 1 for male and 2 for female, these should be the only values for that variable - except perhaps for a few blanks for the missing values.

Data Cleaning (con’t)

- There are two types of missing values in SPSS: system-missing and user-defined.
- System-missing values are assigned by SPSS when, for example, you perform an illegal function, like dividing a number by zero. System-missing values can also be assigned in an input data set.
- User-defined missing values are numeric values that you can specify and SPSS will consider to be missing. For example, you may define -9999 to be a missing value.

Data Cleaning (con't)

- You can assign many different missing values to a given variable, perhaps using the different values to indicate different reasons for the data point to be missing.

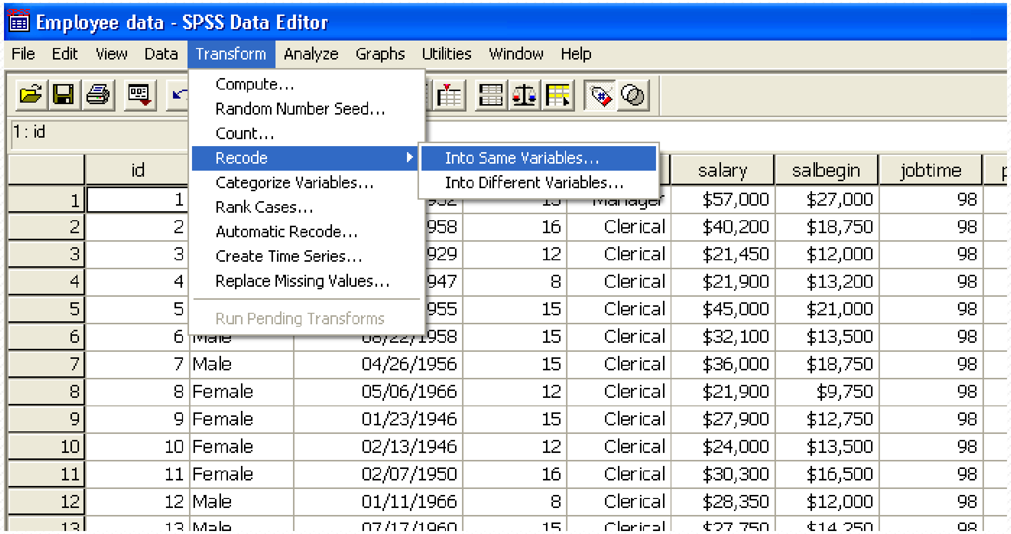
  For example, for an item on a survey, -9999 might indicate that the respondent skipped the item, -8888 might indicate that the item was not answered because it was part of skip pattern, and -7777 might indicate that a note was written in the margin instead of a standard response.

- You can specify up to three unique values for each variable. User-defined missing values can also be a range, such as 5 to 10. This is useful when you want to include only half of a scale, for example.
- String values can also be used as missing values, including a series of blanks (i.e., a null string).

Recode Procedure

Recode is used to

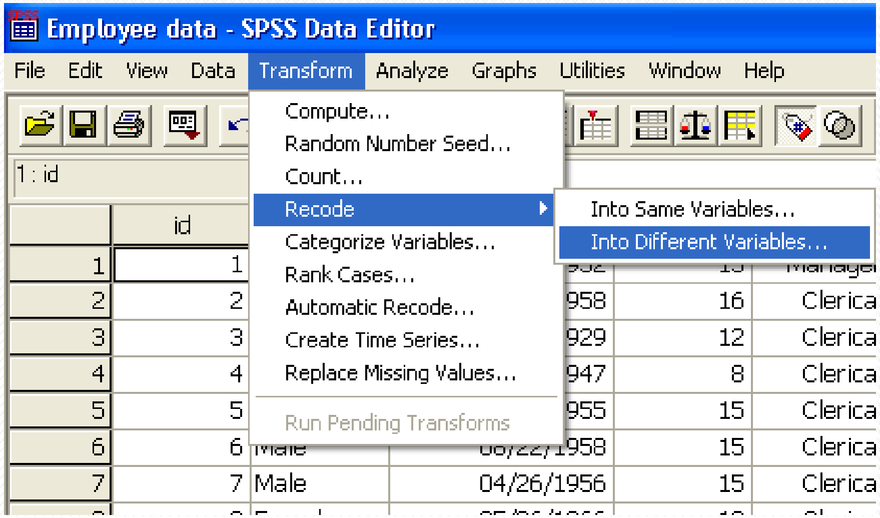
- to change the values of an existing variable
- to create a new variable based on the values an existing variable



	id					salary	salbegin	jobtime
1	1					$57,000	$27,000	98
2	2		958	16	Clerical	$40,200	$18,750	98
3	3		929	12	Clerical	$21,450	$12,000	98
4	4		947	8	Clerical	$21,900	$13,200	98
5	5		955	15	Clerical	$45,000	$21,000	98
6	6			15	Clerical	$32,100	$13,500	98
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98

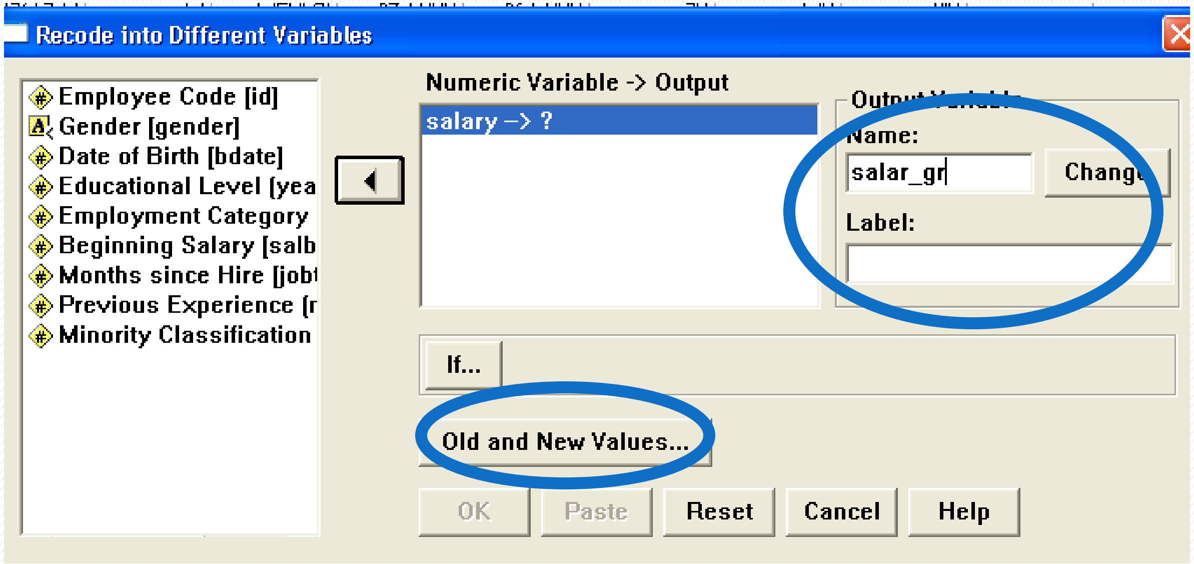
Recode into New Variable

- In the menu, click Transform.
- Select Recode.
- Click Into Different Variable(s)



Recode into New Variable

- Select and move variable(s) over.
- Name and label new variable.
- Click Old and New Values



Recode into Different Variables

Numeric Variable -> Output

salary --> ?

Output Variable

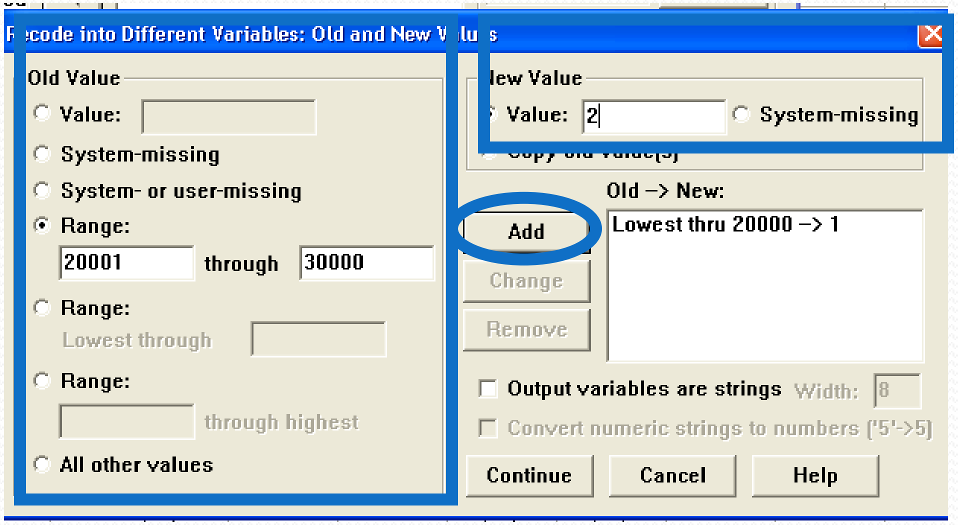
Name: salar_gr

Label:

Recode into New Variable

For each value of the existing variable

- Enter the new value
- Click Add
- Repeat for each value or range of values
- Click Continue



Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

○ System- or user-missing

● Range: 20001 through 30000

○ Range: Lowest through

○ Range: through highest

○ All other values

New Value

● Value: 2 ○ System-missing

Old --> New:

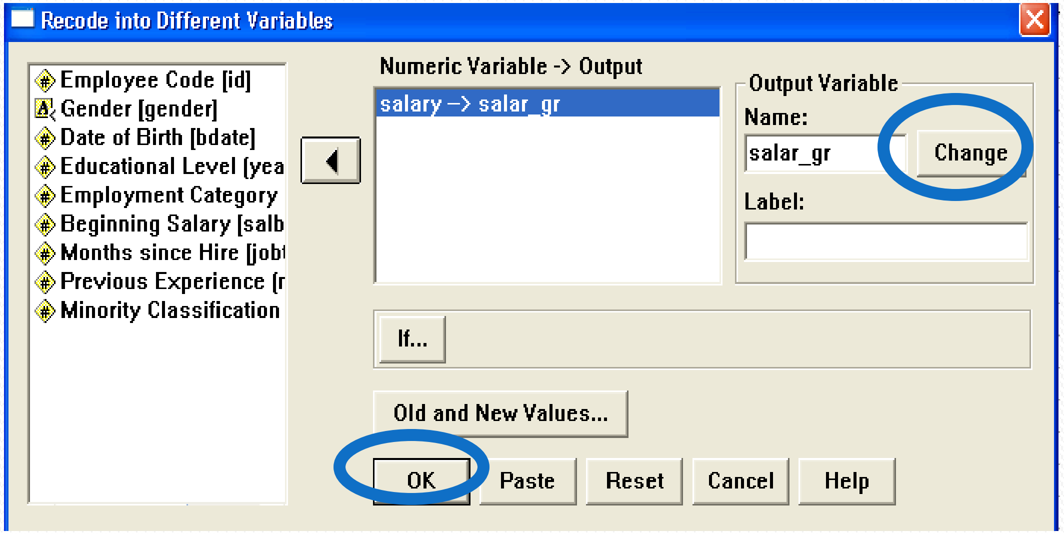
Lowest thru 20000 --> 1

□ Output variables are strings Width: 8

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

Recode into New Variable

- Click Change
- Click OK



Recode into Different Variables

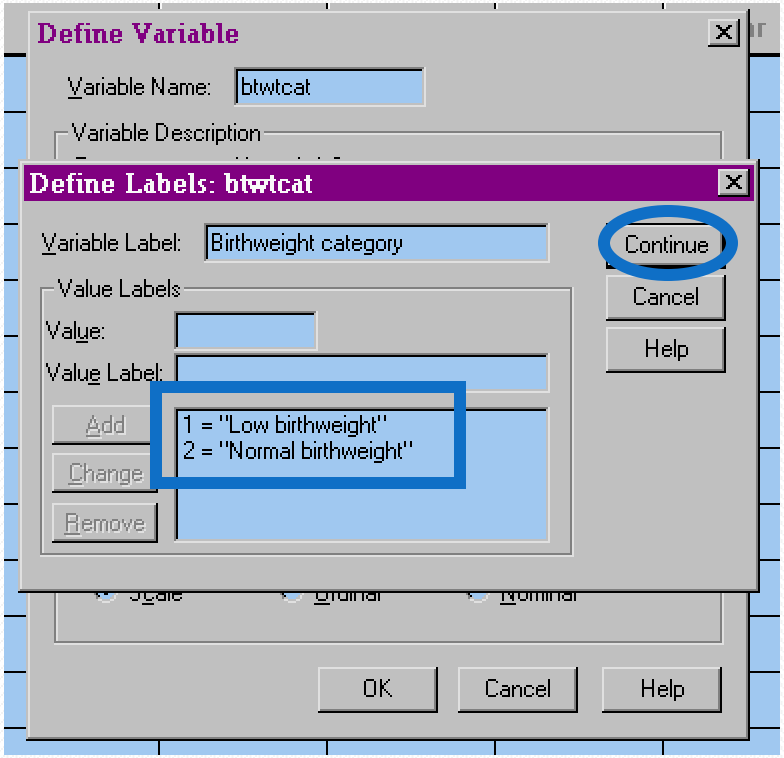
Employee Code [id]
Gender [gender]
Date of Birth [bdate]
Educational Level (yea
Employment Category
Beginning Salary [salb
Months since Hire [jobt
Previous Experience (r
Minority Classification

Numeric Variable -> Output: salary --> salar_gr

Output Variable
Name: salar_gr
Label:

Define Labels for New Variable

- In the Data menu, click Define Variable.
- Click Labels.
- Enter value labels for the new variable.



Define Variable

Variable Name: btwtcat

Define Labels: btwtcat

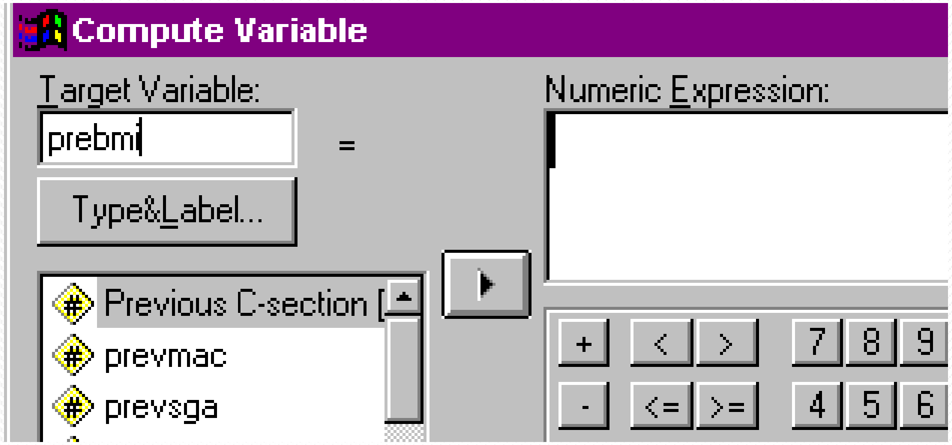
Variable Label: Birthweight category

Value Labels

1 = "Low birthweight"
2 = "Normal birthweight"

Compute Procedure

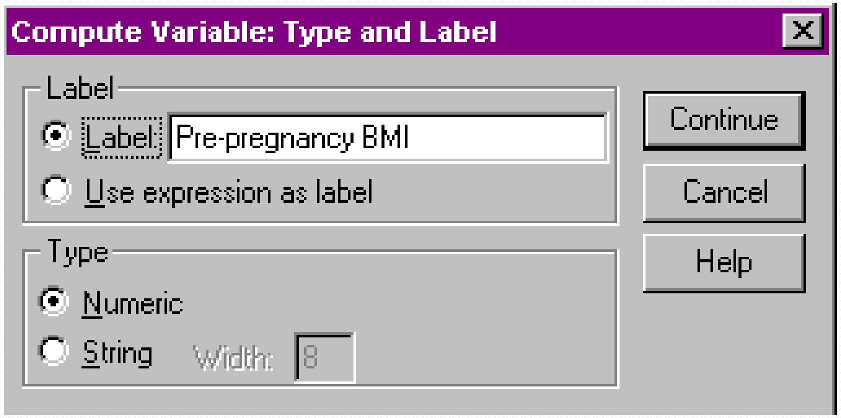
- Name the new variable.
- Click Type&Label to define the characteristics of the new variable.



Target Variable: prebmi

Compute Procedure

- Label the new variable.
- Enter the variable type.



Compute Variable: Type and Label

Label

● Label: Pre-pregnancy BMI

○ Use expression as label

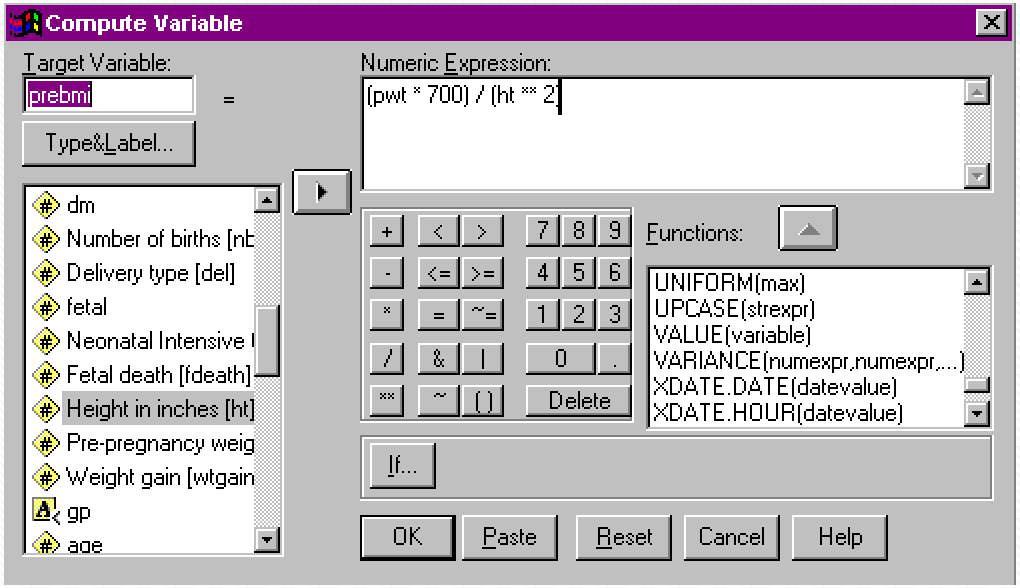
Type

● Numeric

○ String Width: 8

Compute Procedure

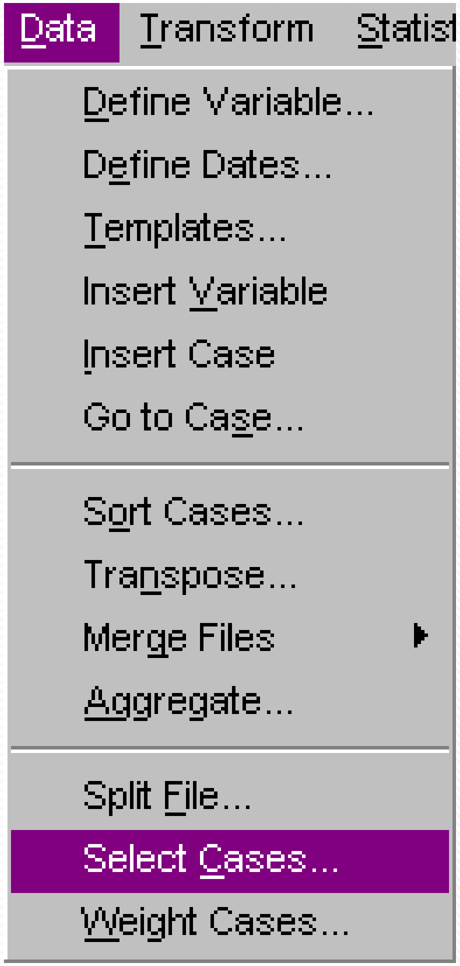
- Enter the numeric expression that will determine the values of the new variable.
- Click OK.



Target Variable: prebmi

Numeric Expression: (pwt * 700) / (ht ** 2

Select Cases

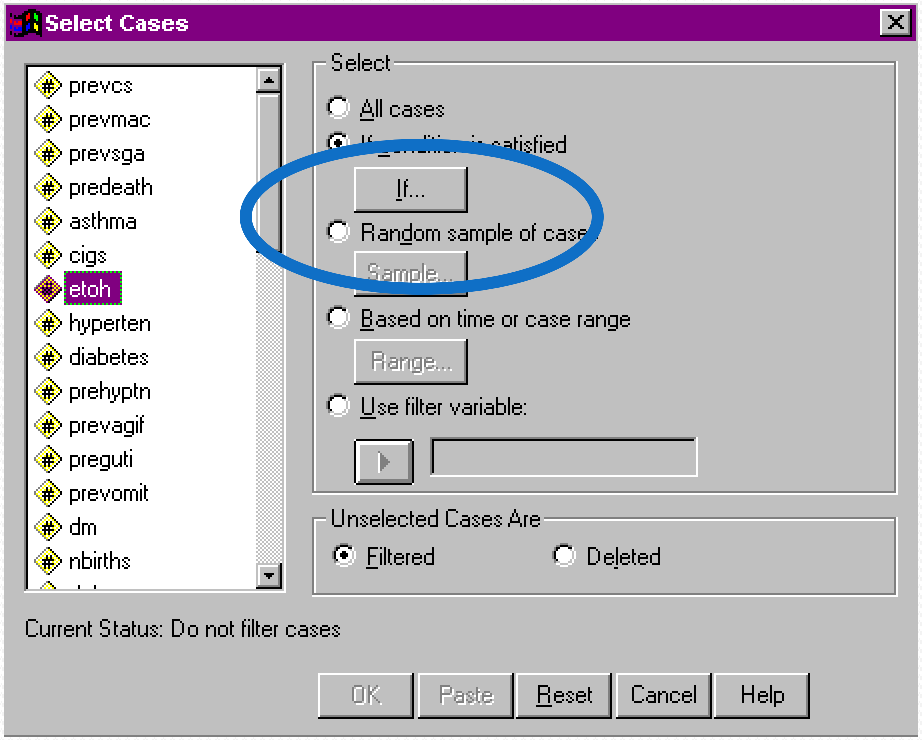


For a subset of the datafile, use Select Cases.

- In the menu, click Data.
- Click
  Select Cases...

Select Cases - Alcohol drinkers only

To select only those cases which meet certain criteria, choose the If option.



Select Cases

Select

○ All cases

● If condition is satisfied

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

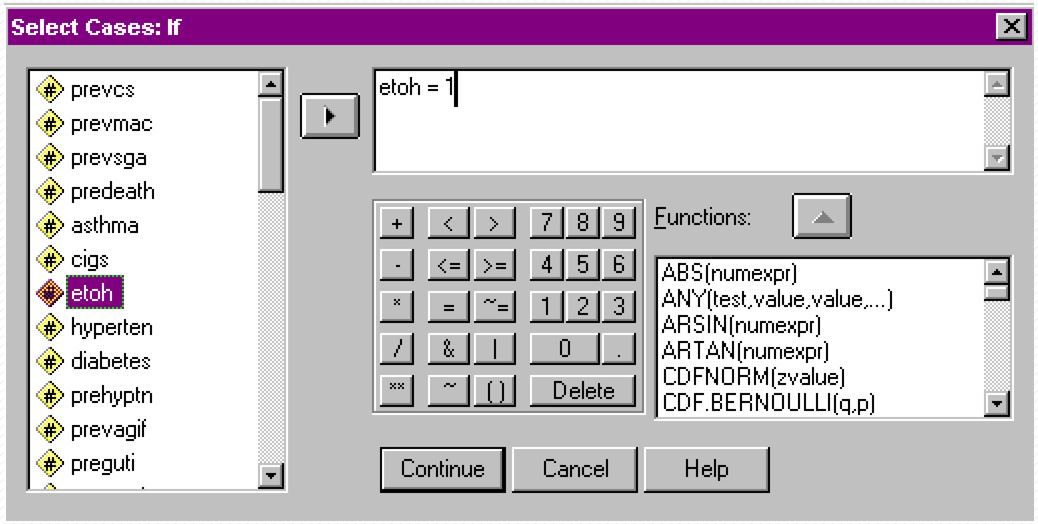
Unselected Cases Are

● Filtered ○ Deleted

Current Status: Do not filter cases

Select Cases - Alcohol drinkers only

- Enter the expression that will determine which variables will be selected.
- Click Continue.

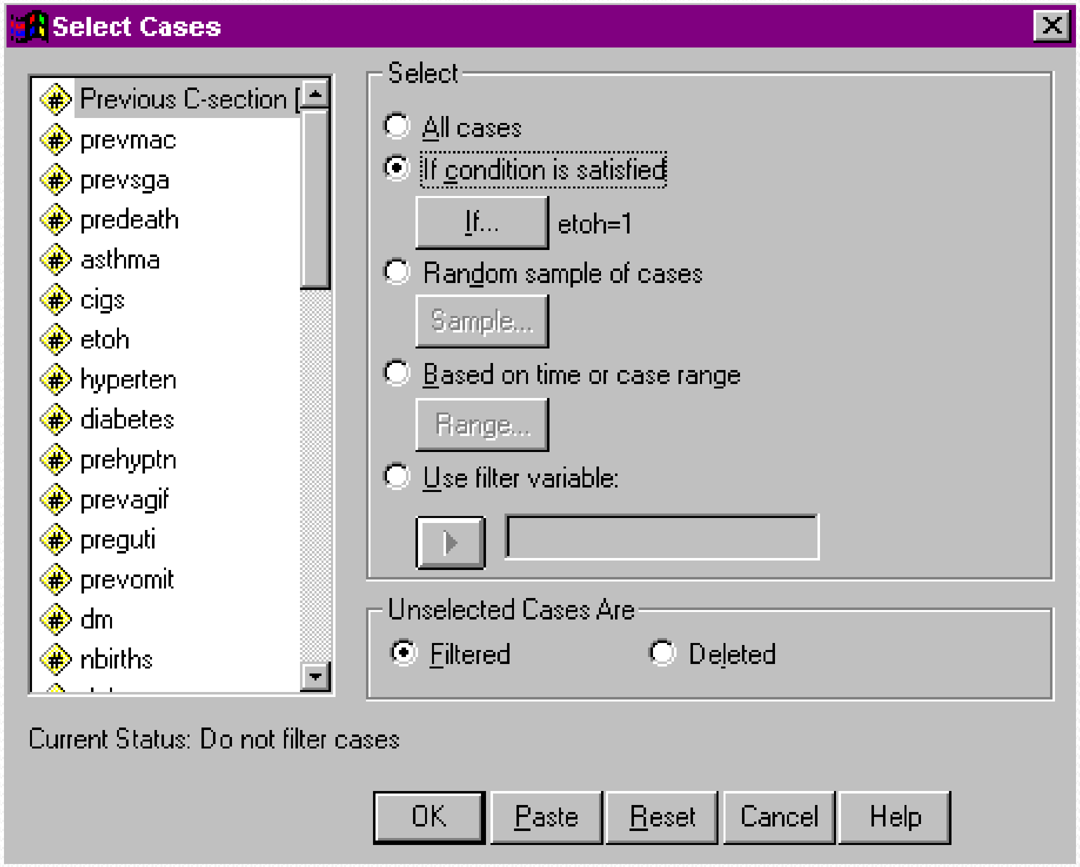


Select Cases: If

etoh = 1

Select Cases - Alcohol drinkers only

When you’ve finished specifying selection criteria, click OK.



Select Cases

Select

○ All cases

● If condition is satisfied

If... etoh=1

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

Unselected Cases Are

● Filtered ○ Deleted

Current Status: Do not filter cases