QUICK & EASY STATISTICS

A Practical and Interactive Approach Using SPSS

FATAI AKEMOKWE

QUICK & EASY STATISTICS

A Practical and Interactive Approach Using SPSS

FATAI AKEMOKWE

ISBN-13: 978-1515333449

ISBN-10: 1515333442

For ’Misan,
for making every moment meaningful.

This page deliberately left blank

CONTENTS

ACKNOWLEDGEMENTS

My thanks go to Dr Afe Orilade and Jafaru Akemokwe, my co-instructors at Enigma Ventures. Dr (Mrs) V. Omuemu and Dr. Frank Imarhiagbe mentored me in clinical research; you are always appreciated.

This book only exists because of all the researchers who have pushed me into seeking simpler ways to explain Statistics using SPSS. Thanks for the pressure.

WELCOME

The aim of this book is to provide you with a focused practical approach to the use of statistics in research with emphasis on use of the IBM® SPSS® Statistics as a tool. At the time of writing, the latest SPSS available is version 22. Differences in version so far have not altered the basic process of using it for analysis.

For many aspiring researchers, learning statistics is a frightful rite of passage. This book sticks to the essentials. I have trimmed away the distracting and often unnecessary aspects that are best left to career statisticians.

Although the emphasis here is on using SPSS, the necessary theoretical basis in statistics will be provided as succinctly as possible. This knowledge can then be adapted to other statistical software.

The sample data files used in this book can be downloaded using this Internet link: is.gd/zMDZ1D

Unless otherwise stated, *myData.sav* is the data file used in the examples.

Visit our website and blog for these and more resources.

Enjoy.

AN INTRODUCTION TO STATISTICAL TERMS

There are 3 types of lies:
lies, damn lies and statistics!
- Attributed to Mark Twain
(or Benjamin Disraeli.
Apparently, history is the fourth type of lie.)

Statistics is the science (an, maybe, art) of collecting, organizing and analyzing data.
Descriptive statistics: aims at describing a population:
Inferential statistics: aims at drawing conclusions from the data.
Data: raw facts and figures.
Population: the entire set of data which one aims to describe or make inferences about.
Sample: A subset of the population
Variable: any piece of data that varies from case to case or varies from one observation to another. (A *constant* is data that remains unchanged)

Independent (or Predictor) variable: a factor whose effect on another variable is to be assessed. This is also called a **Risk Factor**, **Explanatory variable** or **Exposure variable**.

Dependent (or Outcome) variable: a variable whose value is assumed to be influenced by other variables.

For example, does the sex of patients affect their response to antihypertensive medication? In this case, *sex* is the predictor variable while "*response to antihypertensive medication*" is the dependent variable. On the other hand, if the research question is "*Does timing of intercourse affect the sex of the fetus?*", then "*sex of the fetus*" becomes the outcome variable while "*timing of intercourse"* is the independent variable.

Parameter: a number that is used to describe the *population*. It is represented with Greek letters for example, population standard deviation (σ), population mean (μ).

Statistic (without the final "*s"*!): a number used to describe a *sample*. Represented with Roman (normal) letters for example, sample standard deviation (s), sample mean ($\bar{x}$).

STUDY DESIGN

The design of a study determines the analysis to be performed on data that is collected. In turn, choosing a study design depends on the aim of the researcher.

Do you intend to describe the characteristics of one or more populations in terms of rates, proportions and percentages?

Study design: (Descriptive) Cross-sectional/Prevalence study

Example of Research Question: *How many doctors work in Gwagalada? How many of them smoke tobacco? What is the prevalence rate of hepatitis B antigen positivity in nurses in Gwagalada?*

Do you wish to analyze the frequency of occurrence of a suspected risk factor in those with a disease ("cases") compared to its frequency in those without the disease ("controls")? *Controls may be randomly selected or they be matched (in terms of age, sex, race or other variables).*

Study Design: (Retrospective) Case-Control study

Example of Research Question: *When*

compared to a group of healthy persons, are lung cancer patients more likely to have smoked cigarettes?

Do you plan to follow up healthy persons over time to determine the risk factors associated with future occurrence of a disease?

Study Design: (Prospective) Cohort Study

Example of Research Question: *Over a thirty-year period, are smokers in Gwagalada more likely to develop lung cancer compared to non-smokers?*

Do you want to find out how effective an intervention is?

Study Design: (Experimental) Randomized Control Trial

Example of Research Question: *Is there any difference in the efficacy of radiotherapy, chemotherapy, radical surgery, or various combinations of these therapies in the treatment of early-stage breast cancer?*

PREPARING YOUR DATA FOR ANALYSIS

TYPES OF DATA/VARIABLES

All data boils down to two *types*:

Numeric: any data that is represented with numbers only. Formats available in SPSS are:

(Plain) *Numeric*

Date: Allows dates to be entered in various formats

Dot: Decimal points represented with dots, every thousand separated by commas (as used in Nigeria) for example, 999,999,999.99

Comma: Decimal points represented with commas, every thousand separated by dots (as used in France) for example, 999.999.999,99

Scientific: for example, 3.2E3 representing 3.2×10^3 (3200) while 3.2E-3 represents 3.2×10^{-3} (0.0032)

Dollar: used to represent currency in US $

Custom *Currency*: Other currency

String: any combination of letters, symbols or numbers String and ordinal data make up **categorical** data. Data may also be "**Missing**" where

the variable does not apply to the subject (for example, prostate size in a female) or, for some reason, was not obtained *ab initio*. "Missing" data is excluded from most analyses.

Measures of variables in SPSS may be:

Nominal: data belonging to categories that are mutually exclusive for example, sex, marital status, occupation. They are usually represented by text (**String**) but may be represented by numbers for ease of analysis. However, these numbers have no quantitative importance (for example, male=1, female=2 or male=0, female=1). Data with only two (2) categories (examples: "Yes/No" questions, Sex) is called *binary* data. In SPSS, nominal variables are preceded by the icon 

Scale: Measurable data with fractions for example, weight, height, dose of medications. These values can be added, subtracted, divided or multiplied. They may also be negative or positive in value. In SPSS, scale variables are preceded by the icon 

Ordinal: data that can be ranked for example, position in class, degree of heart block, stage of cancer. One rank is higher than the next but there is

no way of dividing, multiplying, adding or subtracting (someone with Class II heart failure cannot be said to have twice as much heart failure as someone with Class I!). Ordinal data is represented by whole numbers in SPSS; it does not allow for decimal points or fractions. Ordinal variables in SPSS are preceded by the icon 

CREATING A DATASET

There are two options in creating a dataset (also called a database) in SPSS. You may choose to enter the data directly into SPSS or import the data from another source.

ENTERING DATA DIRECTLY INTO SPSS

Prior to entering data directly into SPSS, ensure that the following characteristics of each variable have been written out in a **code dictionary** based on your research tool/questionnaire.

Name	Label	Type (Width*)	Codes for Values	Codes for Missing
Serial_No	Serial Number	Numeric (3)	None	None permitted
Sex	Gender	Numeric (1)	0 = Male 1 = Female	9
Age	Age (Years)	Numeric (3)	None	999
MarStat	Marital Status	Numeric (1)	0 = Married 1 = Never Married 2 = Divorced 3 = Widowed	9

An example of a coding dictionary.

**Width=number of characters SPSS will accept in the cells (a width of "3" will truncate a value of 4567 to 456, causing confusion!)*

RULES FOR CODING

1. Variable names must start with a letter. They should NOT contain spaces or special characters (like (, *, &, %, $, #, @).
2. Each variable name must be unique (no duplication).
3. The Variable Label is what appears in the tables, graphs and other output after analysis. It can contain spaces and any other character. for example, “Age of Respondents (Months)”
4. As much as possible, use numbers to encode data. This allows for ease of analysis using SPSS.
5. When coding, the positive response in a predictor variable should have the higher numerical value for example, Smoker = 1, Non-smoker = 0 (or Smoker=2, Non-smoker = 1). If you are looking at male sex as risk factor, then code male = 1, Female =0. On the other hand, if female sex appears to be predictive, then code female=1 and male=0.
6. Similarly, when coding for outcome variables, the outcome we are interested in should have the higher numerical value for example, does excess television viewing increase the risk of in-patient death at the 30th day of admission? Here, the

outcome variable name could be **Death_30**, with "**Dead**" coded as 1 and "**Alive**" coded as 0 since we are primarily concerned with those that died.

7. SPSS is case-sensitive. If you are going to use text (for "**String**" variables), it is best to ensure that the data values are entirely in small-letters or capital letters. "Cat", "cAT", "CAT", "cat" and "CAt" would be recognized as three different values by SPSS.

8. The essence of coding is to simplify data entry. Avoid long codes.

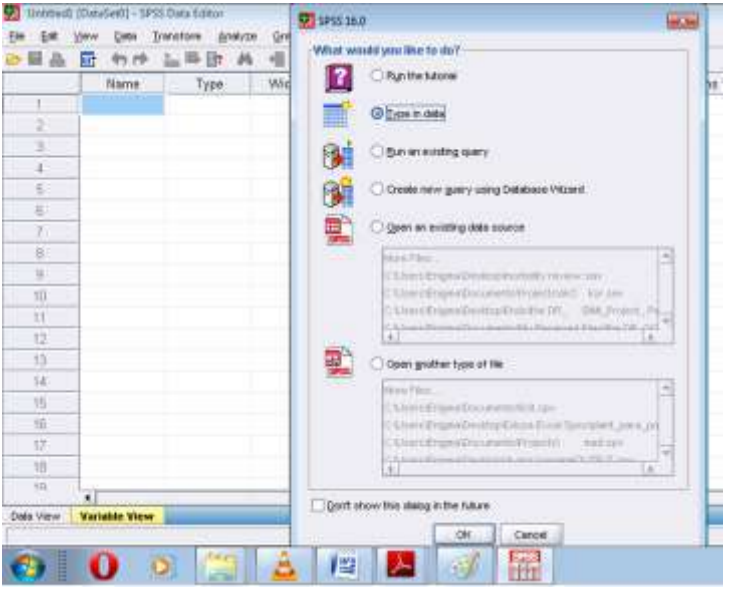
9. Select an impossible value as the code for missing values for example, negative or extremely large values for "Age".

10. Where data contains unmatched/unpaired groups you want to compare (cases vs. controls), a special *grouping* variable can be created (for example, variable name: Group1, Cases coded as 1, Controls as 0). This allows for easy comparison of such groups using hypothesis testing.

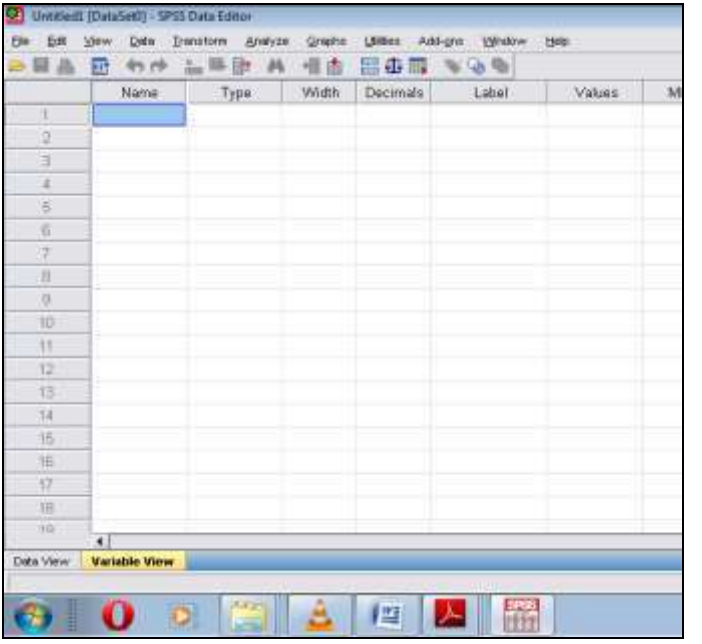
11. For paired/matched data, different columns should be created for each variable. An example is where blood pressure is measured for subjects and subsequently the measurement is repeated. The readings can occupy two columns as distinct variable (probably named BP1 and BP2)

THE SPSS INTERFACE

On running the SPSS program, typically a dialog window appears.

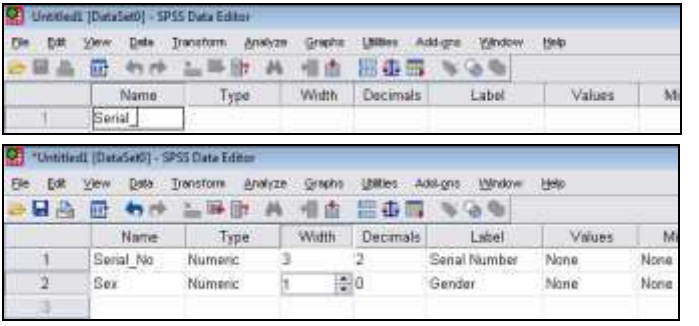


If you choose the “Type in data” option, a Data Editor interface follows.

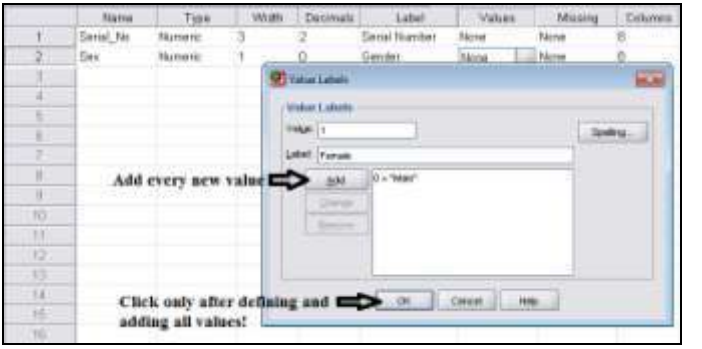


The **SPSS Data Editor** interface has two windows: a **Data View** and a **Variable View**. The next step is to transcribe the variables defined in the coding dictionary into the **Variable View**.

- Click on the tab marked "**Variable View**". Enter the variable names, label, type, decimal (number of decimals, default value is 2) and width (default value is 8).

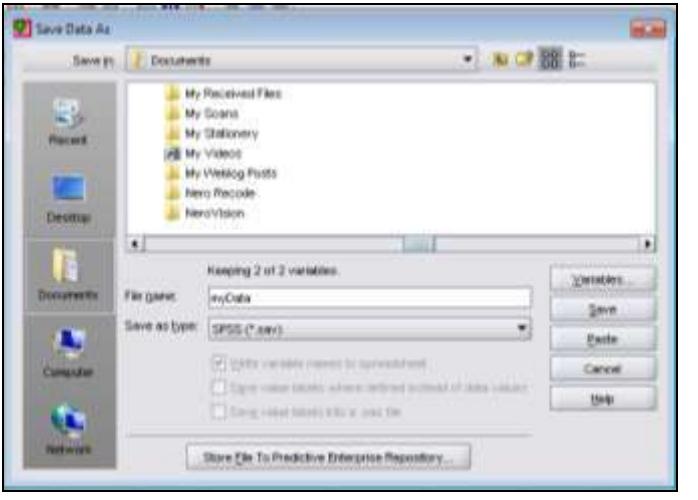


- Define the values for each variable.

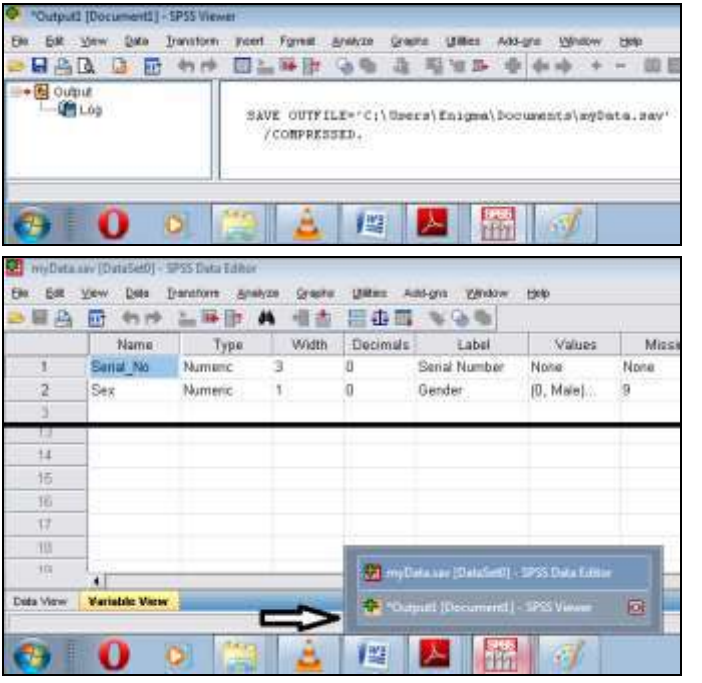


- Define missing values.

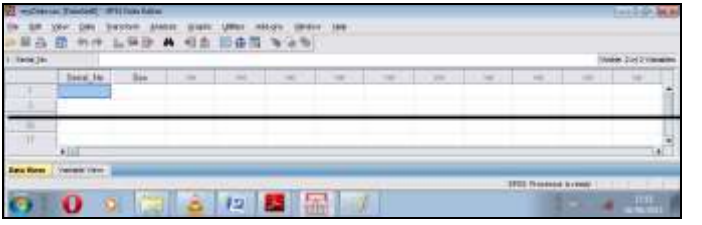
After defining your variables, save the data set. Do this frequently to prevent data loss. Go to "**File**" >>"**Save**", or simply use the  icon.



Datasets are saved in the *.sav* format. After saving the dataset, another SPSS window opens. This is the **SPSS Viewer**. This window shows a log of all SPSS actions and it is where all results of analysis will appear. Viewer logs can be saved as Output files in *.spv* format.



Return to the Data Editor Window and click the "**Data View**" tab at the bottom-left corner. The Data View is a spreadsheet with vertical columns ("**Variable**") and horizontal rows ("**Cases**") which intersect at cells. This is similar to the spreadsheet in Microsoft Excel® but there are a few important differences. Unlike Excel, cells can only contain values, not formulas. Hence, values within cells do not update automatically when changes are made to related cells. Unlike Excel where variable names are entered in the topmost row, in SPSS, this contains the data of the first subject. You can then type in the data into the appropriate cells using the codes.

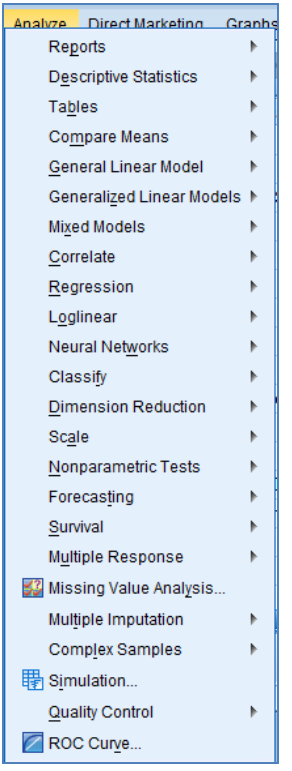


ELEMENTS OF THE SPSS INTERFACE

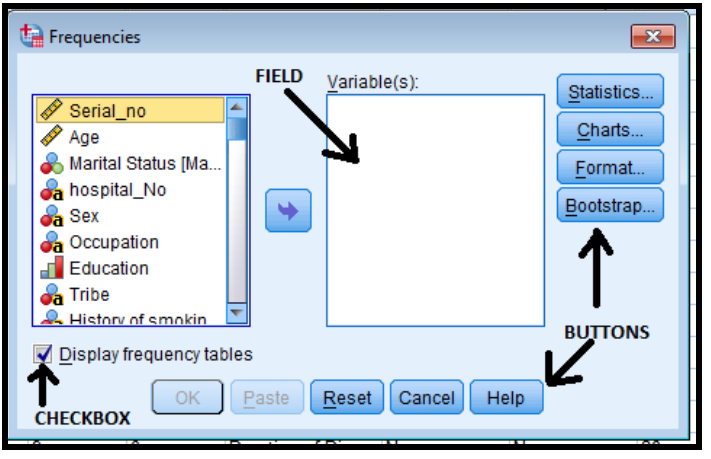
The visual elements we come across when using SPSS include:

1. Main Menu

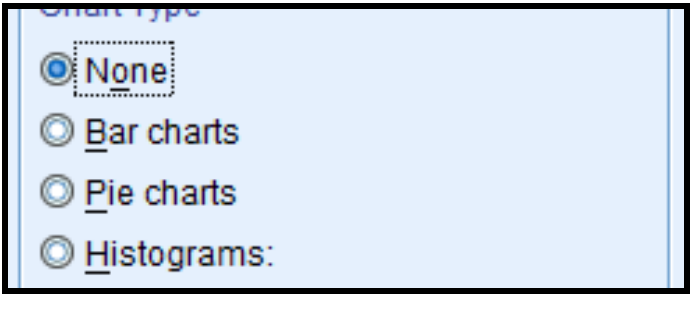
2. Drop Down Menus



3. Dialog Window



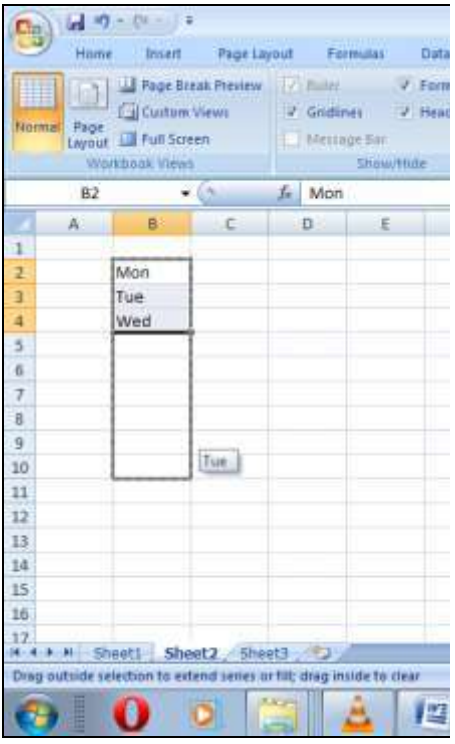
4. Radio Buttons

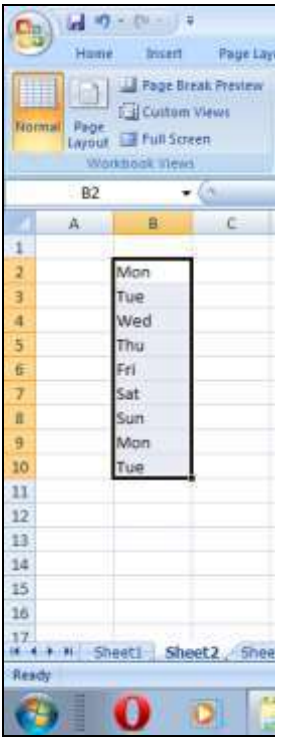


IMPORTING DATA FROM OTHER SOURCES

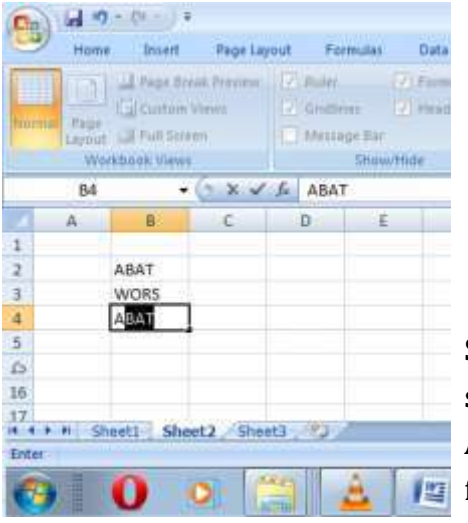
SPSS is able to use data stored in several other data formats. This book will restrict its scope to using data stored in the Microsoft Excel® format.

Entering data initially into MS Excel® has several advantages. The cells in the spreadsheet allow use of formula linking them to contents of other cells. The value of a cell containing a formula changes automatically when that of a linked cell is altered. Excel also has AutoFill, AutoSum and AutoCorrect functions which can be time-saving during data entry.

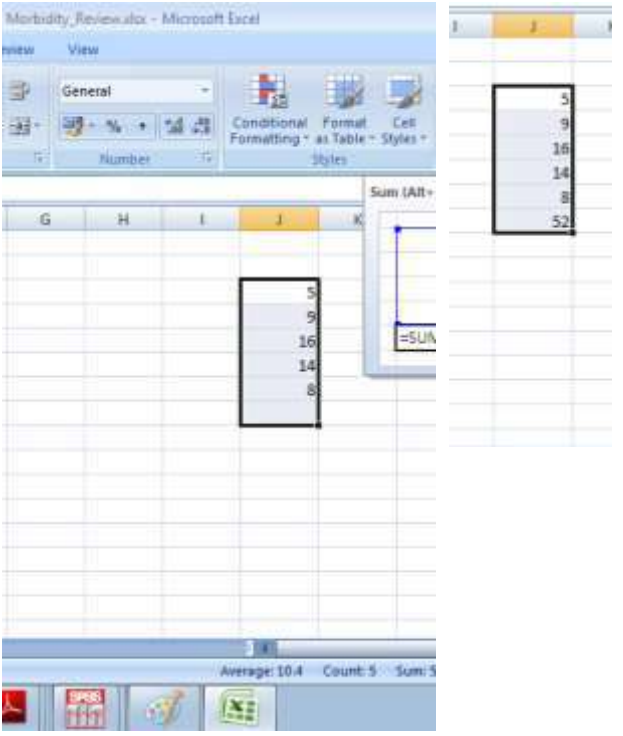




AutoFill in MS Excel®



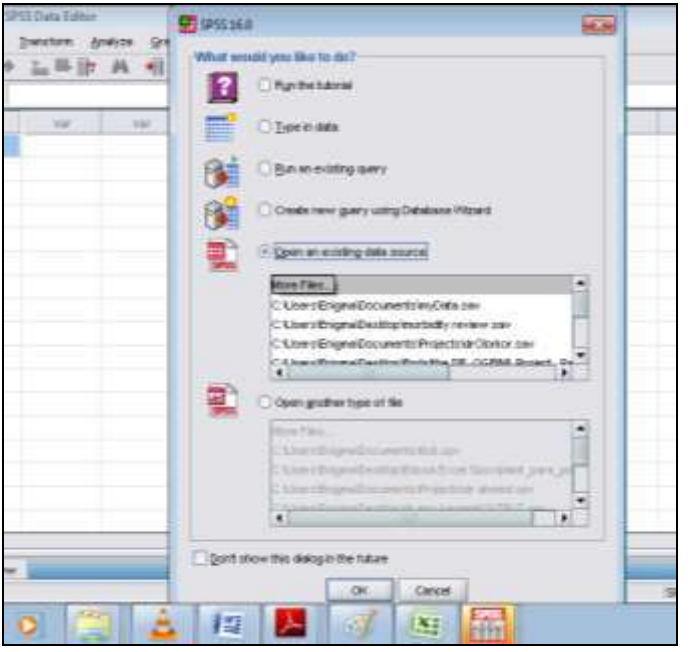
Screen capture showing AutoComplete function in MS Excel®



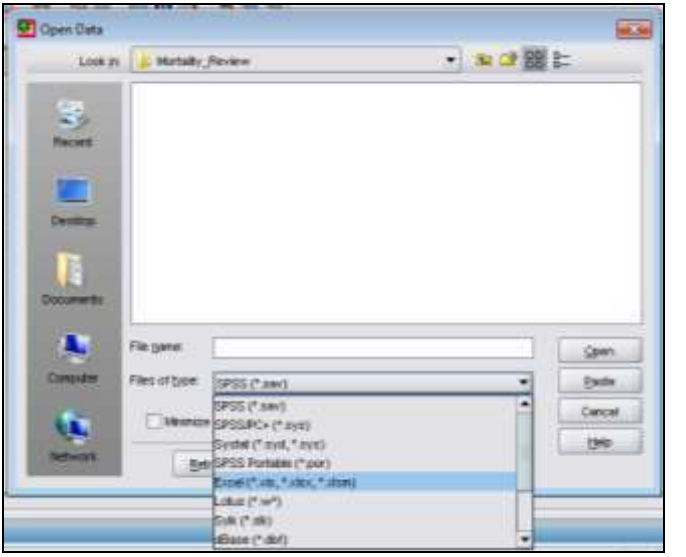
AutoSum function in MS Excel®

To import existing data from MS Excel®:

1. On first opening SPSS Select “Open an existing data source” >> “More Files…”

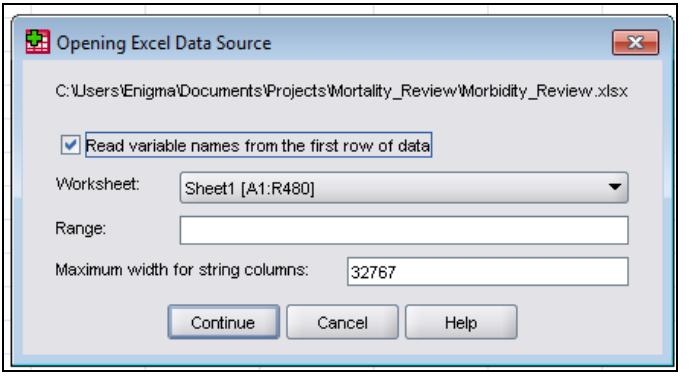


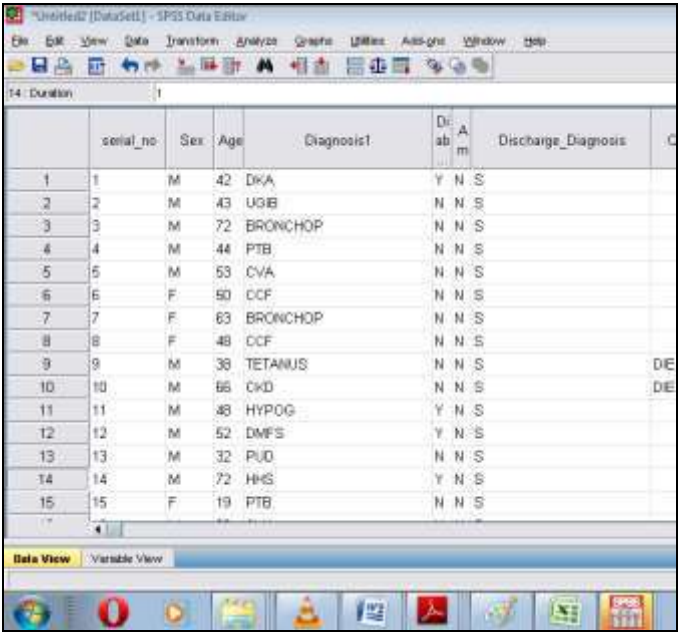
2. When the dialog window opens, navigate to the folder containing the desired Excel file. Click the drop-down menu labelled "**Files of type:**" and select "**Excel (*.xls, *.xlsx, *.xlsm**)". Double-click the desired file.




Open Data
Look in:
Mortality_Review
Mortality_Review.xlsx
~$Mortality_Review.xlsx
Recent
Desktop
Documents
Computer
Network
File name:
Mortality_Review.xlsx
Files of type:
Excel (*.xls, *.xlsx, *.xlsm)
Open
Paste
Cancel
Help
Minimize string widths based on observed values
Retrieve File From Predictive Enterprise Repository...

3. Another dialog box appears. Ensure you have ticked the checkbox on "**Read variable names from the first row of data**", then click "**Continue**"





	serial_no	Sex	Age	Diagnosis1	Di ab	A m	Discharge_Diagnosis	C
1	1	M	42	DKA	Y	N	S	
2	2	M	43	UGIB	N	N	S	
3	3	M	72	BRONCHOP	N	N	S	
4	4	M	44	PTB	N	N	S	
5	5	M	53	CVA	N	N	S	
6	6	F	50	CCF	N	N	S	
7	7	F	63	BRONCHOP	N	N	S	
8	8	F	48	CCF	N	N	S	
9	9	M	38	TETANUS	N	N	S	DIE
10	10	M	66	CKD	N	N	S	DIE
11	11	M	48	HYPOG	Y	N	S	
12	12	M	52	DMFS	Y	N	S	
13	13	M	32	PUD	N	N	S	
14	14	M	72	HHS	Y	N	S	
15	15	F	19	PTB	N	N	S	

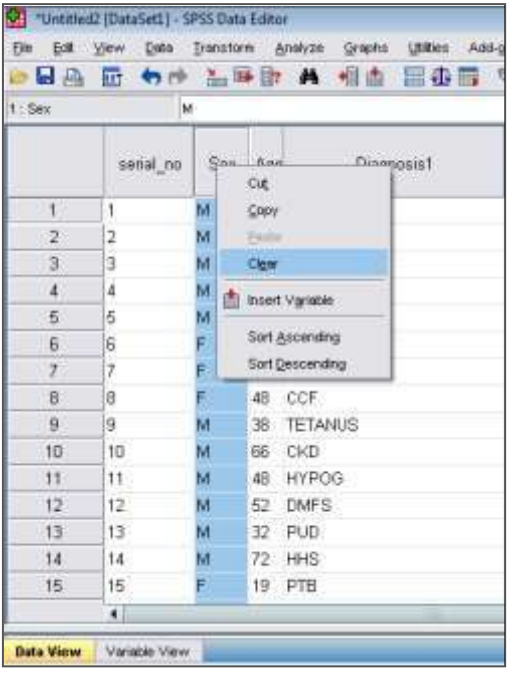
4. Save the dataset as earlier discussed. Note that if the variable name in Excel is invalid, it will be replaced with "**VAR00000XX**" where X is a number. If the variable name in Excel is a special SPSS command (for example, GET, FREQUENCY), the variable data will not be imported and an error message will be displayed in the viewer. Blank cells in Excel become "**System-Missing**" values in SPSS, displayed as full-stops.

5. Use Variable View to define the label, variable type, values codes and missing values code.

*An alternative method is to go to "**Files**" >> "**Open**" >> "**Data**". Then, go through steps 2 to 5 listed above***.**

OTHER TASKS IN DATA HANDLING IN SPSS

ALTERING DATA ENTRIES



To change a data entry, select the cell, type in the new value and press the **Enter** key. This overwrites the previous content of the cell.

To delete the content of a cell, select the cell and press the **Delete** button on your keyboard. Entire cases or variables can be deleted by selecting the grey headers and pressing the **Delete** key. Alternatively, right-click the header and select "**Clear**".

Any changes can be reversed using "**Undo**" icon  or holding **CTRL+Z.** Changes do not become permanent until the file is saved.

COPYING AND PASTING DATA

Data in single rows, single columns or across both can be copied or cut from one part of the worksheet and pasted in other parts of the spreadsheet. Tabular data from other programs (including Excel) can also be pasted into SPSS. Note that pasted data will overwrite the contents of the destination cells.

To copy or cut data,

1. Select source cells
2. Hold **CTRL+C** (to copy) or **CTRL+X** (to cut). Alternatively right-click and choose "**Copy**" or "**Cut**". A third option is to go to "**Edit**" > "**Copy**"/ "**Cut**"

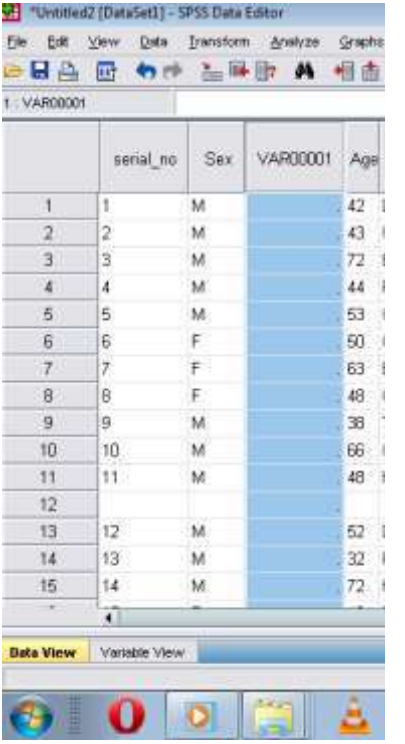
Select the cell at the upper left corner of target destination and hold **CTRL+V**. alternatively right-click and choose "**Paste**". A third option is to go to "**Edit**" > "**Paste**"

INSERTING NEW CASES/ VARIABLES

To insert a new variable (i.e. a new column), select the column that will be on the right of the new variable. Right-click and select "**Insert Variable**". Alternatively go to "**Edit**" >> "**Insert Variable**". The third option is to click on the icon  .

To insert a new case (i.e. row), select the row that will be below the new case. Right-click and select "**Insert Case**". Alternatively go to "**Edit**" >> "**Insert Case**". The third option is to click on the icon  .

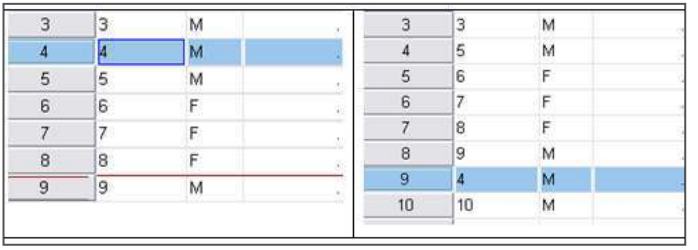

*Untitled2 [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze
1 : Age
42
serial_no
Sex
Age
Cut
Copy
Clear
Insert V
Sort As
Sort De
1 1 M 4
2 2 M 4
3 3 M 7
4 4 M 4
5 5 M 5
6 6 F 5
7 7 F 6
8 8 F 48 CCF
9 9 M 38 TETANI
10 10 M 66 CKD
11 11 M 48 HYPOC
12
13 12 M 52 DMFS
14 13 M 32 PUD
15 14 M 72 HHS
Data View
Variable View


*Untitled2 [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs
1 : VAR00001
serial_no
Sex
VAR00001
Age
1 1 M 42
2 2 M 43
3 3 M 72
4 4 M 44
5 5 M 53
6 6 F 50
7 7 F 63
8 8 F 48
9 9 M 38
10 10 M 66
11 11 M 48
12
13 12 M 52
14 13 M 32
15 14 M 72
Data View
Variable View

MOVING DATA

To move a variable, *left*-click the column header and *drag* to the target location using the red vertical line as the marker.

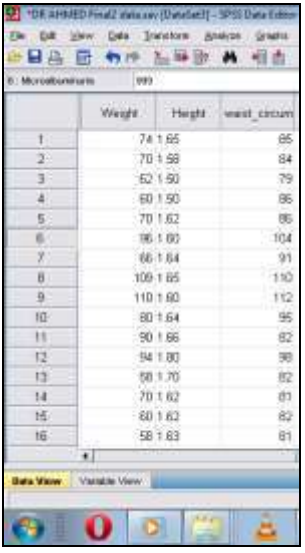
To move a case, left the row header and drag similarly using the horizontal red line as a marker.



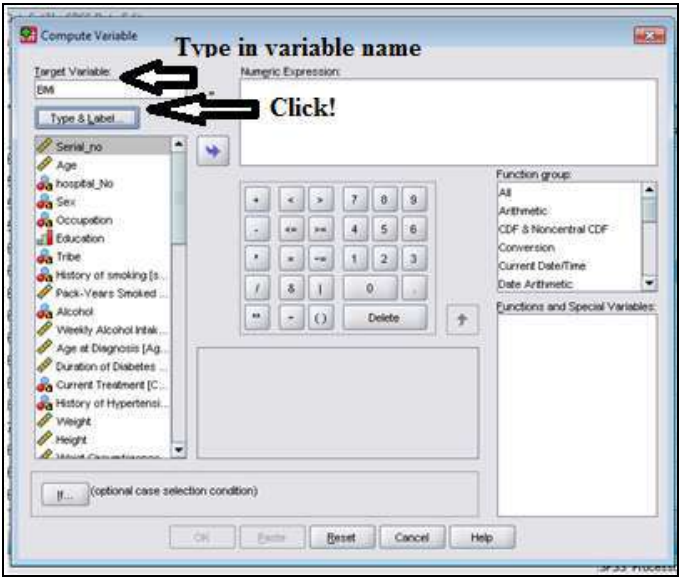
COMPUTING A NEW VARIABLE

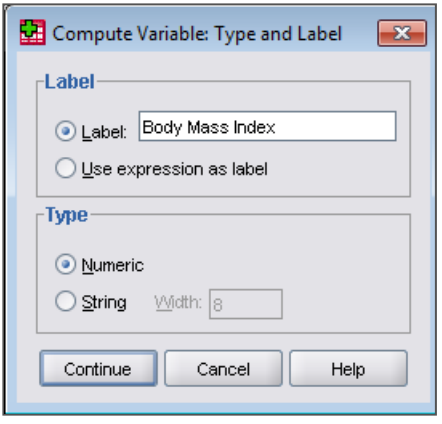
The aim here is to create a new numeric variable from previously existing numerical variables. In this example, we shall create a new variable "**BMi**" (Label = Body Mass Index) from two previously defined variables called "**Weight**" and "**Height**" using the formula BMi = Weight/Height^2.

To do this, go to "**Transform**" >> "**Compute Variable…**" >> Type in Variable name in the text field "**Target Variable**"


Weight
Height
waist_circum
Data View
Variable View


Transform
Compute Variable...
Count Values within Cases...
Recode into Same Variables...
Recode into Different Variables...
Automatic Recode...
Visual Binning...
Optimal Binning...
Rank Cases...
Date and Time Wizard...
Create Time Series...
Replace Missing Values...
Random Number Generators...
Data View
Variable View


Compute Variable
Type in variable name
Target Variable:
BMI
Type & Label
Click!
Numeric Expression:
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Serial_no
Age
hospital_No
Sex
Occupation
Education
Tribe
Alcohol
Height
Delete
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help


Compute Variable: Type and Label
Label
Label:
Body Mass Index
Use expression as label
Type
Numeric
String
Width:
8
Continue
Cancel
Help


Compute Variable
Target Variable:
BMI
Type & Label...
Numeric Expression:
Height / (Height ** 2)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Delete
(optional case selection condition)
OK
Paste
Reset
Cancel
Help


1 : Serial_no 1
BMI
1 27.18
2 28.04
3 27.56
4 26.67
5 26.67
6 37.50
7 24.64
8 40.04
9 42.97
10 29.74
11 32.66
12 29.01
13 20.07
14 26.48
15 22.66
16 21.83
17 26.09
Data View
Variable View

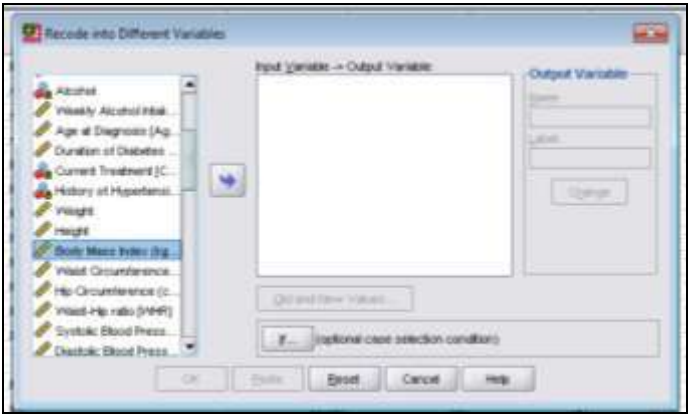
In the field labelled "**Numeric Expression**", type in the formula using the BODMAS rule. Note that "**" implies "raised to the power of". For complex formulas, use brackets freely (just be sure you close all of them). The formulas also allow for use of conditional (logic) operators such as OR ("**|**"), AND ("**&**") and NOT (**~=**). Unlike Excel, subsequent alterations in either "**Weight**" or "**Height**" in SPSS will not affect the values in "**BMi**" unless it is re-computed.

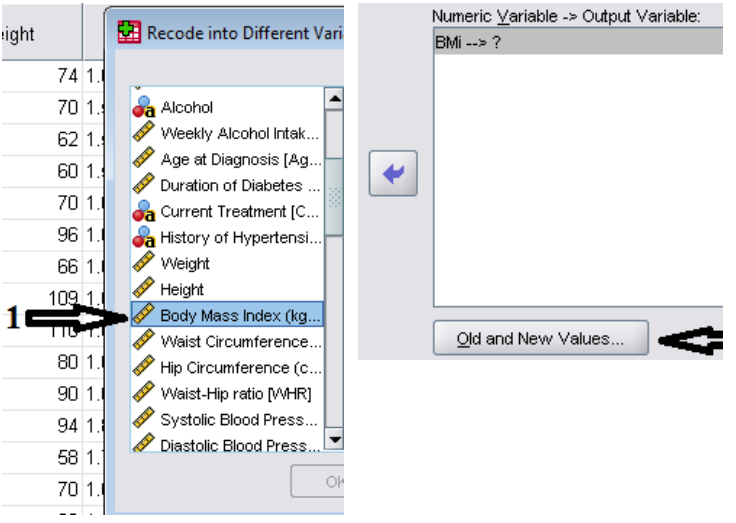
RECODING A VARIABLE

The purpose of this is to transform one variable into a new variable ("**Recode into Different Variable**") or to replace its values ("**Recode into same variable**") using certain criteria. Please note that "**Recoding into** *same* **variable**" causes the previous values to be discarded leading to possible loss of important data.

In this example, our newly generated variable "BMi" will be recoded into a different variable called "**DegObese**", for "**Degree of Obesity**". The values will be 1 (BMi <18.5), 2 (BMi between 18.5 and 25.9), 3 (BMi between 26 and 29.9) and 4 (BMI of 30 and above).

Go to "**Transform**" >> "**Recode into different variable**". The following dialog window appears. From to the list to your left, select the variable to be recoded, and transfer it to the field marked "**Input variable→Output Variable**". Type in the desired name and label of the output variable in the text fields under "**Output Variable**", click "**Change**" and "**Old and New Values…**"


Recode into Different Variables
Input Variable -> Output Variable:
Output Variable:
Old and New Values...
OK
Paste
Reset
Cancel
Help


Recode into Different Vari
Alcohol
Weekly Alcohol Intak...
Age at Diagnosis [Ag...
Duration of Diabetes ...
Current Treatment [C...
History of Hypertensi...
Weight
Height
Body Mass Index (kg...
Waist Circumference...
Hip Circumference (c...
Waist-Hip ratio [WHR]
Systolic Blood Press...
Diastolic Blood Press...
1
Numeric Variable -> Output Variable:
BMi --> ?
Old and New Values...

At this stage, another dialog window opens for the actual recoding. To recode **BMi** <18.5 into **DegObese**=1. (Note the use of *18.4* instead of 18.5. This is because typing 18.5 would cause subjects with BMI of 18.5 to be included and wrongly categorized as "1" instead of "2".)

The "**New Value**" field, by default, accepts only numbers. If the desired output variable is "**String**" (for example, **DegObese**= "**Underweight**" instead of 1), tick the first checkbox to allow typing of text.

If numbers were imported as **String** (probably unintentionally), then they can be automatically recoded to numeric data by ticking the second checkbox.

Thanks for reading these sample chapters.

To read more chapters, please buy the eBook from

www.leanpub.com/quickeasystat

OR

Get the Print Book from Amazon