An Introduction to SPSS

Dr. Krishan K. Pandey

1

What is in this initial introduction Class

- SPSS interface: data view and variable view

- How to enter data in SPSS

- How to import external data into SPSS

- How to clean and edit data

- How to transform variables

- How to sort and select cases

- How to get descriptive statistics

Data used for this Lecture

- We use 2009 Youth Risk Behavior Surveillance System (YRBSS, CDC) as an example.
  - YRBSS monitors priority health-risk behaviors and the prevalence of obesity and asthma among youth and young adults.
  - The target population is high school students
  - Multiple health behaviors include drinking, smoking, exercise, eating habits, etc.
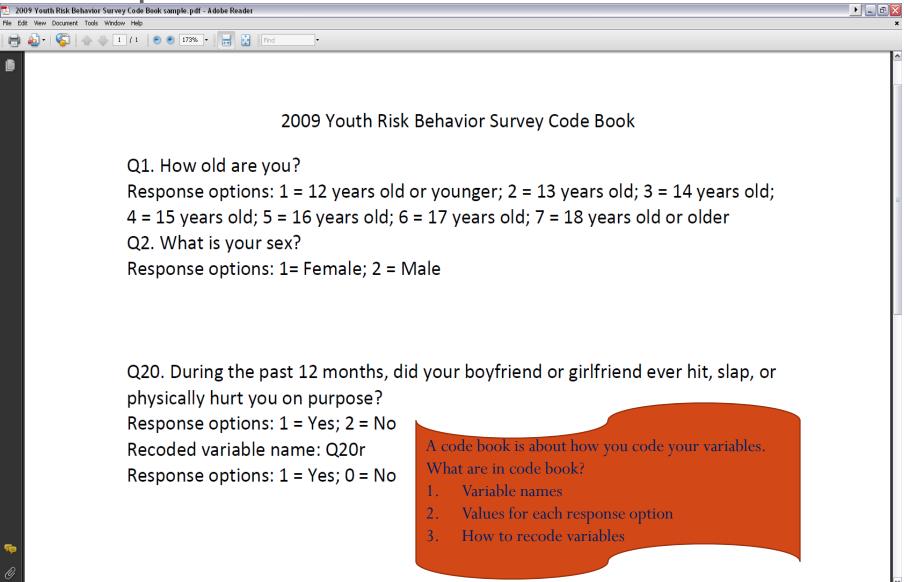
SPSS interface

- Data view
  - The place to enter data
  - Columns: variables
  - Rows: records
- Variable view
  - The place to enter variables
  - List of all variables
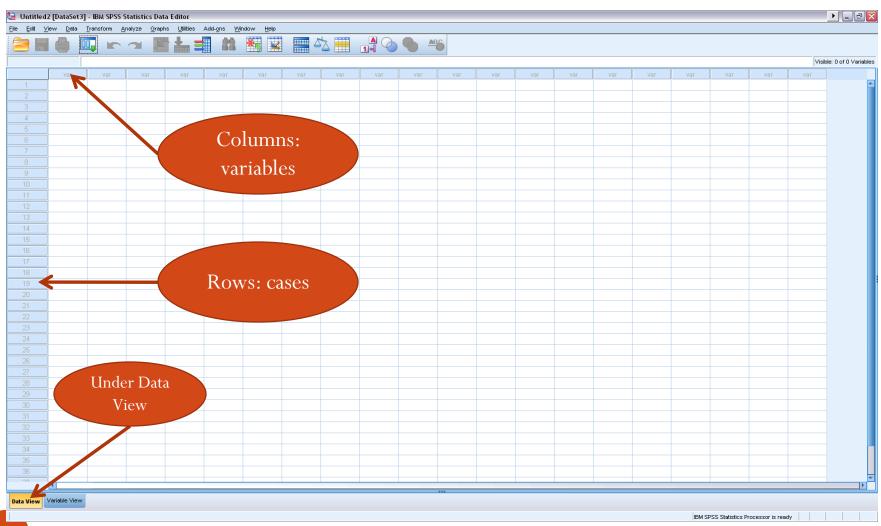 - Characteristics of all variables

Before the data entry

- You need a code book/scoring guide
- You give ID number for each case (NOT real identification numbers of your subjects) if you use paper survey.
- If you use online survey, you need something to identify your cases.
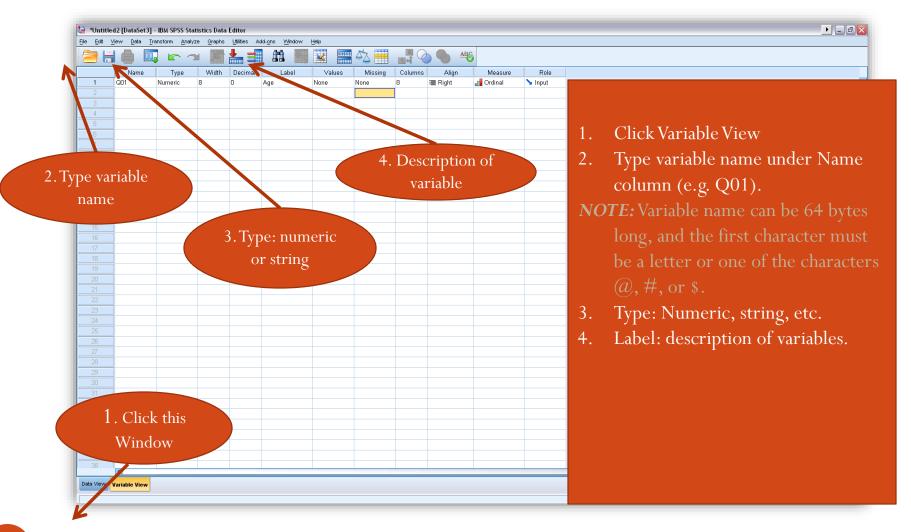- You also can use Excel to do data entry.

Example of a code book

2009 Youth Risk Behavior Survey Code Book

Q1. How old are you?

Response options: 1 = 12 years old or younger; 2 = 13 years old; 3 = 14 years old; 4 = 15 years old; 5 = 16 years old; 6 = 17 years old; 7 = 18 years old or older

Q2. What is your sex?

Response options: 1= Female; 2 = Male

Q20. During the past 12 months, did your boyfriend or girlfriend ever hit, slap, or physically hurt you on purpose?

Response options: 1 = Yes; 2 = No

Recoded variable name: Q20r

Response options: 1 = Yes; 0 = No

A code book is about how you code your variables. What are in code book?

1. Variable names
2. Values for each response option
3. How to recode variables

6

Enter data in SPSS 19.0

- Columns: variables
- Rows: cases
- Under Data View

Enter variables

2. Type variable name

3. Type: numeric or string

4. Description of variable

1. Click this Window

1. Click Variable View
2. Type variable name under Name column (e.g. Q01).
NOTE: Variable name can be 64 bytes long, and the first character must be a letter or one of the characters @, #, or $.
3. Type: Numeric, string, etc.
4. Label: description of variables.

8

Enter variables

Based on your code book!

Enter cases

A SPSS Statistics Data Editor window showing entered data:

Code	Q01
1001	1
1002	2
1003	2
1004	3
1005	2

1. Two variables in the data set.
2. They are: Code and Q01.
3. Code is an ID variable, used to identify individual case (NOT people's real IDs).
4. Q01 is about participants' ages: 1 = 12 years or younger, 2 = 13 years, 3 = 14 years…
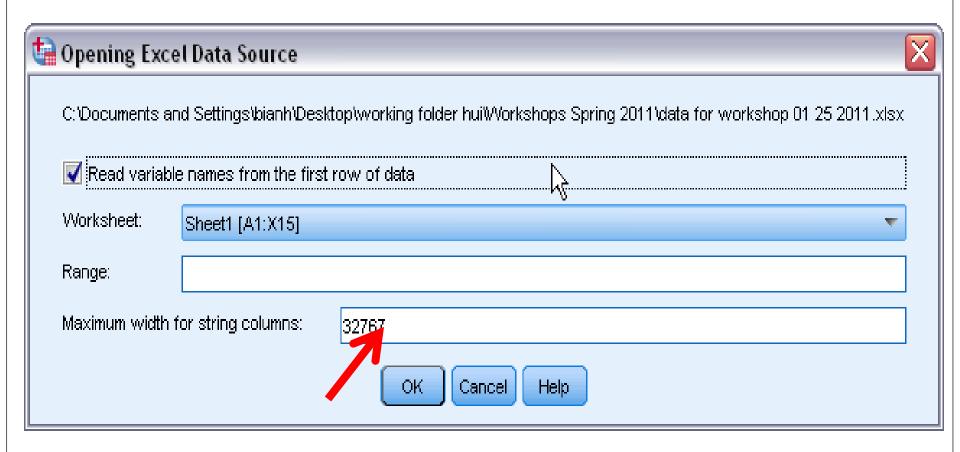
Under Data View

Import data from Excel

- Select File ⟶ Open ⟶ Data

- Choose **Excel** as file type
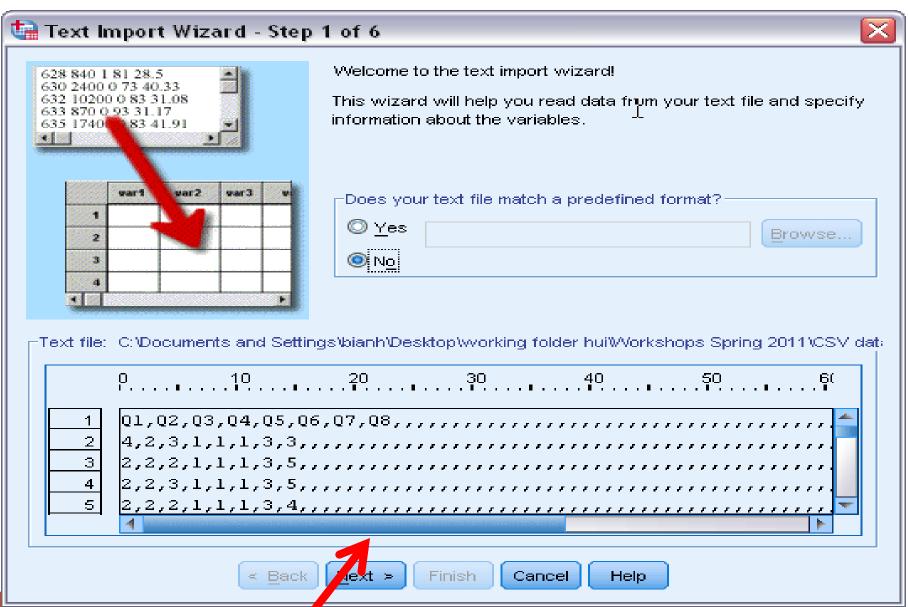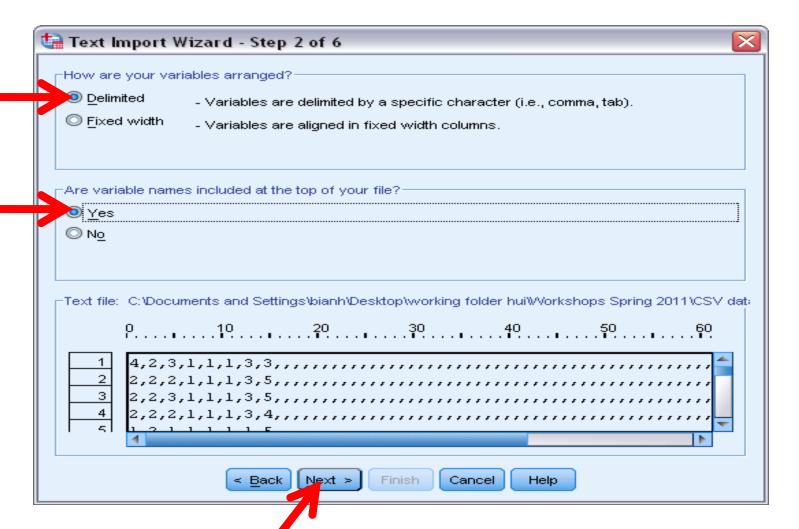
- Select the file you want to import

- Then click Open

Opening Excel Data Source

C:\Documents and Settings\bianh\Desktop\working folder hui\Workshops Spring 2011\data for workshop 01 25 2011.xlsx

☑ Read variable names from the first row of data

Worksheet: Sheet1 [A1:X15]

Range:

Maximum width for string columns: 32767

OK   Cancel   Help

Open Excel files in SPSS

Opening Excel Data Source

C:\Documents and Settings\bianh\Desktop\working folder hui\Workshops Spring 2011\data for workshop 01 25 2011 .xlsx

☑ Read variable names from the first row of data

Worksheet: Sheet1 [A1:X15]

Range:

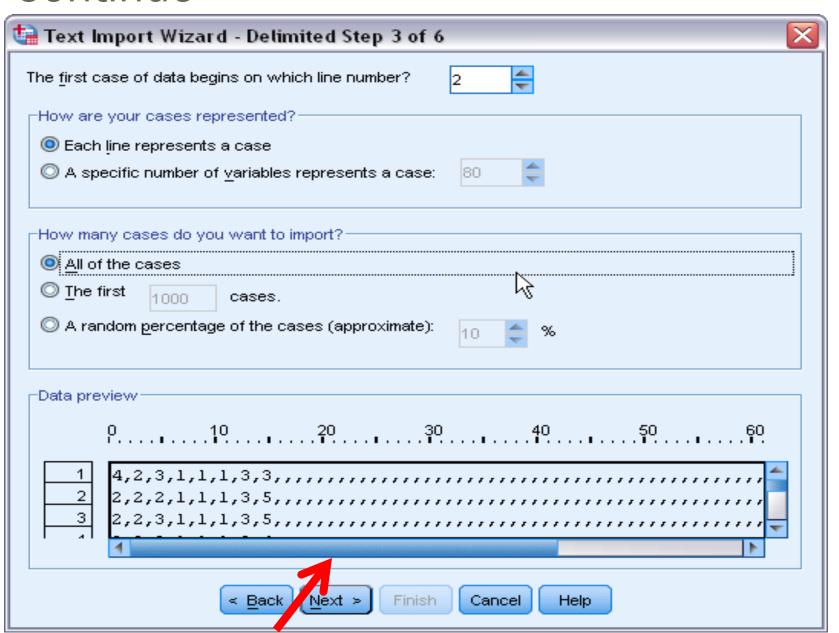Maximum width for string columns: 32767
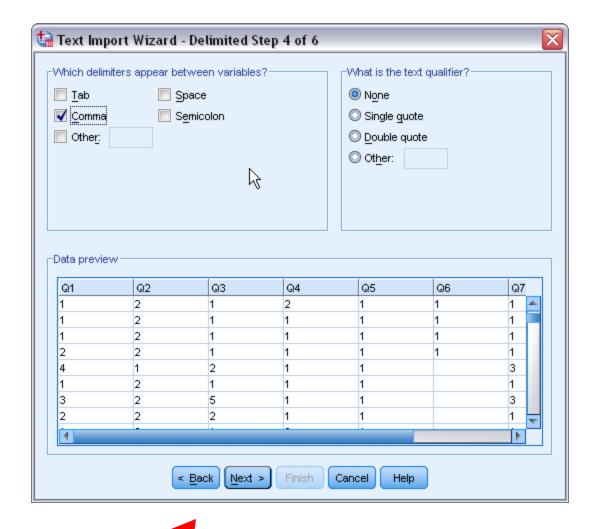
[OK] [Cancel] [Help]

12

Import data from CVS file

- CVS is a comma-separated values file.

- If you use Qualtrics to collect data (online survey), you will get a CVS data file.

- Select File ⟶ Open ⟶ Data

- Choose **All files** as file type

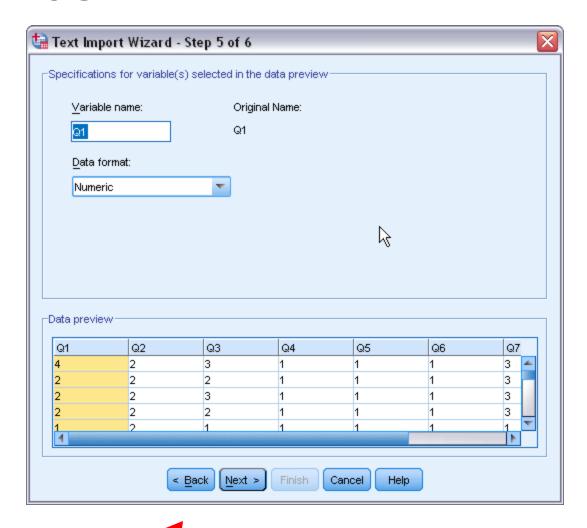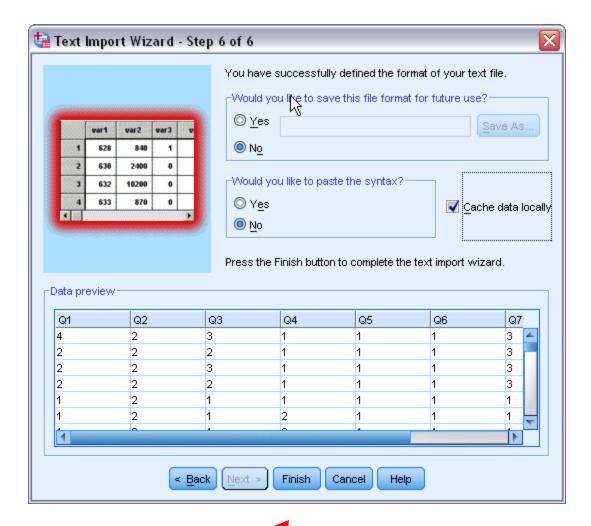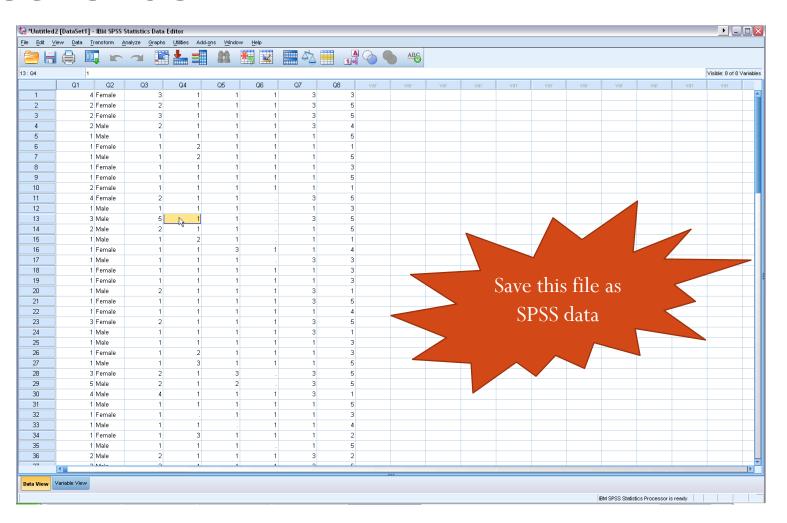- Select the file you want to import

- Then click Open

Continue

Text Import Wizard - Step 1 of 6

628 840 1 81 28.5
630 2400 0 73 40.33
632 10200 0 83 31.08
633 870 0 93 31.17
635 1740 0 83 41.91

Welcome to the text import wizard!

This wizard will help you read data from your text file and specify information about the variables.

Does your text file match a predefined format?

○ Yes

● No

Browse...

Text file: C:\Documents and Settings\bianh\Desktop\working folder hui\Workshops Spring 2011\CSV dat:

```
1  Q1,Q2,Q3,Q4,Q5,Q6,Q7,Q8,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
2  4,2,3,1,1,1,3,3,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
3  2,2,2,1,1,1,3,5,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
4  2,2,3,1,1,1,3,5,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
5  2,2,2,1,1,1,3,4,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

< Back | Next > | Finish | Cancel | Help

14

Continue

Continue

Text Import Wizard - Delimited Step 3 of 6

The first case of data begins on which line number? [2]

How are your cases represented?

- ⦿ Each line represents a case
- ◯ A specific number of variables represents a case: [80]

How many cases do you want to import?

- ⦿ All of the cases
- ◯ The first [1000] cases.
- ◯ A random percentage of the cases (approximate): [10] %

Data preview

```
     0.........10.........20.........30.........40.........50.........60
  1  4,2,3,1,1,1,3,3,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
  2  2,2,2,1,1,1,3,5,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
  3  2,2,3,1,1,1,3,5,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

[< Back] [Next >] [Finish] [Cancel] [Help]

Continue

Continue

Continue

Continue

Clean data after import data files

- Key in values and labels for each variable

- Run frequency for each variable

- Check outputs to see if you have variables with wrong values.

- Check missing values and physical surveys if you use paper surveys, and make sure they are real missing.

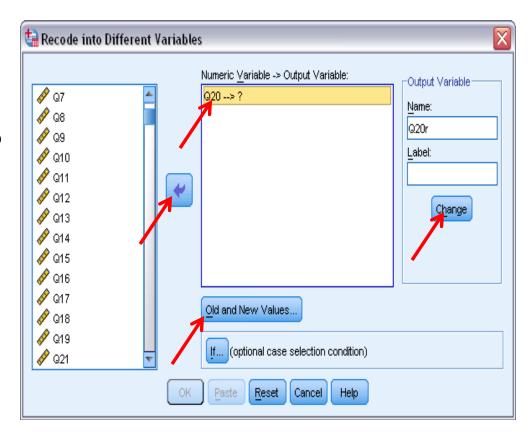- Sometimes, you need to recode string variables into numeric variables

Continue

Q1 How old are you

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 12 years old or younger	19	.1	.1	.1
	2 13 years old	17	.1	.1	.2
	3 14 years old	1851	11.3	11.3	11.5
	4 15 years old	4045	24.7	24.7	36.3
	5 16 years old	4234	25.8	25.9	62.2
	6 17 years old	3963	24.2	24.2	86.4
	7 18 years old or older	2215	13.5	13.6	100.0
	8	1	.0	.0	100.0
	9	1	.0	.0	100.0
	Total	16344	99.6	100.0	
Missing	System	66	.4		
Total		16410	100.0		

Wrong entries
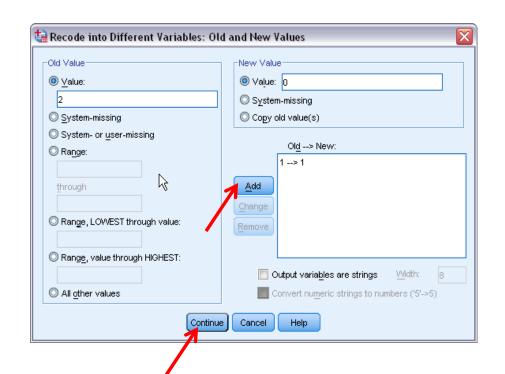
Variable transformation

- Recode variables

1. Select Transform ⟶ Recode into Different Variables
2. Select variable that you want to transform (e.g. Q20): we want 1= Yes and 0 = No
3. Click Arrow button to put your variable into the right window
4. Under Output Variable: type name for new variable and label, then click Change
5. Click Old and New Values

Continue

6. Type 1 under Old Value and 1 under New Value, click Add. Then type 2 under Old Value, and 0 under New Value, click Add.

7. Click Continue after finish all the changes.

8. Click Ok
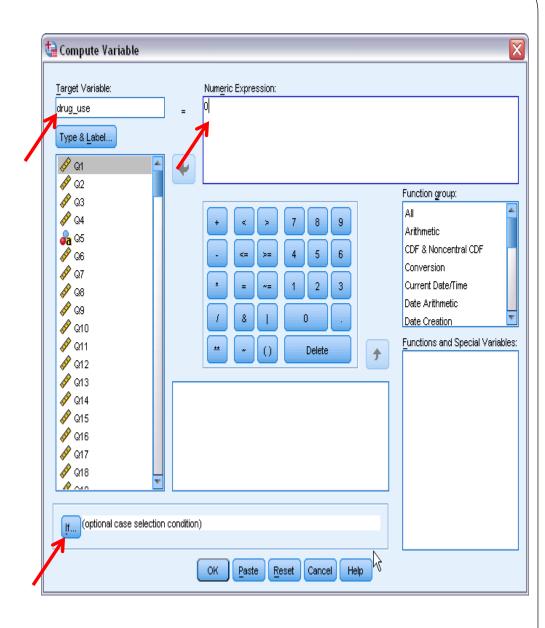
Variable transformation

- Compute variable (use YRBSS 2009 data)
  - Example 1. Create a new variable: drug_use (During the past 30 days, any use of cigarettes, alcohol, and marijuana is defined as use, else as non-use). There are two categories for the new variable (use vs. non-use). Coding: 1= Use and 0 = Non-use
    1. Use Q30, Q41, and Q47 from 2009 YRBSS survey
    2. Non-users means those who answered **0** days/times to all three questions.
    3. Go to Transform ⟶ Compute Variable

Continue

4. Type "drug_use" under Target Variable

5. Type "0" under Numeric Expression. 0 means Non-use
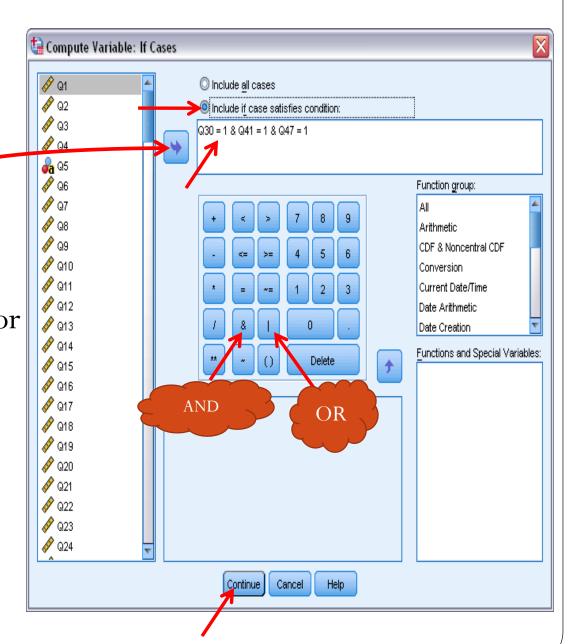
6. Click If button.

Continue

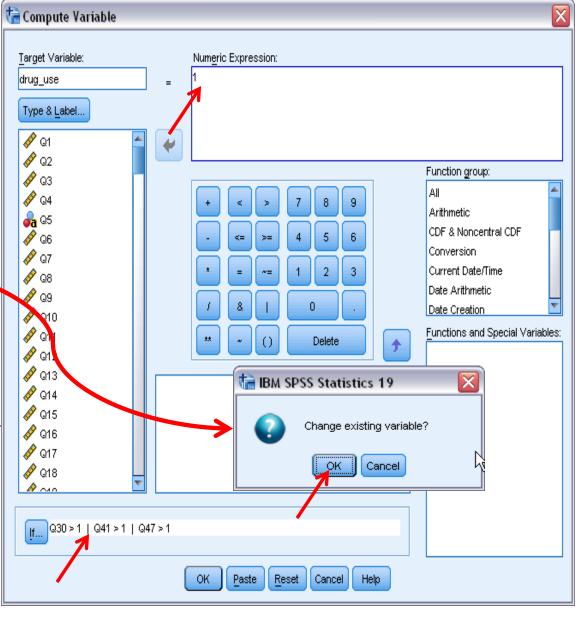7. With help of that Arrow button, type **Q30= 1 & Q41 = 1 & Q47= 1** then click Continue

8. Do the same thing for Use, but the numeric expression is different: **Q30> 1 | Q41> 1 | Q47>1**
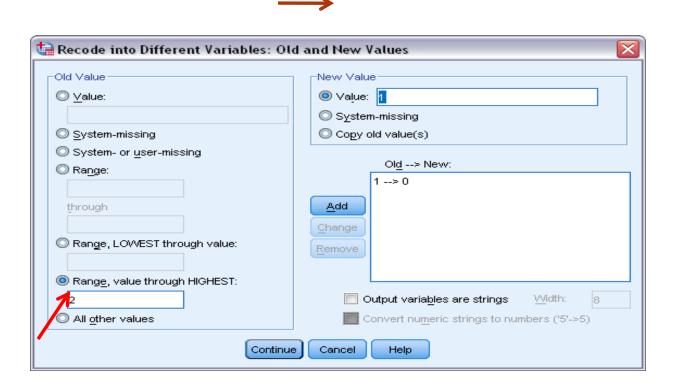
Continue

9. Click OK

10. After click OK, a small window asks if you want to change existing variable because drug_use was already created when you first define non-use.
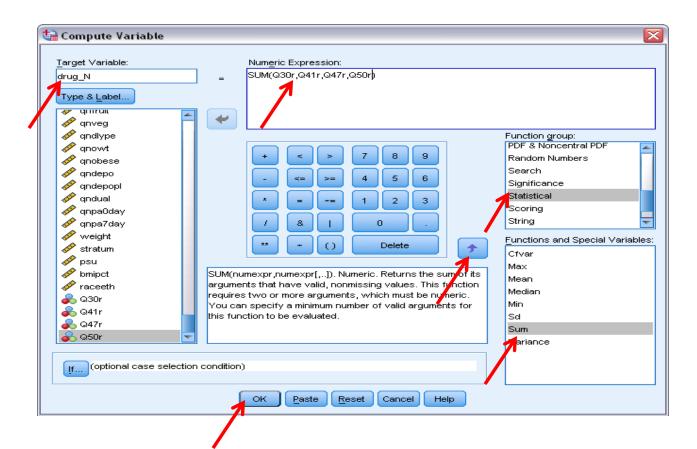
11. Click ok.

Continue

- Compute variables
  - Example 2. Create a new variable drug_N that assesses total number of drugs that adolescents used during the last 30 days.
    1. Use Q30 (cigarettes), 41 (alcohol), 47 (marijuana), and 50 (cocaine). The number of drugs used should be between 0 and 4.
    2. First, recode all four variables into two categories: 0 = non-use (0 days), 1 = use (at least 1 day/time)
    3. Four variables have 6 or 7 categories

Continue

4. Recode four variables: 1 (old) = 0 (new), 2-6/7 (old) = 1 (New).

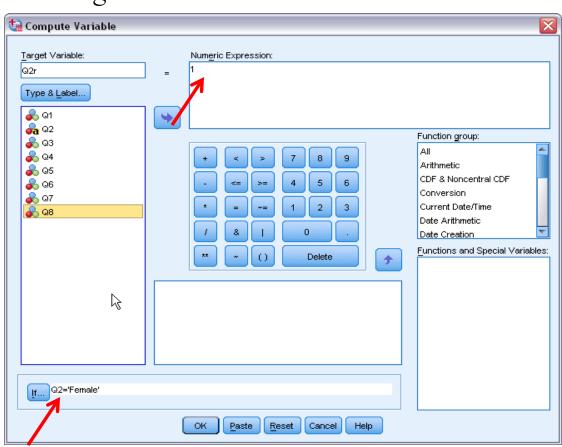5. Then select Transform   Compute Variable

Continue

6. Type drug_N under Target Variable

7. Numeric Expression: SUM (Q30r,Q41r,Q47r,Q50r)

8. Click OK

Continue
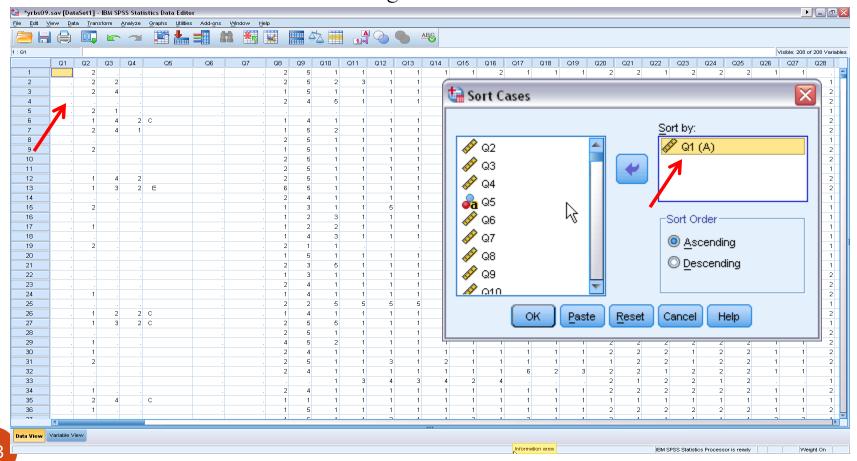
- Compute variables
  - Example 3: Convert string variable into numeric variable

1. Enter 1 at Numeric Expression.
2. Click If button and type Q2 = 'Female'
3. Then click Ok.
4. Enter 2 at Numeric Expression.
5. Click If button and type Q2 = 'Male'
6. Then click Ok
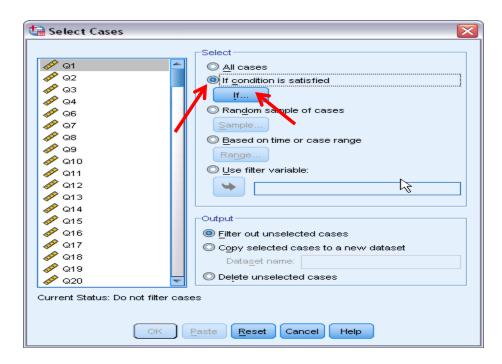
Sort and select cases

- Sort cases by variables: Data → Sort Cases

- You can use Sort Cases to find missing.

Sort and select cases

- Select cases
  - Example 1. Select Females for analysis.
  1. Go to Data ⟶ Select Cases
  2. Under Select: check the second one
  3. Click If button

Continue
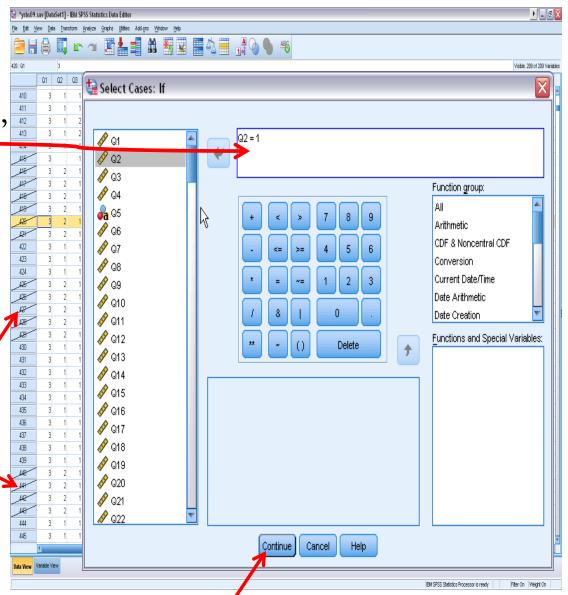
4. Q2 (gender) = 1,

1 means Female
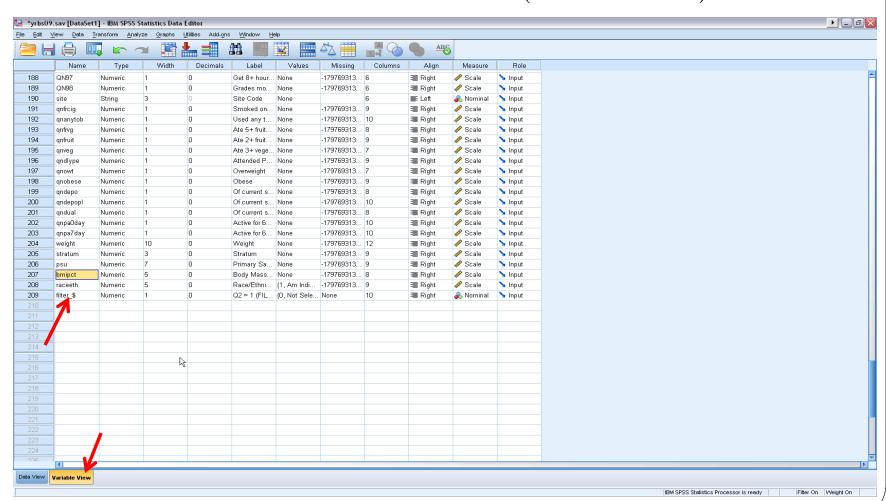
5. Click Continue

6. Click Ok

Unselected cases : Q2 = 2
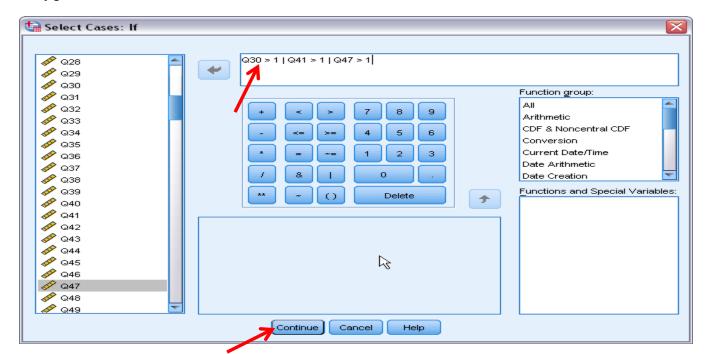
Sort and select cases

7. You will see a new variable: filter_$ (Variable view)

Sort and select cases
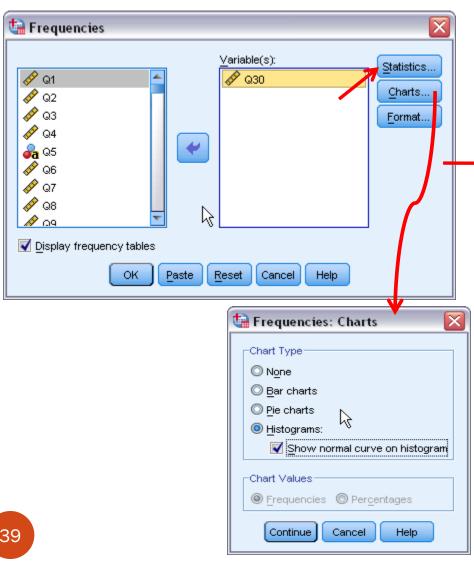
- Select cases
  - Example 2. Select cases who used any of cigarettes, alcohol, and marijuana during the last 30 days.

  1. Data ⟶ Select Cases
  2. Click If button
  3. Type Q30 > 1 | Q41 > 1 | Q47 > 1, click Continue
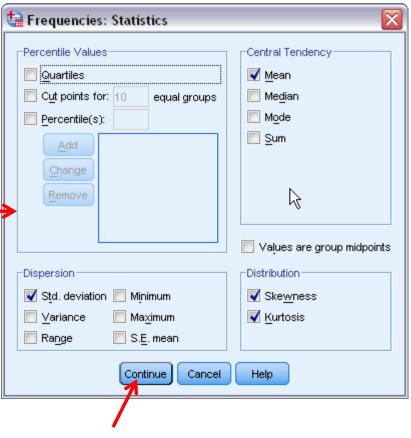
Basic statistical analysis
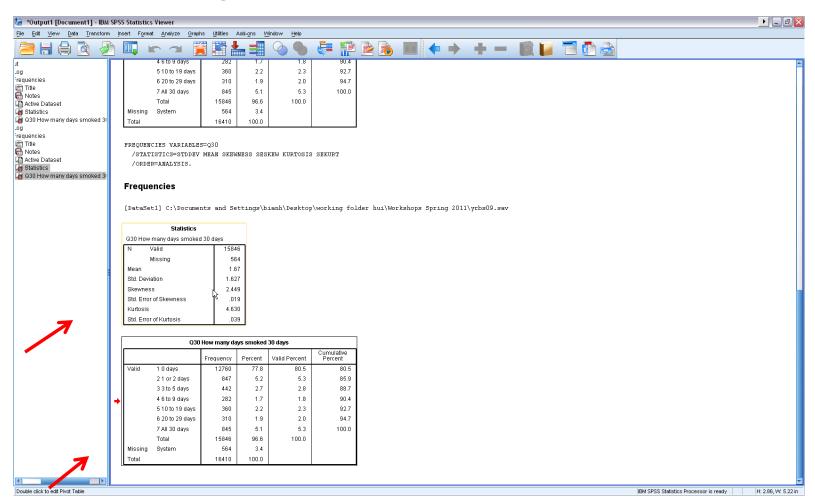
- Descriptive statistics
  - Purposes:
    1. Find wrong entries
    2. Have basic knowledge about the sample and targeted variables in a study
    3. Summarize data
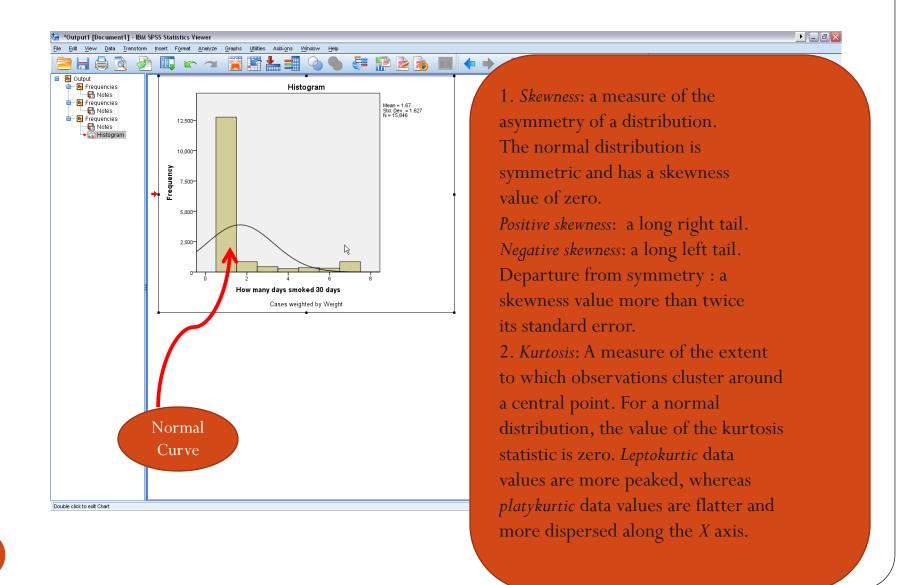
Analyze $\longrightarrow$ Descriptive statistics $\longrightarrow$ Frequency

Continue

Frequency table

File   Edit   View   Data   Transform   Insert   Format   Analyze   Graphs   Utilities   Add-ons   Window   Help

		4 6 to 9 days	282	1.7	1.8	90.4
		5 10 to 19 days	360	2.2	2.3	92.7
		6 20 to 29 days	310	1.9	2.0	94.7
		7 All 30 days	845	5.1	5.3	100.0
		Total	15846	96.6	100.0	
	Missing	System	564	3.4		
	Total		16410	100.0		

```
FREQUENCIES VARIABLES=Q30
  /STATISTICS=STDDEV MEAN SKEWNESS SESKEW KURTOSIS SEKURT
  /ORDER=ANALYSIS.
```

Frequencies

[DataSet1] C:\Documents and Settings\bianh\Desktop\working folder hui\Workshops Spring 2011\yrbs09.sav

Statistics

Q30 How many days smoked 30 days

|		
N	Valid	15846
	Missing	564
Mean		1.67
Std. Deviation		1.627
Skewness		2.449
Std. Error of Skewness		.019
Kurtosis		4.630
Std. Error of Kurtosis		.039

Q30 How many days smoked 30 days

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 0 days	12760	77.8	80.5	80.5
	2 1 or 2 days	847	5.2	5.3	85.9
	3 3 to 5 days	442	2.7	2.8	88.7
	4 6 to 9 days	282	1.7	1.8	90.4
	5 10 to 19 days	360	2.2	2.3	92.7
	6 20 to 29 days	310	1.9	2.0	94.7
	7 All 30 days	845	5.1	5.3	100.0
	Total	15846	96.6	100.0	
Missing	System	564	3.4		
Total		16410	100.0		

Double click to edit Pivot Table

IBM SPSS Statistics Processor is ready   H: 2.86, W: 5.22 in

Normal Curve

1. *Skewness*: a measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of zero.
Positive skewness: a long right tail.
Negative skewness: a long left tail. Departure from symmetry : a skewness value more than twice its standard error.
2. *Kurtosis*: A measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is zero. *Leptokurtic* data values are more peaked, whereas *platykurtic* data values are flatter and more dispersed along the *X* axis.

41

THANK YOU