



## **404-Data Analysis with SPSS**

M.Com (CA) Semester-IV

Introduction to SPSS – Constructing Variables – BiVariate Analysis – Non-Parametric Procedures – multivariate Analysis.

**M.C.Mouli, Department of Commerce & Computer**

**FACULTY OF COMMERCE, SATAVAHANA UNIVERSITY, KARIMNAGAR**  
**MASTER OF COMMERCE (COMPUTER APPLICATIONS) - FOURTH SEMESTER**  
**404 – DATA ANALYSIS WITH SPSS**

(For M.Com-Computer Applications - under CBCS)

Class Hours: 4 ppw Credits: 4

---

**UNIT-I: SPSS Window Processes:** Menu Bar – File Menu, Edit Menu, View Menu, Data Menu, Transform Menu, Analyze Menu, Graphs Menu, Utilities Menu, Add-ons Menu, Window Menu and Help Menu – **Creating and Editing a Data File:** Structure of Data View and Variable View – Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure and Role. **Managing Data:** Dropping and Adding Variables – Listing Cases – Replacing Cases - Missing Cases — Computing New Variables – Recoding Variables – Selecting Cases – Sorting Cases and Merging and Importing Files.

**UNIT-II: Constructing Variables:** Recoding Existing Variables – Computing the Variables  
**Univariate Analysis:** Descriptive Statistics – Frequencies: Listing, summarizing and Sorting Cases – Mean, Media, Mode, Variance and Standard Deviation, Skewness, Maximum, Minimum, Range, Sum and Standard Error- **Creating and Editing Graphs and Charts:** Bar, 3-D Bar, Line, Area, Pie, Box-plot, Scatter Dot and Histogram.

**UNIT-III: Bi-variate Analysis:** Hypothesis and Significance Tests – Concept of p value - Significance Levels – Relationships between Two Variables – Cross Tabulations – Bar Charts – Correlation – Simple Linear Regression - Scatter plots. **Comparing Means through Bi-variate Analysis:** One-way Analysis of Variance – t-tests – Independent Sample t-test – Paired Sample t-test

**UNIT-IV: Non-parametric Procedures: Two Independent Sample Tests:** Mann-Whitney U-test – Two related Samples Test: Wilcoxon Test, Sign Test – The Runs Test – One - **Sample Test:** Kolmogorov-Smirnov Test – One-Sample Chi-Square Test – **Test for Several Related Samples:** Friedman One-way ANOVA - K-Sample Median Test.

**UNIT-V: Multivariate Analysis:** Factor Analysis – Opening Dialog Window – Descriptive Window – Kaiser-Meyer – Olkin(KMO) Measure of Sampling Adequacy – Bartlett's Test of Sphericity – Extraction of Factors – Principle Component Analysis – Communalities – Total Variance Explained – Eigen Values – Scree Plot – Component Transformation Matrix - Rotated Component Matrix– Interpretation of Output.

### **Suggested Readings**

1. Darren George and Paul Mallery, SPSS for Windows Step by Step – A Simple Guide and Reference, 7th Edition, Pearson Education, New Delhi, 2007
2. Sabine Landau and Brian S Everitt, A Handbook of Statistical Analyses using SPSS, Chapman & Hall/CRC, Washington DC, 2014 - (for e-book: [http://www.academia.dk/BiologiskAntropologi/Epidemiology/PDF/SPSS\\_Statistical\\_Analyses\\_using\\_SPSS.pdf](http://www.academia.dk/BiologiskAntropologi/Epidemiology/PDF/SPSS_Statistical_Analyses_using_SPSS.pdf))

### **References**

1. Stephen Sweet and Karen Grace-Martin, Data Analysis with SPSS – A First Course in Applied Statistics, Newyork, 2010.
2. Arthur Griffith, SPSS for Dummies, Wiley Publishing, Hoboken, New Jersey, 2007.
3. Robert B Burns and Richard A Burns, Business Research Methods and Statistics using SPSS, Sage Publications, New Delhi, 2008.

**FACULTY OF COMMERCE, SATAVAHANA UNIVERSITY, KARIMNAGAR**  
**MASTER OF COMMERCE (COMPUTER APPLICATIONS) - FOURTH SEMESTER**  
**404: DATA ANALYSIS WITH SPSS**

(For M.Com-Computer Applications - under CBCS) Lab: 2 PPW

Lab – Students are required to undergo Lab Sessions with SPSS Software

\*\*\*\*\*

**UNIT-I:**

1. Exercise on Understanding SPSS menus
2. Exercise on Understanding Structure of Data and Variable View
3. Exercise on Creating and Editing a Data File
4. Exercise on Adding and Dropping Variables
5. Exercise on Recoding Variables
6. Exercise on Sorting Cases
7. Exercise on Merging Files is
8. Exercise on Importing Files

**UNIT-II:**

9. Exercise on Computing Variable
10. Exercise on Computation of Mean, Median and Mode
11. Exercise on Computation of Standard Deviation, Variance and Skewness.
12. Exercise on Computation of Range, Sum, Minimum and Maximum
13. Exercise on Creating Bar and Line Diagrams
14. Exercise on Creating Histogram, Pie-Chart and Area Chart

**UNIT-III:**

15. Exercise on Cross Tabulations
16. Exercise on Computing Correlation
17. Exercise on Computing Linear Regression
18. Exercise on Comparing Means
19. Exercise on One-way Analysis of Variance
20. Exercise on Computation of Independent Sample t-test
21. Exercise on Computation of Paired t-test

**UNIT-IV:**

22. Exercise on Mann-Whitney U-test
23. Exercise on Wilcoxon Test
24. Exercise on Sign Test
25. Exercise on Runs Test
26. Exercise on Kolmogorov-Smirnov Test
27. Exercise on One-Sample Chi-Square Test
28. Exercise on Friedman One-way ANOVA
29. Exercise on K-Sample Median Test

**UNIT-V:**

30. Exercise on Factor Analysis
31. Exercise on Interpretation of Output of Factor Analysis.

\*\*\*\*\*

UNIT-I: **SPSS Window Processes:** Menu Bar – File Menu, Edit Menu, View Menu, Data Menu, Transform Menu, Analyze Menu, Graphs Menu, Utilities Menu, Add-ons Menu, Window Menu and Help Menu – **Creating and Editing a Data File:** Structure of Data View and Variable View – Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure and Role. **Managing Data:** Dropping and Adding Variables – Listing Cases – Replacing Cases - Missing Cases — Computing New Variables – Recoding Variables – Selecting Cases – Sorting Cases and Merging and Importing Files.

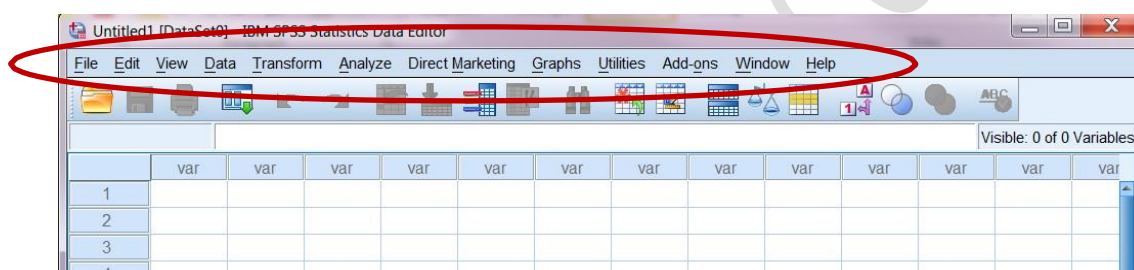
\*\*\*\*\*

### 1. Exercise on Understanding SPSS Menus?

**SPSS:** Statistical Package for Social Science: it can be to be presented with so many different options and functions... so we have created this menu tutorial to try to make things a little clearer and easier. This tutorial aims to provide you with an overview of the main menus that you can use in SPSS, and point out the important menu options that you will need to use for your own work.

SPSS has a number of menu options located at the top of the screen (as will any other computer program). This tutorial will now go through them, one-by-one.

To use this PDF version of the Menus tutorial, open SPSS and select each of the menu options one at a time. As you go through each of the menus, only the options that might be relevant to you are shown in the images below.



#### The 'File' Menu (shortcut **Alt + F**)

You will probably be familiar with a few of the options here already. Essentially this menu allows you to **Open** existing files, create **New** ones and **Print** or **Save** anything that you are working on.

The **Recently Used Data** and **Recently Used Files** lists are useful as they allow you to quickly access files that you have recently opened or worked on.

The main options you might be likely to use are highlighted in the dropdown menu displayed on the left. Feel free to ignore the others for now.

This menu should be quite familiar if you have used word processors before.

**Undo** and **Redo** can help rectify any mistakes you make.

**Cut**, **Copy** and **Paste** allow you to move blocks of numbers from one area of the spreadsheet to another.

**Find...** and **Go to Case...** allows you to locate a particular data score or participant, which comes in very handy when you are dealing with a large set of data.

#### The 'View' menu (shortcut **Alt + V**)

The View menu deals with the visual aspects of the spreadsheet, specifically: what **Toolbars** are displayed, which **Fonts** are used, whether you can see the **Grid Lines** on the spreadsheet, or whether

Value Labels are shown for your variables.

This menu allows you to organize your data file. You are unlikely to use most of the options on this menu initially; however a few of the options may come in handy.

For example, you can identify some potential mistakes made in data entry, by flagging possible duplicate entries of data, using **Identify Duplicate Cases**.

You can **Sort Cases** in your dataset (e.g. by numerical or date order) or **Sort Variables**.

You can **Transpose** your data set, so your rows become columns and vice versa.

Different data sets can be merged using **Merge Files**.

You can use **Split File** to separate your data into groups for analysis. And you can also use this menu to create an identical version of your file with **Copy Dataset**.

This menu allows you to manipulate your variables.

The **Recode** options allow you to change the values of specific variables (e.g. if you wanted to change the coding system you were using).

**Compute** allows you to create a new variable from existing variables (e.g. if you wanted to add or average several individual scores, which you might do when scoring a questionnaire).  
Meaning less output, but that's just part of learning how it works!

## 2. Exercise on understanding Structure of Data and Variable View?

The data view: The data view displays your actual data and any new variables you have created (we'll discuss creating new variables later on in this session).

From the menu, select File > Open > Data.

In the Open File window, navigate to C:\SPSSTutorialData\Employee data.sav and open it by double-clicking. SPSS opens a window that looks like a standard spreadsheet. In SPSS, columns are used for variables, while rows are used for cases (also called records).

Press **Ctrl-Home** to move to the first cell of the data view.

Press **Ctrl-End** to move to the last cell of the data view.

Press **Ctrl-Home** again to move back to the first cell.

The variable view

Variable types: SPSS uses (and insists upon) what are called *strongly typed* variables. *Strongly typed* means that you must define your variables according to the type of data they will contain. You can use any of the following types, as defined by the SPSS Help file.

- **Numeric:** A variable whose values are numbers. Values are displayed in standard numeric format. The Data Editor accepts numeric values in standard format or in scientific notation.
- **Comma:** A numeric variable whose values are displayed with commas delimiting every three places, and with the period as a decimal delimiter. The Data Editor accepts numeric values for comma variables with or without commas; or in scientific notation.
- **Dot:** A numeric variable whose values are displayed with periods delimiting every three places, and with the comma as a decimal delimiter. The Data Editor accepts numeric values for dot variables with or without dots; or in scientific notation. (*Sometimes known as European notation*).
- **Scientific notation:** A numeric variable whose values are displayed with an embedded E and a signed power-of-ten exponent. The Data Editor accepts numeric values for such variables with or without an exponent. The exponent can be preceded either by E or D with an optional sign, or by the sign alone—for example, 123, 1.23E2, 1.23D2, 1.23E+2, and even 1.23+2.
- **Date:** A numeric variable whose values are displayed in one of several calendar-date or clock-time formats. Select a format from the list. You can enter dates with slashes, hyphens, periods, commas, or blank spaces as delimiters. The century range for 2-digit year values is determined by your Options settings (from the Edit menu, choose Options and click the Data tab).

- **Custom currency:** A numeric variable whose values are displayed in one of the custom currency formats that you have defined in the Currency tab of the Options dialog box. Defined custom currency characters cannot be used in data entry but are displayed in the Data Editor.
- **String:** Values of a string variable are not numeric, and hence not used in calculations. They can contain any characters up to the defined length. Uppercase and lowercase letters are considered distinct. Also known as an alphanumeric variable.

At the bottom of the data window, you'll notice a tab labeled **Variable View**. The variable view window contains the definitions of each variable in your data set, including its name, type, label, size, alignment, and other information.

Click the **Variable View** tab.

Review the information in the rows for each variable.

### 3. Exercise on creating and editing a data file?

Creating a new data set \_\_\_\_\_

If you're doing original research or, in our case, creating databases for clients, there comes a time when you have to create your own data file from scratch. In this task, you'll create a new data set, define a set of variables, and then enter some data in the variables. You'll also create some automatic data entry constraints to improve the accuracy of your data entry.

In this task, you will create four types of variables: numeric, date, string, and binary.

1. From the menu, select **File > New > Data**. If you're asked to save the contents of the current file, click **No**.
2. When the new file opens in the Data View, click the **Variable View** tab at the bottom of the window.
3. With the cursor in the Name column on the first row (referring to the name of the variable) type: **clientid**

In the Type column, click the build button ("build button" is actually a Microsoft term, but since SPSS's documentation doesn't give the button a name, we'll use "build") to open the Variable Type dialog box.

Figure-4: Variable Type Dialog Box

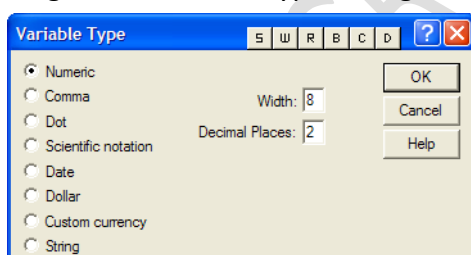
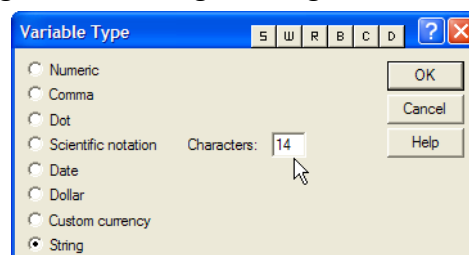


Figure-5: Defining the length of a string variable



Select (click) **String**. Notice that you can now define the length of the variable (Figure 5).

Select all the text in the Characters field and type: **14**

Click **OK**. The dialog box closes and the variable is now set to a length of 14 with no decimal places.

Press **tab** or **enter** three times to move to the label column.

Type: **Client ID**

This is the label that will appear on all output and in dialog boxes like those you used in crosstabs and charts.

Press **tab** or **enter** three times to move to the "columns" column. "Columns" defines the width of the **display** of the variable, not its actual contents. The display width affects how the column will be displayed in output like crosstabs and pivot tables.

Select all the text in the "columns" column and type: **14**

Leave the remaining columns as they are, with left alignment and "nominal" as the measure.

On the next row, click in the name column and type: **gender**



Press **tab** or **Enter** to move to the next column.

Click the build button to open the Variable Type dialog box.

Select **String** and click **OK** to accept the width.

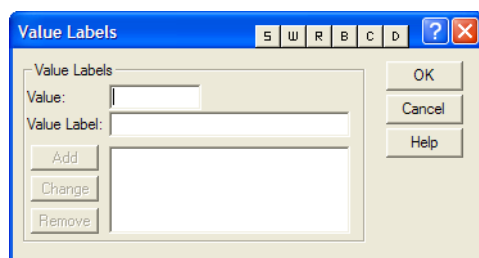
Click in the Label column for gender and type: **Gender**

Notice that in the variable labels, you can use upper and lower case as well as spaces and punctuation.

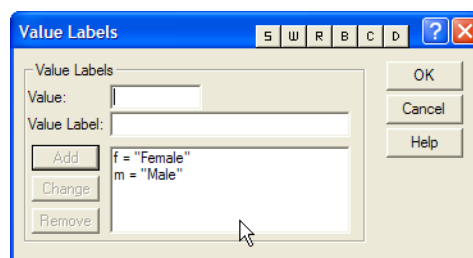
Press **tab** or **Enter** to move to the Values column.

Click the build button in the Values column to open the Value Labels window (Figure 6).

**FIGURE 6. Value Labels Window**



**FIGURE 7. Completed Value Labels window**



**Note:** *Variable* labels determine how the name of the variable is displayed in out- put. *Value* labels determine how each value is displayed. Thus, setting a label of “Female” for “f” in the gender variable instructs SPSS to display “Female” as a column heading for all cases with a value of *f* in gender.

In the Value field, type: **f**

In the Value Label field, type: **Female**

Click **Add**.

In the Value field, type: **m**

In the Value Label field, type: **Male**

Click **Add**.

The Value Labels window should now look like Figure 7.

Click **OK**.

On the next row, click in the Name column and type: **employed**

Press **tab** or **Enter** or click in the Type field.

Click the build button to open the Variable Type window.

Employed is going to be a numeric, binary variable, so leave numeric selected, but change **Width** to **1** and **Decimal Places** to **0**.

Click **OK**.

Tab to or click in the Label field and type:

### **Employed year-end**

Press **tab** or **Enter** to move to the Values field and click the Build Button.

In the Value field, type: **1**

In the Value Label field, type: **Yes**

Click **Add**.

In the Value field, type: **0**

In the Value Label field, type: **No**

Click **Add**.

Click **OK**.

On the next row, click in the Name field and type: **nextelig**

Press **tab** or **Enter** or click in the Type field and click the build button to open the Variable Type window.

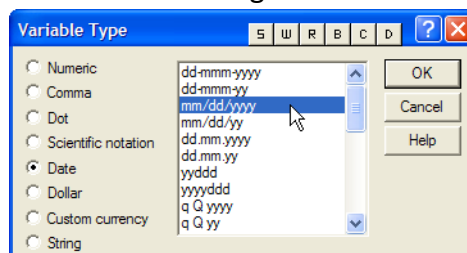
Select Date by clicking it.

In the pane to the right, select the date format mm/dd/yyyy as in Figure 8.

Click **OK**.

Tab to or click in the Label field and type:

**FIGURE 8.** Selecting a Date Format



#### 4. Exercise on adding and dropping variables?

To Enter Data: Open SPSS. A spreadsheet called UNTITLED DATA will open. In the first column insert numbers from 1 to 25 as shown on the data sheet. Clicking within any cell will select it for data entry. To move around the spreadsheet use the ENTER key to go down, TAB to go across to the right, and SHIFT TAB to move left. Or, you can use the arrows on the extended keyboard to move around within the spreadsheet.

To Input Variable Information (example):

Go to the **variable view** and click on the first variable name.

Replace var00001 with ID where it says **Name**.

Continue to move across the row and fill in the information for that variable.

**Type:** the default is numeric. Do not change for this example.

**Width:** the default is 8 spaces. Do not change for this example.

**Decimal places:** the default is 2. **Change to 0** for this example.

**Label:** In the "variable label" bar, type your variable name, in this example: **Identification number**

**Missing:** This is to identify variables that you want treated as missing. For example, if a respondent had put in not applicable on a survey item you would want to code it as a missing variable. Leave blank for this example.

**Columns:** The default is 8. Do not change for this example.

**Align:** This changes how the variables appear on your screen. Do not change for this example.

**Measure:** There are three different types of levels of measurement you can choose from -nominal, ordinal, and scale. Choose the appropriate measure. In this example, choose **nominal**.

For the categorical variables, **Province, Gender, Ethnicity, and Religion**, you will need to define the categories.

For example, variable 2: Province:

Under **values**-double click on the box labeled: **None...**

type: **1** in value bar

type: Alberta in value label bar

Click Add

type: **2** in value bar

type: British Columbia in value label bar

Click Add

Continue

OK

Enter all the data and name all the variables in this manner, according to the description provided.

Repeat this sequence for all the variables.



**Data Description:**

- Var1 Respondent's identification number (ID)
- Var2 Province the respondents lives in (PROVINCE)
- 1 Alberta
- 2 British Columbia
- Var3 Respondent's gender (GENDER)
- 1 male
- 2 female
- Var4 Respondent's ethnicity (ETHNICITY)
- 1 Caucasian
- 2 Black
- Var5 Respondent's age (AGE)
- Var6 Respondent's religious affiliation (RELIGION)
- 1 Protestant
- 2 Catholic
- 3 Jewish
- 4 None
- 5 Other

Var7 Respondent's mother's education - years of schooling (MAEDUC)

Name your data set and save it the data (either to disk or to your student file)

Note: Use this data to complete Exercise 1 Part B. **DATA SET**

	id	province	gender	ethnicity	age	religion	maeduc
1	1	1	1	1	32	1	16
2	2	1	2	1	37	2	13
3	3	1	2	2	72	2	20
4	4	1	2	1	86	3	12
5	5	1	1	1	30	1	5
6	6	1	1	1	32	2	10
7	7	1	2	2	29	1	18
8	8	1	1	2	29	1	4
9	9	1	2	2	53	1	6
10	10	1	1	2	68	1	9
11	11	1	2	1	19	2	2
12	12	1	2	2	43	2	14
13	13	2	2	2	38	4	12
14	14	2	1	1	45	2	17
15	15	2	1	2	24	4	1
16	16	2	1	1	53	2	3
17	17	2	2	1	20	2	7
18	18	2	1	1	27	2	11
19	19	2	1	1	54	2	8
20	20	2	2	2	25	1	15
21	21	2	1	2	20	2	1
22	22	2	2	2	38	2	7
23	23	2	1	1	20	2	5
24	24	2	2	2	34	2	10
25	25	2	2	1	67	1	19

## 5. Exercise on recoding variables?

### Frequencies

#### Task 1:

Open your data file from Exercise1.

Imagine that you need to classify your respondents into five categories in terms of their ages. To do so you will need to create a new categorical variable.

Recode the continuous variable Respondent's Age (**age**) into a new categorical variable (**agegroup**).

The values for the new variable will be as follows:

New Values (agegroup)	Old values (age):
1 – late adolescent	18-20
2 – young adult	21-40
3 – middle adult	41-60
4 – late adult	61-90

**Note:** The width of these intervals are not equal. In a *true* study, we would want the interval widths to be consistent!

In the menu bar go to Transform

Recode

Into different variable...

Transfer "**age**" into Output variable box

Type the name of a new variable - **agegroup**

Click on Change

Click on Old and New Values

In Old values select Range and type the first range of the old values: 18-20

In New value type 1

Click on Add

Repeat these steps for all old and new values

Continue

OK

You should have a new variable (agegroup) with the values 1 to 4.

Now define the new variable and its value levels. (You do this under **variable view**)

Now obtain the frequencies for agegroup:

What age group category has the least number of participants/people? \_\_\_\_\_

What age group category has the most number of participants/people? \_\_\_\_\_

What % of the sample is late adult? \_\_\_\_\_

What % of the sample is young adult? \_\_\_\_\_

What % of the sample is middle adult? \_\_\_\_\_

## 6. Exercise on Sorting Cases?

**Sort Cases:** This dialog box sorts cases (rows) of the active dataset based on the values of one or more sorting variables. You can sort cases in ascending or descending order.

If you select multiple sort variables, cases are sorted by each variable within categories of the preceding variable on the Sort list. For example, if you select *gender* as the first sorting variable and *minority* as the second sorting variable, cases will be sorted by minority classification within each gender category.

The sort sequence is based on the locale-defined order (and is not necessarily the same as the numerical order of the character codes). The default locale is the operating system locale. You can control the locale with the Language setting on the General tab of the Options dialog box (Edit menu).

To Sort Cases  
From the menus choose:  
Data > Sort Cases...  
Select one or more sorting variables.  
Optionally, you can do the following:

**Index the saved file.** Indexing table lookup files can improve performance when merging data files with STAR JOIN.

**Save the sorted file.** You can save the sorted file, with the option of saving it as encrypted. Encryption allows you to protect confidential information stored in the file. Once encrypted, the file can only be opened by providing the password assigned to the file.

To save the sorted file with encryption:

Select Save file with sorted data and click File.

Select Encrypt file with password in the Save Sorted Data As dialog box.

Click Save.

In the Encrypt File dialog box, provide a password and re-enter it in the Confirm password text box. Passwords are limited to 10 characters and are case-sensitive.

*Warning:* Passwords cannot be recovered if they are lost. If the password is lost the file cannot be opened.

Creating strong passwords

Use eight or more characters.

Include numbers, symbols and even punctuation in your password.

Avoid sequences of numbers or characters, such as "123" and "abc", and avoid repetition, such as "111aaa".

Do not create passwords that use personal information such as birthdays or nicknames.

Periodically change the password.

## 7. Exercise on merging files? Merge two files in SPSS

To implement the merge:

Use the SORT CASE command to sort each file you wish to merge on the index variable.

After sorting, use the SAVE OUTPUT command to save each file to a systems file.

Use the MATCH FILES command to merge the files and create a single, final version of the file.

The following example of this process merges two files, merge1.sav and merge2.sav, with the shared index variable v1, into a final file, sort.sav:

```
GET FILE = merge1.sav.
```

```
SORT CASE BY v1.
```

```
SAVE OUTFILE = merge1.sav.
```

```
GET FILE = merge2.sav.
```

```
SORT CASE BY v1.
```

```
SAVE OUTFILE = merge2.sav.
```

```
MATCH FILES
```

```
/FILE=merge1.sav
```

```
/FILE=merge2.sav
```

```
/BY v1.
```

```
SAVE OUTFILE = sort.sav.
```

## 8. Exercise on importing files?

Import SPSS Files

To open an SPSS file (Windows)

Select file > Open

From the list next to file name, Select SPSS data Files(\*.sav)

Select the SPSS file

(Optional) to specify the column headings, select one of the following, select one of the following

Set JMP column names from options: SPSS Labels creates column headings from SPSS labels.

SPSS Variable Names Creates column headings from variable names.

Click Open

JMP opens the file as a data table.

To open an SPSS file (Macintosh)

1.Select File > Open.

2.Select the SPSS file.

3.(Optional) To specify the column headings, do one of the following

- Deselect Use SPSS Labels as Headings to convert variable names to column headings.
- Select Use SPSS Labels as Headings to convert labels to column headings.

4.Click Open.

JMP opens the file as a data table.

## UNIT-II

### 1. Exercise on computing variables?

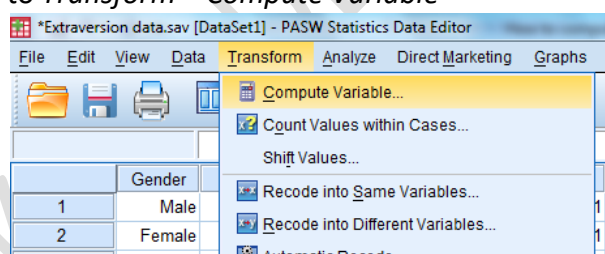
It is sometimes necessary to compute a new variable, condensing several raw data points into one. For example, when using standardized questionnaires you often need to calculate a total and/or several sub- scale scores rather than analyzing every question separately.

Enter all your raw data and complete all data recoding (eg. reverse scoring) if required before you begin computing scores. The compute variable function does not automatically update when you enter new data or modify existing data.

For this example, I have 10 questions, two of which have been reverse scored (Q4Rev, Q7Rev). Questions 2, 4, 6, 8 and 10 form one sub-scale. Questions 1, 3, 5, 7 and 9 form a second sub-scale.

	Gender	Age	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q4Rev	Q7Rev
1	Male	19	2	2	3	1	2	1	3	2	1	2	5	3
2	Female	20	3	4	3	1	2	5	3	4	4	-9999	5	3
3	Male	19	5	5	4	3	4	5	5	4	4	4	3	1
4	Male	19	1	1	1	1	2	1	2	1	2	1	5	4
5	Male	20	2	3	2	3	4	2	3	3	3	2	3	3
6	Female	23	4	5	4	1	4	4	4	5	4	5	5	2
7	Female	19	1	2	1	5	1	1	2	1	2	2	1	4
8	Male	25	2	3	2	2	2	5	2	3	2	5	4	4
9	Male	18	1	3	2	5	2	1	2	2	4	3	1	4
10	Male	38	1	3	5	-9999	2	3	1	4	2	2	-9999	5
11	Male	20	2	3	4	2	3	3	3	3	2	2	4	3
12	Male	20	2	4	3	2	5	2	3	2	1	2	4	3
13	Female	21	2	4	1	2	2	2	2	1	1	4	4	4
14	Female	20	3	4	2	3	5	2	3	4	2	3	3	3
15	Male	21	5	1	2	3	2	1	1	1	1	2	3	5
16	Female	19	4	3	5	2	3	4	5	-9999	-9999	5	4	1
17	Female	19	2	3	5	2	3	5	5	4	4	4	4	1
18	Female	18	4	3	4	2	1	2	1	4	2	3	4	5
19	Female	19	4	3	2	5	5	5	5	5	5	5	1	1
20	Female	18	2	1	2	5	4	5	4	3	2	3	1	2

### Go to Transform – Compute Variable



In the Target Variable box enter a name for the variable you will be calculating. The Numeric Expression box is like a calculator in so far as you enter all the variables included in the calculation.

A summed score: If you need to add up all the answers within a sub-scale...

In the Numeric Expression box, type *SUM*. Then in brackets, move across all the variables included in this calculation, using the reverse scored variables where necessary (eg. Q4Rev in this example), and insert a comma in between each variable.

## 2. Exercise on computation of mean, median, mode?

**Mean, median, and mode:** Mean, median, and mode are different measures of center in a numerical data set. They each try to summarize a dataset with a single number to represent a "typical" data point from the dataset.

**Mean:** The "average" number; found by adding all data points and dividing by the number of data points.

**Example:** The mean of 444, 111, and 777 is  $(4+1+7)/3 = 12/3 = 4$  because when the numbers are put in order (1, 444, 777), the number 444 is in the middle.

**Median:** The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

**Example:** The median of 444, 111, and 777 is 444 because when the numbers are put in order (1, 444, 777), the number 444 is in the middle.

**Mode:** The most frequent number—that is, the number that occurs the highest number of times.

**Example:** The mode of {4, 222, 444, 333, 222, 2} is 222 because it occurs three times, which is more than any other number.

*Want to learn more about mean, median, and mode? Check out the more in-depth examples below, or check out this video explanation.*

**Calculating the mean:** There are many different types of mean, but usually when people say mean, they are talking about the arithmetic mean.

The arithmetic mean is the sum of all of the data points divided by the number of data points.

$\text{mean} = \frac{\text{sum of data}}{\text{number of data points}}$

Here's the same formula written more formally:

$\text{mean} = \frac{\sum x_i}{n}$  where  $\sum$  is the sum symbol,  $x_i$  is the data point, and  $n$  is the number of data points.

**Example:** Find the mean of this data:

111, 222, 444, 555

Start by adding the data:

$1+2+4+5=12$

There are 4 data points.

$\text{mean} = \frac{12}{4} = 3$

**The mean is 3.**

**Finding the median:** The median is the middle point in a dataset—half of the data points are smaller than the median and half of the data points are larger.

To find the median:

Arrange the data points from smallest to largest.

If the number of data points is odd, the median is the middle data point in the list.

If the number of data points is even, the median is the average of the two middle data points in the list.

**Example:** Find the median of this data:

111, 444, 222, 555, 000

Put the data in order first:

000, 111, 222, 444, 555

There is an odd number of data points, so the median is the middle data point.

000, 111, 222, 444, 555

**The median is 222.**



Finding the mode: The mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.

### Example 1

Ms. Norris asked students in her class how many siblings they each had.

#### Find the mode of the data:

000, 000, 111, 111, 111, 111, 111, 111, 222, 222, 222, 333, 555

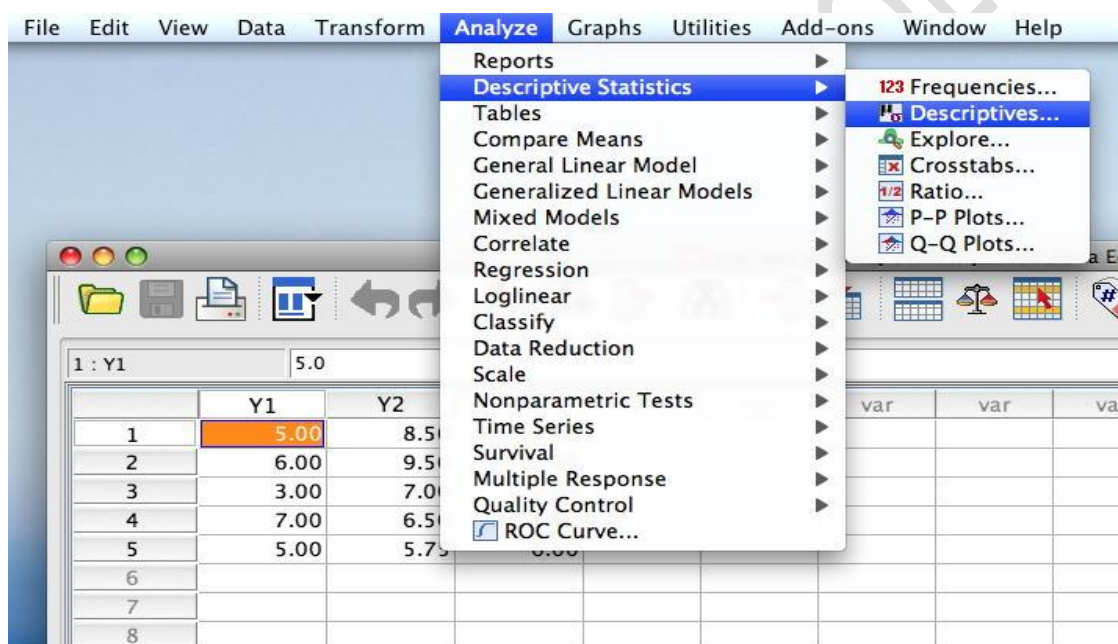
Look for the value that occurs the most:

000, 000, \large111, \large111, \large111, \large111, \large111, \large111, 222, 222, 222, 333, 555

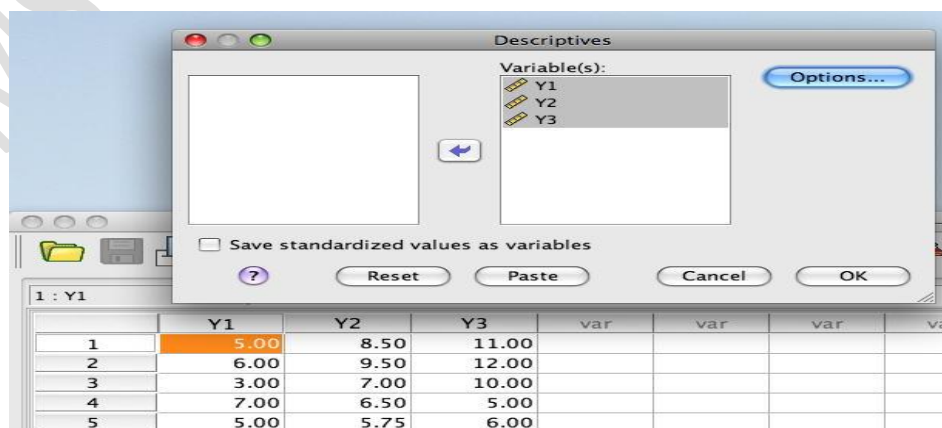
### 3. Exercise on computation of standard deviation, variance and skewness?

#### Computing the Standard Deviation in SPSS

The standard deviation is a measure of variability. In SPSS, you compute it by choosing Analyze/Descriptive Statistics/Descriptives...



You then specify the variables you want for which you want to compute the standard deviation:



Here is the result. Note that "Std. Deviation" is used to stand for "standard deviation.":

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Y1	5	3.00	7.00	5.2000	1.48324
Y2	5	5.75	9.50	7.4500	1.52480
Y3	5	5.00	12.00	8.8000	3.11448
Valid N (listwise)	5				

Notice that, by default, you get N, the minimum, the maximum, and the mean in addition to the standard deviation. You could have chosen more or fewer statistics by clicking the "option" button. The syntax for computing the standard deviation is:

DESCRIPTIVES VARIABLES=Y1 Y2 Y3

/STATISTICS=MEAN STDDEV MIN MAX.

#### 4. Exercise on computation of range sum, minimum and maximum?

How to Compute Total Scale Scores in SPSS

There are different ways to compute total scale scores, some of which are discussed in your book. Many scale scores are computed by summing the responses to all of the items included in the scale. For example, in the RAINN data, you may want to compute a "Total Satisfaction" score by summing all of the five satisfaction items. To compute a total scale score by summing the items follow these steps:

Open the SPSS data file.

Select **Transform** from the SPSS main toolbar.

Select **Compute Variable**.

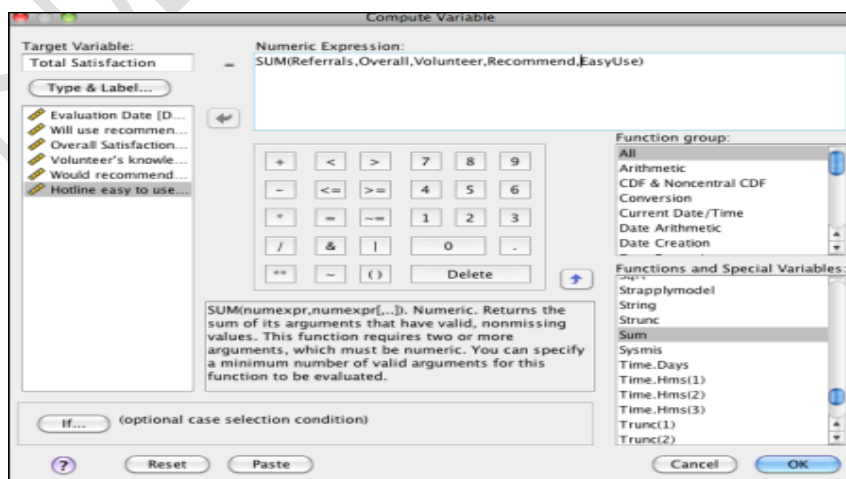
Type the name of the new variable (the variable name you will assign to the scale score) in the target variable box in the upper left corner (e.g. Total Satisfaction).

In the **Function group** box click **All** or **Statistical**. Scroll down under **Functions and Special Variables** and click on **SUM ( )**.

Click on the up-arrow to move the function into the number expression box. It should read, SUM(?,?)

From the Type and Label box, select the variables that are included in the scale by clicking on each variable and then on the arrow to move it into the parentheses.

There should be a comma between each variable in the parentheses. Once you have confirmed this, click on OK and you are finished.



A new variable with the variable name you assigned it will be found at the end of your data file. Scroll to the variable and then down the column to look at the values for Total Scale score.

## 5. Exercise on creating bar and line diagrams?

### References:

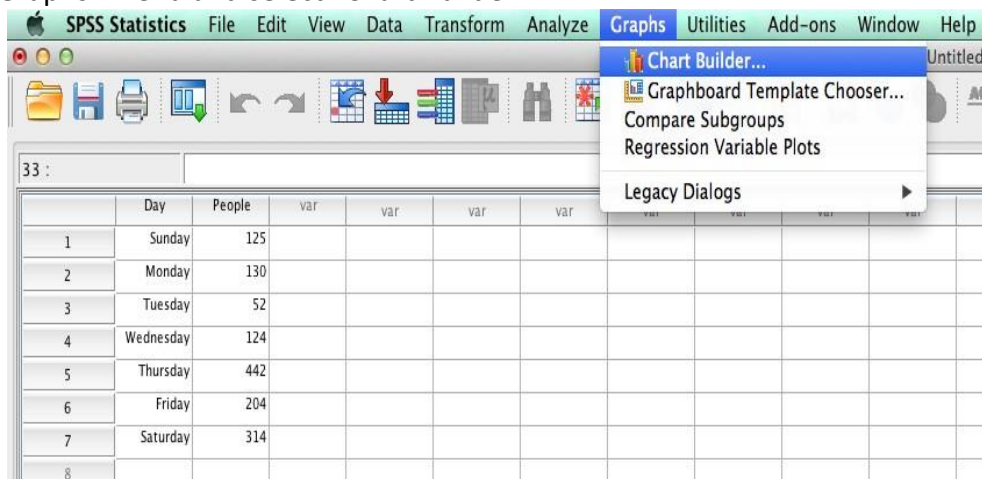
IBM Corp. Released 2013. IBM SPSS Statistics for Macintosh, Version 22.0. Armonk, NY: IBM Corp.

### Objective:

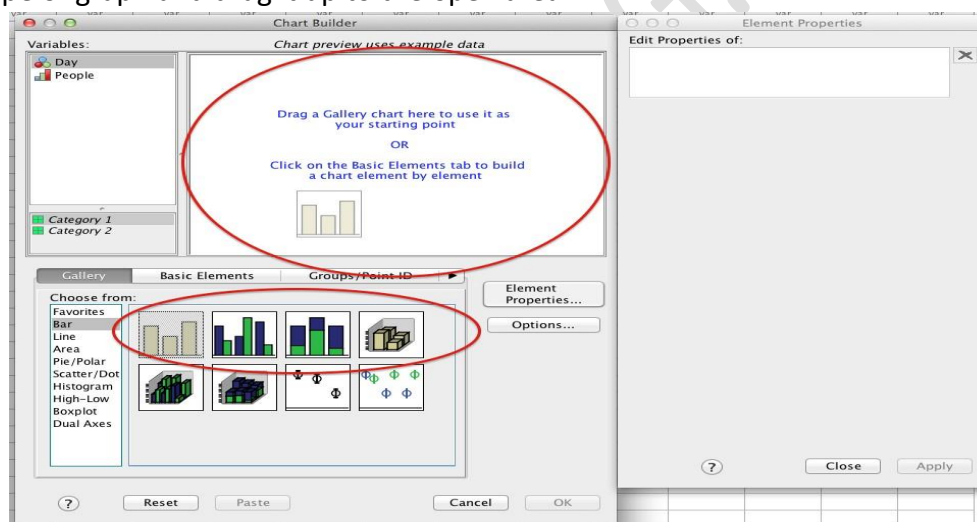
Create a bar graph for a categorical variable for which values do not add up to 100%.

Steps to Creating a Bar Graph in SPSS

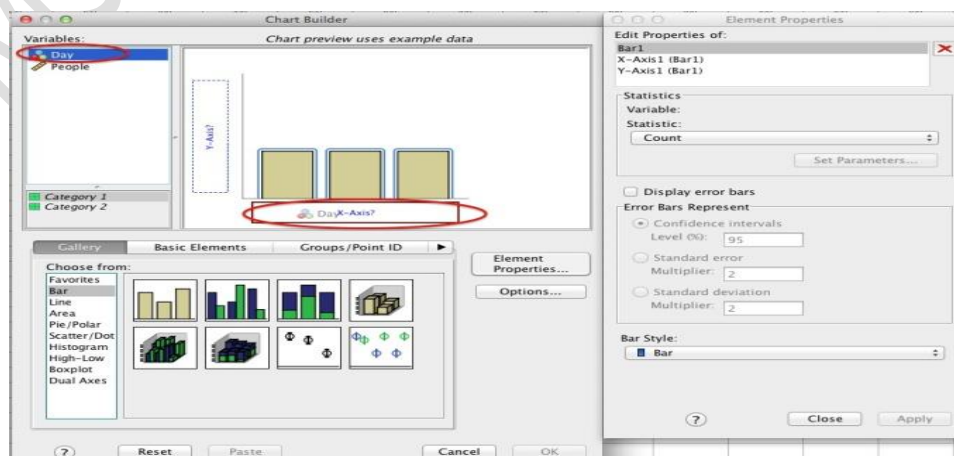
Go to the “Graphs” menu and select “Chart Builder.”



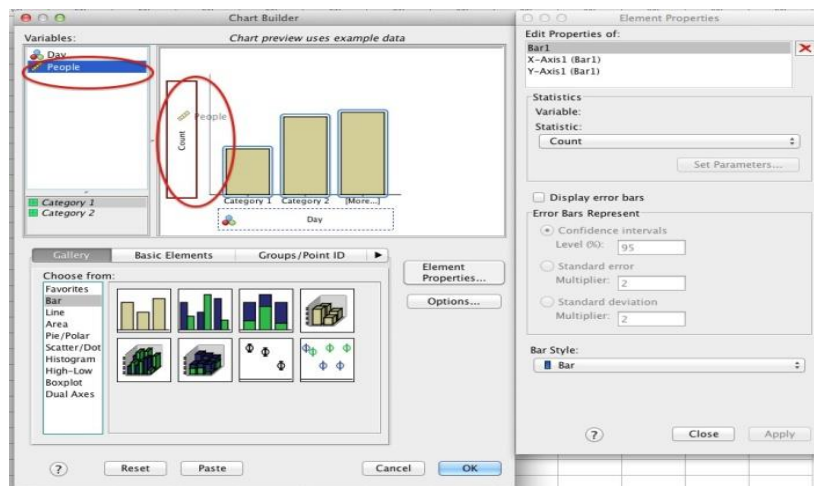
Select the type of graph and drag it up to the open area.



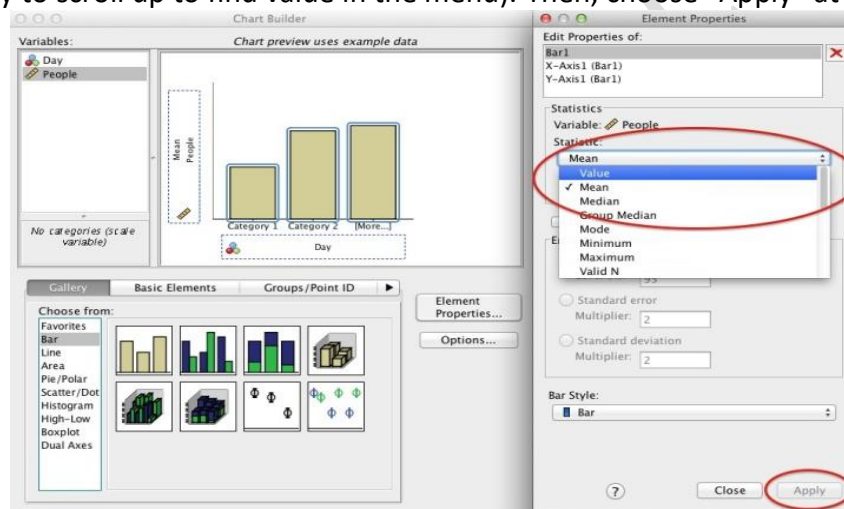
Drag the categorical variable from the column on the left to the area below the graph labeled “X-Axis?”



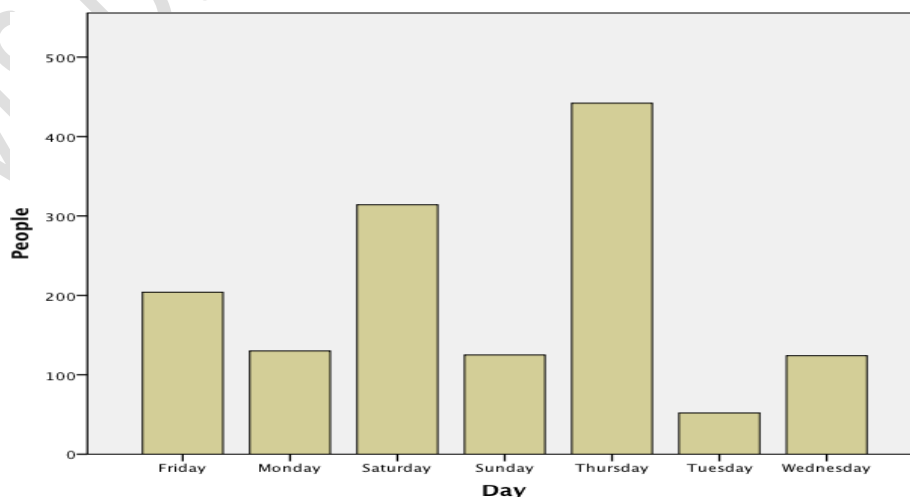
Drag the count or frequency variable that goes along with the selected categorical variable into the space now labeled “Count.”



This will automatically change “Count” to “Mean.” So the next step is to change this back to count or frequency by choosing the “Statistic” menu in the window on the left and selecting “Value.” (It may be necessary to scroll up to find value in the menu). Then, choose “Apply” at the bottom left.



Finally, click “OK.” The graph will appear in the Output window.



## 6. Exercise on creating histogram and pie chart and area chart?

### References:

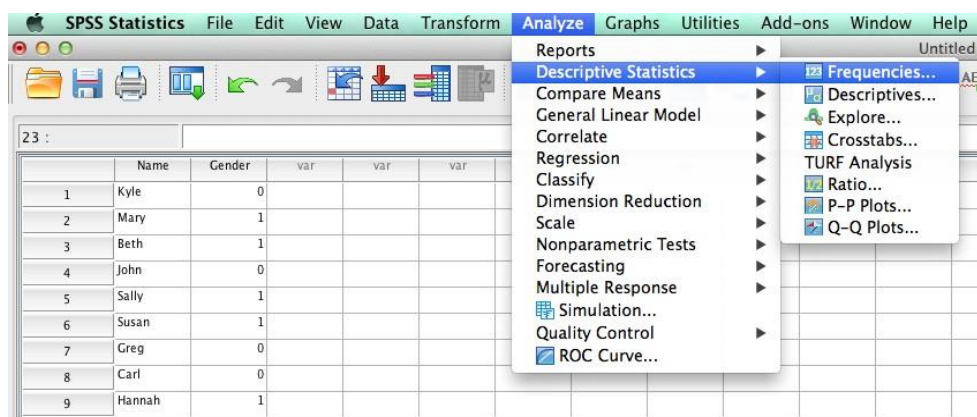
IBM Corp. Released 2013. IBM SPSS Statistics for Macintosh, Version 22.0. Armonk, NY: IBM Corp.

### Objective:

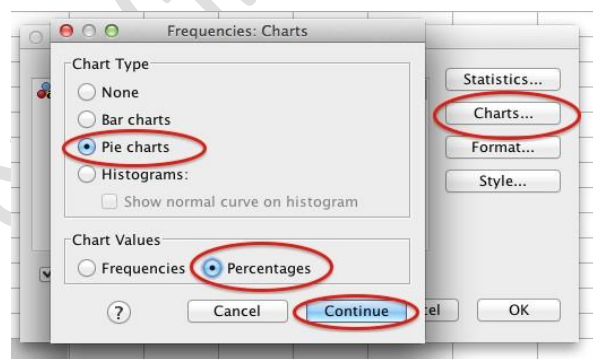
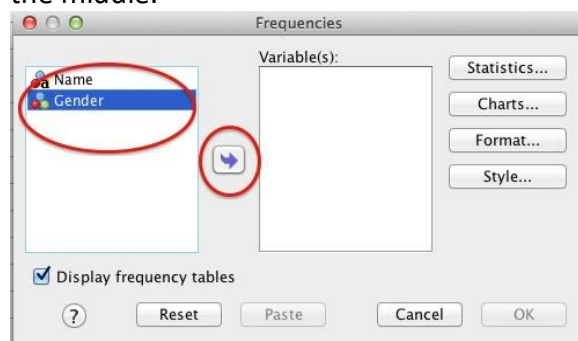
Create a pie chart for a categorical variable for which values add up to 100%.

Steps to Creating a Pie Chart in SPSS

Go to the “Analyze” menu and select “Descriptive Statistics,” then “Frequencies.”

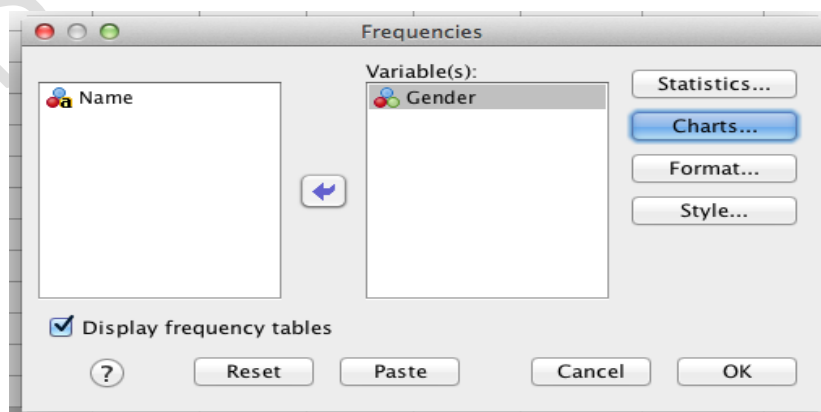


Select the variable of interest for the pie chart from the list on the left, then click on the arrow in the middle.



Click “Charts” on the right. Then choose “Pie Chart.” Under “Chart Values,” choose “Frequencies” or “Percentages.” In this case, “Percentages will be chosen. Finally, click “Continue.”

Finally, choose “OK.”



The results will be displayed in a separate window called the “Output” window.

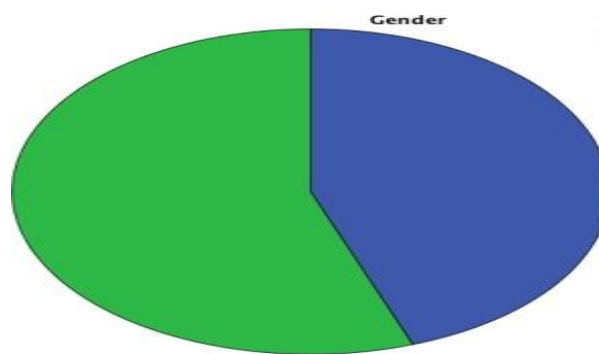
Statistics

Gender

N	Valid	9
	Missing	0

Gender

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 0	4	44.4	44.4	44.4
1	5	55.6	55.6	100.0
Total	9	100.0	100.0	



To use part of the output, simply right-click and choose copy. Then, paste into a word document or another place



## UNIT-III

### 1. Bivariate Analysis: Hypothesis and significance test

Bivariate analysis involves analyzing the relationship between two variables to determine if there is a significant association between them. Here's how you typically approach it:

#### Hypothesis

**Null Hypothesis (H<sub>0</sub>):** There is no significant relationship between the two variables.

**Alternative Hypothesis (H<sub>1</sub>):** There is a significant relationship between the two variables.

#### Significance Test

The most common significance test used for bivariate analysis depends on the type of variables being analyzed:

##### For Categorical Variables:

**Chi-Square Test:** This test determines whether there is a significant association between two categorical variables.

##### Hypothesis:

H<sub>0</sub>: There is no association between the two variables.

H<sub>1</sub>: There is an association between the two variables.

**Significance Level:** Typically, a significance level ( $\alpha$ ) of 0.05 is used. If the p-value is less than 0.05, the null hypothesis is rejected, indicating a significant association.

##### For Continuous Variables:

**Correlation Coefficient (Pearson's correlation for linear relationship, Spearman's rank correlation for monotonic relationship):** This measures the strength and direction of the linear or monotonic relationship between two continuous variables.

##### Hypothesis:

H<sub>0</sub>: There is no linear (or monotonic) relationship between the two variables.

H<sub>1</sub>: There is a linear (or monotonic) relationship between the two variables.

**Significance Level:** Again, a significance level ( $\alpha$ ) of 0.05 is commonly used. If the p-value is less than 0.05, the null hypothesis is rejected, indicating a significant relationship.

#### Steps for Significance Testing

**Data Collection:** Collect data for the two variables.

**Data Preparation:** Clean and preprocess the data, ensuring it is suitable for analysis.

**Choose a Test:** Depending on the type of variables, select an appropriate significance test (e.g., chi-square test for categorical variables, correlation coefficient for continuous variables).

**Calculate Test Statistic:** Calculate the test statistic (e.g., chi-square statistic, correlation coefficient) using the collected data.

**Determine Significance:** Compare the obtained p-value with the chosen significance level ( $\alpha$ ). If  $p < \alpha$ , reject the null hypothesis; otherwise, fail to reject it.

**Interpretation:** Based on the test result, interpret whether there is a significant relationship between the two variables or not.

**Conclusion:** Conclude the analysis based on the interpretation of the test results.

#### Example

Let's say you want to analyze the relationship between gender (male/female) and smoking status (smoker/non-smoker) using a chi-square test. Your hypothesis would be:

H<sub>0</sub>: There is no association between gender and smoking status.

H<sub>1</sub>: There is an association between gender and smoking status.

After performing the chi-square test, if you obtain a p-value of 0.003, which is less than the significance level of 0.05, you would reject the null hypothesis. This suggests that there is a significant association between gender and smoking status.

## 2. Concept of p value - significance Levels - Relationships between Two Variables

The concept of p-value, significance levels, and relationships between two variables are fundamental in statistical analysis, especially in bivariate analysis. Let's break down each concept:

### P-value

The p-value is a measure that helps us determine the significance of results obtained from a statistical test. It indicates the probability of observing the data or more extreme results under the assumption that the null hypothesis is true.

**Low p-value:** If the p-value is low (typically less than a predetermined significance level, often 0.05), it suggests that the observed data is unlikely to have occurred if the null hypothesis were true. This leads to the rejection of the null hypothesis in favor of the alternative hypothesis.

**High p-value:** If the p-value is high, it indicates that the observed data is reasonably likely to have occurred even if the null hypothesis were true. In this case, we fail to reject the null hypothesis.

### Significance Levels

Significance levels, denoted by  $\alpha$ , represent the threshold below which the p-value is considered statistically significant. Common significance levels include 0.05, 0.01, and 0.1.

**$\alpha = 0.05$ :** This is the most commonly used significance level. It means that if the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a significant effect.

**$\alpha = 0.01$ :** A stricter significance level. If the p-value is less than 0.01, we reject the null hypothesis. This level is used when a higher level of confidence is required.

**$\alpha = 0.1$ :** A less stringent significance level. If the p-value is less than 0.1, we reject the null hypothesis. This level is sometimes used for exploratory analysis.

### Relationships between Two Variables

The relationship between two variables in bivariate analysis can be characterized in various ways depending on the types of variables:

#### For Categorical Variables:

**Chi-Square Test:** Determines whether there is a significant association between two categorical variables. It measures the strength of association between the variables.

#### For Continuous Variables:

**Correlation Coefficient:** Measures the strength and direction of the linear relationship between two continuous variables.

Pearson's correlation coefficient for linear relationships.

Spearman's rank correlation coefficient for monotonic relationships (not necessarily linear).

#### For Ordinal Variables:

**Spearman's Rank Correlation Coefficient:** Measures the strength and direction of the monotonic relationship between two ordinal variables.

### Interpretation

**Significant Relationship:** If the p-value is less than the chosen significance level, we reject the null hypothesis and conclude that there is a significant relationship between the two variables.

**No Significant Relationship:** If the p-value is greater than the chosen significance level, we fail to reject the null hypothesis, indicating that there is no significant relationship between the variables. Understanding these concepts is crucial for interpreting statistical analyses and drawing meaningful conclusions about the relationships between variables.

## 3. Exercise on cross tabulations?

### Steps:

**Sample question:** Make an SPSS Crosstab for age (2,3,4,6,8,44,34,33,45,56,57,57) vs. healthcare type (1,1,2,3,4,5,6,1,2,4,5,5). Show percentages of healthcare type in the contingency table.

Step 1: **Type your data into an SPSS worksheet.** Contingency tables require at least two variables (columns) of data. For this sample question, type ages (2,3,4,6,8,44,34,33,45,56,57,57) into the first column and then type Healthcare type (1,1,2,3,4,5,6,1,2,4,5,5) into the second column. Change the variable names (the column headers) by clicking the “Variable” button at the bottom of the sheet and typing over the variable name.

Step 2: **Click “Analyze,” then hover over “Descriptive Statistics” and then click “Crosstabs.”** The Crosstabs dialog window will open.

Step 3: **Select one variable in the left window** and then click the top arrow to populate the “Row(s)” box. Select a variable to populate the “Column(s)” box and then click the center arrow.

For this sample problem, “Age” was selected for “Row(s)” and “Healthcare Type” was selected for “Column(s).” Once you have made your selection, click “Cells.”



Step 4: **Check which percentages you want to see (rows or columns).** What you select will depend upon what variables you put in rows and what you put in columns. For this sample problem, “Healthcare Type” was placed in the columns, so check “Column” under percentages.

Step 5: **Click “Continue” and then click “OK.”** The Crosstabs window will appear.

## Crosstabs

[DataSet1]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age * HealthcareType	12	100.0%	0	0.0%	12	100.0%

Age \* HealthcareType Crosstabulation

			HealthcareType					
			1.00	2.00	3.00	4.00	5.00	6.00
Age 2.00	Count		1	0	0	0	0	0
	% within HealthcareType		33.3%	0.0%	0.0%	0.0%	0.0%	0.0%
3.00	Count		1	0	0	0	0	0
	% within HealthcareType		33.3%	0.0%	0.0%	0.0%	0.0%	0.0%
4.00	Count		0	1	0	0	0	0
	% within HealthcareType		0.0%	50.0%	0.0%	0.0%	0.0%	0.0%
6.00	Count		0	0	1	0	0	0
	% within HealthcareType		0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
8.00	Count		0	0	0	1	0	0
	% within HealthcareType		0.0%	0.0%	0.0%	50.0%	0.0%	0.0%
33.00	Count		1	0	0	0	0	0
	% within HealthcareType		33.3%	0.0%	0.0%	0.0%	0.0%	0.0%

IBM SPSS Statistics Processor is ready

**Tip:** Typically in a contingency table, the independent variable is placed in rows and the dependent variable is placed in columns. This is just a tradition and has no mathematical basis, so don't over analyze which variable to place in columns and which to place in rows when you are creating crosstabs in SPSS.

## 4. Exercise on computing correlation?

### Simple Correlation

#### Task 1: Correlation between Two Variables

Use the 1991 U.S. General Social Survey.dat data set (ITS website) to find the strength of the relationship between fathers' education level (highest year of school completed, father: paeduc) and mother's education level (highest year of school completed, mother: maeduc).

In the main menu bar go to:

Analyze

Correlate

Bivariate (meaning 2 variables)...

Transfer the appropriate set of variables to the Variable box

The default options selected are Pearson Correlation Coefficient, 2 tailed significance test, flag significant correlations

OK

	paeduc	maeduc
paeduc	r=1.00	
maeduc		r=1.00

Is the correlation significant? Yes / No      If yes, at what significance level? \_\_\_\_\_

How many people are in the data set? \_\_\_\_\_

What proportion of variance in maeduc is explained by paeduc? \_\_\_\_\_

Note about interpreting significant correlations: With larger samples, small correlations may be deemed significant because of the power. A better way of interpreting correlations is to consider the proportion of variance ( $r^2$ ). For example, a correlation of 0.2 may be significant, but accounts for only 4 percent of the variance.

Scatter plot: The scatterplot enables you to see whether a correlation will accurately summarize the relationship between 2 variables. Correlations are appropriate only for linear relationships. The  $r$  will be an underestimation if the relationship is curvilinear. It is important to examine scatterplots when studying relationships between variables.

To produce a scatterplot for the pair of variables, in the main menu bar go to:

Graphs

Chart Builder - OK

Select "Scatter" from the gallery

Select "Simple" or the first graph presented – running your mouse over each example graph will tell you what they are.

Select the variables from the list on the upper left and drag and drop the variable on the selected axis

Transfer maeduc to the Y-axis and paeduc to the X-axis

OK (The graph will then be entered into your viewer folder)

SPSS produces simple scatterplots this way. To obtain a line of best fit (more on this next lab)

Double click on your graph

Chart Editor window will open

From the menu bar in the Chart Editor window select ELEMENTS - Fit Line at Total

OK

Close the Chart Editor window

Describe the relationship between the maeduc and paeduc. \_\_\_\_\_

#### Task 2: Correlations for a Subset of the Sample

Determine the relationship between education (educ) and mothers education (maeduc) for male students.

Reduce your output. To select a subsample of students you need to select cases. In the main menu bar:

Data

Select Cases

If condition is satisfied

If

Move Sex into empty box on the right and create statement specifying the gender of interest (i.e., sex = 1 will specify males)

Continue

OK

Now run the correlation (analyze, correlate, bivariate) and produce the scatterplot.

Male respondents	Education	Mother's Education
Education		
Mother's Education		

What proportion of variance in Education is explained by Mother's Education for male students?

Ans:

What do you conclude?

Ans:

Before running further analyses, you need to unselect the cases (Data, Select Cases, All Cases, OK).

### 5. Exercise on computing linear regression?

Ten Corvettes between 1 and 6 years old were randomly selected from last year's sales records in Virginia Beach, Virginia. The following data were obtained, where  $x$  denotes age, in years, and  $y$  denotes sales price, in hundreds of dollars.

$x$	6	6	6	4	2	5	4	5	1	2
$y$	125	115	130	160	219	150	190	163	260	260

Graph the data in a scatterplot to determine if there is a possible linear relationship.

Compute and interpret the linear correlation coefficient,  $r$ .

Determine the regression equation for the data.

Graph the regression equation and the data points.

Identify outliers and potential influential observations.

Compute and interpret the coefficient of determination,  $r^2$ .

Obtain the residuals and create a residual plot. Decide whether it is reasonable to consider that the assumptions for regression analysis are met by the variables in questions.

At the 5% significance level, do the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and, hence, that age is useful as a predictor of sales price for Corvettes?

Obtain and interpret a 95% confidence interval for the slope,  $\beta$ , of the population regression line that relates age to sales price for Corvettes.

Obtain a point estimate for the mean sales price of all 4-year-old Corvettes.

Determine a 95% confidence interval for the mean sales price of all 4-year-old Corvettes.

Find the predicted sales price of Jack Smith's 4-year-old Corvette.

Determine a 95% prediction interval for the sales price of Jack Smith's 4-year-old Corvette.

**Note** that the following steps are not required for all analyses...only perform the necessary steps to complete your problem.

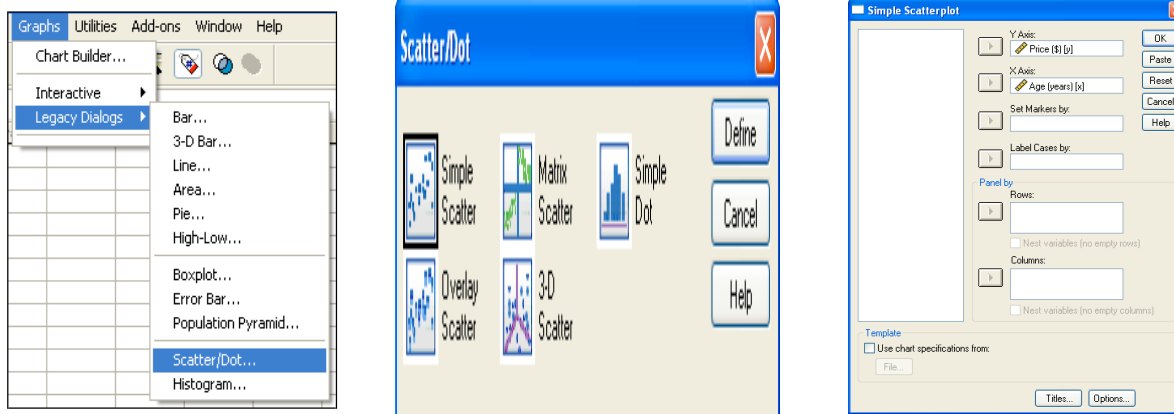
Use the above steps as a guide to the correct SPSS steps.

Enter the age values into one variable and the corresponding sales price values into another variable (see figure, below).

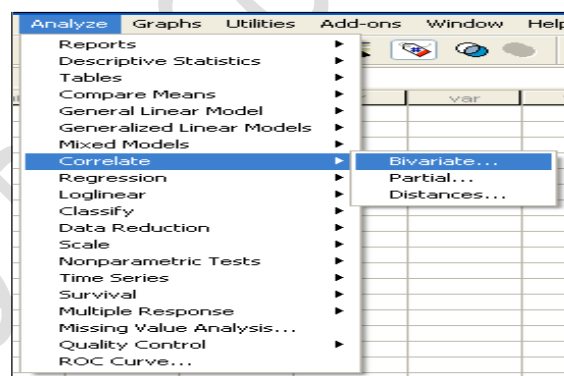
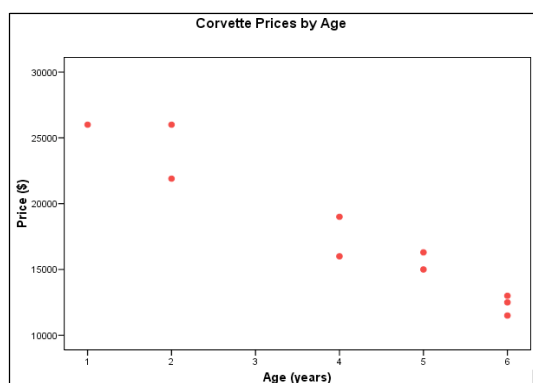
	$x$	$y$
1	6	12500
2	6	11500
3	6	13000
4	4	16000
5	2	21900
6	5	15000
7	4	19000
8	5	16300
9	1	26000
10	2	26000



Select Graphs - Legacy Dialogs -Scatter/Dot... (select Simple then click the Define button) with the Y

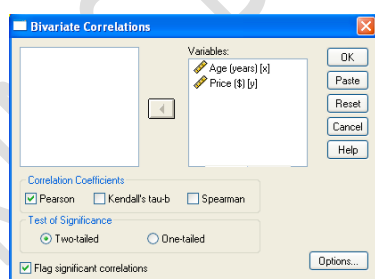


Axis variable (Price) and the X Axis variable (Age) entered (see figures, below). Click “Titles...” to Enter a Descriptive title for your graph, and click “Continue”. Click “OK”. Your output should look similar to the figure below.



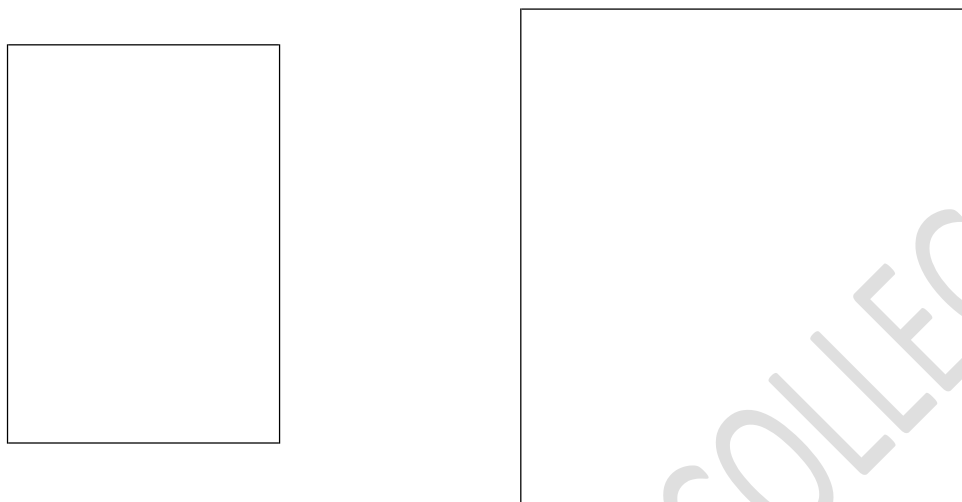
Graph the data in a scatterplot to determine if there is a possible linear relationship. The points seem to follow a somewhat linear pattern with a negative slope. Select Analyze > Correlate > Bivariate... (see figure, below).

Select “Age” and “Price” as the variables, select “Pearson” as the correlation coefficient, and click “OK” (see the left figure, below).



Compute and interpret the linear correlation coefficient,  $r$ . The correlation coefficient is  $-0.9679$  (see the right figure, above). This value of  $r$  suggests a strong negative linear correlation since the value is negative and close to  $-1$ . Since the above value of  $r$  suggests a strong negative linear correlation, the data points should be clustered closely about a negatively sloping regression line. This is consistent with the graph obtained above. Therefore, since we see a strong negative linear relationship between Age and Price, linear regression analysis can continue.

Since we eventually want to predict the price of 4-year-old Corvettes (parts j–m), enter the number “4” in the “Age” variable column of the data window after the last row. Enter a “.” for the corresponding “Price” variable value (this lets SPSS know that we want a prediction for this value and not to include the value in any other computations) (see left figure, below).



Select Analyze > Regression > Linear... (see right figure, above).

Select “Price” as the dependent variable and “Age” as the independent variable (see upper-left figure, below).

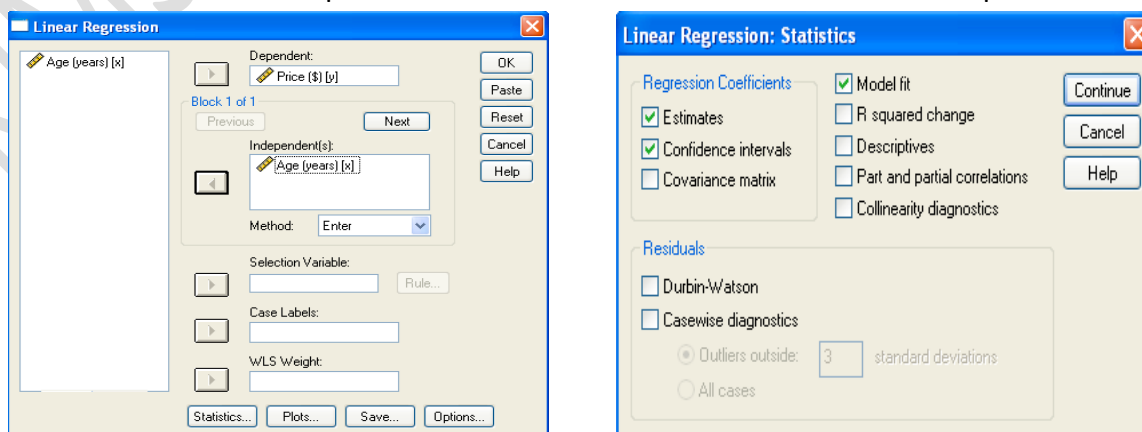
Click “Statistics”, select “Estimates” and “Confidence Intervals” for the regression coefficients, select “Model fit” to obtain  $r^2$ , and click “Continue” (see upper-right figure, below). Click “Plots...”, select “Normal Probability Plot” of the residuals, and click “Continue” (see lower-left figure, below). Click “Save...”, select “Unstandardized” predicted values, select “Unstandardized” and “Studentized” residuals, select “Mean” (to obtain a confidence interval...output in the Data Window) and “Individual” (to obtain a prediction interval...output in the Data Window) at the 95% level (or whatever level the problem requires), and click “Continue” (see lower-right figure, below). Click “ok”

The output from this procedure is extensive and will be shown in parts in the following answers. ere between \$14,552.9173 (LICI\_1) and \$21,445.1410 (UICI\_1)

Click “OK”.

## 6. Exercise on comparing means? How to Compute Means in SPSS?

This tutorial shows how to compute means over both variables and cases in a simple but solid way.



We encourage you follow along by downloading and opening restaurant.sav, part of which is shown below.

	name	gender	v1	v2	v3	v4	v5
1	Ellie Phillips	0	4	6	6	3	3
2	Audrey Nelson	0	5	2	3	4	4
3	Leah Hernandez	0	.	.	.	4	5
4	Justin Rodriguez	1	3	5	5	5	5
5	Brianna Garcia	0	4	3	3	3	3

#### Quick Data Check

Before computing anything whatsoever, we always need to know what's in our data in the first place. Skipping this step often results in ending up with wrong results as we'll see in a minute. Let's first inspect some frequencies by running the syntax below.

#### \*Show data values and value labels in output tables.

set tnumbers both.

#### \*Quick data check.

frequencies v1 to v5.

Result

Right, now there's two things we need to ensure before proceeding. Firstly, do all variables have **similar coding schemes**? For the food rating, higher numbers (4 or 5) reflect more positive attitudes ("Good" and "Very good") but does this hold for all variables? If we take a quick peek at our 5 tabs, we see this holds.

Second, do we have any **user missing values**? That is, do we want to include all data values in our computations? In this case, we don't. We need to exclude 6 ("No answer") from all computations. We'll do so with the syntax below.

Setting Missing Values

#### \*Set 6 as user missing value.

missing values v1 to v5 (6).

#### \*Check again.

frequencies v1 to v5.

### 7. Exercise on one way analysis and variance?

#### How to Perform a One-Way ANOVA in SPSS

##### Purpose of ANOVA

The ANOVA is a statistical technique which compares different sources of variance within a data set. The purpose of the comparison is to determine if significant differences exist between two or more groups.

##### Why ANOVA and not T-test?

Comparing three groups using t-tests would require that 3 t-tests be conducted. Group 1 vs. Group 2, Group 1 vs. Group 3, and Group 2 vs. Group 3. This increases the chances of making a type I error. Only a single ANOVA is required to determine if there are differences between multiple groups.

The t-test does not make use of all of the available information from which the samples were drawn. For example, in a comparison of Group 1 vs. Group 2, the information from Group 3 is neglected. An ANOVA makes use of the entire data set.

It is much easier to perform a single ANOVA than it is to perform multiple t-tests. This is especially true when a computer and statistical software program are used.

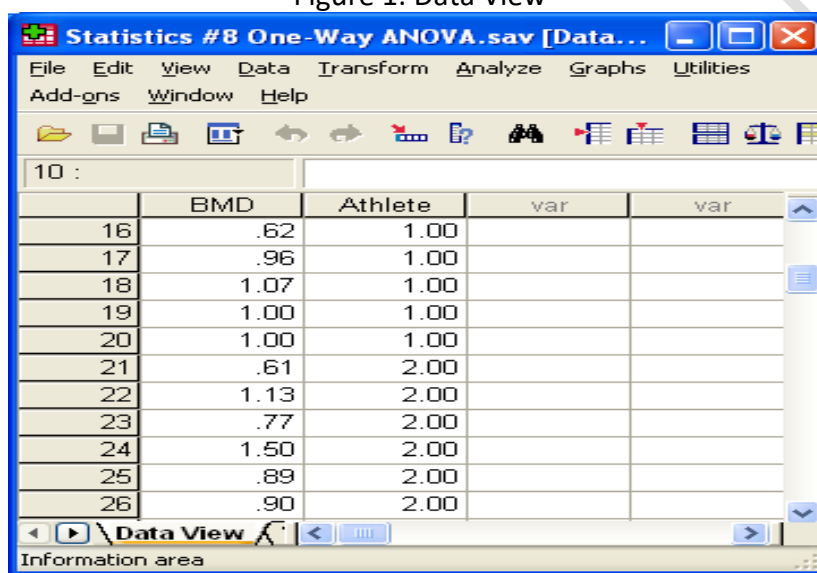
## The Theory in Brief

Like the *t*-test, the ANOVA calculates the ratio of the actual difference to the difference expected due to chance alone. This ratio is called the *F ratio* and it can be compared to an *F distribution*, in the same manner as a *t ratio* is compared to a *t distribution*. For an *F ratio*, the actual difference is the variance between groups, and the expected difference is the variance within groups. Please read the ANOVA handout for more information.

## Let's Roll

Just like with the independent *t*-test, you'll need two columns of information. One column should be whatever your dependent variable is (**BMD** in Figure 1 below), and the other should be whatever you want to call your grouping variable (that is, your independent or quasi-independent variable; this is **Athlete** in Figure 1). Notice that each score in the **BMD** column is classified as being in group 1, group 2, or group 3; SPSS needs to know which scores go with which group to be able to carry out the ANOVA.

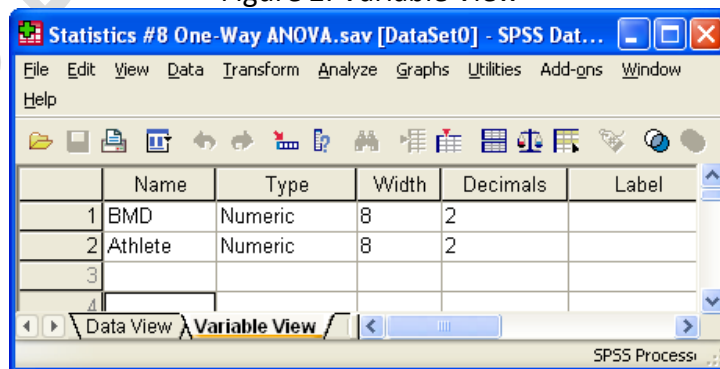
Figure 1: Data View



	BMD	Athlete	var	var
16	.62	1.00		
17	.96	1.00		
18	1.07	1.00		
19	1.00	1.00		
20	1.00	1.00		
21	.61	2.00		
22	1.13	2.00		
23	.77	2.00		
24	1.50	2.00		
25	.89	2.00		
26	.90	2.00		

How did I name the variables **BMD** and **Athlete**? There are two tabs at the bottom of the Data Editor, one labeled Data View, and the other labeled Variable View, as shown in Figure 1: You can toggle back and forth between the Data View (see Figure 1) and the Variable View, which is illustrated in Figure 2:

Figure 2: Variable View



	Name	Type	Width	Decimals	Label
1	BMD	Numeric	8	2	
2	Athlete	Numeric	8	2	
3					
4					

In the Name column, you can type whatever labels you wish for your variables. If you don't type in labels, SPSS will use labels like **VAR001** and **VAR002** by default.

When viewing the results of the ANOVA it will be helpful to know what each **Athlete** number represents. In our case, 1.00 is the Control group, 2.00 is the Swimmer group, and 3.00 is the

Weight Lifter group. While in the Variable View, click on the Values cell of the **Athlete** variable to enter labels for each condition number. See Figure 3.

Figure 3: Name the Grouping Variable

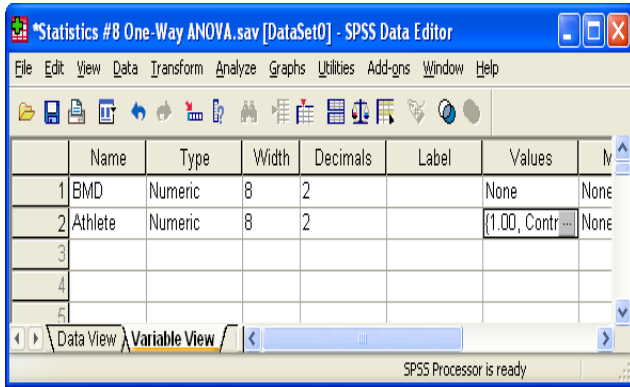
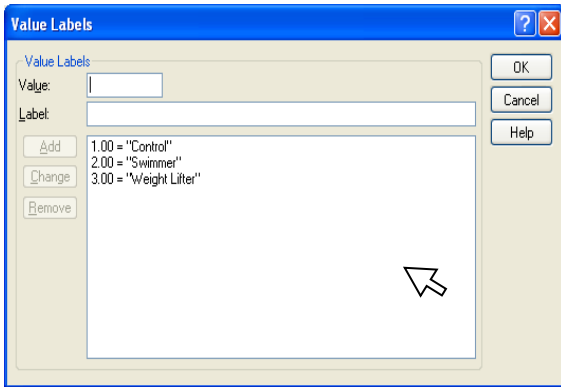
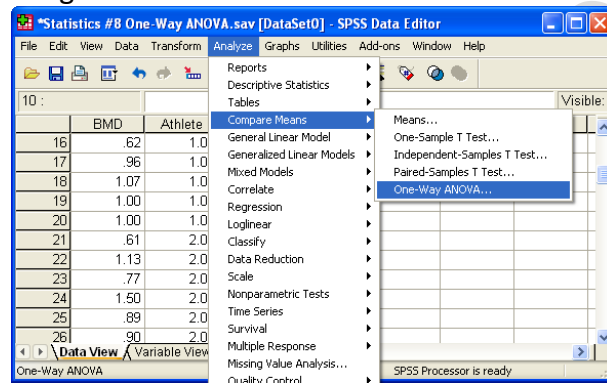


Figure 4: Starting the ANOVA

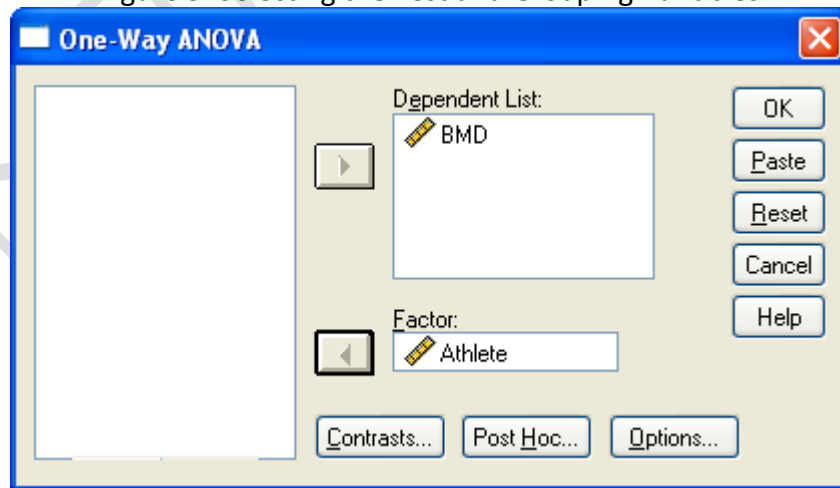


To actually perform the ANOVA, you need to click on the Analyze menu, select Compare Means and then One-Way ANOVA, as in Figure 4.



After this, a dialog box will appear. In this box, you'll be able to select a variable for the "Dependent List" (this is what SPSS calls a dependent variable for this kind of analysis) and a "Factor" (this is the independent variable). I've selected **BMD** for the Dependent List and **Athlete** as the Factor, as show in Figure 5.

Figure 5: Selecting the Test and Grouping Variables



Notice also that there's an Options button. You should click this, and then, in the new dialog box that appears, check the *Descriptive* box (as illustrated in Figure 6a). This tells SPSS to give you descriptive statistics for your groups (things like means and standard deviations). You should also check the *Homogeneity of variance test* box, and the *Means plot* box

## 8. Exercise on computation of Independent Samples and Dependent Samples t-tests?

### Task 1: Confidence Intervals

Using the GSS93 **subset.sav** data set (located within the SPSS program or on the ITS website), you will make interval estimates (confidence intervals) of the parameters for the adult population of the United States.

To get the confidence intervals you will need to go to:

Analyze

Descriptive Statistics

Explore

Transfer your variable to the Dependent List.

Select Statistics: Descriptive and specify the appropriate Confidence Interval for the Mean.

Continue, OK

### 1. What is the average number of years of education (mean highest level of education) of the female's adult population?

	Number of Cases	Mean	95% Confidence Interval	
			Lower Bound	Upper Bound
Years of Education (educ)				

### Task 2: Testing a Hypothesis About Two Related Means

Use the Anxiety2.sav data set (ITS website).

1. Create a new variable that is the difference between trial 1 and trial 4 anxieties (variables trial1 and trial4)

Go to:

Transform

Compute

Type "diff" in target variable box

Click on " trial1" and transfer it into the numeric expression box.

Click on the subtract sign or type in "-"

Click on " trial4" and transfer it into the numeric expression box,

OK

Now make a Histogram of the variable "diff" (go to Graphs, Chart Builder, Select Histogram, and put diff on X axis, OK) and examine the distribution.

1. Does the distribution appear to be normal? \_\_\_\_\_

2. Conduct a t-test for dependent samples

Analyze

Compare Means

Paired Samples T-test

Highlight both trial1 and trial4 variables and transfer them into the Paired variables box.

Under Options specify the confidence intervals as 95%

Continue

OK

3. Now answer the following questions:

What is the correlation between trial1 and trial4? \_\_\_\_\_

Using the 0.05 level of significance, do you reject or retain the null hypothesis? \_\_\_\_\_



### Task 3: Testing Hypothesis about Two Independent Means.

**Problem:** A researcher is interested in the effect of an approach to teaching graduate statistics on statistics anxiety. The statistics course offered by the Educational Psychology department is a lecture based course and a computer based course with no lectures. The content of both courses is exactly the same. There are twelve students in each class. At the end of the course students were asked to fill out the Statistics Anxiety Questionnaire. The results are presented below:

EDPY 500	EDPY 500
Lecture Based Approach	Computer Based Approach
10	27
23	24
11	15
17	19
7	17
4	21
18	26
11	17
11	20
14	29
10	27
19	22

Please enter this data into SPSS. (HINT: To do this, you will have to enter two rows of data: one for the class (the first 12 rows will have an indicator 1 to indicate lecture and the second 12 rows will have an indicator 2 to indicate computer) and one column for the respective anxiety scores).

Test the null hypothesis that the difference between the mean anxiety score of the students taking the lecture based course and the mean anxiety score of the students taking the computer based course is zero.

1. Enter the data into the SPSS file and define the variables.
2. Produce the histograms and examine the distribution of the anxiety scores for both groups.

To do this go to:

Data

Split File

Click on "Organize Output by Groups"

Click on Groups Based On:

Enter Class

Sort File By Grouping Variable

OK

3. Do the scores in both populations appear to be normally distributed?
4. Go back and UNSPLIT the file. Remove "class" from Groups Based On, Click on Analyze All Cases and then select OK.

5. Conduct a t-test for two independent samples:

Analyze

Compare Means

Independent Samples t test

Transfer your dependent variable (anxiety) to Test Variable(s) and the independent variable (teaching approach) to the Grouping Variable bar.

Define the groups...

Type the numerical values for the two groups  
Continue  
Under options select the 95% confidence interval  
Continue  
OK

**6. Examine your output and answer the following questions:**

a) What are the mean anxiety scores for the two groups?

Ans:

b) Is the assumption of homogeneity of variance met? For Levine's test for equality of variances, if the test is non-significant, do not reject the hypothesis that the two population variances are equal.

Ans:

c) What is the mean difference for the two samples?

Ans:

d) What is the value of the t test?

Ans:

e) How many degrees of freedom are there?

Ans:

f) What is the obtained p value?

Ans:

g) Using the 0.05 level of significance, do you reject or retain the null hypothesis?

Ans:

**9. t-test -independent Sample t-test - Paired Sample t-test**

Certainly! The t-test is a statistical test used to determine whether there is a significant difference between the means of two groups. There are two main types of t-tests: the independent sample t-test and the paired sample t-test.

**Independent Sample t-test**

The independent sample t-test, also known as the two-sample t-test, is used to compare the means of two independent groups to determine if they are significantly different from each other.

**Hypotheses:**

**Null Hypothesis (H<sub>0</sub>):** The means of the two groups are equal.

**Alternative Hypothesis (H<sub>1</sub>):** The means of the two groups are not equal.

**Assumptions:**

Both groups are independent.

The data in each group are approximately normally distributed.

The variances of the two groups are approximately equal (homogeneity of variances).

**Example:**

Let's say we want to compare the exam scores of two groups of students, one who attended tutoring sessions and another who didn't. We want to see if there is a significant difference in their mean scores.

**Paired Sample t-test**

The paired sample t-test, also known as the dependent sample t-test, is used when there is a natural pairing between observations in the two groups. It compares the means of two related groups to determine if there is a significant difference between them.

**Hypotheses:**

**Null Hypothesis (H<sub>0</sub>):** The mean difference between the two groups is zero.

**Alternative Hypothesis (H<sub>1</sub>):** The mean difference between the two groups is not zero.

**Assumptions:**

The data within each group are approximately normally distributed.

The paired observations are independent of each other.

The differences between paired observations are normally distributed.

**Example:**

Suppose we want to determine if a new teaching method improves students' test scores. We measure the test scores of the same group of students before and after the teaching intervention.

**Key Differences:**

**Independence of Samples:** The independent sample t-test compares the means of two separate groups, while the paired sample t-test compares means within the same group.

**Sample Relationship:** In the independent sample t-test, the samples are independent of each other, while in the paired sample t-test, the samples are related or paired.

**Assumptions:** The assumptions for each test slightly differ, especially in terms of independence and distribution of data.

Both tests provide a t-statistic and a p-value, which can be used to determine whether the observed difference between the groups is statistically significant. If the p-value is less than the chosen significance level (typically 0.05), we reject the null hypothesis and conclude that there is a significant difference between the groups.

\*\*\*\*\*

**UNIT-IV: Non-parametric Procedures: Two Independent Sample Tests:** Mann-Whitney U-test – Two related Samples Test: Wilcoxon Test, Sign Test – The Runs Test – One - **Sample Test:** Kolmogorov-Smirnov Test – One-Sample Chi-Square Test – **Test for Several Related Samples:** Friedman One-way ANOVA - K-Sample Median Test.

\*\*\*\*\*

### 1. Non-parametric Procedures:

Non-parametric procedures are statistical methods used when the data does not meet the assumptions of parametric statistics, particularly when the data is not normally distributed or when it's difficult to estimate parameters. Here are some common non-parametric procedures:

**Mann-Whitney U Test:** Used to compare two independent groups when the dependent variable is ordinal or continuous, but not normally distributed.

**Wilcoxon Signed-Rank Test:** Compares two related groups when the dependent variable is ordinal or continuous, but not normally distributed.

**Kruskal-Wallis Test:** An extension of the Mann-Whitney U Test, used to compare three or more independent groups.

**Friedman Test:** The non-parametric equivalent of repeated measures ANOVA, used to compare three or more related groups.

**Sign Test:** Compares the medians of two related groups.

**Runs Test:** Checks for randomness in a sequence of data.

**Rank Correlation (Spearman's and Kendall's):** Measures the strength and direction of association between two ranked variables.

Non-parametric procedures are robust to outliers and do not assume specific distributions, making them useful when data does not meet parametric assumptions. However, they may have less statistical power compared to parametric tests, especially with smaller sample sizes.

### 2. Two independent Sample Tests :

Two common non-parametric tests for independent samples are the Mann-Whitney U Test and the Kolmogorov-Smirnov Test. Here's a brief overview of each:

#### **Mann-Whitney U Test:**

Also known as the Wilcoxon rank-sum test, it is used to compare the distributions of two independent samples.

**Assumptions:** Data from both groups are independent, ordinal, or continuous but not normally distributed.

#### **Procedure:**

Rank all the data from both groups combined.

Calculate the sum of ranks for each group.

Use the ranks to calculate the test statistic U.

**Interpretation:** If the p-value is less than the significance level, you can reject the null hypothesis, concluding that there is a statistically significant difference between the two groups.

#### **Kolmogorov-Smirnov Test:**

It tests whether two samples come from the same distribution or not.

**Assumptions:** Data from both groups are independent and continuous.

#### **Procedure:**

Compute the empirical cumulative distribution function (ECDF) for each sample.

Calculate the maximum vertical difference between the two ECDFs.

Use this difference to compute the Kolmogorov-Smirnov statistic.

**Interpretation:** If the p-value is less than the significance level, you can reject the null hypothesis, concluding that there is a statistically significant difference between the two groups.

Both tests are useful when data are not normally distributed or when you have ordinal data. However, the Mann-Whitney U Test specifically compares the medians of the two groups, while the Kolmogorov-Smirnov Test focuses on the overall distributions.

### 3. Two related Sample Test: Wilcoxon Test, Sign Test

Certainly! When dealing with related samples, two common non-parametric tests are the Wilcoxon signed-rank test and the sign test:

#### **Wilcoxon Signed-Rank Test:**

It's used to compare the median of paired (or related) samples.

Assumptions: Data are paired, ordinal, or continuous but not normally distributed.

#### **Procedure:**

Calculate the differences between paired observations.

Discard zero differences (or assign them a rank of zero).

Rank the absolute differences.

Sum the ranks of the positive differences (or negative differences, depending on the direction of the alternative hypothesis).

Use the signed-rank statistic to calculate the p-value.

**Interpretation:** If the p-value is less than the significance level, you can reject the null hypothesis, concluding that there is a statistically significant difference between the two related samples.

#### **Sign Test:**

It's used to test whether the median of the differences between two related samples is zero.

Assumptions: Data are paired and ordinal.

#### **Procedure:**

Calculate the differences between paired observations.

Count the number of positive differences and the number of negative differences.

Use the binomial distribution to calculate the p-value.

**Interpretation:** If the p-value is less than the significance level, you can reject the null hypothesis, concluding that there is a statistically significant difference between the two related samples.

The Wilcoxon Signed-Rank Test is more powerful when the data are continuous, but not normally distributed, and when the sample size is moderate to large. The Sign Test is simpler and more robust to extreme values but may have less power, especially with small sample sizes.

### 4. The Runs Test - One Sample Test

The Runs Test is a non-parametric test used to determine whether a sequence of data appears to be random or not. It's typically used as a one-sample test, meaning you're testing a single sample against a hypothesized distribution or against randomness. Here's how it works:

#### **Runs Test Procedure:**

##### **Data Preparation:**

Convert your data into a sequence of binary outcomes, often by dichotomizing the data into two categories (e.g., "success" or "failure").

Label each observation as a "run" of either success or failure.

Calculate the Number of Runs:

A run is a consecutive series of the same outcome.

Count the number of runs in your data sequence.

Calculate the Expected Number of Runs:

The expected number of runs,  $E(R)$ , under randomness is calculated as:

$$E(R) = \frac{2n_1n_2}{n_1+n_2+1}$$

where  $n_1$  is the number of successes,  $n_2$  is the number of failures, and  $n = n_1 + n_2$  is the total number of observations.

Calculate the Standard Deviation of Runs:

The standard deviation of runs,  $SD(R)$ , is given by:

$$SD(R) = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

**Calculate the Z-Score:**

The Z-score is calculated as:

$$Z = \frac{R - E(R)}{SD(R)}$$

where  $R$  is the observed number of runs.

**Interpretation:**

If the absolute value of the calculated Z-score is greater than the critical value for your chosen significance level (usually  $\alpha = 0.05$ ), you reject the null hypothesis.

**Null Hypothesis ( $H_0$ ):** The sequence is random.

**Alternative Hypothesis ( $H_1$ ):** The sequence is not random.

Considerations:

If the sequence is long, the Runs Test may approximate a normal distribution, allowing you to use a Z-test.

If the sequence is short, you may need to use tables of critical values for the Runs Test.

The Runs Test is useful for detecting patterns or deviations from randomness in a sequence of data, such as in quality control, finance, or time series analysis.

## 5. Kolmogorov-Smirnov Test - One Sample Chi-square Test

The Kolmogorov-Smirnov (K-S) test can also be used as a one-sample test to determine if a sample comes from a specific distribution. Here's how it works:

**Kolmogorov-Smirnov Test Procedure:**

Null and Alternative Hypotheses:

Null Hypothesis ( $H_0$ ): The sample follows a specified distribution.

Alternative Hypothesis ( $H_1$ ): The sample does not follow the specified distribution.

**Data Preparation:**

Arrange your sample data in ascending order.

Calculate the Empirical Cumulative Distribution Function (ECDF):

For each data point, calculate the proportion of data points less than or equal to it.

Calculate the Cumulative Distribution Function (CDF):

For each data point, calculate the cumulative probability from the specified distribution.

Compute the Maximum Absolute Difference (D):

Find the maximum absolute difference between the empirical CDF and the CDF from the specified distribution.

**Calculate the Test Statistic (D):**

The test statistic  $D$  is the maximum absolute difference.

**Compare to Critical Value or P-value:**

For small sample sizes or when the specific distribution is unknown, you can use a table of critical values.

For larger sample sizes or when the specific distribution is known, you can use the K-S test to directly compute the p-value.

If the computed test statistic  $D$  exceeds the critical value or if the p-value is less than the significance level ( $\alpha$ ), reject the null hypothesis.

**One Sample Chi-square Test:** If you're interested in comparing the distribution of your sample to a specified distribution, you can also use a chi-square goodness-of-fit test. Here's how:



**Null and Alternative Hypotheses:**

Null Hypothesis ( $H_0$ ): The sample follows the specified distribution.

Alternative Hypothesis ( $H_1$ ): The sample does not follow the specified distribution.

Data Preparation:

Group your data into bins or categories based on the specified distribution.

Calculate the expected frequencies for each bin.

**Calculate the Chi-square Statistic ( $\chi^2$ ):**

Compute the sum of squared differences between observed and expected frequencies, divided by the expected frequencies for each bin:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \chi^2 = \sum_{i=1}^k \frac{O_i^2}{E_i} - \sum_{i=1}^k E_i$$

where  $O_i$  is the observed frequency,  $E_i$  is the expected frequency, and  $k$  is the number of bins.

**Determine Degrees of Freedom:**

Degrees of freedom ( $df$ ) is  $k-1$ , where  $k$  is the number of bins.

Compare to Critical Value or P-value:

Use the chi-square distribution table or software to find the critical value.

If the computed chi-square statistic exceeds the critical value or if the p-value is less than the significance level ( $\alpha$ ), reject the null hypothesis.

Considerations:

The K-S test is sensitive to the distribution's shape and location, while the chi-square test compares observed and expected frequencies.

The K-S test is generally used when you want to compare the entire distribution, while the chi-square test focuses on comparing specific categories or bins.

Both tests can be used to assess whether a sample follows a specified distribution, but they approach the problem from different perspectives. Choose the appropriate test based on your specific needs and assumptions.

**6. Test for Several Related Samples: Friedman One-way ANOVA - K-Sample Median Test**

When you have several related samples and want to compare their central tendencies, two common non-parametric tests are the Friedman test and the K-sample median test. Let's explore each:

**Friedman One-way ANOVA:**

The Friedman test is used to determine whether there are statistically significant differences between the central tendencies of multiple related samples. Here's how it works:

Null and Alternative Hypotheses:

**Null Hypothesis ( $H_0$ ):** There are no differences between the central tendencies of the related samples.

**Alternative Hypothesis ( $H_1$ ):** There are differences between the central tendencies of the related samples.

Data Preparation:

Arrange your data in a matrix where rows represent the subjects or experimental units, and columns represent the treatments or conditions.

Rank the data within each row, ignoring ties.

Calculate the Friedman Test Statistic ( $\chi^2_{\text{Friedman}}$ ):

**Compute the Friedman test statistic:**

$$\chi^2_{\text{Friedman}} = \frac{12}{n(n+1)} \left( \sum_{j=1}^k R_j^2 - \frac{n^2(k+1)}{4} \right) \quad \chi^2_{\text{Friedman}} = \frac{12}{n(n+1)} \left( \sum_{j=1}^k R_j^2 - 4n \left( \frac{k+1}{4} \right)^2 \right)$$

where  $k$  is the number of treatments,  $n$  is the number of subjects, and  $R_j$  is the sum of ranks for the  $j$ th treatment.

Determine Degrees of Freedom:

Degrees of freedom ( $df$ ) is  $k-1$ .

Compare to Critical Value or P-value:

Use the chi-square distribution table or software to find the critical value.

If the computed Friedman test statistic exceeds the critical value or if the p-value is less than the significance level ( $\alpha$ ), reject the null hypothesis.

### **K-Sample Median Test:**

The K-sample median test (also known as the Mood's median test) compares the medians of multiple related samples. Here's how it works:

#### **Null and Alternative Hypotheses:**

Null Hypothesis ( $H_0$ ): The medians of all groups are equal.

Alternative Hypothesis ( $H_1$ ): At least one median differs from the others.

#### **Data Preparation:**

Arrange your data in a matrix similar to the Friedman test.

#### **Calculate the Median of All Ranks ( $M$ ):**

Compute the median of all ranks.

Calculate the Chi-square Statistic ( $\chi^2$ ):

Compute the test statistic:

$$\chi^2 = \frac{12}{n(n+1)} \sum_{j=1}^k (T_j - n(M+1))^2$$

where  $T_j$  is the sum of ranks for the  $j$ th treatment.

### **Determine Degrees of Freedom:**

Degrees of freedom ( $df$ ) is  $k-1$ .

Compare to Critical Value or P-value:

Use the chi-square distribution table or software to find the critical value.

If the computed test statistic exceeds the critical value or if the p-value is less than the significance level ( $\alpha$ ), reject the null hypothesis.

### **Considerations:**

Friedman test is more appropriate when the assumptions for parametric ANOVA are not met (e.g., normality, homogeneity of variances).

K-sample median test is simpler and can be used when comparing medians across groups without relying on assumptions about the underlying distribution.

Choose the appropriate test based on your data characteristics and research question.

\*\*\*\*\*

**UNIT-V: Multivariate Analysis:** Factor Analysis – Opening Dialog Window – Descriptive Window – Kaiser-Meyer – Olkin(KMO) Measure of Sampling Adequacy – Bartlett’s Test of Sphericity – Extraction of Factors – Principle Component Analysis – Communalities – Total Variance Explained – Eigen Values – Scree Plot – Component Transformation Matrix - Rotated Component Matrix– Interpretation of Output.

\*\*\*\*\*

### 1. Multivariate Analysis: Factor Analysis:

Factor analysis is a statistical technique used to identify patterns in the relationships among variables. It's often employed when dealing with a large number of variables to understand the underlying structure or dimensions that explain the correlations between them.

Here's how it works:

- **Data Collection:** You start with a dataset containing observations on multiple variables. These variables could be anything from survey responses to physical measurements.
- **Correlation Matrix:** Calculate the correlation matrix of the variables. This matrix shows how each variable relates to every other variable in the dataset.
- **Factor Extraction:** Using methods like principal component analysis (PCA) or maximum likelihood estimation, factor analysis extracts a smaller number of underlying factors from the correlation matrix. These factors represent the common variance shared among the variables.
- **Factor Rotation:** After extraction, factors are typically rotated to achieve a simpler, more interpretable solution. Orthogonal rotation methods (such as Varimax) or oblique rotation methods (such as Promax) are commonly used.
- **Interpretation:** Interpret the factors based on the loadings of variables onto them. Variables with high loadings on a factor are considered to be strongly related to that factor.
- **Naming and Application:** Finally, factors are often named based on the variables that load heavily on them. These factors can then be used in further analysis or interpretation.

Factor analysis is widely used in social sciences, psychology, market research, and other fields to reduce the complexity of data and identify underlying structures or dimensions. It helps in understanding the relationships between variables and in simplifying data interpretation.

### 2. Opening Dialog Window – Descriptive Window – Kaiser-Meyer

Certainly! Let's break down what you might expect when opening a dialog window for conducting a Kaiser-Meyer-Olkin (KMO) test and a descriptive window for factor analysis:

#### 1. Opening Dialog Window:

Title: Factor Analysis

Options:

**KMO Test:** Checkbox to include KMO test.

**Descriptive Statistics:** Checkbox to include descriptive statistics.

**Factor Extraction Method:** Dropdown menu to select the method (e.g., Principal Component Analysis, Maximum Likelihood).

**Factor Rotation:** Dropdown menu to select rotation method (e.g., Varimax, Promax).

**Number of Factors:** Input field to specify the number of factors to extract.

**Other Options:** Any other relevant options like scaling methods or handling missing data.

**Buttons:**

**OK:** Confirm selections and proceed with analysis.

**Cancel:** Close the window without performing any action.

## 2. Descriptive Window:

**Title:** Descriptive Statistics

**Content:**

**Mean:** Average value of each variable.

**Standard Deviation:** Measure of the amount of variation or dispersion of each variable.

**Skewness:** Measure of the asymmetry of the distribution of each variable.

**Kurtosis:** Measure of the "tailedness" of the distribution of each variable.

**Range:** The difference between the maximum and minimum values of each variable.

**Correlation Matrix:** Matrix showing correlations between each pair of variables.

**Buttons:**

**OK:** Close the window.

## 3. Kaiser-Meyer-Olkin (KMO) Test:

**Purpose:** The KMO test assesses the sampling adequacy for factor analysis. It measures the proportion of variance among variables that might be common variance.

**Output:**

**KMO Measure of Sampling Adequacy:** A value between 0 and 1, where values closer to 1 indicate that the variables are suitable for factor analysis.

Interpretation: Values above 0.6 or 0.7 are generally considered acceptable.

**Dialog Window:**

**Title:** Kaiser-Meyer-Olkin Test

**Content:** Instructions on the purpose of the test and interpretation of the results.

**Buttons:**

**OK:** Close the window after the user has seen the results.

**Cancel:** Close the window without performing any action.

These windows provide users with the necessary tools to set up and conduct factor analysis, including assessing the suitability of their data through the KMO test and examining descriptive statistics.

## 3. Olkin (KMO) Measure of Sampling Adequacy

The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy is a statistic used to assess the suitability of data for factor analysis. It measures the proportion of variance among variables that might be common variance. In other words, it indicates whether the variables are appropriate for performing factor analysis.

The KMO statistic ranges from 0 to 1. Higher values indicate that the variables are more suitable for factor analysis. Here's how the KMO statistic is interpreted:

**KMO Value near 1:** Indicates that the variables are highly correlated and therefore are likely to be suitable for factor analysis.

**KMO Value near 0:** Suggests that the variables are not highly correlated and may not be appropriate for factor analysis.

Typically, a KMO value above 0.6 or 0.7 is considered acceptable for factor analysis. However, values closer to 1 are preferable.

When conducting factor analysis, you would typically look for the KMO value in the output of your statistical software or package. If the KMO value is provided, you interpret it to determine whether your dataset is appropriate for factor analysis.

If you're using software like SPSS, R, or Python's stats models library, you can find the KMO statistic in the output of the factor analysis function or as a separate result specifically for the KMO test.

For example, in SPSS, after running a factor analysis, the KMO statistic is typically reported in the "KMO and Bartlett's Test" table, where a value above 0.6 or 0.7 would indicate sampling adequacy.

If you're calculating the KMO statistic manually, you'd typically use the correlation matrix of your variables to compute it. Then, you'd interpret the result to determine whether your dataset is suitable for factor analysis.

#### 4. Bartlett's Test of Sphericity

Bartlett's Test of Sphericity is another statistical test used in factor analysis to determine whether the correlation matrix of the variables is significantly different from an identity matrix. In simpler terms, it helps to assess whether there is enough correlation between variables to warrant conducting a factor analysis.

##### Here's how Bartlett's Test of Sphericity works:

**Null Hypothesis:** The null hypothesis states that the correlation matrix is an identity matrix, which means there are no correlations between variables.

**Alternative Hypothesis:** The alternative hypothesis states that the correlation matrix is not an identity matrix, indicating that there are correlations between variables.

**Test Statistic:** Bartlett's test calculates a chi-square statistic based on the difference between the observed correlation matrix and the identity matrix.

**Significance Level:** The chi-square statistic is compared to a critical value from the chi-square distribution at a specified significance level (e.g., 0.05).

**Decision:** If the calculated chi-square statistic is greater than the critical value, the null hypothesis is rejected, indicating that there are significant correlations between variables and factor analysis may be appropriate.

In summary, if Bartlett's Test of Sphericity returns a significant result (i.e., the p-value is less than the chosen significance level), it suggests that the variables are correlated enough to proceed with factor analysis.

You would typically interpret Bartlett's Test of Sphericity along with the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy. While KMO assesses the overall suitability of the dataset for factor analysis, Bartlett's test specifically evaluates the assumption of sphericity (i.e., that variables are uncorrelated in the population).

You can usually find the result of Bartlett's Test of Sphericity in the output of statistical software or packages like SPSS, R, or Python's statsmodels library after conducting a factor analysis. If the p-value is significant, it supports the decision to proceed with factor analysis.

#### 5. Extraction of Factors – Principle Component Analysis – Communalities

In factor analysis, after performing the Kaiser-Meyer-Olkin (KMO) test and Bartlett's Test of Sphericity, the next step is often the extraction of factors. One of the commonly used methods for factor extraction is Principal Component Analysis (PCA).

Here's how factor extraction typically proceeds, along with a discussion of communalities:

##### 1. Extraction of Factors:

**Purpose:** The purpose of factor extraction is to identify the underlying factors that explain the correlations between variables in the dataset.

**Principal Component Analysis (PCA):** PCA is a popular method for factor extraction. It identifies factors that account for the maximum variance in the data.

**Process:** PCA involves eigenvalue decomposition of the correlation matrix of the variables. The resulting eigenvalues represent the amount of variance explained by each factor, and the corresponding eigenvectors represent the factor loadings.

## **2. Communalities:**

**Definition:** Communalities represent the proportion of variance in each variable that is accounted for by the extracted factors. In other words, communalities indicate how much of each variable's variance is explained by the factors.

**Calculation:** Communalities are typically calculated as the sum of squared factor loadings for each variable. For PCA, this is equal to the squared multiple correlations between each variable and all the factors.

**Interpretation:** Higher communalities indicate that a larger proportion of the variable's variance is explained by the factors, suggesting that the variable is well represented by the extracted factors. Lower communalities may indicate that the variable is not well explained by the factors and may need to be reconsidered or excluded from further analysis.

## **3. Output:**

After performing PCA, you would typically get a factor loading matrix showing the relationships between variables and factors, along with communalities for each variable.

The factor loading matrix displays how much each variable contributes to each factor. Higher absolute values indicate stronger relationships between variables and factors.

Communalities are usually presented alongside the factor loading matrix, providing insight into how well each variable is explained by the extracted factors.

In summary, Principal Component Analysis (PCA) is a method commonly used for factor extraction in factor analysis. Communalities provide information about how well each variable is represented by the extracted factors. Together, they help in understanding the underlying structure of the data and identifying meaningful factors.

## **6. Total Variance Explained – Eigen Values – Scree Plot**

In factor analysis, after extracting factors using a method like Principal Component Analysis (PCA), it's important to assess how much of the total variance in the data is explained by the extracted factors. This is typically done through the Total Variance Explained, Eigenvalues, and the Scree Plot.

### **1. Total Variance Explained:**

**Definition:** Total Variance Explained represents the cumulative amount of variance in the original variables that is accounted for by the extracted factors.

**Calculation:** It is computed by summing the eigenvalues of the extracted factors. Each eigenvalue represents the amount of variance explained by its corresponding factor.

**Interpretation:** Higher values indicate that more variance in the data is explained by the factors. Generally, you aim to extract enough factors to capture a substantial portion of the total variance, while avoiding over-extraction.

### **2. Eigenvalues:**

**Definition:** Eigenvalues represent the amount of variance explained by each factor extracted.

**Calculation:** In PCA, eigenvalues are obtained as the diagonal elements of the covariance matrix of the factors.

**Interpretation:** Large eigenvalues suggest that the corresponding factor explains a substantial amount of variance in the data. However, small eigenvalues (close to 1 or less) indicate that the corresponding factor might not be very informative.



### 3. Scree Plot:

**Definition:** A Scree Plot is a graphical representation of the eigenvalues, typically plotted against the number of factors.

**Interpretation:** The Scree Plot helps in determining the number of factors to retain. Eigenvalues are plotted on the y-axis, while the number of factors is plotted on the x-axis. A sharp drop-off in eigenvalues (forming "scree") indicates the optimal number of factors to retain. Factors before the drop-off point are considered significant, while those after are considered negligible.

#### Why it's important:

Total Variance Explained, Eigenvalues, and the Scree Plot help in deciding how many factors to retain from the analysis. Retaining too few factors might lead to underrepresentation of the data, while retaining too many can lead to over-fitting.

They provide insights into the underlying structure of the data and help researchers make informed decisions about the dimensionality of their data.

In summary, these metrics are crucial for assessing the quality of the factor analysis and determining the appropriate number of factors to retain. They provide a balance between capturing enough variance in the data and avoiding over-fitting.

### 7. Component Transformation Matrix - Rotated Component Matrix– Interpretation of Output.

In factor analysis, after extracting factors and optionally rotating them for better interpretability, the output typically includes a Component Transformation Matrix and a Rotated Component Matrix. Here's how to interpret each:

#### 1. Component Transformation Matrix:

**Definition:** The Component Transformation Matrix represents the relationship between the original variables and the extracted (un-rotated) factors.

**Interpretation:** Each cell in the matrix shows the correlation between the original variable and the corresponding un-rotated factor. These correlations are called "factor loadings."

**Purpose:** This matrix helps in understanding which variables are most strongly associated with each factor before rotation.

**Example Interpretation:** A high positive loading (close to 1) indicates a strong positive relationship between the variable and the factor, while a high negative loading (close to -1) indicates a strong negative relationship. Loadings close to 0 suggest weak or no relationship.

#### 2. Rotated Component Matrix:

**Definition:** The Rotated Component Matrix represents the relationship between the original variables and the rotated factors.

**Interpretation:** Like the Component Transformation Matrix, each cell in the Rotated Component Matrix shows the correlation between the original variable and the rotated factor. These correlations are called "rotated factor loadings."

**Purpose:** This matrix is easier to interpret than the un-rotated matrix because the rotated factors are often more clearly defined and easier to understand.

**Example Interpretation:** After rotation, variables tend to load more strongly on fewer factors, making interpretation simpler. Variables with high loadings on a particular factor are considered to be strongly related to that factor, while those with low loadings are less related.

#### **Interpretation of the Output:**

**Identifying Strong Loadings:** Look for variables with high loadings (either positive or negative) on specific factors in the Rotated Component Matrix. These variables are most strongly associated with those factors.

**Naming Factors:** Based on the variables with high loadings on each factor, you can assign a name or interpret the underlying meaning of each factor. For example, if variables related to education, income, and occupation load highly on one factor, you might interpret that factor as "socioeconomic status."

**Interpreting Rotation:** If rotation was used, the rotated factors might be easier to interpret than the unrotated factors because rotation aims to simplify the factor structure.

**Assessing Model Fit:** Evaluate the overall model fit using metrics such as Total Variance Explained, Eigenvalues, and the Scree Plot.

In summary, the Component Transformation Matrix and the Rotated Component Matrix provide insight into the relationship between variables and factors in factor analysis. They help in interpreting the underlying structure of the data and identifying meaningful factors.

\*\*\*\*\*