

Basic Statistical Analysis Using SPSS

Awol Seid
St. Paul's Hospital Millennium Medical College
Addis Ababa, Ethiopia

© October 2019

Contents

1	Review of Basic Statistics	1
1.1	Common Terms	1
1.2	Classification of Statistics	2
1.2.1	Descriptive Statistics	2
1.2.2	Inferential Statistics	2
1.3	Variables	3
1.3.1	Types of Variables: Quantitative vs Qualitative	3
1.3.2	Measurement Scales: Nominal, Ordinal, Interval and Ratio	4
1.3.3	Role of Variables: Dependent vs Independent	6
1.4	Rules of Coding Variables	8
1.4.1	Coding Binary Variables	8
1.4.2	Coding Nominal Variables	8
1.4.3	Coding Ordinal Variables	9
2	Introduction to SPSS	10
2.1	Exploring SPSS	10
2.2	The Data Editor Window	11
2.2.1	The Variables View Sheet	12
2.2.2	The Data View Sheet	13
2.3	Creating a New Dataset in SPSS	13
2.3.1	STEP 1: Defining Variables in the Variables View	14
2.3.2	STEP 2: Entering the Data in the Data View	15
2.3.3	STEP 3: Saving the Data	17
2.4	Accessing Data Files	17
2.4.1	Opening an SPSS Data File	17
2.4.2	Importing and Exporting Data	18
3	Basic Data Management	20
3.1	Preliminary Data Analysis	21
3.2	Manipulating Data	23
3.2.1	Protecting the Original Data	23
3.2.2	Inserting and Deleting Cases	23
3.2.3	Inserting and Deleting Variables	24
3.2.4	Sorting Cases	25
3.2.5	Sorting Variables	26
3.2.6	Selecting Cases	26

3.2.7	Creating New Variable	29
3.2.8	Recoding a String Variable	31
3.2.9	Recoding a Categorical Variable	32
3.2.10	Recoding a Continuous Variable	34
3.3	Combining Datasets	35
3.3.1	Adding Cases (Observations)	35
3.3.2	Adding Variables	37
4	Descriptive Analysis	39
4.1	Frequency Tables	39
4.2	Constructing Pie and Bar Charts	42
4.2.1	Pie Chart	42
4.2.2	Bar Charts	43
4.3	Graphs for Scale Variables	48
4.3.1	Histogram	48
4.4	Basic Summary Statistics	50
4.4.1	Central Tendency and Variation	51
4.4.2	Exploring Descriptive Statistics by Group	52
5	Inferential Statistics	56
5.1	Estimation of a Parameter	56
5.1.1	Point Estimation	56
5.1.2	Interval Estimation	57
5.2	Hypothesis Testing for Parameters	57
5.2.1	Types of Errors	57
5.2.2	Statistical Significance	58
5.2.3	Interpretations	58
6	Hypothesis Testing	59
6.1	Testing about a Single Population Mean	59
6.2	Comparing Paired Samples	61
6.3	Comparing Independent Samples	62
6.4	Comparing Several Population Means: ANOVA	65
6.5	Chi-Square Test of Association	69
7	Regression Analysis	71
7.1	Linear Correlation	71
7.1.1	Scatter Plot	71
7.1.2	Covariance	74
7.1.3	Correlation Coefficient	74
7.2	Linear Regression	77
8	Logistic Regression Models	83
8.1	Binary Logistic Regression	83
8.2	Multinomial Logistic Regression	85
8.3	Ordinal Logistic Regression	88

9 Survival Models	91
9.1 Cox Regression	91

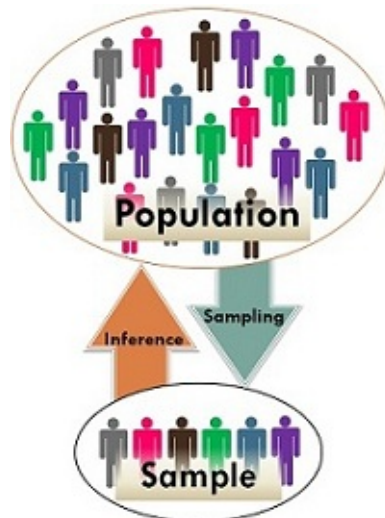
Chapter 1

Review of Basic Statistics

1.1 Common Terms

Before getting involved in the subject matter in detail, let us define some of the terms used extensively in the field of statistics.

- **Population:** A statistical population consists of *all objects* under study. The total number of objects in a population is called *population size* (N).
- **Sample:** A sample is the subset of a population. The number of objects in a sample is also called *sample size* (n).



Example 1.1. For each of the following situations, identify the population and the sample.

1. A study of 300 households in Addis Ababa city administration showed that 99% of them have comprehensive knowledge of HIV/AIDS.
 - Population: All households in Addis Ababa city administration.
 - Sample: The 300 households in Addis Ababa city administration.

2. A study of 250 patients admitted to St. Paul's Hospital during the past year revealed that, on the average, the patients lived 150 kms away from the hospital.
 - Population: All patients admitted to St. Paul's Hospital during the past year.
 - Sample: The 250 patients admitted to St. Paul's Hospital during the past year.
 - **Datum:** Datum is an observed value representing one or more characteristics of an object. It is also known as an *observation* or an *item* or a *case* or a *unit*.
 - The height of an individual: 1.72m.
 - The height and age of a person given as: 1.65m, 27yrs.
 - **Data:** Data is a collection of observed values (observations or cases) of some objects.
 - The heights of two individuals: 1.72m, 1.69m.
 - The height and age of two persons: (1.65m, 27yrs), (1.79m, 35yrs).

1.2 Classification of Statistics

Based on the scope of the decision, statistics can be classified into two; *descriptive* and *inferential* statistics.

1.2.1 Descriptive Statistics

Descriptive statistics is concerned with *organizing and summarizing* the most important features of the collected data *without going beyond the data itself*. That is, descriptive statistics describes only the data that we have, without attempting to conclude anything that goes beyond the data. It includes the *methods of data organization* like classification, tabulation and frequency distributions; *methods of data presentation* like diagrammatic and graphical displays; and certain indicators of data like *measures of central tendency* and *measures of variation*.

1.2.2 Inferential Statistics

Inferential statistics is concerned with drawing statistically *valid conclusions* about the characteristics of the population based on the results obtained from the sample. In this form of statistical analysis, *descriptive statistics is linked with probability theory* in order to generalize the results of the sample to the population. Performing *hypothesis testing*, determining *relationships between characteristics* and making *predictions (forecasting)* are also inferential statistics.

Example 1.2. Suppose a researcher is interested to know the average mark of a certain class in "Statistics" course. From a class of size 150, s/he took a random sample of 9 students and gave them an exam out of 100. Then, s/he got the average score 76. Consider the following statements.

- The average score of the 9 selected students is 76.
- The average score of the class is 76.

Example 1.3. We want to compare the average mark of boys and girls. Suppose we took a random sample of 7 boys of the total 85 boys and 6 girls of the total 60 girls, and gave them all the same exam. The average score of the 7 boys became 87 and that of the 6 girls became 92. Consider the following statements.

- The average score of the 7 boys is lower than that of the 6 girls.
- The 6 selected girls did better than the 7 selected boys.
- Girls did better in the exam than boys.

Exercise 1.1. Classify each of the following statements as *descriptive* or *inferential* statistics.

1. The average age of the students in this class is 21 years.
2. There is a strong association between smoking and lung cancer.
3. The price of wheat will be increased by 5% in the coming year.
4. Of the students enrolled in St.PHMMC this year, 74% are female and 26% are male.
5. The chance of winning the Ethiopian National Lottery in any day is 1 out of 167000.

1.3 Variables

A variable is a characteristic or an attribute that can assume different values. For example: height, family size, gender, marital status \dots .

1.3.1 Types of Variables: Quantitative vs Qualitative

Based on the values that variables assume, variables can be classified as *quantitative* or *qualitative* (*categorical*).

- **Quantitative variables:** Quantitative variables are those variables which assume numeric values. These variables are numeric in nature. Height and family size are examples of quantitative variables.

Quantitative variables are again further classified into two; *discrete* and *continuous* variables.

- **Discrete variables:** Discrete variables are those variables that assume a countable number of distinct and recognizable whole number values. Family size, number of children in a family, number of cars at the traffic light, \dots are some examples of discrete variables. Discrete variables can assume a *finite* number of possible values or an *infinite countable* number of values. The values of these variables are obtained by counting (0, 1, 2, \dots).
- **Continuous variables:** Continuous variables can take any value including decimals. These variables theoretically assume an *infinite* number of possible values. Their values are obtained by measuring. Examples of continuous variables are height, weight, time, temperature, \dots

- **Qualitative variables:** Qualitative variables are, on the other hand, those variables that assume non-numeric values called *categories* or *levels* or *groups*. For example, gender is a qualitative variable with two categories (levels): male and female. Marital status is also qualitative with, say, four categories: single, married, divorced, other.

Based on the number of values that qualitative variables assume, they can be classified as *binary* (*dichotomous*) or *multinomial* (*polytomous*).

- **Binary variables:** Binary variables often consist of 'either-or' type responses. That is, these variables have only *two* categories (levels). For example, gender, exam result of a student (pass, fail), smoking (smoker, non-smoker) are binary variables.
- **Multinomial variables:** Multinomial variables are those qualitative variables with *three or more* categories. For example, blood type (A, B, AB, O), marital status (single, married, divorced, other), religion (orthodox, muslim, protestant, ...), color (blue, red, green, black, ...) are multinomial variables.

Example 1.4. Classify each of the following variable as qualitative or quantitative and if it is quantitative classify as discrete or continuous.

1. Color of automobiles in a dealer's show room
2. Age of patients seen in a dental clinic
3. Number of seats in a movie theater
4. Blood pressure of a patient
5. The distance between a hospital to a house
6. Classification of patients based on nursing care needed (complete, partial, safers)
7. Temperature in the class room
8. Number of tomatoes on each plant on a field
9. Weight of newly born babies in a hospital during a year
10. Number of heart attacks
11. Temperature (very cold, cold, hot, very hot).
12. Heart rate
13. Cholesterol level

1.3.2 Measurement Scales: Nominal, Ordinal, Interval and Ratio

Consider the following two cases.

Case 1:

- Mr A wears 5 when he plays foot ball.

- Mr B wears 6 when he plays foot ball.

Who plays better? What is the average t-shirt number?

Case 2:

- Mr A scored 5 in Stat quiz.
- Mr B scored 6 in Stat quiz.

Who did better? What is the average score?

Based on the number on the t-shirts, it is not possible to judge whether Mr B plays better. But, by using the test score, it is possible to judge that Mr B did better in the exam. Also it not possible to find the average t-shirt numbers because the numbers on the t-shirts are simply codes but it is possible to obtain the average test score.

In general, a scale of measurement shows the *amount of information* contained in the value of a variable, and what *mathematical operations* and *statistical analysis* are permissible to be done on the values of the variable. There are four levels of measurement. These levels, from the weakest to the strongest, in order are: *nominal*, *ordinal*, *interval* and *ratio*.

1. **Nominal variables:** Nominal variables are qualitative variables which show classification of individuals into *mutually exclusive (non-overlapping)* and *exhaustive* categories without any associated ranking. For example; gender, religion, ethnicity, eye color (black, brown, etc), ... are nominal variables. Numbers may be assigned to the categories of these variables for coding purposes. But, it is not possible to compare individuals based on the numbers assigned to the categories. The only mathematical operation permissible on these variables is counting.
2. **Ordinal variables:** Ordinal variables are also qualitative variables whose values can be ordered and ranked. However, the ranks only indicate as to which category *greater* or *better* but there is *no precise difference* between the categories of the variable. Example: grade scores (A, B, C, D, F), academic qualifications (B.Sc., M.Sc., Ph.D.), strength (very weak, weak, strong, very strong), health status (very sick, sick, cured), strength of opinions in likert scales (strongly agree, agree, neutral, disagree, strongly disagree)
3. **Interval variables:** Interval variables are quantitative variables and identify not only as to which category is greater or better but also *by how much*. It is the stronger form of measurement but, there is no true (absolute) zero. Zero indicates *low* than *empty*. Examples: temperature, 0°C does not mean there is no temperature but, rather, it is too cold. Similarly, if a student scores 0 in a certain course, it does not mean the student has no knowledge in the course at all.
4. **Ratio variables:** These scales are the highest form of measurements. Ratio variables are quantitative variables but, unlike the interval variables, zero shows absence of a characteristic. All mathematical operations are allowed to be operated on the values of these variables. Examples: height, weight, income, expenditure, consumption, ...

Summary:

Scale				
Nominal	Numbers Assigned to Runners	7	8	3
Ordinal	Rank Order of Winners	Third place	Second place	First place
Interval	Performance Rating on a 0 to 10 Scale	8	9	10
Ratio	Time to Finish, in Seconds	15.2	14.1	13.4

- All quantitative variables are either interval or ratio scales where as all qualitative variables are either nominal or ordinal scales.
- Most statistical analyses do not distinguish interval and ratio scale variables. As a result, in most practical aspects, they are grouped under metric (scale) variables.
- If a characteristic has only one value in a particular study it is not a variable; it is a constant.
 - Thus, marital status is not a variable if all participants are married.
 - Gender is not a variable if all participants in a study are female.

1.3.3 Role of Variables: Dependent vs Independent

Based on the role of variables in a statistical analysis, variables can be classified as *dependent* or *independent* variables.

- A *dependent* variable is a variable, that is, of primary interest to be determined as an outcome. For example, the outcome of a certain treatment or the educational achievement level can be considered dependent variables. The terms *outcome*, *response* and *dependent* are used interchangeably.
- An *independent* variable is a variable to be used to determine the dependent variable. It is also called a *factor*, an *exposure*, a *predictor* or a *covariate*. There are two types of independent variables: *attribute (measured)* and *active (manipulated)* variables.
 - An *attribute* independent variable is a variable whose values are *preexisting attributes* of objects under study. The values of such a variable cannot be systematically changed or manipulated. For example, education, sex, socio-economic status, ...
 - An *active* independent variables can be experimentally manipulated. Such an active independent variable is a necessary (but not sufficient) condition to make *cause-and-effect* conclusions. For example, a researcher might investigate a new kind of therapy compared to the traditional treatment (the treatment group each person is assigned to). A second example could be a design to evaluate the effect

of different fertilizers on crop yields. A third example might be to study the effect of a new teaching method, such as cooperative learning, on student performance. Studies with active independent variables are experimental studies.

Even though a statistical analysis does not differentiate whether the independent variable is an attribute or active, there is a crucial difference in interpretation. For scientific researches in applied disciplines, the need to demonstrate that a given intervention or treatment causes change in behaviour or performance is extremely important. Only the approaches that have an active independent variable can provide data that allow one to infer that the independent variable caused the change or difference in the dependent variable. In contrast, a significant difference between or among persons with different values of an attribute independent variable should not lead one to conclude that the attribute independent variable caused the dependent variable to change.

In choosing a statistical method/model, there are some questions to be considered: What is the research question/objective of the study? What are the variables to be included in the data? What is the type of each variable? Every research has at least one outcome (dependent) variable. Thus, the dependent and independent variables should be identified.

Based on the type and role of variables, the common statistical methods are shown in the following table.

Dependent Variable	Independent Variable	Method
Continuous	Binary	<i>t</i> test
Continuous	Multinomial	ANOVA
Continuous	Quantitative/Categorical/Both	Linear Regression
Categorical	Categorical	χ^2 test
Binary	Quantitative/Categorical/Both	Binary Logistic Regression
Multinomial	Quantitative/Categorical/Both	Multinomial Logistic Regression
Ordinal	Quantitative/Categorical/Both	Ordinal Logistic Regression
Discrete	Quantitative/Categorical/Both	Poisson Regression

- Note: For correlation and χ^2 test, there is no need to differentiate variables as dependent and independent.

Exercise 1.2. For each of the following objectives of a statistical investigation, identify the dependent and independent variables.

1. The effect of gender on blood pressure.
2. The effect of age on hypertension.
3. The effect of gender on hypertension.
4. Comparing the number of sexual partners between urban and rural resident women.
5. Investigating the effectiveness of a new therapy compared to the traditional treatment.
6. Evaluating the effect of different fertilizers on crop yields.
7. Comparing the blood pressure of male and female individuals.

8. Studying the effect of a new teaching method, such as cooperative learning, on student performance.
9. The effect of age (classified as < 18 , $18 - 24$, $25 - 34$, $35 - 64$ and > 65 years) on hypertension.

1.4 Rules of Coding Variables

Coding is the process of assigning numbers to the categories (levels) of a qualitative (categorical) variable. All codes (data values) should be numeric. Even though, it is possible to use letters or words as codes, it is not desirable to do with most statistical software packages. Also, all codes for a variable must be mutually exclusive. That is, two or more categories of a variable should not have the same code. Below are some rules of coding a binary, multinomial and ordinal variables.

1.4.1 Coding Binary Variables

For a binary variable, the two levels can be represented by any two numbers that are separated by one: 0 and 1, 1 and 2, 0 and -1, etc. For example, gender may be coded as 1 for males and 2 for females.

Variable	Value	Value Label
Gender	1	Male
	2	Female

The sign of the parameter estimate may change depending on whether the higher value is assigned for the Male or Female category. But the parameter estimate and significance level will be the same. However, the parameter estimate will be different if there is more than one point between the two codes. For example, coding schemes like +1 and -1 will give a different result because there is more than one unit between the two values.

1.4.2 Coding Nominal Variables

The actual values taken on by nominal variables are irrelevant. The starting category for coding of a nominal variable is arbitrary. For example, if occupation has four categories, it can be coded as

Variable	Value	Value Label
Occupation	0	Government Employee
	1	NGO Employee
	2	Private Employee
	3	Other

or

Variable	Value	Value Label
Occupation	2	Government Employee
	3	NGO Employee
	1	Private Employee
	4	Other

Nominal independent variables are included in a statistical model in the form of multiple binary (called *dummy*) variables (will be discussed later).

1.4.3 Coding Ordinal Variables

The actual values of the levels of ordinal variables are irrelevant so long as the numeric difference between the levels is one unit. For example, it does not matter whether a 5-level ordinal variable (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied) is coded as 0, 1, 2, 3, 4, or 1, 2, 3, 4, 5 or 9, 10, 11, 12, 13. It is better to use higher values for the higher categories ("agree", "good" or "positive" end) of an ordinal variable.

Variable	Value	Value Label
Satisfaction level	1	Very dissatisfied
	2	Dissatisfied
	3	Neutral
	4	Satisfied
	5	Very satisfied

Some codings might use 1 for "strongly agree" and 5 for "strongly disagree" which would matter the interpretation of the result. This is not wrong as long as we are clear and consistent. For example, if education level has, say, 4 categories: no education, primary, secondary and tertiary education, it can be coded as:

Variable	Value	Value Label
Education	0	Tertiary Education
	1	Secondary Education
	2	Primary Education
	3	No Education

Here, we have to remember an increase in the codes indicates a decrease in the education level. However, we are less likely to get confused when interpreting the results if high values have positive meaning.

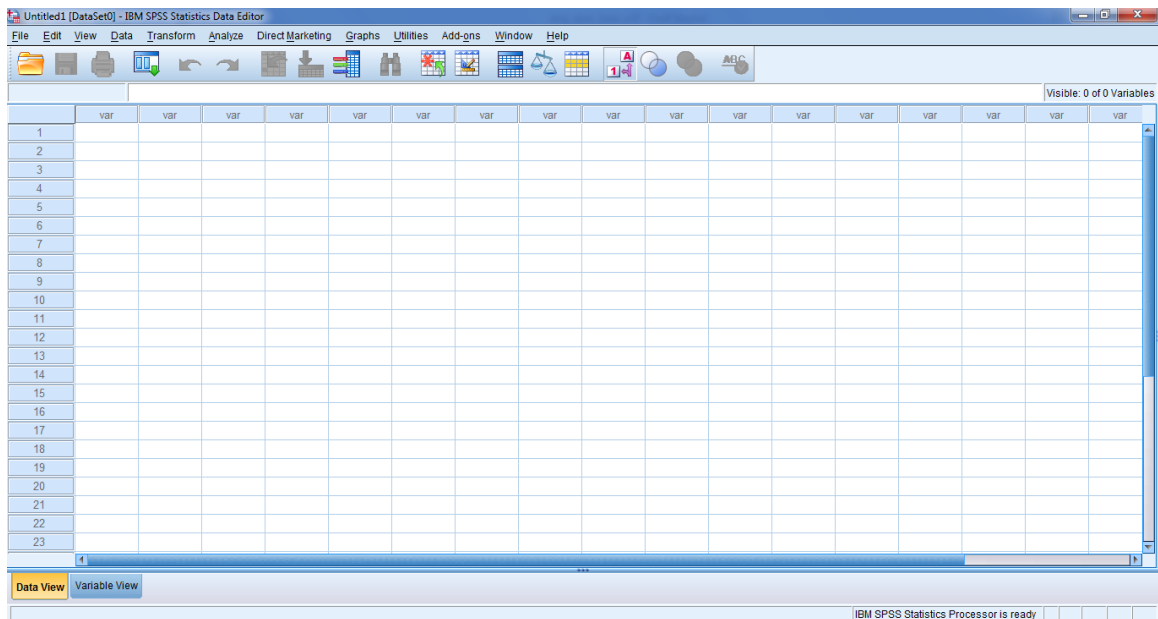
Chapter 2

Introduction to SPSS

SPSS was originally developed for conducting statistical analysis in the field of social sciences only and the abbreviation SPSS was used for "Statistical Package for the Social Sciences". But, the current development of the software is applicable for a variety of disciplines that stands for change to "Statistical Product and Service Solutions".

2.1 Exploring SPSS

To open SPSS, click on the **Start** → **IBM SPSS Statistics 20**. Then, the main SPSS interface looks the following.



In the main interface, the **Title** bar displays the name of the opened data file if any, or "Untitled1[DataSet0]" if empty or if the file has not yet been saved. Next, the *Menu* bar lists different pull down menus which provides easy access to most SPSS features. Also, the *Status* bar at the bottom of each SPSS window tells what SPSS is currently doing. Typical messages one will see are "IBM SPSS Statistics Processor is ready", "Running procedure...".

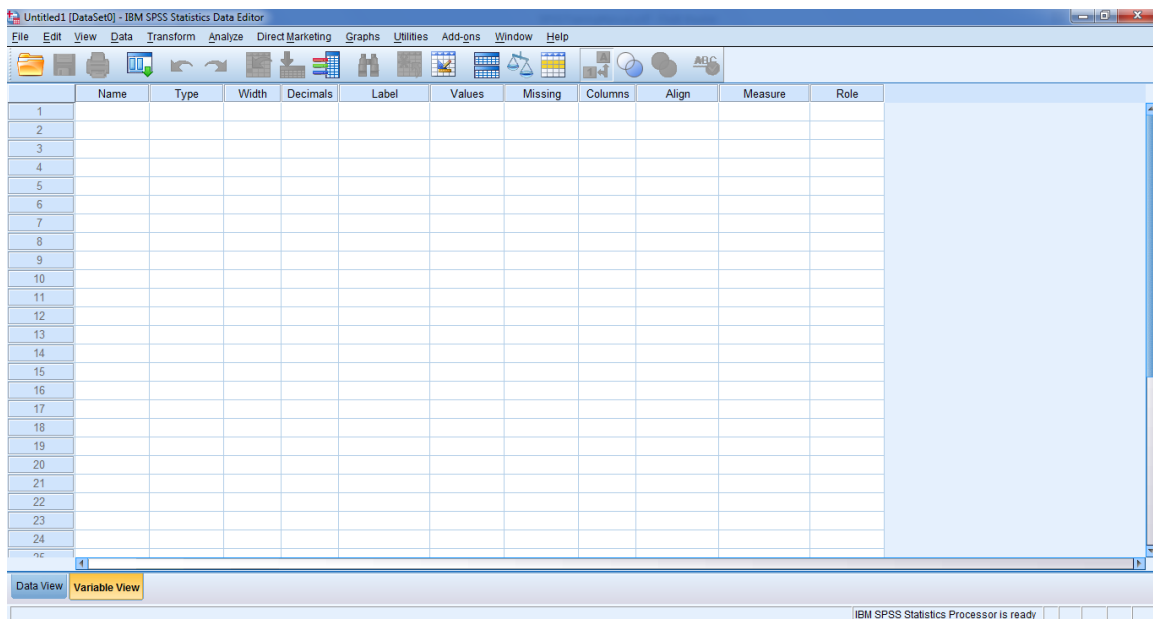
Of the several windows that can be opened when using SPSS, the three common windows are the *Data Editor* Window, the *Output* Window and the *Chart Editor* Window.

- **The Output (Viewer) Window:** The *Output* window displays the statistical results, tables, and charts from the analysis performed. An output window opens automatically when a procedure that generates output is run. In the output window, the results can be edited, moved, deleted and copied in a Microsoft Explorer like environment. This window is not accessible until output has been generated. A file with an extension of *.spo* is assumed to be a *Viewer* file containing statistical results and graphs.
- **The Chart Editor Window:** The *Chart Editor* window is only displayed after SPSS has been requested to produce a plot (chart). In this window, the plots can be edited, i.e., the colors can be changed, different type fonts or sizes can be selected, axes can be rotated (switch the horizontal and vertical axes), the chart type can be changed and the like.

2.2 The Data Editor Window

The *Data Editor* window opens automatically when SPSS is started. It is a spreadsheet in which the variables are to be defined and the data are to be entered. Each row corresponds to a case or an observation while each column represents a variable. There is no limit to the number of variables and/or cases that can be used.

Notice, the *Data Editor* window has two sheets, labelled at the bottom **Data View** and **Variable View**. The **Data View** sheet shows the data just as an Excel worksheet does. The **Variable View** is used define each variable in the dataset (defining variables in SPSS is described in detail in Section 2.2.1), that is, the **Variable View** sheet contains the definitions of each variable in the dataset. Now to open the **Variable View** sheet, just click on the **Variable View** tab. Then, the sheet looks:



While the variables are listed as columns in the **Data View** sheet, they are listed as rows in the **Variable View**. In the **Variable View**, each row is a variable, each column is an attribute (characteristic) associated with that variable.

2.2.1 The Variables View Sheet

The **Variable View** sheet is used create variable names and define the attributes of each variable. The entries of this sheet are:

- **Name** - Variable names can be up to 64 characters, always beginning with a letter and not end with a period. They can contain numbers (also @, #, and \$ characters) but no funny characters like spaces/blanks/hyphens or special characters. It does not matter if a variable is called WEIGHT, weight, or WEiGhT since variable names are not case sensitive. But, they must be unique.
- **Type** - It indicates what type the variable is. There are several variable types in SPSS. The most common are described below.
 - **Numeric**: A variable whose values are numbers. The *Data Editor* accepts numeric values both in standard format and scientific notation.
 - **Comma**: A numeric variable whose values are displayed with commas delimiting every three places, and with the period as a decimal delimiter, example 76,721.05. The *Data Editor* accepts numeric values for comma variables with or without commas; or in scientific notation.
 - **Dot**: A numeric variable whose values are displayed with periods delimiting every three places, and with the comma as a decimal delimiter, example 76.721,05. The *Data Editor* accepts numeric values for dot variables with or without dots; or in scientific notation.
 - **Scientific notation**: A numeric variable whose values are displayed with an embedded E and a signed power-of-ten exponent. The *Data Editor* accepts numeric values for such variables with or without an exponent. The exponent is preceded by E, for example, 123, 1.23E2 or 1.23E+2.
 - **Date**: A numeric variable whose values are displayed in one of several calendar date. Dates can be entered with slashes, hyphens, periods, commas, or blank spaces as delimiters.
 - **String**: Some numbers are not really numbers. That is, they are numbers but we cannot use them in mathematical calculations. Take a phone number, for example, or an account number or a postal code. We can sort them, but we cannot add or subtract or multiply them. Well, we could, but the result would be meaningless. In essence, these numbers are actually just text which happens to be numeric. A good example is the recent plate number of cars in Ethiopia, which contains both numbers and text. We refer to these variables as string (text) or alphanumeric. Uppercase and lowercase letters are considered distinct.
- **Width** - The numerical entry in this box gives how many spaces the entries in the **Data View** will be for this variable.

- **Decimals** - For numeric data, this entry gives how many decimal places will be shown for this variable in the **Data View**.
- **Label** - It stands for a descriptive title for the variable.
- **Values** - This is used to specify a label for each numerical value of a categorical variable.
- **Missing** - This allows us to specify which values for a variable indicate missing data. Blanks are recommended for missing values because SPSS is designed to handle blanks as missing values and, by default, it assigns a period for them. However, sometimes datasets we might receive use special numerical values (for example: 99, 999, 9999) to indicate missing values. Hence, unless we tell SPSS that these values are codes for missing, SPSS will treat them as actual data.
- **Columns** - The numerical value in this item gives how many spaces will be allocated for the variable in the **Data View**. This is different from **Width** in that **Width** limits the number of spaces for the actual number. **Columns** limits how many spaces will be visible in the **Data View**.
- **Align** - This entry either left aligns, centers, or right aligns the entries for the variable.
- **Measure** - This indicates what measurement scale of variable is. The available measures are scale, ordinal and nominal. A scale in SPSS is a quantitative variable.

2.2.2 The Data View Sheet

After defining the variables, the next task is to enter the data in the **Data View** sheet directly from the questionnaires or data entry form. Remember that variables are listed as columns in the **Data View** sheet while they are listed as rows in the **Variable View**.

2.3 Creating a New Dataset in SPSS

There are three steps that must be followed to create a new dataset in SPSS: defining variables in the **Variable View**, entering the data in the **Data View** and saving the data in the hard drive of our computer.

Example 2.1. Enter the following data: three variables (Weight, Sex and Marital Status (1=Single, 2=Married, 3=Divorced, 4=Other)) with five observations.

Weight	Sex	Marital Status
60.0	M	1
58.5	F	2
53.0	F	3
56.5	M	2
70.0	M	4

2.3.1 STEP 1: Defining Variables in the Variables View

Now note that **Weight** is a quantitative variable (numeric in nature). And **Sex** and **Marital Status** are both qualitative. **Sex** will be treated as **String**. But since the categories of **Marital Status** are coded, it will be treated as **Numeric**.

Having the above note in mind, first each variable should be defined with its characteristics in the **Variable View** sheet. To start, open a new SPSS and go to the **Variable View**. Then, on the first row, define the **Weight** variable as follows:

1. Write **Wei** standing for the **Weight** variable in the **Name** column.
2. Change the **Type** column to **Numeric** which is the default.
3. Change the **Decimals** column to 1.
4. In the **Label** column, write "Weight of a Student".
5. Change the **Align** column to **Center**.
6. In the **Measure** column, change it to **Scale**.

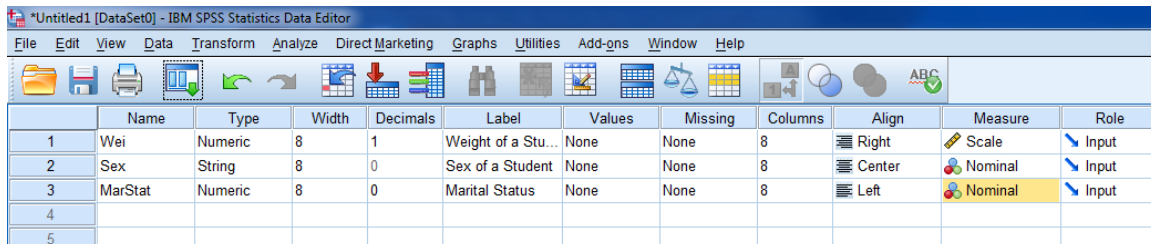
On the second row of the **Variables View** sheet, define **Sex** as follows.

1. Write the name **Sex** standing for the **Sex** variable in the **Name** column.
2. Change the **Type** column to **String**.
3. In the **Label** column, write "Sex of a Student".
4. Change the **Align** column to **Center**.
5. In the **Measure** column, change it to **Nominal**.

Similarly, define **Marital Status** in the third row of the **Variables View** sheet:

1. Write the name **MarStat** standing for **Marital Status** in the **Name** column.
2. Change the **Type** column to **Numeric**.
3. Change the **Decimals** column to 0 since there is no decimal place in the values.
4. In the **Label** column, write "Marital Status".
5. Change the **Align** column to **Left**.
6. In the **Measure** column, change it to **Nominal**.

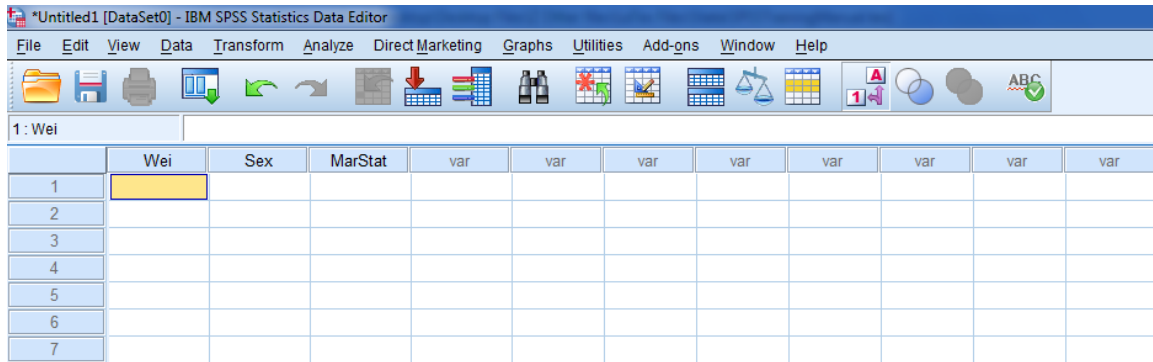
Now the **Variable View** looks as follows.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Wei	Numeric	8	1	Weight of a Stu...	None	None	8	Right	Scale	Input
2	Sex	String	8	0	Sex of a Student	None	None	8	Center	Nominal	Input
3	MarStat	Numeric	8	0	Marital Status	None	None	8	Left	Nominal	Input
4											
5											

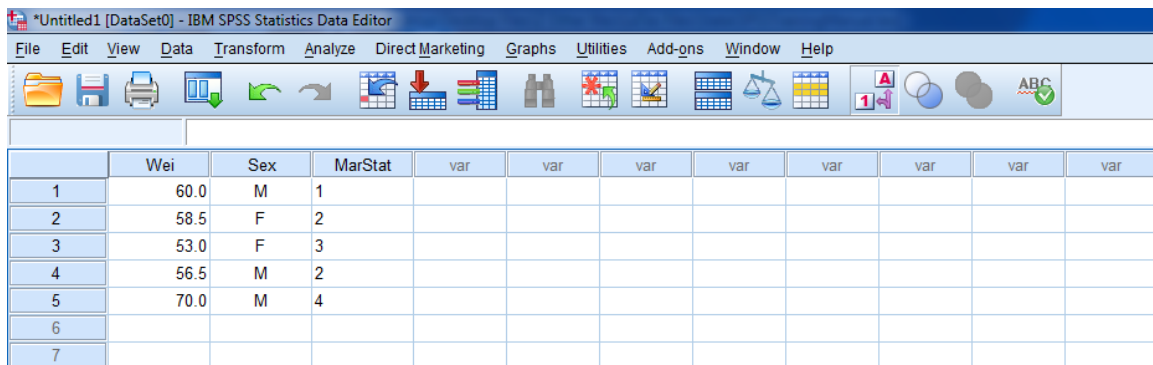
2.3.2 STEP 2: Entering the Data in the Data View

Once all of the variables are defined, the data can be entered manually in the **Data View** sheet. Now go to the **Data View** which shows the variable names as the column names. It looks like:



	Wei	Sex	MarStat	var	var	var	var	var	var	var	var
1											
2											
3											
4											
5											
6											
7											

The data is then typed into one cell at a time. The information is entered into the cell when the active cell is changed.



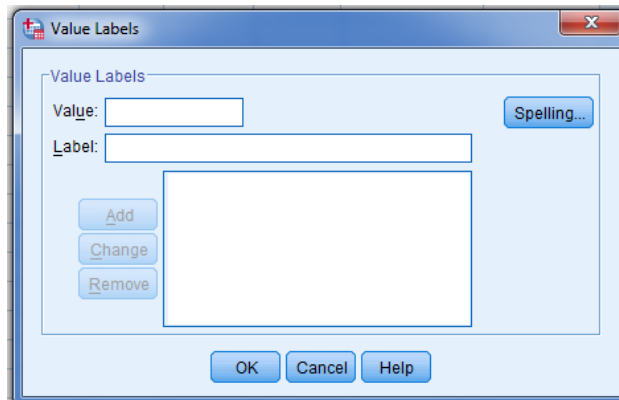
	Wei	Sex	MarStat	var	var	var	var	var	var	var	var
1	60.0	M	1								
2	58.5	F	2								
3	53.0	F	3								
4	56.5	M	2								
5	70.0	M	4								
6											
7											

Note: Had we entered the data into the **Data View** prior to defining the variables, SPSS assigned the default variable names VAR00001, VAR00002, VAR00003. To change the variable names, click on the **Variable View** tab. Then change VAR00001 to Wei and similarly the other two.

Creating Value Labels

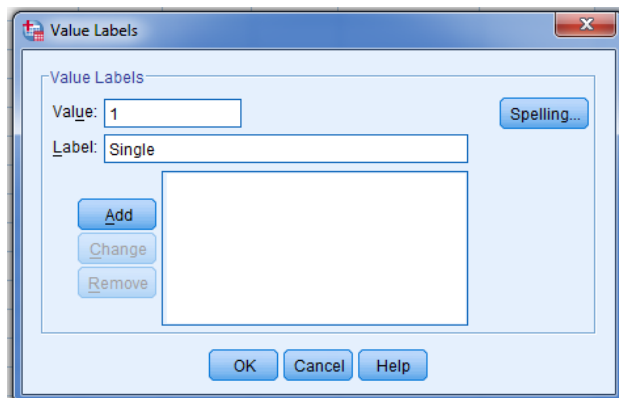
It is nice to have the values of a categorical variable labeled with their meaning. For example, Marital Status should have labeled as Single, Married, Divorced and Other rather than 1, 2, 3, and 4.

To create a label for the MarStat variable, click on the **Values** column of the MarStat variable in the **Variable View** sheet. Then the **Value Labels** dialogue box appears.

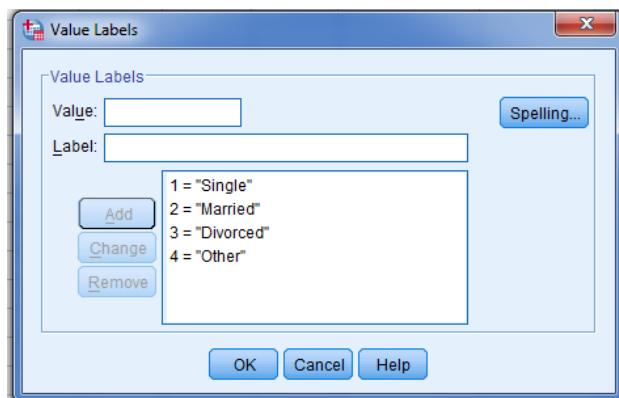


Then,

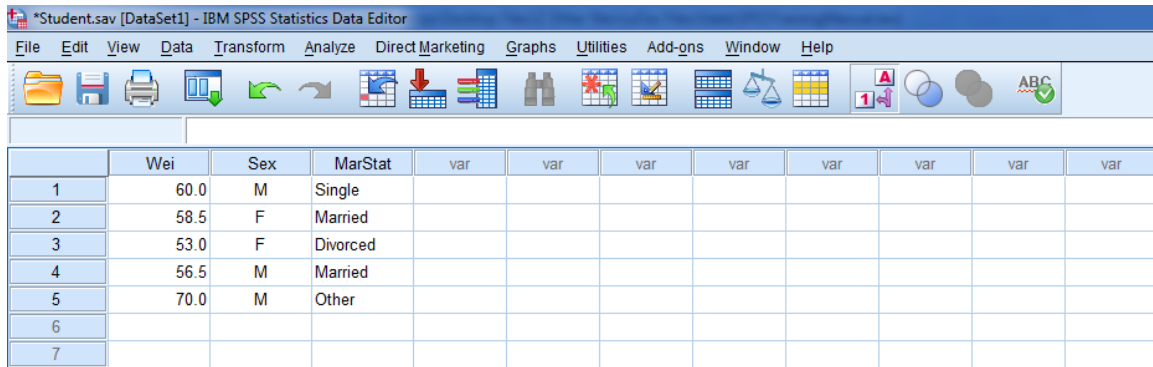
- Type 1 in the **Value** field.
- Write Single in the **Label** field.
- Click the **Add** tab to have this label added to the list.



In a similar way, continue labeling until all the values are labeled. When all values are labeled, the **Value Labels** window becomes:



Now click on **OK**. Then go back to the **Data View** and observe the difference.



	Wei	Sex	MarStat	var	var	var	var	var	var	var	var
1	60.0	M	Single								
2	58.5	F	Married								
3	53.0	F	Divorced								
4	56.5	M	Married								
5	70.0	M	Other								
6											
7											

2.3.3 STEP 3: Saving the Data

To retain the current dataset, it must be saved to a file.

1. From the *Menu* bar, click on **File** → **Save As**.
2. In the **Save Data As** dialogue box, in the **File name:** field, write a data file name, say, **Student**. Since, from the **Save as type:** drop-down list, the default extension is 'SPSS Statistics (.sav)', then SPSS by default saves the data file by adding the extension **.sav**, that is, **Student.sav**.
3. From the **Look in:** drop-down list, select the location path where the file will be saved.
4. Then, click on the **Save** tab.

Now the *Title* bar looks:



The saved data file is **Student.sav**. Such a file with an extension of **.sav** is assumed to be an SPSS data file for Windows format.

Exercise 2.1. Enter the following data in SPSS: height in meter, blood type (0=Type A, 1=Type B, 2=Type AB, 3=Type O) and gender (0=Male, 1=Female). Then save the data using the file name **Blood**.

Height	Blood Type	Gender
1.55	1	0
1.6	0	1
1.72	1	0
1.5	2	1
1.85	3	0

2.4 Accessing Data Files

2.4.1 Opening an SPSS Data File

To open an SPSS data file,

1. From the *Menu* bar, click on **File** → **Open** → **Data**.

2. In the **Open Data** dialogue box, from the **Look in:** drop-down list, select the location path where the file is saved.
3. Of the list of data files, if any, click on the data file name to be opened, for example, `Student.sav`.
4. Then, click on the **Open** tab.

2.4.2 Importing and Exporting Data

Data can be imported-from (read) and exported-to (saved) a number of different sources. Some of the common data files SPSS supports are: excel data files: `.xls`, `.xlsx`; text files: `.txt`, `.csv`; stata data files: `.dta`.

Example 2.2. Opening a non-SPSS data file: Let us open an excel data file, `CD4.xlsx`, into SPSS.

1. From the *Menu* bar, click on **File** → **Open** → **Data**.
2. In the **Open Data** dialogue box, from the **Files of type:** drop-down list, select an extension, 'Excel (`.xls`, `.xlsx`, `.xslm`)', (possibly 'All Files (`*.*`)'). If not changed, SPSS automatically searches the default extension 'SPSS Statistics (`.sav`)'.
3. From the **Look in:** drop-down list, select the location path where the file is saved.
4. Of the list of data files, if any, click on the data file name to be opened, that is, `CD4.xlsx`.
5. Click on the **Open** tab and then the **OK** tab.

Example 2.3. Saving into a different format: Let us save the `Student.sav` data in an excel file. The procedure is as follows.

1. From the *Menu* bar, click on **File** → **Save As**.
2. In the **Save Data As** dialogue box, from the **Save as type:** drop-down list, select the extension 'Excel 2007 through 2010 (`.xlsx`)'. If not changed, SPSS by default saves in 'SPSS Statistics (`.sav`)' format.
3. From the **Look in:** drop-down list, select your preferred location path where the file will be saved.
4. Then, in the **File name:** field, write a data file name, say, `Stud`.
5. Lastly, click on the **Save** tab.

Example 2.4. In the folder given to you, there is an SPSS data file named `JUSH_HAART`. In most of the illustrations in this training, we'll use this data. The descriptions of the variables are given below.

Table 2.1: Variable Descriptions of the JUSH_HAART data

Variable Name	Variable Label	Value Label
CardNum	Patient's Card Number	
Age	Age in Years	
Sex	Sex	
Wei	Weight in Kilograms	
MarStat	Marital Status	0=Never Married, 1=Married, 2=Divorced, 3=Separated, 4=Widowed
EducLev	Education Level	0=No Education, 1=Primary, 2=Secondary, 3=Tertiary
Emp	Employment Condition	0=Full-time, 1=Part-time, 2=Not Working, 3=Unemployed
ClinStag	Clinical Stage	1=Stage I, 2=Stage II, 3=Stage III, 4=Stage IV
FunStat	Functional Status	0=Working, 1=Ambulatory, 2=Bedridden
CD4	Number of CD4 Counts	
Status	Survival Outcome	0=Active, 1=Dead, 2=Transferred, 3=Lost-to-follow
Defaulter	Dropped Out Patient	0=Active, 1=Defaulted
SurvTime	Survival Time (Months)	

Chapter 3

Basic Data Management

Data management includes all activities associated with data other than the direct use of the data. It takes place during all stages of a study which includes all aspects in planning the data needs of the study (e.g., specifying the objectives of the study), designing data collection sheets, data collection, data entry, data cleaning (validation and checking), data manipulation, data analysis and interpretation, data files backup and data documentation. The two main objectives of data management are to create a reliable database containing high quality data, without introducing *data processing errors*¹; and to manipulate and process the data so as to make it ready for the required analysis.

Data processing errors are errors that occur after the data have been collected. Therefore, after creating a new data file or opening an existing data file, it is typically essential to examine the data and identify possible data processing problems (errors). Examples of data processing errors include:

- Coding errors (e.g., groups of marital status gets improperly coded because of changes in the coding scheme)
- Routing errors (e.g., the interviewer asks the wrong question or asks questions in the wrong order)
- Consistency errors (contradictory responses, such as the reporting of a pregnancy after the respondent has identified himself as a male)
- Range errors (responses outside of the range of plausible answers, such as a reported age of 290)
- Duplicating errors (a single case might be entered accidentally more than once)
- Transpositions (e.g., 19 becomes 91 during data entry)
- Copying errors (e.g., 0 (zero) becomes O during data entry)

To prevent such data processing errors, the stage at which the errors occurred must be identified and then the problem should be corrected. Some of the mechanisms are:

¹This is distinct from measurement errors, which are differences between the true state of affairs and what appears on the data collection form.

- Manual checking (e.g., checks for completeness, handwriting legibility)
- Range and consistency checking during and after data entry
- Double entry and validation following data entry
- Data analysis screening for outliers during data analysis
- Identifying and deleting duplicate cases

Data manipulation is the process of changing and organizing data to make it ready for the required analysis. For instance, new variables may be created from the existing quantitative variables. For example: the BMI an individual is calculated from the Weight and Height measurements.

Also, existing qualitative variables may be recoded during data manipulation. Example: If marital status is already coded as 1=single, 2= married, 3=divorced, 4=widowed, 5=separated; it can be recoded as 0=single, 1=married, 2=other (divorced, widowed or separated).

New variables may be generated by combining the existing variables. Example: watching TV (Yes, No) and listening to Radio (Yes, No) might be combined to generate a new variable called Media exposure (Yes - if exposed to either media, No- if not exposed to both medias). Similarly, attitude towards something may be determined from a set of likert scale measurement items.

3.1 Preliminary Data Analysis

Before directly going to analyse data, it is essential to look at the details of the data at a glance. The question is "Does the data make sense?" out of range, missing, illogical/implausible values, consistency with other variables. The basic rule is printing frequencies for categorical variables and summary statistics for quantitative variables. These two printouts can be used as a primary references and give a picture of the overall shape of the data.

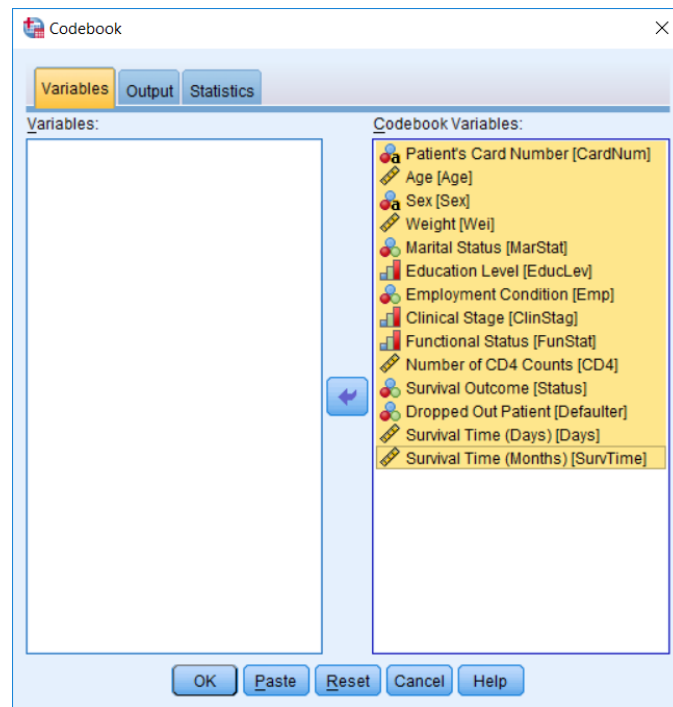
The Codebook Procedure

The **Codebook** procedure reports descriptive statistics for the specified variables, particularly counts and percentages for nominal and ordinal variables (including all string variables), and measures of central tendency and variation (mean, standard deviation and quartiles) for scale variables.

From the *Menu* bar, click on **Analyze** → **Reports** → **Codebook**. In the **Codebook** dialogue box, enter at least one variable to the **Codebook Variables:** box.

Example 3.1. Let us examine and observe the overall shape of the JUSH_HAART.sav data using the **Codebook** procedure.

As shown below, all variables are entered into the **Codebook Variables:** box. You can also specify the information you need under the **Output** and **Statistics** tabs.



Then, click on the **OK** tab and examine the results on the *Output* window. The output for *Age* of the patients is:

Age		Value
Standard Attributes	Position	2
	Label	Age
	Type	Numeric
	Format	F8
	Measurement	Scale
	Role	Input
N	Valid	1464
	Missing	0
Central Tendency and Dispersion	Mean	34.01
	Standard Deviation	9.160
	Percentile 25	28.00
	Percentile 50	32.00
	Percentile 75	39.00

The above table for *Age* shows the mean age is 34.01 years with a standard deviation of 9.16 years. In addition, 25% of the patients were below 28 years, 50% of them were below 32 years and 75% of the patients were below 39 years.

The table for *Sex*, shown below, displays 930 (63.5%) of the patients were female and the remaining 534 (36.5%) of the patients were male.

Sex				
		Value	Count	Percent
Standard Attributes	Position	3		
	Label	Sex		
	Type	String		
	Format	A1		
	Measurement	Nominal		
	Role	Input		
Valid Values	F		930	63.5%
	M		534	36.5%

Similarly, the table for MarStat is as follows, showing there are 4 (0.3%) missing values.

MarStat				
		Value	Count	Percent
Standard Attributes	Position	5		
	Label	Marital Status		
	Type	Numeric		
	Format	F8		
	Measurement	Nominal		
	Role	Input		
Valid Values	0	Never Married	293	20.0%
	1	Married	739	50.5%
	2	Divorced	134	9.2%
	3	Separated	140	9.6%
	4	Widowed	154	10.5%
Missing Values	System		4	0.3%

3.2 Manipulating Data

3.2.1 Protecting the Original Data

A data file can be marked as read-only so as to prevent the accidental modification of the original data. From the *Menu* bar, click on **File** → **Mark File Read Only**. If subsequent modifications are made to the data, the modified data can be saved *only with a different file name*; so the original data are not affected. The file permissions can be changed back to read/write by selecting **Mark File Read Write** from the **File** menu.

3.2.2 Inserting and Deleting Cases

To insert a new case, right click on the row number in the **Data View** sheet in which the new case is to be inserted and then click on **Insert Cases**. (Entering data in an empty row of the **Data View** automatically creates a new case.)

To delete a case, right click on the row number (case) to be deleted and click on **Clear**.

For example, to insert a new case on the second row or to delete the second case, just right click on row number 2 and then click on **Insert Cases** or **Clear**, respectively.

JUSH_HAART.sav [DataSet2] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

2 : CardNum 1203

	CardNum	Age	Sex	Wei	MarStat	EducLev	Emp	ClinStag	FunStat	CD4	Status	Defaulter	Days	SurvTime
1	1202	40	F	43	Separated	Primary	Unemployed	Stage IV	Working	365	Active	Active	809	26.97
2	1203	50	M	58	Married	Primary	Part-time	Stage II	Bedridden	434	Active	Active	1000	33.33
3	1207	35	F	55	Married	Primary	Unemployed	Stage IV	Working	270	Active	Active	837	27.90
4	1209	45	M	60	Married	Tertiary	Full-time	Stage I	Working	450	Active	Active	1530	51.00
5	1211	44	M	65	Married	Secondary	Not Working	Stage II	Working	1352	Loss-to-foll...	Defaulted	91	3.03
6	1212	30	F	73	Widowed	Secondary	Full-time	Stage III	Working	120	Active	Active	1770	54.00
7	1214	63	F	72	Widowed	Primary	Full-time	Stage III	Working	217	Active	Active	1496	49.87
8	1221	25	F	44	Separated	No Educati...	Unemployed	Stage III	Ambulatory	72	Loss-to-foll...	Defaulted	3	.10
9	1222	24	F	55	Married	Secondary	Unemployed	Stage I	Working	305	Active	Active	35	1.17

3.2.3 Inserting and Deleting Variables

Inserting and deleting a variable can be done using both sheets of the *Data Editor*. On the **Data View** sheet, click on the column that the new variable is to be inserted and then click on **Insert Variable**. Or in the **Variable View** sheet, right click on the row number that the new variable is to be inserted, and then click on **Insert Variable**. (Entering data in an empty column in the **Data View** or in an empty row in the **Variable View** automatically creates a new variable with a default name.)

To delete a variable, on the **Data View** sheet, right click on the variable name to be deleted and then click on **Clear**. Or in **Variable View** sheet, right click on the row number to be deleted, and then click on **Clear**.

For example, to insert a new variable between Sex and Wei, or to delete the Wei variable, just right click on Wei and then click on **Insert Variable** or **Clear**, respectively.

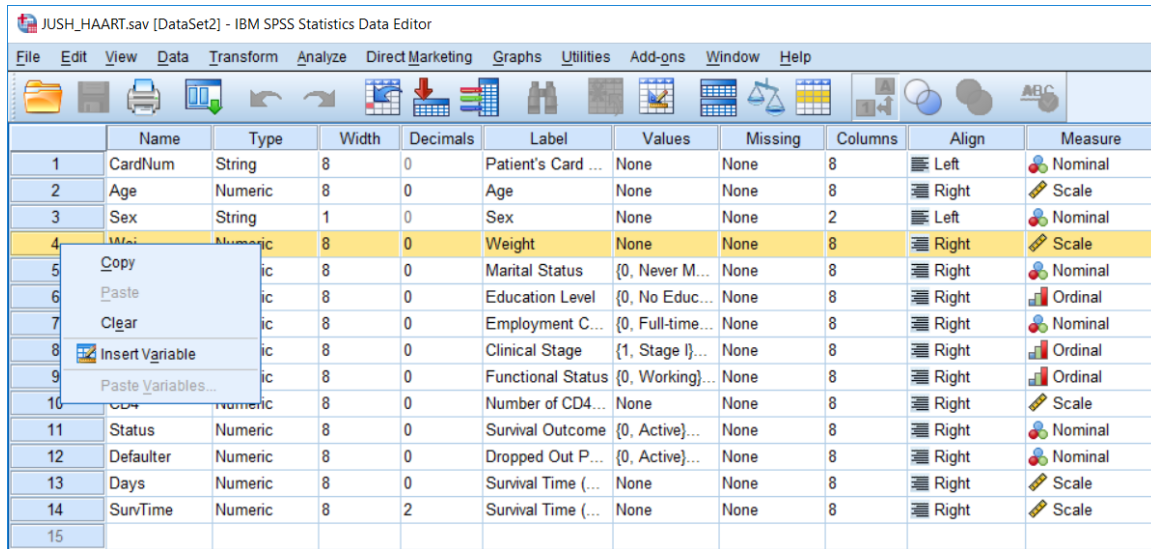
JUSH_HAART.sav [DataSet2] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 : Wei 43

	CardNum	Age	Sex	Wei	EducLev	Emp	ClinStag	FunStat	CD4	Status	Defaulter	Days	SurvTime	
1	1202	40	F	43	Primary	Unemployed	Stage IV	Working	365	Active	Active	809	26.97	
2	1203	50	M	58	Primary	Part-time	Stage II	Bedridden	434	Active	Active	1000	33.33	
3	1207	35	F	55	Primary	Unemployed	Stage IV	Working	270	Active	Active	837	27.90	
4	1209	45	M	60	Tertiary	Full-time	Stage I	Working	450	Active	Active	1530	51.00	
5	1211	44	M	65	Secondary	Not Working	Stage II	Working	1352	Loss-to-foll...	Defaulted	91	3.03	
6	1212	30	F	73	Secondary	Full-time	Stage III	Working	120	Active	Active	1770	54.00	
7	1214	63	F	72	Primary	Full-time	Stage III	Working	217	Active	Active	1496	49.87	
8	1221	25	F	44	Educati...	Unemployed	Stage III	Ambulatory	72	Loss-to-foll...	Defaulted	3	.10	
9	1222	24	F	55	Secondary	Unemployed	Stage I	Working	305	Active	Active	35	1.17	
10	1224	40	M	61	Married	Tertiary	Full-time	Stage I	Ambulatory	279	Transferred	Defaulted	46	1.53

OR



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	CardNum	String	8	0	Patient's Card ...	None	None	8	Left	Nominal
2	Age	Numeric	8	0	Age	None	None	8	Right	Scale
3	Sex	String	1	0	Sex	None	None	2	Left	Nominal
4	Weight	Numeric	8	0	Weight	None	None	8	Right	Scale
5	Marital Status	Nominal	8	0	Marital Status	{0, Never M...	None	8	Right	Nominal
6	Education Level	Ordinal	8	0	Education Level	{0, No Educ...	None	8	Right	Ordinal
7	Employment C...	Nominal	8	0	Employment C...	{0, Full-time...	None	8	Right	Nominal
8	Clinical Stage	Ordinal	8	0	Clinical Stage	{1, Stage I}...	None	8	Right	Ordinal
9	Functional Status	Ordinal	8	0	Functional Status	{0, Working}...	None	8	Right	Ordinal
10	Number of CD4...	Scale	8	0	Number of CD4...	None	None	8	Right	Scale
11	Status	Nominal	8	0	Survival Outcome	{0, Active}...	None	8	Right	Nominal
12	Defaulter	Nominal	8	0	Dropped Out P...	{0, Active}...	None	8	Right	Nominal
13	Days	Scale	8	0	Survival Time (...)	None	None	8	Right	Scale
14	SurvTime	Scale	8	2	Survival Time (...)	None	None	8	Right	Scale
15										

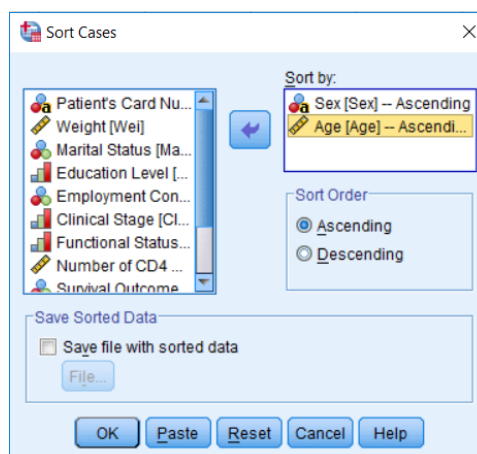
3.2.4 Sorting Cases

In order to sort the data, from the *Menu* bar, click on **Data** → **Sort Cases**. Then, in the **Sort Cases** dialogue box, enter the sorting variable(s) in the **Sort by:** box. If two or more variables are sorting variables, then the cases are sorted by each variable within categories of the preceding variable on the sort list.

Note: For **String** variables, uppercase letters precede their lowercase counterparts in sort order. For example, the string value "Yes" comes before "yes" in sort order.

Example 3.2. Sort the JUSH_HAART.sav data by Sex and Age.

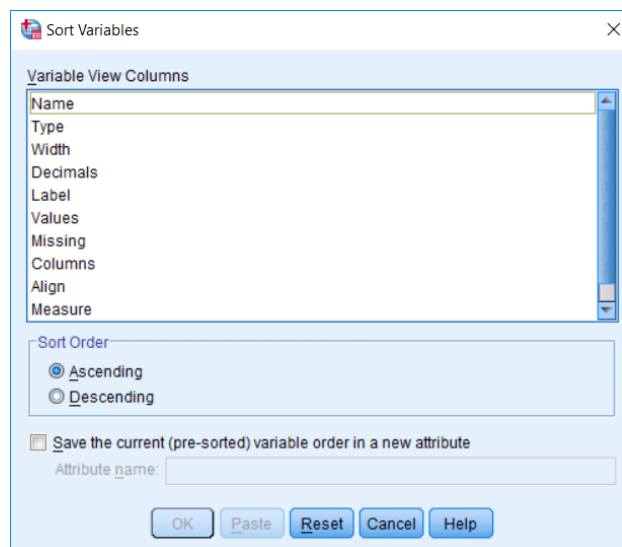
Both Sex and Age should be entered in the **Sort by:** box.



Here, Sex was the first variable entered, followed by Age; accordingly, the data will first be sorted by Sex, then, within each Sex category, the data will be sorted by Age.

3.2.5 Sorting Variables

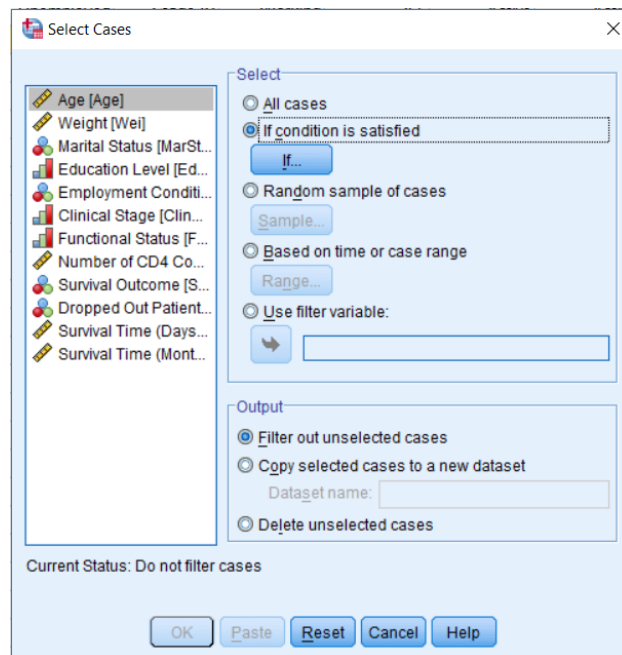
In order to sort the variables, from the *Menu* bar, click on **Data** → **Sort Variables**. In addition to sorting the variables by their names, they can also be sorted by the other **Variable View** characteristics.



3.2.6 Selecting Cases

Instead of just wanting to look at all possible values for a particular variable, we can analyze a specific subset of the data by selecting only certain cases in which we are interested. How would we do that? We would use a conditional statement.

To select cases, from the *Menu* bar, click on **Data** → **Select Cases**. Then, the **Select Cases** dialog box comes. Under the **Select** option, check on the **If condition is satisfied**.

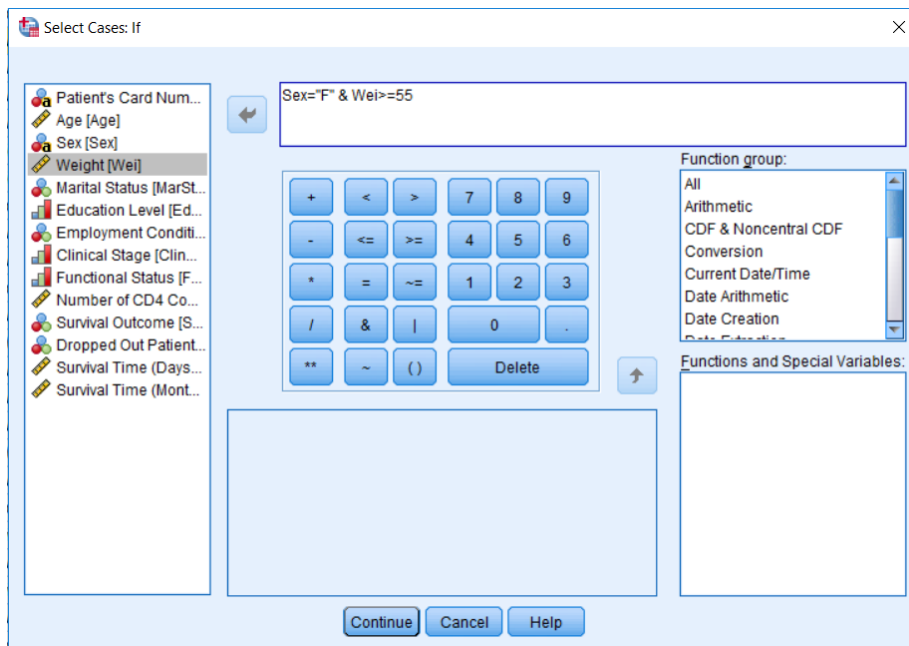


Under the **Output** options:

- **Filter out unselected cases** indicates the unselected cases in the *Data Editor* by placing a slash over the row numbers.
 - The unselected cases are removed from subsequent analyses.
 - The **All Cases** under the **Select** option should be reset.
- **Copy selected cases to a new dataset** saves the selected cases to a new dataset.
- **Delete unselected cases** removes unselected cases from the working dataset.

Example 3.3. Select (filter) female patients whose weight is greater than or equal to 55 and remove the unselected cases.

When the **If** button of the **Select Cases** dialogue box is clicked, the **Select Cases: If** dialogue box opens. Then, to select females whose weight is greater than or equal to 55, we do as follows.



Note: String variable values must be enclosed in double quotation marks.

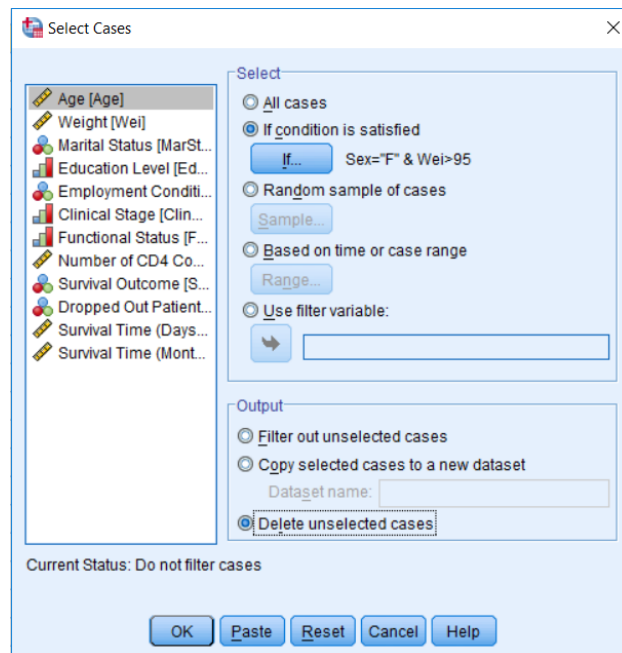
Click on **Continue** and then **OK**. Then, the *Data Editor* window looks:

	CardNum	Age	Sex	Wei	MarStat	EduLev	Emp	ClinStag	FunStat	CD4	Status	Defaulter	Days	SurvTime	filter_\$	var
1	1202	40	F	43	Separated	Primary	Unemployed	Stage IV	Working	365	Active	Active	809	27	Not Selected	
2	1203	50	M	58	Married	Primary	Parttime	Stage II	Bedridden	434	Active	Active	1000	33	Not Selected	
3	1207	35	F	55	Married	Primary	Unemployed	Stage IV	Working	270	Active	Active	837	28	Selected	
4	1209	45	M	60	Married	Tertiary	Fulltime	Stage I	Working	450	Active	Active	1530	51	Not Selected	
5	1211	44	M	65	Married	Secondary	Not Working	Stage II	Working	1352	Loss-to-foll...	Defaulted	91	3	Not Selected	
6	1212	30	F	73	Widow	Secondary	Fulltime	Stage III	Working	120	Active	Active	1770	54	Selected	
7	1214	63	F	72	Widow	Primary	Fulltime	Stage III	Working	217	Active	Active	1496	50	Selected	
8	1221	25	F	44	Separated	No Educati...	Unemployed	Stage III	Ambulatory	72	Loss-to-foll...	Defaulted	3	0	Not Selected	
9	1222	24	F	55	Married	Secondary	Unemployed	Stage I	Working	305	Active	Active	35	1	Selected	
10	1224	40	M	61	Married	Tertiary	Fulltime	Stage I	Ambulatory	279	Transferred	Defaulted	46	2	Not Selected	
11	1228	53	F	50	Separated	Secondary	Unemployed	Stage III	Ambulatory	436	Active	Active	1255	42	Not Selected	
12	1229	39	F	50	Married	Primary	Fulltime	Stage I	Ambulatory	213	Loss-to-foll...	Defaulted	33	1	Not Selected	
13	1230	33	M	62	Never Mari...	Secondary	Fulltime	Stage III	Working	68	Transferred	Defaulted	226	8	Not Selected	
14	1231	30	F	61	Married	Primary	Fulltime	Stage I	Working	1003	Active	Active	573	19	Selected	
15	1240	30	F	48	Married	No Educati...	Not Working	Stage I	Working	516	Active	Active	1022	34	Not Selected	
16	1242	27	F	63	Married	Secondary	Fulltime	Stage I	Working	368	Active	Active	986	33	Selected	
17	1244	43	F	46	Never Mari...	No Educati...	Not Working	Stage III	Ambulatory	38	Loss-to-foll...	Defaulted	1302	43	Not Selected	
18	1246	31	M	57	Married	Secondary	Fulltime	Stage III	Working	407	Dead	Defaulted	445	15	Not Selected	
19	1249	25	F	48	Separated	No Educati...	Unemployed	Stage II	Bedridden	122	Loss-to-foll...	Defaulted	6	0	Not Selected	
20	1250	30	F	56	Divorced	No Educati...	Fulltime	Stage II	Working	504	Loss-to-foll...	Defaulted	614	20	Selected	
21	1251	27	F	55	Married	Secondary	Fulltime	Stage II	Working	197	Transferred	Defaulted	921	31	Selected	
22	1252	30	F	36	Divorced	Primary	Unemployed	Stage III	Ambulatory	57	Loss-to-foll...	Defaulted	4	0	Not Selected	
23	1253	27	F	58	Married	No Educati...	Parttime	Stage IV	Working	474	Loss-to-foll...	Defaulted	1498	50	Selected	

From now onwards, those cases having a slash over on the row numbers are deactivated, that means, they will not be included in the subsequent analysis. Hence, do not forget to reset **All Cases** under the **Select** option of **Select Cases** dialogue box which makes all cases active for analyses.

Example 3.4. Select females whose weight is greater than 95 and remove unselected cases.

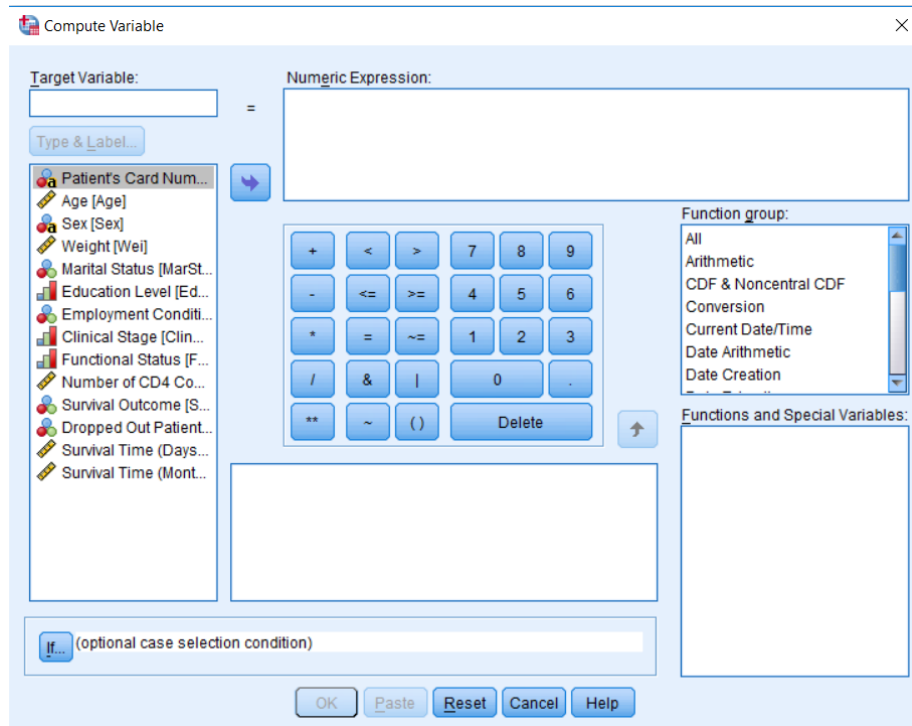
To select female patients whose weight is greater than 95 and remove the unselected cases, the **Select Cases** dialogue box looks:



After clicking the **OK** tab, you can easily observe that there are only 2 cases as shown in the **Data View** sheet of the *Data Editor*. All the unselected cases are removed from the working file.

3.2.7 Creating New Variable

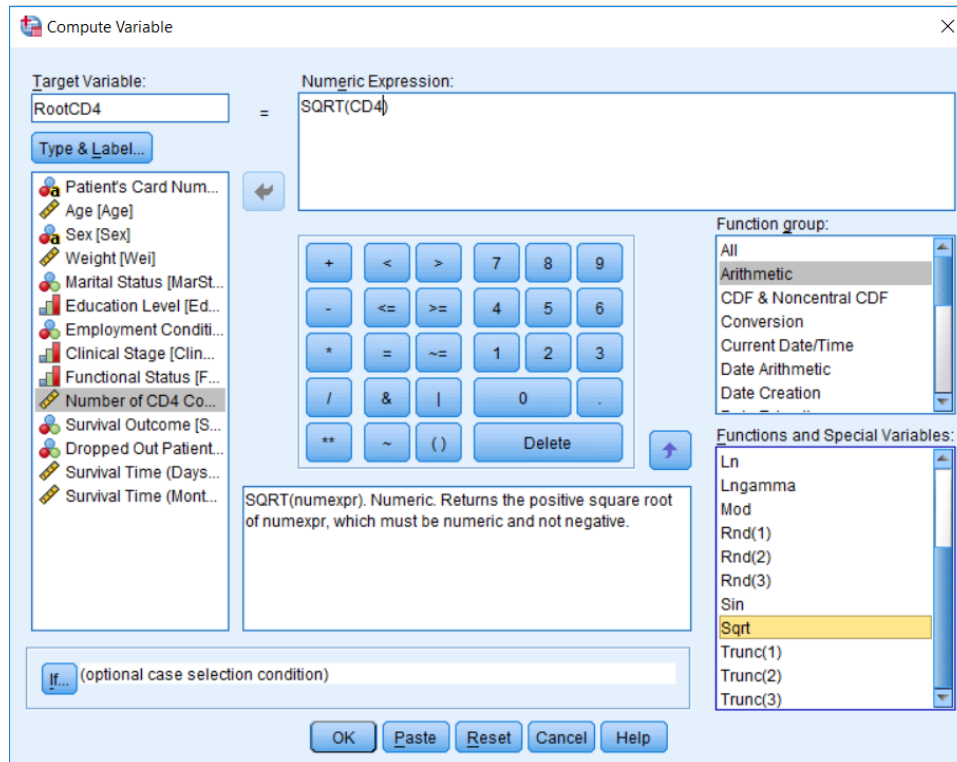
Variable transformation is a way of creating new variables using existing continuous variables and formulae. To create a new variable, from the *Menu* bar, click on **Transform** → **Compute**. This opens the **Compute Variable** dialog box.



Example 3.5. Create new variable by taking the square root of the CD4 variable.

Now to compute the square root of the CD4 variable,

1. First, write the new variable name, RootCD4 in the **Target Variable:** field.
2. Next from the **Function group:** list select **Arithmetic**.
3. And then from **Functions and Special Variables:** list select **Sqrt** and enter into the **Numeric Expression:** box (any formula can be written as function of existing variables and/or numbers).
4. Lastly, inside the brackets of the square root, enter the CD4 variable.



You can also specify the type and label for the new variable. The **If** button can be used to case selection condition.

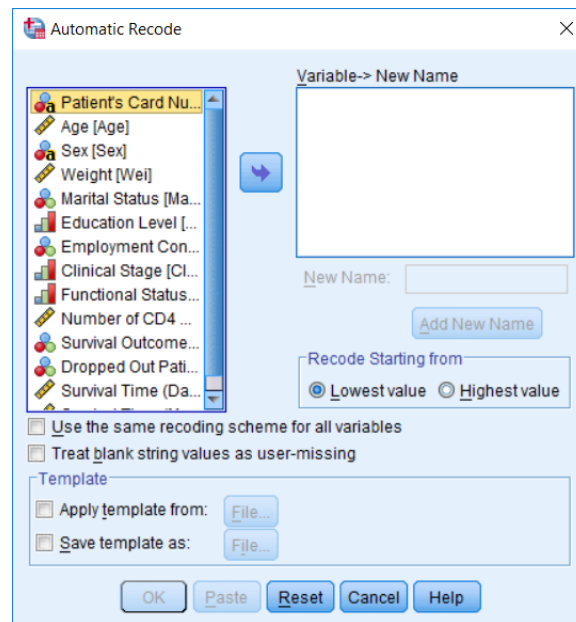
Exercise 3.1. Generate a new variable *AgeSquare* by squaring the *Age*.

Exercise 3.2. Generate a new variable *LogCD4* by taking the logarithm of *CD4*.

3.2.8 Recoding a String Variable

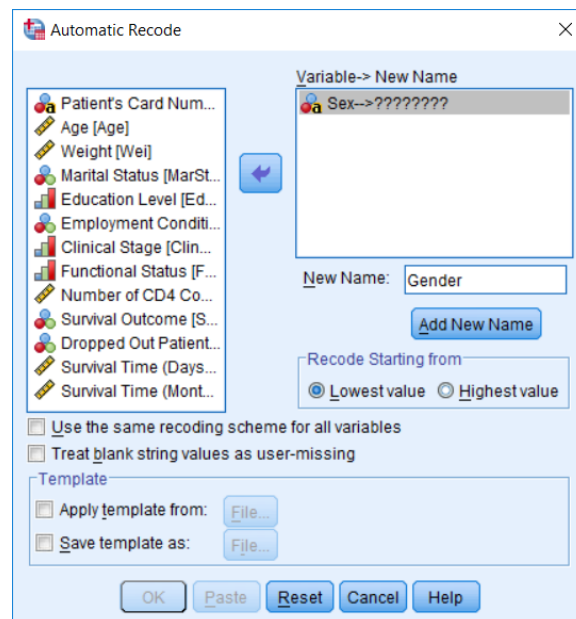
If there is a string variable, such as *Sex* where the values are "F" and "M" in our case, we may want to assign numeric values to them.

From the *Menu* bar, by clicking on **Transform** → **Automatic Recode**, the **Automatic Recode** dialogue box opens. In the **Variable** → **New Name** box, enter the string variable.



Example 3.6. Automatically recode the string variable Sex.

In the **Automatic Recode** dialogue box, enter **Sex** in the **Variable–>New Name** box.



Next in the **New Name:** field write **Gender** and click on the **Add New Name** button to add the new name into **Variable–>New Name** box. Click on **OK**. This automatically assigns the values 1 for 'F' and 2 for 'M' and labels to the categories of Gender.

3.2.9 Recoding a Categorical Variable

For Gender, the numeric values for 'Female' and 'Male' patients are 1 and 2, respectively. Suppose, that is not what we want. We want the typical 0='F', 1='M' setup. How might we

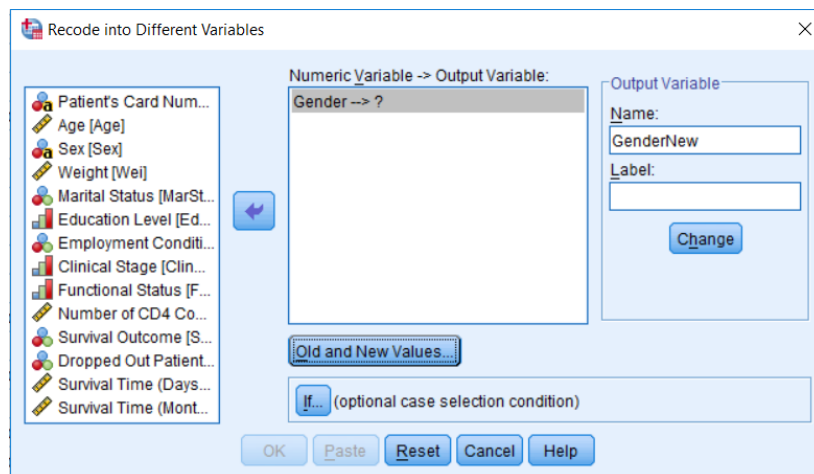
do this? We will be looking how to recode this categorical variable. There are two options available for recoding variables. The first option is *recoding values into the same variable* which eliminates all record of the original values and the second one is *creating a new variable* containing the recoded values.

From the *Menu* bar, by clicking on **Transform** → **Recode into Different Variables**, then a dialogue box opens.

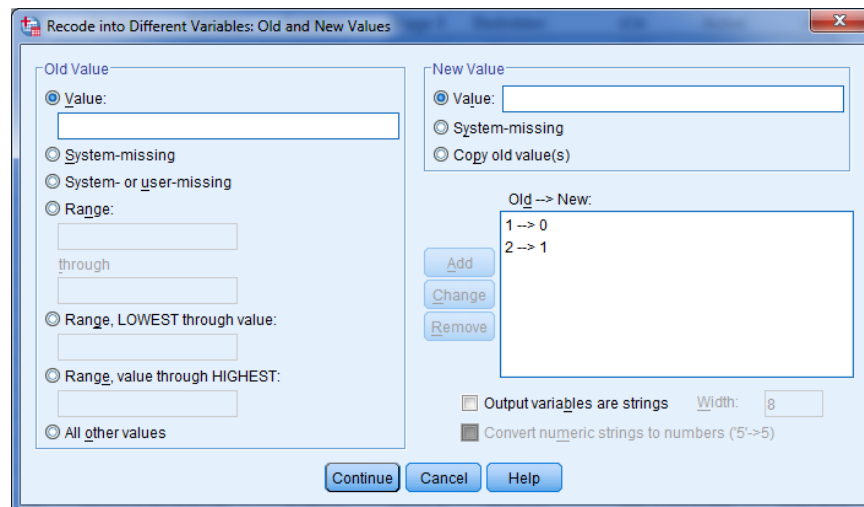
Example 3.7. Recode Gender so that 0 stands for 'F' and 1 stands for 'M'.

In the **Recode into Different Variables** dialogue box,

1. In the **Numeric Variable**–>**Output Variable** box, enter Gender.



2. Next in the **Output Variable** option **Name:** field write a new variable name, say, GenderNew and click on the **Change** button to add the new name into **Numeric Variable**–>**Output Variable** box.
3. When you click on the **Old and New Values** button, the following dialogue box comes.
 - (a) In the **Old Value** option **Value:** field, type 1 and in the **New Value** option **Value:** field, type 0 then click on the **Add** tab.
 - (b) Next in the **Old Value** option **Value:**, field, type 2 and in the **New Value** option **Value:** field type 1 then click on the **Add** tab.



(c) Click on the **Continue** tab and **OK**.

Exercise 3.3. Recall Employment Condition (Emp) is coded as 0= Full-time, 1= Part-time, 2= Not Working and 3= Unemployed. Now recode this variable using 0= Working (Full-time and Part-time), 1= Not Working and 2= Unemployed setup.

Exercise 3.4. Survival outcome (Status) is coded as 0= Active, 1= Dead, 2= Transferred, and 3= Loss-to-follow. Now recode this variable using 0= Alive (Active and Transferred), 1= Dead and 2= Lost-to-follow setup.

3.2.10 Recoding a Continuous Variable

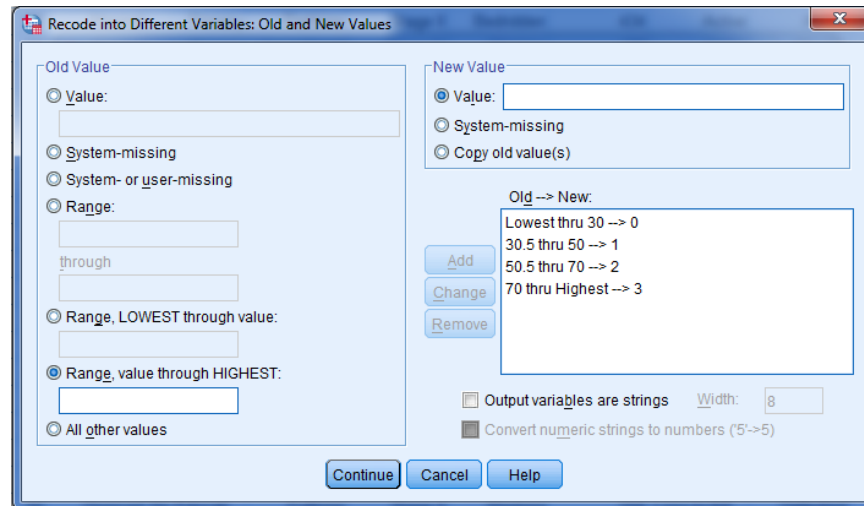
Some times, it is also useful to collapse a continuous variable into categorical groups. From the *Menu* bar, by clicking on **Transform** → **Recode into Different Variables**, then the usual dialogue box opens.

Example 3.8. Categorize the Wei variable into four categories: 0=Min-30, 1=30.5-50, 2=50.5-70 and 3=70.5-Max. Then, create their value labels.

In the **Recode into Different Variables** dialog box,

1. In the **Numeric Variable** → **Output Variable** box, enter Wei.
2. Next in the **Output Variable** option **Name:** field write a new variable name, say, WeiCat and click on the **Change** button to add the new name into **Numeric Variable** → **Output Variable** box.
3. Click on the **Old and New Values** button.
 - (a) In the **Old Value** option in the **Range, LOWEST through Value:** field, type 30 and in the **New Value** option **Value:** field, type 0 then click on the **Add** tab.
 - (b) Next in the **Old Value** option **Range:** fields, type 30.5 and 50, respectively and in the **New Value** option **Value:** field, type 1 then click on the **Add** tab.
 - (c) Thirdly, in the **Old Value** option **Range:** fields, type 50.5 and 70, respectively and in the **New Value** option **Value:** field, type 2 then click on the **Add** tab.

- (d) Lastly, in the **Old Value** option in the **Range, value through HIGHEST:** field, type 70 and in the **New Value** option **Value:** field, type 3 then click on the **Add** tab.



- (e) Click on **Continue** and **OK**.

Exercise 3.5. Categorize Age into three categories: 0=Min-29, 1=30-39 and 2=40-Max. Then, create their value labels.

3.3 Combining Datasets

It is often needed to merge several datasets into one. There are two main ways of merging datasets. The first situation is when we have two datasets with the *same variables but different cases* and the second situation is when we have two datasets with *same cases but different variables*.

The data file in memory (the one that is currently opened) is referred to as the 'master' or 'working' file. The file that is to be joined with the 'master' file is known as the 'using' data file.

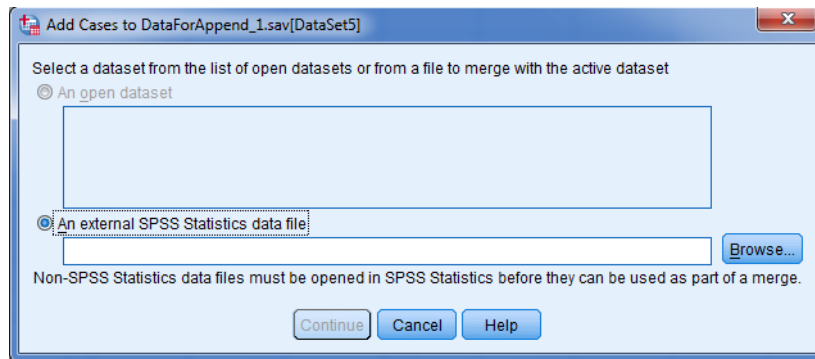
3.3.1 Adding Cases (Observations)

Here, the two data files are considered to have *different observations but same variables*. Hence, observations from the using file are added to the end of the working data file, that is, the files are stacked vertically. But, be sure that each variable should have *same name and same data type* in both data files.

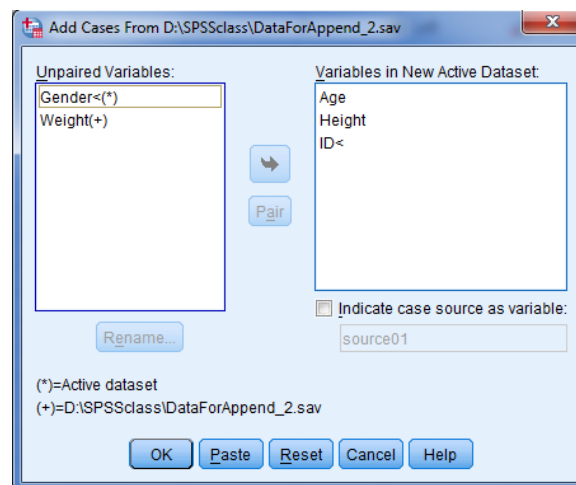
From the *Menu* bar, click on **Data** → **Merge Files** → **Add Cases**. This opens the **Add Cases to ...** dialogue box having the **Browse** button that helps to locate where the using file is saved.

Example 3.9. In the folder given to you, there are two data files named **DataForAppend_1.sav** and **DataForAppend_2.sav** having some same variables but different cases. Add the cases from **DataForAppend_2.sav** to **DataForAppend_1.sav**.

First, open the `DataForAppend_1.sav` data and click on **Data** → **Merge Files** → **Add Cases**. In the **Add Cases to ...** dialogue box, click on the **Browse** button to locate where the using file (`DataForAppend_2.sav`) is saved.



After browsing and selecting the using file, click on the **Continue** tab. Then the **Add Cases From ...** dialogue box with the list of variables in two boxes appears as shown below. The **Unpaired Variables:** box contains variables to be excluded from the new appended data file (due to the difference in names and/or type) while the **Variables in New ...:** box contains variables to be included in the new appended data file.



Click on the **OK** tab.

Note:

- If the same information is recorded under different variable names in the two files, you can create a pair from the **Unpaired Variables:** list. Select the two variables on the list and click on **Pair**.
- To include an unpaired variable from one file without pairing it with a variable from the other file, select the variable from the **Unpaired Variables:** list and add it to **Variables in New ...:** list. Any unpaired variable included in the merged file will contain missing data for cases from the file that does not contain that variable.

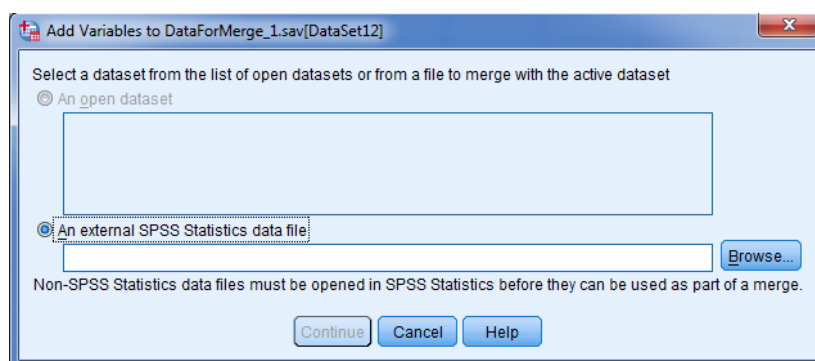
3.3.2 Adding Variables

Unlike the previous merging, the two data files are now considered to have *different variables but same observations*. The additional variables from the using file are added to the data file in memory (the files are stacked horizontally). A necessary condition is that both datasets should have a unique identifier of each observation which might consist of a single variable or a series of variables. In other words, both files should have a matching variable (or variables) that is (are) used to associate an observation from the master file with an observation in the using file. Before trying to merge, both datasets should be sorted by the matching variable. If two or more variables are used to match cases, the two data files must be sorted by ascending order of these key variables.

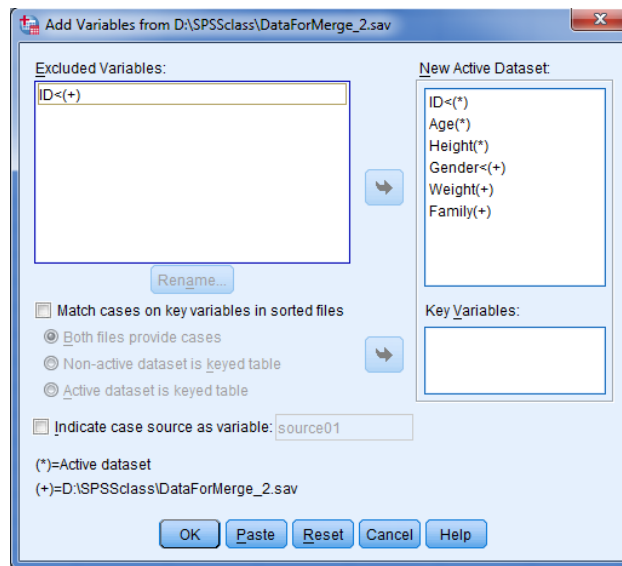
From the *Menu* bar, click on **Data** → **Merge Files** → **Add Variables**. This opens the **Add Variables to ...** dialogue box having the **Browse** button that helps to locate where the using file is saved.

Example 3.10. There are two data files named `DataForMerge_1.sav` and `DataForMerge_2.sav`, in the folder given to you. The **ID** variable is an identifier (matching variable) in both data files. Merge the variables from `DataForMerge_2.sav` to `DataForMerge_1.sav`.

First, open `DataForMerge_1.sav` data. In the **Add Variables to ...** dialogue box, click on the **Browse** button to locate where the using file (`DataForMerge_2.sav`) is saved.



After browsing and selecting the file to be merged, click on the **Continue** tab. Then the **Add Variables from ...** dialogue box with the list of variables appears.



The **Excluded Variables:** box contains variables to be excluded from the new merged data file. Variable names in the second data file that duplicate variable names in the working data file are excluded by default because it assumes that these variables contain duplicate information. To include an excluded variable with a duplicate name to the merged file, rename it and add it to the **New Active Dataset:** box.

Next, check on the **Match cases on key** ... and enter the key variable(s) to **Key Variables:** list. In the case of two or more key variables, the order of these variables on the **Key Variables:** list must be the same as their sort sequence.

Chapter 4

Descriptive Analysis

After the data have entered in the data editor, the first step to complete is to do descriptive analysis, which involves computing various summary statistics (for example; min, max, mean, standard deviation, skewness, kurtosis) and graphical displays (for example; frequency tables, boxplots, stem and leaf plots, histogram). In the previous chapter, we have already started describing data using the **Codebook** procedure and this chapter focuses on more descriptive analysis. Descriptive statistical analysis is important for several reasons:

- To see if there are problems in the data such as outliers, coding problems, missing values and/or other errors.
- To get basic information regarding the demographic variables of the data to report in the Results section of a research report.
- To examine the extent to which the statistical assumptions (like normality, homogeneity of variances) are fulfilled.
- To examine the relationships between variables, and determine the strength and direction of the association.

A statistical analyses will logically proceed from simple to complex. Mostly, a statistical analysis will be conducted in three phases: univariable, bivariabile, and multivariable data analyses.

- A univariable analysis is when analyzing one variable at a time.
- A bivariabile analysis is when analyzing the relationship between two variables.
- A multivariable analysis is when analyzing several variables together.

The results of univariable and bivariabile analyses will be used to select methods and variables to be used in the multivariable analysis(es).

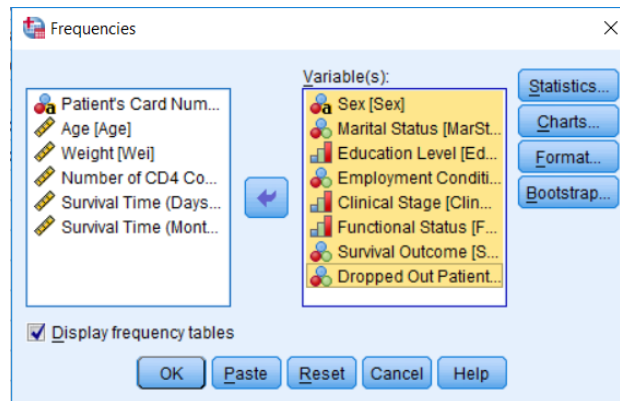
4.1 Frequency Tables

Frequency tables are used to summarize categorical variables. These could help us to understand how many observations are in each category (level) of a variable and how much are missing.

One-Way Frequency Tables

To construct one-way frequency tables, from the *Menu* bar, click on **Analyze** → **Descriptive Statistics** → **Frequencies**. In the **Frequencies** dialogue box, enter at least one categorical variable in the **Variable(s):** box.

Example 4.1. Obtain the one-way frequency tables for all categorical variables in the JUSH_HAART.sav data.



The first two frequency tables of the output are:

Sex					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	930	63.5	63.5	63.5
	M	534	36.5	36.5	100.0
	Total	1464	100.0	100.0	

Marital Status					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Never Married	293	20.0	20.1	20.1
	Married	739	50.5	50.6	70.7
	Divorced	134	9.2	9.2	79.9
	Separated	140	9.6	9.6	89.5
	Widowed	154	10.5	10.5	100.0
	Total	1460	99.7	100.0	
Missing	System	4	.3		
	Total	1464	100.0		

The result indicates that 930 (63.5%) of the patients were females and the remaining 534 (36.5%) were males. Regarding the marital status of the patients, 293 (20%), 739 (50.5%), 134 (9.2%), 140 (9.6%) and 154 (10.5%) were never married, married, divorced, separated and widowed, respectively. But note that there are 4 (0.3%) missing observations.

Multi-Way Frequency Tables

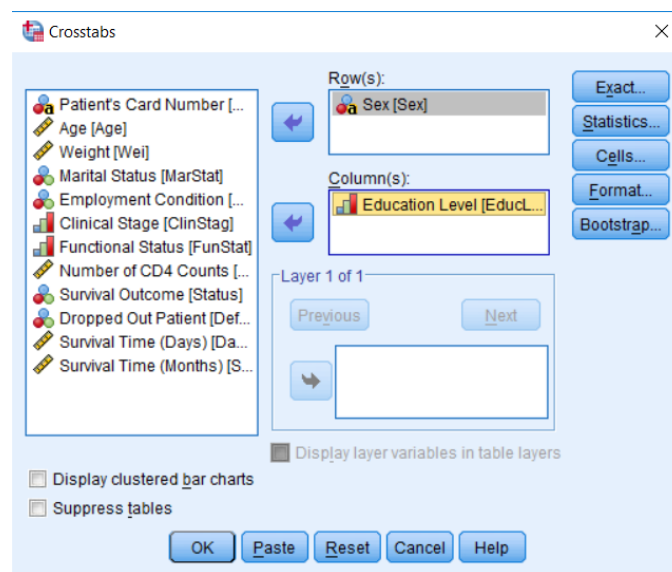
For two or more categorical variables, the data can be summarized in a tabular form in which the cells of the table contain number of observations (frequencies) in the intersection categories of the variables. Such a table is called *contingency table* (*cross-tabulation*).

The **Crosstabs** procedure in SPSS forms two-way and multi-way contingency tables, and provides a variety of tests and measures of association for two-way contingency tables only.

For constructing cross-tabulations, from the *Menu* bar, click on **Analyze** → **Descriptive Statistics** → **Crosstabs**. In the **Crosstabs** dialogue box, enter at least one categorical variable in the **Row(s):** box and another one in the **Column(s):** box.

Example 4.2. Construct the cross-tabulation of Sex and Education Level, and examine the frequencies.

In the **Crosstabs** dialogue box, enter Sex in **Row(s):** box and Education Level in **Column(s):** box as show below and click the **OK** tab.

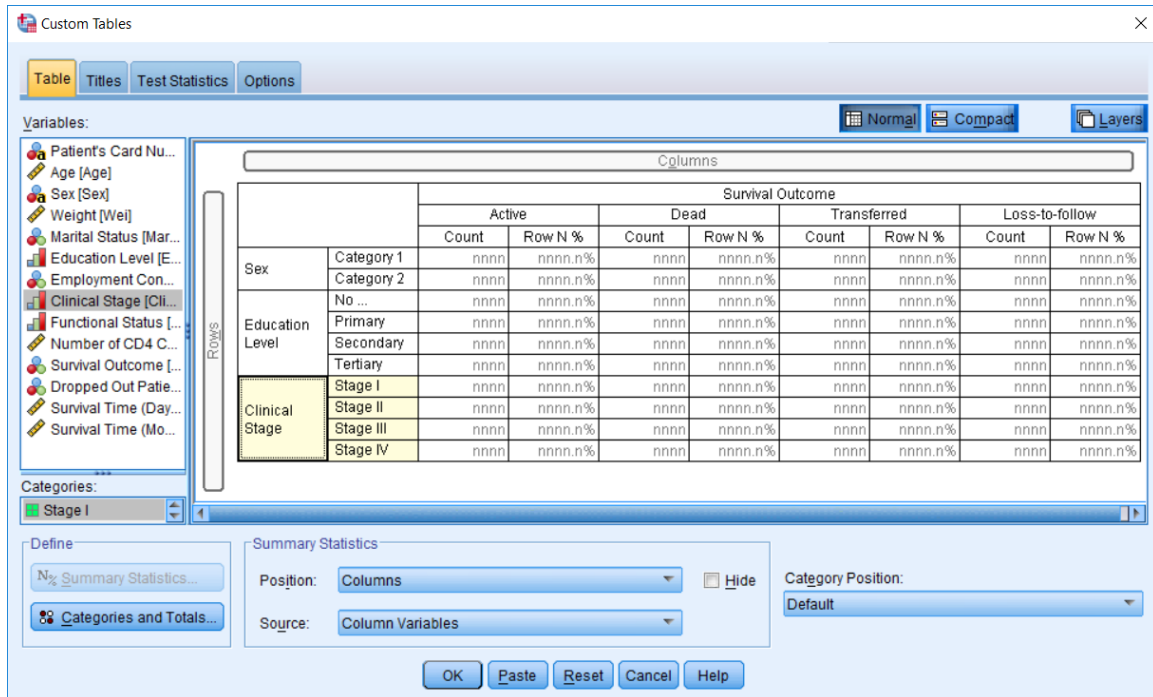


The crosstab result is:

Count		Education Level				Total
		No Education	Primary	Secondary	Tertiary	
Sex	F	211	325	316	73	925
	M	86	190	176	82	534
Total		297	515	492	155	1459

There were 211 female patients who did not have education, 325 females with primary education, Of the total 534 males, 86 of them had no education, 190 of them were having primary education,

SPSS can also do custom tables which describe the relationship between variables in a table of frequencies. These tables can either be simple two-way tables or multi-way tables. To do so, from the *Menu* bar, click on **Analyze** → **Tables** → **Custom Tables**. In the **Custom Tables** dialogue box, select and drag the variable(s) to be in the row and column and click the **OK** tab.

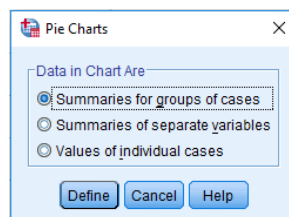


4.2 Constructing Pie and Bar Charts

4.2.1 Pie Chart

Pie chart is popularly used in practice to show the percentage break down of a single qualitative variable data. It is a circle divided into a number of slices whose size corresponds to the relative frequency of each class.

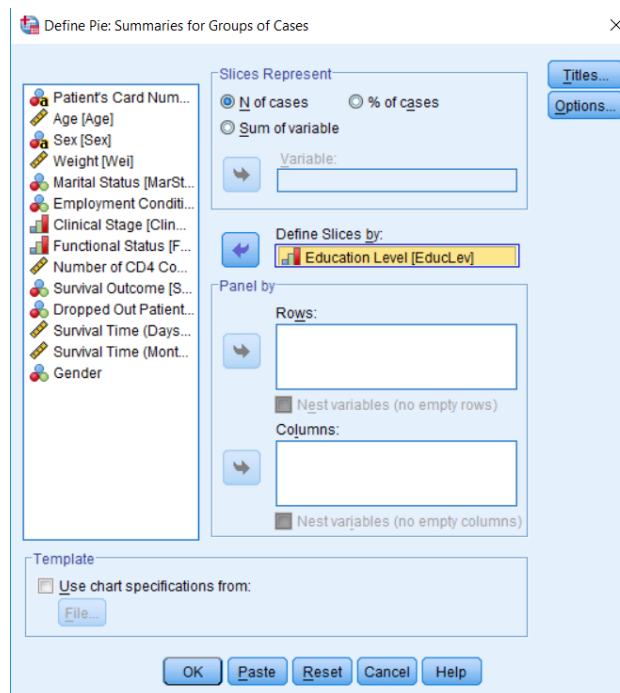
From the *Menu* bar, click on **Graphs** → **Legacy Dialogs** → **Pie**. Then, the **Pie Charts** dialogue box appears.



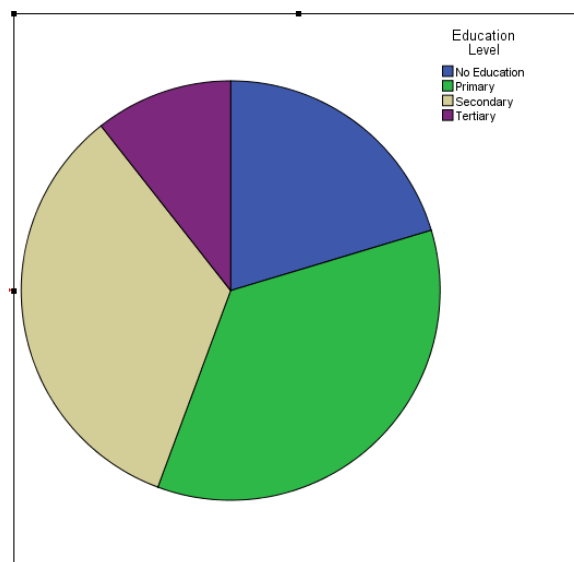
Click on the **Define** button which will open the **Define Pie: ...** dialogue box.

Example 4.3. Construct pie chart for education level (EducLev).

In the **Define Pie: ...** dialogue box, enter EducLev in the **Define Slices by:** box.

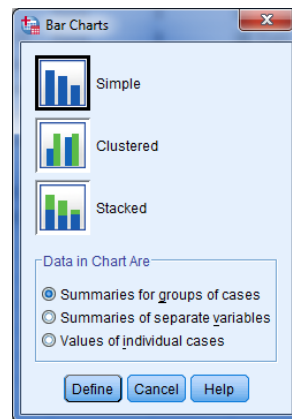


Then click on **OK**. Then, the pie chart looks as follows.



4.2.2 Bar Charts

From the *Menu* bar, click on **Graphs** → **Legacy Dialogs** → **Bar**. Then, the **Bar Charts** dialog box appears.

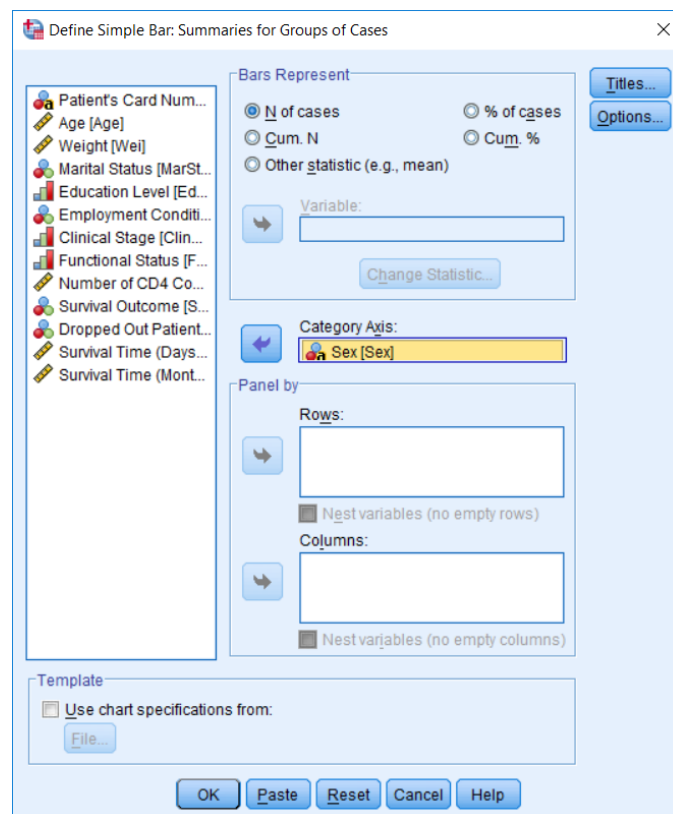


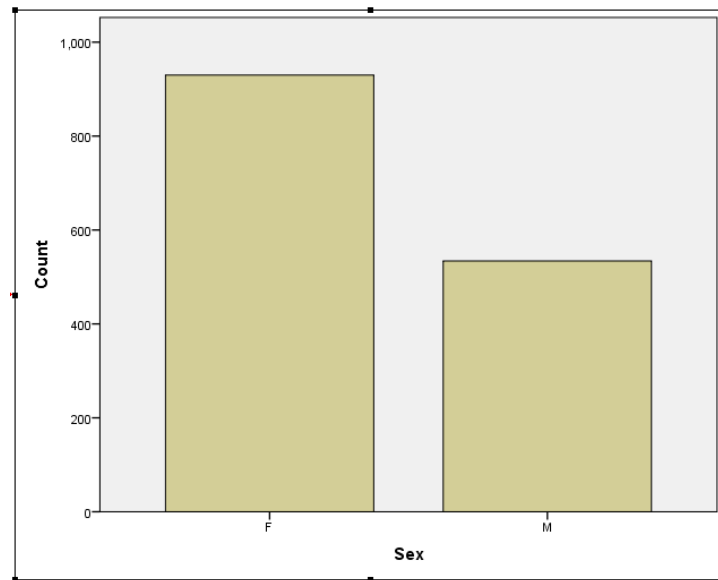
Simple Bar Chart

In a simple bar chart, the categories of a qualitative variable (possibly a discrete variable with a small number of values) are marked on the X axis and the frequencies (magnitude) of the categories are marked on the Y axis.

Example 4.4. Construct simple bar chart for Sex.

Of the three types of bar charts in the **Bar Charts** dialogue box, select the first one, that is, the **Simple** option. Then click on the **Define** button. In the **Category Axis:** box enter Sex and then **OK**.



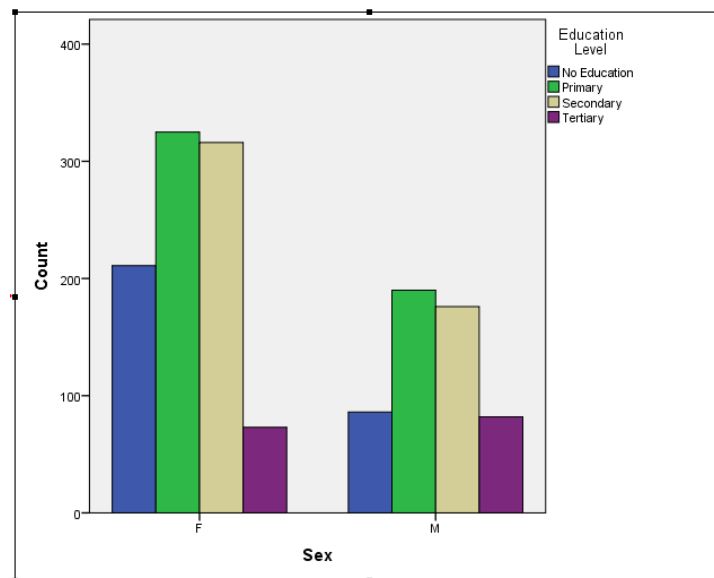
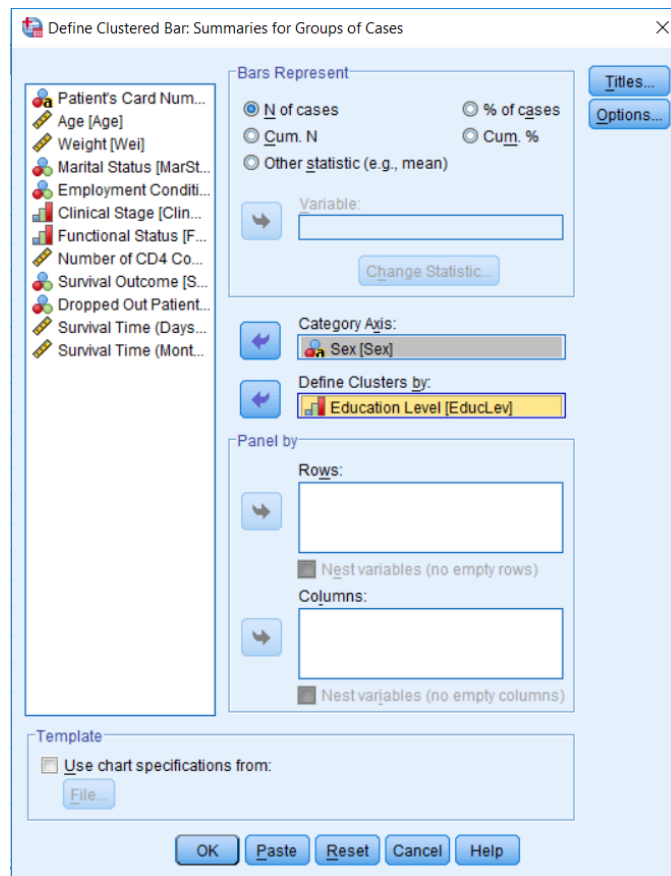


Multiple (Clustered) Bar Chart

Multiple (clustered) bar chart is used to display data on more than one variable. In the multiple bars diagram two or more sets of inter-related data are interpreted.

Example 4.5. Construct multiple bar chart for Sex and Education Level.

Again, of the three types of bar charts in the **Bar Charts** dialogue box, select the second one, that is, the **Clustered** option. Then click on the **Define** button. In the **Category Axis:** box enter Sex and in the **Define Clusters by:** enter Education Level. Then **OK**.

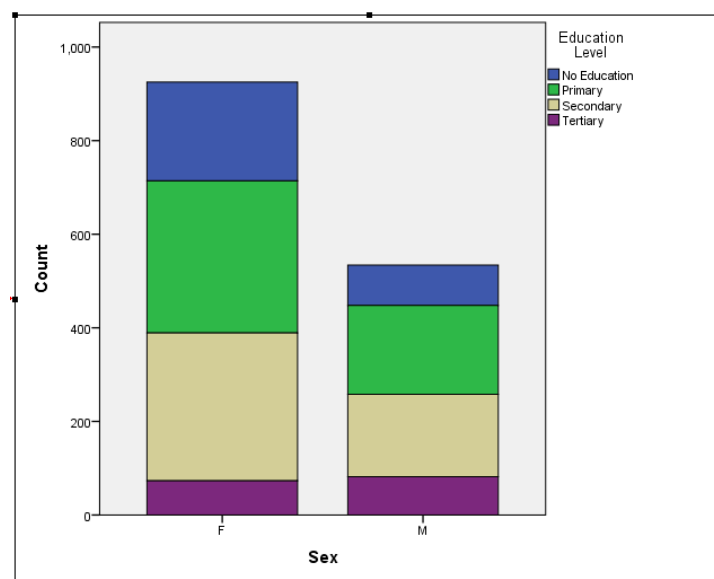
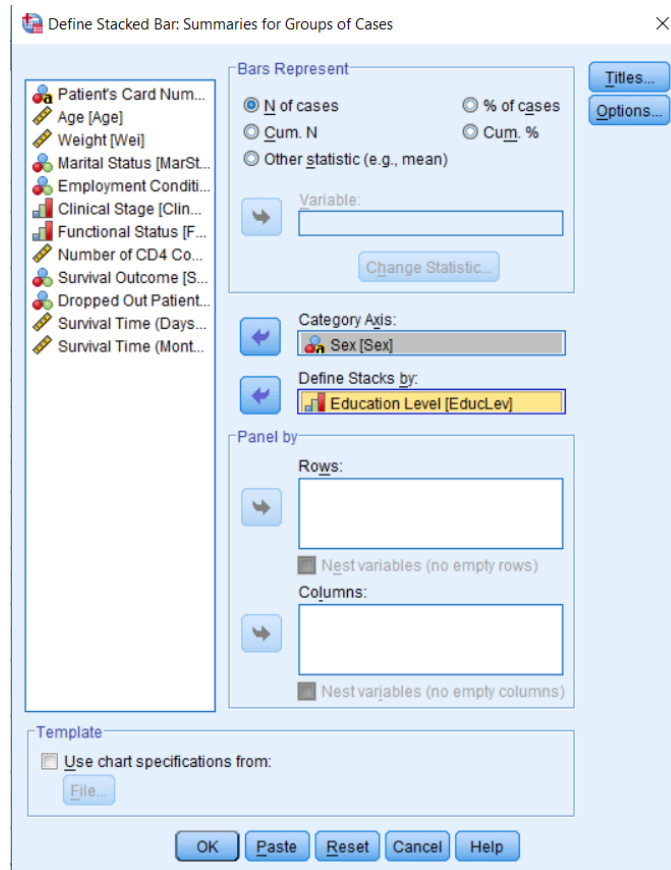


Component (Stacked) Bar Chart

Component (stacked) bar chart is used when there is a desire to show a total or aggregate is divided into its component parts.

Example 4.6. Construct multiple bar chart for Sex and Education Level, and compare it with the clustered bar chart above.

Lastly, of the three types of bar charts in the **Bar Charts** dialogue box, select the third one, that is, the **Stacked** option. Then click on the **Define** button. In the **Category Axis:** box enter **Sex** and in the **Define Stacks by:** enter **Education Level**. Then click on **OK**.



4.3 Graphs for Scale Variables

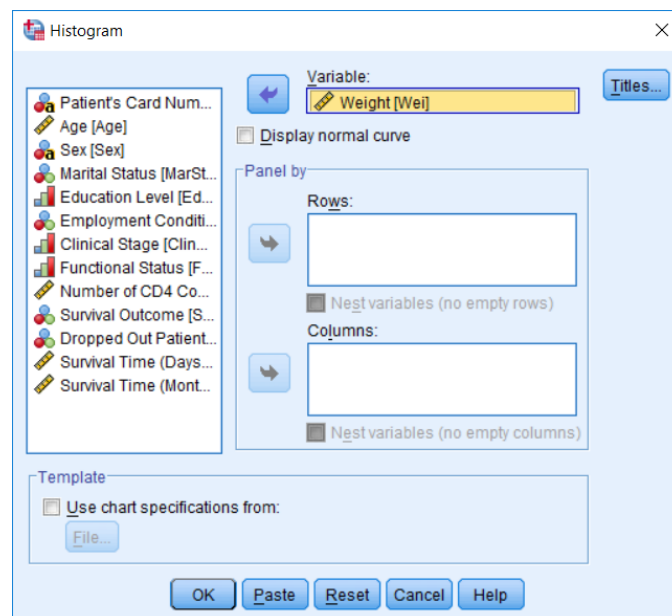
4.3.1 Histogram

Histogram is the most common graphical presentation used for quantitative variables for assessing normality. The trouble is that visual inspection of histograms can be deceiving because some approximate normal distributions do not look like a normal curve.

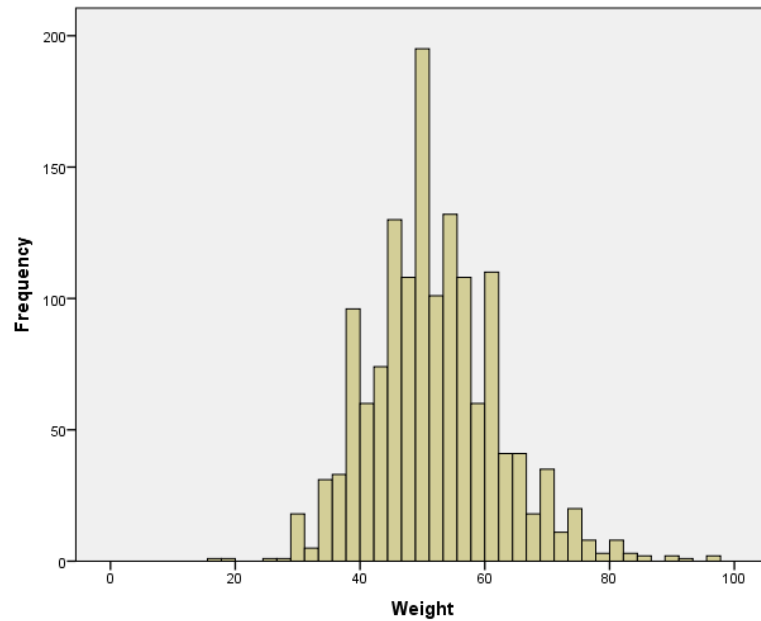
From the *Menu* bar, click on **Graphs** → **Legacy Dialogs** → **Histogram**.

Example 4.7. Construct histogram for Wei and say something about the distribution.

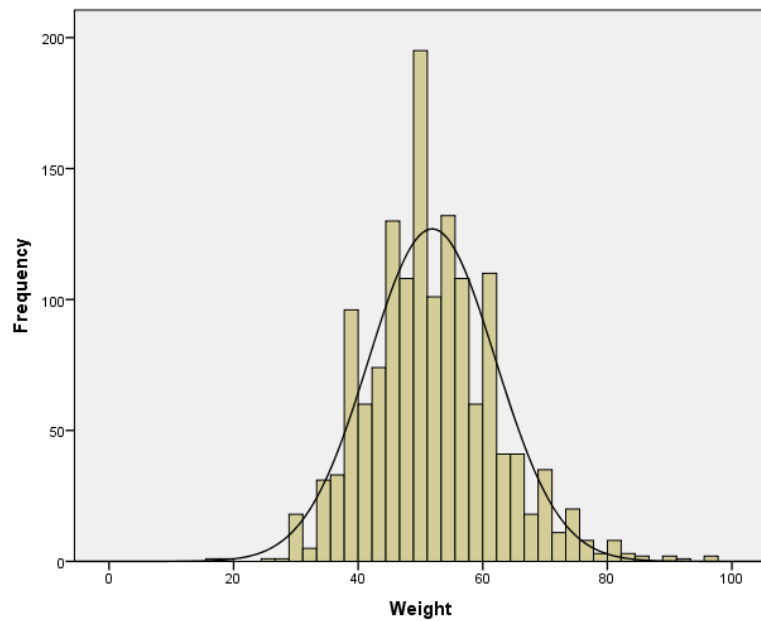
In the **Variable:** box of the **Histogram** dialogue box enter Wei.



Then, after clicking on **OK**, the histogram of weight of the patients is displayed as follows.



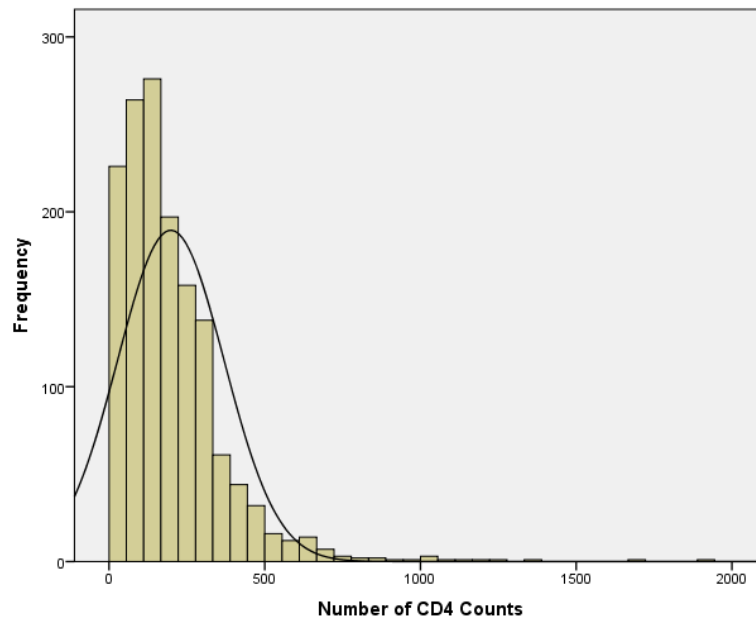
Or by checking on the **Display normal curve** of the **Histogram** dialogue box, the theoretical normal distribution curve will be superimposed in the histogram as shown below.



From this histogram, it seems the weight of the patients approximately follows a normal distribution.

Example 4.8. Construct histogram for CD4.

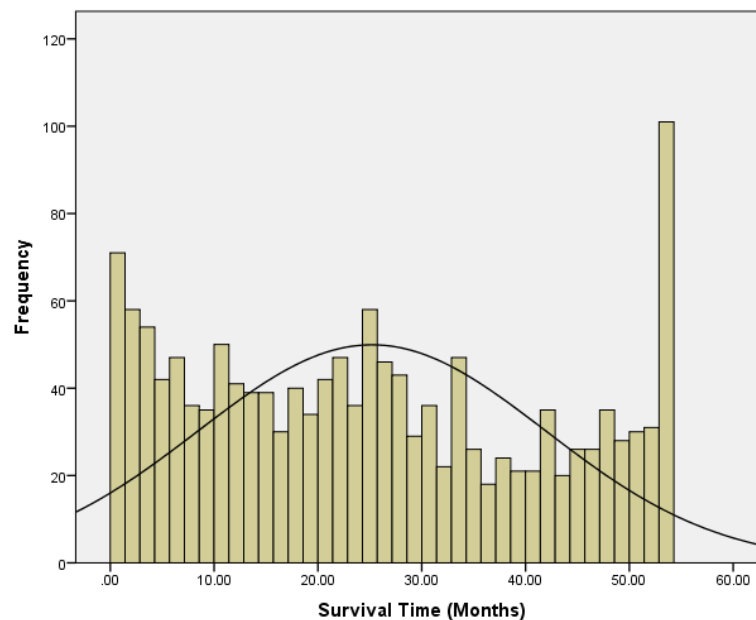
Using similar procedure as above, the histogram of the CD4 is as follows.



This plot shows CD4 count is not normally distributed (positively skewed).

Example 4.9. Construct histogram for the survival time `SurvTime` of patients.

The histogram of `SurvTime` shown below indicates `SurvTime` does not follow a normal distribution.



4.4 Basic Summary Statistics

For quantitative variables, it is necessary to calculate certain indicators like measures of central tendency and measures of variation.

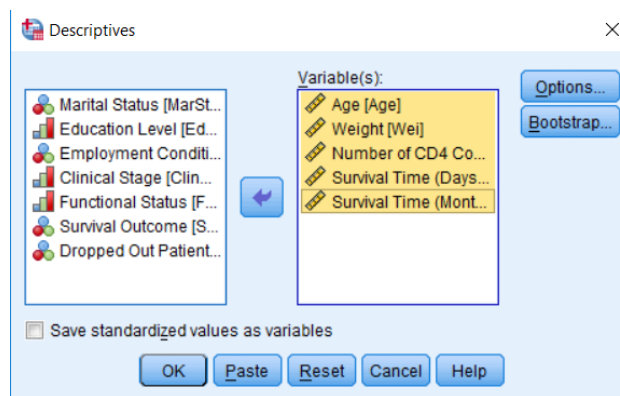
4.4.1 Central Tendency and Variation

The **Descriptives** procedure displays univariate summary statistics for quantitative (scale) variables in a single table.

From the *Menu* bar, click on **Analyze** → **Descriptive Statistics** → **Descriptives**. In the **Descriptives** dialogue box, enter at least one quantitative variable in the **Variable(s):** box in the usual manner.

Example 4.10. Obtain descriptive statistics for all scale variables of the JUSH_HAART.sav data.

In the **Descriptives** dialogue box, all quantitative variables are entered in **Variable(s):** box as follows.



Here is the output:

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Age	1464	18	85	34.01	9.160
Weight	1460	16	96	51.87	10.193
Number of CD4 Counts	1464	1	1914	198.19	171.240
Survival Time (Days)	1464	1	2141	761.66	511.852
Survival Time (Months)	1464	.03	54.00	25.1977	16.70695
Valid N (listwise)	1460				

The minimum and maximum ages of the 1464 patients are 18 and 85 years, respectively. The average age is 34.01 years with a standard deviation of 9.16 years. The average weight is 51.87 kilograms with a standard deviation of 10.19 kilograms (note that these values are calculated from 1460 patients which means there are 4 weight missing values) with a minimum weight of 16 and a maximum weight of 96 kilograms. The average number of CD4 counts is 198.19 with a standard deviation of 171.240. The same is true for the remaining variables.

Also, by clicking on the **Options** button of **Descriptives** dialogue box, we can select additional measures like range, variance, standard error for the mean, kurtosis and skewness.

4.4.2 Exploring Descriptive Statistics by Group

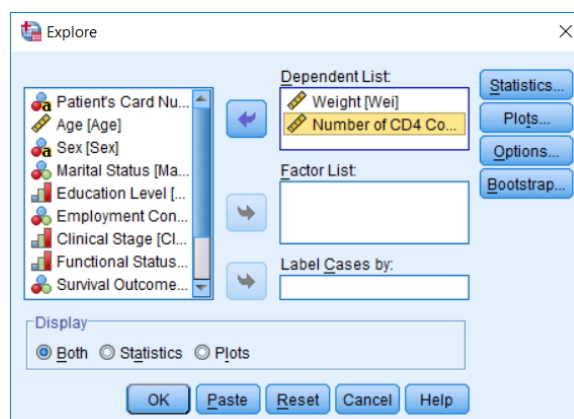
The **Explore** procedure produces summary statistics and graphical displays, either for *all cases* or separately for *groups of cases* (because, sometimes, it is also necessary to explore the scale variable(s) within each category of a categorical variable). It is used for data screening, outlier identification, normality assumption checking, and characterizing differences among groups of cases. Such data screening may help to identify the presence of unusual values, extreme values, gaps in the data or other peculiarities.

From the *Menu* bar, click on **Analyze** → **Descriptive Statistics** → **Explore**. In the **Explore** dialogue box, enter at least one quantitative variable in the **Dependent List:** box.

Example 4.11. Obtain descriptive statistics using the **Explore** procedure for the Wei and CD4 variables.

In the **Explore** dialogue box, enter both Wei and CD4 in the **Dependent List:** box. It is also possible to select an identification variable to label cases and enter to **Label Cases By:** box.

Also, the five largest and five smallest values with case labels can be displayed if the **Outliers** option is selected under the **Statistics** tab.

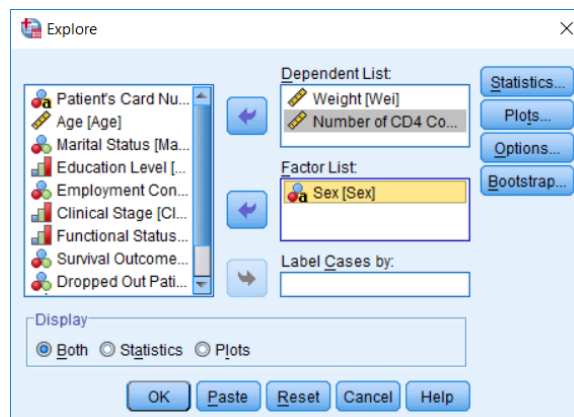


The partial result looks:

Descriptives				
			Statistic	Std. Error
Weight	Mean		51.87	.267
	95% Confidence Interval for Mean	Lower Bound	51.34	
		Upper Bound	52.39	
	5% Trimmed Mean		51.55	
	Median		51.00	
	Variance		103.905	
	Std. Deviation		10.193	
	Minimum		16	
	Maximum		96	
	Range		80	
	Interquartile Range		13	
	Skewness		.519	.064
	Kurtosis		.909	.128
Number of CD4 Counts	Mean		198.69	4.481
	95% Confidence Interval for Mean	Lower Bound	189.90	
		Upper Bound	207.48	
	5% Trimmed Mean		180.60	
	Median		159.00	
	Variance		29311.335	
	Std. Deviation		171.206	
	Minimum		1	
	Maximum		1914	
	Range		1913	
	Interquartile Range		180	
	Skewness		2.877	.064
	Kurtosis		16.429	.128

Example 4.12. Explore the Wei and CD4 variables within each category of Sex.

Now in the **Explore** dialogue box, enter both Wei and CD4 in the **Dependent List:** box and enter Sex in the **Factor List:** box.

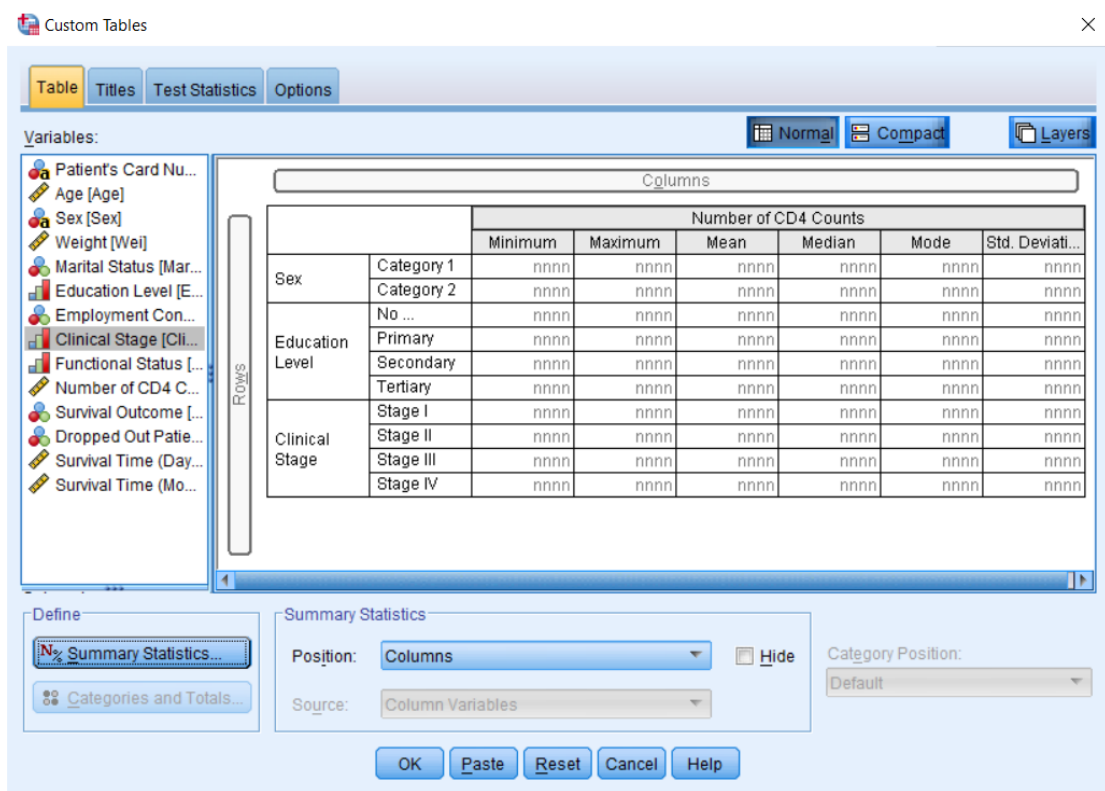


After clicking on the **Statistics** tab, the partial outputs are:

Descriptives					
Sex			Statistic	Std. Error	
Weight	F	Mean		49.68	.326
		95% Confidence Interval for Mean	Lower Bound	49.04	
			Upper Bound	50.32	
		5% Trimmed Mean		49.24	
		Median		49.00	
		Variance		98.344	
		Std. Deviation		9.917	
	Minimum		16		
	Maximum		96		
	Range		80		
	Interquartile Range		12		
	Skewness		.741	.080	
	Kurtosis		1.615	.160	
	M	M	Mean		55.67
95% Confidence Interval for Mean			Lower Bound	54.86	
			Upper Bound	56.48	
5% Trimmed Mean				55.49	
Median				55.00	
Variance				90.992	
Std. Deviation				9.539	
Minimum			29		
Maximum			92		
Range			63		
Interquartile Range			11		
Skewness			.347	.106	
Kurtosis			.828	.211	

Number of CD4 Counts	F	Mean		209.71	5.842	
		95% Confidence Interval for Mean	Lower Bound	198.25		
			Upper Bound	221.18		
		5% Trimmed Mean		191.43		
		Median		173.00		
		Variance		31739.389		
		Std. Deviation		178.156		
		Minimum		2		
		Maximum		1914		
		Range		1912		
		Interquartile Range		192		
		Skewness		2.929	.080	
		Kurtosis		17.435	.160	
		M	M	Mean		178.13
	95% Confidence Interval for Mean			Lower Bound	164.81	
			Upper Bound	191.44		
	5% Trimmed Mean			161.08		
	Median			140.00		
	Variance			24532.012		
	Std. Deviation			156.627		
	Minimum			1		
	Maximum			1352		
	Range			1351		
	Interquartile Range		172			
Skewness		2.692	.106			
Kurtosis		12.675	.211			

The **Custom Tables** procedure can also be used to produce summary of a quantitative variable for different groups.



The output is:

		Number of CD4 Counts					
		Minimum	Maximum	Mean	Median	Mode	Standard Deviation
Sex	F	2	1914	210	173	139	178
	M	1	1352	178	140	308	157
Education Level	No Education	6	1718	200	157	42	173
	Primary	1	1914	210	167	138	182
	Secondary	4	1352	194	150	308	170
Clinical Stage	Tertiary	2	832	165	137	93	126
	Stage I	3	1718	252	218	308	179
	Stage II	1	1914	204	164	138	171
	Stage III	1	1204	168	128	36	158
	Stage IV	2	832	141	80	36	154

Chapter 5

Inferential Statistics

The primary objective of a statistical analysis is drawing statistically valid conclusions about the characteristics of the population based on the results obtained from sample. There are two important facts that are key to statistical inference. These are (population) *parameter* and (sample) *statistic*. A *parameter* is a fixed (but usually unknown) summary measure of the characteristic of a population. For example, population mean (μ), population variance (σ^2), population proportion (π) are parameters.

On the other hand, a *statistic* is a known summary measure of the characteristic of a sample. For example, sample mean (\bar{x}), sample variance (s^2), sample proportion (p) are statistics. These measures, unlike parameters, are random because their values vary from sample to sample.

Statistical inference can be defined as the process of making conclusions for population parameters using sample statistics. It generally takes two forms, namely, *estimation* of a parameter and *testing of a hypothesis*.

5.1 Estimation of a Parameter

For the purpose of general discussion, let θ be the population parameter and $\hat{\theta}$ be the corresponding statistic. The statistic $\hat{\theta}$ intended for estimating a parameter θ is called an *estimator* of θ . A specific numerical value of an estimator calculated from the sample is called the *estimate*. The process of obtaining an estimate of the unknown value of a parameter by a statistic is called *estimation*. There are two types of estimations. One is *point* estimation and the other is *interval* estimation.

5.1.1 Point Estimation

Point estimation is the process of obtaining a *single* sample value (point estimate) that is used to estimate the desired population parameter. The estimator is known as point estimator. For example, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is a point estimator of μ , $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is a point estimator of σ^2 .

5.1.2 Interval Estimation

Point estimator has some drawbacks. First, a point estimator from the sample *may not exactly locate the population parameter*, that is, the value of point estimator *is not likely to be exactly equal to the value of the parameter*, resulting in some margin of uncertainty. If the sample value is different from the population value, the point estimator *does not indicate the extent of the possible error*. Second, a point estimate *does not specify as to how confident we can be that the estimate is close to the parameter* it is estimating. That is, we *cannot attach any degree of confidence* to such an estimate as to what extent it is closer to the value of the parameter. Because of these limitations of point estimation, interval estimation is considered desirable. *Interval estimation* involves the determination of an *interval (a range of values)* within which the population parameter must lie with a *specified degree of confidence*. It is the *construction of an interval on both sides of the point estimate* within which we can reasonably be confident that the true parameter will lie.

5.2 Hypothesis Testing for Parameters

A statistical hypothesis is an assumption (a conjecture) about a population parameter. Such an assumption usually results from speculation concerning observed behavior, natural phenomena, or established theory. Hence, hypothesis testing is a statistical procedure which leads to take a decision about a statistical hypothesis for being supported or not by the sample data. It starts by making a set of two mutually-exclusive and exhaustive hypotheses about the parameter(s) in question.

The first hypothesis is called a *null hypothesis* (denoted by H_0) which states there is no difference between a parameter and a hypothesized value. For any parameter θ and an assumed value θ_0 , the null hypothesis is written as $H_0 : \theta = \theta_0$.

The second hypothesis, is called an *alternative hypothesis* (denoted by H_1), contradicts the null hypothesis, that is, it states there is a difference between a parameter and a hypothesized value. This hypothesis may have three different forms:

- Two-sided test: $H_1 : \theta \neq \theta_0$.
- One-sided test: $H_1 : \theta > \theta_0$ (right tailed test)
- One-sided test: $H_1 : \theta < \theta_0$ (left tailed test)

5.2.1 Types of Errors

There are two types of errors in hypothesis testing.

- **Type I Error:** It is an error occurred if one rejects the null hypothesis which is actually true. The probability of making such an error is called *significance level* (denoted by α).
- **Type II Error:** It is an error occurred if one failed to reject the null hypothesis which is actually false. The probability of making type II error is denoted by β . The probability of correctly rejecting the null hypothesis which is actually false, called *power* of a test, is, therefore, $1 - \beta$.



In statistical hypothesis testing, the maximum acceptable probability of rejecting a true null hypothesis, the significance level (α), is specified first.

5.2.2 Statistical Significance

If the p -value is less than the specified significance level (α), the null hypothesis (H_0) can be rejected. Then, the test is said to be significant. If we failed to reject the null hypothesis, then the test would be non-significant (insignificant).

But note that a significant test does not indicate practical (clinical) significance or importance. With large samples, it is possible to find statistical significance even when the difference is very small (i.e., has a small effect size). A statistically significant result with a small effect size means that it is sure that there is at least a little difference, but it may not be of any practical significance.

5.2.3 Interpretations

In testing a statistical hypothesis, first, decide whether to reject the null hypothesis. If the test is found significant, we need to answer two or more questions.

1. What is the direction of the effect? Difference inferential questions (t test or analysis of variance) compare two or more groups so it is necessary to state which group performed better. For associational inferential questions (eg, correlation, regression), the sign is very important, so we must indicate whether the relationship is positive or negative.
2. What is the size of the effect? We should include the effect size, confidence intervals or both in the description of our results.
3. The researcher or consumer of the research should make a judgment whether the result has practical or clinical significance or importance. To do so, they need to take into account the effect size, the cost of implementing the change, and the probability and severity of any side effect or unintended consequence.

Chapter 6

Hypothesis Testing

6.1 Testing about a Single Population Mean

A one-sample t -test helps to determine whether the population mean (μ) is equal to a hypothesized value (μ_0). The underlying assumption of the t -test is that the observations are random samples drawn from normally distributed populations.

Steps:

1. The null hypothesis to be tested is $H_0 : \mu = \mu_0$ and the alternative hypothesis can be $H_1 : \mu \neq \mu_0$, $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$.
2. Choose a level of significance (α): common choices are 0.01, 0.05 and 0.10.
3. The test statistic is: $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$ where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the sample mean, $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ is the sample variance (hence s is the sample standard deviation), n is the sample size and μ_0 is the assumed value. The test statistic has a t distribution with $n - 1$ degrees of freedom.
4. Decision: If the p -value is less than the specified α , H_0 should be rejected, otherwise do not.
5. Conclusion.

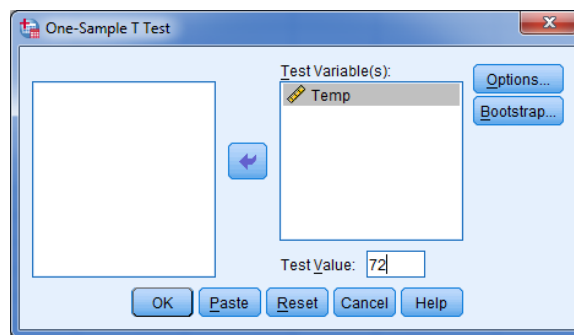
For testing a single population mean using SPSS, from the *Menu* bar, click on **Analyze** → **Compare Means** → **One-Sample T Test**.

Example 6.1. The thermostat in your classroom is set at 72°F, but you think the thermostat is not working well. On seven randomly selected days, you measure the temperature at your seat. Your measurements (in degrees Fahrenheit) are 71, 73, 69, 68, 69, 70, and 71. Test whether the mean temperature at your seat is different from 72°F. Conduct the analysis using SPSS.

Here, the hypothesis to be tested is $H_0 : \mu = 72$ vs $H_1 : \mu \neq 72$. Now, enter the data into SPSS under the variable name **Temp** as follows.

	Temp	var	var	var	var
1	71				
2	73				
3	69				
4	68				
5	69				
6	70				
7	71				
8					
9					
10					

In the **One-Sample T Test** dialogue box, enter **Temp** in the **Test Variable(s):** box. Then, in the **Test Value:** box, enter 72 which is the value assumed under H_0 .



If you want to change the confidence level (default is 95%), click on the **Options** button, and then specify in the **Confidence Interval Percentage:** box. Then click on the **Continue** tab and **OK**.

The output provides two table results. The first table (**One-Sample Statistics**) provides the number of observations, mean, standard deviation and standard error of the sample mean. The sample mean and standard deviation of the 7 observations is 70.14°F and 1.68°F, respectively.

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
Temp	7	70.1429	1.67616	.63353

One-Sample Test						
Test Value = 72						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Temp	-2.931	6	.026	-1.85714	-3.4073	-.3070

The second table (**One-Sample Test**) provides the test statistic value, the degrees of freedom, the p -value, the difference of the sample mean from the assumed value and the confidence interval for the population mean. Thus, since the p -value is **Sig. (2-tailed) = 0.026** which is smaller than the (default) level of significance ($\alpha = 0.05$), the null hypothesis that the mean temperature at your seat is 72°F should be rejected. Therefore, the mean temperature in the classroom is significantly different from 72°F.

6.2 Comparing Paired Samples

For two paired variables, the difference of the two variables, $d_i = Y_{1i} - Y_{2i}$, is treated as if it were a single sample. This test is appropriate for pre-post treatment responses. The null hypothesis is that the true mean difference of the two variables is D_0 , $H_0 : \mu_d = D_0$. The difference is typically assumed to be zero unless explicitly specified.

Steps:

1. The null hypothesis to be tested is $H_0 : \mu_d = 0$ and the alternative hypothesis may be $H_1 : \mu_d \neq 0$, $H_1 : \mu_d < 0$ or $H_1 : \mu_d > 0$.
2. Choose a level of significance (α)
3. The test statistic is: $t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \sim t(n-1)$ where $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ is the sample mean of the differences, $s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$ is the sample variance of the differences and n is the sample size. This test statistic has a t distribution with $n-1$ degrees of freedom.
4. Decision: If the p -value is less than the specified α , H_0 should be rejected otherwise do not.
5. Conclusion.

In SPSS, the procedure for paired test is: **Analyze** → **Compare Means** → **Paired-Samples T Test**.

Example 6.2. A researcher is interested in investigating whether alcohol has a positive or negative effect on the heart beat of individuals. S/he has measured the heart beat (per minute) of six persons before and after drinking Alcohol. The data is:

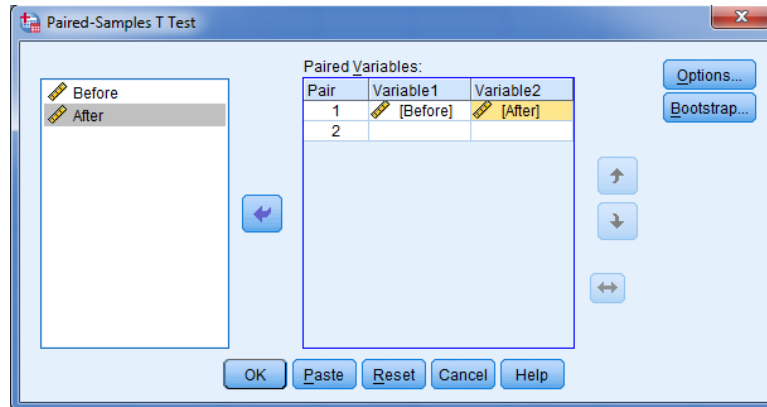
Before Drinking Alcohol	86	90	75	72	78	68
After Drinking Alcohol	97	96	80	76	77	73

Test the hypothesis using SPSS.

Enter the paired data, naming the first variable of the pair **Before** and the second **After**, as follows.

	Before	After	var	var	var
1	86	97			
2	90	96			
3	75	80			
4	72	76			
5	78	77			
6	68	73			
7					
8					
9					
10					

Then, in the **Paired-Samples T Test** dialogue box, enter **Before** under **Variable1** and **After** under **Variable2** of the **Paired Variables:** box.



The **Paired-Samples T Test** procedure provides three table of results. The first table, **Paired Samples Statistics**, contains descriptive statistics for each of the two variables. The second table, **Paired Samples Correlations**, displays the correlation between the two variables as 0.943 and its corresponding p -value as **Sig.** = 0.005. Since the p -value is less than $\alpha = 0.05$, it can be concluded that there is a strong positive relationship between the before and after measurements of the heart beat of individuals.

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Before	78.1667	6	8.40040	3.42945
	After	83.1667	6	10.57198	4.31599

Paired Samples Correlations				
		N	Correlation	Sig.
Pair 1	Before & After	6	.943	.005

Paired Samples Test									
		Paired Differences				t	df	Sig. (2-tailed)	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower				Upper
Pair 1	Before - After	-5.00000	3.84708	1.57056	-9.03726	-.96274	-3.184	5	.024

As can be seen from the third table, (**Paired Samples Test**), since p -value is **Sig. (2-tailed)** = 0.024 which smaller than $\alpha = 0.05$, we can conclude that alcohol has an increasing effect in the heart beat of individuals.

6.3 Comparing Independent Samples

1. The null hypothesis to be tested is $H_0 : \mu_1 = \mu_2$ and the alternative hypothesis may be $H_1 : \mu_1 \neq \mu_2$, $H_1 : \mu_1 < \mu_2$ or $H_1 : \mu_1 > \mu_2$.
2. Choose a level of significance (α).

3. The test statistic is: $t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^n y_{1i}$ is the sample

mean of the first group and $\bar{y}_2 = \frac{1}{n_2} \sum_{i=1}^n y_{2i}$ is the sample mean of the second group, $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is the pooled variance of the both groups (note $s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^n (y_{1i} - \bar{y}_1)^2$ is the sample variance of the first group and $s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^n (y_{2i} - \bar{y}_2)^2$ is the sample variance of the second group), n_1 is sample size of the first group and n_2 is sample size of the second group. The test statistic has a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

4. Decision: If the p -value is less than the specified α , H_0 should be rejected otherwise do not.
5. Conclusion.

The above test statistic is only used when the two distributions are assumed to have the same variance. If the two population variances are different, then they must be estimated separately and the test statistic is a little bit modified as

$$t = \frac{(\bar{y} - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

This modified test, also known as Welch's t -test, has a t distribution with v degrees of freedom where

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Note that the true distribution of the test statistic actually depends (slightly) on the two unknown variances. Therefore, to determine which test statistics to be used for comparing two population means, first the equality of variances should be checked.

Comparing Two Population Variances

1. The null and alternative hypotheses to be tested are:

$$H_0 : \sigma_1 = \sigma_2$$

$$H_1 : \sigma_1 \neq \sigma_2$$

2. Choose a level of significance (α).

3. The test statistic is: $F = \frac{s_1^2}{s_2^2}$ where $s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^n (y_{1i} - \bar{y}_1)^2$ is the sample variance

of the first group and $s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^n (y_{2i} - \bar{y}_2)^2$ is the sample variance of the second group, n_1 is sample size of the first group and n_2 is sample size of the second group. This statistic has an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

4. Decision: If the p -value is less than the specified α , then H_0 is rejected indicating that the common variance assumption does not hold.
5. Conclusion.

In the case of two independent populations, the procedure in SPSS is: **Analyze** → **Compare Means** → **Independent-Samples T Test**. Here, the response variable should be stacked in one column and a grouping variable should be in another column.

Example 6.3. Company officials were concerned about the length of time a particular drug product retained its toxin's potency. A random sample of 8 bottles of the product was drawn from the production line and measured for potency. A second sample of 10 bottles was obtained and stored in a regulated environment for a period of one year. The readings obtained from each sample are given below.

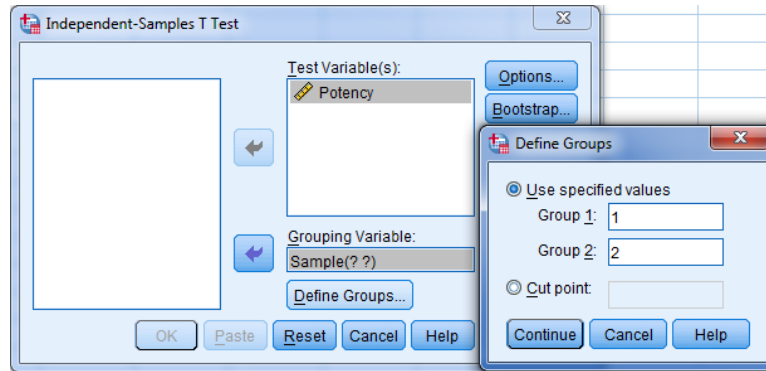
Sample 1	10.2	10.5	10.3	10.8	9.8	10.6	10.7	10.2		
Sample 2	9.8	9.6	10.1	10.2	10.1	9.7	9.5	9.6	9.8	9.9

Using SPSS, test the null hypothesis that the drug product retains its potency. Also, construct the 95% confidence interval for the difference of the population means.

To enter the above data in SPSS, enter the grouping variable by naming **Sample** in one column and the stacked response in another column naming as **Potency**.

	Sample	Potency	var	var	var	var
1	1	10.2				
2	1	10.5				
3	1	10.3				
4	1	10.8				
5	1	9.8				
6	1	10.6				
7	1	10.7				
8	1	10.2				
9	2	9.8				
10	2	9.6				
11	2	10.1				
12	2	10.2				
13	2	10.1				
14	2	9.7				
15	2	9.5				
16	2	9.6				
17	2	9.8				
18	2	9.9				
19						
20						

Then, in the **Independent-Samples T Test** dialogue box, enter **Potency** in the **Test Variable(s):** box and **Sample** in the **Grouping Variable:** box. Next, click on the **Define Groups** button, and then type 1 in the **Group 1:** field and 2 in the **Group 2:** field. Click on the **Continue** tab and then **OK**.



This procedure results two tables; one a table of descriptive statistics for both variables and the other the table of test results.

	Sample	N	Mean	Std. Deviation	Std. Error Mean
Potency	1	8	10.388	.3271	.1156
	2	10	9.830	.2406	.0761

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Potency	Equal variances assumed	.941	.347	4.172	16	.001	.5575	.1336	.2742	.8408
	Equal variances not assumed			4.028	12.545	.002	.5575	.1384	.2574	.8576

As shown in the above result in the **Independent Samples Test** table, SPSS by default conducts **Levene’s Test for Equality of Variances** in the two groups. The test shows the equal variance assumption is not rejected as its p -value (**Sig** = 0.347) is larger than $\alpha = 0.05$.

Then, the **t-test for Equality of Means** is calculated under both assumptions (equal variance assumption and different variances). The test statistic has a p -value of, **Sig. (2-tailed)**, 0.001 implying the rejection of the null hypothesis of no difference in the mean potency of the two samples. Therefore, there is a significant difference in the mean potency of the two samples.

Exercise 6.1. A quick but impressive method of estimating the concentration of a chemical in a rat has been developed. The sample from this method has 8 observations and the sample from the standard method has 4 observations. Assuming different population variances, test whether the quick method gives under-estimate result. The data in the two samples are:

Standard Method	25	24	25	26				
Quick Method	23	18	22	28	17	25	19	16

6.4 Comparing Several Population Means: ANOVA

Despite its name, analysis of variance (ANOVA) is used to compare the means of more than two groups based on the variance ratio test. The principle underlying the ANOVA is that

the total variability in a dataset is partitioned into its component parts. The sources of variation comprise one or more factors, each resulting in variability which can be accounted for (explained by the levels or categories of the factor), and also unexplained (residual) variation which results from uncontrolled biological variation and technical error.

Note that the null hypothesis is that the all group means are equal. That is, if there are g groups, then $H_0 : \mu_1 = \mu_2 = \dots = \mu_g$. The alternative hypothesis is at least one of the means is significantly different from the other.

Assumptions of the one-way ANOVA

1. The samples are independently and randomly drawn from source population(s).
2. The source populations are reasonably normal distributions.
3. The samples have approximately equal variances.

If the samples are equal size, no main worry about these assumptions because one-way ANOVA is quite robust (relatively unperturbed by violations of its assumptions). But if the samples are different size and the assumption of equal variance does not hold, an appropriate non-parametric alternative for one-way ANOVA, which is called the Kruskal-Wallis test.

The procedure in SPSS for one-way ANOVA is: **Analyze** → **Compare Means** → **One-Way ANOVA**. Similar to the **Independent-Samples T Test**, the response should be stacked in one column and the grouping variable should be in another column.

Example 6.4. Suppose a university wishes to compare the effectiveness of four teaching methods (Slide, Self-Study, Lecture and Discussion) for a particular course. Twenty four students are randomly assigned to the teaching methods. At the end of teaching the students with their assigned method, a test (out of 20%) was given and the performance of the students were recorded as follows:

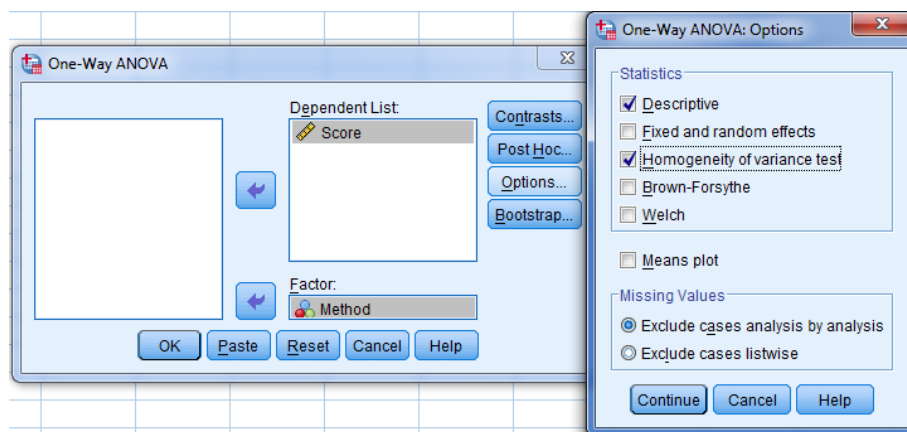
Slide	Self-Study	Lecture	Discussion
9	10	12	9
12	6	14	8
14	6	11	11
11	9	13	7
13	10	11	8
	5	16	6
			7

Examine whether there is any difference among the teaching methods.

After entering the teaching **Method** in one column and the **Score** in other column of the *Data Editor*.

	Method	Score	var	var	var	var
1	Slide	9				
2	Slide	12				
3	Slide	14				
4	Slide	11				
5	Slide	13				
6	Self-Study	10				
7	Self-Study	6				
8	Self-Study	6				
9	Self-Study	9				
10	Self-Study	10				
11	Self-Study	5				
12	Lecture	12				
13	Lecture	14				
14	Lecture	11				
15	Lecture	13				
16	Lecture	11				
17	Lecture	16				
18	Discussion	9				
19	Discussion	8				
20	Discussion	11				
21	Discussion	7				
22	Discussion	8				
23	Discussion	6				
24	Discussion	7				

Then, in the **One-Way ANOVA** dialogue box, enter **Score** in the **Dependent List:** box and enter **Method** in the **Factor:** box.



This procedure, by default, does not provide descriptive statistics per group. To obtain descriptives for each group, click on the **Options** button and check on the **Descriptive** box. Also, one of the assumptions of ANOVA is that the variances are the same across groups. To examine it, the **Homogeneity of variance test** must be checked as above.

The larger the p -value (that is, a p -value of 0.461) for the **Levene Statistic** for the **Test of Homogeneity of Variances** indicates that the common variance assumption holds. Hence, the ANOVA test is appropriate.

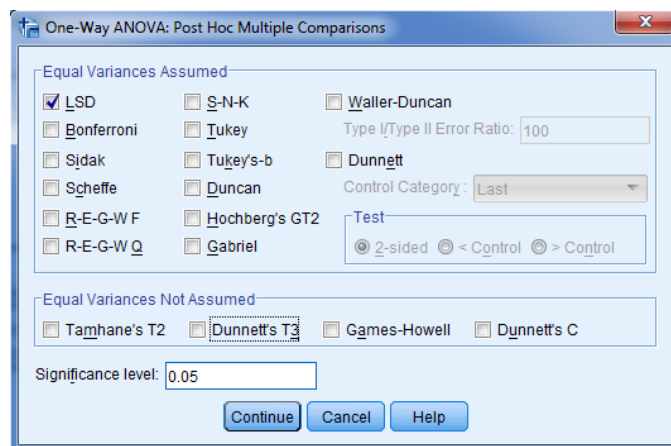
Descriptives									
Score									
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	
					Lower Bound	Upper Bound			
Slide	5	11.80	1.924	.860	9.41	14.19	9	14	
Self-Study	6	7.67	2.251	.919	5.30	10.03	5	10	
Lecture	6	12.83	1.941	.792	10.80	14.87	11	16	
Discussion	7	8.00	1.633	.617	6.49	9.51	6	11	
Total	24	9.92	2.948	.602	8.67	11.16	5	16	

Test of Homogeneity of Variances			
Score			
Levene Statistic	df1	df2	Sig.
.894	3	20	.461

ANOVA					
Score					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	124.867	3	41.622	11.104	.000
Within Groups	74.967	20	3.748		
Total	199.833	23			

In the ANOVA test, the significant F statistic (a p -value of < 0.0001) tells us that the means are not all equal, that means, at least one of the teaching methods differs from the other.

However, the one-way ANOVA does not tell us where the differences are. To examine the differences in the teaching methods, mean separation method should be used. To do so, click on the **Post Hoc** button of the **One-Way ANOVA** dialogue box and then check on at least one comparison methods like **LSD**, **Bonferroni** or **Scheffe**.



The output from the **LSD** mean separation method is as follows.

Multiple Comparisons						
Dependent Variable: Score						
LSD						
(I) Method	(J) Method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Slide	Self-Study	4.133*	1.172	.002	1.69	6.58
	Lecture	-1.033	1.172	.389	-3.48	1.41
	Discussion	3.800*	1.134	.003	1.44	6.16
Self-Study	Slide	-4.133*	1.172	.002	-6.58	-1.69
	Lecture	-5.167*	1.118	.000	-7.50	-2.84
	Discussion	-.333	1.077	.760	-2.58	1.91
Lecture	Slide	1.033	1.172	.389	-1.41	3.48
	Self-Study	5.167*	1.118	.000	2.84	7.50
	Discussion	4.833*	1.077	.000	2.59	7.08
Discussion	Slide	-3.800*	1.134	.003	-6.16	-1.44
	Self-Study	.333	1.077	.760	-1.91	2.58
	Lecture	-4.833*	1.077	.000	-7.08	-2.59

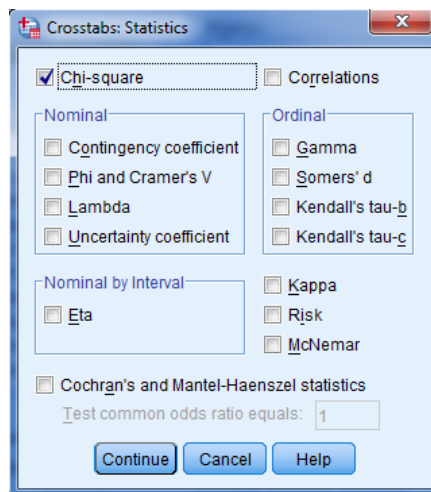
*. The mean difference is significant at the 0.05 level.

The significant pairs are Slide > Self-Study, Slide > Discussion, Self-Study < Lecture and Lecture > Discussion.

6.5 Chi-Square Test of Association

The χ^2 test is used for testing the independence of two categorical variables. The null hypothesis, H_0 :, states there is no statistical association between the two categorical variables.

The Pearson χ^2 test in SPSS is found as an option in **Statistics** tab of the **Crosstabs** dialogue box. As usual, if the p -value is less than the specified level of significance α , H_0 will be rejected.



Example 6.5. Is there a statistical association between the Sex of a patients and Education Level of the JUSH_HAART data.

The result is as follows. The expected frequencies are obtained by selecting **Expected** under the **Counts** option of the **Cells** tab in the **Crosstabs** dialogue box.

Sex * Education Level Crosstabulation							
			Education Level				Total
			No Education	Primary	Secondary	Tertiary	
Sex	F	Count	211	325	316	73	925
		Expected Count	188.3	326.5	311.9	98.3	925.0
	M	Count	86	190	176	82	534
		Expected Count	108.7	188.5	180.1	56.7	534.0
Total		Count	297	515	492	155	1459
		Expected Count	297.0	515.0	492.0	155.0	1459.0

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	25.397 ^a	3	.000
Likelihood Ratio	24.929	3	.000
N of Valid Cases	1459		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 56.73.

The significance of the Pearson chi-square test (the smaller the p -value) reveals that there is an association between the Sex of the patient and Education Level.

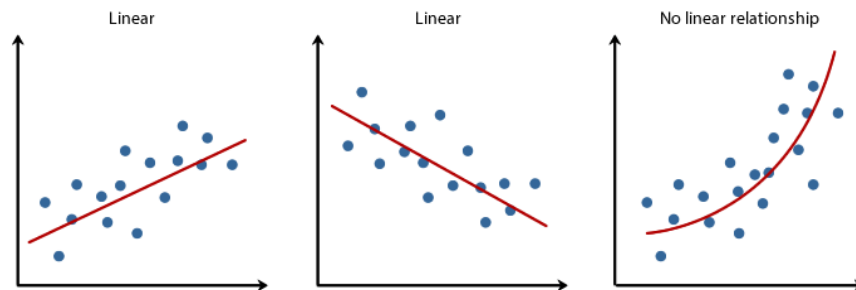
Chapter 7

Regression Analysis

7.1 Linear Correlation

Correlation is a statistical tool desired towards measuring the degree of linear relationship (association) between two quantitative variables. If the change in one variable affects the change in the other variable, then the variables are said to be *correlated*.

The simplest way to examine the correlation between two quantitative variables is to plot the pair of values $(x_i, y_i), i = 1, 2, \dots, N$ on the xy plane, known as *scatter plot*. If the relationship between the two variables can be described by a straight line, then the relationship is known as *linear* other wise it is called *non-linear*.

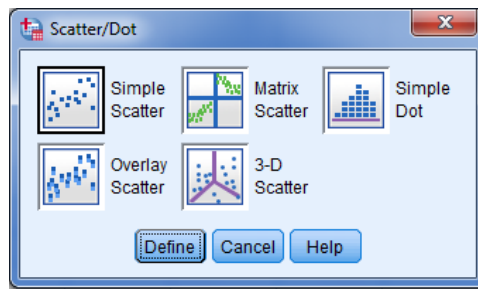


Such a plot gives some idea about the presence and absence of correlation, and the nature (direct or indirect) of correlation. But, it will not indicate about the strength or degree of relationship between two variables.

7.1.1 Scatter Plot

Correlation that involves only two variables is called simple correlation. The simplest way to present bivariable data is to plot the values $(x_i, y_i), i = 1, 2, \dots, n$ on the xy plane. This is known as *scatter plot*. This gives an idea about the correlation of the two variables. But, it will give only a vague idea about the presence and absence of correlation and the nature (direct or inverse) of correlation. It will not indicate about the strength or degree of relationship between two variables.

From the *Menu* bar, click on **Graphs** → **Legacy Dialogs** → **Scatter/Dot**. Then, click on the **Simple Scatter** option of the **Scatter/Dot** dialogue box.



Then, click on the **Define** button.

Example 7.1. A researcher wants to find out if there is a relationship between the heights of sons with the heights and weights of fathers. In other words, do taller fathers have taller sons? The researcher took a random sample of 8 fathers and their 8 sons. Their height in centimeters and the weight of fathers in kilograms are given below.

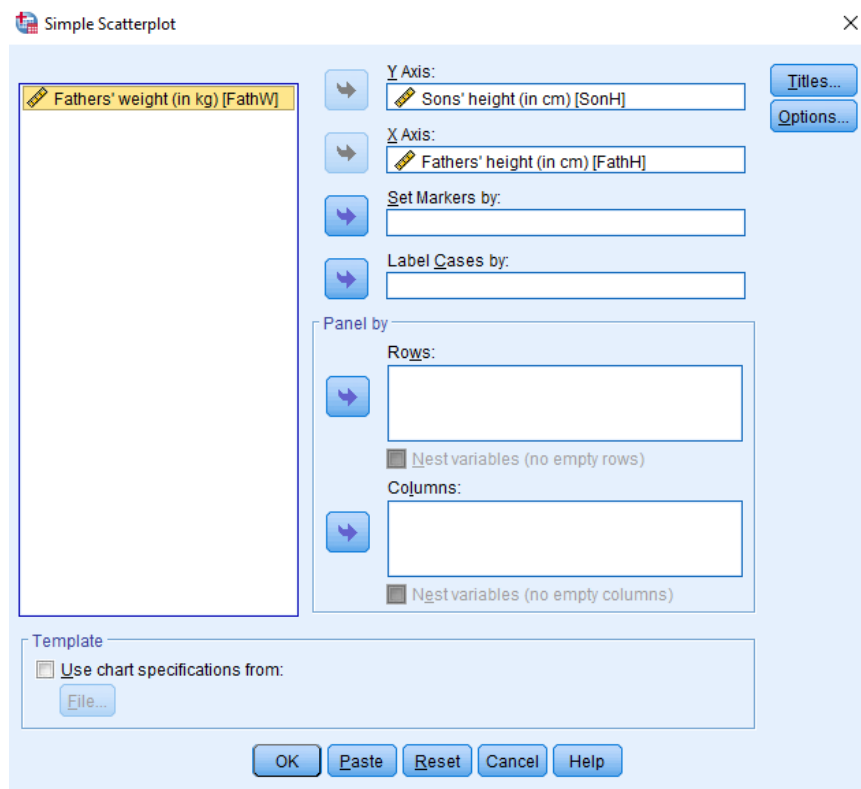
Father Height (x_1)	168	173	165	170	175	178	180	164
Father Weight (x_2)	65	67	66	66	68	67	66	65
Son Height (y)	165	170	168	170	173	175	174	167

Obtain the scatter plot of son’s height and father’s height, and son’s height and father’s weight.

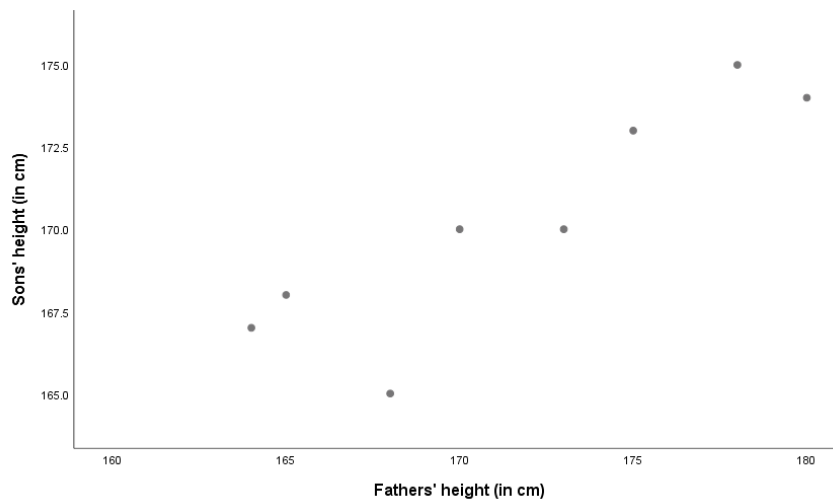
The data should be entered as follows.

	FathH	FathW	SonH	var	var	var	var	var
1	168	65	165					
2	173	67	170					
3	165	66	168					
4	170	66	170					
5	175	68	173					
6	178	67	175					
7	180	66	174					
8	164	65	167					
9								
10								

In the **Simple Scatterplot** dialogue box, enter **SonH** in the **Y Axis:** box and enter **FathH** in the **X Axis:** box.

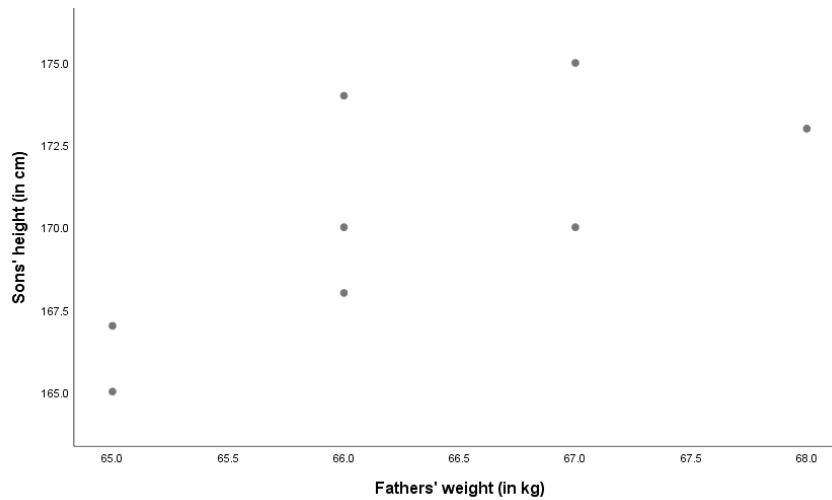


The scatter plot of son’s height and father’s height is:



As can be seen from the plot, it is clear that there is a linear relationship between son’s height and father’s height.

Similarly, in the **Simple Scatterplot** dialogue box, by entering SonH in the **Y Axis:** box and FathW in the **X Axis:** box, the scatter plot of son’s height and father’s weight is shown below.



From this scatter plot, it seems there is a linear relationship between son's height and father's weight.

7.1.2 Covariance

Covariance is a measure of the joint variation between two variables, i.e., it measures the way in which the values of the two variables vary together.

Recall the population variance of a certain variable x is defined as $\sigma_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \sigma_{xx}$

and is estimated by the sample variance given by $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = s_{xx}$. Similarly the population covariance between two variables x and y is defined as

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N} \left(\sum_{i=1}^N x_i y_i - \frac{1}{N} \sum_{i=1}^N x_i \sum_{i=1}^N y_i \right)$$

and is estimated by the sample covariance:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right).$$

The value of a covariance can be negative, zero or positive. If the covariance is zero, there is no linear relationship between the two variables. If the covariance is positive, there is a direct linear relationship between the variables. If it is negative, there is an indirect linear relationship.

7.1.3 Correlation Coefficient

The *coefficient of correlation*, which was developed by Karl Pearson, is a measure of the *degree* or *strength* of the linear association between two variables. It is defined as a ratio of the *covariance* between the two variables and the *product of the standard deviations* of each

variable. The population correlation coefficient is denoted by the Greek letter ρ , rho:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

Depending on the sign of a covariance, a correlation coefficient can be positive or negative. But, the value lies between the limits -1 and +1; that is $-1 \leq \rho \leq 1$.

- If ρ is zero, there is no linear relationship between the two variables.
- If ρ is approximately -0.5 or +0.5, there is a medium inverse (indirect) or positive (direct) linear relationship between the variables, respectively.
- If ρ is approximately -1 or +1, there is a strong inverse (indirect) or positive (direct) linear relationship between the variables, respectively.

The sample correlation coefficient is denoted by r :

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

This can also be written as:

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}}$$

Notes:

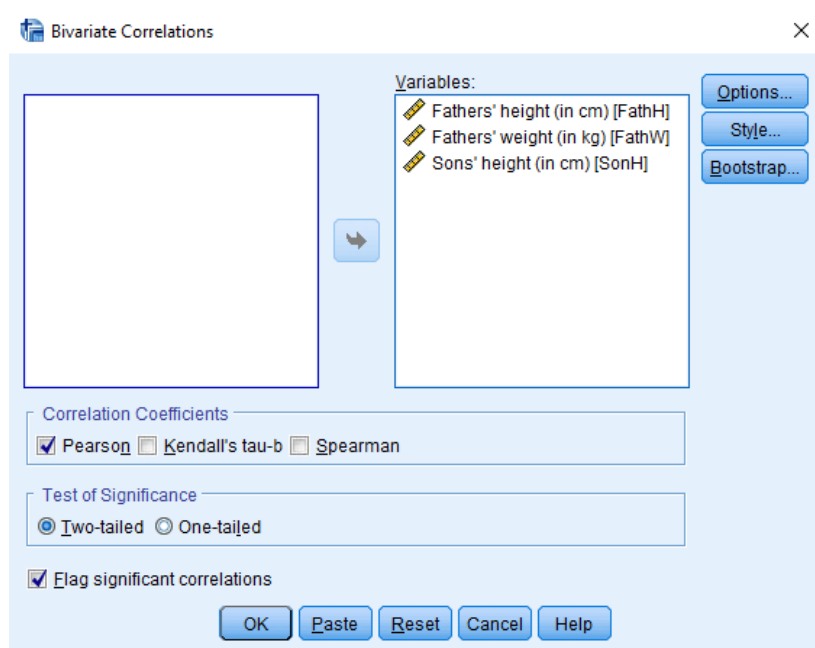
- Although, the sign of the correlation and covariance are the same, the correlation is ordinarily easier to interpret as:
 - its magnitude is bounded, that is, $-1 \leq r \leq 1$.
 - it is unit less.
 - it takes the variability into account.
- But, it has also some disadvantages:
 - Correlation does not measure nonlinear relationships. If x and y are statistically *independent*, the correlation coefficient between them will be zero; but the converse is not always true. In other words, *zero correlation does not necessarily imply independence*. Thus, for example, even if $y = x^2$; $-4 < x < 4$ is an exact relationship, yet r is zero. (Why?)

- Although, correlation is a measure of the linear association between variables, it does not necessarily imply any *cause and effect* relationship.

Using SPSS, to determine the correlation between quantitative variables, from the *Menu* bar, click **Analyze** → **Correlate** → **Bivariate**. Then, in the **Bivariate Correlations** dialogue box, enter at least two quantitative variables in the **Variable(s):** box.

Example 7.2. Using the data given on example 7.1, perform correlation analysis of among son's height, father's height and father's weight.

Enter the three quantitative variables in the **Variable(s):** box of the **Bivariate Correlations** dialogue box.



Then, click on **OK**. The output of the correlation matrix is:

		Correlations		
		Fathers' height (in cm)	Fathers' weight (in kg)	Sons' height (in cm)
Fathers' height (in cm)	Pearson Correlation	1	.605	.892**
	Sig. (2-tailed)		.112	.003
	N	8	8	8
Fathers' weight (in kg)	Pearson Correlation	.605	1	.722*
	Sig. (2-tailed)	.112		.043
	N	8	8	8
Sons' height (in cm)	Pearson Correlation	.892**	.722*	1
	Sig. (2-tailed)	.003	.043	
	N	8	8	8

** . Correlation is significant at the 0.01 level (2-tailed).
* . Correlation is significant at the 0.05 level (2-tailed).

The result shows two pairs of correlations are significant. Hence, it can be concluded there is a strong positive correlation between son's height and father's height, and son's height and father's weight. But, there is no significant linear relationship found between father's height and father's weight.

7.2 Linear Regression

Regression may be defined as the estimation of the unknown value of one variable from the known value(s) of one or more variables. The variable whose values are to be estimated is known as *dependent* variable while the variable(s) which is (are) used in determining the value of the dependent variable is (are) called *independent* variable(s). The dependent variable can take negative, zero or positive values, and it is assumed to have a normal distribution with mean μ and constant variance σ^2 .

If the relationship between the two variables can be described by a straight line, then the regression is known as *linear regression* other wise it is called *non-linear*.

A linear regression involving only two variables (one dependent and one independent) is called *simple linear regression* and a linear regression analysis that involves more than two variables (one dependent and two or more independents) is called *multiple linear regression*. A *multiple linear regression* model is a linear function of a set of a set of independent variables (quantitative, qualitative or both).

In a linear regression, the *mean* of the outcome is modeled. The model has the form: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \cdots + \hat{\beta}_kx_k + \cdots + \hat{\beta}_px_p$ where

- \hat{y} is the estimated mean of the dependent variable.
- $\hat{\beta}_0$ is the estimated intercept of the linear model (it is the mean of the dependent variable when the value of the independent variable is zero).
- $\hat{\beta}_k; k = 1, 2, \dots, p$ is the k^{th} slope parameter estimate.

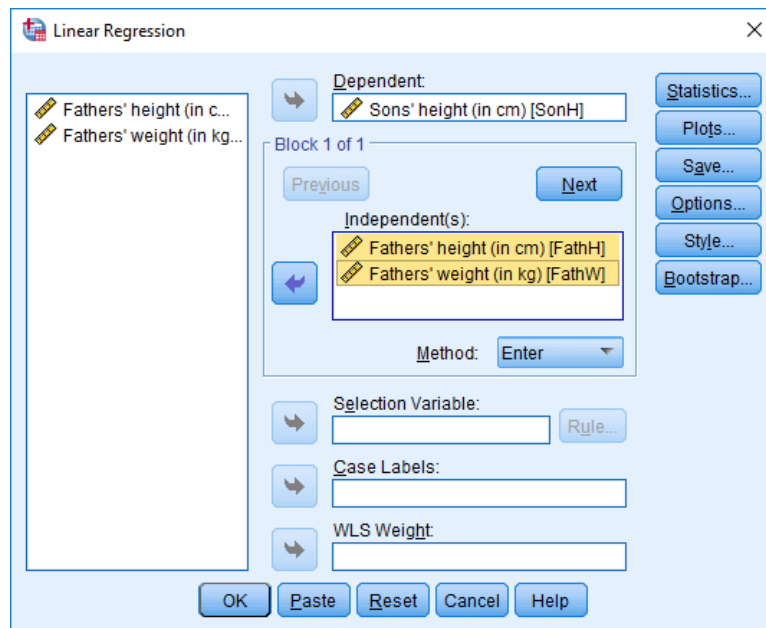
The parameter estimates are interpreted as the slope of a line describing the relationship of the independent variable to the outcome.

- Given $\hat{\beta}_k > 0$. For each unit increase in the k^{th} independent variable, the mean of the outcome increases by $\hat{\beta}_k$.
- Similarly, given $\hat{\beta}_k < 0$. The mean of the outcome decreases by $\hat{\beta}_k$ for each unit increase in the k^{th} independent variable.

The procedure to do regression analysis in SPSS is: **Analyze** → **Regression** → **Linear**. Then, in the **Linear Regression** dialogue box, enter the dependent (response) variable in the **Dependent:** box and enter all the independent (explanatory) variables in the **Independent(s):** box.

Example 7.3. Recall example 7.1. Perform regression analysis of son's height on father's height and father's weight.

Enter the SonH in the **Dependent:** box, and enter both FathH and FathW in the **Independent(s):** box of the **Linear Regression** dialogue box. Then click **OK**.



The **Linear Regression** procedure of SPSS, delivers the main results in three tables as shown below. The **Model Summary** provides the coefficient of determination and the adjusted coefficient of determination as 0.848 and 0.787, respectively. This means, 78.7% of the variation in the height of sons is explained by both the father's height and father's weight.

The next table is the **ANOVA** which is used to determine the overall significance of the model. Since the p -value is 0.009 which is smaller than $\alpha = 0.05$, it is clear that the model is significant (that means, at least one of the explanatory variable is significant).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.921 ^a	.848	.787	1.630

a. Predictors: (Constant), Fathers' weight (in kg), Fathers' height (in cm)

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	74.218	2	37.109	13.970	.009 ^b
	Residual	13.282	5	2.656		
	Total	87.500	7			

a. Dependent Variable: Sons' height (in cm)
b. Predictors: (Constant), Fathers' weight (in kg), Fathers' height (in cm)

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	30.930	40.110		.771	.475
	Fathers' height (in cm)	.432	.132	.718	3.280	.022
	Fathers' weight (in kg)	.985	.747	.288	1.318	.245

a. Dependent Variable: Sons' height (in cm)

Lastly, the **Coefficients** table contains parameter estimates of of the model together with their standard errors. The estimated model is $\widehat{\text{SonH}}_i = 30.930 + 0.432\text{FathH}_i + 0.985\text{FathW}_i$; $i = 1, 2, \dots, 8$. Looking at the p -value of each parameter estimate, only **FathH** is significant. Therefore, son's height is positively associated with father's height (taller fathers have taller sons). Specifically, a one centimeter increment in the height of fathers leads to a 0.432 centimeters in height of sons.

But, even if the effect of **FathW** is large, it is not significant at even 10% significance level in the presence of **FathH**. Hence, the next procedure is to remove the **FathW** variable from the model and refit the model with only **FathH**. The the final model indicates 76.1% of the variation in the height of sons is explained by the height of their father's. In addition, a one centimeter increment in the height of fathers leads to a 0.536 centimeters in height of sons.

Example 7.4. A short survey was conducted on a random sample of 23 patients to know the percentage level of their satisfaction by the medical treatment they were given. The patients were asked about three additional variables: their age, gender and education level. The recorded data are presented in the following table where y = patient satisfaction in percentage, x_1 = age in years, x_2 = gender (0=male, 1=female) and x_3 = education level (1= uneducated, 2=primary, 3=secondary, 4=tertiary).

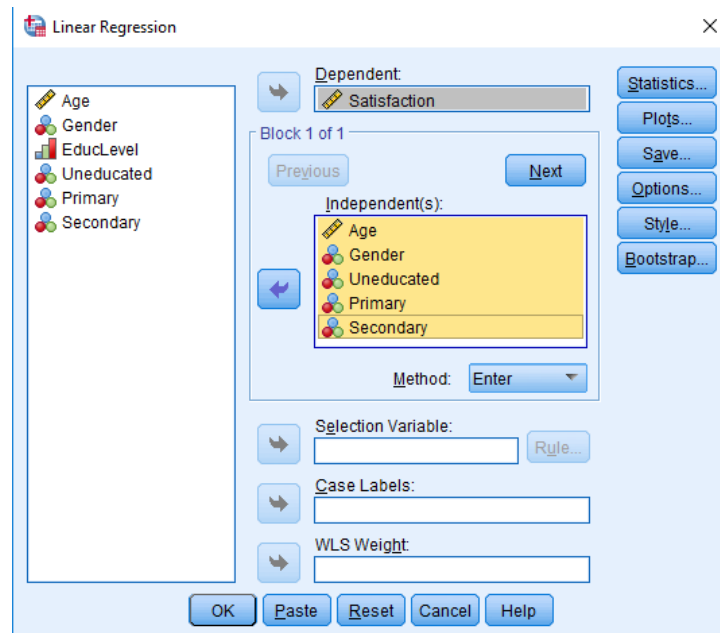
No	y_i	x_{i1}	x_{i2}	x_{i3}	Indicator variables for education level (x_{i3})		
					Uneducated (d_{i31})	Primary (d_{i32})	Secondary (d_{i33})
1	26.1	52	0	1	1	0	0
2	36.5	49	0	1	1	0	0
3	46.1	42	0	1	1	0	0
4	47.2	38	0	1	1	0	0
5	49.0	55	0	1	1	0	0
6	51.0	34	0	1	1	0	0
7	52.5	44	0	2	0	1	0
8	66.4	36	0	2	0	1	0
9	48.0	50	0	2	0	1	0
10	54.6	45	0	2	0	1	0
11	66.7	40	1	2	0	1	0
12	57.9	36	0	3	0	0	1
13	57.0	53	1	3	0	0	1
14	60.5	43	1	3	0	0	1
15	89.4	28	1	3	0	0	1
16	89.1	29	1	3	0	0	1
17	60.3	33	1	4	0	0	0
18	67.5	43	0	4	0	0	0
19	70.7	41	0	4	0	0	0
20	77.7	29	1	4	0	0	0
21	77.0	29	1	4	0	0	0
22	79.2	33	1	4	0	0	0
23	88.6	29	1	4	0	0	0

Estimate a regression model of patient satisfaction on their age, gender and education level, and interpret the results.

The data in the data editor looks as follows:

	Satisfaction	Age	Gender	EducLevel	Uneducated	Primary	Secondary
1	26.1	52	Male	Uneducated	1	0	0
2	36.5	49	Male	Uneducated	1	0	0
3	46.1	42	Male	Uneducated	1	0	0
4	47.2	38	Male	Uneducated	1	0	0
5	49.0	55	Male	Uneducated	1	0	0
6	51.0	34	Male	Uneducated	1	0	0
7	52.5	44	Male	Primary	0	1	0
8	66.4	36	Male	Primary	0	1	0
9	48.0	50	Male	Primary	0	1	0
10	54.6	45	Male	Primary	0	1	0
11	66.7	40	Female	Primary	0	1	0
12	57.9	36	Male	Secondary	0	0	1
13	57.0	53	Female	Secondary	0	0	1
14	60.5	43	Female	Secondary	0	0	1
15	89.4	28	Female	Secondary	0	0	1
16	89.1	29	Female	Secondary	0	0	1
17	60.3	33	Female	Tertiary	0	0	0
18	67.5	43	Male	Tertiary	0	0	0
19	70.7	41	Male	Tertiary	0	0	0
20	77.7	29	Female	Tertiary	0	0	0
21	77.0	29	Female	Tertiary	0	0	0
22	79.2	33	Female	Tertiary	0	0	0
23	88.6	29	Female	Tertiary	0	0	0

As before click on **Analyze** → **Regression** → **Linear**. Then, move **Satisfaction** into the **Dependent:** box, and move **Age**, **Gender**, and the three design variables of education level (**Uneducated**, **Primary** and **Secondary**) into the **Independent(s):** box of the **Linear Regression** dialogue box. Note that the reference category for education level is **Tertiary**. Then click **OK**.



The three tables of the output are presented below. The first table presents the coefficient of multiple determination and adjusted coefficient of multiple determination as 81.8% and 76.5%. This indicates that about 76.5% of the variation in the patients satisfaction is explained by the three variables (age, gender and education level) jointly.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.904 ^a	.818	.765	8.1411

a. Predictors: (Constant), Secondary, Age, Primary, Gender, Uneducated

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5067.459	5	1013.492	15.291	.000 ^b
	Residual	1126.731	17	66.278		
	Total	6194.190	22			

a. Dependent Variable: Satisfaction
b. Predictors: (Constant), Secondary, Age, Primary, Gender, Uneducated

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	103.041	10.796		9.544	.000
	Age	-.956	.258	-.482	-3.706	.002
	Gender	5.245	4.909	.158	1.068	.300
	Uneducated	-17.382	5.912	-.465	-2.940	.009
	Primary	-5.353	5.572	-.135	-.961	.350
	Secondary	-.330	4.919	-.008	-.067	.947

a. Dependent Variable: Satisfaction

The estimated model is $\hat{y}_i = 103.041 - 0.956x_{i1} + 5.245x_{i2} - 17.382d_{i31} - 5.353d_{i32} - 0.331d_{i30}$; $i = 1, 2, \dots, 23$. Looking at the parameter estimates table, age is significant but gender is not, at $\alpha = 0.05$. Also, since one of the three design (dummy) variables of education level is significant, education level is a significant factor of patient satisfaction at $\alpha = 0.05$.

Controlling for all other variables included in the model:

- As the age of a patient increases by 1 year, his/her mean satisfaction level decreases by 0.956%.
- For the given sample, the mean satisfaction level of female patients (relative to male patients) increases by 5.245%. But, this difference is not significant for the population in general at $\alpha = 5\%$.
- The interpretation of each design variable of education level is made relative to the tertiary education which is the reference category.
 - The mean level of satisfaction between uneducated patients and tertiary educated patients is significantly different, but the other two design variables representing primary vs tertiary and secondary vs tertiary are not significant at $\alpha = 5\%$.
 - Therefore, it can be concluded the mean level of satisfaction of uneducated patients decreases by 17.382% relative to educated (primary, secondary or tertiary) patients.

Chapter 8

Logistic Regression Models

8.1 Binary Logistic Regression

A *binary* variable has two categories, for example: alive or dead; development of cancer: yes or no. One of the categories is labeled as "success" and the other as "failure". Mostly, the success is coded by 1 and the failure is coded by 0. A statistical model used for binary outcome variable is *binary logistic regression*. This technique is simply referred as logistic regression. But, the binary is used here, to distinguish it from other (multinomial and ordinal) logistic regression models.

Binary logistic regression models the *logit* of the probability of the success outcome. The *logit* is the natural logarithm of the *odds* of success (log-odds). An *odds* is the *probability* of success divided by the *probability* of failure.

The model is of the form:

$$\text{logit}(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \cdots + \hat{\beta}_p x_p$$

where

- $\hat{\pi}$ is the estimated probability of success.
- $\hat{\beta}_0$ is the estimated intercept of the logit model.
- $\hat{\beta}_k$; $k = 1, 2, \dots, p$ is the k^{th} slope parameter estimate.

The parameter estimates are interpreted in terms of odds ratio, $\text{OR} = \exp(\beta_k)$.

- Given $\hat{\beta}_k > 0 \Rightarrow \text{OR} > 0$. For each unit increase in the k^{th} independent variable, the odds of success increases by a factor of $\exp(\hat{\beta}_k)$.
- Similarly, given $\hat{\beta}_k < 0 \Rightarrow \text{OR} < 0$. For each unit increase in the k^{th} independent variable, the odds of success decreases by a factor of $\exp(\hat{\beta}_k)$.

From the *Menu* bar, click on **Analyze** → **Regression** → **Binary Logistic...** Then, in the **Logistic Regression** dialogue box, enter the dependent (response) variable in the **Dependent:** box and enter all the independent (explanatory) variables in the **Covariates:** box.

- Unlike the **Linear Regression** procedure, there is no need to create dummy variables for qualitative (categorical) independent variables in the **Logistic Regression** procedure.
- By clicking on the **Categorical** button, all the categorical independent variables can be moved into the **Categorical Covariates:** box. This will internally create dummy variables for all those categorical variables.

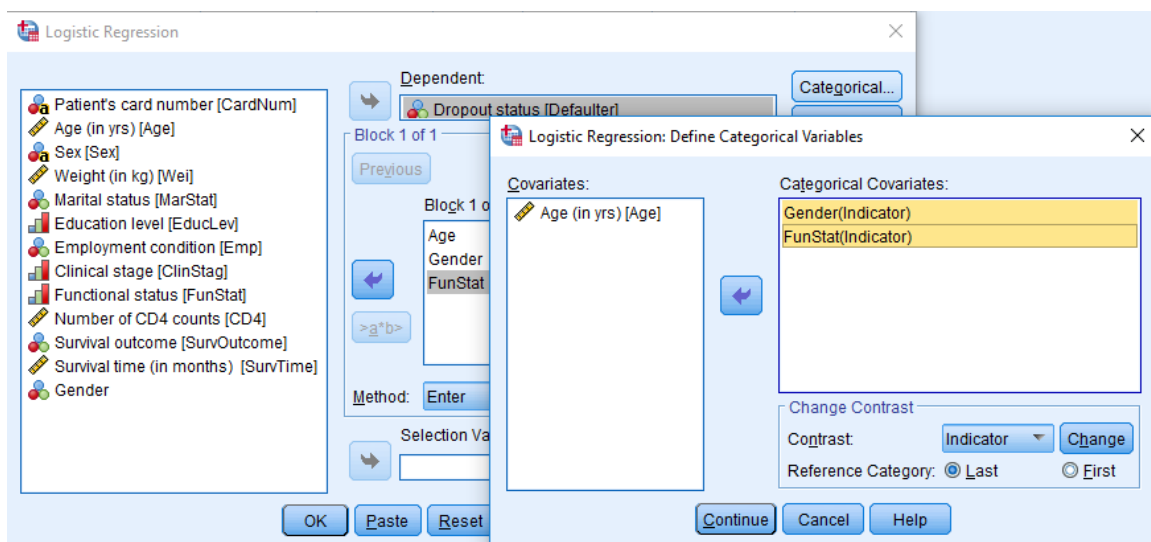
Example 8.1. Using the JUSH_HAART data, suppose, we are interested in identifying the risk factors associated with HAART treatment defaulter patients. Let's consider three explanatory variables: age in years (*Age*), sex of the patients (*Gender*: 1=Male, 2=Female) and functional status (*FunStat*: 0=Working, 1=Ambulatory, 2=Bedridden). Let us fit the logit model and interpret.

The response variable takes the value $y_i = 1$ if the patient was defaulted and $y_i = 0$ otherwise (if the patient was on the treatment). The design variables for Functional Status are:

Functional Status	Design Variables	
	Ambulatory (d_{31})	Bedridden (d_{32})
Working	0	0
Ambulatory	1	0
Bedridden	0	1

Now the model can be written as

$$\text{logit } \pi = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Gender}_i + \beta_{31} \text{Ambulatory}_i + \beta_{32} \text{Bedridden}_i$$



The parameter estimates of the model are provided in the **Variables in the Equation** table of the output.

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	Age (in yrs)	-.029	.008	13.841	1	.000	.971	.956	.986
	Gender(1)	.531	.137	14.947	1	.000	1.700	1.299	2.224
	Functional status			34.322	2	.000			
	Functional status(1)	-1.357	.283	23.045	1	.000	.257	.148	.448
	Functional status(2)	-.789	.292	7.286	1	.007	.454	.256	.806
	Constant	.671	.376	3.180	1	.075	1.957		

a. Variable(s) entered on step 1: Age (in yrs), Gender, Functional status.

As it can be seen from this result, Age, Gender and Functional Status (since both of the design variables are significant) are significant at 5% level of significance.

- When the age of a patient increases by one year, the odds of being defaulted decreases by a factor of 0.971 (AOR=0.971, 95%CI: 0.956-0.986) assuming all other variables are same.
- Also, males are 1.700 times more likely (AOR=1.700, 95%CI: 1.299-2.224) to default than females, that is, the odds of being defaulted is 70.0% higher (AOR=1.700, 95%CI: 1.299-2.224) for males than for females, assuming the other variables constant.
- Again, assuming all other variables constant, working and ambulatory patients are 74.3% (AOR=0.257, 95%CI: 0.148-0.448) and 54.6% (AOR=0.454, 95%CI: 0.256-0.806) times less likely to be defaulted than bedridden patients, respectively.

8.2 Multinomial Logistic Regression

A *multinomial* variable has multiple categories with no sensible ordering, for example: cause of death, type of cancer. A statistical model used for multinomial outcome variable is *multinomial* (also called *nominal*) *logistic regression*.

One of the categories is taken as a *reference* (either the most frequent category or the one to which we wish to draw contrast to). Then, multinomial logistic regression models the *logit* of each outcome category instead of the *reference* category. If the number of outcome categories is J , then there will be $J - 1$ comparisons (logit models).

Given four categories, A, B, C and D, and the reference category is A. Then, the multinomial logistic regression calculates the logit for being in category B *versus* A, category C *versus* A, and category D *versus* A.

The model is of the form:

$$\log\left(\frac{\hat{\pi}_j}{\hat{\pi}_J}\right) = \hat{\beta}_{0j} + \hat{\beta}_{1j}x_1 + \hat{\beta}_{2j}x_2 + \cdots + \hat{\beta}_{kj}x_k \cdots + \hat{\beta}_{pj}x_p; j = 1, 2, \dots, J - 1$$

where

- $\hat{\pi}_j$ is the estimated probability of the j^{th} category.

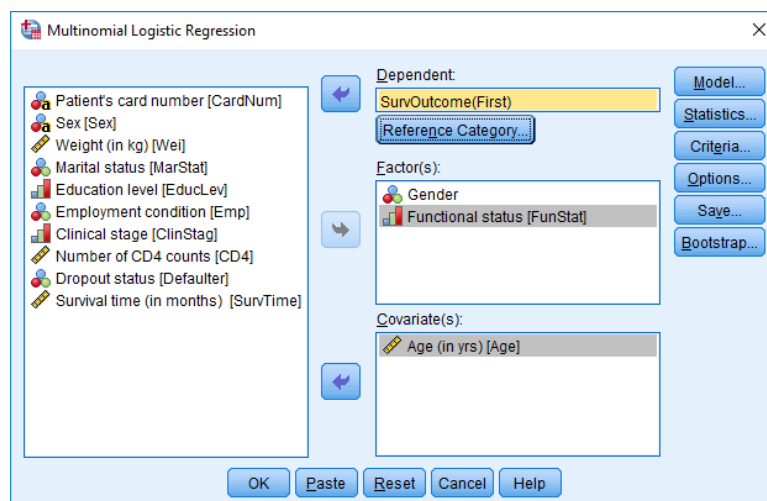
- $\hat{\beta}_{0j}$ is the estimated intercept of the j^{th} logit model.
- $\hat{\beta}_{kj}$; $k = 1, 2, \dots, p$ is the k^{th} slope estimate of the j^{th} logit model.

Like a binary logistic regression, the parameter estimates are interpreted in terms of odds ratio, $OR = \exp(\beta_{kj})$.

- Given $\hat{\beta}_{kj} > 0 \Rightarrow OR > 0$. For each unit increase in the k^{th} independent variable, the odds of category j (instead of category J) increases by a factor of $\exp(\hat{\beta}_{kj})$.
- Given $\hat{\beta}_{kj} < 0 \Rightarrow OR < 0$. For each unit increase in the k^{th} independent variable, the odds of category j (instead of category J) decreases by a factor of $\exp(\hat{\beta}_{kj})$.

From the *Menu* bar, click on: **Analyze** → **Regression** → **Multinomial Logistic...** Then, in the **Multinomial Logistic Regression** dialogue box, enter the dependent (response) variable in the **Dependent:** box, and move all the qualitative and quantitative independent (explanatory) variables into the **Factor(s):** and **Covariate(s):** box, respectively.

Example 8.2. Based on the survival outcome of HAART treatment, HIV/AIDS patients were classified into four categories (0= Active, 1= Dead, 2= Transferred to other hospital, 3= Lost-to-follow). To identify factors associated with these survival outcomes, let us fit a multinomial logit model with three explanatory variables: Age, Gender (1= Male, 2=Female) and Functional Status (0= Working, 1= Ambulatory, 2= Bedridden).



The output is:

		Parameter Estimates						95% Confidence Interval for Exp (B)	
Survival outcome ^a		B	Std. Error	Wald	df	Sig.	Exp(B)	Lower Bound	Upper Bound
Dead	Intercept	-.991	.734	1.822	1	.177			
	Age (in yrs)	-.020	.018	1.164	1	.281	.981	.946	1.016
	[Gender=1]	.564	.325	3.002	1	.083	1.757	.929	3.325
	[Gender=2]	0 ^b	.	.	0
	[Functional status=0]	-2.280	.479	22.668	1	.000	.102	.040	.261
	[Functional status=1]	-1.340	.487	7.574	1	.006	.262	.101	.680
	[Functional status=2]	0 ^b	.	.	0
Transferred	Intercept	-.298	.545	.300	1	.584			
	Age (in yrs)	-.030	.012	5.943	1	.015	.970	.947	.994
	[Gender=1]	.635	.211	9.040	1	.003	1.887	1.247	2.854
	[Gender=2]	0 ^b	.	.	0
	[Functional status=0]	-1.584	.393	16.277	1	.000	.205	.095	.443
	[Functional status=1]	-.751	.401	3.511	1	.061	.472	.215	1.035
	[Functional status=2]	0 ^b	.	.	0
Lost-to-follow	Intercept	-.288	.510	.320	1	.572			
	Age (in yrs)	-.032	.010	9.127	1	.003	.969	.949	.989
	[Gender=1]	.455	.178	6.516	1	.011	1.575	1.111	2.233
	[Gender=2]	0 ^b	.	.	0
	[Functional status=0]	-.828	.395	4.386	1	.036	.437	.201	.948
	[Functional status=1]	-.536	.411	1.702	1	.192	.585	.262	1.309
	[Functional status=2]	0 ^b	.	.	0

a. The reference category is: Active.
 b. This parameter is set to zero because it is redundant.

• Effect of Age:

- The odds of being dead (instead of active) is independent of the age of the patient at $\alpha = 5\%$.
- As the age of the patient increases by 1 year, the odds of being transferred to other hospital (instead of active) decreases by 3% (AOR=0.970, 95%CI: 0.947-0.994).
- An increase by 1 year in the age of the patient decreases the odds of being lost-to-follow (instead of active) decreases by 3.1% (AOR=0.969, 95%CI: 0.949-0.989).

• Effect of Gender:

- The odds of being dead (instead of active) is independent of the gender of the patient at $\alpha = 5\%$.
- The odds that male patients being dead (instead of active) is 1.887 times that of females, or male patients are 1.887 times more likely (AOR=1.887, 95%CI: 1.247-2.854) to be dead (instead of active) than female patients.
- The odds of being lost-to-follow (instead of active) for male patients is 1.575 times odds of being lost-to-follow (instead of active) for females, or male patients are 57.5% more likely (AOR=1.575, 95%CI: 1.111-2.233) to be lost-to-follow (instead of active) than female patients.

• Effect of Functional Status:

- Working and ambulatory patients are 0.102 (AOR=0.102, 95%CI: 0.040-0.261) and 0.262 (AOR=0.262, 95%CI: 0.101-0.680) times less likely, respectively, to be dead (instead of active) than those bedridden patients.
- Similarly, working patients are 79.5% (AOR=0.205, 95%CI: 0.095-0.443) and 56.3% (AOR=0.437, 95%CI: 0.201-0.948) times less likely to be transferred to other hospital and lost-to-follow, respectively, (instead of active) than bedridden patients.
- But, the odds of being transferred to other hospital and lost-to-follow (instead of active) are not significantly different between ambulatory and bedridden patients at 5% level of significance.

8.3 Ordinal Logistic Regression

An *ordinal* variable has multiple categories that can be ordered or ranked, for example; anemia level (mild, moderate, severe). Mostly "larger" values are assumed to correspond to "higher" outcome categories. A statistical model used for ordinal outcome variable is *ordinal logistic regression*.

Ordinal logistic regression models the *cumulative logits* at all possible dichotomization cut-points of the ordinal outcome. If the number of outcome categories is J , then there will be $J - 1$ ways of dichotomizations (comparisons). For example, in a 5-level ordinal outcome, there are 4 possible cut-points: 1 *versus* 2-5, 1-2 *versus* 3-5, 1-3 *versus* 4-5, and 1-4 *versus* 5.

There is an assumption called *proportional odds*. Proportional odds means that the effect is about the same regardless of the cut-point of the ordinal variable.

The model is:

$$\log \left[\frac{\hat{\pi}(y \leq j)}{\hat{\pi}(y > j)} \right] = \hat{\beta}_{0j} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \cdots + \hat{\beta}_p x_p; j = 1, 2, \dots, J - 1$$

where

- $\hat{\pi}(y \leq j)$ is the estimated cumulative probability of category j and below.
- $\hat{\beta}_{0j}$ is the estimated intercept of the ordinal logit model.
- $\hat{\beta}_k; k = 1, 2, \dots, p$ is the k^{th} cumulative slope estimate of the ordinal logit model.

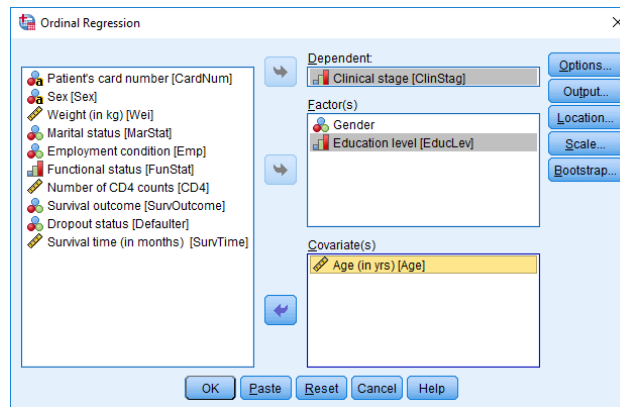
The parameter estimates are interpreted in terms of (cumulative) odds ratio, $COR = \exp(\beta_k)$.

- Given $\hat{\beta}_k > 0 \Rightarrow COR > 1$. For each unit increase in the k^{th} independent variable, the odds of higher categories increases by a factor of $\exp(\hat{\beta}_k)$.
- Given $\hat{\beta}_k < 0 \Rightarrow COR < 1$. For each unit increase in the k^{th} independent variable, the odds of higher categories decreases by a factor of $\exp(\hat{\beta}_k)$.

An ordinal logit model has a proportionality assumption which means the distance between each category is equivalent (proportional odds assumption).

From the *Menu* bar, click on: **Analyze** → **Regression** → **Ordinal...** Then, in the **Ordinal Regression** dialogue box, enter the dependent (response) variable in the **Dependent:** box, and move all the qualitative and quantitative independent (explanatory) variables into the **Factor(s):** and **Covariate(s):** box, respectively.

Example 8.3. To determine the effect of Age, Gender (0= Female, 1=Male) and Education Level (0= No Education, 1= Primary, 2= Secondary, 3= Tertiary) on the baseline Clinical Stage of HIV/AIDS patients (1= Stage I, 2= Stage II, 3= Stage III and 4= Stage IV) at the time of starting the HAART treatment, the following parameter estimates of ordinal logistic regression are obtained.



The parameter estimates are presented below (SPSS provides only the parameter estimates, it does not provide cumulative odds ratios).

Parameter Estimates								
		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[ClinStag = 1]	-.705	.242	8.521	1	.004	-1.178	-.232
	[ClinStag = 2]	.833	.242	11.862	1	.001	.359	1.306
	[ClinStag = 3]	3.042	.259	138.329	1	.000	2.535	3.549
Location	Age	.003	.005	.276	1	.599	-.008	.014
	[Gender=1]	.220	.105	4.431	1	.035	.015	.425
	[Gender=2]	0 ^a	.	.	0	.	.	.
	[EduLev=0]	.489	.183	7.161	1	.007	.131	.847
	[EduLev=1]	.251	.168	2.234	1	.135	-.078	.581
	[EduLev=2]	.315	.169	3.458	1	.063	-.017	.647
	[EduLev=3]	0 ^a	.	.	0	.	.	.
Link function: Logit.								
a. This parameter is set to zero because it is redundant.								

- The cumulative estimates associated with age $\hat{\beta}_1 = 0.003$ (corresponding cumulative odds ratio $\exp(0.003) = 1.003$) suggests that the cumulative probability starting at the clinical stage IV end of the scale increases as the age of the patient increases (an increase in the age of the patient leads to be in higher clinical stages) given the gender and education level. But, it is not significant at 5%.
- The estimate $\hat{\beta}_2 = -0.220$ (corresponding cumulative odds ratio $\exp(-0.220) = 0.803$) indicates the estimated odds of being in the clinical stage below any fixed level for males

are 0.803 times the estimated odds for female patients (females are more likely to be in lower clinical stages as compared to males or males are more likely to be in higher clinical stages as compared to females) given the age and education level of the patient.

- Patient with no education are 1.631 (cumulative odds ratio $\exp(0.489) = 1.631$) times more likely to be in higher clinical stages than those patients with tertiary education. But, the effects of primary and secondary education compared to tertiary education are not significant.

Chapter 9

Survival Models

9.1 Cox Regression

Cox regression is used for modeling for time-to-occurrence of an outcome of interest, such as time-to-death, time-to-development of cancer. If the outcome of interest occurs, it is called an *event*. If the outcome of interest does not occur, it is called a *censored*.

Cox regression models the natural logarithm of the *relative hazard* of the event of interest. A hazard is the probability of the event during any given time point. Calling the probability of a bad event of interest, like death, "hazard" might not be strange. But, it feels strange to think of the hazard of a positive outcome, like recurring from a disease. But technically, it is the same thing.

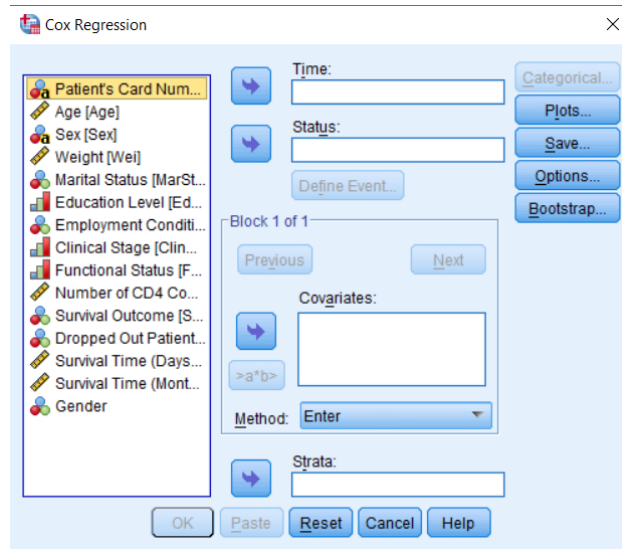
The model is: $\log \hat{h}(t) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k + \dots + \hat{\beta}_p x_p$. where

- $\hat{h}(t)$ is the estimated hazard rate at time t .
- $\hat{\beta}_0$ is the estimated intercept of the log hazard model.
- $\hat{\beta}_k; k = 1, 2, \dots, p$ is the k^{th} slope parameter estimate.

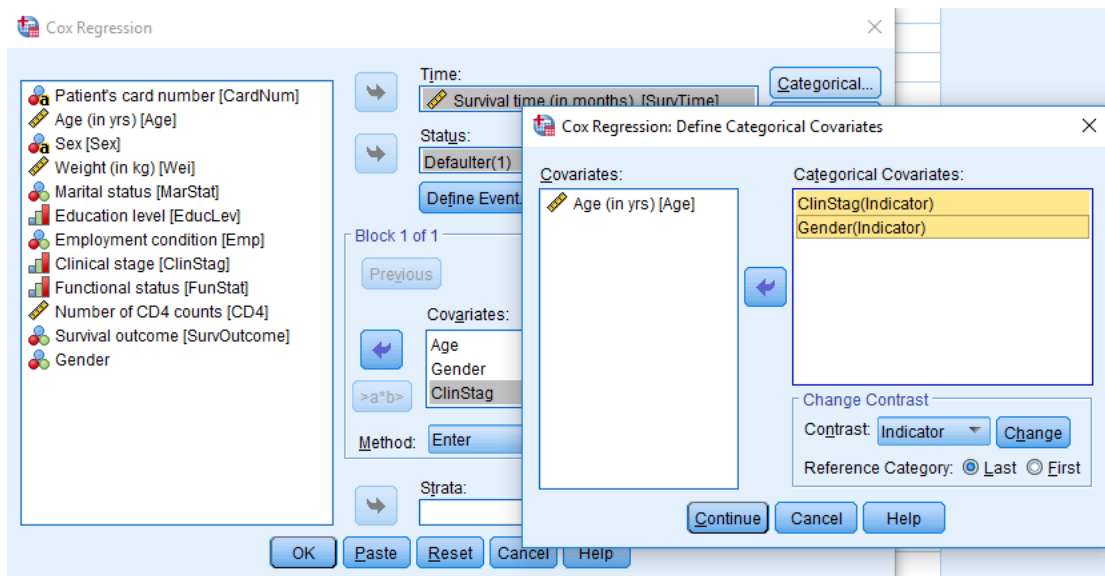
Estimates are interpreted in terms of hazard ratio, $HR = \exp(\beta_k)$.

- Given $\hat{\beta}_k > 0 \Rightarrow HR > 1$. For each unit increase in the k^{th} independent variable, the relative hazard increases by a factor of $\exp(\hat{\beta}_k)$.
- Given $\hat{\beta}_k < 0 \Rightarrow HR < 1$. For each unit increase in the k^{th} independent variable, the relative hazard decreases by a factor of $\exp(\hat{\beta}_k)$.

From the *Menu* bar, click on: **Analyze** → **Survival** → **Cox Regression**. Next, in the **Cox Regression** dialogue box, enter the time-to-event response in the **Time:** box, and move all the independent (explanatory) variables into the **Covariate(s):** box. Then, click on the **Categorical** button to move all the categorical independent variables into the **Categorical Covariates:** box.



Example 9.1. To determine the effect of Age, Gender (0= Female, 1=Male) and Clinical Stage (1= Clinical Stage I, 2= Clinical Stage II, 3= Clinical Stage III, 4= Clinical Stage IV) on the time-to-defaulting from the HAART treatment, Cox regression were fitted.



The parameter estimates are provided in the **Variables in the Equation** table of the output.

Variables in the Equation								
	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age (in yrs)	-.031	.007	18.546	1	.000	.969	.956	.983
Gender	.489	.118	17.262	1	.000	1.631	1.295	2.054
Clinical stage			12.684	3	.005			
Clinical stage(1)	-.737	.225	10.707	1	.001	.478	.308	.744
Clinical stage(2)	-.545	.208	6.885	1	.009	.580	.386	.871
Clinical stage(3)	-.370	.207	3.205	1	.073	.691	.460	1.036

- As the age of the patient increases by 1 year, the hazard of defaulting decreases by a factor of 0.969 (AHR=0.969, 95%CI: 0.956-0.983).
- The hazard of defaulting for male patients increases by 63.1% (AHR=1.631, 95%CI: 1.295-2.054) relative to female patients.
- The hazard for patients who were in clinical stage I and II when starting the treatment are 0.478 (AHR=0.478, 95%CI: 0.308-0.744) and 0.580 (AHR=0.580, 95%: 0.386-0.871) times the hazard of those patients who were in clinical stage IV. But, the hazard of defaulting for patients who were in clinical stage III is not significantly different from those patients who were in clinical stage IV.

Note: In the Cox modelling, there is an assumption of *proportional hazards*. The assumption is that the hazards for persons with different patterns of factors (covariates) are constant over time. For example, if the relative hazard of heart attack among diabetics is three times higher than among nondiabetics in the first year of the study, the relative hazard of heart attack must also be (about) three times higher among diabetics than nondiabetics in the second year of the study. Note that the hazard for a heart attack can be very different in the first year than in the second year (e.g., much higher in the first year than in the second year), but the difference between the hazards for diabetics and nondiabetics must be constant throughout the study period.