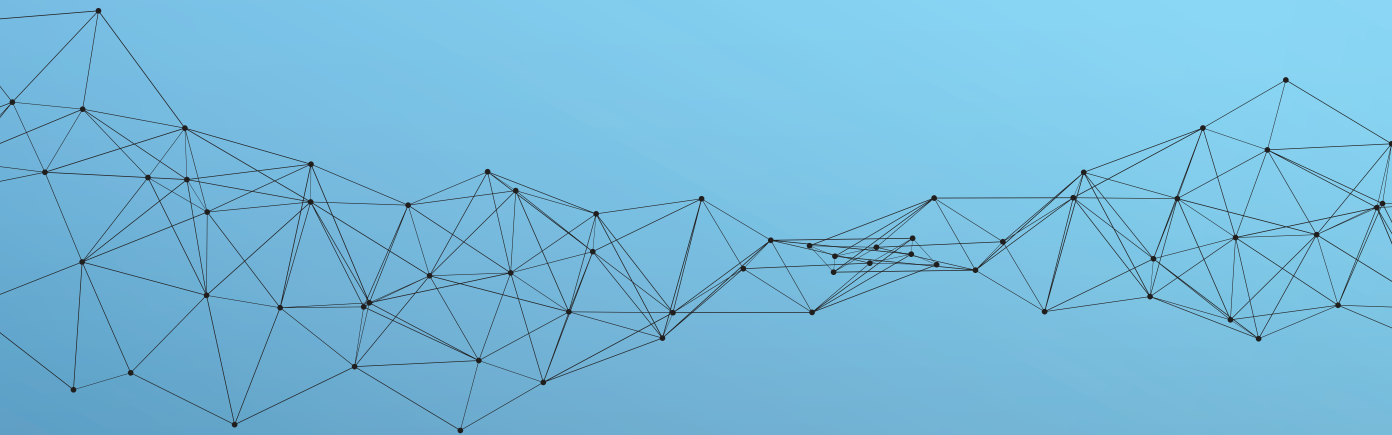


LEARNING
SPSS
WITHOUT
PAIN

MOHAMMAD TAJUL ISLAM



Learning SPSS without Pain
A Comprehensive Manual for Data Analysis and Interpretation of Outputs

First Edition

Mohammad Tajul Islam

MBBS (Dac), DTM&H (Thai), MSc (CTM) (Thai), MPH (Thai)

Professor (Adjunct), Department of Public Health
North South University and State University of Bangladesh

ASA Publications

Dhaka, Bangladesh

ISBN 978-984-34-8254-9

Copyright © 2020 by Author

First published 2020

This e-manual is free for everyone. Anyone can download this manual from the links, print it out for personal use and share with others, but strictly prohibited to use it for any kind of commercial interest. The datasets provided in the links and used in this manual are hypothetical. It is also free for everyone to download and use the datasets for practice.

Publisher

ASA Publications

Dhaka, Bangladesh

Email: asapublications7@gmail.com

Distributor

Altaf Medical Book Center

Shop: 121 & 128; Lane: 3; Islamia Market, Nilkhet, Dhaka 1205

Cell: +880-1711-985991; +880-1611-985991; +880-1511-985991

Price: Tk. 250

Production credits

Publisher: ASA Publications

Production Director: Md Altaf Hossain

Marketing Manager: Md Altaf Hossain

Composition: Zariath Al-Mamun Badhon and Md Mamunur Rasel

Cover Design: Zariath Al-Mamun Badhon

Printing: ASA Publications

Binding: Rahim Bindings

Printed in Bangladesh

To all my family members, students and young researchers in
health and social sciences

Preface

This manual is intended for the students (MPH, FCPS, MD, MS, MPhil, and others), teachers and young researchers in health and social sciences. It is written in a very simple language and used health related data as examples. This manual answers three basic questions related to data analysis. They are: a) what to do (what statistics to be used for the data analysis to achieve the objectives); b) how to do (how to analyze data by SPSS); and c) what do the outputs mean (how to interpret the outputs)? All these questions are answered in a very simple and understandable manner with examples. This manual covers the basic statistical methods of data analysis used in health and social science research. It is the gateway to learn SPSS and will help the users to go further. This manual is organized in 22 sections that covers the data management, descriptive statistics, hypothesis testing using bivariate and multivariable analysis and others. It is easier to learn through exploration rather than reading. The users can explore further once the basics are known. From my understanding, using the statistics as covered in this manual, the students and researchers will be able to analyze most of their data on epidemiological studies and publish them in the international peer review journals.

I am optimistic that this manual will make the students' and researchers' life easier for analyzing data and interpreting the outputs meaningfully. *The users can download the datasets used in this manual from the following links.* If you have any comments about the manual, feel free to write at the e-mail address below.

Links for the e-manual and data files:

Link 1: <https://github.com/rubyrider/Learning-SPSS-without-Pain>

Link 2: <https://jmp.sh/F65fcni>

Link 3: <https://drive.google.com/drive/folders/1t9QjNMBV-bI-oyAwEBQEm8auuYv3sP7KW?usp=sharing>

M. Tajul Islam
abc.taj@gmail.com
smpp.taj@gmail.com

Acknowledgements

I am particularly indebted and grateful to Harun Or Rashid, Monjura Khatun Nisha and Shakil Ahmed, who reviewed all the sections and provided their critical views and constructive suggestions. Nisha and Shakil helped me in editing the final manuscript. Shakil took the lead role for publishing this manual. Many of my students, who continually pushed and encouraged me to write this manual deserve a share of the credit, including SM Anwar Sadat, Md. Naymul Hasan, Md. Golam Kibria, Sinthia Mazumder, Humayun Kabir and many others. Finally, I express my sincerest thanks for the assistance that I have received from the ASA Publications for publishing the manual. Being the author, I accept full responsibility for any deficiencies that the manual may have.

To the users

Of course, this e-manual is free for all. However, if you can afford, please donate Tk.100 only (US\$ 2 for the users outside Bangladesh) to any charity or a needy person. This little amount is sufficient to offer a meal to an orphan in the developing countries.

Contents

Section 1	Introduction	1
1.1	Steps of data analysis	1
Section 2	Generating Data File	3
2.1	Generating data file	3
2.1.1	Defining variables	6
2.1.2	Data entry in SPSS	9
2.2	Data used in this manual	9
Section 3	Data Cleaning and Data Screening	11
3.1	Checking for out-of-range errors	11
3.2	Checking for outliers	12
3.3	Assessing normality of a dataset	13
Section 4	Data Analysis: Descriptive Statistics	14
4.1	Frequency distribution	14
4.2	Central tendency and dispersion	16
4.2.1	Outputs	18
4.2.2	Interpretation	18
4.3	Alternative method of getting measures of central tendency and dispersion	19
4.3.1	Outputs	19
4.3.2	Interpretation	21
4.4	Descriptive statistics and histogram disaggregated by Sex	22
4.5	Checking for outliers	23
Section 5	Checking Data for Normality	25
5.1	How to understand that the data have come from a normally distributed population	25
5.1.1	Outputs	27
5.1.2	Interpretation	27
Section 6	Data Management	30
6.1	Recoding of data	30
6.1.1	Recoding into same variable	30
6.1.2	Recoding into different variable	32
6.2	Making class intervals	33
6.3	Combine data into a new variable	36
6.4	Data transformation	39

6.5	Calculation of total score	41
6.6	Calculation of duration	44
6.7	Selecting a sub-group for analysis	45
Section 7	Testing of Hypothesis	46
7.1	Association between quantitative and qualitative or quantitative variables	47
7.2	Association between two qualitative variables	48
7.3	Multivariable analysis	49
7.4	Agreement analysis	50
Section 8	Student's t-test for Hypothesis Testing	51
8.1	One-sample t-test	51
8.1.1	Outputs	52
8.1.2	Interpretation	52
8.2	Independent samples t-test	52
8.2.1	Commands	53
8.2.2	Outputs	53
8.2.3	Interpretation	54
8.3	Paired t-test	55
8.3.1	Commands	56
8.3.2	Outputs	56
8.3.3	Interpretation	56
Section 9	Analysis of Variance (ANOVA): One-way ANOVA	58
9.1	One-way ANOVA	58
9.1.1	Commands	59
9.1.2	Outputs	59
9.1.3	Interpretation	60
9.1.3.1	Interpretation of multiple comparisons table	61
9.1.4	Graph on distribution of medians/means	61
9.1.5	What to do if the variances are not homogeneous	62
9.1.5.1	Outputs	62
9.1.5.2	Interpretation	63
Section 10	Two-way ANOVA	64
10.1	Two-way ANOVA	64
10.1.1	Commands	65
10.1.2	Outputs	65
10.1.3	Interpretation	67

Section 11	Repeated Measures ANOVA: One-way	69
11.1	One-way repeated measures ANOVA	69
11.1.1	Commands	70
11.1.2	Outputs	70
11.1.3	Interpretation	73
Section 12	Repeated Measures ANOVA: Within and Between-Subjects	75
12.1	Within and between-subjects ANOVA	75
12.1.1	Commands	76
12.1.2	Outputs	77
12.1.3	Interpretation	82
Section 13	Association between Two Categorical Variables: Chi-Square Test of Independence	85
13.1	Chi-square test of Independence	85
13.1.1	Commands	85
13.1.2	Outputs	86
13.1.3	Interpretation	86
Section 14	Association between Two Continuous Variables: Correlation	89
14.1	Pearson correlation	89
14.1.1	Commands for scatter plot	90
14.1.2	Commands for Pearson correlation	91
14.1.3	Outputs	92
14.1.4	Interpretation	92
14.2	Spearman's correlation	93
14.2.1	Commands for Spearman's correlation	93
14.2.2	Outputs	93
14.2.3	Interpretation	93
14.3	Partial correlation	93
14.3.1	Commands	94
14.3.2	Outputs	94
14.3.3	Interpretation	94
Section 15	Linear Regression	96
15.1	Simple linear regression	97
15.1.1	Commands	97
15.1.2	Outputs	98
15.1.3	Interpretation	99

15.2	Multiple linear regression	100
15.2.1	Creating dummy variables	101
15.2.2	Changing string variable into numeric variable	102
15.2.3	Sample size for multiple regression	103
15.2.4	Commands for multiple linear regression	103
15.2.5	Outputs	103
15.2.6	Interpretation	104
15.2.7	Regression equation	105
15.2.8	Problem of multicollinearity	106
15.2.9	Checking for assumptions	108
15.2.9.1	Checking for outliers and independent data points	108
15.2.9.2	Checking for normality assumption of the residuals and constant variance	110
15.2.10	Variable selection for the model	111
15.2.10.1	Outputs	113
15.2.10.2	Interpretation	113
Section 16	Logistic Regression	115
16.1	Logistic regression analysis	115
16.1.1	Commands	116
16.1.2	Outputs	116
16.1.3	Interpretation: Basic tables	117
16.1.4	Interpretation: Outputs under Block 0	118
16.1.5	Interpretation: Outputs under Block 1	120
16.1.6	ROC curve	123
16.1.7	Sample size for logistic regression	125
16.1.8	Variable selection for a model	125
Section 17	Survival Analysis	128
17.1	Survival analysis: Kaplan-Meier method	129
17.1.1	Commands	129
17.1.2	Outputs	130
17.1.3	Interpretation	132
Section 18	Cox Regression	135
18.1	Cox Regression or Proportional Hazards Regression	135
18.1.1	Commands	135
18.1.2	Outputs	136
18.1.3	Interpretation	137
Section 19	Non-parametric Methods	140
19.1	Mann-Whitney U test	140

19.1.1	Commands	141
19.1.2	Outputs	141
19.1.3	Interpretation	141
19.2	Wilcoxon Signed Ranks test	142
19.2.1	Commands	142
19.2.2	Outputs	142
19.2.3	Interpretation	143
19.3	Kruskal-Wallis test	143
19.3.1	Commands	143
19.3.2	Outputs	144
19.3.3	Interpretation	144
19.4	Friedman test	144
19.4.1	Commands	145
19.4.2	Outputs	145
19.4.3	Interpretation	145
19.5	Chi-square test for goodness-of-fit	146
19.5.1	Commands	146
19.5.2	Outputs	146
19.5.3	Interpretation	147
Section 20	Checking Reliability of Scale: Cronbach's Alpha	148
20.1	Cronbach's alpha	148
20.1.1	Outputs	149
20.1.2	Interpretation	150
Section 21	Analysis of Covariance (ANCOVA): One-way ANCOVA	151
21.1	One-way ANCOVA	151
21.1.1	Commands	153
21.1.2	Outputs: Homogeneity of regression slopes	153
21.1.3	Interpretation: Homogeneity of regression slopes	153
21.1.4	Outputs: One-way ANCOVA	154
21.1.5	Interpretation: One-way ANCOVA	156
Section 22	Two-way ANCOVA	158
22.1	Two-way ANCOVA	158
22.1.1	Commands	159
22.1.2	Outputs	159
22.1.3	Interpretation	162
Annex		165
References		166

Section 1

Introduction

SPSS stands for Statistical Package for Social Sciences. It is a powerful window-based statistical data analysis software. The menu and dialog-box system of SPSS have made the program user-friendly. SPSS is particularly useful to the researchers in public health, medicine, social science and other disciplines. It supports a wide range of univariate, bivariate, multivariable and multivariate data analysis procedures.

This manual is intended for the students, teachers and researchers involved in health and social science research. It provides practical guidance for using SPSS for basic statistical analysis of data. Once the data file is loaded in SPSS, the users can select items from a dropdown menu to analyze data, make graphs, transform variables, and others. SPSS, in general, has made the life of the researchers easier for data analysis. This manual is based on SPSS version 16.0. Although higher versions are available, they offer little extra advantage to the users for commonly used statistical methods of data analysis. Rather, if the user does not have the right computer to support higher version, the execution of the program may become difficult.

This manual is primarily developed targeting the Master of Public Health (MPH) and post-graduate students in medicine (FCPS, MD, MS, and MPhil), keeping in mind their needs. The aim of this manual is to provide a concise but clear understanding on how to conduct a range of statistical analyses using SPSS and interpret the outputs. Special emphasis is given on understanding the SPSS outputs, which is a problem for many of the users. For better understanding of the users, examples and data related to health research are used. However, to use the manual effectively and to understand the outputs, it requires basic knowledge on biostatistics and epidemiology. The users will find it easier if they review the relevant statistical concepts and procedures, and epidemiological methods before using this manual.

1.1 Steps of data analysis

We collect data for our studies using various tools and methods. The commonly

used tools for data collection are questionnaire and record sheet, while the commonly used data collection methods are face-to-face interview, observation, physical examination, lab test and others. Sometimes we use the available data (secondary data) for our research studies, for example, hospital records, and data of other studies (e.g., Bangladesh Demographic and Health Survey data). Once data is collected, the steps of data analysis are:

- Data coding, if pre-coded questionnaire or record sheet is not used
- Development of data file and data entry
- Data cleaning (checking for errors in data entry)
- Data screening (checking assumptions for statistical tests)
- Data analysis
- Interpretation of results

In the following sections, I have discussed the development of data file, data management, data analysis and interpretation of the outputs.

Section 2

Generating Data File

Like other data analysis programs, SPSS has to read a data file to analyze data. We, therefore, need to develop a data file for the use of SPSS. The data file can be generated by SPSS itself or by any other program. Data files generated in other programs can be easily transferred to SPSS for analysis. Here, I shall discuss how to generate a data file in SPSS.

2.1 Generating data file

The first step in creating a data file is to give a “name” and “define” variables included in the questionnaire/ record sheet. The next step is entering data in SPSS. Suppose, we have collected data using a pre-coded questionnaire (codes are shown in the parenthesis) with the following variables.

Categorical variables:

- Sex (m= male; f = female)
- Religion (1= Islam; 2= Hindu; 3= Others)
- Occupation (1= Business; 2= Government job; 3= Private job; 4= Others)
- Marital status (1= Married; 2= Unmarried; 3= Others)
- Have diabetes mellitus (0= No; 1= Yes; 3= Don’t know)

Quantitative variables (numerical variables):

- Age of the respondent
- Monthly family income
- Systolic blood pressure (BP)
- Diastolic BP

Suppose, we have decided to use V1 as the SPSS variable name for age, V2 for sex, V3 for religion, etc. (table 2.1). Instead of V1, V2, V3, you can use any other variable name (e.g., age for age, sex for sex, etc.) for your variables. It is always better to develop a codebook in MS Word or MS Excel before entering data, as shown in table 2.1. This is helpful during data analysis.

Table 2.1. Codebook

SPSS variable name	Actual variable name	Variable code
V1	Age in years	Actual value
V2	Sex	m= Male f= Female
V3	Religion	1= Islam 2= Hindu 3= Others
V4	Occupation	1= Business 2= Government job 3= Private job 4= Others
V5	Monthly family income	Actual value
V6	Marital status	1= Married 2= Unmarried 3= Others
V7	Have diabetes mellitus	1= Yes 2= No
V8_a	Systolic blood pressure	Actual value
V8_b	Diastolic blood pressure	Actual value

Note: Instead of V1, V2, etc., you can use any other name as SPSS variable name. For example, you can use the variable name “age” instead of V1, “sex” instead of V2, etc.


Now, open the SPSS program by double clicking the SPSS icon. You will see the following dialogue box (fig 2.1). Click on cancel box () to close the “SPSS for Windows”. Now we have the dialogue box as shown in fig 2.2 (SPSS Data Editor).

Figure 2.1. Dialogue box for defining variables

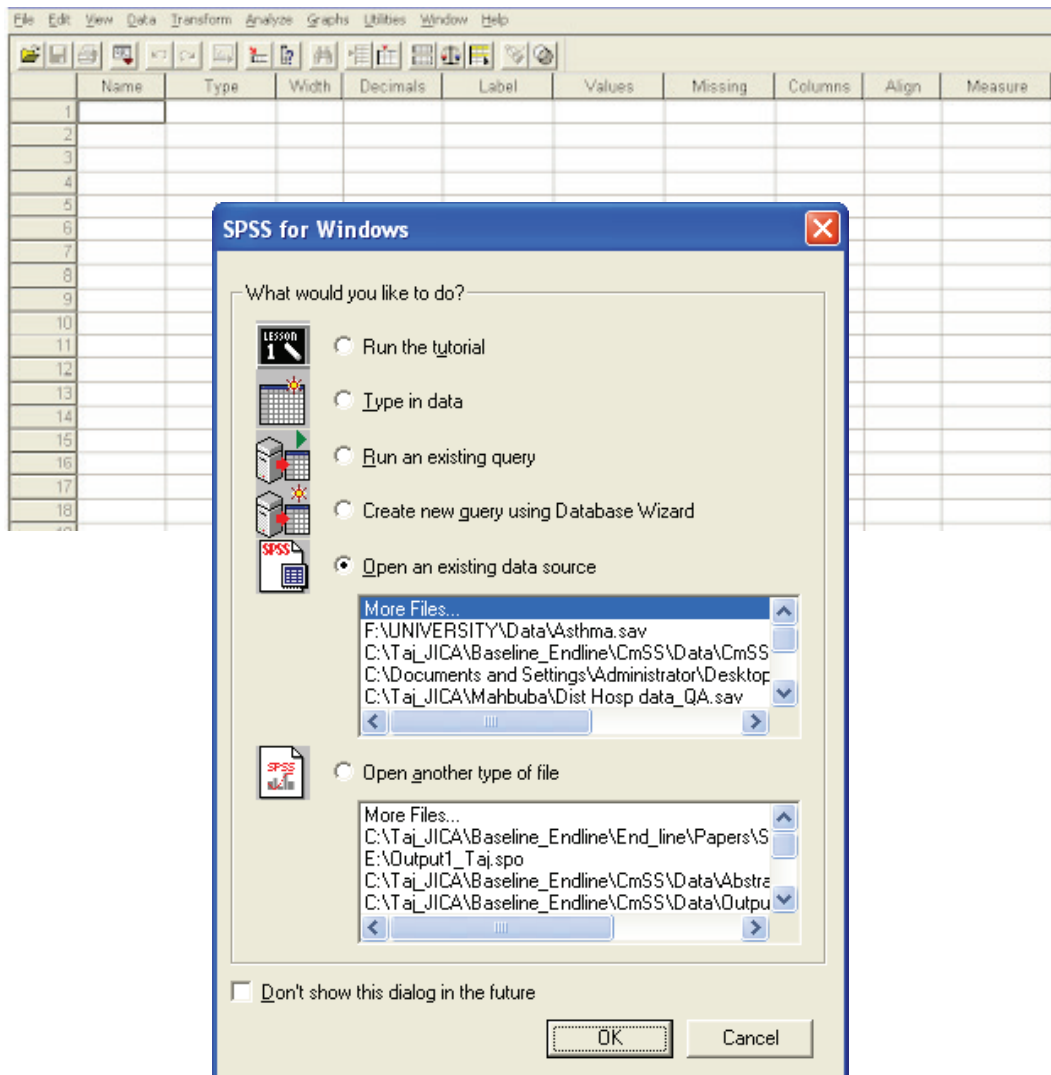
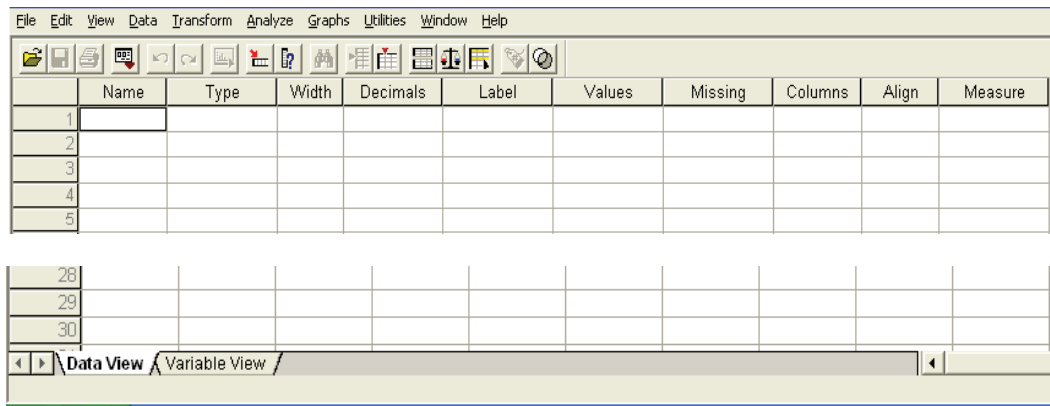


Figure 2.2. SPSS data editor: dialogue box for defining variables



The “SPSS Data Editor (fig 2.2)” shows Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure, at the top row. If you do not see this, click on “Variable View” at the left-bottom corner of the window.

2.1.1 Defining variables:

We shall use SPSS data editor to define the variables. Before entering data, all the study variables need to be defined including their coding information. The lower versions of SPSS (version 12 and below) allow only 8 characters to name a variable. The higher versions (13 and above) allow up to 64 characters to name a variable. While writing the variable names, we need to follow certain rules. They are:

- The variables must be unique (all variables should have different names)
- Variables must begin with a letter (small or capital) rather than a number
- Cannot include full stop (.), space, or symbols like, ?, *, μ , λ , etc.
- Cannot include words that are used as commands by SPSS, such as ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH, etc.
- Cannot exceed 64 characters

To define the variables, follow the instructions below:

Name: In this column, type the brief SPSS variable name as shown in the code-book. For example, type V1 (you can also use “age” as the variable name) for the first variable, age. Note that this short name will be used to identify the variable in the data file.

Type: This column indicates the characteristic of the variable, whether it is numeric or string. Numeric means expressed by numbers (e.g., 1, 2, 3, etc.), while string means expressed by alphabets (e.g., m, f, y, n, etc.). In SPSS, the default value for “Type” is numeric. If the nature of the variable is string (alphabet or text variable), we need to change it. To change the variable type into string, follow the following steps:

Click on the cell under the column “Type” (you will see a box with three dots)
> Click on the “three-dot box” (you will see the options in a separate dialogue box) > Select “String” from the options > Click OK

Similarly, if it is a date variable (e.g., date of hospital admission), you have to change the variable type into a date format in the same manner.

Width: The default value for width is 8. In most of the cases, it is sufficient and serves the purpose. However, if the variable has very large value, then we need to increase it using the arrow (up and down) button in the box. *For practical purpose, keep the width 8 unless the variable values are larger than 8 characters.*

Decimals: This is applicable for the numeric variables and the default value is 2. If your data does not have any decimal value, you can make it “0” using the down arrow or keep it as it is.

Label: This is the space where we write the longer description of the variable (actual variable name as shown in the codebook). For example, we have used “V1” to indicate age in years. We should, therefore, write “Age in years” in the label column for the variable name “V1”.

Values: This is applicable for the variables to define their levels using code numbers (such as 1, 2 or m, f, etc). This allows the SPSS to retain the meaning of values (code numbers) you have used in the data set. For example, our variable 2 is sex and is defined by “V2”. It has two levels, male (coded as “m”) and female (coded as “f”). Follow the commands below to put the value labels.

Click on cell under the column “Value” (you will see a box with three dots) > Click on “three-dot box” > Click in the box “Value” > Type “m” > Click in the box “Value label” > Type “male” > Click on “Add” > Repeat the same process

for female (value “f”, value label “female”, add) > OK

In this way, complete value labels for all the variables, if applicable. *Note that value labels are needed only for the variables that have been coded.*

Missing: If there is any missing value in the dataset, SPSS has the option to indicate that. If we want to set a missing value for a variable, we have to select a value that is not possible (out of range) for that variable. For example, we have conducted a study and the study population was women aged 15-49 years. There are several missing values for age in the data (i.e., age was not recorded on the questionnaire or respondent did not tell the age). First, we have to select a missing value for age. We can select any value which is outside the range 15-49 as the missing value. Say, we have decided to use 99 as the missing value for age. Now, to put the missing value for age in SPSS, use the following commands.

Click on the cell under the column “Missing” (you will see a box with three dots) > Click on the “three-dot box” > Select “Discrete missing values” > Click on the left box > type “99” > Ok

However, you may omit this. Just keep the cell blank while entering data in the data file. SPSS will consider the blank cells in the data file as missing values (system missing).

Columns: The default value for this is 8, which is sufficient for most of the cases. If you have a long variable name, then only change it as needed. For practical purpose, just keep it as it is.

Align: You do not need to do anything for this.

Measure: This cell indicates the measurement scale of the data. If the variable is categorical use “Nominal” for nominal or “Ordinal” for ordinal scale of measurement. Otherwise use “Scale” for interval or ratio scale of measurement. You can also keep it as it is.

In this way, define all the variables of your questionnaire/record sheet in the SPSS data editor. The next step is data entry.

2.1.2 Data entry in SPSS:

Once all the variables are defined, click on the “Data View” tab at the bottom-left corner of the window. You will see the following dialogue box (fig 2.3) with the variable names at the top row. This is the spreadsheet for data entry. Now, you can enter data starting from row 1 for each of the variables. Complete your data entry in this spreadsheet and save the data file at your desired location/folder (save the file as you save your file in MS Word, such as click on File> click on Save as etc.).

If you want to open the data file later, then use the following steps:

Click on File > Open > Data > Select the folder you have saved your SPSS data file > Select the file > Click “Open”

2.2 Data used in this manual

Following data files have been used in this manual as examples. All these datasets are available at the following links. The users can download them for practice. The data files (with hypothetical data) used in this manual include:

- Data_3.sav
- Data_4.sav
- Data_HIV.sav
- Data_repeat_anova_2.sav
- Data_survival_4.sav
- Data_cronb.sav

Links for the data files:

You can download the e-manual and data files from any of the links below.

Link 1: <https://github.com/rubyriders/Learning-SPSS-without-Pain>

Link 2: <https://jmp.sh/F65fcni>

Link 3: <https://drive.google.com/drive/folders/1t9QjNMBV-bI-oyAwEBQEm8auuYv3sP7KW?usp=sharing>

Figure 2.3 Spreadsheet for data entry (SPSS data editor)

SPSS Data Editor window titled "Untitled - SPSS Data Editor".

Menu bar: File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Window, Help.

Toolbar: Standard SPSS icons for file operations, editing, and analysis.

Current view: Data View. Variable View is also visible.

Variable names: v1, v2, v3, v4, v5, v6, v7, v8_a, v8_b, var.

Row numbers: 1, 2, 3, 4, 5, 6, 28, 29, 30.

	v1	v2	v3	v4	v5	v6	v7	v8_a	v8_b	var
1										
2										
3										
4										
5										
6										
28										
29										
30										

Section 3

Data Cleaning and Data Screening

Once data is entered into the SPSS, we need to be sure that there are no errors in the dataset (i.e., there were no errors during data entry). Data cleaning is commonly done by generating frequency distribution tables of all the variables to see the out-of-range values, and by cross tabulations (or by other means) for checking the conditional values. If errors are identified, they need to be corrected. Simultaneously, we also need to check the data if it fulfils the assumptions of the desired statistical test (data screening), e.g., is data normally distributed to do a t-test? The users may skip this section for the time being and go to section 4. Once the users develop some skills in data analysis, they can come back to this section. Use the data file <Data_3.sav> for practice. The codebook of this data file can be seen in the annex (table A.1).

3.1 Checking for out-of-range errors

We can check out-of-range errors by making a frequency distribution of the variable or using the statistics option for minimum and maximum values. For example, you want to see if there are any out-of-range errors in the variable “religion” (note that the variable “religion” has 3 levels/values: 1= Islam; 2= Hindu; 3= Others). To do this, use the following commands:

Analyze > Descriptive statistics > Frequencies > Select the variable “religion” & push it into the “Variable(s)” box > Statistics > Select “minimum” and “maximum” > Continue > Ok

Look at the first table (table 3.1) of SPSS output. If there is any value which is out of the range 1-3 in the dataset, you can see it in the table (3.1) as shown below.

Table 3.1. Statistics

religion		
N	Valid	210
	Missing	0
Minimum		1
Maximum		3

The table shows that the values range from 1 to 3 (minimum 1 and maximum

3), which are within the range of our code numbers. Therefore, there is no out-of-range error in this variable.

3.2 Checking for outliers

Outliers can be checked by constructing the box and plot chart. Outliers are indicated by ID numbers on the chart. Outliers are 1.5 box length distance from the edge (upper or lower) of the box. The extreme values are indicated by “*” (3 box length distance from the edge of the box). To construct the box and plot chart for the variable “systolic blood pressure” (SPSS variable name: sbp), use the following commands.

Analyze > Descriptive statistics > Explore > Select “sbp” and push it into the “Dependent List” box > Ok

You will find the box and plot chart (fig 3.1) along with other outputs (table 3.2). The figure 3.1 shows that there are 3 outliers in “systolic blood pressure” as indicated by the ID numbers (20, 54 & 193).

We can also examine the influence of outliers in the data comparing the 5% trimmed mean (mean of the data after excluding upper 5% and lower 5% of the values) with the mean of the whole dataset. If these two means are close together, there is no influence of outliers in the dataset. Look at the table 3.2. The mean of the systolic blood pressure (BP) is 127.7, while the 5% trimmed mean is 126.5. Since the values are not that different (close to each other), there is no influence of outliers in the data of systolic BP.

Figure 3.1. Box and Plot chart of systolic blood pressure

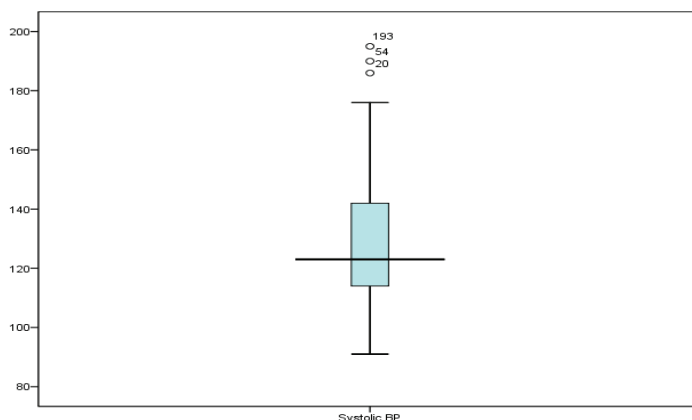


Table 3.2. Descriptives

Systolic BP			Statistics	Std. Error
	Mean		127.73	1.384
	95% Confidence Interval for Mean	Lower Bound	125.00	
		Upper Bound	130.46	
	5% Trimmed Mean		126.58	
	Median		123.00	
	Variance		402.321	
	Std. Deviation		20.058	
	Minimum		91	
	Maximum		195	
	Range		104	
	Interquartile Range		28	
	Skewness		.736	.168
	Kurtosis		.336	.334

3.3 Assessing normality of a dataset

One of the major assumptions for parametric tests is that the dependent quantitative variable is normally distributed. Whether the data has come from normally distributed population or not, can be checked in different ways. Commonest methods of checking normality of a dataset are through:

- Histogram
- Q-Q plot
- Formal statistical test (Kolmogorov Smirnov (K-S) test or Shapiro Wilk test)

This issue is discussed in detail in section 5.

Section 4

Data Analysis: Descriptive Statistics

Descriptive statistics are always used at the beginning of data analysis. The objective of using the descriptive statistics is to organize and summarize data. Commonly used descriptive statistics are frequency distribution, measures of central tendency (mean, median, and mode) and measures of dispersion (range, standard deviation, and variance). Measures of central tendency convey information about the average value of a dataset, while a measure of dispersion provides information about the amount of variation present in the dataset. Other descriptive statistics include quartile and percentile. Use the data file <Data_3.sav> for practice.

4.1 Frequency distribution

Suppose, you want to find the frequency distribution of the variables “sex” and “religion”. To do this, use the following commands:

Analyze > Descriptive Statistics > Frequencies > Select the variables “sex” and “religion” and push them into the "Variable(s)" box > OK (fig 4.1 & 4.2)

Figure 4.1. Commands for frequency distribution of variables

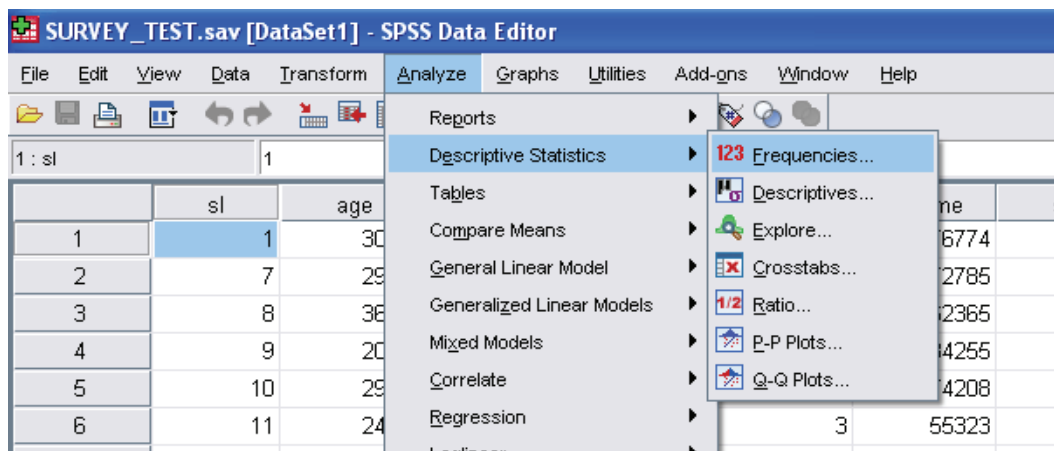
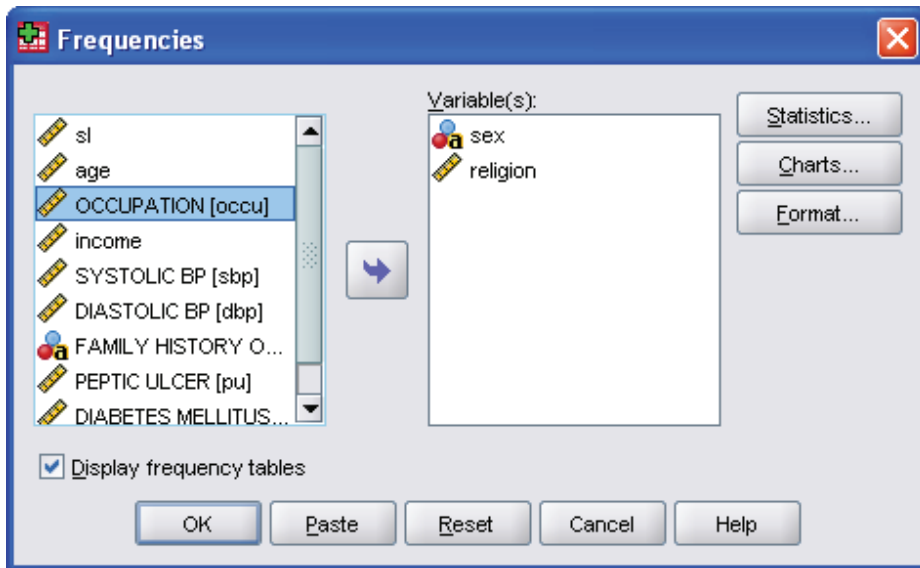


Figure 4.2. Selection of variables for frequency distribution



You will see the following outputs (I have shown only the table of sex) (table 4.1). The table indicates that there are in total 210 subjects, out of which 133 or 63.3% are female and 77 or 36.7% are male. If there is any missing value, the table will show it. In that case, use the “valid percent” instead of “percent” for reporting. For example, table 4.2 shows 4 missing values. You should, therefore, report 130 or 63.1% are female and 76 or 36.9% are male. *Note that the “Percent” and “Valid Percent” will be the same, if there is no missing value.*

Table 4.1. Frequency distribution of sex with no missing value

Sex					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	133	63.3	63.3	63.3
	Male	77	36.7	36.7	100.0
	Total	210	100.0	100.0	

Table 4.2. Frequency distribution of sex with 4 missing values

Sex					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	130	61.9	63.1	63.1
	Male	76	36.2	36.9	100.0
	Total	206	98.1	100.0	
Missing	9	4	1.9		
Total		210	100.0		

4.2 Central tendency and dispersion

We calculate the central tendency and dispersion for the quantitative variables. Suppose, you want to find the mean, median, mode, standard deviation (SD), variance, standard error (SE), skewness, kurtosis, quartile, percentile (e.g., 30th and 40th percentile), minimum and maximum values of the variable “age” of the study subjects. All these statistics can be obtained in several ways. However, using the following commands is the easiest way to get them together (fig 4.3-4.5).

Analyze > Descriptive statistics > Frequency > Select the variable “age” and push it into the "Variable(s)" box > Statistics > Select all the descriptive measures you desire (mean, median, mode, SD, SE, quartile, skewness, kurtosis) > Select "Percentiles" > Write “30” in the box > Add > Write “40” in the box > Add > Continue > OK

Figure 4.3. Commands for obtaining central tendency and dispersion

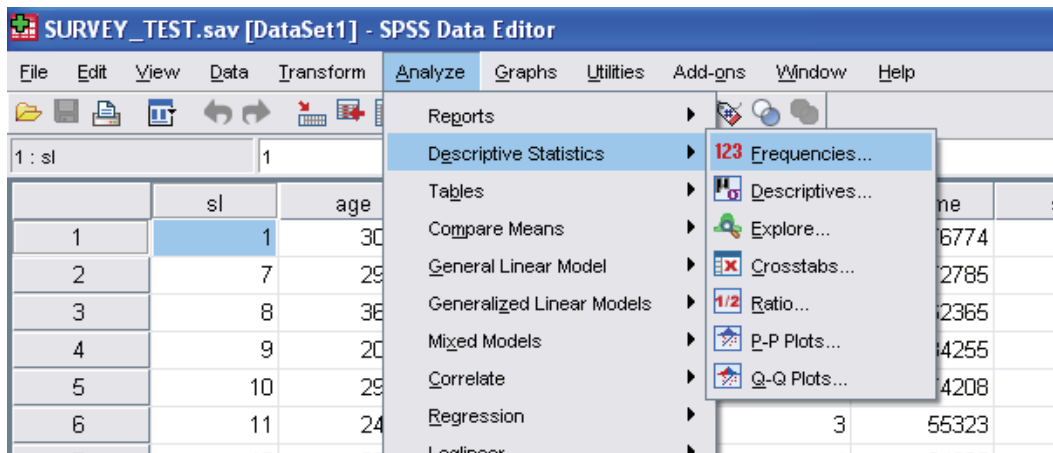


Figure 4.4. Selection of variable(s)

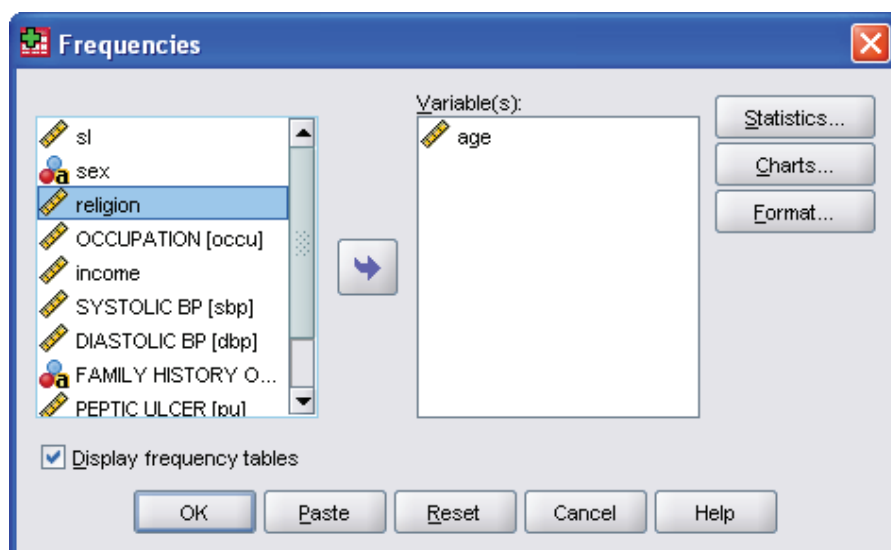
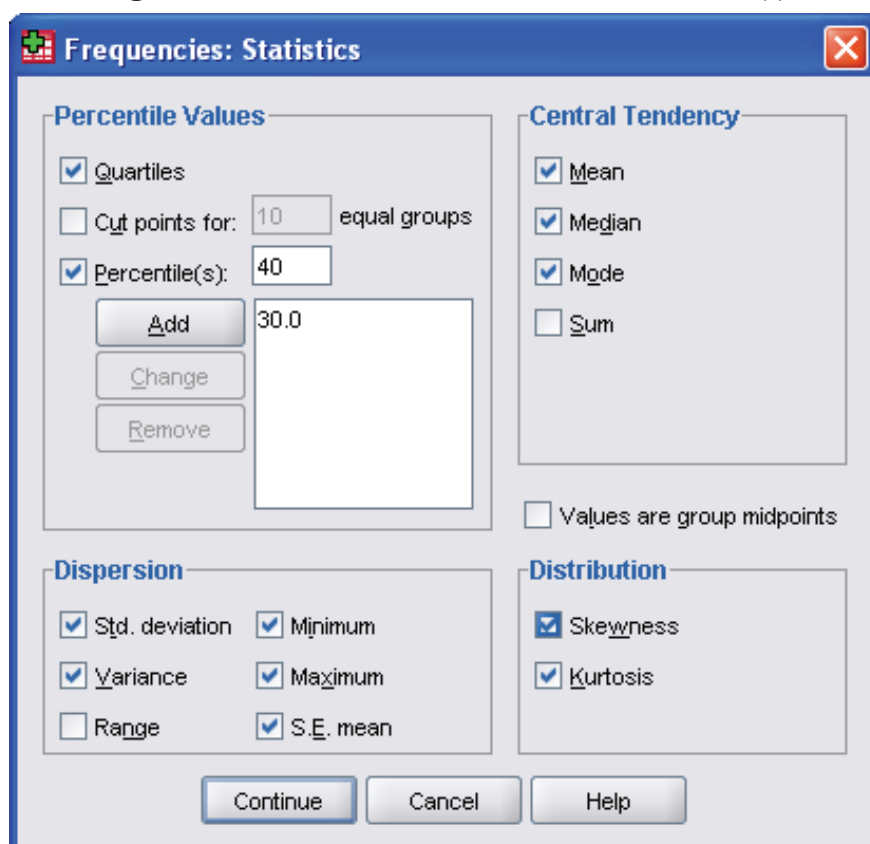


Figure 4.5. Selection of statistics for the variable(s)



4.2.1 Outputs:

The SPSS will produce the following output (table 4.3).

Table 4.3. Descriptive statistics of age

AGE		
N	Valid	210
	Missing	0
Mean		26.5143
Std. Error of Mean		.51689
Median		27.0000
Mode		26.00
Std. Deviation		7.49049
Variance		56.10745
Skewness		-.092
Std. Error of Skewness		.168
Kurtosis		-.288
Std. Error of Kurtosis		.334
Minimum		6.00
Maximum		45.00
Percentiles	25	21.0000
	30	22.3000
	40	25.0000
	50	27.0000
	75	32.0000

4.2.2 Interpretation:

We can see all the descriptive statistics (central tendency and dispersion) that we have selected for the variable “age” including the statistics for Skewness and Kurtosis in table 4.3. Hope, you understand the mean (average), median (middle value of the data set), mode (most frequently occurring value), SD (average difference of individual observation from the mean), variance (square of SD) and SE of the mean. As presented in table 4.3, the mean age is 26.5 and SD is 7.49 years. Let me discuss the other statistics provided in table 4.3, especially the skewness, kurtosis, quartile and percentile.

Skewness and Kurtosis: These two statistics are used to judge whether the data have come from a normally distributed population or not. In table 4.3, we can see the statistics for Skewness (-.092) and Kurtosis (-.288). Skewness indicates the spreadness of the distribution. Skewness “>0” indicates data is skewed to the right; skewness “<0” indicates data skewed to the left, while skewness “~0” indicates data is symmetrical (normally distributed). The acceptable range for normality of

a data set is skewness lying between “-1” and “+1”. However, normality should not be judged based on skewness alone. We need to consider the statistics for kurtosis as well. Kurtosis indicates “peakness” or “flatness” of the distribution. Like skewness, the acceptable range of kurtosis for a normal distribution is between “-1” and “+1”. Data for “age” has skewness -.092 and kurtosis -.288, which are within the normal limits of a normal distribution. We may, therefore, consider that the variable “age” in the population may be normally distributed.

Quartile and Percentile: When a dataset is divided into four equal parts after arranging into ascending order, each part is called a quartile. It is expressed as Q1 (first quartile or 25th percentile), Q2 (second quartile or median or 50th percentile) and Q3 (third quartile or 75th percentile). On the other hand, when data is divided into 100 equal parts (after ordered array), each part is called a percentile. We can see in table 4.3 that, Percentile 25 (means Q1), Percentile 50 (Q2) and Percentile 75 (Q3) for age are 21, 27 and 32 years, respectively. Q1 or the first quartile is 21 years, means that 25% of the study subjects’ age is less than or equal to 21 years. On the other hand, 30th percentile (P_{30}) is 22.3 years, which means that 30% of the study subjects’ age is less than or equal to 22.3 years. Hope, you can now interpret the P_{40} .

4.3 Alternative method of getting measures of central tendency and dispersion

If you want to get all the descriptive statistics (central tendency and dispersion) and charts (such as histogram, stem and leaf, and box and plot charts) of the variable “age”, use the following commands:

Analyze > Descriptive statistics > Explore > Select the variable “age” and push it into the "Dependent List" box > Plots > Select "Stem and leaf" and “Histogram” > Continue > OK

4.3.1 Outputs:

The outputs are shown in table 4.4 and figs 4.6 to 4.9.

Table 4.4. Descriptive statistics of age

			Statistics	Std. Error
AGE	Mean		26.5143	.51689
	95% Confidence Interval for Mean	Lower Bound	25.4953	
		Upper Bound	27.5333	
	5% Trimmed Mean		26.5608	
	Median		27.0000	
	Variance		56.107	
	Std. Deviation		7.49049	
	Minimum		6.00	
	Maximum		45.00	
	Range		39.00	
	Interquartile Range		11.0000	
	Skewness		-.092	.168
	Kurtosis		-.288	.334

Figure 4.6. Histogram of age

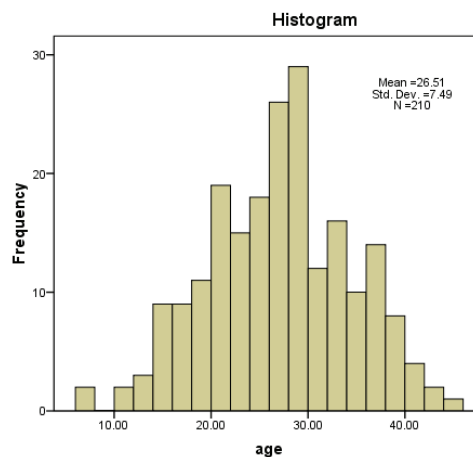


Figure 4.7. Stem and leaf chart of age

age Stem-and-Leaf Plot

Frequency Stem & Leaf

```

2.00  0 . 66
10.00  1 . 0122344444
24.00  1 . 55556666677778888889999
44.00  2 . 0000000000000111112222222333333444444444
63.00  2 . 5555555666666666666666677777777888888888889999999999999999
34.00  3 . 00001111111112222223333333333444444
26.00  3 . 55556666666667777788888999
6.00   4 . 001133
1.00   4 . 5

```

Stem width: 10.00

Each leaf: 1 case(s)

Figure 4.8. Box and plot chart of age

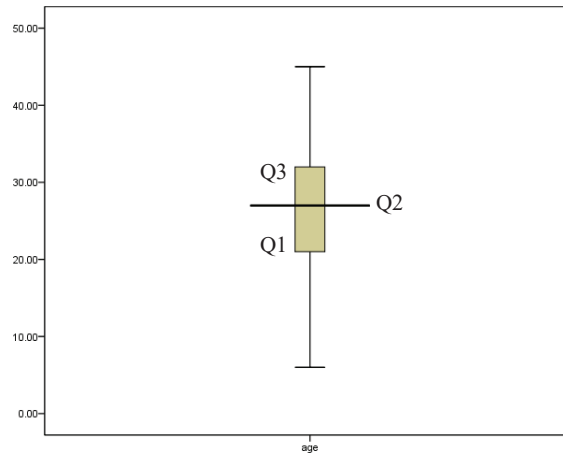
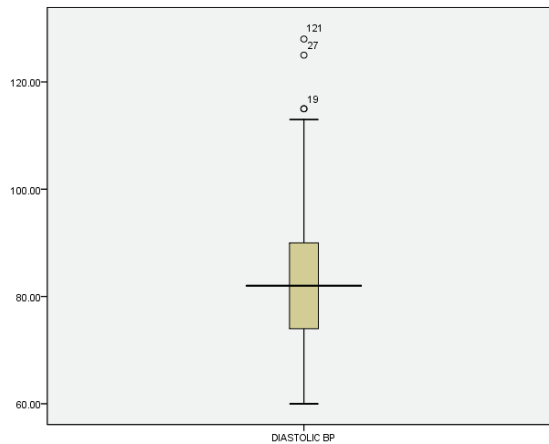


Figure 4.9. Box and plot chart of diastolic BP



4.3.2 Interpretation:

Before we understand the graphs, let us see the descriptive statistics provided in table 4.4. We can see that the SPSS has provided mean and 5% trimmed mean of age. Five percent trimmed mean is the mean after discarding 5% of the upper and 5% of the lower values of age. The extent of the effect of outliers can be checked by comparing the mean with the 5% trimmed mean. If they are close together (as we see in table 4.4; mean= 26.51 and 5% trimmed mean= 26.56), there is no significant influence of the outliers (or there are no outliers) on age in the dataset. If they are very different, it means that the outliers have significant influence on the mean value, and suggests for checking the outliers and extreme values in the dataset. The

table 4.4 also shows the 95% Confidence Interval (CI) for the mean of “age”, which is 25.49-27.53. The 95% CI for the mean indicates that we are 95% confident/sure that the mean age of the population lies between 25.49 and 27.53 years.

The SPSS has provided several graphs (figs 4.6 to 4.9), such as histogram, stem and leaf, and box and plot charts. Histogram gives us information about – a) distribution of the dataset (whether symmetrical or not); b) concentration of values; and c) range of values. Looking at the histogram (fig 4.6), it seems that the data is more or less symmetrical. This indicates that age may be normally (approximately) distributed in the population.

Stem and leaf chart (fig 4.7) provides information similar to a histogram, but retains the actual information on data values. Looking at the stem and leaf chart, we can have an idea about the distribution of the dataset (whether symmetrical or not). Data displayed in figure 4.7 shows that the data is more or less symmetrical. *Stem and leaf charts are suitable for small datasets.*

The box and plot chart (fig 4.8) provides information about the distribution of a dataset. It also provides summary statistics of a variable, like Q1 (first quartile or 25th percentile), median (second quartile or Q2) and Q3 (third quartile or 75th percentile) as well as information about outliers/extreme values. The lower boundary of the box indicates the value for Q1, while the upper boundary indicates the value for Q3. The median is represented by the horizontal line within the box. The smallest and largest values are indicated by the horizontal lines of the whiskers.

In the box and plot chart, presence of outliers is indicated by the ID number and circle, while the presence of extreme values is indicated by “*”. Outliers are the values lying between 1.5 and <3 box length distance from the edge (upper or lower) of the box. On the other hand, the extreme values are 3 or more box length distance from the upper or lower edge of the box. Fig 4.8 shows that there is no outlier in the data for age. I have provided another box and plot chart, which is for the variable “diastolic BP” (fig 4.9). Figure 4.9 shows that there are 3 outliers (ID no. 19, 27 and 121) in the data of diastolic BP, but does not have any extreme value.

4.4 Descriptive statistics and histogram disaggregated by Sex

If you want to get the outputs (measures of central tendency and dispersion of age) by sex (males and females separately), use the following commands. The SPSS will produce the outputs separately for males and females (table 4.5). *Note that*

there are other ways of doing this.

Analyze > Descriptive statistics > Explore > Select “age” and push it into the "Dependent List" box > Select “sex” and push it into the "Factor List" box > Plots > Deselect "Stem and leaf”, and select “Histogram” > Continue > OK

Only the table with descriptive statistics is provided below (table 4.5).

Table 4.5. Descriptive statistics of age by sex

Descriptives					
Age	Sex		Statistics		Std. Error
	Female	Mean		26.8872	.58981
		95% Confidence Interval for Mean	Lower Bound	25.7205	
			Upper Bound	28.0539	
		5% Trimmed Mean		26.8413	
		Median		27.0000	
		Variance		46.267	
		Std. Deviation		6.80202	
		Minimum		10.00	
		Maximum		45.00	
		Range		35.00	
		Interquartile Range		9.50	
		Skewness		.074	.210
		Kurtosis		-.212	.417
	Male	Mean		25.8701	.97549
		95% Confidence Interval for Mean	Lower Bound	23.9273	
			Upper Bound	27.8130	
		5% Trimmed Mean		26.0144	
		Median		26.0000	
		Variance		73.272	
		Std. Deviation		8.55993	
		Minimum		6.00	
		Maximum		41.00	
		Range		35.00	
		Interquartile Range		13.00	
		Skewness		-.153	.274
Kurtosis		-.606	.541		

4.5 Checking for outliers

Outliers and extreme values can be checked looking at the box and plot chart, as discussed earlier. We can also check the presence of outliers using the following commands. For example, we want to understand if there are any outliers present in the variable “age”.

Analyze > Descriptive Statistics > Explore > Select the variable “age” and

push it into the "Dependent List" box > Select "ID_no" (ID no.) and push it into the "Label cases by" box > Select "Statistics" under "Display" > Statistics > Select "Outliers" > Continue > OK

The SPSS will provide 5 upper and 5 lower values with the ID (serial no.) numbers, as shown in table 4.6.

Table 4.6. Extreme values of the variable age

Extreme Values					
			Case Number	id no.	Value
age	Highest	1	210	210	45.00
		2	209	209	43.00
		3	208	208	43.00
		4	207	207	41.00
		5	206	206	41.00
	Lowest	1	3	3	10.00
		2	1	1	10.00
		3	4	4	11.00
		4	2	2	11.00
		5	6	6	12.00
a. Only a partial list of cases with the value 12.00 are shown in the table of lower extremes.					

Section 5

Checking Data for Normality

It is important to know the nature of distribution of a continuous random variable before using statistical methods for hypothesis testing. To use parametric methods for testing hypotheses (e.g., t-test, ANOVA), one of the important assumptions is that the data of the dependent variable are normally distributed. It is, therefore, necessary to check whether the data have come from a normally distributed population or not, before we use the parametric methods. Use the data file <Data_3.sav> for practice.

5.1 How to understand that the data have come from a normally distributed population

This is an important assumption for doing a parametric test. Whether the data have come from a normally distributed population or not, can be assessed in three different ways. They are by:

- a) Graphs, such as histogram and Q-Q chart;
- b) Descriptive statistics, using skewness and kurtosis; and
- c) Formal statistical tests, such as 1-sample Kolmogorov Smirnov (K-S) test and Shapiro Wilk test.

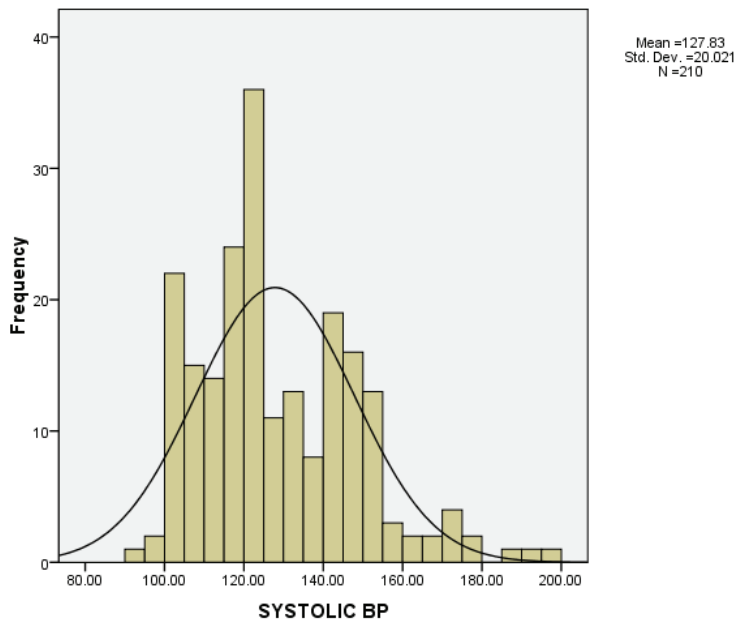
Now, let us see how to get the histogram and Q-Q chart, and do the formal statistical tests (K-S test and Shapiro Wilk test).

Suppose, we want to know whether the variable “systolic BP (SPSS variable name: sbp)” is normally distributed in the population or not. We shall first construct the histogram and Q-Q chart. To construct a histogram for systolic BP, use the following commands:

Graphs > Legacy dialogs > Histogram > Select the variable “sbp” and push it into the “Variable” box > Select “Display normal curve” clicking at the box > Ok

The SPSS will produce a histogram of systolic BP, as shown in fig 5.1.

Figure 5.1. Histogram of systolic BP

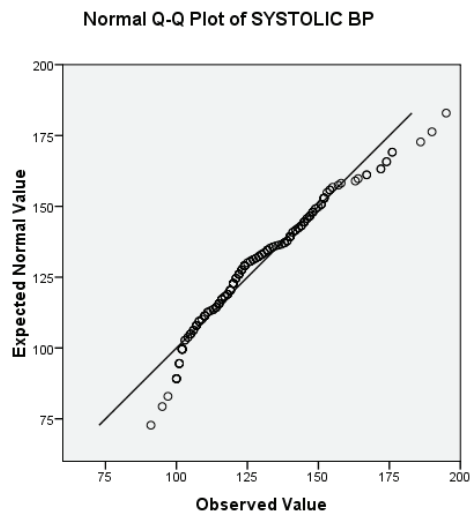


To get the Q-Q plot for systolic BP, use the following commands:

Analyze > Descriptive statistics > Q-Q Plots > Select the variable “sbp” and push it into the “Variables” box > For Test Distribution select “Normal” (usually remains as default) > Ok

The computer will produce a Q-Q plot of systolic BP, as shown in figure 5.2.

Figure 5.2. Q-Q plot of Systolic BP



To do the formal statistical tests (K-S test and Shapiro Wilk test) to understand the normality of the data, use the following commands.

Analyze > Descriptive Statistics > Explore > Select the variable “sbp” and push it into the "Dependent List" box > Plots > Deselect "Stem-and-leaf" and select “Histogram” > Select “Normality plots with test” > Continue > OK

Note that these commands will also produce the histogram and Q-Q plot. You may not need to develop histogram and Q-Q plot separately as mentioned earlier.

5.1.1 Outputs:

You will get the following tables (table 5.1 and 5.2) along with the histogram, Q-Q plot and box and plot chart. The histogram, Q-Q plot and box and plot chart, generated by the commands, have been omitted to avoid repetition.

Table 5.1 Descriptive statistics of Systolic BP

Descriptives				
			Statistics	Std. Error
SYSTOLIC BP	Mean		127.8333	1.38161
	95% Confidence Interval for Mean	Lower Bound	125.1097	
		Upper Bound	130.5570	
	5% Trimmed Mean		126.6878	
	Median		123.0000	
	Variance		400.857	
	Std. Deviation		20.02142	
	Minimum		91.00	
	Maximum		195.00	
	Range		104.00	
	Interquartile Range		28.00	
	Skewness		.728	.168
	Kurtosis		.343	.334

Table 5.2. Statistical tests for normality (of Systolic BP)

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SYSTOLIC BP	.119	210	.000	.956	210	.000
a. Lilliefors Significance Correction						

5.1.2 Interpretation:

The SPSS has generated the histogram and Q-Q plot for “systolic BP” (fig 5.1 and 5.2) and tables 5.1 and 5.2. While getting the specific statistical tests (KS test and

Shapiro Wilk test) to check the normality of a dataset, the SPSS automatically provides the descriptive statistics of the variable (systolic BP) (table 5.1). I have already discussed the measures of skewness and kurtosis to assess the normality of a dataset earlier (section 4).

Histogram (fig 5.1) provides an impression about the distribution of the dataset (whether symmetrical or not). If we look at the histogram of systolic BP, it seems that the data is slightly skewed to the right (i.e., distribution is not symmetrical).

The Q-Q plot (fig 5.2) also provides information on whether data have come from a normally distributed population or not. The Q-Q plot compares the distribution of data with the standardized theoretical distribution from a specified family of distribution (in this case normal distribution). If data are normally distributed, all the points (dots) lie on the straight line. Note that our interest is in the central portion of the line. Deviation from the central portion of the line means non-normality. Deviations at the ends of the plot indicate the existence of outliers. We can see (in fig 5.2) that there is a slight deviation at the central portion as well as at the ends. This may indicate that the data may not have come from a normally distributed population.

The specific tests (objective tests) to assess if the data have come from a normally distributed population are the K-S (Kolmogorov-Smirnov) test and Shapiro Wilk test. The results of these two tests are provided in table 5.2.

Look at the Sig (significance) column of table 5.2. Here, Sig (indicates the p-value) is 0.000 for both the tests. A p-value of <0.05 indicates that the data have not come from normally distributed population. In our example, the p-value is 0.000 for both the tests, which is <0.05 . This means that the data of systolic BP have not come from a normally distributed population. The null hypothesis here is “data have come from a normally distributed population”. The alternative hypothesis is “data have not come from a normally distributed population”. We will reject the null hypothesis, since the p-value is <0.05 .

Note that the K-S test is very sensitive to sample size. The K-S test may be significant for slight deviations of a large sample data ($n > 100$). Similarly, the likelihood of getting a p-value <0.05 for a small sample ($n < 20$, for example) is low. Therefore, the rules of thumb for normality checking are:

- 1) Sample size < 30 : Assume non-normal;
- 2) Moderate sample size (30-100): If the formal test is significant ($p < 0.05$), consider non-normal distribution, otherwise check by other methods, e.g.,

histogram, Q-Q plot, etc.; and

- 3) Large sample size ($n > 100$): If the formal test is not significant ($p > 0.05$), accept normality, otherwise check with other methods.

However, for practical purposes, just look at the histogram. If it seems that the distribution is approximately symmetrical, consider that the data have come from a normally distributed population.

Section 6

Data Management

While analyzing data, you may require to make class intervals, classify a group of people with specific characteristic using a cutoff value (e.g., you may want to classify people who have hypertension using a cutoff value of either systolic or diastolic BP), and recode data for other specific purposes. In this section, I shall discuss data manipulations that are commonly needed during data analysis. For example,

- Recoding of data
- Making class intervals
- Combine data to form an additional variable
- Data transformation
- Calculation of total score
- Extraction of time
- Selection of a subgroup for data analysis

Use the data file <**Data_3.sav**> for practice.

6.1 Recoding of data

For example, you have the variable “sex” coded as “m” and “f”. You want to replace the existing code “m” by 1 and “f” by 2. There are two options for recoding data:

- a) Recoding into same variable; and
- b) Recoding into different variable.

My suggestion would be to use “recoding into different variable” option all the times. This will keep the original data of the variable as it is.

6.1.1 Recoding into same variable:

Note that if you recode data into same variable, the original data/coding would be lost. To do this follow the following commands:

Transform > Recode into Same Variables > Select “sex” and push it into the “Variables” box > Click on “Old and new values” > Select “Value” (usually

default) under “Old Value” > Type **m** in the box below > type 1 in the “Value” area under “New Value” > Click “Add” > Type **f** in the value area under “Old Value” > type 2 in the “Value” area under “New Value” > Click Add > Continue > Ok (Fig 6.1 and 6.2)

Figure 6.1

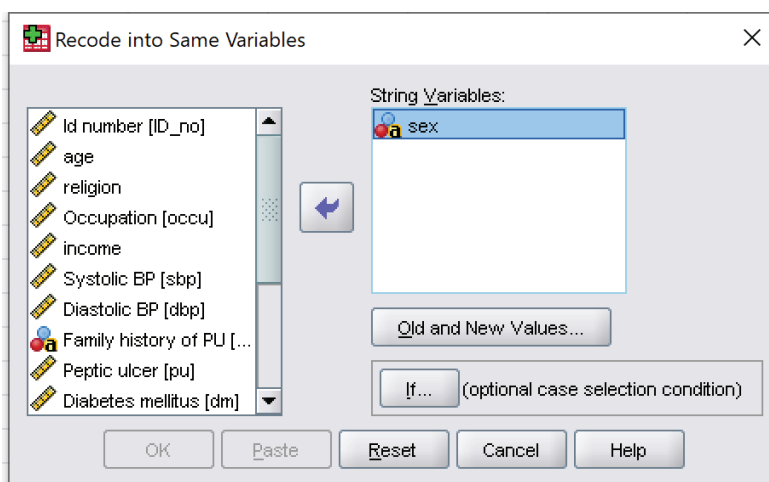
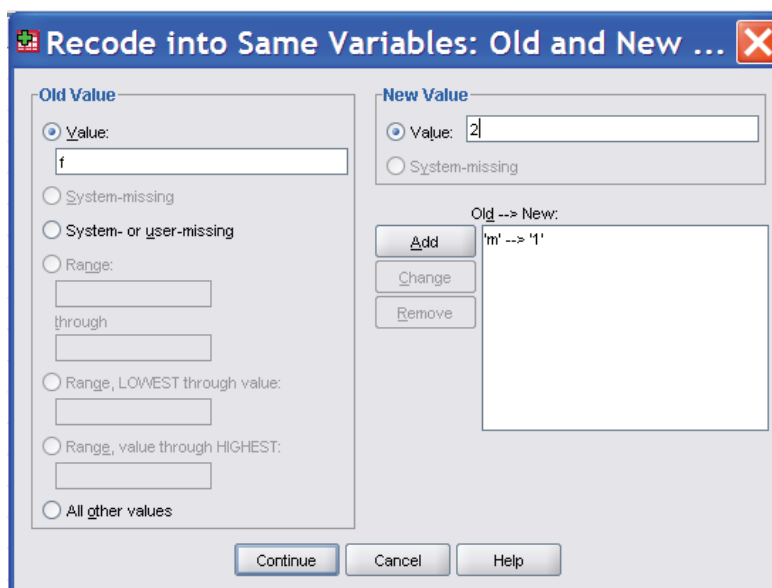


Figure 6.2



Check the data file in the “data view” option. You will notice that all the “m” has been replaced by 1 and “f” by 2. Now, go to the “variable view” of the data file to replace the codes. Click in the “Values” box against the variable “sex”. Replace the codes as 1 is “Male” and 2 is “Female”.

6.1.2 Recoding into different variable:

This option of recoding requires formation of a new variable. The original variable and data will remain intact. To do this, follow the following commands:

Transform > Recode into Different Variables > Select “sex” and push it into “Input Variable –Output Variables” box > Type “**sex1**” in the “Name” box and type “**Gender**” in the “Label” box under “Output Variable” > Click “Change” > Click “Old and New Values” > Type **m** in the “Value” box under the “Old Value” > Type 1 in the “Value” box under the “New Value” > Click “Add” > Type **f** in the “Value” box under the “Old Value” > Type 2 in the “Value” box under the “New Value” > Click “Add” > Continue > Ok (Fig 6.3 and 6.4)

Here, the new variable generated is “sex1” (do not give any space between sex and 1, while typing the variable name in the Name box). Follow the rules of writing variable names as mentioned in section 2 (2.1.1).

Figure 6.3

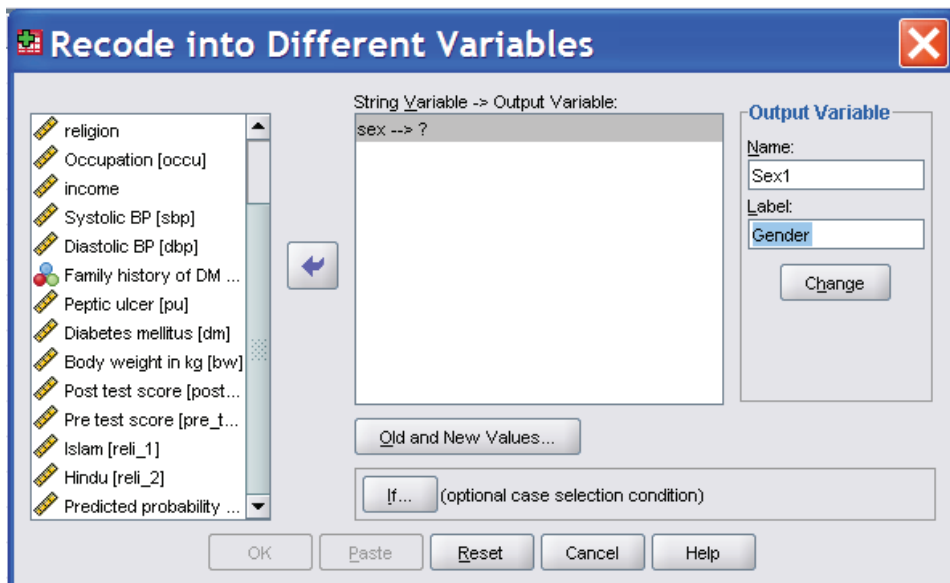
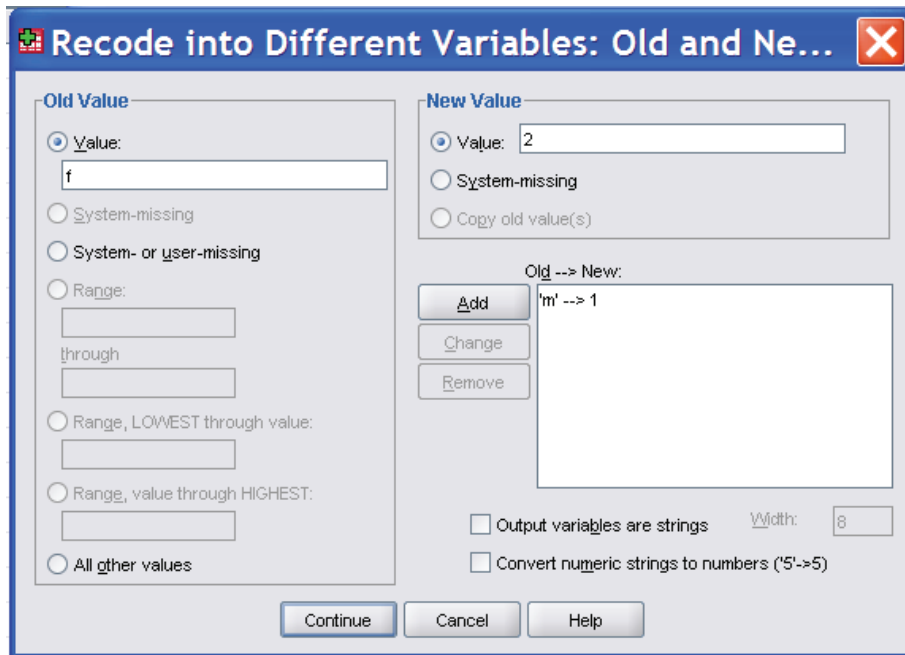


Figure 6.4



Check on the “variable view” option of the data file. You will notice that SPSS has generated a new variable “sex1” (the last variable both in the variable view and data view options). Like before, in the variable view, define the value labels of the new variable sex1 as 1 is “male” and 2 is “female”.

6.2 Making class intervals (categories)

Suppose, you want to categorize the variable “age” into the following categories:

- ≤ 20 years (to be coded as 1),
- 21-30 years (to be coded as 2),
- 31-40 years (to be coded as 3), and
- > 40 years (to be coded as 4)

We shall use the option “Recode into Different Variable” for this exercise. As before, we need to generate a new variable. Suppose, the new variable we want to generate is “age1”. *I would suggest the users to use the option “Recode into Different Variable” all the times.* If you use the option “Recode into Same Variable”, you will loss the original data that cannot be recovered once the data file is saved. To do this, use the following commands:

Transform > Recode into Different Variable > Select “age” and push it into the “Input Variable –Output Variables” box > Type “**age1**” in the “Name” box and type “**age group**” in the “Label” box under “Output Variable” > Click “Change” > Click on “Old and New Values” > Select “System-missing” under “Old value” > Select “System-missing” under “New Value” > Click “Add” > Select “Range, LOWEST through value” and type “20” in the box below > Select “Value” under “New Value” and type “1” > Add > Now select “Range” under “Old Value” > Type “21” in the upper box and “30” in the lower box > Select “Value” under “New Value” and type “2” > Add > Again, type “31” in the upper box and “40” in the lower box > Select “Value” under “New Value” and type “3” > Add > Select “All other values” under “Old Value” > Select “Value” under “New Value” and type “4” > Add > Continue > Ok (fig 6.5 and 6.6)

Figure 6.5

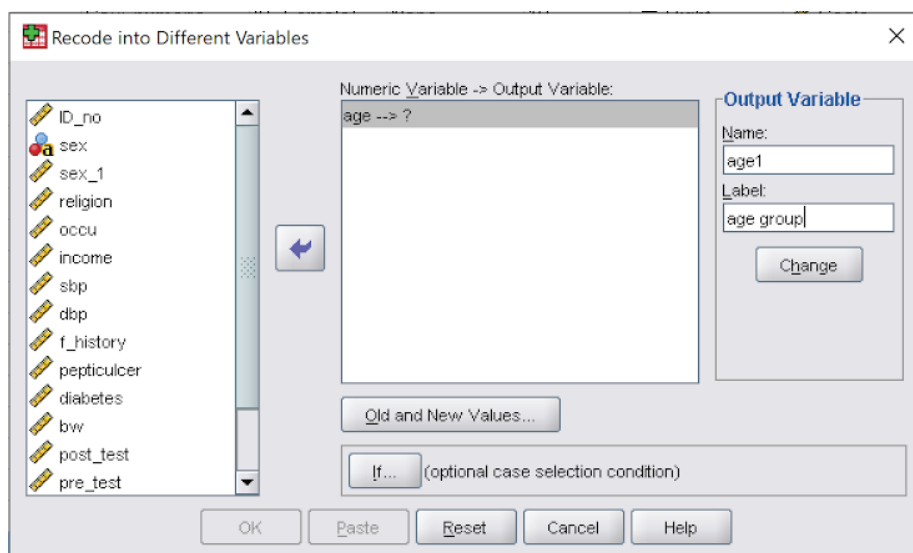
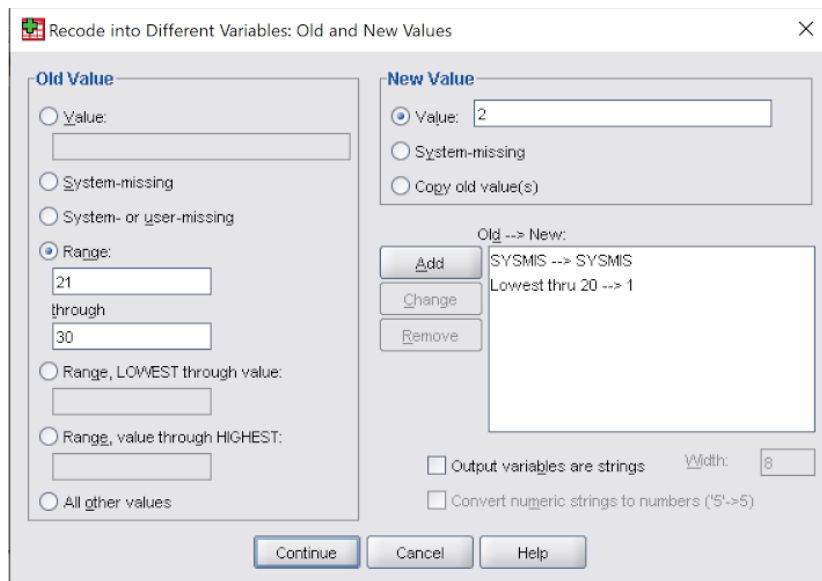


Figure 6.6



Go to the data file in the “data view” option and then to the “variable view” option. You will notice that SPSS has generated a new variable “age1” (the last variable both in the data view and variable view options). Like before, in the “variable view” option, define the value labels of the variable “age1” as 1 is “≤ 20 years”, 2 is “21-30 years”, 3 is “31-40 years” and 4 is “>40 years”.

Using transformation, you can also classify people who have hypertension and who do not have hypertension (for example). To do this, you shall have to use a cutoff point to define hypertension. For example, we have collected data on diastolic BP (SPSS variable name is “dbp”). We want to classify those as “hypertensive” if the diastolic BP is >90 mmHg. Now, recode the variable diastolic BP into a new variable (say, d_hyper) using “Recode into Different Variables” option as ≤ 90 (normal BP) and > 90 (as hypertensive). Hope, you can do it now. If you cannot, use the following commands:

Transform > Recode into Different Variable > Select “dbp” and push it into the “Input Variable –Output Variables” box > Type “**d_hyper**” in the “Name” box and type “**diastolic hypertension**” in the “Label” box under “Output Variable” > Click “Change” > Click on “Old and New Values” > Select “System-missing” under “Old value” > Select “System-missing” under “New Value” > Add > Select “Range, LOWEST through value” under “Old value” and type “90” in

the box below > Select “Value” under “New Value” and type “1” > Add > Select “All other values” under “Old value” > Select “Value” under “New Value” and type “2” > Add > Continue > Ok

This will create a new variable “d_hyper” with code numbers 1 and 2 (the last variable both in the variable and data view options). Code 1 indicates the persons without hypertension (diastolic BP ≤ 90) and code 2 indicates the persons with hypertension (diastolic BP >90). As done before, in the “variable view” option, define the value labels of the new variable “d_hyper” as 1 is “Do not have hypertension” and 2 is “Have hypertension”.

6.3 Combine data into a new variable

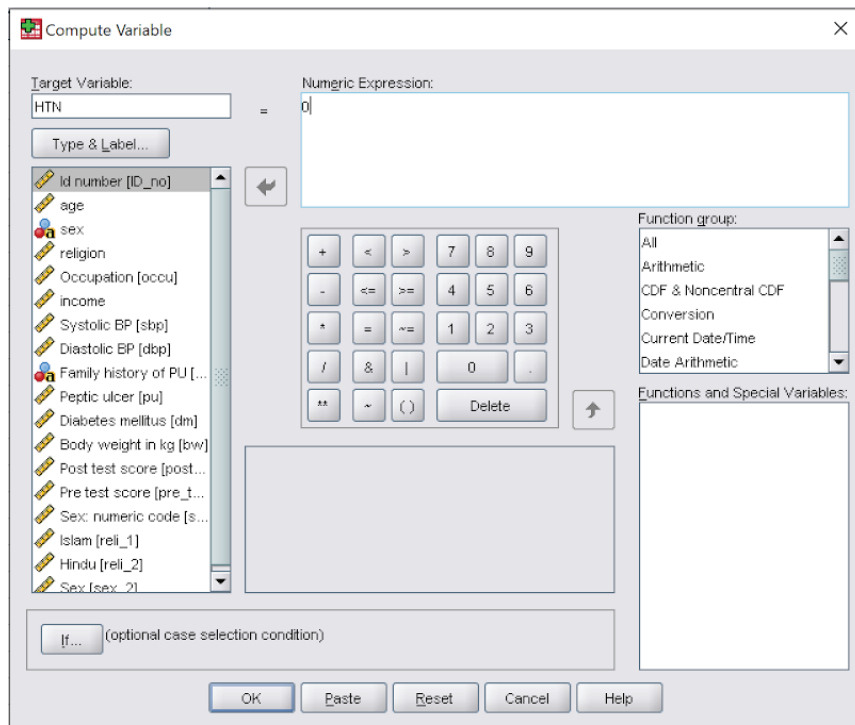
Sometimes, the cutoff point of a measurement (e.g., hemoglobin, blood pressure, etc.) for defining a condition (e.g., anemia, hypertension) may vary according to gender. In such a situation, a single cutoff point for defining a condition is not appropriate.

For example, we have collected data on diastolic BP (SPSS variable name is dbp) both for males and females. We have defined hypertension as diastolic BP >85 mmHg if it is a female, and diastolic BP >90 mmHg if it is a male. Now, how to classify those who have hypertension considering the gender?

To do this, first, we shall create a new variable, say “HTN” for which all the values would be 0 (zero). Use the following commands to do this:

Transform > Compute Variable > Type “HTN” in “Target Variable” box > Click in the box under “Numeric Expression” > Click “0” (zero) from the number pad or keyboard > Ok (fig 6.7)

Figure 6.7



This will generate the new variable “HTN” with all the values 0 (you can check it in the “data view” option; the last variable). Now use the following commands:

Transform > Compute Variable > Click in the box under “Numeric Expression” > Delete 0 > Click on 1 > Click If (optional case selection condition) > Select “Include if case satisfies condition” > Select “dbp” and push it into the box > Click “greater than sign (>)” then write “90” (always use the number pad) > Click “&” on the “number pad” > Select “sex_1” and push it into the box > Click on “=” and then “1” (note: 1 is the code no. for male) > Continue > Ok > (SPSS will provide the message “Change existing variable” > Click on “Yes” (fig 6.8 and 6.9)

Again,

Transform > Compute Variable > Click “If (optional case selection condition)” > Delete “90” and write “85” (for dbp) and delete “1” and click “0” (for sex_1, since 0 is the code for female) > Continue > Ok > (SPSS will give you the message “Change existing variable” > Click “Yes”

Go to the “data view” option of the data file. You will notice that the new variable “HTN” (the last variable both in the “data view” and “variable view” options) has values either “0” or “1”. “0” indicates “no hypertension”, while “1” indicates “have hypertension”. Like before, go to the “variable view” option and define the value labels of the variable “HTN” as “0” is “No hypertension” and “1” is “Have hypertension”.

Figure 6.8

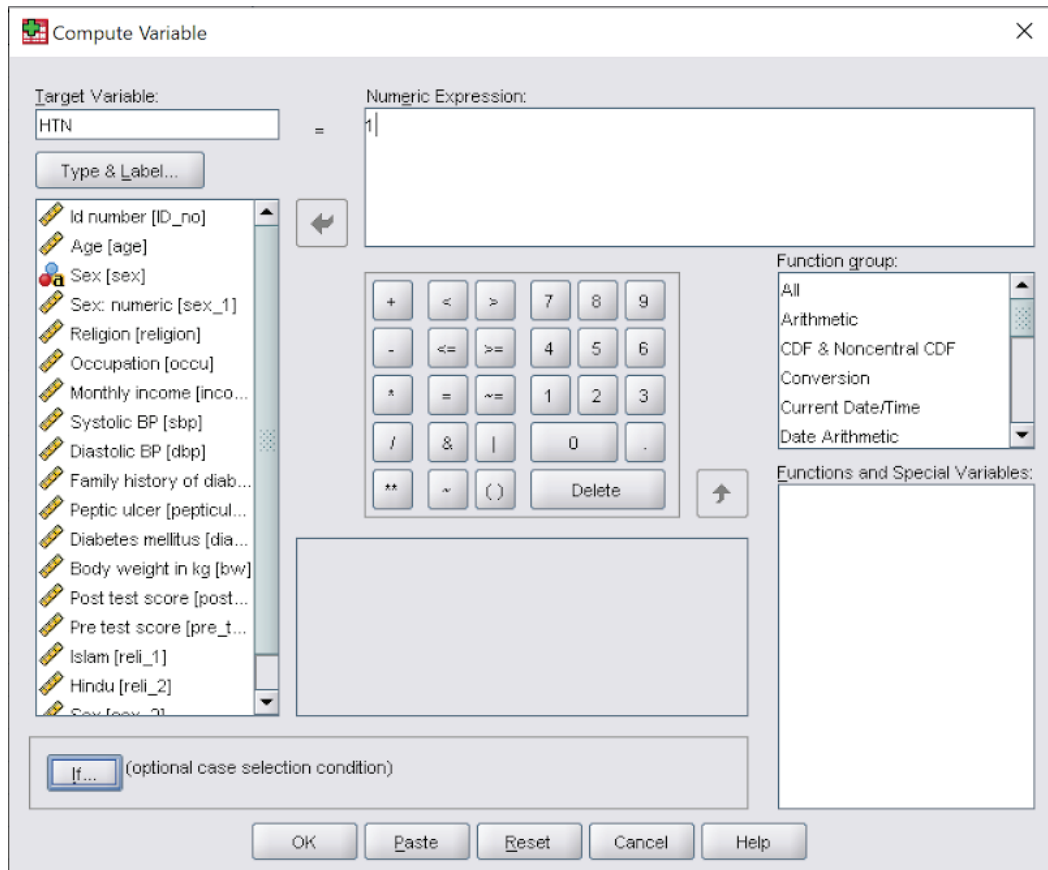
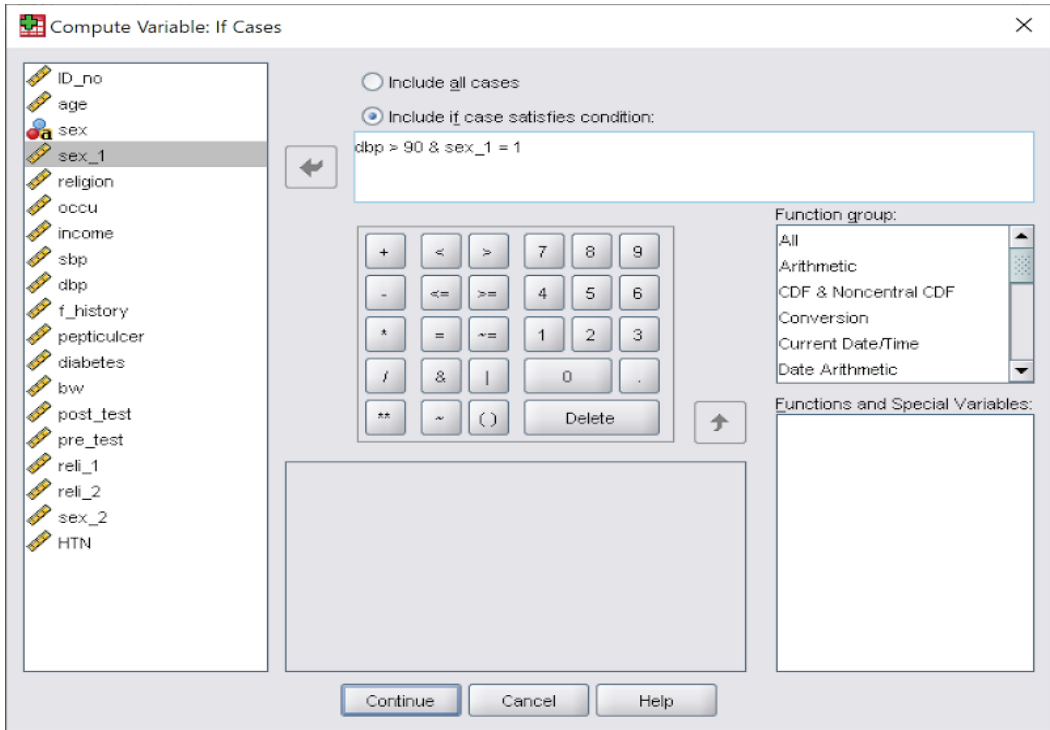


Figure 6.9



6.4 Data transformation

In many situations, the data that have been collected for a study are not normally distributed. Since parametric methods (in general) for testing hypotheses are better than the non-parametric methods, data transformations are occasionally needed to make the distribution normal and to meet the assumptions for a parametric test. Depending on the shape of the data distribution, there are several transformation options. Following table (table 6.1) shows some of the options for data transformation.

Table 6.1. Data transformation options

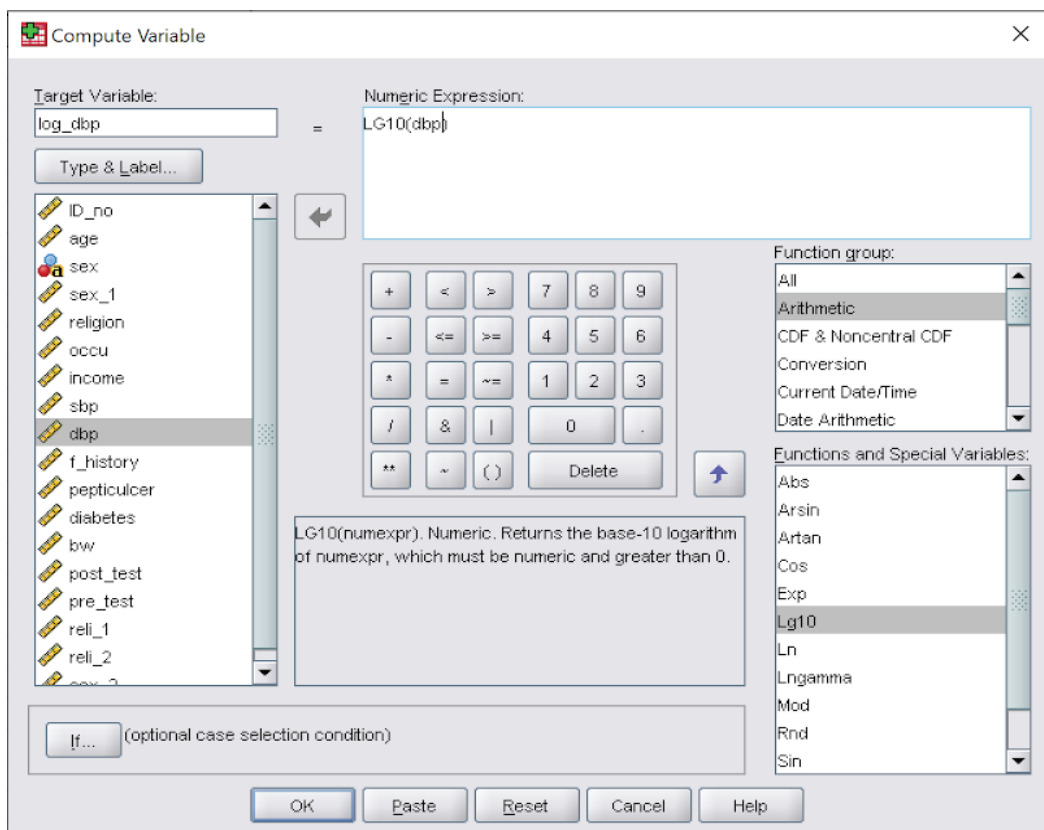
Method	Good for	Bad for
Log	Right skewed data	Zero values and negative values
Square root	Right skewed data	Negative values
Square	Left skewed data	Negative values
Reciprocal	Making small values bigger and big values smaller	Zero values and negative values

Commonly used method of data transformation is Log transformation. Let us see how to get the Log transformation of data. Suppose, you want to transform diastolic BP (variable name is dbp) into Log of diastolic BP. Use the following commands:

Transform > Compute Variable > Type “log_dbp” under “Target Variable” > Click on “Arithmetic” in the “Function Group” box > In “Functions and special variables” box select “Lg10” > Click on the “up arrow” (left side of the box. You will see LG10(?) appears in the Numeric Expression box) > Select “dbp” in the “Type and Label” box > push it into the “Numeric Expression” box > Ok (fig 6.10)

This will create a variable “log_dbp”, with the values “log of diastolic BP” (the last variable). Similarly, you can transform your data into square root using the option “sqrt” in the “Functions and special variables” box.

Figure 6.10



6.5 Calculation of total score

Suppose, you have conducted a study to assess the knowledge of the secondary school children on how HIV is transmitted. To assess their knowledge, you have set the following questions (**data file: HIV.sav**).

HIV is transmitted through:

- | | | |
|--|--------|-------|
| 1) Sexual contact (variable name: k1) | 1. Yes | 2. No |
| 2) Transfusion of unscreened blood (variable name: k2) | 1. Yes | 2. No |
| 3) Sharing of syringe (variable name: k3) | 1. Yes | 2. No |
| 4) Accidental needle stick injury (variable name: k4) | 1. Yes | 2. No |

Note: All the correct answers are coded as 1.

To calculate the total knowledge score, use the following commands:

Transform > Count values within cases > Write “t_know” in the box under “Target variable” > Write “total knowledge on HIV” in the box under “Target level” > Select “k1, k2, k3 and k4” and push them into the “Variables” box > Click “Define values” > Select “Value” and write “1” (since 1 is the correct answer) in the box below > Click “Add” > Continue > OK (fig 6.11 and 6.12)

Figure 6.11

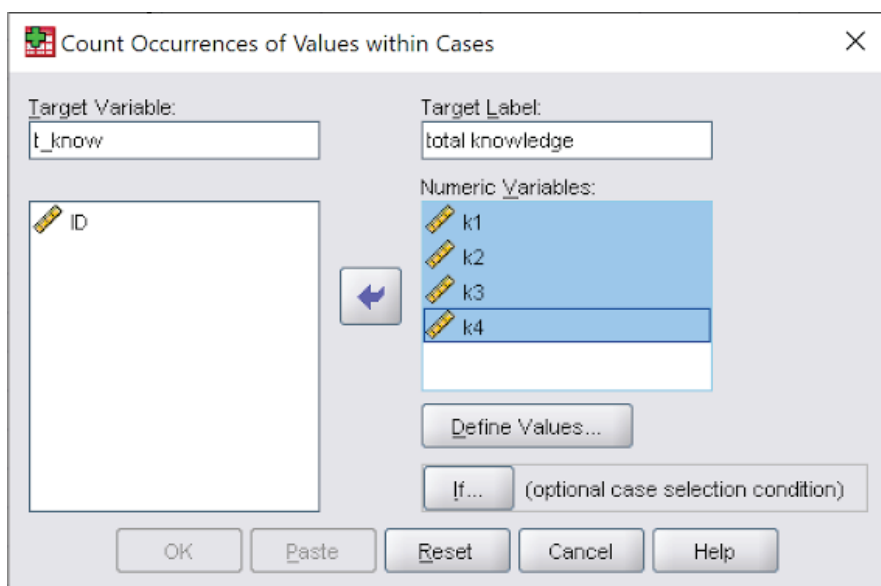
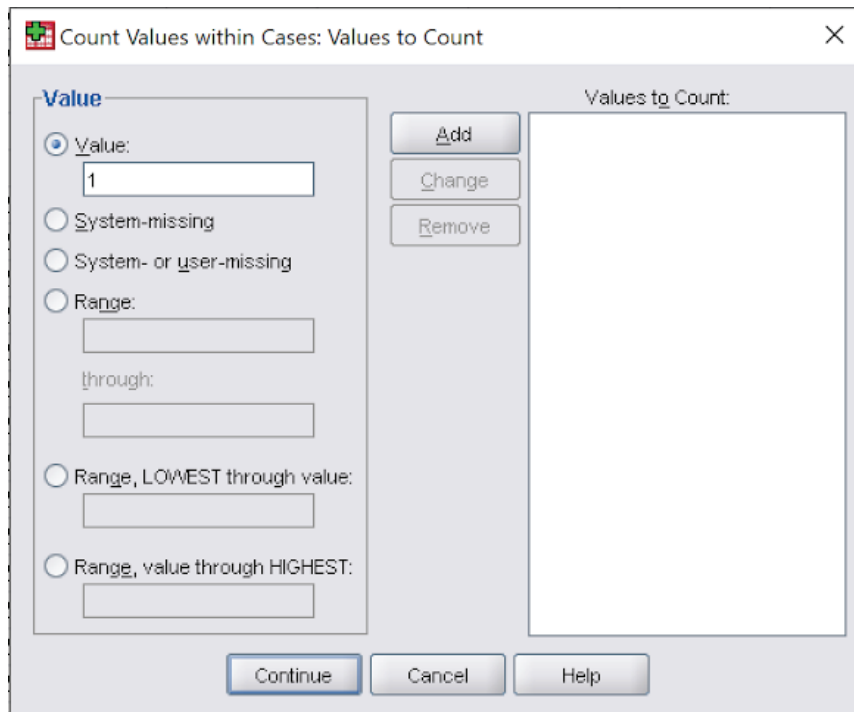


Figure 6.12



SPSS will generate a new variable “t_know (total knowledge on HIV)” (look at the “variable view”). This variable has the total score of knowledge of the students. Now, you can get the descriptive statistics and frequency of the variable “t_know” by using the following commands:

Analyze > Descriptive statistics > Frequencies > Select “t_know” and push it into the “Variables” box > Statistics > Select “Mean, Median and Std. deviation” > Continue > Ok

You will get the tables (table 6.2 and 6.3), showing the descriptive statistics (mean, median, etc.) and frequency distribution of total knowledge of the students. Table 6.2 shows that the mean of the total knowledge is 2.18 (SD 0.63) and the median is 2. Table 6.3 shows that there are 2 (1%) students who do not have any knowledge on HIV transmission (since the score is 0, i.e., could not answer any question correctly). One hundred and twenty five (63.8%) students know 2 ways of HIV transmission, while only 1.5% of the students know all the ways of HIV transmission. You can also classify the students as having “Good” or “Poor” knowledge using a cutoff value based on the total score.

Table 6.2. Descriptive Statistics

total knowledge		
N	Valid	196
	Missing	0
Mean		2.18
Median		2.00
Std. Deviation		.638

Table 6.3. Total knowledge

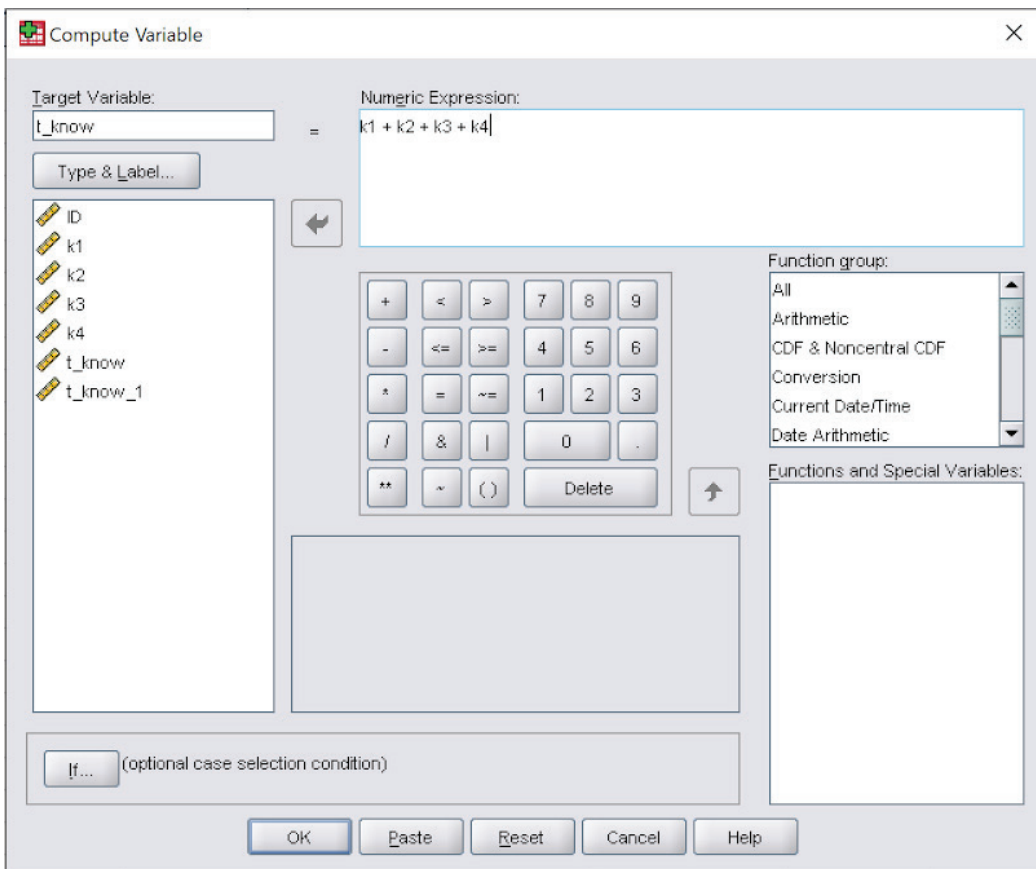
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2	1.0	1.0	1.0
	1	16	8.2	8.2	9.2
	2	125	63.8	63.8	73.0
	3	50	25.5	25.5	98.5
	4	3	1.5	1.5	100.0
	Total	196	100.0	100.0	

There is an alternative way of getting the total score. In that case, the correct answers have to be coded as 1, while the incorrect answers must be coded as 0 (zero). The commands are as follows:

Transform > Compute variable > Write “t_know” in the “Target variable” box > Select “k1” under ‘Type and label’ and push it into the “Numeric expression” box > From the key pad click “+” > Select “k2” and push it into the “Numeric expression” box > From the key pad click “+” > Select “k3” and push it into the “Numeric expression” box > From the key pad click “+” > Select “k4” and push it into the “Numeric expression” box > OK (fig 6.13)

You will get the same results.

Figure 6.13



6.6 Calculation of duration

SPSS can extract time duration from dates. Suppose, you have the data on date of admission (variable name is date_ad) and date of discharge (date_dis) of patients admitted in a hospital. Now, you want to calculate the duration of hospital stay (date of discharge minus date of admission). SPSS can calculate this for you. Use the following commands:

Transform > Compute Variable > Type “**dura**” under “Target Variable” > Click on “Time Duration Extraction” in the “Function Group” box > From “Functions and special variables” box select “Ctime.Days” > Click on the up arrow (at the left side of the box. You will see CTIME.DAYS(?) appears in the “Numeric Expression” box > Select “date_dis” from “Type and Label” box > Push it into the “Numeric Expression” box > Click on – (minus sign from the pad) >

Select “date_ad” from “Type and Label” box and push it into the “Numeric Expression” box > Ok

You will notice that SPSS has generated a new variable “dura” (the last variable) that contains the duration of hospital stay of each subject in the dataset.

6.7 Selecting a sub-group for analysis

You can select a specific sub-group for the analysis of your data. Suppose, you want to analyze your data only for those who have diabetes mellitus. In the dataset, the variable “diabetes” is coded as “1= yes (have diabetes)” and “2= no (do not have diabetes)”. To select the group who have diabetes (i.e., diabetes=1), use the following commands:

Data > Select cases > Select “If condition is satisfied” > Click on “If” > Select the variable “diabetes” and push it into the empty box > Click “=” and then “1” from the number pad > Continue > OK

This would exclude the subjects who do not have diabetes from the analysis. The analysis will be only for those who have diabetes (n=45). If you make a frequency distribution for “sex”, you will see that n=45 (table 6.4). Now, to get all the subjects for the analysis (i.e., to deselect the subgroup), use the commands:

Data > Select cases > Select “All cases” > Ok

Table 6.4. Distribution of sex among diabetic patients

Sex: string					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	20	44.4	44.4	44.4
	Male	25	55.6	55.6	100.0
	Total	45	100.0	100.0	

Section 7

Testing of Hypothesis

The current and following sections provide basic information on how to select statistical tests for testing hypotheses, perform the statistical tests in SPSS and interpret the results of common problems related to health and social science research. Before I proceed, let me discuss a little bit about the hypothesis.

A hypothesis is a statement about one or more populations. The hypothesis is usually concerned with the parameters of the populations about which the statement is made. There are two types of statistical hypothesis, Null (H_0) and Alternative (H_A) hypothesis. The null hypothesis is the hypothesis of equality or no difference. The null hypothesis always says that the two or more quantities (parameters) are equal. *Note that, we always test the null hypothesis, not the alternative hypothesis.* We either reject or do not reject the null hypothesis. If we can reject the null hypothesis, then only we can accept the alternative hypothesis. It is, therefore, necessary to have very clear understanding about the null hypothesis.

Suppose, we are interested to determine the association between coffee drinking and stomach cancer. In this situation, the null hypothesis is "there is no association between coffee drinking and stomach cancer (or, coffee drinking and stomach cancer are independent)", while the alternative hypothesis is "there is an association between coffee drinking and stomach cancer (or, coffee drinking and stomach cancer are not independent) ". If we can reject the null hypothesis by a statistical test (i.e., if the test is significant; $p\text{-value} < 0.05$), then only we can say that there is an association between coffee drinking and stomach cancer.

Various statistical tests are available to test hypothesis. Selecting an appropriate statistical test is the key to analyze data. What statistical test to be used to test the hypothesis depends on study design, data type, distribution of data, and objective of the study. It is, therefore, important to understand the nature of the variable (categorical or quantitative), measurement type, as well as the study design. Following table (table 7) provides basic guideline about the use of statistical tests depending on the type of data and situation.

Table 7. Selecting statistical test for hypothesis testing

7.1 Association between quantitative and qualitative or quantitative variables

	Situation for hypothesis testing	Data normally distributed	Data non-normal
1	<p>Comparison with single population mean (with a fixed value)</p> <p>Example: You have taken a random sample from a population of diabetic patients to assess the mean age. Now, you want to test the hypothesis whether the mean age of diabetic patients in the population is 55 years or not.</p>	1-sample t-test	Sign test/ Wilcoxon Signed Rank test
2	<p>Comparison of means of two related samples</p> <p>Example: You want to test the hypothesis whether the drug “Inderal” reduces blood pressure (BP) or not. To do this study, you have selected a group of subjects and measured their BP before administration of the drug (measurements before treatment; or pre-test). Then you have given the drug “Inderal” to all the subjects and measured their BP after one hour (measurements after treatment; or post-test). Now you want to compare if the mean BP before (pre-test) and after (post-test) administration of the drug is same or not.</p>	Paired t-test	Sign test/ Wilcoxon Signed Rank test
3	<p>Comparison between two independent sample means [association between quantitative and qualitative variable with 2 levels]</p> <p>Example: You have taken a random sample of students of a university. Now, you want to test the hypothesis if the mean systolic blood pressure of male and female students is same or not.</p>	Independent samples t-test	Mann-Whitney U test (also called Wilcoxon Rank-Sum test)

	Situation for hypothesis testing	Data normally distributed	Data non-normal
4	<p>Comparison of more than two independent sample means [association between quantitative and a categorical variable with <i>more than 2 levels</i>]</p> <p>Example: You have taken a random sample from a population. You want to test the hypothesis if the mean income of different religious groups (e.g., Muslim, Hindu and Christian) is same or not. Another example, you have three drugs, A, B & C. You want to investigate whether all these three drugs equally reduce the BP or not.</p>	One way ANOVA	Kruskal Wallis test
5	<p>Association between two quantitative variables</p> <p>Example: You want to test the hypothesis if there is a correlation between systolic BP and age.</p>	Pearson's correlation	Spearman's correlation (Also valid for ordinal qualitative data)

7.2 Association between two qualitative variables

	Situation for hypothesis testing	Test statistics
1	<p>Association between two qualitative variables (independent samples)</p> <p>Example: You have taken a random sample from a population and want to test the hypothesis if there is an association between sex and asthma. Another example, you want to assess the association between smoking and stomach cancer.</p>	Chi-square test/ Fisher's Exact test
2	<p>Association between two qualitative variables (related samples, such as data of a matched case-control study design)</p> <p>Example: You want to test the hypothesis if there is an association between diabetes mellitus and heart disease, when the data is matched for smoking (a matched case-control study design).</p>	McNemar test

7.3 Multivariable analysis

	Type of outcome/dependent variable	Type of multivariable analysis
1	Outcome variable (also called dependent variable) is in interval or ratio scale – e.g., blood pressure, birth weight, blood sugar, etc.	Multiple linear regression; Analysis of variance (ANOVA)
2	Dependent variable is a dichotomous categorical variable (i.e., nominal categorical variable with two levels) – e.g., disease (present or absent); ANC (taken or not taken); outcome (cured or not cured), etc.	Multiple logistic regression
3	Dependent variable is a nominal categorical variable with more than two levels – e.g., treatment seeking behaviour (e.g., treatment not received; treatment received from un-qualified doctor; treatment received from qualified doctor); cause of death (cancer, heart disease, pneumonia), etc.	Multi-nominal logistic regression
4	Dependent variable is an ordinal categorical variable – e.g., severity of anaemia (no anaemia, mild to moderate anaemia, severe anaemia); stage of cancer (stage 1, stage 2, stage 3); severity of pain (mild, moderate, severe), etc.	Proportional odds regression (Ordinal regression)
5	Dependent variable is time to outcome (time to death, time to recurrence, time to cure), etc.	Proportional hazards analysis (Cox regression)
6	Dependent variable is the counts – e.g., number of post-operative infections; number of MI patients admitted in a hospital, etc.	Poisson regression
7	Incidence rates – incidence rate of tuberculosis; incidence rate of pneumonia; incidence rate of car accidents, etc.	Poisson regression

7.4 Agreement analysis

	Situation for hypothesis testing	Test statistics
1	Agreement between two quantitative variables Example: You want to test the hypothesis if two methods of blood sugar measurements agree with each other or not.	Bland Altman test/plots
2	Agreement between two categorical variables Example: You want to test the hypothesis if diagnosis of cataract agree between two physicians.	Kappa estimates

Section 8

Student's t-test for Hypothesis Testing

Student's t-test is commonly known as t-test. It is a commonly used statistical method to test hypothesis. There are several types of t-tests used in different situations (table 7 of section 7), such as: a) one-sample t-test; b) Independent samples t-test; and c) Paired t-test. In this section, I shall discuss all these t-tests and interpretation of the results. Use the data file <Data_3.sav> for practice.

8.1 One-sample t-test

One-sample t-test is done to compare the mean with a hypothetical value. For example, we have collected data on diastolic BP (variable name: dbp) of students of the State University of Bangladesh taking a random sample. We are interested to know if the mean diastolic BP of the students is 80 mmHg or not. Here,

Null hypothesis (H_0): The mean diastolic BP of students is equal to 80 mmHg in the population.

Alternative hypothesis (H_A): The mean diastolic BP of students is different from (not equal to) 80 mmHg in the population.

Assumptions:

1. The distribution of diastolic BP in the population is normal;
2. The sample is a random sample from the population.

The first job, before hypothesis testing, is to check whether the distribution of diastolic BP is normal or not in the population (assumption 1). To do this, check the histogram and/or Q-Q plot of diastolic BP and do the formal statistical test of normality (K-S test or Shapiro Wilk test) as discussed in section 5. If the assumption is met (diastolic BP is at least approximately normal), do the 1-sample t-test, otherwise we have to use the non-parametric test (discussed later). Suppose, diastolic BP is normally distributed in the population. Use the following commands to do the 1-sample t-test:

Analyze > Compare means > One sample t-test > Select the variable “dbp” and push it into the “Test variable(s)” box > Click in the “Test value” box and write “80” > OK

8.1.1 Outputs:

The computer will provide the following tables (table 8.1 and 8.2).

Table 8.1. Descriptive statistics of diastolic BP

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
DIASTOLIC BP	210	83.0429	12.45444	.956

Table 8.2. One-sample t-test results

One-Sample Test						
Test Value = 80						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
DIASTOLIC BP	3.541	209	.000	3.04286	1.3486	4.7371

8.1.2 Interpretation:

In this example, we have tested the *null hypothesis* “the mean diastolic BP of the students is equal to 80 mmHg in the population”. Data shows that the mean diastolic BP of the sample of the students is 83.04 mmHg with an SD of 12.4 mmHg (table 8.1). One-sample t-test results are shown in table 8.2. The calculated value of “t” is 3.541 and the p-value (Sig. 2-tailed) is 0.000. Since the p-value is <0.05, we can reject the null hypothesis at 95% confidence level. This means that the mean diastolic BP of the students (in the population) from where the sample is drawn is different from 80 mmHg ($p < 0.001$). The SPSS has also provided the difference between the observed value (83.04) and hypothetical value (80.0) as mean difference (which is 3.042) and its 95% confidence interval (1.34 – 4.73) (table 8.2).

8.2 Independent samples t-test

Independent samples t-test involves one categorical variable with two levels (2 categories) and one quantitative variable. This test is done to compare the means of two categories of the categorical variable. For example, we are interested to

know if the mean age of diabetic and non-diabetic patients is same or not. Here, the test variable (dependent variable) is age (quantitative) and the categorical variable is diabetes, which has two levels/categories (have diabetes and do not have diabetes).

Hypothesis:

H_0 : The mean age of the diabetic and non-diabetic patients is same in the population.

H_A : The mean age of the diabetic and non-diabetic patients is different (not same) in the population.

Assumptions:

1. The dependent variable (age) is normally distributed at each level of the independent (diabetes) variable;
2. The variances of the dependent variable (age) at each level of the independent variable (diabetes) are same/equal;
3. Subjects represent random samples from the populations.

8.2.1 Commands:

Use the following commands to do the independent samples t-test. Before doing the test, we have to remember/check (from code book or variable view) the category code numbers of diabetes. In our example, we have used code “1” for defining “have diabetes” and “2” for “do not have diabetes”.

Analyze > Compare means > Independent samples t-test > Select “age” and push it into the “test variable(s)” box and select “diabetes” for “grouping variable” box > Click on “define groups” > Type 1 in “Group 1” box and type 2 in “Group 2” box > Continue > OK

Note: You shall have to use exactly the same code numbers as it is in the dataset for the grouping variable. Otherwise, SPSS cannot analyze the data.

8.2.2 Outputs:

The SPSS will produce the outputs as shown in table 8.3 and 8.4.

Table 8.3. Descriptive statistics of age by grouping variable (having diabetes)

Group Statistics					
	DIABETES MELLITUS	N	Mean	Std. Deviation	Std. Error Mean
AGE	Yes	45	27.9111	8.46335	1.26164
	No	165	26.1333	7.18360	.55924

Table 8.4. Independent sample t-test results

Independent Samples Test										
		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
AGE	Equal variances assumed	3.218	.074	1.415	208	.159	1.7778	1.25671	-.699	4.255
	Equal variances not assumed			1.288	62.34	.202	1.7778	1.38003	-.980	4.536

8.2.3 Interpretation:

Table 8.3 shows the descriptive measures of age by grouping variable (diabetes). We can see that there are 45 persons with diabetes and 165 persons without diabetes. The mean age of the diabetic persons is 27.9 (SD 8.46) and that of the non-diabetic persons is 26.1 (SD 7.18) years.

Table 8.4 shows the t-test results. The first portion of the table indicates the *Levene's test* results. This test is done to understand if the variances of age in the two categories of diabetes are homogeneous (equal) or not (assumption 2). Look at the p-value (Sig.) of the Levene's test, which is 0.074. Since the p-value is >0.05 , it indicates that the variances of age of the diabetic and non-diabetic persons are equal (assumption 2 is fulfilled).

Now, look at the other portion of the table, the *t-test for equality of means*. Here, we have to decide which p-value we shall consider. If Levene's test p-value is >0.05 , take the t-test results at the upper row, i.e., t-test for "Equal variances assumed". If the Levene's test p-value is ≤ 0.05 , take the t-test results at the lower

row, i.e., t-test for “Equal variances not assumed”.

In this example, as the Levene’s test p-value is >0.05 , we shall consider the t-test results of “Equal variances assumed”, i.e., the upper row. Table 8.4 shows that the t-value (calculated) is 1.415, and the p-value (2-tailed) is 0.159 (which is >0.05) with 208 degrees of freedom. We cannot, therefore, reject the null hypothesis. This means that the mean age of diabetic and non-diabetic persons in the population from where samples are drawn is not different ($p=0.159$).

8.3 Paired t-test

The paired t-test is done to compare the difference between two means of related samples. Related samples indicate measurements taken from the same subjects in two or more different times/situations. For example, you have organized training for 32 staff of your organization. To evaluate the effectiveness of the training, you have taken a pre-test before the training to assess the current knowledge of the participants. At the end of the training, you have again taken an examination (post-test). Now you want to compare if the training has increased their knowledge or not. Another example is “You want to understand the effectiveness of a drug (e.g., Inderal) in reducing the systolic blood pressure (BP). To do this you have selected a random sample from a population. You have measured the systolic BP of all the individuals before giving the drug (pre-test or baseline). You have again measured their systolic BP one-hour after giving the drug (post-test or endline)”. Paired t-test is the appropriate test to compare the means in both these situations.

Hypothesis:

H_0 : There no difference of the mean scores before and after the training (for example 1).

H_A : The mean scores are different before and after the training.

Assumptions:

1. The difference between two measurements (pre- and post-test) of the dependent variable (examination scores) is normally distributed;
2. Subjects represent a random sample from the population.

8.3.1 Commands:

Analyze > Compare means > Paired-samples t-test > Select the variables “post-test” and “pre-test” and push them into the “Paired variables” box > OK

8.3.2 Outputs:

The SPSS will produce the following outputs (tables 8.5-8.7).

Table 8.5. Descriptive statistics of pre- and post-test results

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Post test score	90.9844	32	8.44096	1.49216
	Pre test score	53.5781	32	15.42835	2.72737

Table 8.6. Correlation between pre- and post-test results

Paired Samples Correlations				
		N	Correlation	Sig.
Pair 1	Post test score & Pre test score	32	.433	.013

Table 8.7. Paired samples t-test results

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	Post test score - Pre test score	37.406	14.02040	2.47848	32.3514	42.4611	15.092	31	.000

8.3.3 Interpretation:

Table 8.5 shows the means of both the pre- (53.5) and post-test (90.9) scores along with the standard deviations and SEs. Looking at the mean scores, we can have an impression on whether the training has increased the mean score or not. We can see that the post-test mean is 90.9, while the pre-test mean is 53.5. To understand, if the difference between post-test mean and pre-test mean is significant or not, we have to check the paired samples t-test results (table 8.7). Table 8.7 shows that the mean difference between the post- and pre-test scores is 37.4. The calculated t-test

value is 15.09 and the p-value (sig.) is 0.000. As the p-value is <0.05 , reject the null hypothesis. This indicates that the mean knowledge score has been increased significantly after the training ($p<0.001$). Note that for conclusion, we do not need table 8.6.

Section 9

Analysis of Variance (ANOVA): One-way ANOVA

Analysis of variance or ANOVA is a commonly used statistical method for testing hypothesis. ANOVA is done to compare the means when the categorical independent variable has more than 2 levels. There are several types of ANOVA tests, such as one-way ANOVA, two-way ANOVA, repeated-measures ANOVA and others. In this section, I shall discuss the one-way ANOVA. Use the data file <Data_3.sav> for practice.

9.1 One-way ANOVA

The one way-ANOVA test is done to compare the means of more than two groups, while t-test compares the means of two groups. The ANOVA test involves two variables, one categorical variable with more than two levels/categories (for example, in our data the variable “religion” [variable name “religion_2”] has 4 categories – Muslim, Hindu, Christian and Buddhism) and a quantitative variable (e.g., income, age, blood pressure, etc.). Suppose, you want to assess if the mean income of all the religious groups is same or not in the population. One-way ANOVA is the appropriate test for this, if the assumptions are met.

Hypothesis:

H_0 : The mean income of all the religious groups is same/equal.

H_A : Not all the means (of income) of religious groups are same.

Assumptions:

1. The dependent variable (income) is normally distributed at each level of the independent variable (religion);
2. The variances of the dependent variable (income) for each level of the independent variable (religion) are same; and
3. Subjects represent random samples from the populations.

If the variances of the dependent variable in all the categories are not equal (violation of assumption 2), but sample size in all the groups is large and similar,

ANOVA can be used.

9.1.1 Commands:

Analyze > Compare means > One-way ANOVA > Select “income” and push it into the “Dependent list” box > Select “religion_2” for the “Factor” box > Options > Select “Descriptive” and “Homogeneity of variance test” > Continue > OK

9.1.2 Outputs:

The SPSS will generate the following outputs (table 9.1-9.3).

Table 9.1. Descriptive statistics of income by religious groups

Descriptives								
INCOME								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
MUSLIM	126	88180.90	17207.61	1532.976	85146.95	91214.85	55927	117210
HINDU	36	79166.03	17804.63	2967.439	73141.81	85190.25	53435	110225
CHRISTIAN	26	79405.62	19857.02	3894.282	71385.19	87426.04	52933	114488
BUDDHISM	22	84796.59	14447.34	3080.185	78391.00	91202.19	56249	109137
Total	210	85194.49	17724.03	1223.074	82783.34	87605.63	52933	117210

Table 9.2. Levene's test for homogeneity of variances of income in different religious groups

Test of Homogeneity of Variances			
INCOME			
Levene Statistic	df1	df2	Sig.
2.056	3	206	0.107

Table 9.3. ANOVA test results

ANOVA					
INCOME					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3306848581.156	3	1102282860.385	3.642	.014
Within Groups	62348694323.301	206	302663564.676		
Total	65655542904.457	9			

9.1.3 Interpretation:

In this example, I have used “income” as the dependent variable and “religion” as the independent (or factor) variable. The independent variable (religion) has 4 categories (levels) – Muslim, Hindu, Christian and Buddhism.

Table 9.1 provides all the descriptive measures (mean, SD, SE, 95% CI, etc.) of income by religion. For example, the mean income of Muslims is 88,180.9 with SD of 17,207.6.

The second table (table 9.2) shows the test results of homogeneity of variances (Levene’s test). This test was done to understand if all the group variances of income are equal or not (assumption 2). Look at the p-value (Sig.), which is 0.107. The p-value is >0.05 , which means that the variances of income in all the religious groups are equal (i.e., assumption 2 is not violated).

Now, look at the ANOVA table (table 9.3). The value of F-statistic is 3.642 and the p-value is 0.014. Since the p-value is <0.05 , reject the null hypothesis. *This means that, not all group means (of income) are same.*

However, the ANOVA test does not provide information about which group means are different. To understand which group means are different, we need to use the post hoc multiple comparison test, such as *Tukey’s test* or *Bonferroni test*. Use the following commands to get the post-hoc test results. *Note that if the ANOVA test (F-test) is not significant (i.e., p-value is >0.05), we do not need the post-hoc test.*

Analyze > Compare means > One-way ANOVA > Select “income” and push it into the “Dependent list” box > Select “religion_2” for the “Factor” box > Options > Select “Descriptive”, and “Homogeneity of variance test” > Continue > Post Hoc > Select “Tukey” (or Bonferroni) > Continue > OK

The SPSS will produce the following table (table 9.4) in addition to others.

Table 9.4. Comparisons of mean income between the religious groups

Multiple Comparisons						
Dependent Variable: INCOME Tukey HSD						
(I) RELIGION	(J) RELIGION	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
MUSLIM	HINDU	9014.88(*)	3287.767	.033	499.15	17530.60
	CHRISTIAN	8775.29	3747.399	.092	-930.94	18481.52
	BUDDHISM	3384.31	4019.891	.834	-7027.71	13796.33
HINDU	MUSLIM	-9014.88(*)	3287.767	.033	-17530.60	-499.15
	CHRISTIAN	-239.59	4477.525	1.000	-11836.93	11357.76
	BUDDHISM	-5630.56	4707.946	.630	-17824.73	6563.60
CHRISTIAN	MUSLIM	-8775.29	3747.399	.092	-18481.52	930.94
	HINDU	239.59	4477.525	1.000	-11357.76	11836.93
	BUDDHISM	-5390.98	5039.677	.708	-18444.37	7662.41
BUDDHISM	MUSLIM	-3384.31	4019.891	.834	-13796.33	7027.71
	HINDU	5630.56	4707.946	.630	-6563.60	17824.73
	CHRISTIAN	5390.98	5039.677	.708	-7662.41	18444.37
* The mean difference is significant at the .05 level.						

9.1.3.1 Interpretation of multiple comparisons table:

Table 9.4 shows the mean difference of income in different religious groups. We can see that the difference of mean income of Muslims and Hindus is 9,014.88 (the minus sign in row 4 indicates that Muslims have greater income than Hindus). The p-value (Sig.) of this difference is 0.033, which is <0.05 . This indicates that the mean income of Muslims and Hindus may be different in the population (Muslims have higher mean income compared to Hindus). The difference of means of other religious groups are not significant as p-values are >0.05 . The table has also provided the 95% CI of the mean differences.

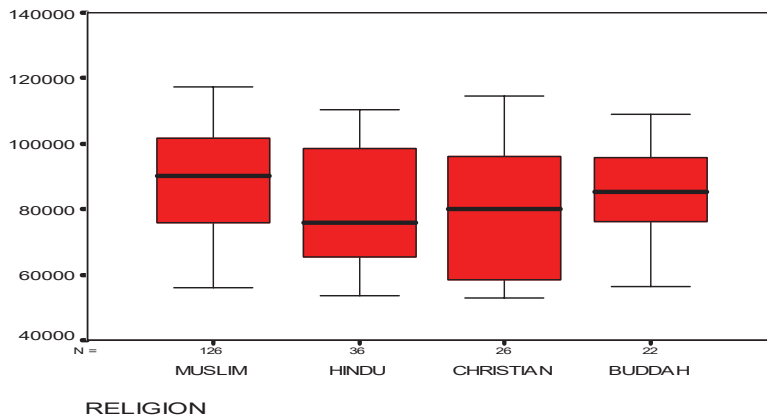
9.1.4 Graph on distribution of medians/means:

You can generate the box and plot charts to see the distribution of medians/means of the dependent variable across the groups (at each level of the independent variable, i.e., in different religious groups). To have the box and plot chart, use the following commands:

Graphs > Legacy dialogs > Boxplots... > Simple > Select “Summaries for groups of cases (already selected by default)” > Define > Select “income” for “Variable” box and select “religion_2” for “Category axis” box > Ok

The SPSS will generate the following box and plot chart (fig 9.1). The horizontal line within the box indicates the median.

Figure 9.1. Box and plot chart of income by religious groups



9.1.5 What to do if the variances are not homogeneous?

If the group variances are not equal (i.e., Levene's test p-value is <0.05), for the comparison of group means, we have to use the Welch test (*or Browne-Forsythe test*). Similarly, for the comparison of individual group means, instead of Tukey's (or Bonferroni) test use the Games-Howell test. Use the following commands to get these test results:

Analyze > Compare means > One-way ANOVA > Select "income" and push it into the "Dependent list" box > Select "religion_2" for the "Factor" box > Options > Select "Descriptive", "Homogeneity of variance test" and "Welch" > Continue > Post Hoc > Select "Games-Howell" under the "Equal Variances not Assumed" > Continue > OK

9.1.5.1 Outputs:

The SPSS will produce the following additional tables (table 9.5 and 9.6).

Table 9.5. Welch test for equality of means

Robust Tests of Equality of Means				
INCOME				
	Statistic ^a	df1	df2	Sig.
Welch	3.292	3	56.236	.027
a. Asymptotically F distributed.				

Table 9.6. Comparison of means of income between the religious groups

Multiple Comparisons						
income Games-Howell						
(I) religion	(J) religion	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
MUSLIM	HINDU	9014.877*	3340.016	.044	166.37	17863.38
	CHRISTIAN	8775.289	4185.146	.175	-2541.95	20092.53
	BUDDHISM	3384.314	3440.575	.760	-5931.90	12700.53
HINDU	MUSLIM	-9014.877*	3340.016	.044	-17863.38	-166.37
	CHRISTIAN	-239.588	4896.032	1.000	-13248.23	12769.05
	BUDDHISM	-5630.563	4277.059	.557	-16986.14	5725.02
CHRISTIAN	MUSLIM	-8775.289	4185.146	.175	-20092.53	2541.95
	HINDU	239.588	4896.032	1.000	-12769.05	13248.23
	BUDDHISM	-5390.976	4965.176	.700	-18635.83	7853.88
BUDDHISM	MUSLIM	-3384.314	3440.575	.760	-12700.53	5931.90
	HINDU	5630.563	4277.059	.557	-5725.02	16986.14
	CHRISTIAN	5390.976	4965.176	.700	-7853.88	18635.83

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

9.1.5.2 Interpretation:

Table 9.5 shows the Welch test results of comparison of means (Robust Tests of Equality of Means). Just look at the p-value (Sig.). The p-value is 0.027, which is <0.05 . This means that the mean income of all the religious groups is not same in the population.

Table 9.6 conveys the same information as of Tukey's test that I have discussed earlier. Here, the difference of mean income of Muslims and Hindus is significantly different as indicated by the p-value (Sig.) ($p=0.044$). The difference of means among the other religious groups is not significant. The table has also provided the 95% CI of the differences.

Section 10

Two-way ANOVA

Two-way ANOVA is like one-way ANOVA except that it examines an additional independent categorical variable. Therefore, the two-way ANOVA involves three variables – one quantitative (dependent variable) and two categorical variables. This test is not commonly used in health research. Use the data file <Data_3.sav> for practice.

10.1 Two-way ANOVA

Suppose, we want to compare the mean systolic BP (SPSS variable name is “sbp”) in different occupation and sex (male and female) groups. Here, the dependent variable is *systolic BP* and the independent variables are *occupation and sex*.

Since we have 4 levels/categories in occupation (govt. job; private job; business and others) and two categories in sex (male and female), we have a factorial design with 8 (4X2) data cells. The two-way ANOVA test answers the following 3 questions:

1. Does occupation influence the systolic BP (i.e., is mean systolic BP among the occupation groups same)?
2. Does sex influence the systolic BP (i.e., is the mean systolic BP same for males and females)?
3. Does the influence of occupation on systolic BP depends on sex (i.e., is there interaction between occupation and sex)?

Questions 1 and 2 refer to the *main effect*, while question 3 explains the *interaction* of two independent variables (occupation and sex) on the dependent variable (systolic BP).

Assumptions:

1. The dependent variable (systolic BP) is normally distributed at each level of the independent variables (occupation and sex);
2. The variances of the dependent variable (systolic BP) at each level of the independent variables are same; and

3. Subjects represent random samples from the populations.

First of all, we have to check for normality of data (systolic BP) in different categories of occupation and sex separately using histogram, Q-Q plot and Shapiro Wilk test (or, K-S test). We also need to check the homogeneity of variances in each group of the independent variables (occupation and sex) using the Levene's test.

10.1.1 Commands:

Analyze > General linear model > Univariate > Select “sbp” for "Dependent variable" box and select “occupation” and “sex” for “Fixed factors” box > Options > Select “Descriptive statistics, Estimates of effect size and Homogeneity test” > Continue > OK

10.1.2 Outputs:

The SPSS will give you the following outputs (table 10.1-10.4).

Table 10.1. Frequency distribution of independent variables

Between-Subjects Factors			
		Value Label	N
OCCUPATION	1	GOVT JOB	60
	2	PRIVATE JOB	49
	3	BUSINESS	49
	4	OTHERS	52
SEX	f	FEMALE	133
	m	MALE	77

Table 10.2. Descriptive statistics of systolic BP by occupation and sex

Descriptive Statistics				
Dependent Variable: SYSTOLIC BP				
OCCUPATION	SEX	Mean	Std. Deviation	N
GOVT JOB	FEMALE	130.84	21.264	38
	MALE	126.86	19.548	22
	Total	129.38	20.574	60
PRIVATE JOB	FEMALE	131.26	21.534	31
	MALE	117.89	13.394	18
	Total	126.35	19.894	49
BUSINESS	FEMALE	131.10	24.023	31
	MALE	123.44	14.448	18
	Total	128.29	21.178	49
OTHERS	FEMALE	125.73	18.772	33
	MALE	129.26	19.084	19
	Total	127.02	18.778	52
Total	FEMALE	129.73	21.309	133
	MALE	124.56	17.221	77
	Total	127.83	20.021	210

Table 10.3. Levene's test result for equality of variances

Levene's Test of Equality of Error Variances(a)			
Dependent Variable: SYSTOLIC BP			
F	df1	df2	Sig.
1.794	7	202	.090

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+OCCUPATION+SEX+OCCUPATION * SEX

Table 10.4. The two-way AVOVA table

Tests of Between-Subjects Effects						
Dependent Variable: SYSTOLIC BP						
Source	Type III Sum of Squares	df	Mean Square	f	Sig.	Partial Eta Squared
Corrected Model	3370.426(a)	7	481.489	1.210	.299	.040
Intercept	3127267.594	1	3127267.594	7856.211	.000	.975
OCCUPATION	470.413	3	156.804	.394	.758	.006
SEX	1394.678	1	1394.678	3.504	.063	.017
OCCUPATION * SEX	1769.393	3	589.798	1.482	.221	.022
Error	80408.741	202	398.063			
Total	3515465.000	210				
Corrected Total	83779.167	209				

a R Squared = .040 (Adjusted R Squared = .007)

10.1.3 Interpretation:

Table 10.1 (between-subjects factors) shows the frequencies of occupation and sex. Table 10.2 (descriptive statistics) shows the descriptive measures of systolic BP at different levels of occupation and sex. For example, the mean systolic BP of females doing the government job is 130.84 (SD 21.2) and that of males doing the government job is 126.8 (SD 19.5).

Table 10.3 shows the Levene's test results for homogeneity of variances. The p-value (Sig.) of the test, as shown in the table, is 0.090. A p-value >0.05 indicates that the variances of systolic BP at each level of the independent variables (occupation and sex) are not different. Thus, the assumption 2 is *not* violated.

The table of "Tests of between-subjects effects" (table 10.4) shows the *main effects* of the independent variables. Look at the p-values (Sig.) of occupation and sex. They are 0.758 and 0.063, respectively. This indicates that the mean systolic BP is not different in different occupation groups as well as sex (males and females). Now, look at the p-value for "*occupation*sex*", which indicates the significance of the interaction between these two variables on systolic BP. A p-value ≤ 0.05 indicates the presence of interaction, that means that the systolic BP of different occupation groups is influenced by (depends on) sex. In our example, the p-value is 0.221 (>0.05), which means that there is no interaction between occupation and sex to influence the systolic BP. The *Partial Eta Squared* (last column of the table) indicates the effect size. The Eta statistics for occupation and sex are 0.006 and 0.017, which are very small. These values are equivalent to R^2 (Coefficient of Determination). Eta 0.006 indicates that only 0.6% variance of systolic BP can be explained by occupation (and 1.7% by sex). However, most of the researchers do not report this in their publications.

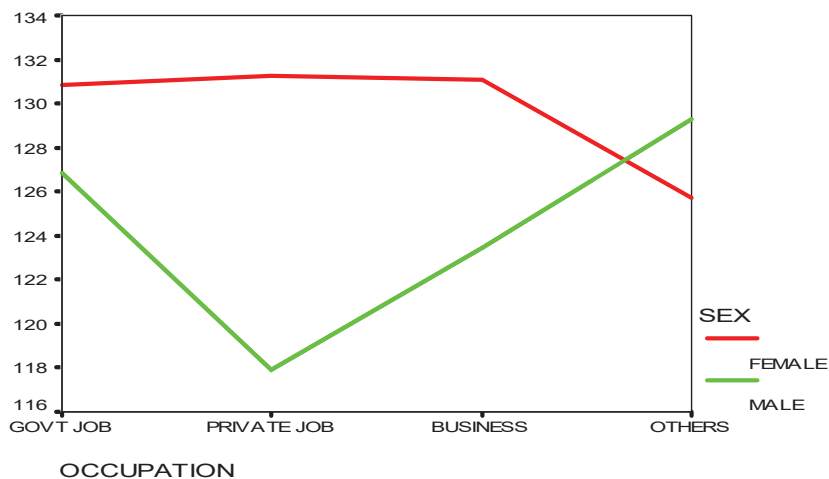
The Post-hoc test (as discussed under one-way ANOVA) is performed if the main effect is significant (i.e., the p-values for occupation and/or sex are <0.05), otherwise it is not necessary. To have a clearer picture of the presence of interaction, it is better to get a graph of the mean systolic BP for occupation and sex. Use the following commands to get the graph.

Graphs > Line > Select "Multiple" > Select "Summarizes for groups of cases" > Define > Select "Other summary function" > Move the dependent variable (sbp) into the "Variable" box > Move one independent variable (occupation) with most categories (here occupation has more categories than sex) into the

"Category axis" box > Move the other independent variable (sex) into "Define line by" box > OK

This will produce the following graph (fig 10.1). The graph shows that there is a greater difference in mean systolic BP between males (117.89) and females (131.26) among private job holders, compared to other occupations. However, this difference is not statistically significant to show an interaction between occupation and sex. This means that there is no significant variation of systolic BP in the occupation groups by sex.

Figure 10.1. Interaction between occupation and sex on systolic BP



Section 11

Repeated Measures ANOVA: One-way

The one-way repeated measures ANOVA test is analogous to paired samples t-test that I have discussed earlier. The main difference is that, in paired samples t-test we have two measurements at different times (e.g., before and after giving a drug, or pre-test and post-test results) on the same subjects, while in one-way repeated measures ANOVA, there are three or more measurements on the same subjects at different points in time (i.e., the subjects are exposed to multiple measurements over a period of time or conditions). The one-way repeated measures ANOVA is also called one-way *within-subjects ANOVA*. Use the data file <Data_repeat_anova_2.sav> for practice.

11.1 One-way repeated measures ANOVA

Suppose, we are interested to assess the mean blood sugar levels at 4 different time intervals (e.g., at hour-0, hour-7, hour-14 and hour-24) after administration of a drug on 15 study subjects. The objective of this study is to assess whether the drug reduces the blood sugar levels over time.

To conduct this study, we have selected 15 individuals randomly from a population and measured their blood sugar levels at the baseline (hour-0). All the individuals were then provided with the drug (say, drug A) and their blood sugar levels were again measured at hour-7, hour-14 and hour-24. We are interested to know if the blood sugar levels over time, after giving the drug, are same or not (in other words, whether the drug is effective in reducing the blood sugar levels over time). The variables hour-0, hour-7, hour-14, and hour-24 are named in SPSS as sugar_0, sugar_7, sugar_14 and sugar_24, respectively. *Note that, in this example, we have only one treatment group* (received drug A) but have the outcome measurements (blood sugar) at 4 different points in time on the same subjects (i.e., we have one treatment group with 4 levels of measurements).

Hypothesis:

H_0 : The mean blood sugar level is same/equal at each level of measurement (i.e., the population mean of blood sugar at 0, 7, 14 and 24 hours is same).

H_A : The mean blood sugar is not same at different levels of measurement (that is, population mean of blood sugar at 0, 7, 14 and 24 hours is different).

Assumptions:

1. The dependent variable (blood sugar level) is normally distributed in the population at each level of within-subjects factor;
2. The population variances of the differences between all combinations of related groups/levels are equal (called *Sphericity assumption*); and
3. The subjects represent a random sample from the population.

11.1.1 Commands:

Analyze > General linear model > Repeated measures > In “Within subject factor name” box write “time” (give any other name) after deleting factor1 > in “Number of levels” box write “4” (since there are 4 time factors) > Add > Write “blood_sugar” in “Measures Name” box > Add > Define > Select variables “sugar_0, sugar_7, sugar_14 and sugar_24” and push them into "Within-Subjects Variables" box > Options > Select "Descriptive statistics, Estimates of effect size and Homogeneity tests" > Select “time” and push it into the “Display means for” box > Select “Compare main effects” > Select “Bonferroni” in box “Confidence interval adjustment” > Continue > Pots > Select “time” and push it into “Horizontal axis” box > Add > Continue > OK

11.1.2 Outputs:

The SPSS will produce several tables. However, we need only the following tables (table 11.1-11.7) for interpreting the results. The tables are set chronologically for easier interpretation (not in the order as provided by SPSS).

Table 11.1. Codes for different levels of measurements of blood sugar

Within-Subjects Factors	
Measure: Blood_sugar	
Time	Dependent Variable
1	sugar_0
2	sugar_7
3	sugar_14
4	sugar_24

Table 11.2. Descriptive statistics of blood sugar at different levels (times) of measurement

Descriptive Statistics			
	Mean	Std. Deviation	N
Blood sugar at hour 0	109.200	5.12975	15
Blood sugar at hour 7	103.733	3.73146	15
Blood sugar at hour 14	97.8667	4.08598	15
Blood sugar at hour 24	98.1333	5.86596	15

Table 11.3. Descriptive statistics of blood sugar at different levels of measurement with 95% CI

Estimates				
Measure: Blood_sugar				
Time	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	109.200	1.324	106.359	112.041
2	103.733	.963	101.667	105.800
3	97.867	1.055	95.604	100.129
4	98.133	1.515	94.885	101.382

Table 11.4. Mauchly's test for Sphericity assumption

Mauchly's Test of Sphericity ^b						
Measure: Blood_sugar						
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^a	
					Greenhouse-Geisser	Lower-bound
Time	.095	29.998	5	.000	.436	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept

Within Subjects Design: Time

Table 11.5. Test results of within subject effects (alternative univariate tests)

Tests of Within-Subjects Effects							
Measure: Blood_sugar							
Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Time	Sphericity Assumed	1299.933	3	433.311	29.023	.000	.675
	Greenhouse-Geisser	1299.933	1.309	993.198	29.023	.000	.675
	Huynh-Feldt	1299.933	1.389	935.638	29.023	.000	.675
	Lower-bound	1299.933	1.000	1299.933	29.023	.000	.675
Error(Time)	Sphericity Assumed	627.067	42	14.930			
	Greenhouse-Geisser	1299.933	18.324	34.222			
	Huynh-Feldt	1299.933	19.451	32.238			
	Lower-bound	1299.933	14.000	44.790			

Table 11.6. Multivariate test results

Multivariate Tests ^b							
Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Time	Pillai's Trace	.918	44.724 ^a	3.000	12.000	.000	.918
	Wilks' Lambda	.082		3.000	12.000	.000	.918
	Hotelling's Trace	11.181	44.724 ^a	3.000	12.000	.000	.918
	Roy's Largest Root	11.181	44.724 ^a	3.000	12.000	.000	.918

a. Exact statistic

b. Design: Intercept

Within Subjects Design: Time

Table 11.7. Pair-wise comparison of mean blood sugar at different times of measurement

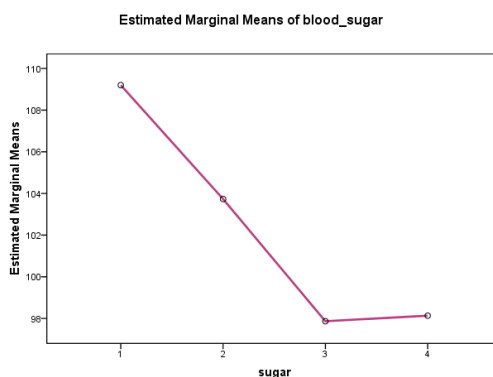
Pairwise Comparisons						
Measure:Blood_sugar						
(I) Time	(J) Time	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval Difference ^a	
					Lower Bound	Upper Bound
1	2	5.467*	1.064	.001	2.202	8.732
	3	11.333*	1.260	.000	7.467	15.200
	4	11.067*	2.256	.001	4.143	17.990
2	1	-5.467*	1.064	.001	-8.732	-2.202
	3	5.867*	.608	.000	4.000	7.734
	4	5.600*	1.492	.013	1.021	10.179
3	1	-11.333*	1.260	.000	-15.200	-7.467
	2	-5.867*	.608	.000	-7.734	-4.000
	4	-.267	1.240	1.000	-4.072	3.539
4	1	-11.067*	2.256	.001	-17.990	-4.143
	2	-5.600*	1.492	.013	-10.179	-1.021
	3	.267	1.240	1.000	-3.539	4.072

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

Figure 11.1. Mean blood sugar at different times of measurement



11.1.3 Interpretation:

The outputs of the analysis are shown in tables 11.1-11.7 and figure 11.1. Table 11.1 shows the value labels (times) of the blood sugar measurements. Table 11.2 and 11.3 (descriptive statistics and estimates) shows the descriptive statistics (mean, standard deviation, no. of study subjects, SE of the means, 95% CI, etc.) of the blood sugar levels at different times of measurement, such as at hour-0, hour-7, hour-14 and hour-24.

One of the important issues for the repeated measures ANOVA test is the *Sphericity assumption*, as mentioned earlier under “assumptions”. Table 11.4 shows the test results of “Mauchly’s test of Sphericity” to understand whether the Sphericity assumption is correct or violated. The table shows that the Mauchly’s W is 0.095 and the p-value is 0.000. Since the p-value is less <0.05, the Sphericity assumption is violated (not correct).

Three types of tests are conducted if within-subjects factors (here, it is the times of measurement of blood sugar, which has 4 levels – hour-0, hour-7, hour-14 and hour-24) have more than 2 levels (here, we have 4 levels). The tests are:

1. Standard univariate test (Sphericity Assumed) [table 11.5];
2. Alternative univariate tests (Greenhouse-Geisser; Huynh-Feldt; Lower-bound) [table 11.5]; and
3. Multivariate tests (Pillai's Trace; Wilks' Lambda; Hotelling's Trace; Roy's Largest Root) [table 11.6]

All these tests evaluate the same hypothesis – i.e., the population means are equal at all levels of the measurement. The standard univariate test is based on the *Sphericity assumption*, i.e., the standard univariate test result is considered, if Sphericity assumption is correct (not violated). *In reality and in most of the cases (also in our example), the Sphericity assumption is violated, and we cannot use the standard univariate test (Sphericity assumed as given in table 11.5) result.*

In our example, we see that the Sphericity assumption is violated, since the Mauchly’s test p-value is 0.000 (table 11.4). Therefore, we shall have to pick up the test results either from alternative univariate tests (table 11.5) or multivariate tests (table 11.6). For practical purpose, it is recommended to use the *multivariate test results* for reporting, since it *does not* depend on Sphericity assumption.

However, for the sake of better understanding, let me discuss table 11.5, which indicates the standard and alternative univariate test results. Table 11.5 shows the

univariate test results of within-subjects effects. The standard univariate ANOVA test result is indicated by the row “Sphericity Assumed”. Use this test result, when Sphericity assumption is *correct/not violated* (i.e., Mauchly’s test p-value is >0.05). Since, our data shows that the Sphericity assumption is violated, we cannot use this test result.

When Sphericity assumption is violated (not correct), you can use the results of one of the alternative univariate tests (i.e., Greenhouse-Geisser, Huynh-Feldt or Lower-bound) for interpretation. It is commonly the Greenhouse-Geisser test, which is reported by the researchers. Table 11.5 shows that the test (Greenhouse-Geisser) provided the same F-value and p-value like other tests. Since the test’s p-value is 0.000, reject the null hypothesis. This means that, the mean blood sugar levels at different time factors (i.e., at different levels of measurement) are not same.

To make it simple, I would suggest to use the multivariate test results, which are not dependent on the Sphericity assumption. Table 11.6 shows the *multivariate test results*. In the multivariate tests table, the SPSS has given several test results, such as Pillai’s Trace, Wilks’ Lambda, Hotelling’s Trace and Roy’s Largest Root. All these multivariate tests have given the same results. It is recommended to use the *Wilks’ Lambda* test results for reporting. In our example, the multivariate test indicates significant time effect on blood sugar levels, as the p-value of Wilks’ Lambda is 0.000. This means that the population means of blood sugar levels at different time factors (different times of measurement) are not same.

The last table (table 11.7) shows pairwise comparison of means at different times of measurement. It shows the results as we have seen under one-way ANOVA (table 9.4; Tukey HSD). Look at the p-values. It is better to assess the differences of adjacent measurements, such as the difference of blood sugar levels between “time 1 & 2”, “time 2 & 3” and “time 3 & 4”. The table shows that all the differences have p-values <0.05 , except for “time 3 and 4” ($p=1.0$). This means that mean blood sugar levels are significantly different in all adjacent time periods except for the time between 3 and 4. The mean blood sugar levels at different times of measurement are depicted in figure 11.1.

Note that if the overall test is not significant (i.e., p-value of Wilks’ Lambda is >0.05), the table for pairwise comparison is not necessary.

Section 12

Repeated Measures ANOVA: Within and Between-Subjects

The within and between-subjects ANOVA is also called two-way repeated measures ANOVA. In the previous section, I have discussed the one-way repeated measures ANOVA, which is also called within-subjects ANOVA. In within-subjects ANOVA, we have *only one* treatment (intervention) group. On the other hand, the within and between-subjects ANOVA is used when there are *more than one* treatment group. In this method, at least 3 variables are involved – one dependent *quantitative* variable, and two independent *categorical* variables with two or more levels. Use the data file <Data_repeat_anova_2.sav> for practice.

12.1 Within and between-subjects ANOVA

Suppose, the researcher wants to do an experiment to compare the efficacy of two drugs (to answer which one is more effective) in reducing the blood sugar levels over time. In such a situation, the researcher may have the following questions to answer:

1. Is there a difference in mean blood sugar levels between drug A and drug B? This is termed *Between-Subjects Factor* – a factor that divides the subjects into two or more distinct subgroups.
2. Is there a reduction in mean blood sugar levels over a time period? This is termed *Within-Subjects Factor* – distinct measurements are made on the same subjects over time. For example, blood sugar levels over time or blood pressure over time, etc.
3. Is there a *group-time interaction*? If there is a time trend, and whether this trend exists for all groups or only for certain groups?

To answer these questions, we have to use *within and between-subjects repeated measures ANOVA*.

Suppose, the researcher has decided to compare the efficacy of Daonil (Glibenclamide) and Metformin (these drugs are used for the treatment of diabetes mellitus) on the reduction of blood sugar levels. In this example, there are 2 treatment groups (SPSS variable name is “treatment”) – Daonil and Metformin. To do the

experiment, the researcher has selected 10 subjects and randomly allocated the treatments (5 in each group). Blood sugar levels of the subjects were measured at the baseline (sugar_0), after 7 hours (sugar_7), after 14 hours (sugar_14) and after 24 hours (sugar_24). Data is provided in the data file <Data_Repeat_anova_2.sav>.

Hypothesis:

We test two hypotheses here. One is for within-subjects effects and the other is for between-subjects effects.

H₀: Daonil and Metformin are equally effective in reducing the blood sugar levels over time (between-subjects effects).

H_A: Both these drugs are not equally effective in reducing the blood sugar level over time (you can also use one sided hypothesis, such as “Daonil is more effective in reducing the blood sugar levels over time compared to Metformin”).

We can also test the hypothesis whether these drugs are effective in reducing the blood sugar levels over time (within-subjects effects; discussed in section 11). The assumptions of two-way repeated measures ANOVA are same as one-way repeated measures ANOVA.

12.1.1 Commands:

Analyze > General linear model > Repeated measures > Write “time” in “Within subject factor name” box > Write “4” in “Number of levels” box (since we have 4 time levels) > Add > Write “blood_sugar” in “Measures name” box > Add > Define > Select variables “sugar_0, sugar_7, sugar_14 and sugar_24” and push them into “Within-Subjects Variables” box > Select “treatment” and push it into “Between-subjects factors” box > Options > Select “treatment” and push it into the “Display means for” box > Select “Compare main effects” > Select “Bonferroni” in “Confidence interval adjustment” box > Select “Descriptive statistics, and homogeneity tests” > Continue > Contrasts > Select “time” > Select “Repeated” in the “Contrast” box under “Change contrast” > Change > Continue > Plots > Select “time” and push it into “Horizontal axis” box > Select “treatment” and push it into the “Separate lines” box

12.1.2 Outputs:

The SPSS will provide many tables, but only the relevant ones are provided below. The outputs are arranged according to – A) Basic tables; B) Tables related to Within-subjects effects; C) Tables related to Between-subjects effects; D) Tables to check the assumptions; and E) Additional tables.

A. Basic tables (table 12.1-12.3):

Table 12.1. Codes for different levels of measurement

Within-Subjects Factors	
Measure: Bloodsugar	
Sugar	Dependent Variable
1	sugar_0
2	sugar_7
3	sugar_14
4	sugar_24

12.2. Codes of different treatment groups

Between-Subjects Factors			
		Value Label	N
treatment groups	1	Daonil	5
	2	Metformin	5

Table 12.3. Descriptive statistics of blood sugar at different levels and treatment groups

Descriptive Statistics				
	treatment groups	Mean	Std. Deviation	N
Blood sugar at hour 0	Daonil	112.8000	2.16795	5
	Metformin	108.4000	7.09225	5
	Total	110.6000	5.46097	10
Blood sugar at hour 7	Daonil	104.0000	4.18330	5
	Metformin	103.0000	4.69042	5
	Total	103.5000	4.22295	10
Blood sugar at hour 14	Daonil	97.4000	3.43511	5
	Metformin	98.6000	3.91152	5
	Total	98.0000	3.52767	10
Blood sugar at hour 24	Daonil	94.4000	2.70185	5
	Metformin	97.6000	2.50998	5
	Total	96.0000	2.98142	10

B. Within-subjects effects (table 12.4-12.6):

Table 12.4. Within-subjects multivariate test results

Multivariate Tests ^b							
Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
time	Pillai's Trace	.955	42.767 ^a	3.000	6.000	.000	.955
	Wilks' Lambda	.045	42.767 ^a	3.000	6.000	.000	.955
	Hotelling's Trace	21.384	42.767 ^a	3.000	6.000	.000	.955
	Roy's Largest Root	21.384	42.767 ^a	3.000	6.000	.000	.955
time * treatment	Pillai's Trace	.452	1.649 ^a	3.000	6.000	.275	.452
	Wilks' Lambda	.548	1.649 ^a	3.000	6.000	.275	.452
	Hotelling's Trace	.825	1.649 ^a	3.000	6.000	.275	.452
	Roy's Largest Root	.825	1.649 ^a	3.000	6.000	.275	.452

a. Exact statistic

b. Design: Intercept + treatment

Within Subjects Design: time

Table 12.5. Descriptive measures of blood sugar at different levels of time

Estimates				
Measure: Bloodsugar				
Time	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	110.600	1.658	106.776	114.424
2	103.500	1.405	100.259	106.741
3	98.000	1.164	95.316	100.684
4	96.000	.825	94.098	97.902

Table 12.6. Pairwise comparisons of blood sugar levels at different time intervals

Tests of Within-Subjects Contrasts							
Measure: bloodsugar							
Source	time	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
time	Level 1 vs. Level 2	504.100	1	504.100	48.010	.000	.857
	Level 2 vs. Level 3	302.500	1	302.500	99.180	.000	.925
	Level 3 vs. Level 4	40.000	1	40.000	4.000	.081	.333
time * treatment	Level 1 vs. Level 2	28.900	1	28.900	2.752	.136	.256
	Level 2 vs. Level 3	12.100	1	12.100	3.967	.082	.332
	Level 3 vs. Level 4	10.000	1	10.000	1.000	.347	.111
Error(time)	Level 1 vs. Level 2	84.000	8	10.500			
	Level 2 vs. Level 3	24.400	8	3.050			
	Level 3 vs. Level 4	80.000	8	10.000			

C. Between-subjects effects (table 12.7-12.9 & fig 12.1):

Table 12.7. Test results of between-subjects effects

Tests of Within-Subjects Contrasts						
Measure:Bloodsugar Transformed Variable:Average						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	416364.025	1	416364.025	9211.593	.000	.999
treatment	.625	1	.625	.014	.909	.002
Error	361.600	8	45.200			

Table 12.8. Descriptive statistics of treatment groups

Estimates				
Measure:Bloodsugar				
treatment groups	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Daonil	102.150	1.503	98.683	105.617
Metformin	101.900	1.503	98.433	105.367

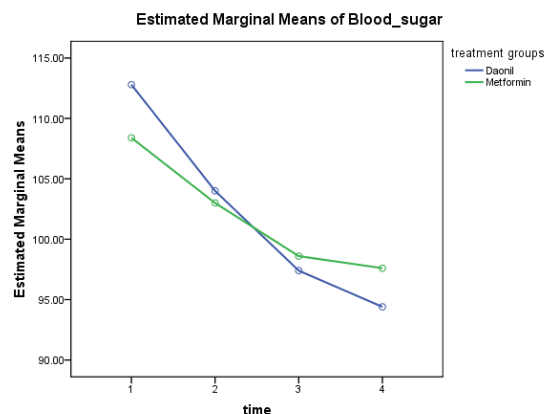
Table 12.9. Pairwise comparisons by treatment groups

Pairwise Comparisons						
Measure:Bloodsugar						
(I) treatment groups	(J) treatment groups	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval Difference ^a	
					Lower Bound	Upper Bound
Daonil	Metformin	.250	2.126	.909	-4.653	5.153
Metformin	Daonil	-.250	2.126	.909	-5.153	4.653

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Figure 12.1. Blood sugar levels by treatment group (Daonil and Metformin)



D. Tables for checking assumptions (table 12.10-12.12):

Table 12.10. Box's M test

Box's Test of Equality of Covariance Matrices ^a	
Box's M	14.734
F	.633
df1	10
df2	305.976
Sig.	.785

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + treatment

Table 12.11. Levene's test of equality of variances

Levene's Test of Equality of Error Variances ^a				
	F	df1	df2	Sig.
Blood sugar at hour 0	3.805	1	8	.087
Blood sugar at hour 7	.076	1	8	.790
Blood sugar at hour 14	.017	1	8	.899
Blood sugar at hour 24	.036	1	8	.855

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + treatment

Table 12.12. Mauchly's test for Sphericity assumption

Mauchly's Test of Sphericity ^b						
Measure: Bloodsugar						
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^b	
					Green-house-Geisser	Huynh-Feldt Lower-bound
Time	.124	14.007	5	.017	.534	.731 .333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + treatment

Within Subjects Design: time

E. Additional tables (12.13-12.17 and fig 12.2):

I have provided some additional tables to show the results, when two treatment groups are *significantly different*. Here, I have compared the blood sugar levels of Daonil with Placebo.

Table 12.13. Descriptive statistics

Descriptive Statistics				
	treatment groups	Mean	Std. Deviation	N
Blood sugar at hour 0	placebo	110.4000	3.36155	5
	Daonil	112.8000	2.16795	5
	Total	111.6000	2.95146	10
Blood sugar at hour 7	placebo	108.6000	2.60768	5
	Daonil	104.0000	4.18330	5
	Total	106.3000	4.08384	10
Blood sugar at hour 14	placebo	108.6000	4.15933	5
	Daonil	97.4000	3.43511	5
	Total	103.0000	6.91215	10
Blood sugar at hour 24	placebo	109.4000	2.60768	5
	Daonil	94.4000	2.70185	5
	Total	101.9000	8.29257	10

Table 12.14. Multivariate test results of within-subjects effects

Multivariate Tests ^b						
Effect		Value	F	Hypothesis df	Error df	Sig.
time	Pillai's Trace	.949	37.505 ^a	3.000	6.000	.000
	Wilks' Lambda	.051	37.505 ^a	3.000	6.000	.000
	Hotelling's Trace	18.752	37.505 ^a	3.000	6.000	.000
	Roy's Largest Root	18.752	37.505 ^a	3.000	6.000	.000
time * treatment	Pillai's Trace	.927	25.566 ^a	3.000	6.000	.001
	Wilks' Lambda	.073	25.566 ^a	3.000	6.000	.001
	Hotelling's Trace	12.783	25.566 ^a	3.000	6.000	.001
	Roy's Largest Root	12.783	25.566 ^a	3.000	6.000	.001

a. Exact statistic

b. Design: Intercept + treatment

Within Subjects Design: time**Table 12.15. Pairwise comparison between time adjacent blood sugar levels**

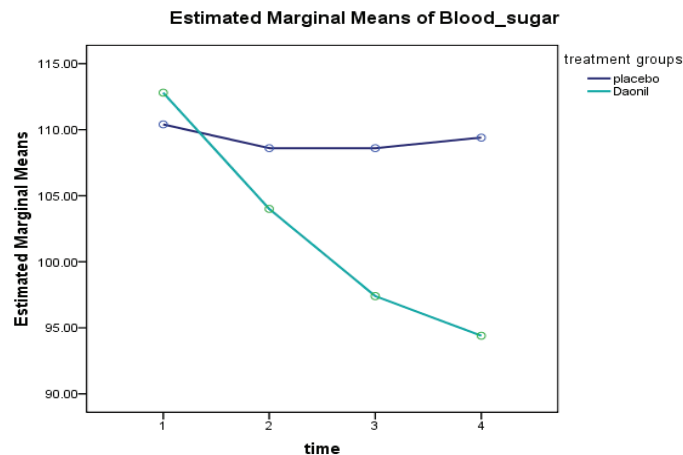
Tests of Within-Subjects Contrasts						
Measure: bloodsugar						
Source	time	Type III Sum of Squares	df	Mean Square	F	Sig.
time	Level 1 vs. Level 2	280.900	1	280.900	66.881	.000
	Level 2 vs. Level 3	108.900	1	108.900	19.274	.002
	Level 3 vs. Level 4	12.100	1	12.100	1.066	.332
time * treatment	Level 1 vs. Level 2	122.500	1	122.500	29.167	.001
	Level 2 vs. Level 3	108.900	1	108.900	19.274	.002
	Level 3 vs. Level 4	36.100	1	36.100	3.181	.112
Error(time)	Level 1 vs. Level 2	33.600	8	4.200		
	Level 2 vs. Level 3	45.200	8	5.650		
	Level 3 vs. Level 4	90.800	8	11.350		

Table 12.16. Test results of between-subjects effects

Tests of Between-Subjects Effects					
Measure:Bloodsugar Transformed Variable:Average					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	111724.900	1	111724.900	15324.461	.000
treatment	126.025	1	126.025	17.286	.003
Error	58.325	8	7.291		

Table 12.17. Descriptive statistics by treatment type

Estimates				
Measure:Bloodsugar				
treatment groups	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
placebo	109.250	1.208	106.465	112.035
Daonil	102.150	1.208	99.365	104.935

Figure 12.2. Blood sugar levels by treatment group (Placebo & Daonil)

12.1.3 Interpretation:

A. Basic tables (outputs under A)

The outputs of the analysis are shown in tables and graphs. Table 12.1 and 12.2 show the value labels of the blood sugar measurements and treatment groups, respectively. Table 12.3 shows the descriptive statistics (mean, standard deviation and no. of study subjects) of blood sugar levels at different times of measurement by treatment groups.

B. Within-subjects effects (outputs under B)

Table 12.4 shows the multivariate test results of within-subjects effects. Since the Sphericity assumption is frequently violated, we would consider the multivariate test results (table 12.4) as discussed in section 11. First, look at the interaction term (time*treatment) in the row of Wilks' Lambda. The p-value (Sig.) is 0.275, which is not significant. This means that there is no interaction between time and treatment (i.e., blood sugar levels over time are not dependent on the treatment groups). Now, look at the row of Wilks' Lambda at time. The p-value is 0.000, which is statistically significant. This means that the mean blood sugar levels measured at different times are significantly different (i.e., there is a significant reduction of blood sugar levels over time in both the treatment groups; fig 12.1). Table 12.5 shows the means and 95% confidence intervals of blood sugar levels at different times of measurement. Table 12.6 shows the difference in blood sugar levels between the adjacent measurements. The table shows that there is significant difference in blood sugar levels between time 1 and 2, and time 2 and 3 ($p=0.000$), but not between time 3 and 4 ($p=0.081$). Note that interaction for any comparison is not significant.

C. Between-subjects effects (outputs under C)

Table 12.7 shows the results of between-subjects effects (between two treatment groups – Daonil and Metformin) on blood sugar levels. The p-value for treatment is 0.909, which is >0.05 . This indicates that there is no significant difference of mean blood sugar levels over time, between Daonil and Metformin groups (also see fig 12.1); i.e., we are unable to reject the null hypothesis. We can see in table 12.8 that the overall mean (without considering the times of measurement) of blood sugar levels for Daonil and Metformin are not that different (102.1 vs 101.9). This indicates that both the drugs are equally effective in reducing the blood sugar levels (no one is better than the other).

The pairwise comparison (table 12.9) shows that the difference in mean blood sugar levels between Daonil and Metformin (0.250) is very small to reject the null hypothesis ($p=0.909$). Figure 12.1 shows the mean blood sugar levels at different times by treatment groups. It shows that the blood sugar levels have been reduced by both the drugs over time, but the difference in reduction between the groups is not significant.

D. Test of assumptions (outputs under D)

Whether the assumptions are violated or not are checked by: a) Box's M test (table 12.10); b) Levene's test (table 12.11); and c) Mauchly's test (table 12.12). If the assumptions are met, the p-values of all these tests would be >0.05 . We can see that the p-values of all these tests are >0.05 except for Mauchly's test ($p=0.017$). Note that the Mauchly's test tests the Sphericity assumption. As discussed earlier, to interpret the results, it is recommended to use the multivariate test, which is not dependent on Sphericity assumption.

E. Additional tables (outputs under E)

Additional tables (table 12.13-12.17) are provided to demonstrate the results, when the treatment groups are different (one is better than the other). Here, I have compared the effectiveness of Daonil compared to placebo. Table 12.13 shows the descriptive statistics of blood sugar levels at different time intervals by treatment group (placebo and Daonil). Though there is no significant difference in mean blood sugar levels at hour 0 (baseline), but they are different over time.

The multivariate test results of within-subjects effects (table 12.14) show that there is interaction (time*treatment) between time and treatment ($p=0.001$) [see the row of Wilks' Lambda under time*treatment]. The p-value of Wilks' Lambda under "time" is also significant ($p=0.000$). This means that there is a significant reduction in mean blood sugar levels over time, and it depends on the treatment group (since there is interaction between time and treatment). Look at fig 12.2. It shows that the blood sugar levels have been reduced significantly over time, among the subjects under the treatment of Daonil, but there is no significant change in the placebo group. The test of between-subjects effects (table 12.16) also conveys the information that the difference in blood sugar levels over time is not same for Daonil and placebo groups ($p=0.003$). We, therefore, conclude that Daonil is effective in reducing the blood sugar level and is superior to (better than) placebo ($p=0.003$).

Section 13

Association between Two Categorical Variables: Chi-Square Test of Independence

The Chi-square test is a commonly used statistical test for testing hypothesis in health research. This test is suitable to determine the association between two categorical variables, whether the data are from cross-sectional, case-control or cohort studies. On the other hand, in epidemiology, cross-tabulations are commonly done to calculate the Odds Ratio (OR) [for case-control studies] or Relative Risk (RR) [for cohort studies] with 95% Confidence Intervals (CI). Odds Ratio and RR are the measures of strength of association between two variables. Use the data file <Data_3.sav> for practice.

13.1 Chi-square test of Independence

Suppose, you have collected data on gender (sex) and diabetes from a group of individuals selected randomly from a population. You are interested to know if there is any association between gender and diabetes. In such a situation, Chi-square test is the appropriate test for testing the hypothesis.

Hypothesis:

H_0 : There is no association between gender and diabetes (it can also be stated as, gender and diabetes are independent).

H_A : There is an association between gender and diabetes (or, gender and diabetes are not independent).

Assumption:

1. Data have come from a random sample drawn from a selected population.

13.1.1 Commands:

Analyze > Descriptive statistics > Crosstabs > Select “sex” and push it into the “Row(s)” box > Select “diabetes” for the “Column(s)” box > Statistics > Select “Chi-square” and “Risk” > Continue > Cells > Select “Row” and “Column” under percentages > Continue > OK

Note: We have selected “risk” to get the OR and RR including their 95% CIs.

13.1.2 Outputs:

Table 13.1. Cross-tabulation between sex and diabetes mellitus

Sex * Diabetes mellitus Crosstabulation					
			Diabetes mellitus		Total
			Yes	No	
Sex	Male	Count	25	52	77
		% within Sex	32.5%	67.5%	100.0%
		% within Diabetes mellitus	55.6%	31.5%	36.7%
	Female	Count	20	113	133
		% within Sex	15.0%	85.0%	100.0%
		% within Diabetes mellitus	44.4%	68.5%	63.3%
Total		Count	45	165	210
		% within Sex	21.4%	78.6%	100.0%
		% within Diabetes mellitus	100.0%	100.0%	100.0%

Table 13.2. Chi-square test result with p-value

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.799 ^a	1	.003		
Continuity Correction ^b	7.795	1	.005		
Likelihood Ratio	8.537	1	.003		
Fisher's Exact Test				.005	.003
Linear-by-Linear	8.758	1	.003		
N of Valid Cases ^b	210				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.50.

b. Computed only for a 2x2 table

Table 13.3. Odds Ratio (OR) and Relative Risk (RR) with 95% Confidence Interval (CI)

Risk Estimate			
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Sex (Male / Female)	2.716	1.385	5.327
For cohort Diabetes mellitus = Yes	2.159	1.288	3.620
For cohort Diabetes mellitus = No	.795	.670	.943
N of Valid Cases	210		

13.1.3 Interpretation:

Table 13.1 is a 2 by 2 table of sex and diabetes with row (% within sex) and column (% within diabetes mellitus) percentages. The question is which percentage should

you report? It depends on the situation and what do you want to report. For the data of a *cross-sectional study*, it may provide better information to the readers if row percentages are reported. For example, one can understand from the table 13.1 that the prevalence of diabetes among males is 32.5% and that of the females is 15.0%, when row percentages are used. However, column percentages can also be reported in cross-sectional studies (most of the publications use column percentages). If column percentage is used, the meaning would be different. In this example (table 13.1), it means that among those who have diabetes, 55.6% are male, compared to 31.5% who do not have diabetes. If data are from a *case-control study*, you must report the column percentages (we cannot use row percentages for case-control studies). On the other hand, for the data of a cohort study, one should report the row percentages. In this case it would indicate the incidence of the disease among males and females.

We can see in table 13.1 (in the row of total) that the overall prevalence (irrespective of sex) of diabetes is 21.4% (consider the data is from a cross-sectional study). Table 13.1 also shows that 32.5% of the males have diabetes compared to only 15.0% among females (i.e., the prevalence among males and females). The Chi-square test actually tests the hypothesis whether the prevalence of diabetes among males and females is same in the population or not.

Table 13.2 shows the Pearson Chi-square test results, including the degree of freedom (df) and p-value (Asymp. Sig). The table also shows other test results, such as Continuity Correction and Fisher's Exact test. Before we look at the Chi-square test results, it is important to check if there is any cell in the 2 by 2 table with expected value <5 . This information is given at the bottom of the table at "a" as "0 cells (0%) have expected count less than 5". For the use of the Chi-square test, it is desirable to have no cell (in a 2 by 2 table) with expected count less than 5. If this is not fulfilled, we have to use the *Fisher's Exact test p-value* to interpret the result (see table 13.2). In fact, to use the Chi-square test, *no more than 20% cells* should have expected frequency <5 . You can have the expected frequencies for all the cells if you select "Expected" under "Count" in "Cell" option during analysis.

For the Chi-square test, consider the Pearson Chi-square value (see table 13.2). In our example, Chi-square value is 8.799 and the p-value is 0.003 (table 13.2). Since the p-value is <0.05 , there is an association between sex and diabetes. It can, therefore, be concluded that the prevalence of diabetes among males is significant-

ly higher than the females, which is statistically significant at 95% confidence level ($p=0.003$).

Table 13.3 shows the OR (2.716) and its 95% CI (1.385-5.327). Use the OR if the data are from a case-control study. Odds Ratio is also sometimes used for cross-sectional data. The table also provided the RR (2.159) and its 95% CI (1.288-3.620) [take the RR and 95% CI for diabetes= yes]. Use RR if the data are from a cohort study. Note that both the OR and RR is statistically significant as they do not include 1 in the 95% CI. Odds ratio 2.71 indicates that males are 2.7 times more likely to have diabetes compared to the females. On the other hand, RR 2.15 indicates that the risk of having diabetes is 2.1 times higher in males compared to females. SPSS will *not* provide the OR and RR, if there are more than 2 categories in any of the variables (e.g., a 2 by 3 table). In such a situation, you have to get the OR and RR in different ways.

Table 13.4. Decision for using Chi-square test

Situation	Right test
Sample size >100 and expected cell value >10	Pearson's Chi-square (uncorrected)
Sample size 31-100 and expected cell count between 5-9	Pearson's Chi-square with Yate's correction (continuity correction row in table 13.2)
Sample size less than 30 and/or expected cell value <5	Fisher's Exact test

Section 14

Association between Two Continuous Variables: Correlation

Nature and strength of relationship between two or more continuous variables can be determined by *regression and correlation* analysis. *Correlation* is concerned with measuring the *strength of relationship* between continuous variables. The correlation model provides information on the relationship between two variables, without distinguishing which is dependent and which is independent variable. But the basic procedure for regression and correlation model is the same.

Under the correlation model, we calculate the “r” value. The “r” is called the sample *correlation coefficient*. It (“r” value) indicates the degree of linear relationship between dependent (Y) and independent (X) variable. Value of “r” ranges from +1 to –1. In this section, I shall discuss the correlation model. Use the data file <Data_3.sav> for practice.

14.1 Pearson correlation

Person correlation is done when the normality assumption is met (i.e., both the dependent and independent variables are normally distributed; assumption 1). Suppose, we want to explore if there is a correlation/association between systolic BP (variable name is “sbp”) and diastolic BP (variable name is “dbp”).

Hypothesis:

H_0 : There is no correlation between systolic and diastolic BP.

H_A : There is a correlation between systolic and diastolic BP.

Assumptions:

1. The variables (systolic and diastolic BP) are normally distributed in the population;
2. The subjects represent a random sample from the population.

The first step, before doing correlation, is to generate a *scatter diagram*. The scatter diagram provides information/ideas about:

- Whether there is any correlation between the variables;

- Whether the relationship (if there is any) is linear or non-linear; and
- Direction of the relationship, i.e., whether it is positive (if the value of one variable increases with the increase of the other variable) or negative (if the value of one variable decreases with the increase of the other variable).

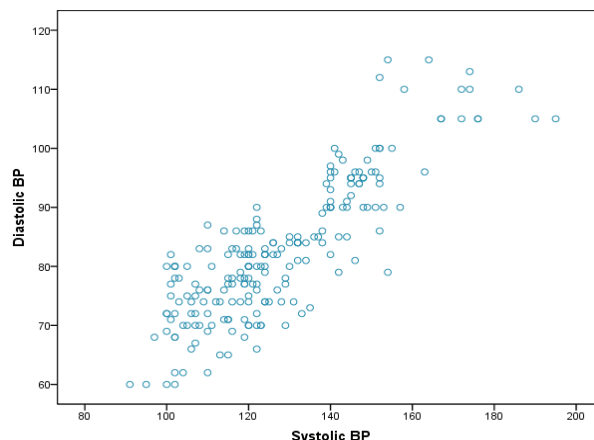
14.1.1 Commands for scatter plot:

To get the scatter plot of systolic and diastolic BP, use the following commands:

Graphs > Legacy dialogues > Scatter... > Select “Simple” > Define > Select “sbp” for “X-axis” and “dbp” for “Y-axis” > Select “ID_no” for “Level cases by” (this would label the outliers, if there is any, by its id number) > Ok

This will give you the following scatter plot (fig 14.1).

Figure 14.1. Scatter plot of systolic and diastolic BP



If you want to get the **regression line** on the scatter plot, use the following commands:

Graphs > Legacy dialogues > Interactive > Scatterplot... > Select “sbp” and drag it into the “X-axis” box > select “dbp” and drag it into the “Y-axis” box > Click on “Fit” tab > Select “Regression” after clicking the dropdown arrow > Ok

The SPSS will produce the following scatter plot (fig 14.2) with the regression line on it. In the same manner, you can produce the scatter diagram of age and diastolic BP (fig 14.3).

Figure 14.2. Scatter diagram of systolic and diastolic BP with regression line

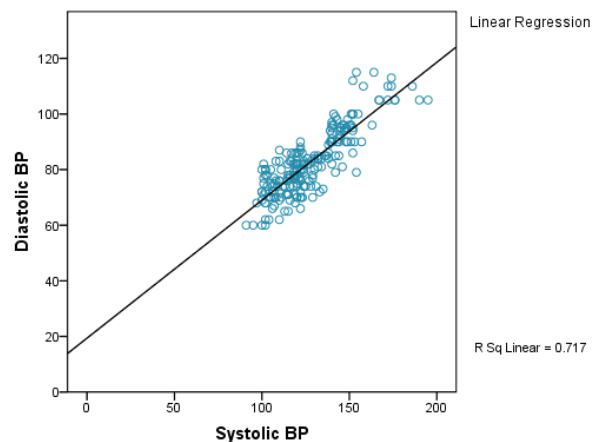
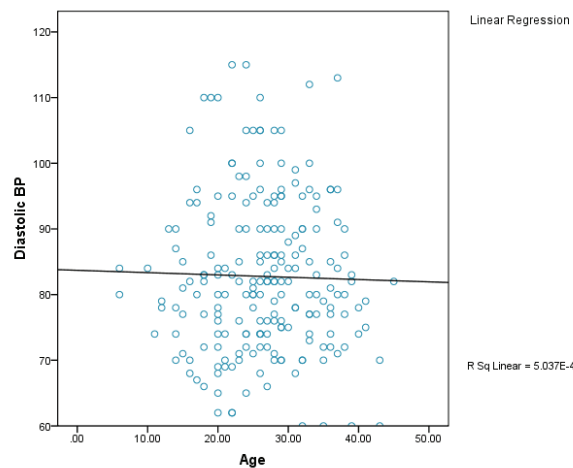


Figure 14.3. Scatter diagram of diastolic BP and age with regression line



14.1.2 Commands for Pearson correlation:

Analyze > Correlate > Bivariate > Select the variables “sbp” and “dbp” and push them into the "Variables" box > Select “Pearson” under the “Correlation coefficients” (usually set as default) > Ok

14.1.3 Outputs:

Table 14.1. Pearson correlation between systolic and diastolic BP

Levene's Test of Equality of Error Variances ^a			
		SYSTOLIC BP	DIASTOLIC BP
SYSTOLIC BP	Pearson Correlation	1	.858(**)
	Sig. (2-tailed)	.	.000
	N	210	210
DIASTOLIC BP	Pearson Correlation	.858(**)	1
	Sig. (2-tailed)	.000	.
	N	210	210

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

14.1.4 Interpretation:

In the first step, I have constructed the scatter plot of systolic and diastolic BP (fig 14.1 and 14.2). Figure 14.1 shows that the data points are scattered around an invisible straight line and there is an increase in the diastolic BP (Y) as the systolic BP increases (X). This indicates that there may have a positive correlation between these two variables. Look at fig 14.2, which shows the regression line in the scatter plot. The regression line has passed from near to the lower left corner to the upper right corner, indicating a positive correlation between systolic and diastolic BP. If the relationship were negative (inverse), the regression line would have passed from the upper left corner to lower right corner. Figure 14.3 shows the scatter plot of diastolic BP and age. It does not indicate any correlation between diastolic BP and age, since the dots are scattered around the regression line, which is more or less parallel to the X-axis.

For correlation, look at the value of correlation coefficient [r] (Pearson Correlation). Table 14.1 shows that the correlation coefficient of systolic and diastolic BP is 0.858 and the p-value is 0.000. Correlation coefficient “r” indicates the *strength/degree of linear relationship* between the two variables (systolic and diastolic BP). As the value of “r” is positive and the p-value is <0.05, there is a significant positive correlation between systolic and diastolic BP.

The value of “r” lies between –1 and +1. Values near to “zero” indicate no correlation, while values near to “+1” or “–1” indicate strong correlation. Negative (–r) value indicates an inverse relationship. A value of $r \geq 0.8$ indicates very strong correlation; “r” value between 0.6 and 0.8 indicates moderately strong correlation; “r” value between 0.3 and 0.5 indicates fair correlation and “r” value < 0.3 indicates poor correlation.

14.2 Spearman's correlation

Spearman's correlation (instead of Pearson correlation) is done when the normality assumption is violated (i.e., if the distribution of either the dependent or independent or both the variables are not normal). Spearman's correlation is also applicable for two categorical ordinal variables, such as intensity of pain (mild, moderate, severe pain) and grade of cancer (stage 1, stage 2, stage 3, etc.).

Suppose, we want to explore if there is any correlation between systolic BP (variable name is "sbp") and income (income is not normally distributed).

14.2.1 Commands for Spearman's correlation:

Analyze > Correlate > Bivariate > Select the variables "sbp" and "income" and push them into the "Variables" box > Select "Spearman" under the "Correlation coefficients" > Ok

14.2.2 Outputs:

Table 14.2. Correlation (Spearman's) between systolic BP and income

Correlations				
			Systolic BP	Monthly income
Spearman's rho	Systolic BP	Correlation Coefficient	1.000	.007
		Sig. (2-tailed)		.919
		N	210	210
	Monthly income	Correlation Coefficient	.007	1.000
		Sig. (2-tailed)	.919	
		N	210	210

14.2.3 Interpretation:

Table 14.2 shows the Spearman's correlation coefficient between systolic BP and income. The results indicate that there is no correlation between systolic BP and income ($r = 0.007$; $p = 0.919$), since the "r" value is very small and the p-value is > 0.05 .

14.3 Partial correlation

The purpose of doing the partial correlation is to assess the correlation (indicated

by the “r” value) between two variables after adjusting for one or more other variables (continuous or categorical). This means that through partial correlation, we get the adjusted “r” value after controlling for the confounding factors. For example, if we assume that the relationship between systolic and diastolic BP may be influenced (confounded) by other variables (such as, age, and diabetes), we should do the partial correlation to exclude the influence of other variables (age, and diabetes). The partial correlation will provide the correlation (r value) between systolic and diastolic BP after controlling/ adjusting for age and diabetes.

14.3.1 Commands:

Analyze > Correlate > Partial > Select “sbp” and “dbp” for “Variables” box > Select “age” and “diabetes” for “Controlling for” box > Ok

14.3.2 Outputs:

Table 14.3. Correlation between systolic and diastolic BP after controlling for age and diabetes mellitus

Correlations				
Control Variables			Systolic BP	Monthly income
Age & Diabetes mellitus	Systolic BP	Correlation	1.000	.847
		Significance (2-tailed)	.	.000
		df	0	206
	Diastolic BP	Correlation	.847	1.000
		Significance (2-tailed)	.000	.
		df	206	0

14.3.3 Interpretation:

Table 14.3 shows the results of partial correlation between systolic and diastolic BP after adjusting for age and diabetes mellitus. We can see that $r=0.847$ and $p=0.000$. This means that these two variables (systolic and diastolic BP) are significantly correlated ($p=0.000$), even after controlling for age and diabetes mellitus. If the relationship between systolic and diastolic BP was influenced by age and diabetes mellitus, the crude (unadjusted) and adjusted “r” values would be different. Look at table 14.1, which shows the crude “r” value (0.858). After adjusting for age and diabetes, the “r” value becomes 0.847 (table 14.3). Since the crude and adjusted “r” values are almost similar, there is no influence of age and

diabetes mellitus in the relationship between systolic and diastolic BP (i.e., age and diabetes mellitus are not the confounding factors in the relationship between systolic and diastolic BP).

Section 15

Linear Regression

Regression analysis is a commonly used statistical method for data analysis. Nature and strength of relationship between two or more continuous variables can be determined by regression and correlation analysis. I have already discussed about correlation in the previous section. While *correlation* is concerned about measuring the *direction and strength of linear relationship* between the variables, *regression* analysis is helpful to *predict or estimate* the value of one variable corresponding to a value of another variable(s) (e.g., to understand whether systolic BP is a good predictor of diastolic BP). In regression analysis, our main interest is in *regression coefficient* (also called slope or β), which indicates the strength of association between dependent (Y) and independent (X) variable. Regression can be done as: a) *Simple linear regression*, and b) *Multiple linear regression*.

In this section, both simple and multiple linear regressions are discussed. Multiple linear regression is a type of *multivariable analysis*. Multivariable analysis is a statistical tool where multiple independent variables are considered for a single outcome. The terms “multivariate analysis” and “multivariable analysis” are often used interchangeably in health research. Multivariate analysis actually refers to the statistical methods for the analysis of multiple outcomes. Multivariable analyses are widely used in observational studies, intervention studies (randomized and nonrandomized trials), and studies of diagnosis and prognosis. The main purposes of multivariable analysis are to:

- a) Determine the relative contribution of independent variables to the outcome variable;
- b) Adjust for the confounding factors;
- c) Predict the probability of an outcome, when several characteristics are present in an individual; and
- d) Assess interaction of multiple variables for the outcome.

There are several types of multivariable analysis methods. The choice of multivariable analysis, for the type of outcome variable, is summarized in table 7.3 (section 7). The commonly used multivariable analysis methods in health research include multiple linear regression, logistic regression and proportional hazards

regression (Cox regression) that are discussed in this manual. Use the data file <Data_4.sav> for practice.

15.1 Simple linear regression

In simple linear regression, there is one dependent and one independent variable. The objective of simple linear regression is to find the *population regression equation*, which describes the true relationship between the dependent variable (Y) and independent variable (X). In simple linear regression model, two variables are involved – one is independent variable (X), placed on X-axis, and the other is dependent variable (Y), placed on Y-axis. Then, we call it “regression of Y on X”.

Suppose, we want to do a simple linear regression analysis of diastolic BP (dependent variable) on systolic BP (independent variable). The objective is to find the population regression equation to predict the diastolic BP by systolic BP.

Assumptions:

- 1. Normality:** For any fixed value of X (systolic BP), the sub-population of Y values (diastolic BP) is normally distributed;
- 2. Homoscedasticity:** The variances of the sub-populations of “Y” are all equal;
- 3. Linearity:** The means of the sub-populations of “Y” lie on the same straight line;
- 4. Independence:** Observations are independent of each other.

The first step in analyzing the data for regression is to construct a scatter diagram, which has already been discussed in section 14. This would give an idea about the linear relationship between the variables, systolic and diastolic BP.

15.1.1 Commands:

Analyze > Regression > Linear > Select “dbp” for “Dependent” box and “sbp” for “Independent(s)” box > Method “Enter” (usually the default) > Statistics > Select “Estimates, Descriptive, Confidence interval, and Model fit” > Continue > Ok

15.1.2 Outputs:

Table 15.1. Mean and standard deviation of the variables

Descriptive Statistics			
	Mean	Std. Deviation	N
DIASTOLIC BP	83.04	12.454	210
SYSTOLIC BP	127.83	20.021	210

Table 15.2. Correlation between systolic and diastolic BP

Correlations			
		DIASTOLIC BP	SYSTOLIC BP
Pearson Correlation	DIASTOLIC BP	1.000	.858
	SYSTOLIC BP	.858	1.000
Sig. (1-tailed)	DIASTOLIC BP	.	.000
	SYSTOLIC BP	.000	.
N	DIASTOLIC BP	210	210
	SYSTOLIC BP	210	210

Table 15.3. Correlation coefficient (R) and coefficient of determination (R-square)

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.858(a)	.737	.736	6.403

a Predictors: (Constant), SYSTOLIC BP

Table 15.4. ANOVA table for significance of "R"

ANOVA(b)						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23890.586	1	23890.586	582.695	.000(a)
	Residual	8528.028	208	41.000		
	Total	32418.614	209			

a Predictors: (Constant), SYSTOLIC BP

b Dependent Variable: DIASTOLIC BP

Table 15.5. Constant (a) and regression coefficient (b)

Coefficients(a)								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	14.779	2.862		5.164	.000	9.136	20.422
	SYSTOLIC BP	.534	.022	.858	24.139	.000	.490	.578

a Dependent Variable: DIASTOLIC BP

15.1.3 Interpretation:

Table 15.1 and 15.2 provides the descriptive statistics (mean and standard deviation) and correlation coefficient (r-value) of the diastolic and systolic BP. The model summary table (table 15.3) shows the Pearson's correlation coefficient "R" ($r = 0.858$) and coefficient of determination "R-square" ($r^2 = 0.737$). The value of "R" is same, as we have seen in section 14.

It is important to note the value of R-square (coefficient of determination) given in the model summary table (table 15.3). R-square indicates the amount of variation in "Y" due to "X", that can be explained by the regression line. Here, the R-square value is 0.737 (~0.74), which indicates that 74% variation in diastolic BP can be explained by the systolic BP. The rest of the variation (36%) is due to other factors (unexplained variation). The adjusted R-square value (0.736), as shown in the table, is the value when the R-square is adjusted for better population estimate.

The ANOVA table (table 15.4) indicates whether the correlation coefficient (R) is significant or not (i.e., whether the linear regression model is useful to explain the dependent variable by the independent variable). As the p-value (Sig.) is 0.000, R is statistically significant at 95% confidence level. We can, therefore, conclude that there is a significant positive correlation (because R value is positive) between the diastolic and systolic BP, and we can use the regression equation for prediction. The table also shows the regression (also called explained) sum of squares (23890.586) and residual (also called error) sum of squares (8528.028). The residual indicates the difference between the observed value and predicted value (i.e., value on the regression line). Residual sum of squares provides an idea about how well the regression line actually fits into the data.

Table 15.5 (coefficients) provides quantification of the relationship between the diastolic and systolic BP. The table shows the values for "a" or Y-intercept (also called constant) and "b" (unstandardized coefficients) or slope (also called regression coefficient, β). Note that for a single independent variable, standardized coefficient (Beta) is equal to Pearson's correlation value.

Here, the value of "a" is 14.779 and "b" is 0.534 (both are positive). The value, $a = +14.78$, indicates that the regression line crosses/cuts the Y-axis above the origin (zero) and at the point 14.78 (a negative value indicates that the regression line crosses the Y-axis below the origin). This value (value for a) does not have any practical meaning, since it indicates the average diastolic BP of individuals, if the systolic BP is 0 (zero).

The value of “b” (the regression coefficient or slope) indicates the amount of variation/change in “Y” (here it is diastolic BP) for each unit change in “X” (systolic BP). Here, the value of “b” is 0.534, which means that if the systolic BP increases (or decreases) by 1 mmHg, the diastolic BP will increase (or decrease) by 0.534 mmHg. The table also shows the significance (p-value) of “b”, which is 0.000. Note that for simple linear regression, if R is significant, “b” will also be significant and will have the same sign (positive or negative).

We know that the simple linear regression equation is, $Y = a + bX$ (“Y” is the predicted value of the dependent variable; “a” is the Y-intercept or constant; “b” is the regression coefficient and “X” is a value of the independent variable). Therefore, the regression/prediction equation for this regression model is

$$Y = 14.737 + 0.534X.$$

With this equation, we can estimate the diastolic BP by the systolic BP. For example, what would be the estimated diastolic BP of an individual whose systolic BP is 130 mmHg? The answer is, the estimated diastolic BP would be equal to $(14.727 + 0.534 \times 130)$ 84.1 mmHg.

Note that, if we want to use the regression equation for the purpose of prediction/estimation, “b” has to be statistically significant ($p < 0.05$). In our example, the p-value for “b” is 0.000, and we can, therefore, use the equation for the prediction of diastolic BP by systolic BP.

Table 15.5 has actually evaluated whether “b” in the population is zero or not by the t-test (*Null hypothesis*: “b” is equal to “zero” in the population; *Alternative hypothesis*: the population regression coefficient is not equal to “zero”). We can reject the null hypothesis, since the p-value is < 0.05 . It can, therefore, be concluded that the systolic BP can be used to predict/estimate the diastolic BP using the regression equation, $Y = 14.737 + 0.534X$.

15.2 Multiple linear regression

In simple linear regression, two variables are involved - one dependent (Y) and one independent (X) variable. The independent variable is also called *explanatory* or *predictor* variable. In multiple regression, there are more than one explanatory (independent) variables in the model. The explanatory variables may be quantitative or categorical. The main purposes of multiple regression analysis are to:

- Get the adjusted estimates of regression coefficients (B) of the explanatory variables in the model;
- Predict or estimate the value of the dependent variable by the explanatory variables in the model; and
- Understand the amount of variation in the dependent variable explained by the explanatory variables in the model.

Suppose, we want to assess the contribution of four variables (age, systolic BP, sex and religion) in explaining the diastolic BP in a sample of individuals selected randomly from a population. Here, the dependent variable is the diastolic BP and the explanatory variables (independent variables) are age, systolic BP, sex and religion. Of the explanatory (independent) variables, two are quantitative (age and systolic BP) and two are categorical variables (sex and religion). Of the categorical variables, sex has two levels (male and female) and religion has 3 levels (Islam, Hindu and Christian). When the independent variable is categorical with more than two levels (e.g., religion), we need to create dummy variables for that variable. For example, if we want to include the variable “religion” in the regression model, we shall have to create dummy variables for religion.

15.2.1 Creating dummy variables:

In our example, the variable “religion” has 3 levels and are coded as 1= Islam; 2= Hindu and 3= Christian. We cannot simply put “religion” as one of the explanatory variables in the regression model, because the coding is arbitrary and the regression estimates obtained for religion would be meaningless. We need to create dummy variables for religion.

The number of dummy variables to be created for “religion” is two (no. of levels minus 1). Before creating the dummy variables, we have to decide about the comparison group. Let us consider “Christian” as the comparison group, and assign “0” (zero) as its code number. We shall create two dummy variables – one is “reli_1” and the other is “reli_2” for religion. To create the dummy variables, we shall have to recode the variable “religion” using the following commands.

Step 1: Create the first dummy variable “reli_1” for religion

Transform > Recode into different variables > Select “religion” and push it into the “Input variable – Output variable” box > Write “reli_1” in the “Output vari-

able name” box > write “Dummy variable 1 for religion” in the “label” box > Change > Click on “Old and New Values..” > Select “Value” under “Old value” and write 1 in the box > Select “Value” under “New value” and write 1 in the box > Add > Select “All other values” under “Old value” > Write 0 (zero) in the box “Value” under the “New value” > Add > Continue > Ok

Step 2: Create the second dummy variable “reli_2” for religion

Transform > Recode into different variables > Select “religion” and push it into the “Input variable – Output variable” box > Write “reli_2” in the “Output variable name” box > Write “Dummy variable 2 for religion” in the “Label” box > Change > Click on “Old and New Values..” > Select “Value” under “Old value” and write 2 in the box > Select “Value” under “New value” and write 1 in the box > Add > Select “All other values” under “Old value” > Write 0 in the box “Value” under the “New value” > Add > Continue > Ok

The above commands will create two dummy variables for religion, the “reli_1” (for which code 1= Islam and 0= other religions, i.e., Hindu and Christian)” and “reli_2” (for which code 1= Hindu and 0= other religions, i.e., Islam and Christian)”. You can see the new variables in the variable view of the data file. *Don't forget to provide the value labels for the dummy variables.*

15.2.2 Changing string variable into numeric variable:

If we want to include the variable “sex” in the model, we need to check its coding. If the variable is coded as string variable (e.g., m= male and f= female, as is done in our data), we need to recode it as a numeric variable, say 0= female and 1= male. In this case, when multiple regression will be performed, the regression estimate in the model will be for males compared to females (lower value will be the comparison group). Use the following commands to create a numeric variable for sex.

Transform > Recode into different variables > Select “sex” and push it into the “Input variable – Output variable” box > Write “sex_1” in the “Output variable name” box > Write “Sex numeric” in the “Label” box > Click on “Change” > Click on “Old and New Values..” > Select “Value” under “Old value” and write f in the box > Select “Value” under “New value” and type 0 in the box > Add > Select “Value” under “Old value” and write m in the box > Write 1 in the box “Value” under the “New value” > Add > Continue > Ok

This would create a new variable “sex_1” (last variable in the variable view) with codes 0= female and 1= male. Go to the “variable view” of the data file and set these code numbers in the column “Value” of the variable sex_1.

15.2.3 Sample size for multiple regression:

Multiple regression should be done if the sample size is fairly large. What would be the minimum sample size, depends on how many independent variables we want to include in the model. Different authors provided different guidelines. One author recommends a minimum of 15 subjects for each of the independent variables in the model. Other authors provided a formula ($n = 50 + 8m$) to estimate the number of subjects required for the model. For example, if we intend to include 5 independent variables in the model, we need to have at least 90 subjects ($50 + 8*5$). For stepwise regression, there should have 40 cases for each of the independent variables in the model.

15.2.4 Commands for multiple linear regression:

Use the following commands for multiple regression analysis, where the dependent variable is dbp (diastolic BP) and the explanatory (independent) variables are age, sbp (systolic BP), sex_1, reli_1 and reli_2.

Analyze > Regression > Linear > Select “dbp” for “Dependent” box and “age, sbp, sex_1, reli_1 and reli_2” for “Independent(s)” box > Method “Enter” (usually the default) > Statistics > Select “Estimates, Confidence interval, and Model fit” > Continue > Ok

15.2.5 Outputs:

Table 15.6. Multiple R, R-square and adjusted R-square values

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.852 ^a	.725	.719	6.233

a. Predictors: (Constant), Reli_2, Systolic BP, age, Sex, Reli_1

b. Dependent Variable: Diastolic BP

Table 15.7. ANOVA table for significance of R

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20925.182	5	4185.036	107.710	.000 ^a
	Residual	7926.385	204	38.855		
	Total	28851.567	209			

a. Predictors: (Constant), Reli_2, Systolic BP, age, Sex, Reli_1

b. Dependent Variable: Diastolic BP

Table 15.8. Adjusted regression coefficients of explanatory variables and their significance

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	20.894	3.420		6.110	.000	14.151	27.637
	age	.004	.058	.003	.070	.944	-.110	.119
	Systolic BP	.490	.022	.836	22.559	.000	.447	.532
	Sex	-2.179	.916	-.090	-2.380	.018	-3.985	-.374
	Islam	.162	1.369	.007	.119	.906	-2.537	2.862
	Hindu	-.271	1.492	-.010	-.182	.856	-3.212	2.670

a. Dependent Variable: Diastolic BP

15.2.6 Interpretation:

Table 15.6 (model summary) shows the values for R (0.852), R-square (0.725) and adjusted R-square (0.719) [adjusted for better population estimation]. In multiple regression, the R measures the correlation between the observed value of the dependent variable and the predicted value based on the regression model. The R-square may overestimate the population value, if the sample size is small. The adjusted R-square gives the R-square value of better population estimation. The R-square value 0.725 indicates that all the independent variables (age, systolic BP, sex and religion) together in the model explains 72.5% variation in diastolic BP, which is statistically significant ($p=0.000$), as shown in the ANOVA table (table 15.7).

The Coefficients table (table 15.8) shows regression coefficients (unstandardized and standardized), p-values (Sig.) and 95% confidence intervals (CI) for regression coefficients of all the explanatory variables in the model along with the constant. This is the most important table for interpretation of results. The unstandardized regression coefficients (B) are shown in the table for age (0.004; $p=0.944$), systolic BP (0.490; $p<0.001$), sex (-2.179; $p=0.018$ for males compared

to females), Islam (0.162; $p=0.906$ compared to Christian) and Hindu (-0.271; $p=0.856$ compared to Christian).

From this output (table 15.8), we conclude that the systolic BP and sex are the factors significantly influencing the diastolic BP (since the p -values are <0.05). The other variables in the model (age and religion) do not have any significant influence in explaining the diastolic BP. The unstandardized coefficient (B) [also called *multiple regression coefficient*] for systolic BP, in this example, is 0.490 (95% CI: 0.45 to 0.53). This means that the average increase (or decrease) in diastolic BP is 0.49 mmHg, if the systolic BP increases (or decreases) by 1 mmHg after adjusting for all other variables (age, sex and religion) in the model. On the other hand, the unstandardized coefficient (B) for sex is -2.179 (95% CI: -3.985 to -0.374), which means that (on an average) the diastolic BP of males is 2.2 mmHg less (since the coefficient is negative. If it is positive, it would be more) than that of the females, given the other variables constant.

The standardized coefficients (Beta) (table 15.8) indicate which independent variables have more influence on the dependent variable (diastolic BP). Bigger the value more is the influence. We can see in table 15.8 that the standardized coefficients for systolic BP and sex are 0.83 and -0.09, respectively. This means that systolic BP has greater influence in explaining the diastolic BP than the sex.

15.2.7 Regression equation:

The regression equation to estimate the average value of the dependent variable with the explanatory variables is as follows:

$$Y = a + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \dots + B_nX_n$$

Here, “Y” is the estimated mean value of the dependent variable; “a” is the constant (or Y-intercept); “B” is the regression coefficient(s) and X is the value of the variable(s) in the model.

Suppose, we want to estimate the average diastolic BP of an individual who is 40 years old, male, Muslim and has systolic BP 120 mmHg. In table 15.8 (coefficients), we see that the regression coefficients are for age ($=0.004$ [B1]), systolic BP ($=0.49$ [B2]), sex ($= -2.179$ [B3] for being male) and Islam ($=0.162$ [B4] for being Muslim) and the constant is 20.894. Therefore, the estimated diastolic BP of an individual would be:

$$Y = 20.894 + 0.004*40 + 0.49*120 + (-2.179*1) + 0.162*1 = 77.84.$$

15.2.8 Problem of multicollinearity:

Before deciding about the multiple regression model, we need to check for *multicollinearity* (inter-correlations among the independent variables) of the independent variables. If there are moderate to high inter-correlations among the independent variables, two situations may occur. *Firstly*, the importance of a given explanatory variable may be difficult to determine because of biased (distorted) p-value; and the *other* is dubious relationships may be obtained. For example, if there is multicollinearity, we may observe that the regression coefficient for sex is not significant and the systolic BP has a negative relationship with the diastolic BP. Another important sign of multicollinearity is a *severe reduction of the Adjusted R Square value*.

To determine the correlations among the independent variables, we can generate the Pearson's correlation matrix. For example, we want to see the correlations among the systolic BP, age, sex and religion. Use the following commands to get the correlation matrix.

Analyze > Correlate > Bivariate... > Select the variables "sbp, age, sex_1, reli_1 and reli_2" for the "Variables" box > Ok

The SPSS will produce the following correlation matrix table (table 15.9).

Table 15.9. Correlation matrix of independent variables

		Correlations				
		Systolic BP	Sex	Reli_1	Reli_2	age
Systolic BP	Pearson Correlation	1	-.125	.026	-.011	-.042
	Sig. (2-tailed)		.071	.710	.870	.542
	N	210	210	210	210	210
Sex	Pearson Correlation	-.125	1	.077	.038	-.066
	Sig. (2-tailed)	.071		.269	.581	.344
	N	210	210	210	210	210
Reli_1	Pearson Correlation	.026	.077	1	-.757*	.073
	Sig. (2-tailed)	.710	.269		.000	.292
	N	210	210	210	210	210
Reli_2	Pearson Correlation	-.011	.038	-.757*	1	-.020
	Sig. (2-tailed)	.870	.581	.000		.776
	N	210	210	210	210	210
Age	Pearson Correlation	-.042	-.066	.073	-.020	1
	Sig. (2-tailed)	.542	.344	.292	.776	
	N	210	210	210	210	210

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

We can see in the table (15.9) that there is a moderately strong correlation between reli_1 and reli_2 ($r = -0.757$), while the correlation coefficients for other variables are low. However, the correlation between reli_1 and reli_2 did not affect our regression analysis.

Pearson's correlation can only check collinearity between any two variables. Sometimes a variable may be multicollinear with a combination of variables. In such a situation, it is better to use the tolerance measure, which gives the strength of the linear relationships among the independent variables (usually the dummy variables have higher correlation). To get the tolerance measure (another measure for multicollinearity), use the following commands:

Analyze > Regression > Linear > Select "dbp" for "Dependent" box and "age, sbp, sex_1, reli_1 and reli_2" for "Independent(s)" box > Method "Enter" > Statistics > Select "Estimates, Confidence interval, Model fit, and **Collinearity diagnostics**" > Continue > Ok

This would provide the collinearity statistics in the coefficients table as shown in table 15.10.

Table 15.10. Collinearity statistics for multicollinearity diagnosis

Coefficients ^a										
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	20.894	3.420		6.110	.000	14.151	27.637		
	age	.004	.058	.003	.070	.944	-.110	.119	.983	1.018
	Systolic BP	.490	.022	.836	22.559	.000	.447	.532	.981	1.019
	Sex	-2.179	.916	-.090	-2.380	.018	-3.985	-.374	.950	1.053
	Islam	.162	1.369	.007	.119	.906	-2.537	2.862	.411	2.432
	Hindu	-.271	1.492	-.010	-.182	.856	-3.212	2.670	.416	2.404

a. Dependent Variable: Diastolic BP

The tolerance value ranges from 0 to 1. A value close to "zero" indicates that the variable is almost in a linear combination (i.e., has strong correlation) with other independent variables. In our example (table 15.10), the tolerance values for age, systolic BP, and sex are more than 0.95. However, the tolerance values of Islam (reli_1) and Hindu (reli_2) [the dummy variables] are a little more than 0.4. *The recommended tolerance level is more than 0.6 before we put the variable in the multiple regression model.* However, a tolerance of 0.4 and above is accept-

able, especially if it is a dummy variable. The other statistics provided in the last column of the table are the VIF (Variance Inflation Factor). This is the inverse of the tolerance value.

If there are variables that are highly correlated (tolerance value is <0.4), one way to solve the problem is to exclude one of the correlated variables from the model. The other way is to combine the explanatory variables together (e.g., taking their sum).

Finally, for developing a model for multiple regression, we should first check for multicollinearity and then the residual assumptions (see below). If they fulfil the requirements, then only we can finalize the regression model.

15.2.9 Checking for assumptions:

For practical purpose, there are three assumptions that need to be checked on the residuals for the linear regression model to be valid. The assumptions are:

- a. There is no outlier;
- b. The data points are independent;
- c. The residuals are normally distributed with mean = 0 and have constant variance.

15.2.9.1 Checking for outliers and independent data points (assumptions a and b):

Use the following commands to check for outliers (casewise diagnostics) and data points are independent (Durbin-Watson statistics):

Analyze > Regression > Linear > Select “dbp” for “Dependent” box and “age, sbp, sex_1, reli_1 and reli_2” for “Independent(s)” box > Method “Enter” > Statistics > Select “Estimates, Confidence interval, Model fit, Casewise diagnostics and Durbin-Watson” > Continue > Ok

The SPSS will produce the Model Summary table (table 15.11), casewise diagnostics table (table 15.12, if there are outliers, otherwise not) and residuals statistics table (table 15.13).

Table 15.11. Durbin-Watson statistics for checking data points are independent

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.852 ^a	.725	.719	6.233	1.701

a. Predictors: (Constant), Religion: dummy var_2, Systolic BP, age, Sex: numeric code, Religion: dummy var_1

b. Dependent Variable: Diastolic BP

Table 15.12. Case number and the outlier

Casewise Diagnostics ^a				
Case Number	Std. Residual	Diastolic BP	Predicted Value	Residual
204	-3.625	62	85.11	-23.115

a. Dependent Variable: Diastolic BP

Table 15.13. Residuals statistics with outliers in the data set

Residuals Statistics ^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	65.46	115.85	82.77	9.917	210
Residual	-23.115	18.678	.000	6.301	210
Std. Predicted Value	-1.745	3.336	.000	1.000	210
Std. Residual	-3.625	2.929	.000	.988	210

a. Dependent Variable: Diastolic BP

Table 15.14. Residuals statistics without any outliers in the data set

Residuals Statistics ^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	65.32	116.16	82.77	10.006	210
Residual	-15.446	18.452	.000	6.158	210
Std. Predicted Value	-1.743	3.338	.000	1.000	210
Std. Residual	-2.478	2.960	.000	.988	210

a. Dependent Variable: Diastolic BP

Look at the residuals statistics table (table 15.13). Our interest is in the *Std. Residual* value. The “minimum” and “maximum” values should not exceed “+3” or “-3”. Table 15.13 shows that the minimum value is “-3.625” (exceeded -3). This means that there are outliers. Now, look at the casewise diagnostics table (table 15.12). The table shows that there is an outlier in the diastolic BP, the value of which is 62 and the case number (ID number) is 204 (if there is no outlier in the data, this table will not be provided). If no outlier is present in the data, we shall get a Residuals Statistics table like table 15.14.

The Durbin-Watson test is done to check whether data points are independent. The Model Summary table (table 15.11) shows the Durbin-Watson statistics results in the last column. *The Durbin-Watson estimate ranges from 0 to 4. Values around 2 indicate that the data points are independent.* Values near zero indicate a strong positive correlation and values near 4 indicate a strong negative correlation. The table shows that the value of the Durbin-Watson statistics, in our example, is 1.701, which is close to 2 (i.e., data points are independent)

15.2.9.2 Checking for normality assumption of the residuals and constant variance:

Commands:

Analyze > Regression > Linear > Select “dbp” for “Dependent” box and “age, sbp, sex_1, reli_1 and reli_2” for “Independent(s)” box > Method “Enter” > Statistics > Select “Estimates, Confidence interval, Model fit” > Continue > Plots > Select **“Histogram and normal probability plot”** under “Standardized residual plots” > Place **“*ZRESID on Y; and *ZPRED on X”** under “Scatter 1 of 1” > Continue > Ok

The SPSS will produce the histogram (fig 15.1), normal probability plot (fig 15.2) and a scatter plot (fig 15.3) for the residuals. The distribution of the residuals is normal as seen in the histogram (fig 15.1) and P-P plot (fig 15.2). The constant variance (homoscedasticity) is checked in the scatter plot (fig 15.3). If the scatter of the points shows no clear pattern (as seen in fig 15.3), we can conclude that the variance is constant.

Figure 15.1. Histogram

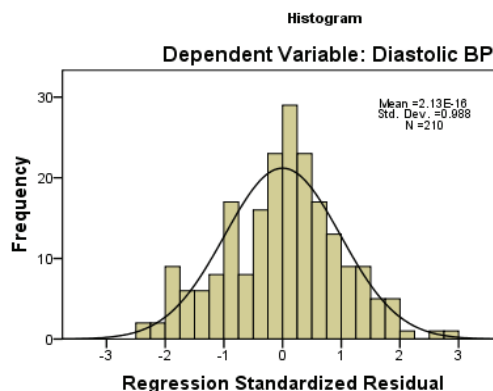


Figure 15.2. Normal probability plot

Normal P-P Plot of Regression Standardized Residual

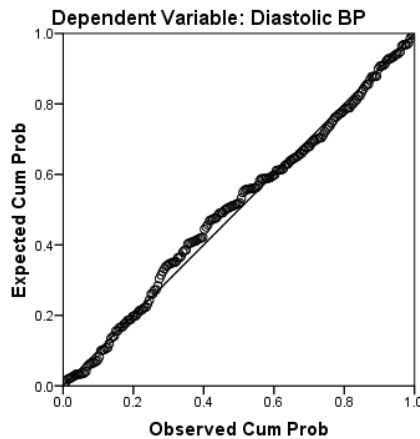
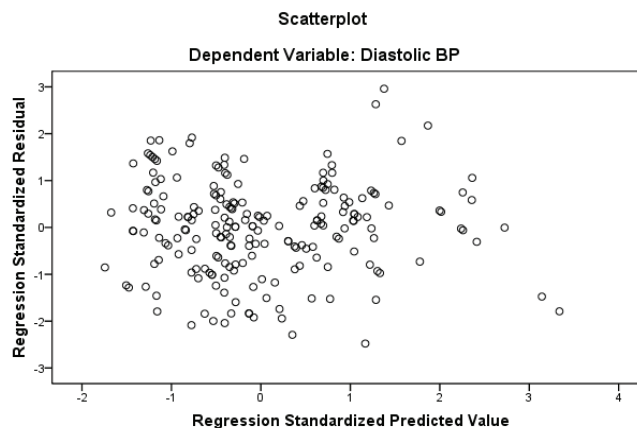


Figure 15.3. Scatter plot of standardized residual vs standardized predicted value



15.2.10 Variable selection for the model:

In general, independent variables to be selected for multivariable analysis should include the risk factors of interest and potential confounders (based on theory, prior research findings and empirical findings), while variables with lots of missing values should be excluded.

We have used the “Enter” method for modelling earlier in this section. The “Enter” method uses all the independent variables in the model included by the researcher. It does not exclude any variable automatically from the model. Automatic procedures can be used to determine which independent variables will be

included in the model. The major reason for using the automatic selection procedure is to minimize the number of independent variables necessary to estimate or predict the outcome. SPSS and other data analysis software have the option to automatically select the independent variables in the model. They use statistical criteria to select the variables and their order in the model. The commonly used variable selection techniques are provided in table 15.15.

Table 15.15. Methods of variable selection

Technique	Method	Advantages and disadvantages
Forward	This method <i>enters</i> variables in the model sequentially. The order is determined by the variable's association (significance) with the outcome (variables with strongest association are entered first) after adjustment for the other variables already in the model.	Best suited for dealing with the studies where the sample size is small. Does not deal well with suppressor (confounding) effects.
Backward	This technique <i>removes</i> variables from the model sequentially. The order is determined by the variable's association with the outcome (variables with weakest association leave first) after adjustment for the variables already in the model.	Better for assessing suppressor effect than the forward selection method.
Stepwise/ Remove	This is the <i>combination</i> of forward and backward methods. In the stepwise method, variables that are entered are checked at each step for removal. Likewise, in the removal method, variables that are excluded will be checked for re-entry.	Has the ability to manage large number of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options.
Enter (all variables)	Enters all the variables at the same time (does not remove any variable automatically from the model).	Including all variables may be problematic, if there are many independent variables and the sample size is small

Let us see how to use the “Stepwise” method (commonly used method in multiple regression analysis) for modelling. To do this, use the following commands (only change is in “Method”):

Analyze > Regression > Linear > Select “dbp” for “Dependent” box and “age, sbp, sex_1, reli_1 and reli_2” for “Independent(s)” box > Method “**Stepwise**” (fig 15.4) > Statistics > Select “Estimates, Confidence interval, and Model fit” > Continue > Ok

15.2.10.1 Outputs (only the relevant table is provided):

Table 15.16. Regression models

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	19.407	2.793		6.948	.000	13.900	24.913
	Systolic BP	.496	.022	.847	22.961	.000	.453	.539
2	(Constant)	21.016	2.838		7.405	.000	15.421	26.611
	Systolic BP	.490	.022	.836	22.768	.000	.447	.532
	Sex: numeric	-2.180	.893	-.090	-2.441	.015	-3.941	-.419

a. Dependent Variable: Diastolic BP

15.2.10.2 Interpretation:

During analysis, we included 5 independent variables (age, systolic BP, sex, and two dummy variables of religion) in the model. The SPSS has provided table 15.16 that shows the adjusted regression coefficients and models. Let us compare the outputs in table 15.16 with those of table 15.8, where we have used the “*enter*” method. In table 15.8, we can notice that the SPSS retained all the independent variables in the model that were included, and only the “systolic BP” and “sex” were found to be significantly associated with the dependent variable (diastolic BP). When we used the “*stepwise*” method, the SPSS has provided two models – model 1 and model 2. In the model 1, there is only one independent variable (systolic BP) and in the model 2, there are two independent variables (systolic BP and sex; others are automatically removed). We consider the last model as the final model.

Sometimes you may need to include certain variable(s) in the model for theoretical or practical reason. In such a situation, after you derive the model with “stepwise” method, add the additional variable(s) of your choice and re-run the model using the “*enter*” method.

For automatic selection method, you can specify the inclusion (entry) and

exclusion (removal) criteria of the variables. Usually, the inclusion and exclusion criteria, set as default in SPSS, are 0.05 and 0.10, respectively (fig 15.5). You can, however, change the criteria based on your requirements. Finally, for model building, the researcher should decide the variables to be included in the final model based on theoretical understanding and empirical findings.

Figure 15.4. Model selection options

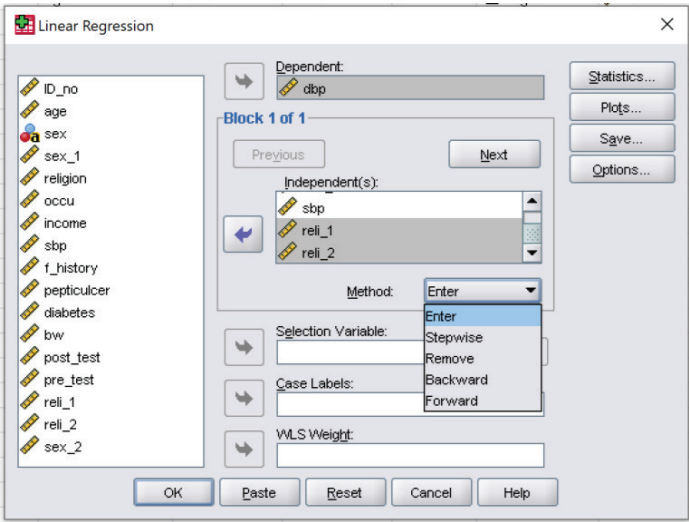
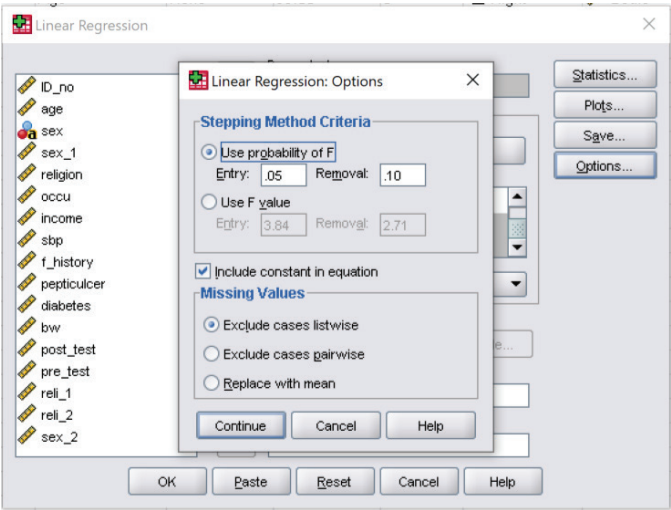


Figure 15.5. Selection criteria for entry and removal of variables



Section 16

Logistic Regression

Logistic regression is a commonly used statistical method for health data analysis. Logistic regression is done when the outcome variable is a dichotomous variable, such as diabetes (present or absent), vaccinated (yes or no) and an outcome (died or did not die). The purposes of logistic regression analysis (and other multivariable analysis) are to: a) Adjust the estimate of risk for a number of confounding factors set in the model; b) Determine the relative contribution of factors to a single outcome; c) Predict the probability of an outcome for a number of independent variables in the model; and d) Assess interaction of multiple variables for the outcome. Use the data file <Data_4.sav> for practice.

16.1 Logistic regression analysis

Logistic regression is appropriate to adjust for multiple confounding factors or model (identify) the predictors of a dichotomous categorical outcome variable (e.g., disease present or absent). For logistic regression analysis in SPSS, the dichotomous outcome variable should be coded as “0= disease absent” and “1= disease present”. SPSS will consider the *higher value to be the predicted outcome and the lower value as the comparison group*. Suppose, we want to predict (or identify the factors associated with) diabetes with sex (variable name: sex_1), age, peptic ulcer (variable name: pepticulcer) and family history of diabetes (variable name: f_history). To perform the logistic regression analysis, recode diabetes as “0= diabetes absent” and “1= diabetes present” (if it is not coded like this). Similarly, it is better to recode the categorical predictor (independent) variables as “0 for no (comparison group)” and “1 for yes”.

Assumptions:

Logistic regression does not make any assumptions concerning the distribution of predictor (independent) variables. However, it is sensitive to high correlation among the independent variables (multicollinearity). The outliers may also affect the results of logistic regression.

16.1.1 Commands:

Analyze > Regression > Binary logistic > Put “diabetes” in the “Dependent” box > Put “sex_1”, age, pepticulcer, and f_history in the “Covariate” box > Categorical > Push “sex, pepticulcer and f_history” in “Categorical covariates” box > Select “f_history” > Select “first” under “Change contrast” (we are doing this because 0 is our comparison group) > Click on “Change” > (do the same thing for all the variables in “Categorical covariates” box) > Continue > Options > Select “Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals, Correlations of estimates, and CI for exp(B)” > Continue > OK

16.1.2 Outputs:

SPSS provides many tables while doing the logistic regression analysis. Only the useful tables are discussed here. After the basic tables (tables 16.1 to 16.3), the outputs of logistic regression are provided under Block 0 and Block 1.

A. Basic tables:

Table 16.1. Case Processing Summary

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	210	100.0
	Missing Cases	0	.0
	Total	210	100.0
Unselected Cases		0	.0
Total		210	100.0

a. If weight is in effect, see classification table for the total number of cases.

Table 16.2. Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Table 16.3. Categorical Variables Codings

		Frequency	Parameter coding
			(1)
Family history of DM	No	114	.000
	Yes	96	1.000
Peptic ulcer	Yes	59	1.000
	No	151	.000
Sex: numeric code	Female	133	.000
	Male	77	1.000

16.1.3 Interpretation: Basic tables

Table 16.1 (case processing summary) shows that the analysis includes all the 210 subjects and there is no missing value. If there is any missing data, this table will show it. Note that the subjects are excluded from analysis, if there are any missing data.

Table 16.2 (dependent variable encoding) tells us which category of the dependent variable (diabetes) is the predicted outcome. The higher value is the predicted outcome. Here, the higher value is 1 (have diabetes) and is the predicted outcome for the dependent variable.

Table 16.3 (categorical variables codings) indicates the comparison groups of the independent (explanatory) variables. Here, the lower value is the comparison group. For example, for family history of diabetes, “No” is coded as 0 (.000), while “Yes” is coded as 1 (1.000). This means that persons who do not have the family history of diabetes is the comparison group. Similarly, not having peptic ulcer (.000) and being female (.000) is the comparison group for peptic ulcer and sex, respectively.

B. Outputs under Block 0:

Table 16.4. Classification Table^{a,b}

	Observed		Predicted		
			DIABETES MELLITUS		Percentage Correct
			No	Yes	
Step 0	DIABETES MELLITUS	No	165	0	100.0
		Yes	45	0	.0
	Overall Percentage				78.6

a. Constant is included in the model.

b. The cut value is .500

16.1.4 Interpretation: Outputs under Block 0

Analysis of data without any independent variable in the model is provided under Block 0. The results indicate the baseline information that can be compared with the results when independent variables are put into the model (provided under Block 1).

Look at the classification table (table 16.4). The table indicates the *overall percentage* of correctly classified cases (78.6%). We will see whether this value has increased with the introduction of independent variables in the model under Block 1 (given in table 16.8). If the value remains the same, it means that the independent variables in the model do not have any influence/contribution to predict diabetes (dependent variable). In our example, the overall percentage has increased after inclusion of the independent variables in the model (90.5%; table 16.8 under Block 1) compared to the value under Block 0 (78.6%). This means that adding independent variables improved the ability of the model to predict the dependent variable.

C. Outputs under Block 1:

Table 16.5. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	101.589	4	.000
	Block	101.589	4	.000
	Model	101.589	4	.000

Table 16.6. Model Summary

Step 1	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	116.635 ^a	.384	.593

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Table 16.7: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14.663	8	.066

Table 16.8. Classification Table^a

	Observed		Predicted		
			Diabetes mellitus		Percentage Correct
			No	Yes	
Step 1	Diabetes mellitus	No	159	6	96.4
		Yes	14	31	68.9
	Overall Percentage				90.5

a. The cut value is .500

Table 16.9. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	age	.230	.039	34.660	1	.000	1.259	1.166	1.359
	sex(1)	1.587	.528	9.031	1	.003	4.891	1.737	13.774
	f_history	1.119	.517	4.688	1	.030	3.062	1.112	8.430
	pepticulcer	1.782	.487	13.363	1	.000	5.942	2.285	15.448
	Constant	-10.684	1.526	49.006	1	.000	.000		

a. Variable(s) entered on step 1: age, sex, f_history, pepticulcer.

Table 16.10. Correlation Matrix

		Constant	age	sex(1)	f_history	pepticulcer
Step 1	Constant	1.000	-.932	-.385	-.302	-.322
	age	-.932	1.000	.161	.054	.156
	sex(1)	-.385	.161	1.000	.388	.155
	f_history	-.302	.054	.388	1.000	.099
	pepticulcer	-.322	.156	.155	.099	1.000

Table 16.11. Casewise List^b

Case	Selected Status ^a	Observed	Predicted	Predicted Group	Temporary Variable	
		Diabetes mellitus			Resid	ZResid
11	S	Y**	.006	N	.994	13.213
25	S	Y**	.026	N	.974	6.177
38	S	Y**	.052	N	.948	4.248
41	S	Y**	.062	N	.938	3.883
62	S	Y**	.099	N	.901	3.009
124	S	Y**	.043	N	.957	4.692
137	S	N**	.890	Y	-.890	-2.839

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

16.1.5 Interpretation: Outputs under Block 1

Omnibus tests of Model Coefficients (table 16.5): This table indicates whether the overall performance of the model is better if independent variables are included in the model compared to the model without any independent variables (given under Block 0). We want this test to be significant ($p\text{-value} < 0.05$). In this example, the p -value of the Omnibus test is 0.000, which indicates that the proposed model is better than the model without any predictor (independent) variables.

Model summary table (table 16.6): This table indicates usefulness of the model. The Cox & Snell R-square and Nagelkerke R-square (called pseudo R-square) values provide indication about the amount of variation in the outcome variable that can be explained by the independent variables in the model. In this example, the values of the pseudo R-square are 0.384 (Cox & Snell R-square) and 0.593 (Nagelkerke R-square), respectively. This means that between 38.4% and 59.3% variation in the outcome variable can be explained by the independent variables set in the model. *This information is not needed if the objective of the analysis is to adjust for the Odds Ratio.*

Hosmer-lemeshow goodness-of-fit test (table 16.7): When the intention of analysis is prediction, i.e., to identify the predictors to predict the outcome, then the question is “How good is the model for prediction”? This is judged based on the Hosmer-Lemeshow goodness-of-fit test, and positive and negative predictive values, given in the classification table (table 16.8).

The Hosmer-Lemeshow test indicates how well the observed and predicted values fit with each other (i.e., observed and predicted probabilities match with each other). The null hypothesis is “the model fits” and the p -value is expected to be >0.05 (non-significant). If the p -value is not significant, it means that the model is a good fit for prediction (i.e., the observed and predicted values are close together). In this example, the p -value is 0.066, indicating that the model is useful for prediction of the outcome variable. If the test is significant ($p < 0.05$), then the model is not good to predict the outcome variable by the independent variables in the model. *Note that this information is not needed if the objective of doing logistic regression is to adjust for the confounding factors.*

Classification table (table 16.8): This table indicates how well the model is able

to predict the correct category in each case (have or do not have the disease). This table shows that the overall accuracy of this model to predict diabetes (with a predicted probability of 0.5 or greater) is 90.5%. This table also shows the *Sensitivity* and *Specificity* of the model as 68.9% ($31 \div 45$) and 96.4% ($159 \div 165$), respectively. *Positive and negative predictive values* can also be calculated from the table, which are 83.8% ($31 \div 37$) and 91.9% ($159 \div 173$), respectively. Interpretation of the findings of this table is a little bit complicated and needs further explanation, especially to explain sensitivity, specificity, and positive and negative predictive values.

However, the information that we need to check is the *overall percentage*. Compare this value with the value under the Block 0 outputs. We expect this value (overall percentage) to be increased, otherwise adding independent variables in the model does not have any impact on prediction. We can see that the overall percentage of the model to correctly classify cases is 90.5% under Block 1 (table 16.8). This value, compared to the value (78.6%; table 16.4) that we have seen under Block 0, has improved. This means that adding independent variables in the model improved the ability of the model to predict the dependent variable. *This information is needed, if the intention of this analysis is prediction. If the objective is adjustment for confounding factors, we can ignore this information.*

Variables in the equation (table 16.9): *This is the most important table to look at.* This table shows the results of logistic regression analysis. This table indicates how much each of the independent variables contributes to predict/explain the outcome variable. This table also indicates the adjusted Odds Ratio (OR) and its 95% confidence interval (CI). The B values (column 3) indicate the logistic regression coefficients for the variables in the model. These values are used to calculate the probability of an individual to have the outcome. The positive values indicate the likelihood for the outcome, while the negative values indicate the less likelihood for the outcome. The exponential of B [Exp(B)] is the adjusted OR.

Let us see how to interpret the results. There are 4 independent (explanatory) variables in the model – age (as a continuous variable), sex, family history of diabetes and peptic ulcer. The table shows the adjusted OR [Exp(B)], 95% CI for the adjusted OR and p-value (Sig.). The adjusted OR for sex is 4.891 (95% CI: 1.737-13.774), which is statistically significant ($p=0.003$). Here, our comparison group is female (see table 16.3). This indicates that males are 4.9 times more likely

to have diabetes compared to females after adjusting (or controlling) for age, family history of diabetes and peptic ulcer. Similarly, persons who have the family history of diabetes are 3.1 times more likely (OR: 3.06; 95% CI: 1.11-8.43; $p=0.03$) to have diabetes compared to those who do not have the family history after adjusting for age, sex, and peptic ulcer. Interpretation of $\text{Exp}(B)$ for age is a little bit different since the variable was entered as a continuous variable. Here, $\text{Exp}(B)$ for age is 1.259. This means that the odds of having diabetes would increase by 25.9% [$\text{Exp}(B) - 1$; i.e., $1.259 - 1$] (95% CI: 16.6-35.9) with each year increase of age, which is statistically significant ($p<0.001$).

If we want to know which variable contributed most in the model, then look at the Wald statistics. Higher the value (of Wald), the greater is the importance. Age is the most important variable contributed to the model since it has the highest Wald value (34.6).

Checking for multicollinearity (table 16.10): It is important to check for multicollinearity of the independent variables in the model. If there is multicollinearity, the model becomes dubious. Multicollinearity is checked in the correlation matrix table (table 16.10). This table shows the correlations between the independent variables (correlation coefficients or r values). If there is multicollinearity, r values would be higher (greater than 0.5). If we look at the correlation matrix table (table 16.10), none of the values are greater than 0.5 except for the correlation between age and constant, which is -0.932.

Now, look at table 16.9 (variables in the equation). If multicollinearity is present (and affects the model), the magnitude of the SEs (standard errors) would be high or low (greater than 5.0 or less than 0.001). Existence of multicollinearity means that the model is not statistically stable. To solve the problem (in general), look at the SE and omit the variable(s) with large (or small) SE, until the magnitude of the *SEs hover between 0.001 and 5.0*.

If there is a high correlation between the constant and any of the predictor variables, you can omit the constant from the model. In our example, there is a high correlation (-.932) between age and constant. However, it did not affect the results as none of the SEs are >5.0 or <0.001 . Therefore, we do not need to do anything. If it affects the results, just omit (deselect) the constant from the model (*deselect “include constant in model” located at the bottom of the fig 16.1*) from “Options template” during analysis as shown in fig 16.1.

Figure 16.1. Option template for logistic regression

Logistic Regression: Options

Statistics and Plots

☒ Classification plots ☒ Correlations of estimates

☒ Hosmer-Lemeshow goodness-of-fit ☐ Iteration history

☒ Casewise listing of residuals ☒ CI for exp(B): 95 %

☒ Outliers outside 2 std. dev. ☐ All cases

Display

☒ At each step ☐ At last step

Probability for Stepwise

Entry: 0.05 Removal: 0.10 Classification cutoff: 0.5

Maximum iterations: 20

☒ Include constant in model

Continue Cancel Help

Casewise list (table 16.11): This table provides information about the cases for which the model does not fit well. Look at the ZResid values (last column). The values above 2.5 are the outliers and do not fit well in the model. The case numbers are shown in the column 1. If present (cases that do not fit the model well), all these cases need to be examined closely. Under the “Predicted Group” column, you may see “Y (means yes)” or “N (means no)”. If it is “Y”, the model predicts that the case (case no. 137, in our example) should have diabetes, but in reality (in the data) the subject does not have diabetes (see the observed column where it is “N”). Similarly, is if it is “N” under the “predicted group”, the model predicts that the case should not have diabetes, but in reality the subject has diabetes.

16.1.6 ROC curve:

We can construct the ROC (Receiver Operating Characteristic) curve to assess the model discrimination. Area under the ROC curve ranges from 0 to 1. A value of 0.5 indicates that the model is useless. Our data shows that the ROC value (area under the curve) is 0.914 (95% CI: 0.86-0.96; $p=0.000$) (table 16.12). A high value of

ROC indicates good model for prediction. To generate the ROC curve, use the following commands:

Analyze > Regression > Binary logistic > Put “diabetes” in the “dependent” box > Put “sex_1”, age, pepticulcer, and f_history in the “Covariate” box > Categorical > Push “sex, pepticulcer and f_history” in “Categorical covariates” box > Select “f-history” > Select “First” under “Change contrast” (we are doing this because 0 is our comparison group) > Click on “Change” > (do the same thing for all the variables in “Categorical covariates” box) > Continue > Options > Select “Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals, Correlations of estimates, and CI for exp(B)” > Continue > **Save** > Select “Probabilities” under “Predicted values” > Continue > OK

This will create a new variable, **Pre_1** (predicted probability) (look at the bottom of the variable view). Now, to get the ROC curve, use the following commands:

Analyze > ROC curve > Select “Pre_1” for the “Test variable” box and “diabetes” for the “State variable” box and put “1” for the value of the state variable (since code 1 indicates individuals with diabetes) > Select “ROC curve” and “Standard error and Confidence interval” under “Display” > OK

Table 16.12. Area Under the Curve

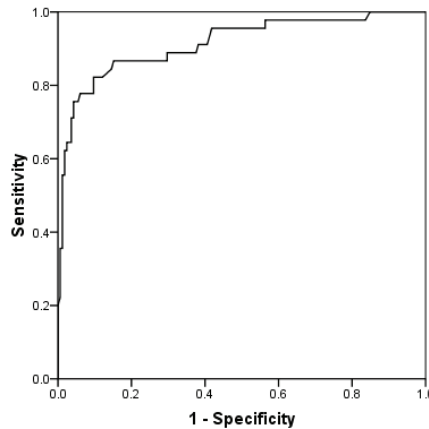
Area	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.914	.027	.000	.861	.967

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Figure 16.2. ROC Curve



Diagonal segments are produced by ties.

16.1.7 Sample size for logistic regression:

Sample size is always a concern for analysis of data. The sample size needed for logistic regression depends on the effect size you are trying to demonstrate and the variability of the data. It is always better to calculate the sample size during the design phase of the study. However, a rule-of-thumb for planning a logistic regression is that for every independent variable in the model you need to have at least 10 outcomes (some authors recommend minimum 15-25 cases for each independent variable).

16.1.8 Variable selection for a model:

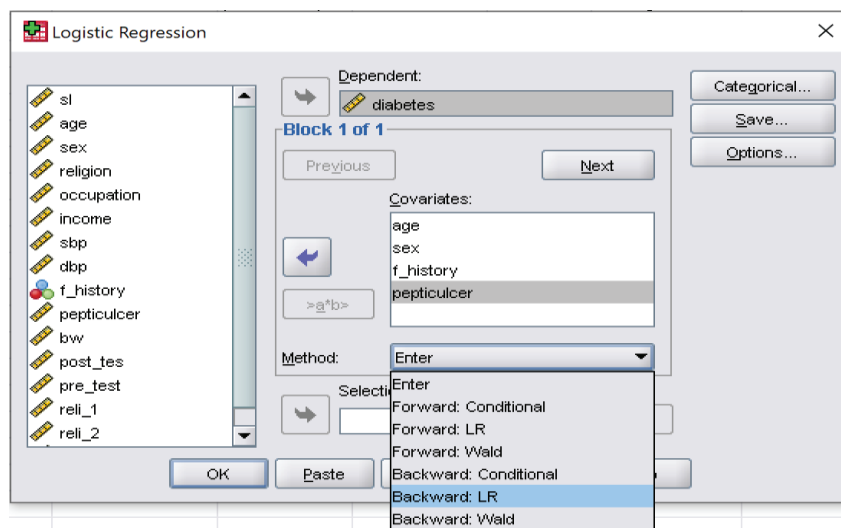
I have discussed various methods of variable selection for a model in the previous section in detail (section 15). Like other multivariable analysis, independent variables to be selected for logistic regression should include the risk factors of interest and potential confounders, while avoiding variables with lots of missing values.

Earlier in this section, I have used the “Enter” method for logistic regression analysis. The “Enter” method uses all the independent variables in the model included by the researcher. We can also use the automatic selection method for analysis. For logistic regression, the commonly used method for automatic selection of variables is the “*Backward LR*” method. However, if there is multicollinearity, you can select the “*Forward LR*” method for data analysis.

Commands for automatic selection of independent variables (use the data file <Data_3.sav>):

Analyze > Regression > Binary logistic > Put “diabetes” in the “Dependent” box > Put “sex_1”, age, pepticulcer, and f_history in the “Covariate” box > **Select “Backward LR” from “Method”** (fig 16.3) > Categorical > Push “sex, pepticulcer and f_history” in the “Categorical covariates” box > Select “f-history” > Select “First” under “Change contrast” (we are doing this because 0 is our comparison group) > Click on “Change” > (do the same thing for all the variables in “Categorical covariates” box) > Continue > Options > Select “Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals, Correlations of estimates, and CI for exp(B)” > Continue > OK

Figure 16.3



SPSS will give the following table (table 16.13) along with others (not shown here as they are not relevant). We can see that analysis is done in 4 steps (Step 1 to 4; the first column of the table). In the first step, all the independent variables are in the model. Gradually, SPSS has removed variables that are not significantly associated with the outcome. Finally, SPSS has provided the final model (Step 4) with a single variable (sex) in it, which is significantly associated with the outcome. If the "Enter" method was used, SPSS would give us only the step 1 (see table 16.9).

The inclusion (entry) and exclusion (removal) criteria, set as default in SPSS, are 0.05 and 0.10, respectively. As discussed in section 15, you can change the “Entry” and “Removal” criteria from the “Option” dialogue box (under the “Probability for Stepwise”). Finally, for model building, you should decide the variables to be included in the final model based on theoretical understanding and empirical findings.

Table 16.13. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	age	-.037	.023	2.502	1	.114	.964	.921	1.009
	sex_1(1)	-1.209	.388	9.707	1	.002	.299	.140	.639
	f_history(1)	-.425	.397	1.143	1	.285	.654	.300	1.425
	pepticulcer(1)	-.118	.430	.075	1	.784	.889	.383	2.064
	Constant	2.997	.737	16.529	1	.000	20.028		
Step 2 ^a	age	-.037	.023	2.660	1	.103	.963	.921	1.008
	sex_1(1)	-1.212	.388	9.772	1	.002	.298	.139	.636
	f_history(1)	-.428	.397	1.163	1	.281	.652	.299	1.419
	Constant	2.998	.737	16.530	1	.000	20.049		
Step 3 ^a	age	-.036	.023	2.521	1	.112	.964	.922	1.009
	sex_1(1)	-1.041	.347	8.980	1	.003	.353	.179	.698
	Constant	2.729	.686	15.830	1	.000	15.316		
Step 4 ^a	sex_1(1)	-.999	.344	8.457	1	.004	.368	.188	.722
	Constant	1.732	.243	50.954	1	.000	5.650		

a. Variable(s) entered on step 1: age, sex_1, f_history, pepticulcer.

Section 17

Survival Analysis

In many situations, researchers are interested to know the progress of a patient (with a disease) from a specific point in time (e.g., from the point of diagnosis or from the point of initiation of treatment) until the occurrence of a certain outcome, such as the death or recurrence of any event (such as, recurrence of cancer). Prognosis is usually assessed by: a) Estimating the *median survival time*, and b) *Cumulative probability of survival* after a certain time interval (e.g., 5-year, 3-year, etc.). For example, the researchers may be interested to know what is the median survival time of colonic cancer if the patient is not treated (or treated), and what is the estimated probability that a patient may survive for more than 5 years (5-year cumulative survival probability) if the patient is treated (or not treated). The methods employed to answer these questions in a follow-up study are known as survival analysis (or life table analysis) methods.

Survival analysis is done in the follow-up studies. To do the survival analysis, we need to have data (information) from each of the patients, at least on:

- *Time*: Length of time the patient was observed in the study (called survival time);
- *Outcome*: Whether the patient developed the outcome of interest (event) during the study period, or the patient was either lost to follow-up or remained alive at the end of the study (censored); and
- *Treatment group*: Which treatment (e.g., treatment A or B) did the patient receive in the study (optional)?

The survival time is of two types – a) Censored time; and b) Event time. The *censored* time is the amount of time contributed by:

- a. The patients who did not develop the outcome and remained in the study up to the end of the study period, or
- b. Patients who were lost to follow-up due to any reason, such as migration, withdraw, etc.; or
- c. Patients who developed outcome (e.g., died due to accident) due to other reasons than the disease of interest.

On the other hand, the event time is the amount of time contributed by the patients who developed the outcome of interest during the study period.

If we have the above information, it is possible to estimate the median survival times and cumulative survival probabilities for two or more treatment groups for comparison. Such a comparison allows us to answer the question “which treatment delays the time of occurrence of the event”. The method commonly used to analyze the survival-time data is the *Kaplan-Meier* method, and SPSS can be used to analyze such data. Use the data file <Data_survival_4.sav> for practice.

17.1 Survival analysis: Kaplan-Meier method

Suppose, a researcher has conducted a follow-up study (clinical trial) on patients with heart failure to compare the effectiveness of a new drug (n=22) compared to placebo (n=22). The outcome of interest in this study is death (event). The objective is to assess whether the “new treatment” delays the time to death (event) compared to placebo among the patients with heart failure. Following variables are included in the data file.

- *Time*: It is the amount of time each patient has spent in the study in days;
- *Treatment*: Which treatment did the patient receive (0= placebo; 1= new treatment);
- *Outcome (event)*: Whether the patient developed the event, i.e., died or not (0= censored; 1= died)

Assumptions:

- The probability of the outcome is similar among the censored and under-observation individuals;
- There is no secular trend over the calendar period;
- The risk is uniform during the interval;
- Losses are uniform over the interval.

17.1.1 Commands:

Analyze > Survival > Kaplan Meier > Push the variable “time” to “Time” box
> Push the variable “outcome” in the “Status” box > Click “Define event” >
Select “Single value” and type “1” (here 1 is the event) in the box > Continue
> Push “treatment” in the “Factor” box > Click Options... > Select “Survival

table(s), Mean and Median survival” under statistics > Select “Survival” under “Plots” > Continue > Click “Compare Factor...” > Select “Log rank” under “Test statistics” > Continue > OK

17.1.2 Outputs:

The SPSS will give the following outputs.

Table 17.1. Case Processing Summary

Treatment group	Total N	N of Events	Censored	
			N	Percent
Placebo	22	16	6	27.3%
New treatment	22	11	11	50.0%
Overall	44	27	17	38.6%

Table 17.2. Means and Medians for Survival Time

Treatment group	Mean ^a				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Placebo	72.545	14.839	43.462	101.629	40.000	12.899	14.719	65.281
New treatment	125.264	13.402	98.996	151.532	146.000	28.786	89.580	202.420
Overall	98.925	10.812	77.733	120.117	89.000	21.232	47.385	130.615

a. Estimation is limited to the largest survival time if it is censored.

Table 17.3. Survival Table

Treatment group		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
Placebo	1	2.000	Died	.955	.044	1	21
	2	3.000	Died	.909	.061	2	20
	3	4.000	Died	.864	.073	3	19
	4	7.000	Died	.818	.082	4	18
	5	10.000	Died	.773	.089	5	17
	6	22.000	Died	.727	.095	6	16
	7	28.000	Died	.682	.099	7	15
	8	29.000	Died	.636	.103	8	14
	9	32.000	Died	.591	.105	9	13
	10	37.000	Died	.545	.106	10	12
	11	40.000	Died	.500	.107	11	11
	12	41.000	Died	.455	.106	12	10
	13	54.000	Died	.409	.105	13	9
	14	61.000	Died	.364	.103	14	8
	15	63.000	Died	.318	.099	15	7
	16	71.000	Died	.273	.095	16	6
	17	127.000	Censored	.	.	16	5
	18	140.000	Censored	.	.	16	4
	19	146.000	Censored	.	.	16	3
	20	158.000	Censored	.	.	16	2
	21	167.000	Censored	.	.	16	1
	22	182.000	Censored	.	.	16	0
New treatment	1	2.000	Died	.955	.044	1	21
	2	6.000	Died	.909	.061	2	20
	3	12.000	Died	.864	.073	3	19
	4	54.000	Died	.818	.082	4	18
	5	56.000	Censored	.	.	4	17
	6	68.000	Died	.770	.090	5	16
	7	89.000	Died	.722	.097	6	15
	8	96.000	Died	.	.	7	14
	9	96.000	Died	.626	.105	8	13
	10	125.000	Censored	.	.	8	12
	11	128.000	Censored	.	.	8	11
	12	131.000	Censored	.	.	8	10
	13	140.000	Censored	.	.	8	9
	14	141.000	Censored	.	.	8	8
	15	143.000	Died	.547	.117	9	7
	16	145.000	Censored	.	.	9	6
	17	146.000	Died	.456	.129	10	5
	18	148.000	Censored	.	.	10	4
	19	162.000	Censored	.	.	10	3
	20	168.000	Died	.304	.151	11	2
	21	173.000	Censored	.	.	11	1
	22	181.000	Censored	.	.	11	0

Table 17.4. Overall Comparisons

	Chi-square	df	Sig.
Log Rank (Mantel-Cox)	4.660	1	.031
Test of equality of survival distributions for the different levels of Treatment group.			

Table 17.5. Overall Comparisons

	Chi-square	df	Sig.
Log Rank (Mantel-Cox)	4.660	1	.031
Breslow (Generalized Wilcoxon)	6.543	1	.011
Tarone-Ware	6.066	1	.014

Test of equality of survival distributions for the different levels of Treatment group.

17.1.3 Interpretation:

Table 17.1 is the summary table indicating the number of study subjects in each group (22 in the placebo and 22 in the new treatment group) and the number of events (no. died) occurred in each group including the number censored. The table shows that in the treatment group, 11 patients died and 11 were censored, while in the placebo group, 16 died and 6 censored.

Table 17.2 shows the mean and median survival times for both the placebo and new treatment groups. We do not consider the mean survival time for reporting. We consider the *median survival time*. The median survival time is the time when the cumulative survival probability is 50%. The table indicates that the median survival time, if the patient is in the placebo group, is 40 days (95% CI: 14.71-65.28), while it is 146 days (95% CI: 89.5-202.42), if the patient is in the new treatment group. This means that the new treatment increases the survival time, i.e., the new treatment is associated with longer time to event (and placebo is associated with shorter time to event). Thus, we conclude that the person lives longer if s/he receives the new treatment compared to the placebo.

Table 17.3 shows the survival probability (Cumulative Proportion Surviving at the Time) at different points of time in the placebo and treatment group. From the table, we can see that the cumulative survival probability at the end of 71 days, in the placebo group, is 0.273 (27.3%). Since there is no death after that, the cumulative survival probability at the end of 182 days will be the same (27.3%).

On the other hand, the cumulative survival probability is 0.304 (30.4%) at the end of 168 days, if the patient is in the new treatment group. As there is no death after that, the cumulative survival probability at the end of 181 days will be the

same (30.4%). In the new treatment group, the cumulative survival probability at the end of 71 days is about 0.722 (72.2%), which is much higher than in the placebo group (27.3%). This indicates that the probability of survival at the end of 71 days is higher among the patients who received new treatment compared to placebo. This may indicate the benefit of the new treatment (i.e., the new treatment is better than the placebo).

However, if we consider the cumulative survival probability of patients in both these groups at the end of 180 days, the outcome is not that different – 27.3% in the placebo group and 30.4% in the treatment group. This information indicates that the survival probability is still higher if the person is on the new treatment than on the placebo.

We can also estimate the median survival time (it is the time when the cumulative survival probability is 50%) in both these groups from this table. The median survival time for placebo group is 40 days and that of the treatment group is 146 days (see table 17.3). Now, the question is whether the survival experiences of both these groups in the population are different or not? To answer this question, we have to use a statistical test (Log Rank test) as given in table 17.4.

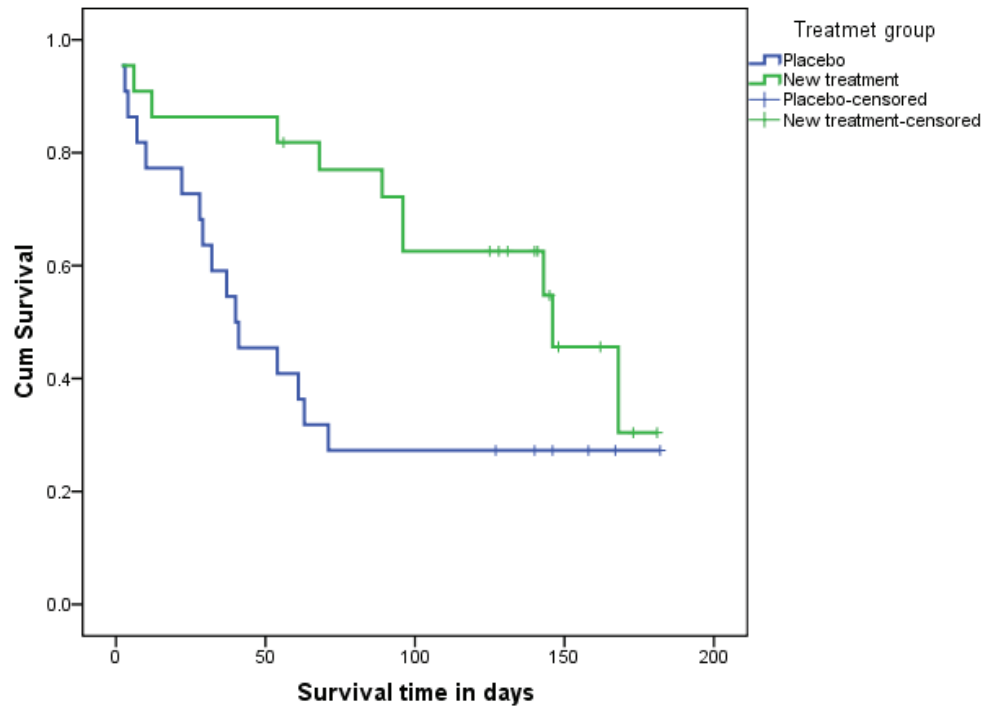
Table 17.4 shows the *Log Rank* test results. For an objective comparison of the survival experience of two groups, it is desirable to use some statistical methods that would tell us whether the difference of the survival experiences in the population is statistically significant or not. Here, the null hypothesis is “there is no difference in the survival experience of these two groups (new treatment and placebo) in the population”. Such a null hypothesis is tested by Log Rank test. The *Log Rank* test results show that the p-value is 0.031, which is <0.05 . This means that, the survival experience of both these groups in the population is not same. In other words, it indicates that the survival probability is better if the patient receives the new treatment (i.e., the new treatment is more effective/better than the placebo in improving the patients’ survival).

Note that there are alternative procedures for testing the null hypothesis that the two survival curves are identical. They are *Breslow test*, *Tarone-Ware test* and *Peto test* (table 17.5). The Log Rank test ranks all the deaths equally, while the other tests give more weight to early deaths. The options are available in SPSS under the “Compare Factor” tab.

Survival curve: The cumulative survival probability is usually portrayed visually

by a graph called survival curve (fig 17.1). The “steps” in the graph represent the time when events (deaths or any other event of interest) occurred. The graph allows us to represent visually the median survival time and the cumulative survival probability for any specific time period (e.g., 30-day; 6-month; 1-year, 3-year, 5-year, 10-year cumulative survival probability, etc.). In general, the line above, indicates better survival probability. We can see that the line for the new treatment is above the line for the placebo.

Figure 17.1. Survival Functions



Section 18

Cox Regression

The Cox Regression is also called *Proportional Hazards Analysis*. In the previous section (section 17), I have discussed the survival analysis using the Kaplan Meier method. Like other regression methods (e.g., multiple linear regression and logistic regression), Cox Regression is a multivariable analysis technique where the dependent measure is a mixture of time-to-event and censored time observations. Use the data file <Data_survival_4.sav> for practice.

18.1 Cox Regression or Proportional Hazards Regression

Returning to our previous example (section 17), we analyzed data to assess the effectiveness of a new treatment compared to the placebo. Our objective was to determine whether the new treatment delays the time to death compared to the placebo among patients with heart failure. We found that the new treatment significantly delayed the time to death compared to placebo, as indicated by the median survival time and Log Rank test. However, the effectiveness of the new treatment might be influenced (confounded) by other factors, such as age, hypertension, diabetes or other characteristics. All these variables, therefore, need to be controlled during analysis for assessing the effectiveness of the new treatment. Cox Regression is a statistical method that is used to control the confounding factors (categorical, continuous or discrete covariates) that may influence the effectiveness of the new treatment.

Cox Regression gives us the *Hazard Ratio*, which is analogous to Relative Risk (RR). Hazard Ratio (also called Relative Hazard) is the ratio of hazards if the person is exposed compared to the person not exposed. In Cox Regression, the dependent variable is the Log of hazard.

18.1.1 Commands:

Let us use the previous example and data for Cox Regression analysis along with the variables sex and age for adjustment. Note that the variable “treatment” has two categories – placebo (coded as “0”) and new treatment (coded as “1”).

Analyze > Survival > Cox Regression > Push “time” into the “Time” box >

Push “outcome” into the “Status” box > Click on “Define event” > In “Single value” box write 1 (since 1 is the code no. of the event) > Continue > Push “treatment, age and sex” into the “Covariate” box > Click on “Categorical” > Push “treatment and sex” into the “Categorical Covariates” box > Select “Last” from “Reference category” (usually the default) under “Change contrast” > Continue > Click on “Options” > Select “CI for Exp(B) and Correlation of estimates” > Continue > Click on “Plot” > Select “Survival and Log minus Log” > Select the variable “treatment” from the “Covariate Values Plotted at” and push into the “Separate Line for” box > Continue > Ok

18.1.2 Outputs:

Only relevant tables are provided below.

Table 18.1. Case Processing Summary

		N	Percent
Cases available in analysis	Event ^a	27	61.4%
	Censored	17	38.6%
	Total	44	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		44	100.0%
a. Dependent Variable: Survival time in days			

Table 18.2. Categorical Variable Codings^{c,d}

		Frequency	(1) ^b
Treatment	0=Placebo	22	1
	1=New treatment	22	0
sex ^a	0=Male	21	1
	1=Female	23	0

a. Indicator Parameter Coding

b. The (0,1) variable has been recoded, so its coefficients will not be the same as for indicator (0,1) coding.

c. Category variable: treatment (Treatment gr)

d. Category variable: sex (Sex)

Table 18.3. Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
							Lower	Upper
treatment	1.004	.458	4.805	1	.028	2.728	1.112	6.693
age	.009	.023	.150	1	.698	1.009	.965	1.055
sex	.889	.436	4.169	1	.041	2.434	1.036	5.716

Figure 18.1. Survival Function of heart failure patients

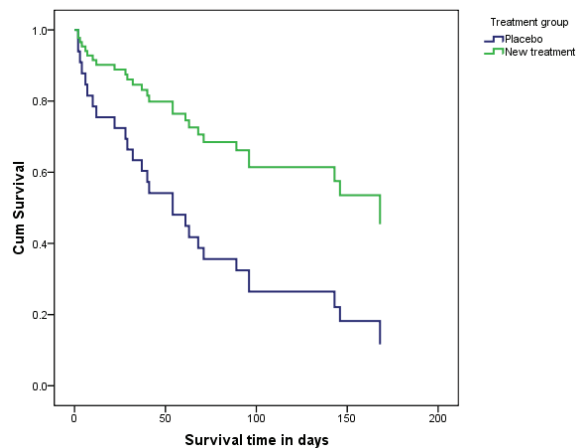
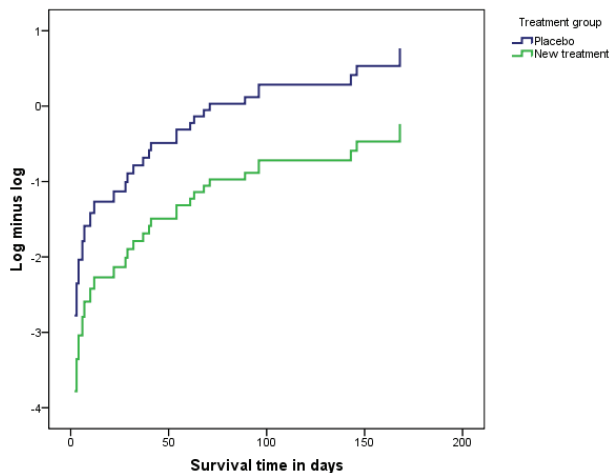


Figure 18.2. LML (Log minus Log) Function



18.1.3 Interpretation:

Table 18.1 shows the number of cases that are analyzed. Table 18.2 is *very* important for interpretation. This table indicates which category of the categorical vari-

ables is the comparison group. Look at the last column [(1)^b]. The value “0” in this column indicates the comparison group. In the table “New treatment” is indicated as “0” in the last column. Therefore, in the analysis, “new treatment” is the comparison group (though “new treatment” is actually coded as “1”). Similarly, “females” are the comparison group in this analysis since the value of being “female” is “0” in the last column. We shall, therefore, get the Hazard Ratio for the “placebo” group compared to the “new treatment” group and for the “males” compared to the “females”, as shown in table 18.3 (Variables in the Equation).

Our main interest is in table 18.3 (Variables in the Equation). The table indicates the *Hazard Ratio* [Exp(B)], p-value (Sig.) and 95% confidence interval (CI) for the Hazard Ratio [95% CI for Exp(B)]. The Hazard Ratio for the variable “treatment” is 2.72 (95% CI: 1.11-6.69) and the p-value is 0.028. This indicates that compared to the “new treatment”, patients in the “placebo” group are 2.72 times more likely to *have shorter time to event* after controlling for “age” and “sex”, which is statistically significant (p=0.028). On the other hand, males are more likely (2.43 times) to have shorter time to event compared to the females after controlling for the variables “treatment” and “age” (p=0.041). Age, independently, does not have any significant effect on the survival time, since the p-value is 0.698.

Figure 18.1 shows the survival plot of the heart failure patients by treatment group. The upper line is for the new treatment group and the lower one is for the placebo group. The figure shows the outcome difference between the new treatment and placebo. The group represented by the upper line has the better survival probability.

However, before we conclude the results, we have to check if: a) there is multicollinearity among the independent variables; and b) relative hazards over the time are proportional (also called the proportionality assumption of the proportional hazards analysis). Look at the SE of the variables in the model (table 18.3). There is no value which is very small (<0.001) or very large (>5.0) (refer to the logistic regression analysis in section 16), indicating that there is no problem of multicollinearity in the model.

For the second assumption, we need to check the *log-minus-log* survival plot (fig 18.2). If there is a constant vertical difference between the two curves (i.e., curves are parallel to each other), it means that the relative hazards over time are proportional. If the curves cross each other, or are much closer together at some

points in time and much further apart at other points in time, then the assumption is violated. In our example, two lines are more or less parallel indicating that the assumption is not violated. When the proportional hazard assumption is violated, it is recommended to use the Cox regression with time dependent covariate to analyze the data.

Section 19

Non-parametric Methods

Non-parametric tests, in general, are done when the quantitative dependent variable is not normally distributed. Non-parametric tests are also used when the data are measured in nominal and ordinal scales. Table 19.1 shows the types of non-parametric methods recommended against the parametric tests, when the dependent variable is not normally distributed in the population. Note that non-parametric tests are less sensitive compared to the parametric tests and may, therefore, fail to detect differences between groups that actually exist. Use the data file <Data_3.sav> for practice.

Table 19.1. Types of non-parametric techniques against the alternative parametric methods

Non-parametric test	Alternative parametric test
Mann-Whitney U test	Independent-samples t-test
Wilcoxon Signed Ranks test	Paired t-test
Kruskal-Wallis test	One-way ANOVA
Friedman test	One-way repeated measures ANOVA
Chi-square test for goodness-of-fit	None
Chi-square test for independence (discussed earlier)	None
Spearman's correlation (discussed earlier)	Pearson correlation

19.1 Mann-Whitney U test

This test is the alternative test for Independent Samples t-test, when the dependent variable is not normally distributed. This test compares the differences between two groups on a continuous measure (variable). This test is based on ranks of observations and is better than the median test. This test, tests the null hypothesis that the two populations have equal medians. For example, we may want to know whether the median systolic BP (where the distribution of systolic BP is non-normal) of males and females is same.

19.1.1 Commands:

Analyze > Nonparametric tests > 2 Independent samples > Select “sbp” and push into the “Test Variable List” box > Select “sex_1” and push into the “Grouping Variable” box > Click on “Define Groups” > Write 0 in “Group1” box and 1 in “Group 2” box (note: our code nos. are 0 for female and 1 for male) > Continue > Select “Mann-Whitney” under “Test Type” > Ok

19.1.2 Outputs:

Table 19.2. Mann-Whitney: Ranks

	Sex: numeric	N	Mean Rank	Sum of Ranks
Systolic BP	Female	133	109.90	14616.50
	Male	77	97.90	7538.50
	Total	210		

Table 19.3. Test Statistics^a

	Systolic BP
Mann-Whitney U	4535.500
Wilcoxon W	7538.500
Z	-1.379
Asymp. Sig. (2-tailed)	.168

a. Grouping Variable: Sex: numeric

19.1.3 Interpretation:

Our interest is in table 19.3. Just look at the p-value of the test. Here, the p-value is 0.168, which is >0.05 . This indicates that the distribution of systolic BP among males and females is not different (or median of systolic BP of males and females is not different). However, with this test result, the median systolic BP of females and males should be reported. To get the *medians*, use the following commands.

Analyze > Compare means > Means > Select “sbp” and push into the “Dependent List” box > Select “sex_1” and push it into the “Independent List” box > Remove “Mean, Number of cases and Standard deviation” from the “Cell Statistics” box > Select “Median” from “Statistics” box and push it into the “Cell Statistics” box > Continue > OK

You will get the following tables (19.4 & 19.5). Tables 19.5 shows the median systolic BP of males and females.

Table 19.4. Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Systolic BP * Sex: numeric	210	100.0%	0	.0%	210	100.0%

Table 19.5. Report

Median	
Sex: numeric	Systolic BP
Female	124.00
Male	122.00
Total	123.00

19.2 Wilcoxon Signed Ranks test

This test is the non-parametric alternative of the paired samples t-test. This test compares the distribution of two related samples (e.g., pre-test and post-test). Wilcoxon test converts the scores into ranks and then compares. Suppose, to evaluate the impact of a training, you have taken the pre- and post-tests, before and after the training. You want to assess if there is any change in the post-test score compared to the pre-test score due to the training.

19.2.1 Commands:

Analyze > Nonparametric tests > 2 Related Samples > Select “post-test and pre-test” together and push into the “Test Pairs” box > Options > Select “Descriptive” and “Quartile” > Continue > Select “Wilcoxon” under “Test Type” > Ok

19.2.2 Outputs:

Table 19.6. Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Post test score	32	90.9844	8.44096	62.00	100.00	87.0000	92.5000	98.1250
Pre test score	32	53.5781	15.42835	23.50	87.50	41.8750	52.0000	64.5000

Table 19.7. Test Statistics^b

	Pre test score - Post test score
Z	-4.938 ^a
Asymp. Sig. (2-tailed)	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

19.2.3 Interpretation:

Table 19.6 shows the descriptive statistics of pre- and post-test scores. The median (50th percentile) score of pre-test is 52.0, while the median score is 92.5 for the post-test. The difference between these scores is quite big. Look at table 19.7. The p-value of the Wilcoxon Signed Ranks test is 0.000, which is highly significant. This indicates that the pre- and post-test scores (medians) are significantly different. We, therefore, conclude that the training has significantly improved the knowledge of the participants (since the median of the post-test score is significantly higher than that of the pre-test score).

19.3 Kruskal-Wallis test

It is the non-parametric equivalent of the one-way ANOVA test. In this test, scores are converted into ranks and the mean rank of each group is compared. Suppose, we want to test the hypothesis whether the systolic BP is different among religious groups (Muslim, Hindu and Christian) [the null hypothesis is systolic BP is not different across the religious groups].

19.3.1 Commands:

Analyze > Nonparametric Tests > K Independent Samples > Select “sbp” and push into “Test Variable List” box > Select “religion” and push it into “Grouping Variable” box > Click “Define Range” > Write 1 in “Minimum” box and 3 in “Maximum” box (the religion has code numbers from 1 to 3) > Continue > Options > Select “Quartile” > Continue > Select “Kruskal Wallis H” > Ok

19.3.2 Outputs:

Table 19.8. Ranks

	Religion	N	Mean Rank
Systolic BP	MUSLIM	126	105.54
	HINDU	58	106.47
	Christian	26	103.13
	Total	210	

Table 19.9. Test Statistics^{a,b}

	Systolic BP
Chi-Square	.054
df	2
Asymp. Sig.	.973

a. Kruskal Wallis Test

b. Grouping Variable: religion

19.3.3 Interpretation:

Table 19.9 shows the Kruskal Wallis test results (dependent variable is systolic BP and grouping variable is religion with 3 levels – Muslim, Hindu and Christian as shown in table 19.8). The p-value (Asymp. Sig.) of the Chi-square test is 0.973, which is >0.05 . Therefore, we are unable to reject the null hypothesis. We conclude that the median of the systolic BP among religious groups is not significantly different. You can get the median of the systolic BP using the commands as mentioned under Mann-Whitney U test. The medians of systolic BP in different religious groups are provided in table 19.10.

Table 19.10. Report

Median	
Religion	Systolic BP
MUSLIM	122.00
HINDU	126.00
Christian	121.50
Total	123.00

19.4 Friedman test

The Friedman test is the non-parametric alternative of the one-way repeated measures ANOVA test. Suppose, we are interested to evaluate the changes in blood sugar levels (if they are different or not) at four different time intervals (e.g., at

hour 0, hour 7, hour 14 and hour 24) after administration of a drug. To conduct this study, we have selected 15 individuals randomly from a population and measured their blood sugar levels at the baseline (hour 0). All the individuals were then given the drug, and their blood sugar levels were measured again at hour 7, hour 14 and hour 24. The blood sugar levels at hour 0, hour 7, hour 14, and hour 24 are named in SPSS as Sugar_0, Sugar_7, Sugar_14 and Sugar_24, respectively. Use the data file <Data_Repeat_anova_2.sav> for exercise.

19.4.1 Commands:

Analyze > Nonparametric Tests > K Related Samples > Select “sugar_0, sugar_7, sugar_14 and sugar_24” and push them into “Test Variables” box > Statistics > Select “Quartile” > Continue > Select “Friedman” > Ok

19.4.2 Outputs:

Table 19.11. Descriptive Statistics

	N	Percentiles		
		25th	50th (Median)	75th
Blood sugar at hour 0	15	106.0000	110.0000	115.0000
Blood sugar at hour 7	15	100.0000	105.0000	110.0000
Blood sugar at hour 14	15	96.0000	100.0000	107.0000
Blood sugar at hour 24	15	95.0000	98.0000	110.0000

Table 19.12. Ranks

	Mean Rank
Blood sugar at hour 0	3.80
Blood sugar at hour 7	2.73
Blood sugar at hour 14	1.63
Blood sugar at hour 24	1.83

Table 19.13. Test Statistics^a

N	15
Chi-Square	27.562
df	3
Asymp. Sig.	.000

a. Friedman Test

19.4.3 Interpretation:

Outputs are provided in tables 19.11-19.13. Table 19.11 shows the median blood

sugar levels at 4 different time periods. Look at the Friedman test results as provided in table 19.13. The Chi-square value is 27.56 and the p-value (Asymp. Sig.) is 0.000, which is <0.05 . This indicates that there is a significant difference in blood sugar levels across the 4 time periods ($p<0.001$). The findings indicate that the drug is effective in reducing the blood sugar levels.

19.5 Chi-square test for goodness-of-fit

The Chi-square test of independence, which is most frequently used to determine the association between two categorical variables, has been discussed in section 13. The Chi-square test for goodness-of-fit is also referred to as one-sample chi-square test. It is often used to compare the proportion of cases with a hypothetical proportion. Suppose, we have conducted a survey taking a random sample from a population, and the data show that the prevalence of diabetes is 21.4%. Now, we want to test the hypothesis, whether the prevalence of diabetes in the population is 18% or not (the null hypothesis is “the prevalence of diabetes in the population is 18%”). To have the answer, we shall do the Chi-square test for goodness-of-fit (seldom we test such a hypothesis).

19.5.1 Commands:

Analyze > Nonparametric tests > Chi-square > Move the variable “diabetes” into the “Test Variable List” box > Select “Values” under “Expected Values” > Write “0.18” in the box > Add > Again write “0.72” (1 minus 0.18) in the box > Add > Click on “Options” > Select “Descriptive” > Continue > Ok

19.5.2 Outputs:

Table 19.14. Diabetes mellitus

	Observed N	Expected N	Residual
Yes	45	42.0	3.0
No	165	168.0	-3.0
Total	210		

19.15. Test Statistics

	Diabetes mellitus
Chi-Square	.268 ^a
df	1
Asymp. Sig.	.605

a. 0 cells (.0%) have expected frequencies less than 5.
The minimum expected cell frequency is 42.0.

19.5.3 Interpretation:

Table 19.14 provides the observed and expected frequencies for those who have diabetes (as “Yes”) and those who do not have diabetes (as “No”). These are the descriptive information, and you do not need to report them. Table 19.15 is the main table to interpret the results. Our interest is at the p-value. The Chi-square value is 0.268 and the p-value is 0.605. Since the p-value is >0.05 , we cannot reject the null hypothesis. This means that the prevalence of diabetes in the population may not be different from 18%.

Section 20

Checking Reliability of Scale: Cronbach's Alpha

When the researchers select a scale (e.g., a scale to measure depression) in their study, it is important to check that the scale is reliable. One of the ways to check the internal consistency (reliability) of the scale is to calculate the Cronbach's alpha coefficient. Cronbach's alpha indicates the degree to which the items in the scale correlate with each other in the group.

Ideally, the Cronbach's alpha value should be above 0.7. However, this value is sensitive to the number of items in the scale. If the number of items in the scale is less than 10, the Cronbach's alpha coefficient tends to be low. In such a situation, it is more appropriate to use the "*mean inter-item correlations*". The optimum range of the mean inter-item correlation is between 0.2 and 0.4. Use the data file <Data_cronb.sav> for practice.

20.1 Cronbach's alpha

Before doing the procedure to get the Cronbach's alpha coefficient, be sure that all the negatively worded values are "reversed" by recoding. If this is not done, it would produce very low (or negative) value of the Cronbach's alpha coefficient. Suppose, we have used a scale to measure the depression. The scale has 4 questions, q1, q2, q3 and q4. To get the Cronbach's alpha coefficient, use the following commands:

Analyze > Scale > Reliability analysis > Select all the items (q1, q2, q3 & q4) that construct the scale and push them into the "Items" box > Make sure that "Alpha" is selected in "Model" section (it is usually the default) > Type name of the scale (e.g., depression or any other name suitable for the data) in the "Scale label" box > Statistics > Select "Item, Scale, & Scale if item deleted" under "Descriptives for" section > Select "Correlations" under "Inter-item" section > Select "Correlation" under "Summaries" section > Continue > Ok

20.1.1 Outputs:

Table 20.1. Case Processing Summary

		N	%
Cases	Valid	60	100.0
	Excluded ^a	0	.0
	Total	60	100.0

a. Listwise deletion based on all variables in the procedure.

Table 20.2. Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.839	.840	4

Table 20.3. Item Statistics

	Mean	Std. Deviation	N
q1	3.1333	1.08091	60
q2	3.2667	1.00620	60
q3	3.0167	1.08130	60
q4	3.2833	1.13633	60

Table 20.4. Inter-Item Correlation Matrix

	q1	q2	q3	q4
q1	1.000	.512	.491	.548
q2	.512	1.000	.635	.630
q3	.491	.635	1.000	.589
q4	.548	.630	.589	1.000

Table 20.5. Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Inter-Item Correlations	.567	.491	.635	.143	1.292	.003	4

Table 20.6. Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
q1	9.5667	7.741	.600	.364	.827
q2	9.4333	7.572	.711	.518	.781
q3	9.6833	7.373	.677	.476	.794
q4	9.4167	6.993	.705	.500	.781

Table 20.7. Scale Statistics

Mean	Variance	Std. Deviation	N of Items
12.7000	12.519	3.53817	4

20.1.2 Interpretation:

The Reliability Statistics table (table 20.2) shows the Cronbach's alpha value. In this example, the value is 0.839. This indicates very good correlation among the items in the scale (scale is reliable).

However, before looking at the value of Cronbach's alpha, look at the table "Inter-item Correlation Matrix" (table 20.4). All the values in the table should be *positive* (all are positive in our example). One or more negative values (if there is any) indicate that some of the items have not been "reverse scored" correctly. This information is also provided in the table "Item-total Statistics (table 20.6)". All the values under "Corrected item - Total Correlation" should be positive (there should not have any negative value).

The corrected item-total correlation (table 20.6) indicates the degree to which each item correlates with the total score. In our example, the values are 0.60, 0.71, 0.67 and 0.70. Low value for any item (<0.3) could be a problem. If the Cronbach's alpha value (<0.7) (table 20.2) and the corrected item-total correlation value (<0.3) is low, one may consider omitting the item from the scale with low value. In our example, there is no such problem.

However, if the number of items is small in the scale (fewer than 10), it may be difficult to get a reasonable Cronbach's alpha value. In such a situation, report the *Mean Inter-item Correlation* value (Summary-item Statistics table; table 20.5). In this example, the Inter-item Correlation values range from 0.491 to 0.635, and the mean is 0.567 (optimum range of the mean is 0.2 to 0.4). This indicates a strong relationship among the items.

Section 21

Analysis of Covariance (ANCOVA): One-way ANCOVA

ANCOVA stands for Analysis of Covariance, which is done to statistically control the extraneous variable(s) [called covariate] for comparison of the mean of two or more groups. It is similar to ANOVA. In ANOVA, one can incorporate only the categorical independent variables to have the main effect and interaction. But in ANCOVA, one can incorporate both the categorical and quantitative variables in the model, including the interaction between the categorical and quantitative independent variables. The ANCOVA can be performed as One-way, Two-way and Multivariate ANCOVA techniques. Use the data file <Data_3.sav> for practice.

21.1 One-way ANCOVA

The purpose of doing the one-way ANCOVA test is to understand the differences of the means of the dependent variable (e.g., systolic BP) against a categorical variable (e.g., sex, or effect of drugs, etc.) after controlling for the quantitative variable(s) [called covariates, such as age, diastolic BP, etc.] in the model. The one-way ANCOVA test involves at least three variables:

- One quantitative *dependent* variable (e.g., systolic BP, post-test score, blood sugar level, etc.);
- Only one categorical *independent* variable with two or more levels (e.g., sex, type of intervention, or type of drug, etc.); and
- One (or more) *covariate* (continuous quantitative variable), e.g., diastolic BP, age, pre-test score, baseline blood sugar level, etc.

The covariates to be selected for the model should be one or more continuous variable(s) and they should significantly correlate with the dependent variable. One can also include categorical variables as covariate in the model.

Suppose, the researcher is interested to compare the effectiveness of 3 drugs (drug A, drug B and drug C) in reducing the systolic BP. To conduct the study, the researcher has randomly selected three groups of people and assigned these drugs, one in each group. In such a situation, one-way ANOVA could be done. However, it was observed that the mean age and pre-treatment systolic BP of these three groups are not same. Since age and pre-treatment systolic BP can influence the

effectiveness of the drugs in reducing the systolic BP, it requires adjustment for these variables (here, age and pre-treatment systolic BP are the covariates) to conclude the results. In such a situation, one-way ANCOVA can be used. Note that for ANCOVA, the independent variable must be a categorical variable (here it is “type of drug”). ANCOVA can adjust for more than one covariate, either continuous or categorical.

Another example, suppose you have organized a training. To evaluate the effectiveness of the training, you have taken the pre- and post-tests of the participants. Now, you want to conclude if males and females (independent variable) have similar performance in the post-test (dependent variable), after controlling for age and pre-test results (covariates). One-way ANCOVA is the appropriate test for both these situations, if the assumptions are met.

Hypothesis:

Suppose, you want to assess if the mean systolic BP (dependent variable) is same among males and females (independent variable) after controlling for diastolic BP (covariate).

H_0 : There is no difference of the mean systolic BP between males and females in the population (after controlling for diastolic BP).

H_A : The mean systolic BP of males and females is different in the population.

Assumptions:

1. The dependent variable is normally distributed at each level of the independent variable;
2. The variances of the dependent variable for each level of the independent variable are same (homogeneity of variance);
3. The covariates (if more than one) are not strongly correlated with each other ($r < 0.8$);
4. There is a linear relationship between the dependent variable and the covariates at each level of the independent variable;
5. There is no interaction between the covariate (diastolic BP) and the independent variable (sex) [called *homogeneity of regression slopes*].

21.1.1 Commands:

A. Homogeneity of regression slopes (Assumption 5): First, we shall have to check the *homogeneity of regression slopes*, using the following commands. Note that, the SPSS variable names for sex is “sex_1 (0= female; 1= male)”, Systolic BP is “sbp” and Diastolic BP is “dbp”.

Analyze > General linear model > Univariate > Push “sbp” into the “Dependent Variables” box > Push “sex_1” into the “Fixed Factor” box > Push “dbp” in the “Covariate box” > Click Model > Select “Custom” under “Specify model” > Confirm that interaction option is showing in the “Build Terms” box > Push “sex_1” and “dbp” into the “Model” box > Click on “sex_1” in “Factors & Covariates” box > Pressing the control button click on the “dbp” in “Factors & Covariates” box > Push them into the “Model” box (you will see “dbp*-sex_1” in the Model box) > Continue > Ok

21.1.2 Outputs: Homogeneity of regression slopes

Table 21.1. Between-Subjects Factors

		Value Label	N
Sex_1	0	Female	133
	1	Male	77

Table 21.2. Tests of Between-Subjects Effects

Dependent Variable: SYSTOLIC BP					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	61902.160 ^a	3	20634.053	194.296	.000
Intercept	385.931	1	385.931	3.634	.058
Sex_1	7.019	1	7.019	.066	.797
Dbp	41714.418	1	41714.418	392.795	.000
Sex_1 * dbp	17.964	1	17.964	.169	.681
Error	21877.006	206	106.199		
Total	3515465.000	210			
Corrected Total	83779.167	209			
a. R Squared = .739 (Adjusted R Squared = .735)					

21.1.3 Interpretation: Homogeneity of regression slopes

Only look at table 21.2 (tests of between-subjects effects). Our interest is on the significance (Sig.) of the interaction (Sex_1*dbp). We can see that the p-value for

interaction is 0.681, which is >0.05 . This indicates that the *homogeneity of regression slopes* assumption is not violated. A p-value of <0.05 indicates that the regression slopes are not homogeneous and the ANCOVA test is inappropriate.

B. One-way ANCOVA:

To perform the one-way ANCOVA, use the following commands:

Analyze > General linear model > Univariate > Push “sbp” into the “Dependent Variables” box > Push “sex_1” into the “Fixed Factor” box > Push “dbp” in the “Covariate” box > Click “Model” > Select “Full Factorial” > Continue > Options > Select “sex_1” and push it into the “Display Means for” box (this would provide the adjusted means) > Select “Compare main effects” > Select “Bonferroni” from “Confidence interval adjustment” > Select “Descriptive Statistics, Estimates of effect size, and Homogeneity” tests under “Display” > Continue > Ok

21.1.4 Outputs: One-way ANCOVA

Table 21.3. Between-Subjects Factors

		Value Label	N
Sex	0	Female	133
	1	Male	77

Table 21.4. Descriptive Statistics

Dependent Variable: SYSTOLIC BP			
Sex	Mean	Std. Deviation	N
Female	129.73	21.309	133
Male	124.56	17.221	77
Total	127.83	20.021	210

Table 21.5. Levene's Test of Equality of Error Variances^a

Dependent Variable: SYSTOLIC BP			
F	df1	df2	Sig.
.365	1	208	.546
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.			

a. Design: Intercept + dbp + sex_1

Table 21.6. Tests of Between-Subjects Effects

Dependent Variable: SYSTOLIC BP						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	61884.196 ^a	2	30942.098	292.534	.000	.739
Intercept	688.574	1	688.574	6.510	.011	.030
Dbp (diastolic BP)	60580.272	1	60580.272	572.740	.000	.735
Sex_1	143.945	1	143.945	1.361	.245	.007
Error	21894.971	207	105.773			
Total	3515465.000	210				
Corrected Total	83779.167	209				
a. R Squared = .739 (Adjusted R Squared = .736)						

Table 21.7. Estimated Marginal

Sex				
Dependent Variable: SYSTOLIC BP				
Sex	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Female	127.191 ^a	.898	125.421	128.962
Male	128.942 ^a	1.186	126.604	131.281

a. Covariates appearing in the model are evaluated at the following values:
 DIASTOLIC BP = 83.04.

Table 21.8. Pairwise Comparisons

Pairwise Comparisons						
Dependent Variable: Systolic BP						
(I) Sex: numeric	(J) Sex: numeric	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval Difference ^a	
					Lower Bound	Upper Bound
Female	Male	-1.751	1.559	.263	-4.825	1.322
Male	Female	1.751	1.559	.263	-1.322	4.825

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Table 21.9. Pairwise Comparisons

Dependent Variable:Systolic BP			Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
(I) Religion	(J) Religion	Mean Difference (I-J)			Lower Bound	Upper Bound
MUSLIM	HINDU	-.448	1.705	1.000	-4.562	3.666
	CHRISTIAN	.948	2.313	1.000	-4.635	6.532
HINDU	MUSLIM	.448	1.705	1.000	-3.666	4.562
	CHRISTIAN	1.397	2.535	1.000	-4.721	7.514
CHRISTIAN	MUSLIM	-.948	2.313	1.000	-6.532	4.635
	HINDU	-1.397	2.535	1.000	-7.514	4.721

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

21.1.5 Interpretation: One-way ANCOVA

Table 21.3 and 21.4 displayed the descriptive statistics. Table 21.4 shows the *unadjusted* means of the systolic BP by sex (female: 129.7 and male: 124.5).

Table 21.5 shows the Levene's test of Equality of Error Variances. This is the test for assumption 2. We expect the p-value (Sig.) to be >0.05 to meet the assumption. In this example, the p-value is 0.54, which is more than 0.05. This means that the variances of the dependent variable (systolic BP) are same at each level of the independent variable (sex).

Table 21.6 (tests of between-subjects effects) is the main table showing the results of one-way ANCOVA test. We tested the hypothesis whether the population mean of the systolic BP of males and females is same after controlling for diastolic BP. Look at the p-value for sex in the table, and it is 0.245. Since the p-value is >0.05 , we cannot reject the null hypothesis. This indicates that the mean systolic BP (in the population) of males and females is not different *after controlling for diastolic BP*. Also look at the value for Partial Eta Squared. Eta indicates the amount of variance (also called effect size) in the dependent variable that is explained by the independent variable (sex). We can see that the effect size is very small (0.007 or 0.7%).

We can also have information about the influence of the covariate (diastolic BP) on the dependent variable (systolic BP). The p-value for diastolic BP is 0.000, which is highly significant. This indicates that there is a significant association between systolic and diastolic BP, after *controlling for sex*. The value of the Partial Eta Squared for diastolic BP is 0.735 (73.5%). This means that 73.5% variance of

systolic BP can be explained by the diastolic BP, after controlling for sex.

Table 21.7 (estimated marginal) shows the adjusted (adjusted for diastolic BP) means of the dependent variable (systolic BP) at different levels of the independent variable (sex). We can see that the mean systolic BP of females is 127.19 mmHg and that of males is 128.94 mmHg, after adjusting for diastolic BP (note that the adjusted means are different from the unadjusted means as shown in table 21.4).

Table 21.8 is the table for pairwise comparison. This table is not necessary in this example, since the independent variable (sex) has two levels. If the independent variable has more than two levels, then the table for pairwise comparison is important to look at, especially if there is a significant association between the dependent and independent variable. Look at table 21.9 [this is an additional table I have provided where the independent variable (religion) has three categories], which shows the pairwise comparison of mean systolic BP by religious groups. The results indicate that there is no significant difference of the mean systolic BP among different religious groups after controlling for diastolic BP, since all the p-values are >0.05 .

Section 22

Two-way ANCOVA

In two-way ANCOVA, there are two independent categorical variables with two or more levels/categories, while in one-way ANCOVA, there is only one independent categorical variable with two or more levels. Therefore, in two-way ANCOVA, four variables are involved. They are:

- One continuous *dependent* variable (e.g., diastolic BP, blood sugar, post-test score, etc.);
- Two categorical *independent* variables (with two or more levels) [e.g., occupation, diabetes, type of drug, etc.]; and
- One or more continuous *covariates* (e.g., age, systolic BP, income, etc.).

Use the data file <Data_3.sav> for practice.

22.1 Two-way ANCOVA

The two-way ANCOVA provides information, *after controlling for the covariate(s)*, on:

1. Whether there is a significant main effect on the dependent variable for the first independent variable (e.g., occupation);
2. Whether there is a significant main effect on the dependent variable for the second independent variable (e.g., diabetes);
3. Whether there is an interaction between the independent variables (e.g., occupation and diabetes).

Suppose, we want to assess, after controlling for age (covariate):

- Whether, occupation influences the diastolic BP (i.e., is mean diastolic BP in different occupation groups same);
- Whether, diabetes influences the diastolic BP (i.e., is the mean diastolic BP same for diabetics and non-diabetics); and
- Does the influence of occupation on diastolic BP depend on the presence of diabetes (i.e., is there interaction between occupation and diabetes)?

The question numbers 1 and 2 refer to the *main effect*, while question 3

explains the interaction of two independent variables (occupation and diabetes) on the dependent variable (diastolic BP). For analysis, we shall use the data file <Data_3.sav>. Note that the SPSS variable names for diastolic BP is “dbp”, occupation is “occupation”, diabetes is “diabetes” and age is “age”.

Assumptions:

All the assumptions mentioned under one-way ANCOVA are applicable for two-way ANCOVA. Look at one-way ANCOVA for the assumptions and how to check them.

22.1.1 Commands:

To perform the two-way ANCOVA, use the following commands:

Analyze > General linear model > Univariate > Push “dbp” into the “Dependent Variables” box > Push “occupation” and “diabetes” into the “Fixed Factor” box > Push “age” into the “Covariate” box > Click “Model” > Select “Full Factorial” > Continue > Options > Push “occupation, diabetes and occupation*diabetes” into the “Display Means for” box (this would provide the adjusted means of the diastolic BP for occupation and diabetes) > Select “Compare main effects” > Select “Bonferroni” from “Confidence interval adjustment” > Select “Descriptive Statistics, Estimates of effect size, and Homogeneity tests” > Continue > Plots > Select “occupation” and push into the “Horizontal” box > Select “diabetes and push it into the “Separate Lines” box > Click “Add” > Continue > Ok

22.1.2 Outputs:

Table 22.1. Between-Subjects Factors

		Value Label	N
OCCUPATION	1	GOVT JOB	60
	2	PRIVATE JOB	49
	3	BUSINESS	49
	4	OTHERS	52
DIABETES MELLITUS	0	No	165
	1	yes	45

Table 22.2. Descriptive Statistics (unadjusted)

Dependent Variable: DIASTOLIC BP				
OCCUPATION	DIABETES MELLITUS	Mean	Std. Deviation	N
GOVT JOB	No	84.58	12.862	50
	yes	82.60	11.559	10
	Total	84.25	12.583	60
PRIVATE JOB	No	82.80	14.128	41
	yes	79.75	9.468	8
	Total	82.31	13.443	49
BUSINESS	No	83.97	12.940	36
	yes	84.69	12.977	13
	Total	84.16	12.818	49
OTHERS	No	80.87	12.036	38
	yes	82.43	7.822	14
	Total	81.29	11.009	52
Total	No	83.15	12.982	165
	yes	82.64	10.410	45
	Total	83.04	12.454	210

Table 22.3. Levene's Test of Equality of Error Variances^a

Dependent Variable: DIASTOLIC BP			
F	df1	df2	Sig.
.963	4	202	.459
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.			

a. Design: Intercept + age + occupation + diabetes + occupation * diabetes

Table 22.4. Tests of Between-Subjects Effects

Dependent Variable: DIASTOLIC BP						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	495.245 ^a	8	61.906	.390	.925	.015
Intercept	101042.759	1	101042.759	636.198	.000	.760
age	34.661	1	34.661	.218	.641	.001
Occupation	229.135	3	76.378	.481	.696	.007
diabetes	12.293	1	12.293	.077	.781	.000
occupation * diabetes	124.301	3	41.434	.261	.854	.004
Error	31923.370	201	158.823			
Total	1480603.000	210				
Corrected Total	32418.614	209				
a. R Squared = .015 (Adjusted R Squared = -.024)						

Table 22.5. Estimated Marginal for occupation (adjusted)

1. OCCUPATION				
Dependent Variable: DIASTOLIC BP				
OCCUPATION	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
GOVT JOB	83.605 ^a	2.183	79.300	87.910
PRIVATE JOB	81.253 ^a	2.436	76.449	86.056
BUSINESS	84.374 ^a	2.041	80.350	88.399
OTHERS	81.708 ^a	1.974	77.815	85.601

a. Covariates appearing in the model are evaluated at the following values: age = 26.5143.

Table 22.6. Estimated Marginal for diabetes (adjusted)

2. DIABETES MELLITUS				
Dependent Variable: DIASTOLIC BP				
DIABETES MELLITUS	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
No	83.037 ^a	.990	81.086	84.988
yes	82.433 ^a	1.930	78.627	86.239

a. Covariates appearing in the model are evaluated at the following values: age = 26.5143.

Table 22.7. Estimated Marginal

3. OCCUPATION * DIABETES MELLITUS					
Dependent Variable: DIASTOLIC BP					
OCCUPATION	DIABETES MELLITUS	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
GOVT JOB	No	84.545 ^a	1.784	81.028	88.063
	yes	82.665 ^a	3.988	74.802	90.528
PRIVATE JOB	No	82.777 ^a	1.969	78.894	86.659
	yes	79.729 ^a	4.456	70.942	88.515
BUSINESS	No	83.979 ^a	2.100	79.837	88.121
	yes	84.769 ^a	3.499	77.870	91.669
OTHERS	No	80.847 ^a	2.045	76.815	84.880
	yes	82.569 ^a	3.381	75.901	89.236

a. Covariates appearing in the model are evaluated at the following values: age = 26.5143.

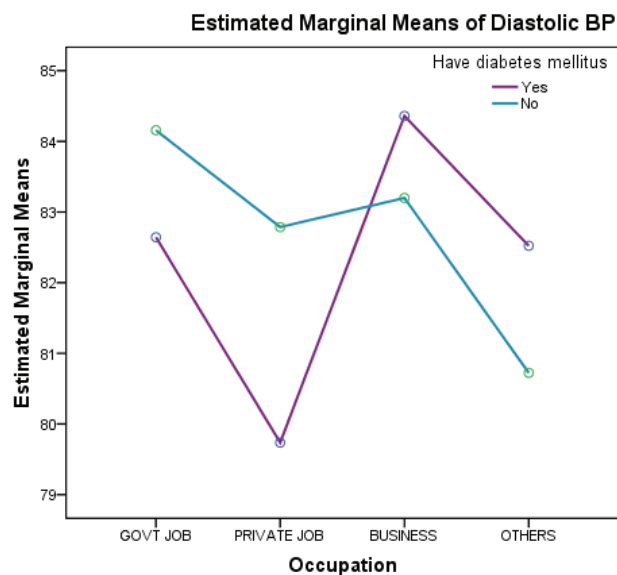
22.8. Pairwise Comparisons

Multiple Comparisons						
Dependent Variable: Diastolic BP						
(I) Occupation	(J) Occupation	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
GOVT JOB	PRIVATE JOB	2.139	3.089	1.000	-6.093	10.371
	BUSINESS	-.379	2.821	1.000	-7.896	7.138
	OTHERS	1.778	2.778	1.000	-5.625	9.180
PRIVATE JOB	GOVT JOB	-2.139	3.089	1.000	-10.371	6.093
	BUSINESS	-2.518	3.002	1.000	-10.519	5.482
	OTHERS	-.361	2.963	1.000	-8.257	7.534
BUSINESS	GOVT JOB	.379	2.821	1.000	-7.138	7.896
	PRIVATE JOB	2.518	3.002	1.000	-5.482	10.519
	OTHERS	2.157	2.678	1.000	-4.978	9.291
OTHERS	GOVT JOB	-1.778	2.778	1.000	-9.180	5.625
	PRIVATE JOB	.361	2.963	1.000	-7.534	8.257
	BUSINESS	-2.157	2.678	1.000	-9.291	4.978

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Figure 22.1 Mean diastolic BP of different occupation groups by diabetes after adjustment for age



22.1.3 Interpretation:

Table 22.1 and 22.2 shows the descriptive statistics. All the means provided in table 22.2 are the crude (unadjusted) means, i.e., without adjusting for age.

Table 22.3 shows the results of Levene's test of Equality of Error Variances. This is the test for homogeneity of variances. We expect the p-value (sig.) to be >0.05 to meet the assumption. In this example, the p-value is 0.459, which is more than 0.05. This means that the variances of the dependent variable (diastolic BP) are same for each level of the independent variables (occupation and diabetes).

Table 22.4 (tests of between-subjects effects) is the main table showing the results of the two-way ANCOVA test. We tested the hypothesis whether:

- Mean diastolic BP (in the population) in different occupation groups is same after controlling for age;
- Mean diastolic BP (in the population) among diabetics and non-diabetics is same after controlling for age; and
- Is there any interaction between occupation and diabetes after controlling for age?

Look at the p-values for occupation, diabetes and occupation*diabetes in table 22.4. They are 0.696, 0.781 and 0.854, respectively, indicating that none of them are statistically significant. This means that occupation and diabetes do not have any influence on the diastolic BP after controlling for age. There is also no interaction between occupation and diabetes after controlling for age. However, we should always check the p-value of interaction first. If the interaction is significant (p-value <0.05), then the main effects (of occupation and diabetes) are not important, because effect of one independent variable is dependent on the level of the other independent variable.

The effect size is indicated by the value of Partial Eta Squared. Eta indicates the amount of variance in the dependent variable that is explained by the independent variable (also called effect size). We can see that the effect sizes are very small both for occupation (0.007) and diabetes (0.000) (table 22.4).

We can also have information about the influence of the covariate (age) on the dependent variable (diastolic BP). We can see (table 22.4) that the p-value for age is 0.641, which is not statistically significant. This indicates that there is no significant association between age and diastolic BP after controlling for occupation and diabetes. The value of Partial Eta Squared for age is 0.001 (0.1%). This means that less than 1% variance in diastolic BP can be explained by age, after controlling for occupation and diabetes.

Tables 22.5 and 22.6 (estimated marginal) show the adjusted means of the

diastolic BP (dependent variable) at different levels of the independent variables (occupation and diabetes) after controlling for age. In this example, the adjusted mean of diastolic BP of government job holders is 83.6 mmHg and that of the diabetics (diabetes mellitus: yes) is 82.4 mmHg, after controlling for age. Similarly, the last table (table 22.7) shows the adjusted mean of diastolic BP of different occupation groups by diabetes.

Table 22.8 is the table of pairwise comparison of the mean diastolic BP in different occupation groups. *This table is necessary when the independent variable has more than two levels, and there is a significant association between the dependent and independent variable.* Look at the p-values (Sig.) in table 22.8. Since all the p-values are >0.05 , there is no significant difference of the mean diastolic BP in the population between different occupation groups after controlling for age.

Figure 22.1 plotted the mean diastolic BP of different occupation groups disaggregated by diabetes. Finally, from the data, we conclude that the diastolic BP is not influenced (there is no association) by occupation and diabetes after controlling for age.

Annex

Table A.1. Codebook of data file <Data_3.sav>

SPSS variable name	Actual variable name	Variable code
ID_no	Identification number	Actual value
age	Age in years	Actual value
sex	Sex: string	m= Male f= Female
sex_1	Sex: numeric	0= Female 1= Male
religion	Religion	1= Islam 2= Hindu 3= Others
religion_2	Religion 2	1= Islam 2= Hindu 3= Christian 4= Buddha
occupation	Occupation	1= Government job 2= Private job 3= Business 4= Others
income	Monthly family income in Tk.	Actual value
sbp	Systolic blood pressure in mmHg	Actual value
dbp	Diastolic blood pressure in mmHg	Actual value
f_history	Family history of diabetes	0= No 1= Yes
pepticulcer	Have peptic ulcer	1= Yes 2= No
diabetes	Have diabetes mellitus	1= Yes 2= No
post_test	Post-test score	Actual value
pre_test	Pre-test score	Actual value
date_ad	Date of hospital admission	Actual value
date_dis	Date of discharge	Actual value

References

1. Daniel WW. (1999). Biostatistics: A Foundation for Analysis in the Health Science. 7th edition. John Wiley & Sons, Inc.
2. Altman DG. (1992). Practical Statistics for Medical Research. 1st Edition. Chapman & Hill.
3. Katz MH. (2011). Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers. 3rd Edition. London, Cambridge University Press.
4. Katz MH. (2009). Study Design and Statistical Analysis: A Practical Guide for Clinicians. Cambridge University Press.
5. Gordis L. (2014). Epidemiology. 5th Edition. ELSEVIER Sounders.
6. Szklo M, Nieto FJ. (2007). Epidemiology: Beyond the Basics. 2nd Edition. Jones and Bartlett Publishers.
7. Field A. (2002). Discovering Statistics Using SPSS for Windows. SAGE Publications: London, California, New Delhi.
8. Reboldi G, Angeli F, Verdecchia P. Multivariable Analysis in Cerebrovascular Research: Practical Notes for the Clinician. Cerebrovasc Dis 2013; 35:187–193. DOI: 10.1159/000345491.
9. Katz MH. Multivariable Analysis: A Primer for Readers of Medical Research. Ann Intern Med 2003; 138:644–650.
10. Pallant J. (2007). SPSS Survival Manual. 3rd edition. Open University Press.
11. Chan YH. Biostatistics 103: Qualitative Data – Tests of Independence. Singapore Med J 2003; Vol 44(10):498-503.
12. Chan YH. Biostatistics 104: Correlational Analysis. Singapore Med J 2003; Vol 44(12):614-619.
13. Chan YH. Biostatistics 201: Linear Regression Analysis. Singapore Med J 2004; Vol 45(2):55-61.
14. Chan YH. Biostatistics 202: Logistic regression analysis. Singapore Med J 2004; Vol 45(4):149-153.
15. Chan YH. Biostatistics 203. Survival analysis. Singapore Med J 2004; Vol 45(6):249-256.
16. Amderson M, Nelson A. Data analysis: Simple statistical tests. FOCUS on Field Epidemiology; Vol 3(6).

About the Author

The author of this manual, *Mohammad Tajul Islam*, is a Professor (Adjunct) at the North South University and State University of Bangladesh. He teaches epidemiology, data analysis and statistical methods in health science for more than 15 years. He is a medical graduate with post-graduation in Tropical Medicine and Epidemiology from the Mahidol University, Bangkok, Thailand. His research interest is maternal and child health. He has authored and co-authored 15 articles published in the international peer review journals. He was involved in more than 25 research projects implemented in Bangladesh, and served as a member of the technical committee for Bangladesh Demographic and Health Survey (BDHS). He is a regular reviewer of the International Journal of Gynaecology and Obstetrics (IJGO) and occasionally reviews articles from other international peer review journals. He has worked at several UN Agencies (WHO, UNICEF, UNFPA) and international development organizations (JICA, Save the Children) including ICDDR,B for more than 20 years.