COMP2322: INTRODUCTION TO HEALTH INFORMATICS

Lab #1

Lab 1: Statistical Package for the Social Science (SPSS)

1.1 Learning Objectives:

- To be familiar with the fundamentals of the SPSS package.
- To know how to insert data using data editor.
- To use descriptive statistics to compute means maximum, and minimum for the whole data and for specific group of data.
- To create new variables using the compute option and assign mathematical values to new variables.
- To draw charts.
- To find frequencies for groups of data.
- Become familiar with the essentials of Google forms.

1.2 SPSS Introduction

SPSS stands for *Statistical Package for Social Sciences.* It was produced by IBM Corporation. According to its name, the main usage is for statistical operations. It provides an easy way to enter data, and then apply some essential dataset functionalities, such as specifying data types, sorting, splitting, generating statistical reports, and much more of such important functions.

Before you begin using the program, get a good impression by being aware of usage scenarios and problems that SPSS provides solutions. Here is a list of such scenarios and problems where SPSS is considered as a good choice to address them:

1. In your graduation project, you have collected a set of questioners from a group of people, and you need to analyze them using descriptive analysis (Minimum, Maximum, Range, Variance, Standard Deviation, mean, *etc.*). Furthermore, you want to generate reports and charts to prove your theory. SPSS provides an easy and convenient way to enter your questioners and apply the mentioned operations on them.

2. Fraud detection and prevention. Give SPSS a set of users profiles correlated with their fraud history, a bank security analyst would be able to detect any potential fraud of any new bank transaction. Thus, giving the bank the ability to take proper actions before damages and loss occur.

3. Medical researchers can use SPSS to convert the data they collect into useful information which helps in viewing of data from a different perspective which may lead to new theories and conclusions.

1.3   SPSS main components

You can find the program in *start menu* or on *desktop*. It has this icon

When you open (run) the program, the following dialog shows up:
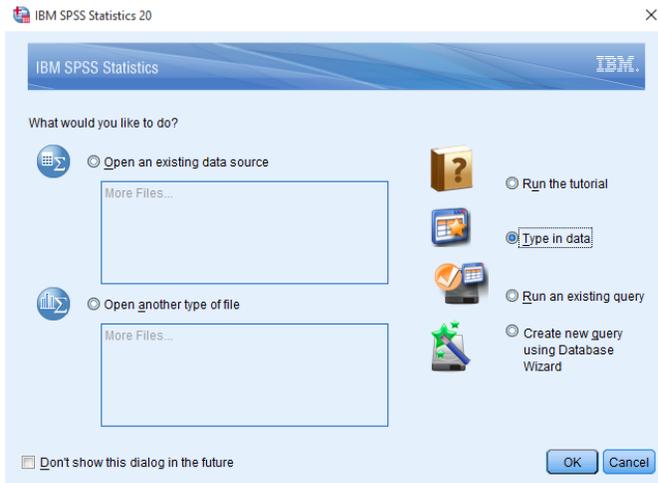
Choose **Type in data** which allows you to enter data manually

Then the main window will show up. As you see in the next picture, the window has two main tabs in the bottom.

Data view and Variable view
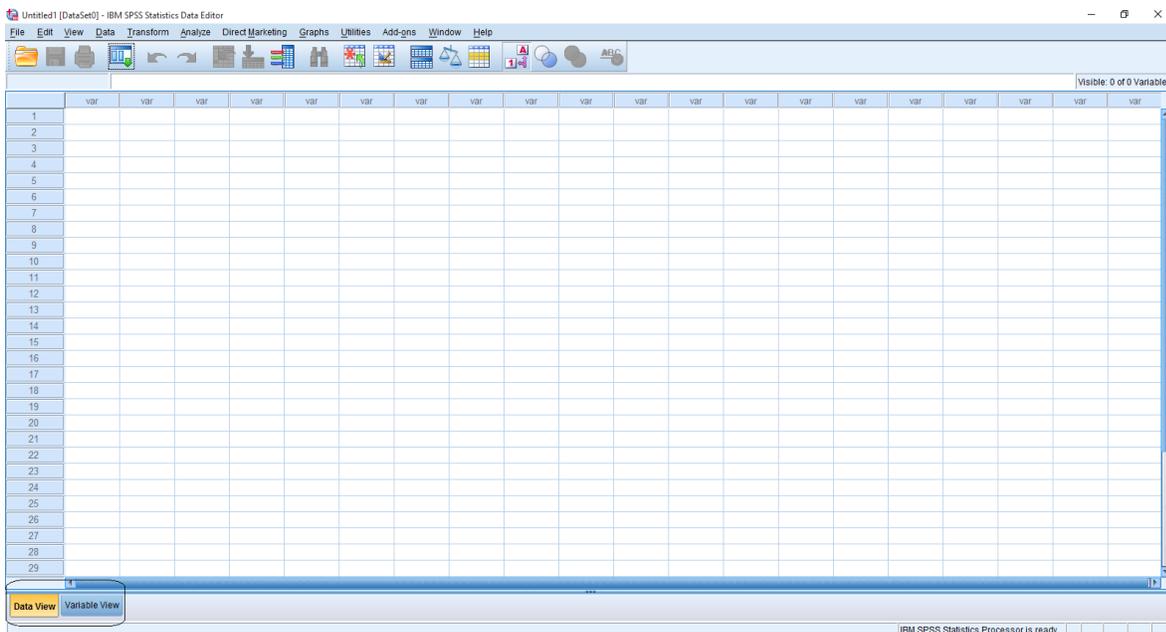
With **Variable view** you can specify the variables used in your dataset, with their attributes. While in **Data view** you can enter the data itself.

To understand them well: Assume that you have gathered the following survey from 10 persons

```
┌─────────────────────────────────────────────────────────────────────┐
│                          Questioner                                   │
│                                                                       │
│  Questionnaire Code  ───────────                                      │
│                                                                       │
│  Age ───────────                                                      │
│                                                                       │
│  Gender      □ Female              □ Male                             │
│                                                                       │
│  Weight ──────────────────                                            │
│                                                                       │
│  Height: ──────────────────                                           │
│                                                                       │
│                                                                       │
└─────────────────────────────────────────────────────────────────────┘
```

Age, Gender, Weight and Height represent the variables. Each has its own attributes. At first, we define these variables in Variable view, then we see them as columns in Data view. While a row in Data view represent an instance or case, which according to the above example represents a person. Therefore, in Data view we will have 4 columns which represent the variables (which are defined in variables view) and 10 rows which represent the persons (the gathered surveys).

Variable view Elements

Variable View is where variables are defined by assigning variable names and specifying the attributes as the following:

Attribute	Default Value
Name- The name of the variable defined. . It is recommended to assign a brief and meaningful name (e.g., "Name," "Gender," and "GPA").	none
Type- The variable type determines how the cases are entered. Generally, text based characters are of "String" type and number-based characters are of "Numeric" type. etc	numeric
Width- Total numbers of characters	8
Decimals- Number of characters beyond the decimal point	2
Label-Value labels allow users to describe what the variable name stands for. For example, if a variable has been defined as "Fav," most likely others may not know what it stands for.	none
Value- allows you to provide labels for the various levels of a variable	none
Missing values- enables you to designate certain scores as missing	none
Columns- the width of the column used to fill this variable but that its effect will appear in the data view only.	8
Align- allows you to determine the alignment of your column	right
Measure- allows you to determine the kind of scale for the particular variable	scale

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										

The Data Editor Menus:

- File: To open an existing file, to read data (from a text file or from an application (for example Excel) into Data Editor Window, to save the data file, and to exit SPSS windows.

- Edit: To make changes like copy, cut and paste to the Data Editor Window.

- Data: To define variables, insert variables or cases, sort cases, merge file, split files, select cases and use a variable to weight cases

- Transform: To do some computations on variables, to create new variables from existing ones or record old variables.

- Analyze: To do statistical analysis, from descriptive statistics to testing hypothesis.

- Graphs: To obtain high resolution plots and graphs, this can be edited in Chart Editor Window.

- Utilities: To run script and to display information on the contents of SPSS data files.

- Windows: To move to any open window or to see which window is active. The window with a check mark is the active one.

- Help: To get help on topics in SPSS and to ask the statistics coach some questions.

Defining the variables

Based on the given Questioner, we are going to do the following in the variable view:

1. In the space **Name**, type the variable name, the first character must be alphabetic; the remaining characters can be alphabetic, and /or numeric, (but the underscore" _ "or the period". "can appear in the name) and no spaces can appear in the name, so for the first variable, let's type "Qcode" (Questioner Code).

2. Under the **Type** column, click the ellipses button ... . The Variable Type dialog box opens. From this figure choose the type that is suitable to   your variable, in our example the type **String** would be suitable type for the Questioner Code variable. Then press **ok**.
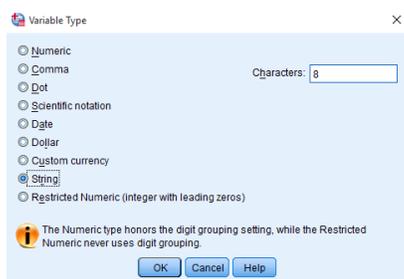
3. When choosing the string type, the maximum **width** of this string must be identified from the column width.

4. In this example the variable "Qcode" is String, so it's never has **Decimal** points

5. Click on **Label** and type in "The Questioner code ". The Label slot usually contains a clear description for the variable to help the user understand what does this variable exactly mean.

6. From the **Align** column, the user choose the alignment of the inserted data; Left, Right or Center.

7. From the **columns** column, the user can modify the width of the column used to fill this variable but that its effect will appear in the data view only.

8. The rest of variables will be defined in the same manner but keep in mind the following points about them:

   a. The second variable would be age which is **Numeric** type and zero Decimals
   b. The third variable would be Gender. This variable is of 2 options (Male and Female) which is string. But for statistical reasons. We define their type to be Numeric, then from the values slot numeric values are assigned to each non-numeric values. When clicking the value slot, the following window will appear. In the Value slot the numeric value is added while in the Value Label slot the label that represents the numeric value. Then the Add button must be selected. In the Data View the Value Label icon 🏷️ is used to represent the labels assigned to each numeric value. Here is how value labels of Gender variable would be:
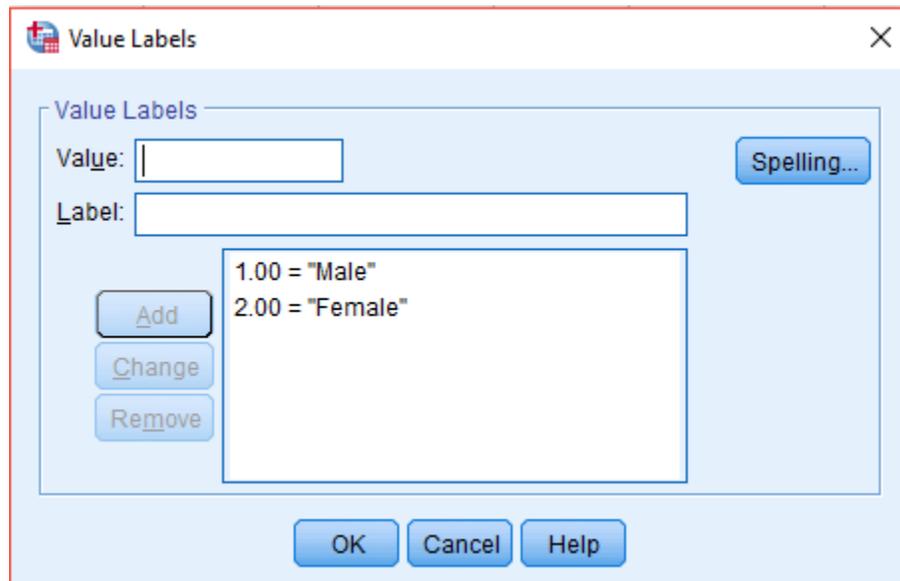
After defining the variables, the **Variable View** will be as the figure below:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Qcode	String	8	0	Qcode Value	None	None	8	≡ Center	Nominal	Input
2	Age	Numeric	8	2	Age Value	None	None	8	≡ Center	Nominal	Input
3	Gender	Numeric	8	2	Gender Value	{1.00, Male}...	None	8	≡ Center	Nominal	Input
4	Height	Numeric	8	2	Height Value	None	None	8	≡ Center	Scale	Input
5	Weight	Numeric	8	3	Weight Value	None	None	8	≡ Center	Scale	Input
6											
7											

Input Data to Data view

After you finish defining the variables, you can switch to Data View to enter the cases (Quested answers of the questioner). **Weight in KG and Height in M**, Let's enter the following data:

	Qcode	Age	Gender	Height	Weight
1	100	25.00	Male	1.80	84.000
2	102	21.00	Male	1.70	72.000
3	103	30.00	Male	1.66	90.000
4	105	27.00	Male	1.90	77.000
5	106	12.00	Male	1.30	40.000
6	109	40.00	Male	1.68	60.000
7	110	34.00	Male	1.58	58.000
8	101	22.00	Female	1.55	60.000
9	104	16.00	Female	1.55	50.000
10	107	29.00	Female	1.67	57.000
11	108	18.00	Female	1.67	47.000
12					
13					

Insert new cases and variables between existing cases or variables:

In the **Data View**, select any cell in the row below the position where you wish to insert the new case or select any cell in the variable to the right of the position where you wish to insert a new variable. From the **Data** menu, click on **insert case** or **insert variable.**

Or use the Icon ⬛ To Insert a case, ⬛ to insert a variable.

Sort Cases

1. From the **Data** Menu, choose " **Sort cases**"

2. You can choose which variables by which you will sort your data. Simply select the desired variables and click on the arrow in the center of the screen to move it to the **Sort by** field. Press **OK**

3. Descriptive Analysis

After data has been entered, users may analyze the data by using descriptive statistics.

Descriptive statistics are the most commonly used statistics for summarizing data **frequency** or **measures of central tendency (mean, median, mode and standard deviation)**

1. **Frequencies**

    The frequencies procedure provides statistics and graphical displays that are useful for describing many types of variables. For a first look at your data, the Frequencies procedure is a good place to start. In our example, we can use frequencies to know the percentage of males and females in our sample. To find this following the coming steps:

    - Click an **Analyze**, then **Descriptive Statistics**, then **Frequencies**

    - Highlight the variable for which you the frequencies are needed and move it into the variable column (Lets choose gender).

    - Click on **Charts** and select the type of the chart required, you can choose either Bar, Pie, or Histogram Chart. Let's choose Pie chart. Then presses continue then **OK**. An output window will show up as the following:
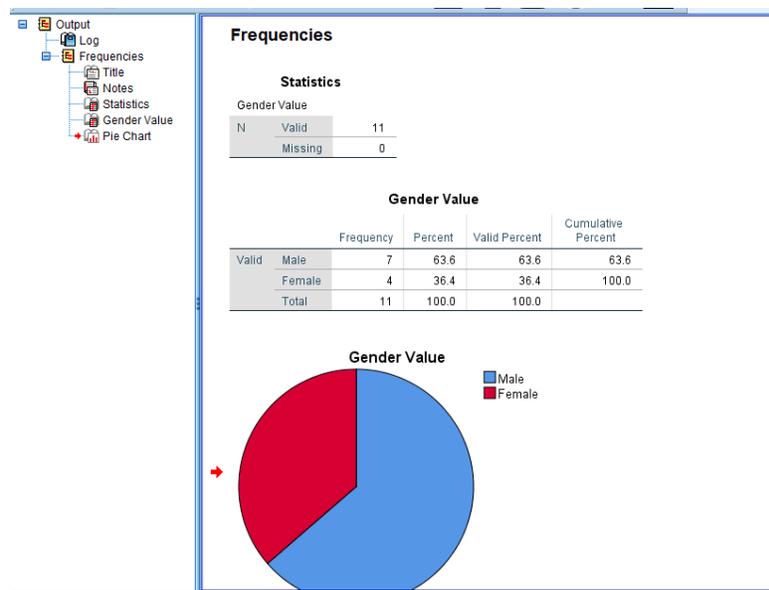
Computing Means and Standard Deviation

Once data is inserted, performing statistical procedure is relatively simple. The mean and standard deviation for each of the variables from Salary study will be computed. Mean, minimum, maximum, and standard deviations will be computed separately for males and females

Computing the Mean and Standard Deviation for the whole data

- Click on **Analyze**, then **Descriptive Statistics**, then **Descriptive.**
- Highlight each of the variables for which you are interested in computing descriptive statistics(e.g. Age, Weight, Height) and move them into the **variable(s)** column. Note that you can move all of these variables over at one time by clicking and dragging over the items that you want to select.
- Click on **Options** and select the desired statistic. At minimum, you should select the Mean, maximum, minimum, and Std. Deviation. Now, click on **Continue** and then **OK**.
- The output should look like that below. You may have fewer or more statistics depending on your selection in the Options menu. Note that the variables labels appear on the printout.
- Output form the analysis: If you would like a hard copy of this output, you can **print** it by clicking on the print icon on the tool bar. Also you can **save** the output by clicking on the disk icon. Note that SPSS automatically adds the ".**spv**" suffix to your output file name.

The following result will be added to the Output window when apply a descriptive analysis on Salary and Age.
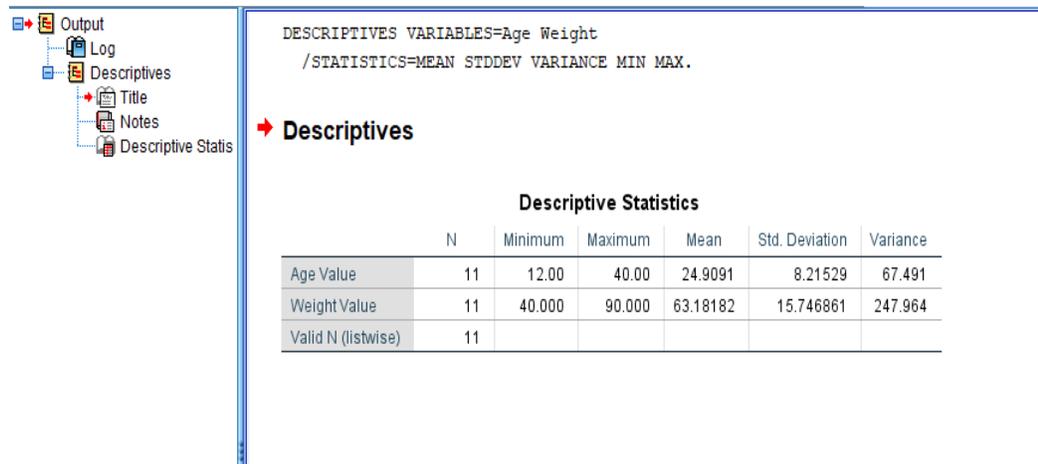
```
DESCRIPTIVES VARIABLES=Age Weight
  /STATISTICS=MEAN STDDEV VARIANCE MIN MAX.
```

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Age Value	11	12.00	40.00	24.9091	8.21529	67.491
Weight Value	11	40.000	90.000	63.18182	15.746861	247.964
Valid N (listwise)	11					

Computing Means and Standard Deviations for males and females separately

It is important to calculate the Mean, Maximum, Minimum and Standard Deviations for groups of data. In our example, it might be important to find them when quested are Males and Females. To do this, the following steps must be followed:

- First click on **Data** and then **Split File**. This allows you to split the file according to a particular variable and conduct separate analysis for each level of the variable.
- Next select **Organize output by groups** and move Gender from the variable list to the **Group Based on** list, then click on **OK**.
- Now, click on **Analyze**, **Descriptive Statistics**, and **Descriptive**. Select the desired variable(s) and Options.
- Note in your output that two sets of data summaries are presented one for females and the other for males.

The following is how the previous analysis would be after splitting the file based on gender variable (male and female):

→ **Descriptives**

Gender Value = Male

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Age Value	7	12.00	40.00	27.0000	9.05539	82.000
Weight Value	7	40.000	90.000	68.71429	17.230925	296.905
Valid N (listwise)	7					

a. Gender Value = Male

Gender Value = Female

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Age Value	4	16.00	29.00	21.2500	5.73730	32.917
Weight Value	4	47.000	60.000	53.50000	6.027714	36.333
Valid N (listwise)	4					

a. Gender Value = Female

4. Computing a New variable from existing ones

- Assume, a new variable "Body Mass Index" is need in our research. To create this variable and compute its numeric value the following steps must be followed:

  1. Click **Transform** and then **Compute Variable.**
  2. Type the name of the variable that you wish to create (in this case, we will name it "BMI") (Weight) / (Height * Height), in the **Target Variable** field. In the **Numeric Expression** field, the required mathematical expression is written. The operations listed on the bottom of this screen can be used in calculations.
  3. On the **Type &label** field .Longer label for the new variable can be added. Then, press **continue** finally press **OK.** The new variable will appear in the data file

This how the compute variable dialog would be:
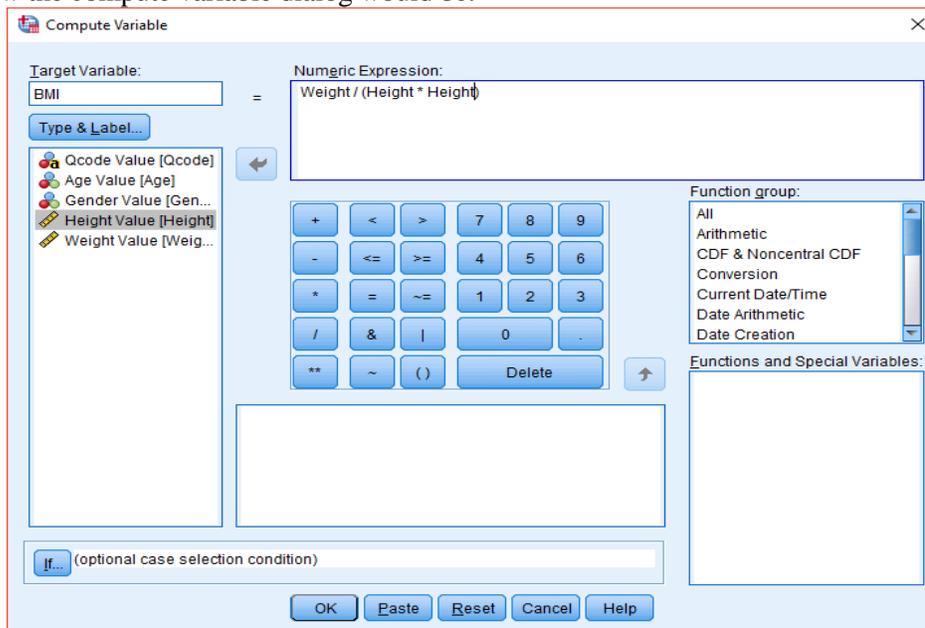
- Assume, a new variable "BMI_Group" is need in our research. To create this variable and compute its numeric value the following steps must be followed:

BMI Groups:

- Underweight: BMI is less than 18.5
- Normal weight: BMI is 18.5 to 24.9
- Overweight: BMI is 25 to 29.9
- Obese: BMI is 30 or more

"If/Then" variables computations

Sometimes you may wish to include only data that fits a certain criteria in your new variable:

a) In compute variable dialog, click on the "**If...** " button.
b) In the "If **Case** "box, click on "**include if case satisfies conditions** ".
c) Next, write an expression that suits your purpose using the calculator and the lists of variables and functions. For example, we will compute "Overweight" IF BMI >= 25 & BMI < 30, after this information has been entered, click "Continue "to return the familiar "Compute variable "window, where you will enter the computation for the new variable.
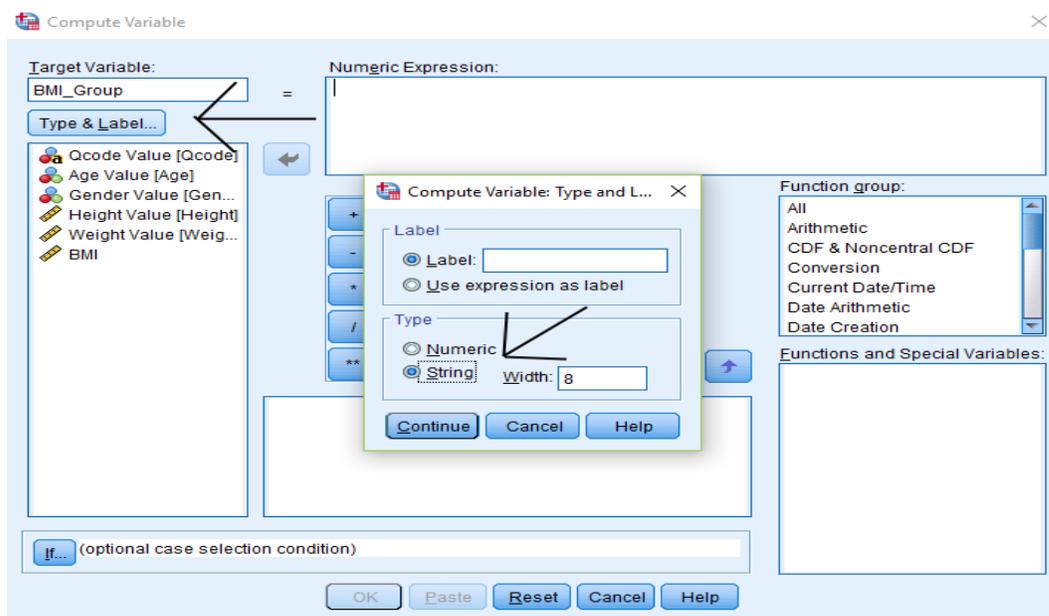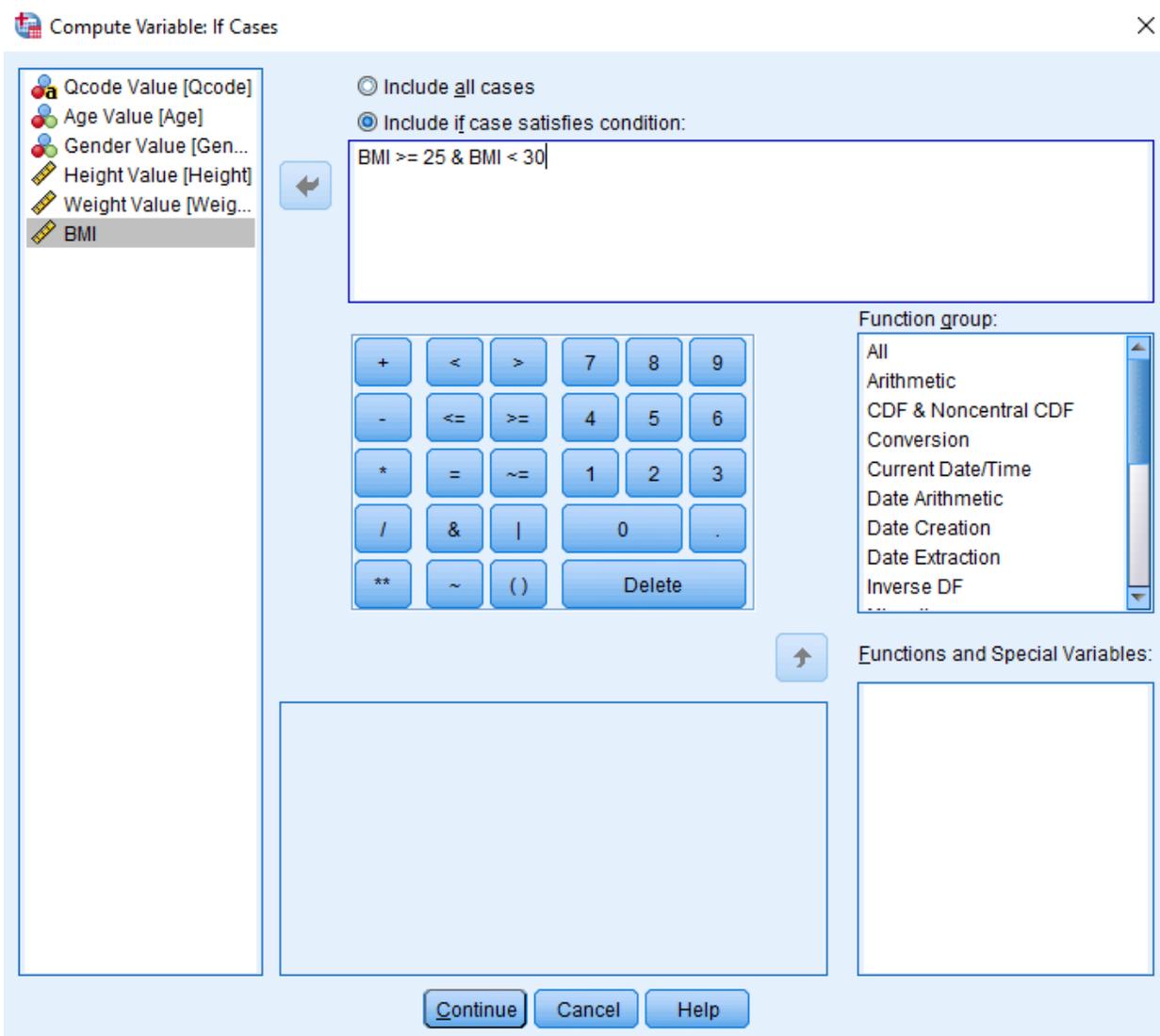
d) Click Continue.

5. Some Questions to solve for the above exercise

1. The average age of overweight people

2. The average age of overweight females and also of males

3. The average age of tall people

4. The average age of tall females and also of males

5. Find how many females are overweight

6. The average height of overweight people

7. The average height of overweight females

8. The average height of overweight males

9. Whom are more overweight, the number of males or females?

10. The average weight of tall people

11. The average weight of short people

12. Whom are more overweight, short or tall people?

13. The average weight of tall females and short females

14. The average weight of tall males and short males

Hint: Use Split File & Frequencies Analysis to solve most of these Questions.

Arithmetic and Statistical functions:

The following are some arithmetic and statistical functions that can be used in the compute window to calculate the mathematical expression for the new variable.

Function Name	Usage
ABS(expr)	Absolute value (-3.7 will become 3.7 ,for instance)
RND(expr)	Rounding
TRUNC(expr)	Truncation (decimal values are simply cut off)
MOD(expr,divid)	Remainder of the first argument (expr) when divided by the second
SQRT(expr)	Square root of expression
SUM (expr1, expr2.....)	Sum of the expression enclosed in the parentheses
MEAN (expr1, expr2....)	Mean of the expressions enclosed in the parentheses
MIN (expr1, expr2.....)	Minimum value of the expressions enclosed in the parentheses
MAX (expr1, expr2....)	Maximum value of the expressions enclosed in the parentheses

5. Creating Simple Bar Graph

Data are often best communicated in visual form. SPSS has a variety of graphs that can be used to display data. A simple bar graph of the data can be drown by following the steps listed below

a) Click on **Graphs**, then **Legacy Dialogs** and choose **Bar.**
b) In the Bar Charts menu, click on **Simple** and then **Define**.
c) In the Define Simple Bar menu, first select **Other Statistics** (this enables you to plot the means of your groups) and then select your dependent variable measure (" Age") and move it into the variable field. The default will be to compute the mean of the variable, but you can modify this (to select the median, mode, or other summary statistics) by clicking Change statistic. Given that we wish to calculate the mean, we will leave it as is.
d) Now move the independent variable ("BMI") into the **Category Axis** field.
e) If you wish to create a title of your figure, click on **Titles**, and type a title.
f) When you are ready, click on **OK**, and compare your figure with the one below

THE END ☺