UNIT II Exploring Data with Descriptive and Inferential Statistics Using SPSS

2.1 INTRODUCTION

Descriptive statistics reflect the nature and characteristics of the data. They reflect information about various aspects of the data and variables. SPSS not only identifies summary statistics but can also explain any error in the data entry. In this unit you would be learning how to use SPSS for finding out summary or descriptive statistics like measures of central tendency, measures of variability around the mean, and measures of deviation from normality. Further, this unit describes how to instruct SPSS to use your data for applying the parametric and nonparametric test.

2.2 OBJECTIVES

After reading this unit you will be able to:

- ❖ Learn about procedure for calculating descriptive statistics for continuous and categorical data.
- ❖ Conversant in using SPSS for applying parametric test.
- ❖ Work with the functions of SPSS for undertaking non parametric test.

2.3 DESCRIPTIVE STATISTICS

Descriptive Statistics are used to precisely describe the data collected. The most common descriptive statistics used are the measures of central tendency (mean, median and mode) and the measures of dispersion (standard deviation, standard error and variance). SPSS can easily bin into the data and can explore various summary statistics.

2.3.1 Frequency Distributions:

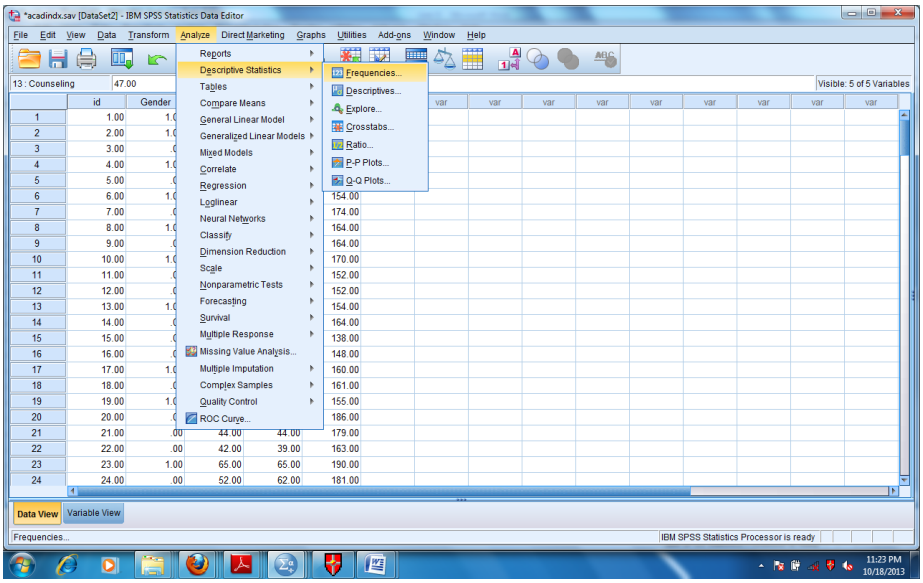
Frequency distribution is essentially the classification of data as per the occurrence of each value. Thus, it counts the number of times a particular value is repeated and gives the number of observations or cases that is denoted for each group and category. Frequency distributions can be calculated using the following steps:

First click go to Analyze and select Descriptive Statistics and then choose Frequencies from the list:

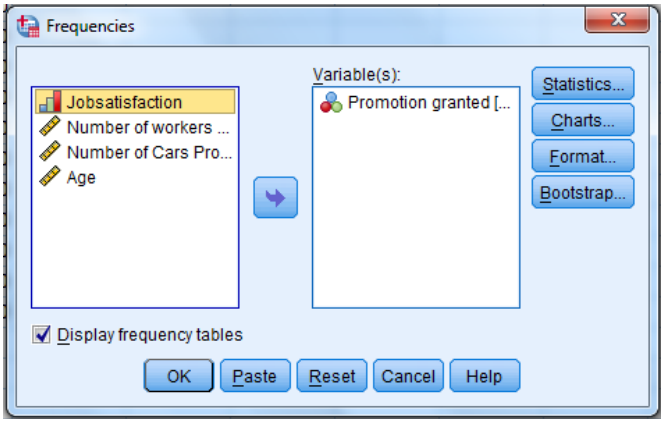
Path:

Analyse > Descriptive Statistics > Frequencies

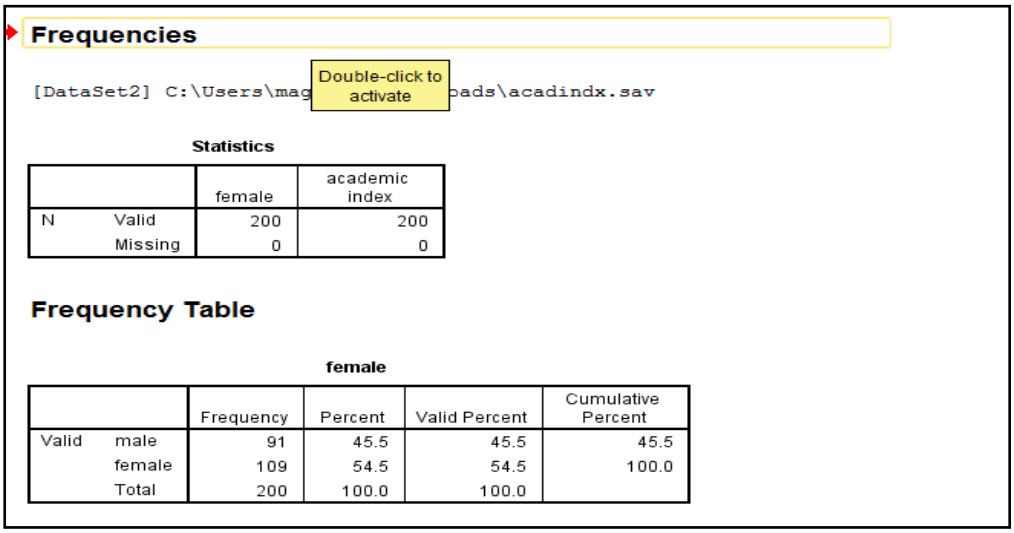
Let us take an example that how many workers are promoted, how many are never promoted and how many are rarely promoted in an organization, for the same refer to the following path-



After choosing the Frequencies from the Analyze Menu a dialogue box would appear as shown below:



Transfer the variable in which you want to compute frequencies in the variables box , keep other options as default and click OK , the output would be generated as under:



Frequencies

[DataSet2] C:\Users\mag Double-click to activate oads\acadindx.sav

Statistics

		female	academic index
N	Valid	200	200
	Missing	0	0

Frequency Table

female

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	91	45.5	45.5	45.5
	female	109	54.5	54.5	100.0
	Total	200	100.0	100.0	

Two tables would be generated in the output; the first table explains the number of female faculty and their academic index. It also displays any missing value in the data.

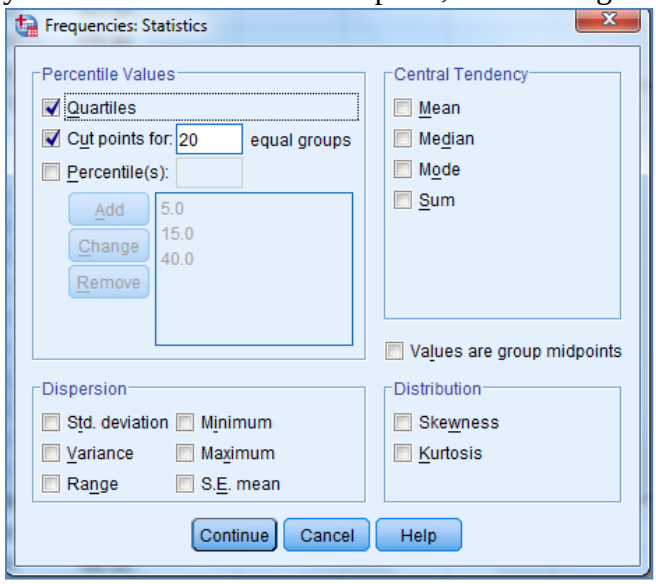
Second table count the number of male and female respondents in the sample along with the percent they constitute in the total. Cumulative percentage of male and female faculty members is also displayed in the output.

2.3.2 Percentiles

The percentiles are used to indicate that how much percentage of a distribution lies below a particular value. SPSS can be used for calculating desired percentiles for a continuous data. Through the percentile option in SPSS data can be divided into different groups. SPSS you can specify number of equal groups in which you can divide the data. These are basically partition values which divides the frequency distribution in equal parts. For Example Q_1, Q_2 and Q_3 are the three quartiles that divides a frequency distribution into four parts so that 25% of the data fall below Q_1, 50% below Q_2 and 75% below Q_3.You can add different percentiles

(for example, the ninetieth percentile), in SPSS. Further, you can also specify cut points for n equal groups.

Lets us compute the percentile for the same data. For computing the percentiles go to Frequencies in the Analyze Menu and select the variable/ variables from the list. Next, when you will click to Statistics Option; a sub dialogue box would appear:



After clicking to the quartile if you wish to calculate percentiles of the total variable then for every 5th percentile value (5th, 10th, 15th etc.) you can divide the percentile scale into 20 equal parts. After specifying 20, you should click continue and OK The table displaying the percentile values would be produced in the output.

Statistics

academic index

N	Valid	200
	Missing	0
Percentiles	5	145.0500
	10	148.1000
	15	152.1500
	20	155.0000
	25	158.2500
	30	161.0000
	35	164.0000
	40	168.0000
	45	171.0000
	50	172.0000
	55	175.5500
	60	179.0000

65	180.0000
70	182.7000
75	184.7500
80	187.0000
85	190.8500
90	196.9000
95	200.0000

This explains that for academic index variable 5% values fall below 145 academic index and 95 % of values are higher than 145 index. 10% of the values fall below 148 academic index and 90 % are higher to it. Similarly you can go on analyzing the data for different percentile value..

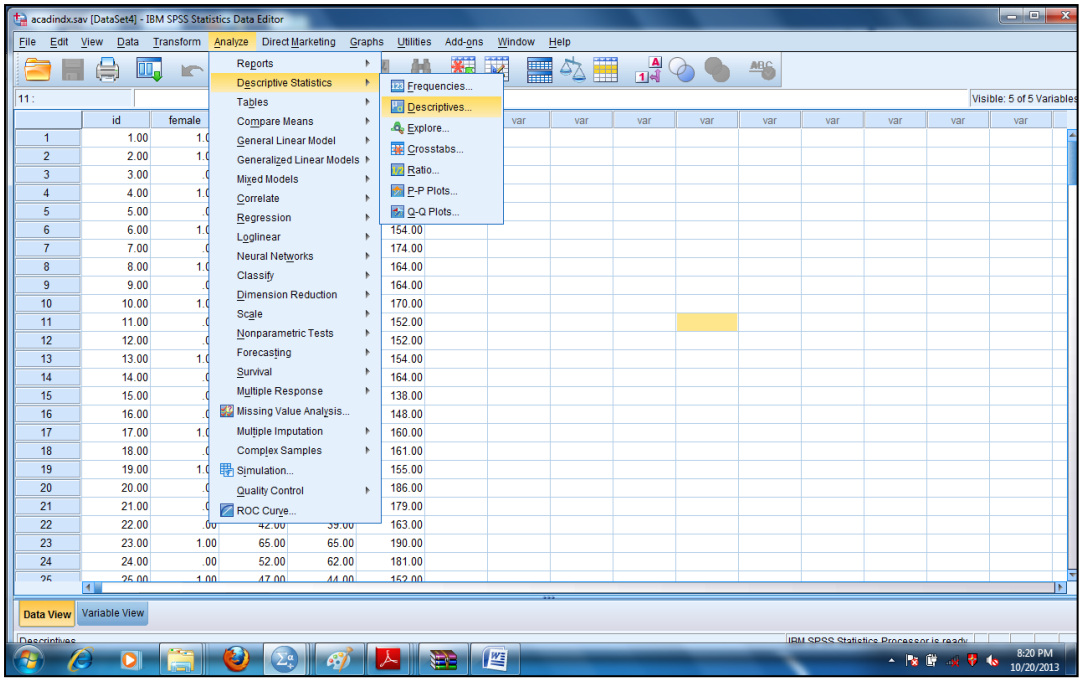
2.3.3 Measures of Central Tendency

Dataset shows a distinguish tendency to cluster around a central point or in the other words central value is the one single value that reflects the nature and characteristics of the entire data series.

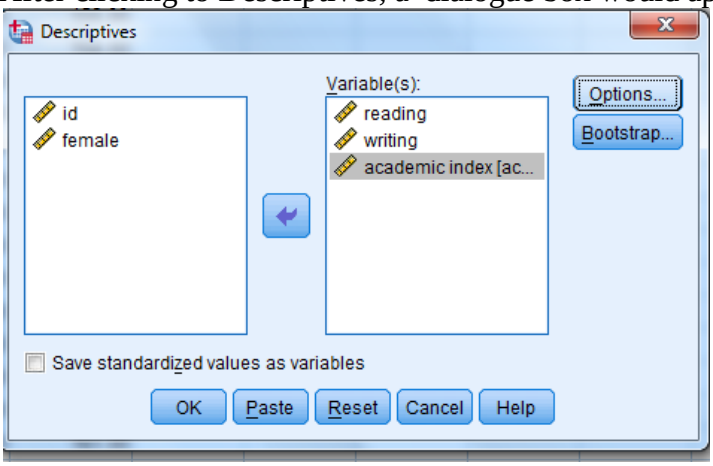
Three measures of central tendency can easily be computed using Descriptive Statistics. Moreover, in SPSS along with various functions or sub functions like correlation, frequencies etc. you have an option of computing these statistics.

2.3.3.1 Mean

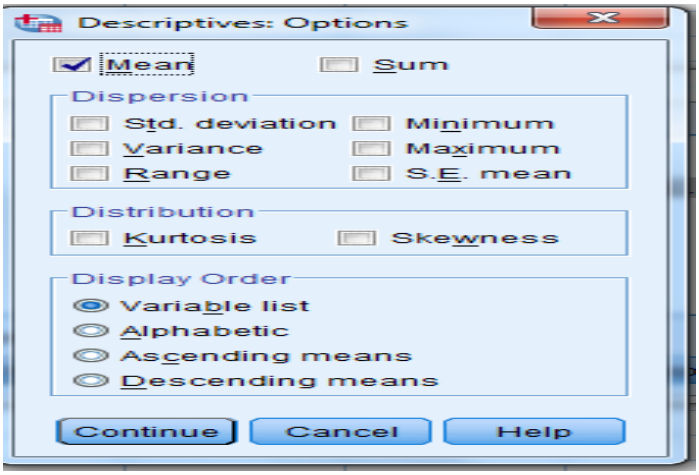
The mean describes a typical or central value of a dataset. This is the most common measure of central tendency. For calculating the mean, first go to analyze and then to Descriptive Statistics . Choose Descriptives from the list.



After clicking to Descriptives, a dialogue box would appear as below:



In the dialogue box you have to click to Options this will results into a sub dialogue box. Select mean in the check boxes. After selecting Click Continue, you would return back to the descriptive dialogue box. You may now click to OK for finding these statistics.



The following output would be generated in the output viewer-

Descriptive Statistics

	N	Mean
reading	200	52.2300
writing	200	52.7750
academic index	200	172.1850
Valid N (listwise)	200	

Thus, from the output it can be deduced that the average readings and writing hours of the faculty members in a month are 52 hours and average academic index of the faculty members is 172.

Note: The mean is a good measure of central tendency when the data set does not contain any exceptionally small or large values.

2.3.3.2 Median

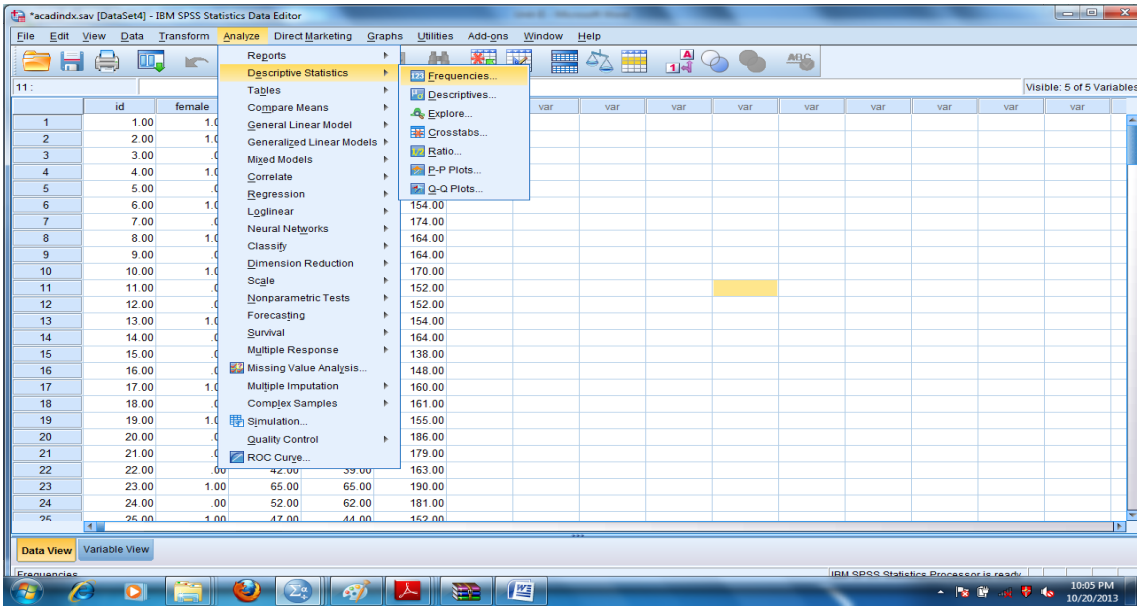
The median is the middle value in a set of data when it is arranged in an array (Smallest to the largest). If the number of values is odd then the median is the middle ranked value and if it is even then it would be the average of two middle ranked observations.

Note: Median is not affected by extreme values.

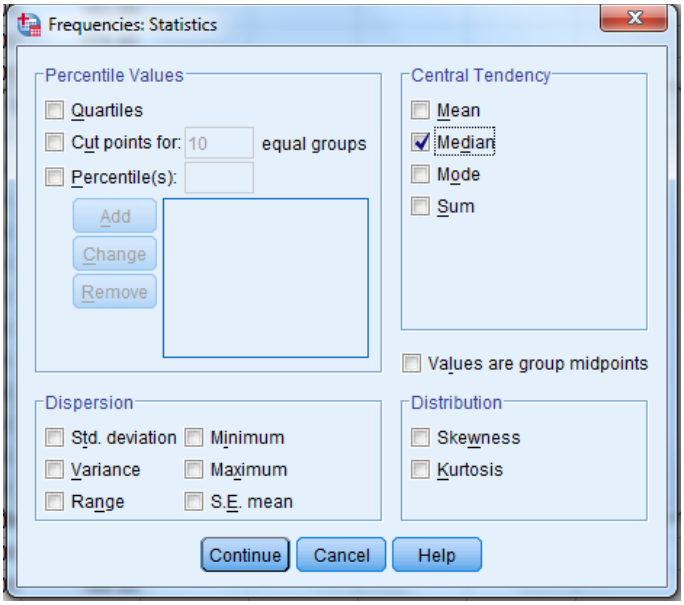
You can compute median using Frequencies.

The procedure for computing Median is as under:

Analyze → Descriptives → Frequencies



After following the above mentioned path, the Frequencies dialogue box would appear then click to statistics and a sub dialogue box would open then select the median in the head Central tendency. Click to continue button and in frequencies dialogue box and then to OK



Refer to the output in the output viewer; the median would be generated along with frequency table in an ascending order.

Statistics

		academic index	writing	reading
N	Valid	200	200	200
	Missing	0	0	0
Median		172.0000	54.0000	50.0000

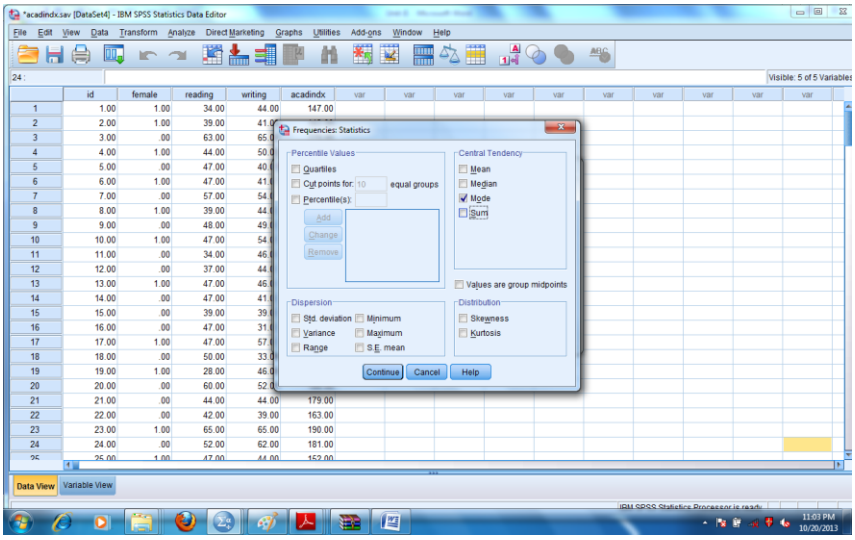
The result conveys that for academic index of 172, half the academic index is equal or below 172 and half the index is equal or above 172.

Further, you can also compute mean using the similar procedure.

2.3.3.3 Mode

The Mode is the value in the set of data that occurs the maximum number of times. The extreme value in the data series does not affect mode.

The Mode can be computed using the similar method as you used for Median.



For the same data using the procedure as used for calculating mean and median, the mode would be produced in output viewer depicted as under:

Statistics

		academic index	writing	Reading
N	Valid	200	200	200
	Missing	0	0	0
Mode		200.00	59.00	47.00

The result conveys that the common academic index is 200 and most of the faculty spends 59 hours in writing and 47 hours in reading.

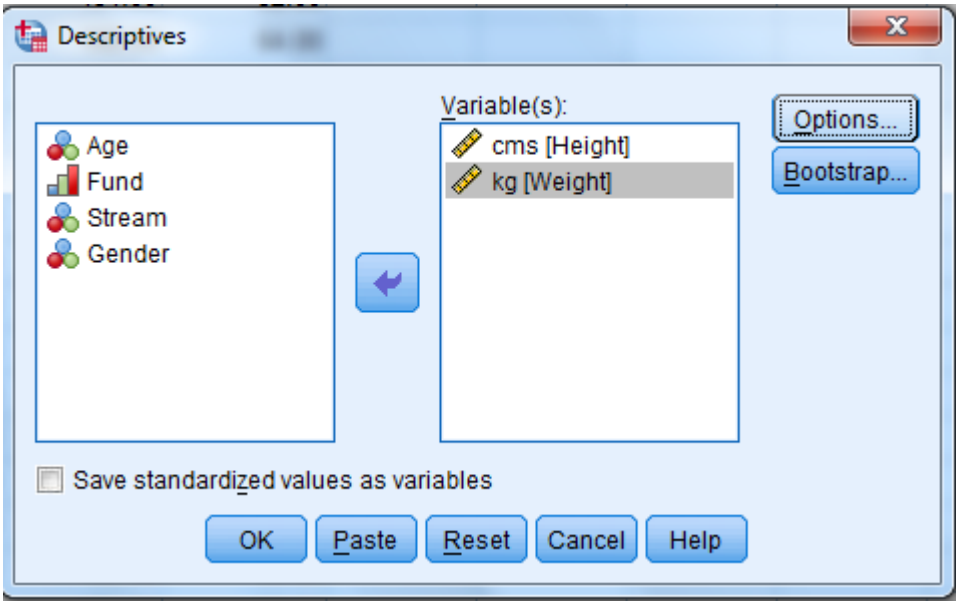
2.3.4 Measures of Variability / Dispersion

Methods of Limits

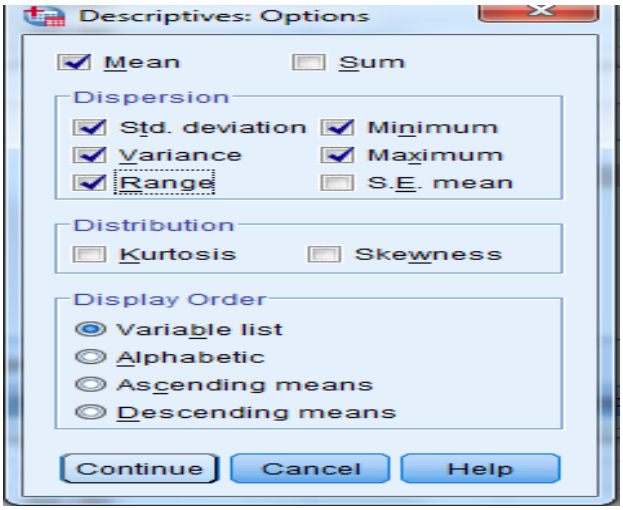
Range is the difference between the highest value and the lowest value and inter quartile range considers middle half of the value of the data. Hence, it measures the spread in the middle 50% of the data. It measures how far from the median one must go either side before it can include one-half the values of the data. It is not influenced by extreme values.

Standard Deviation and Variance explains how the values are distributed or clusters between the two extremes. It assesses how many large values are above the mean value and how many values are below it.

For finding all the above measures first click Analyze and then go to Descriptive you would notice that a dialogue box would appear after clicking to the Descriptives. Move the variables, on which you want to know the patterns, in the variables list box.



Click the Options button and then a sub dialogue box would appear. Select the statistics you want to opt by clicking the check boxes. Further, you can also choose the order /sequence in which variables will be displayed otherwise you can keep the other settings as default. Then select continue and OK



Let us compute these statistics for Height and Weight variable using funds.sav file. The following output would be produced in Output window

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Height in cms	25	27.00	151.00	178.00	168.1600	6.96826	48.557
Weight in kg	25	26.00	50.00	76.00	63.3200	6.58736	43.393

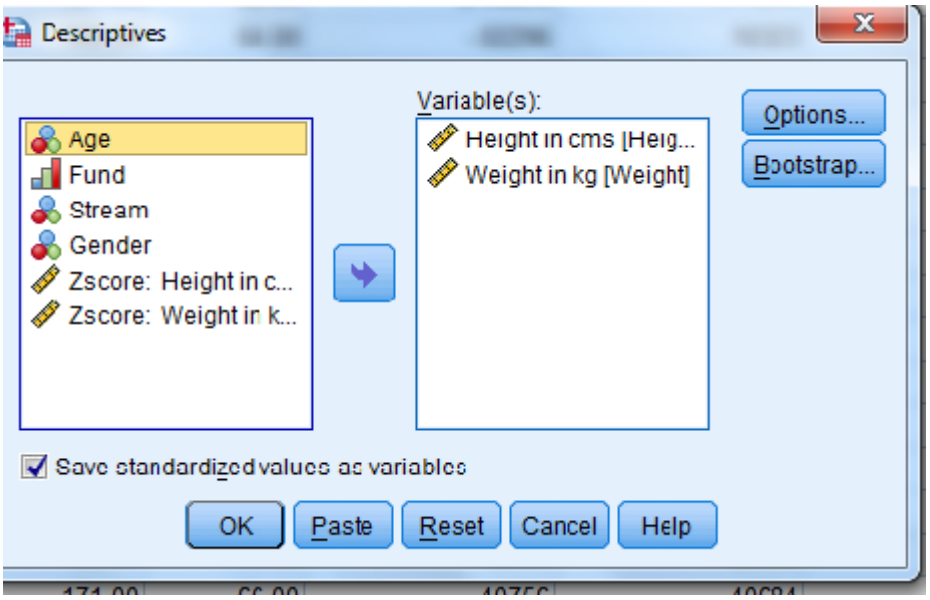
Valid N (listwise)	25						

The **Descriptive Statistics** table consists of the number of observations (**N**) and the **Minimum** and **Maximum** height and weight in the given series. The above table also shows the **Mean** height and weight and the **Standard Deviation**. Height of the respondents has a larger dispersion (6.96) as shown by the standard deviation than the weight of the respondents (6.58).

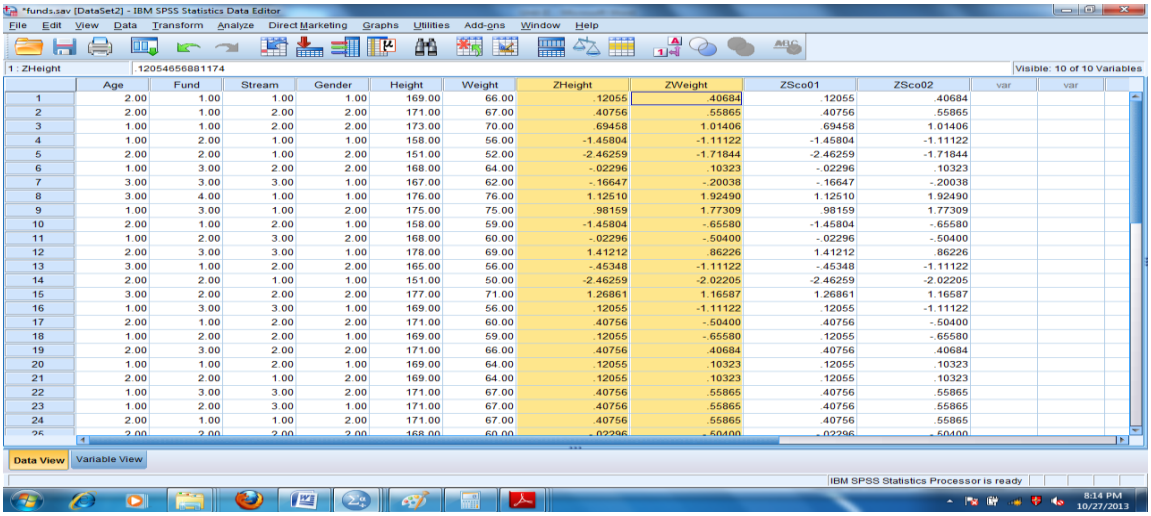
Each column explains the statistics for each variable. If the value of standard deviation is small then it implies that there is high degree of homogeneity or uniformity in the data. This indicates that the Heights and Weights in this sample clusters within 6.96 and 6.68 respectively around the mean of 168.16 and 63.32 respectively (i.e. . This means that 7 out of 25 height lie within the interval ($\bar{X} - 1S = 175.12\ \bar{X} - 1S = 161.2$).

Z Score:

Z scores are useful in finding out the outliers that the value located farthest from the mean. The larger the Z scores the greater the distance from the value to the mean. For finding Z score go to Analyze and then to the Descriptive. Select the variables using the right arrow to the variables list and then click Options select the statistics you want to compute and now click Save Standardized Values as Variables in the check box to save standardized Z-scores.



Z-score for each value would be computed in the data view.



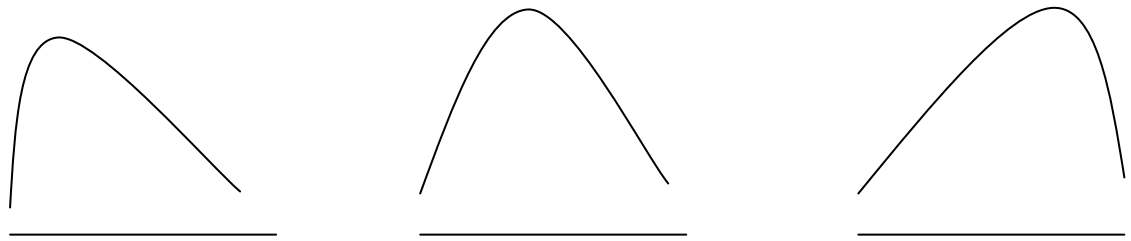
As depicted in the print screen above no apparent outliers are noticed in the data because none of the Z-scores are less than -3 or greater than +3.

Note: As a general rule, a Z-score is considered an outlier if it is less than -3.0 or greater than +3.0

2.3.5 Measures of Deviation from the Normality

Measures of dispersion do not reveal the direction or the pattern of the distribution of the data rather it determines how the values are scattered about the mean. Measures of deviation from the normality assesses whether distribution is symmetrical or not.

Skewness refers to the direction in dispersion that is whether the distribution of values is symmetrical or deviates from symmetry. Measures of skewness indicate the magnitude as well as the direction of skewness in a distribution. When the frequency polygon of a distribution has longer tail to the right of the centre of the distribution it is said to be positively skewed that means a greater number of smaller values. These small values reduce the mean so that mean is less than the median. Similarly, long tail to right explains extremely large values. A value of zero represents symmetry that is half of the curve is the mirror image of another. Skewness value between ± 1.0 is considered appropriate and sometimes value of ± 2.0 are also acceptable in the many cases.



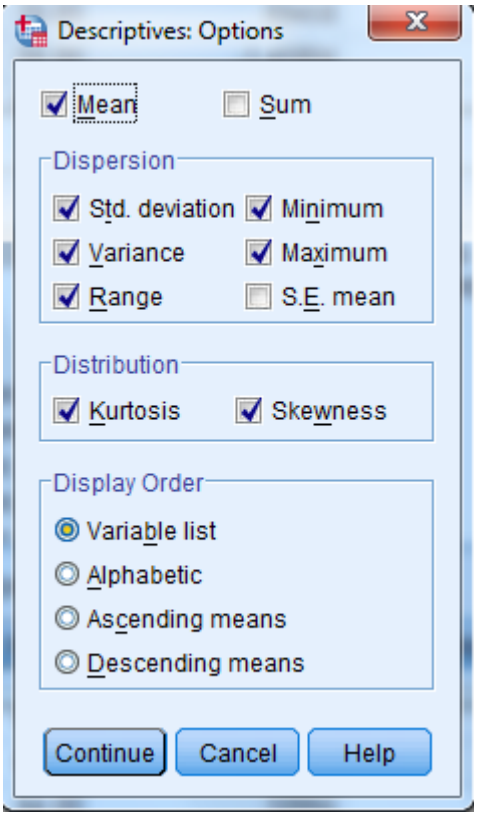
Negative or left skewed Symmetrical Positive or right skewed

Kurtosis is a measure of the " peakedness" or flatness at the top of the distribution. A distribution whose polygon has a high peak is leptokurtic (Positive value for the kurtosis) and those flat at the top is platykurtic distribution (Negative value for the kurtosis). Further, a distribution which neither have a high peak nor have a flat peak at the top is termed as mesokurtic (it denotes normal distribution).

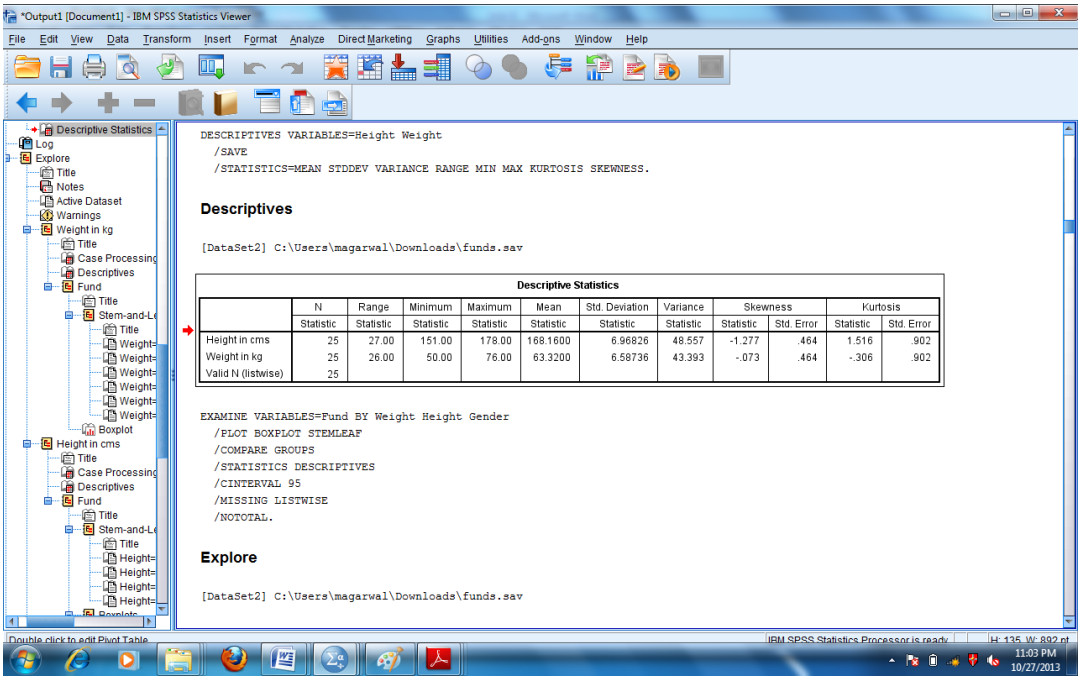
Kurtosis value between ± 1.0 is considered excellent and sometimes value of ± 2.0 are also acceptable in the many cases.

Measures of dispersion can be computed in SPSS by using Descriptives command and also by using Frequencies command. Further, these can also be computed using Explore command which examines normality of the data.

Using Descriptives Command you can select the skewness and kurtosis in the option sub dialogue box.

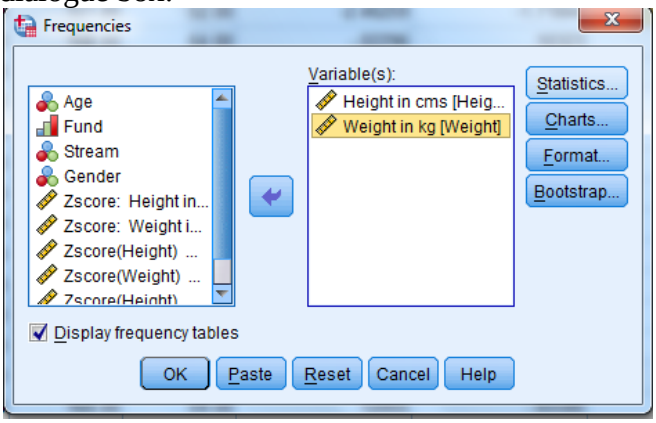


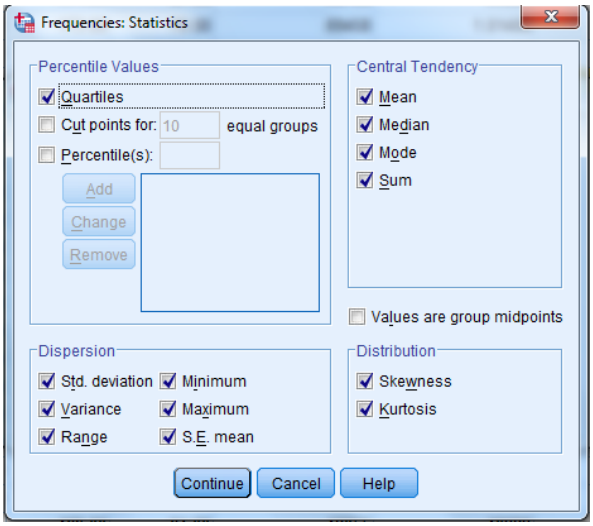
After clicking Kurtosis and Skewness along with the other statistics, the output for fund.sav would appear as under:



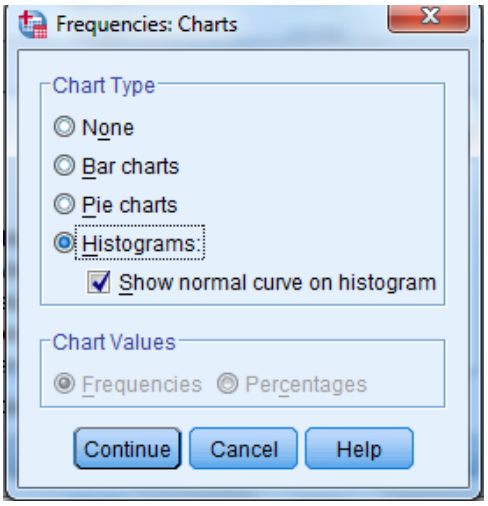
	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Height in cms	25	27.00	151.00	178.00	168.1600	6.96826	48.557	-1.277	.464	1.516	.902
Weight in kg	25	26.00	50.00	76.00	63.3200	6.58736	43.393	-.073	.464	-.306	.902
Valid N (listwise)	25										

Using Frequencies in the Descriptive Statistics select the variables go to the Statistics choose the Skewness and Kurtosis in the Distribution Option and then click continue in the sub dialogue box.





Then in the Frequencies dialogue box again go to Charts select histogram and select show normal curve and then click OK in the Dialogue box.



The output would include descriptive statistics along with the measures of Dispersion with the histogram depicting normal curve.

Frequencies

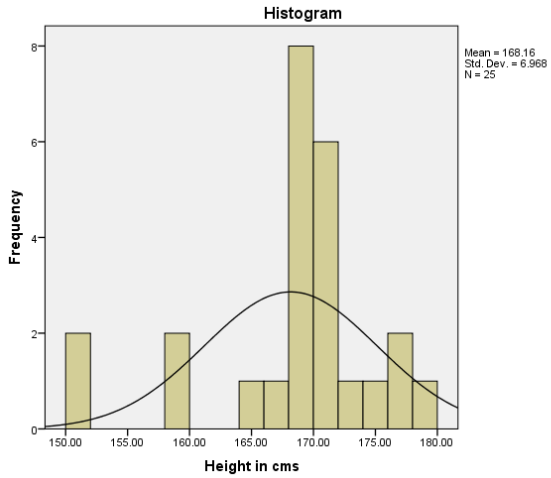
Statistics

Height in cms

N	Valid	25
	Missing	0
Mean		168.1600
Std. Error of Mean		1.39365
Median		169.0000
Mode		171.00
Std. Deviation		6.96826
Variance		48.557
Skewness		-1.277
Std. Error of Skewness		.464
Kurtosis		1.516
Std. Error of Kurtosis		.902
Range		27.00
Minimum		151.00
Maximum		178.00
Sum		4204.00
Percentiles	25	167.5000
	50	169.0000
	75	171.0000

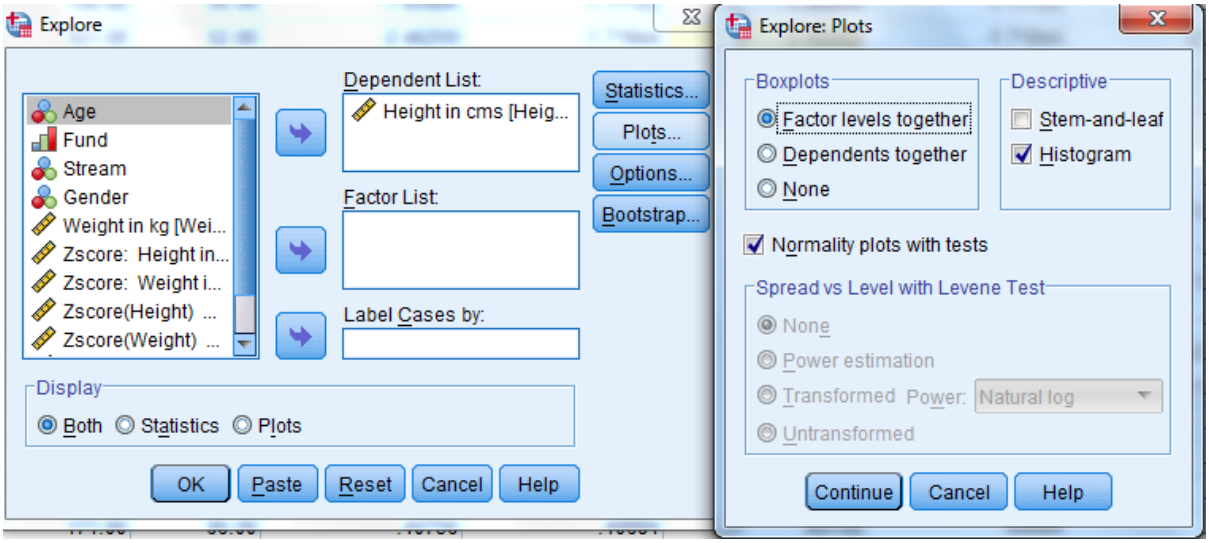
Height in cms

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	151.00	2	8.0	8.0	8.0
	158.00	2	8.0	8.0	16.0
	165.00	1	4.0	4.0	20.0
	167.00	1	4.0	4.0	24.0
	168.00	3	12.0	12.0	36.0
	169.00	5	20.0	20.0	56.0
	171.00	6	24.0	24.0	80.0
	173.00	1	4.0	4.0	84.0
	175.00	1	4.0	4.0	88.0
	176.00	1	4.0	4.0	92.0
	177.00	1	4.0	4.0	96.0
	178.00	1	4.0	4.0	100.0
	Total	25	100.0	100.0	



Using Explore Command

Click to Analyse, Descriptive Statistics and Explore. Select the variable in the Dependent list box in this case let us take height. Then click the Plots button and select Histogram and Normality plots with tests in the check box and then click continue and then OK



The output would contain Case Processing Summaries, Descriptive and Test of Normality. After examining the results for Height and Weight Variable of the respondents, skewness and kurtosis values lies almost in acceptable range. Both of them showed some evidence of negative skewness.

> Refer to the following link to know more about Descriptive Statistics:
> http://en.wikipedia.org/wiki/Descriptive_statistics

2.4 PARAMETRIC TEST

Hypotheses are the statements about the population parameters. During the course of hypothesis testing some inference about the population like mean and proportions are made. Accordingly, for the purpose of decision making, a hypothesis needs to be verified and then accepted or rejected on the basis of the results so derived. This is done with the help of the samples or the observations taken. The process which tests the hypothesis on the basis of sample results enables us to decide whether a hypothesis is to be accepted or rejected.

2.4.1 One –Sample t –test

It is designed to test whether the mean of a distribution differs significantly from some set benchmark or value. For example, let us take the case of data collected about the academic index of the members of the faculty. (acdemicindex.sav)

Question taken for this would be: Does the average reading and writing time devoted by the members of the faculty result in academic index greater than or equal to 174? Does this sample differ significantly from what we consider to be acceptable faculty's performance?

For example, an institution claims that the average academic index of the faculty members is 174. To test it the sample of the academic index of 200 faculty members were collected and analyzed.

On the basis of sample findings the hypothesized value of the population parameter is either accepted or rejected .

In this case null hypothesis would be depicted as under;

$H_0\ \mu = 174$

And alternate hypothesis would be;

$H_1\ \mu \neq 174$

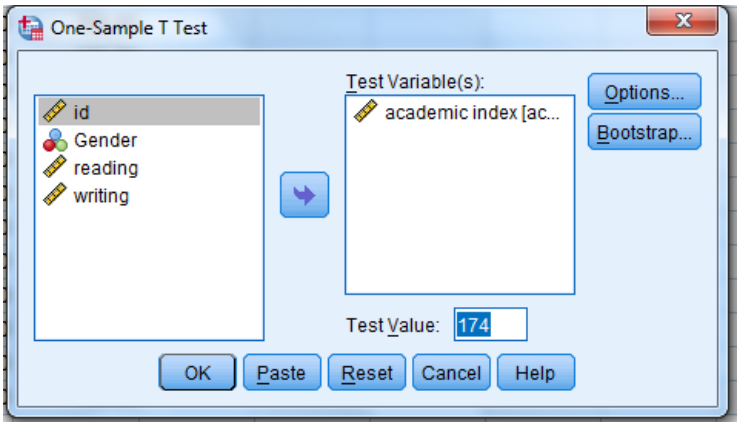
The above hypothesis does not explain whether the academic score is higher or lower. So the hypothesis is non directional i.e. two tailed and if we explain this is higher or lower then it would be one tailed or directional hypothesis.

Process:

You have to first go to analyze and then choose Compare Means and then select one-sample t-test.

A dialogue box would appear, then select the variable and transfer it to the Test variable(s) list box using arrow button. You can also change the Test Value as per your requirement or your benchmark value. Then click OK in the dialogue box. The result would be generated in the output viewer.

Let us take the case of the faculty's academic index variable. Select it in the test variable list and then add 174 in the check box and then click OK



Output would be shown as below:

T-Test

```
[DataSet1] C:\Users\magarwal\Downloads\acadindx.sav
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
academic index	200	172.1850	16.81740	1.18917

One-Sample Test

	Test Value = 174					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
academic index	-1.526	199	.129	-1.81500	-4.1600	.5300

Two tables would be displayed in the output; one is the one sample statistics and the other is one sample test. Sample statistics table gives descriptive measures and one-sample test give the t-statistic along with the 2-tailed significance level i.e. ρ value.

The results of one sample test table shows that there is slightly difference between the hypothesized value and the sample mean, the t-statistics is -1.526 and its associated p-value is .129 which is higher than .005.

This one sample t-test analysis indicates that the mean academic index for this sample of faculty (N=200, M $\bar{X}$=172.18) was slightly lower at the $\rho > .005$ level than the faculty's academic index score (Test value) of 174. The mean difference is simply the observed mean (172.18) minus the test value (174). Therefore, it can be concluded that the difference between the sample-estimated population mean and the compared population mean would not be statistically significantly different.

If p< 0.05 then it can be concluded that the population means are statistically significantly different.

Academic score was statistically significantly higher than the population normal academic score, t (199)=1.526, ρ=.129. Further, the academic score was statistically slightly lower by (95% CI,53 to 4.16) than a normal academic score of 174

On the basis of the first three values (left to right) it can be concluded that

t (199) = -1.526, ρ=.129 → Indicates the probability of obtaining the observed t-

value if null hypothesis is correct

Obtained value of the t-statistic

Degrees of freedom

2.4.2 Independent Sample t-test

Independent sample t-test compares the means of two different samples. It is generally used to study differences in two groups or experiments in two conditions. In this, subjects are randomly selected from the population and then randomly assigned to either experimental or control condition. It is also referred as unpaired or unrelated samples t-test. It helps in comparing the means observed for one variable for two independent samples. Independent sample test is generally used to study the gender differences in performance etc.

SPSS explores descriptive statistics for each group, a Levene's test for equality of variance is calculated and it reports the equal and unequal variance t-values and the 95% confidence interval for the difference in the means.

Let us find out whether there is a difference between the academic score of males and females.

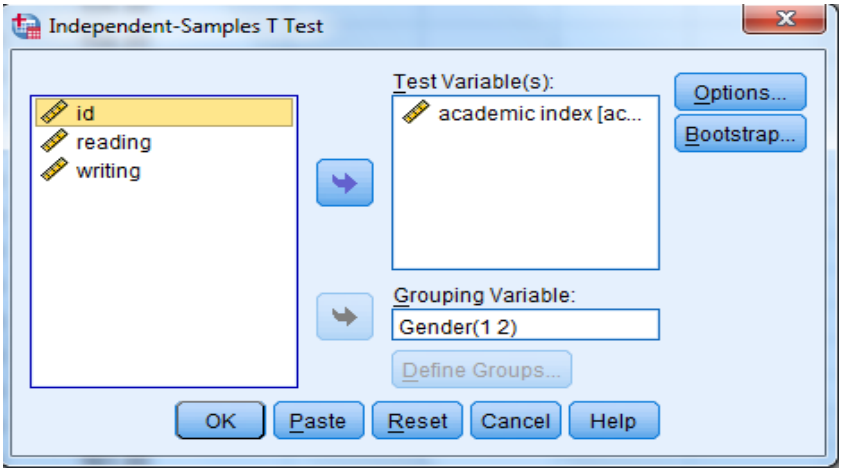
H_0 $\mu\ Academic\ Score\ of\ males\ respondents = \mu\ Academic\ Score\ of\ females\ respondents$

H_1 $\mu\ Academic\ Score\ of males\ respondents \neq \mu\ Academic\ Score\ of\ females\ respondents$

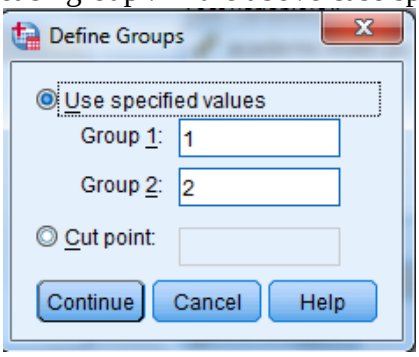
Procedure for calculation

First go to Analyse and then select compare means and independent sample T-test from the Analyze menu.

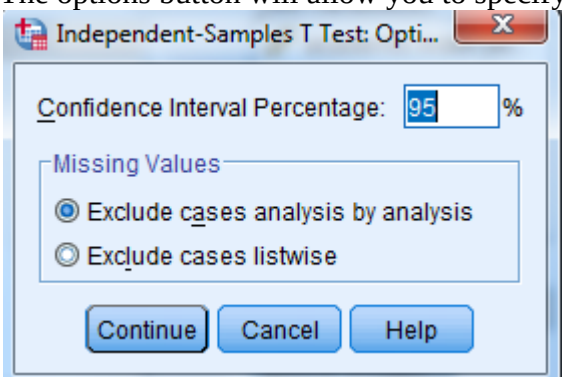
The dialogue box would appear like this ;



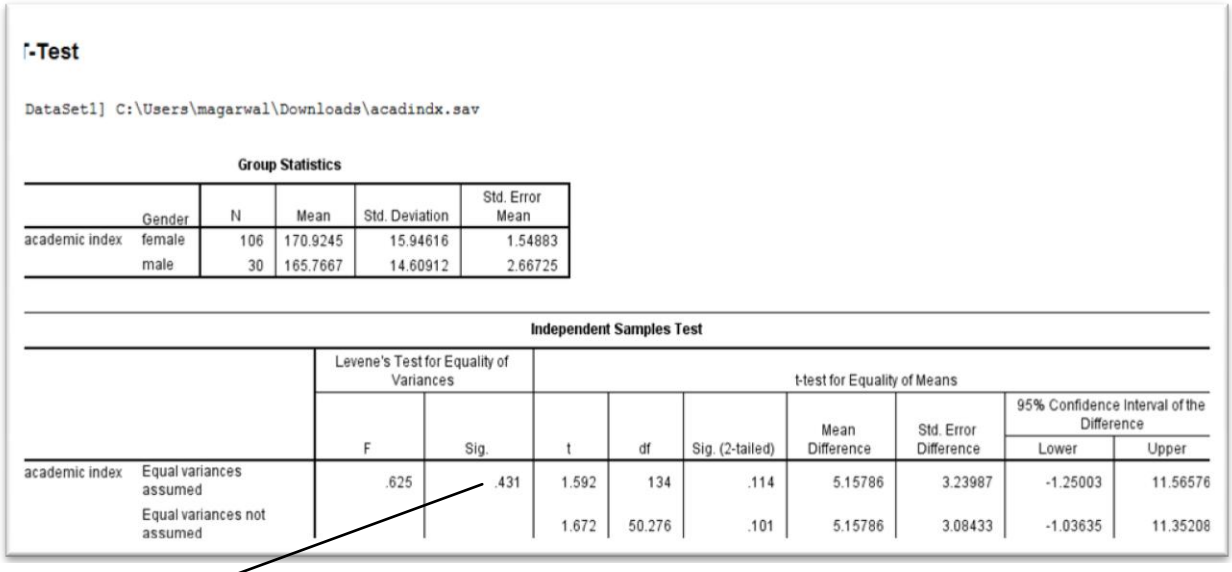
Select academic index variable in the test variable using arrow button. In grouping variables specify the groups to be compared. This can be done by actually entering in specific value for each group . In the above case specify 1 for male and 2 for the female in a sub dialogue box;



The options button will allow you to specify a value for the confidence interval.



After clicking OK the following output would be generated:



T-Test

DataSet1] C:\Users\magarwal\Downloads\acadindx.sav

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
academic index	female	106	170.9245	15.94616	1.54883
	male	30	165.7667	14.60912	2.66725

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
academic index	Equal variances assumed	.625	.431	1.592	134	.114	5.15786	3.23987	-1.25003	11.56576
	Equal variances not assumed			1.672	50.276	.101	5.15786	3.08433	-1.03635	11.35208

Level of significance to determine how t-test should be read.

The output contains two tables – first is Group Statistics and the other is Independent Sample test. The group statistics table explains descriptive measures of the academic score for the independent sample t-test for both the male and female faculty members.

The independent sample test contain Levene's test that determines which t-test to read. That is if the ρ is <0.05 then the variances are not equal and hence you have to read the unequal variances test results in the t-test table to the right. Since equal variances are assumed and the p-value is greater than 0.05 then you will be reading results generated in the first row. When the Levene test shows no significant violations of assumption, you should report the equal variances assumed' version of the t-test. So when the equal variances are assumed then the t test is referred as Student's t-test and in case when equal variances are not assumed then t is termed as Welch t .

The independent t-test analysis indicates that the 106 females have a means of 170.92 academic index, 30 male respondents have a mean of 165.766 and the mean does not differ much significantly from each other and therefore statistically stronger equal variance may be assumed for interpretations. ($\rho = .431$) .

2.4.3 Paired Sample t-test

Paired sample t-test is used to find out the mean difference in related populations to determine whether there is a significant population mean differences or not. Thus, this explains that whether a significant difference occurs between the means of two variables in the same group at different time (before and after the conduct of the event) or in related groups in the same time (students of accounts and finance). It is also referred as the dependent or related sample t-test.

Let us learn using Entrance Test.sav

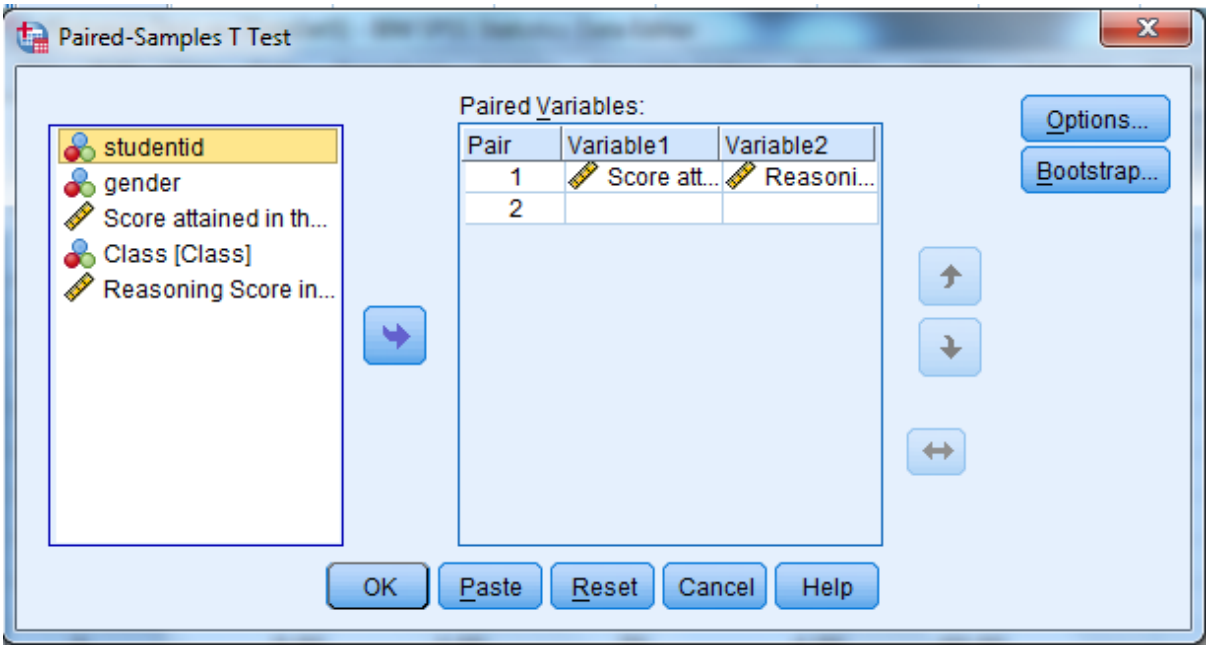
Null hypothesis: There is no significant difference in the score attained in the entrance examination and score attained in reasoning.

Alternate Hypothesis: There is a significant difference in the score attained in the entrance examination and score attained in reasoning.

The following are the steps to follow for calculating paired sample t test.

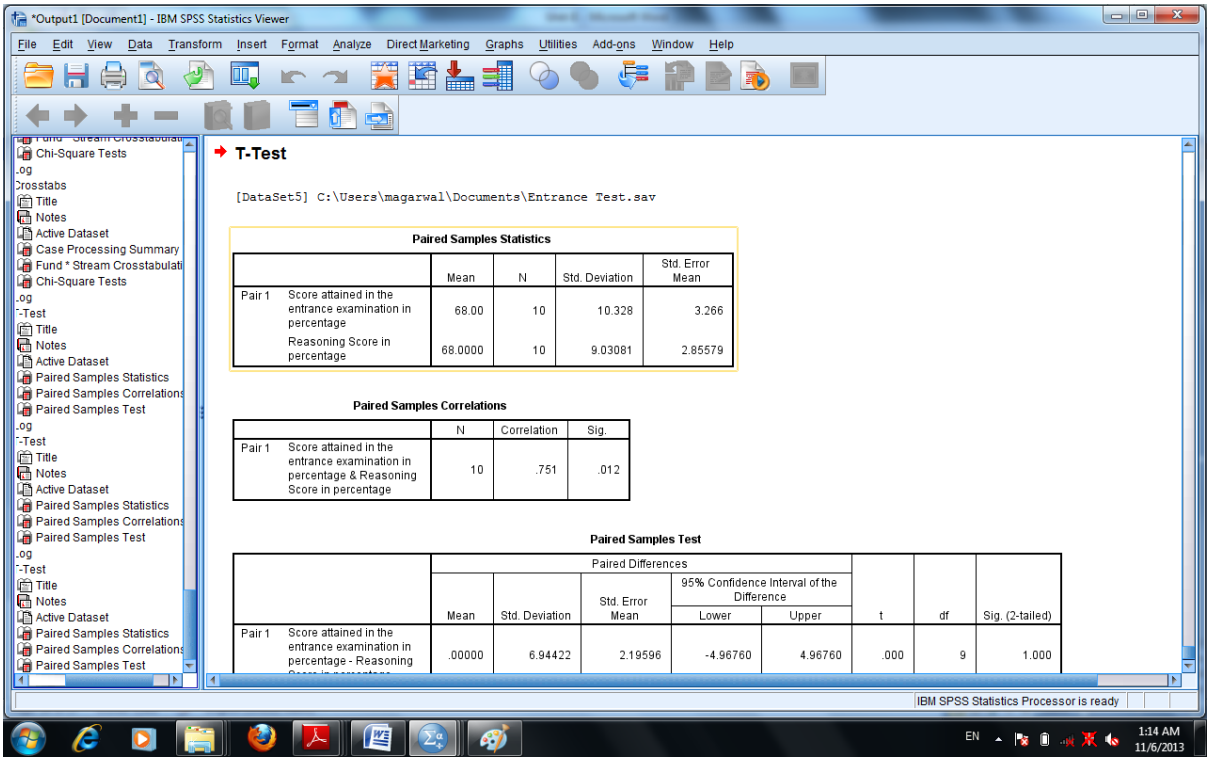
.

First go to Analyze and then to Compare Means and then to paired sample T-test.



Select the pair variables using arrow key. Keeping all the other options in default click OK

The output would be displayed as under:



Three tables would be displayed; the first would depict descriptive statistics for the paired sample. The second would state the correlation between the two variables. In this case it results into highly correlated and is significant. The third explains about paired sample test explaining the difference in mean of overall score and score attained in reasoning (.000). The table generated would depict the value of t-statistic of (.0000) with associated significance

greater than .005. Therefore, we accept the null hypothesis and can conclude that marks attained in reasoning are reflected in the overall score attained in the entrance examination.

Go to the following link to know more about t-test:
http://en.wikipedia.org/wiki/Student%27s_t-test
Do study the assumptions for hypothesis testing before applying these tests.

Check Your Progress-A

Q1. How do you calculate standard deviation in SPSS?

Q2. What is the procedure for testing for differences in means between two groups?

Q3. Distinguish between on sample test and paired sample t-test.

Q4. How do you compute skewness and kurtosis using Analyse command in SPSS?

Q5. Fill in the Blanks with appropriate word or words.

a) A …… hypothesis states that there exists no significant difference between the variables concerned.

b) …….. test is generated in the SPSS while calculating Independent sample t-test. .

c) …… command is used to determine normality of the data which is important for drawing inferences.

d) …….. command is used to describe the data measured on interval or ratio scale.

Q6. Which of the following statements are true or false?

a) For finding Z score go to Analyze and then to the Cross tab option.

b) Median is not affected by extreme values.

c) Frequencies Command does not display the valid and missing frequencies.

2.5 NON PARAMETRIC TEST

SPSS provides a number of non parametric techniques as Chi square, Binomial , Runs etc. In this unit we would only be concentrating on the computation of Chi-square and Mann Whitney test using SPSS.

2.5.1 Chi-Square Test

Chi-square is used to examine the association between two or more variables measured on categorical scale. It test hypothesis whether the two or more samples are drawn from the same population share similar characteristics or not. Thus, chi square test statistic is equal to the squares difference between the observed and expected frequencies, divided by the expected frequency in each cell of the table summed over all cells of the table.

The chi-square is used by the researchers in two ways first as a test of independence and the other is goodness of fit. The test of independence is use to evaluate whether there is any association between two variables and the goodness of fit is used to identify whether there is a significant difference between observed and expected frequencies.

Cross tabulation analysis is used to assess the intersections between independent and dependent variables and to examine the relationship between the two variables. Cross tabs command of SPSS produces a table of different cells with associated frequencies inserted in

each cell by crossing levels. So if you want to cross tabulate in funds.sav and you want to know that how many men and women respondents are from science, commerce and arts stream (funds. Sav) then you can do it by selecting Crosstabs. This would produce a table of 6 cells associated frequencies inserted in each cell by crossing two level of gender with three different streams {3x 2= 6}

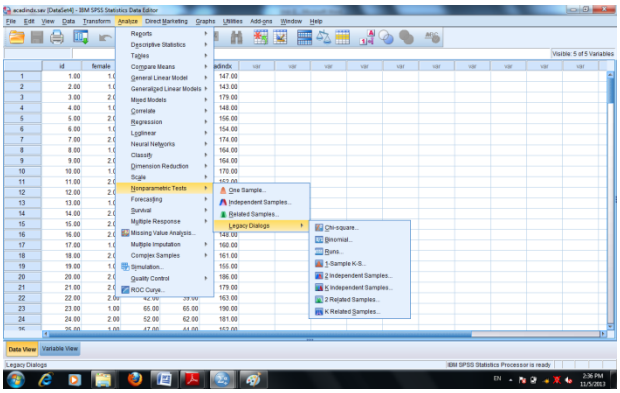
2.5.1.1 Chi square as Goodness of fit

It depicts analysis of single categorical data. It is opted when researcher wanted to know whether there is a significant difference between the observed frequencies and expected frequencies of that variable. Chi-square is used to test whether the fit between the observed and expected distribution is good.

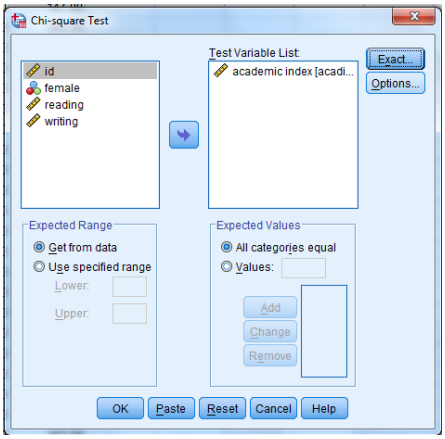
Generally, this procedure is followed when you do not know what the distribution type is, nor do you care. All you are interested in is testing whether the "measured" relative frequencies/proportions are similar to the "expected" relative frequencies/proportions.
Let us formulate the following hypothesis for finding this-

H_0= There is not much difference in the academic index of faculty members in the institution.
H1= There is a significant difference in the academic index of faculty members in the institution.
The procedure to conduct the Chi-square in this case is as follows. First go to Analyze menu select Non parametric test and then go to Legacy Dialogue and then to Chi-square option.

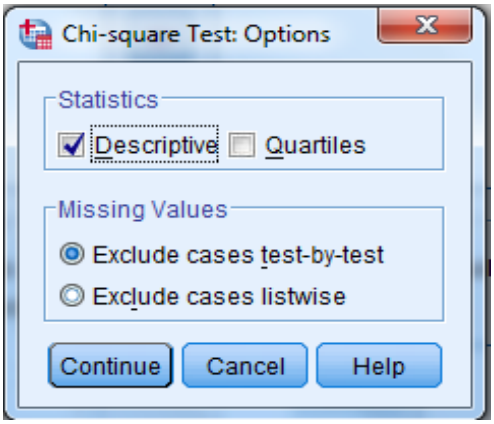
Analyse → Non parametric Tests → Legacy Dialogue → Chi-square



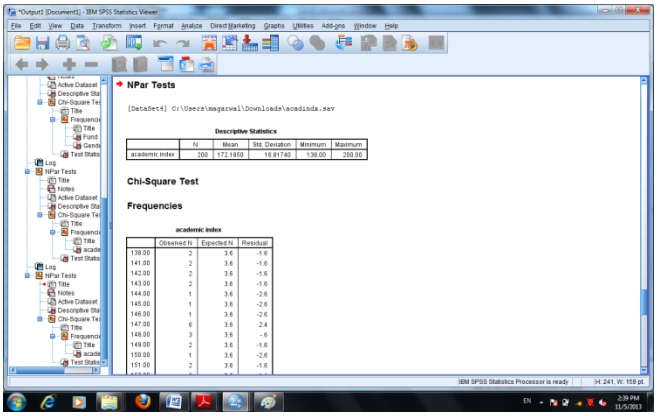
A dialogue box would appear as below:

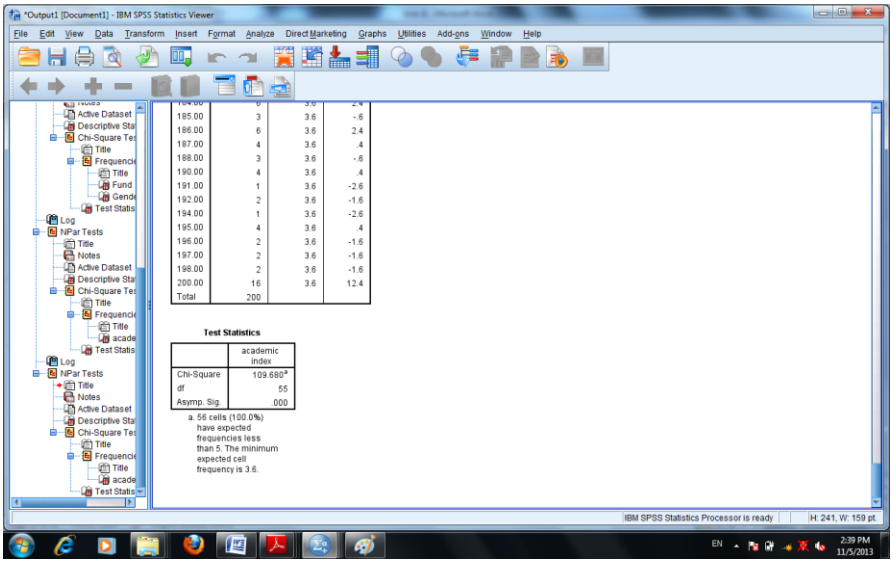


Select the variables on which you want to find goodness of fit in this case *Funds* using arrow button to the Test Variable List. Then click on the Options button, a sub dialogue box would appear.



Click on the Descriptive check box and then click continue and then click OK in the dialogue box. The results of the chi-square would be generated in the output.





The output contains three tables, first one shows the summary or descriptive statistics as it depicts minimum value, maximum vale, mean and standard deviation. The next table identifies the observed, expected and residual values of the academic index. The last table shows the Chi square value of 109.68 at 55 degrees of freedom. Further, $p<0.05$ is significant at 55 degree of freedom, showing that there is a significant difference in the expected and observed frequencies. As such we reject Null hypothesis and accepts the alternate hypothesis that there is a significance difference in the academic index of faculty members in the institution.

Test Statistics

	academic index
Chi-Square	109.680[a]
Df	55
Asymp. Sig.	.000

a. 56 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 3.6.

If there is a large difference between observed and expected values then the chi square would also be large which narrates that there is significant difference in observed and expected frequencies. If $p< 0.05$ then fit is considered as good one which means divergence between observed and expected frequencies is attributable to fluctuations in the sample.

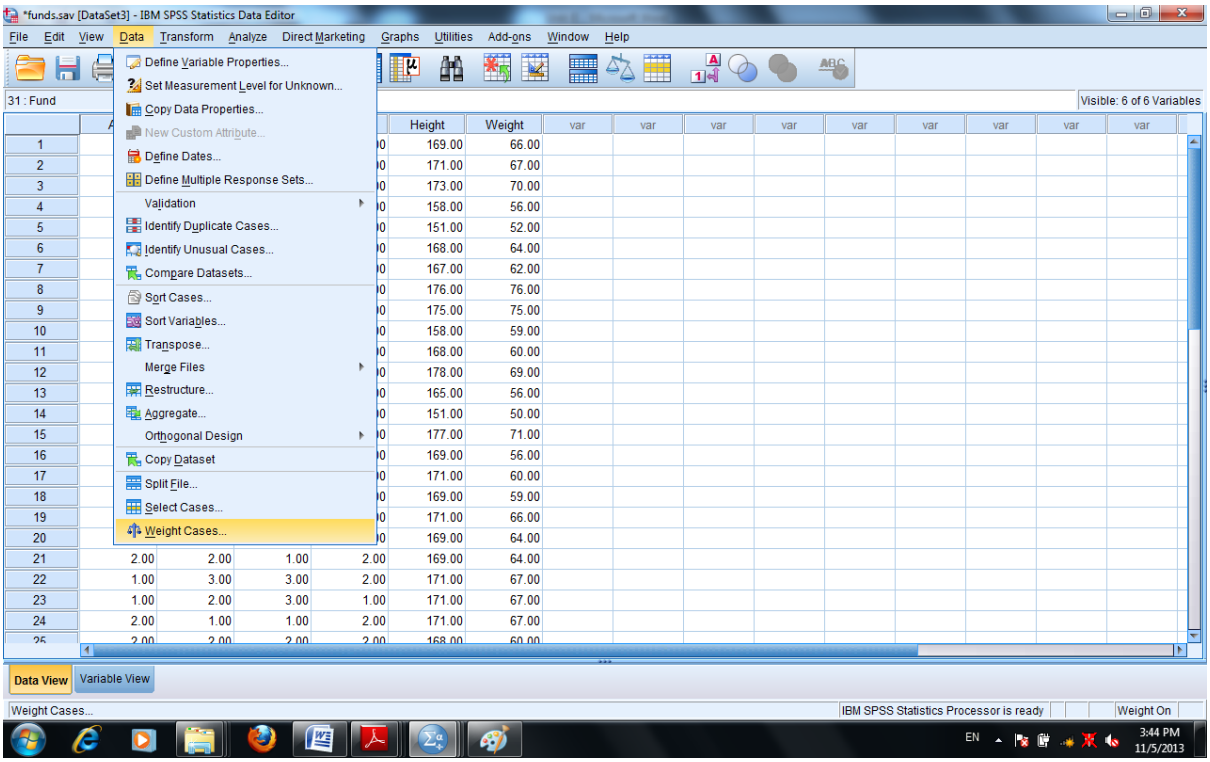
2.5.1.2 Chi square test for Goodness of fit (based on weigh cases)

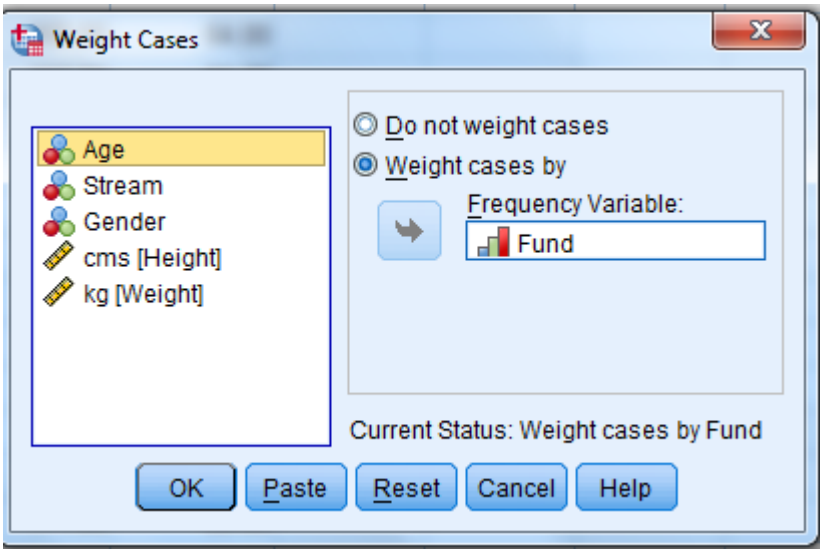
As a researcher you want to know that from funds.sav whether there is a difference in the funds selected on the basis of the cost. You assume level of significance at 5%. The hypothesis for the same is formulated as under:

H0= There is no difference in the funds invested on the basis of cost.

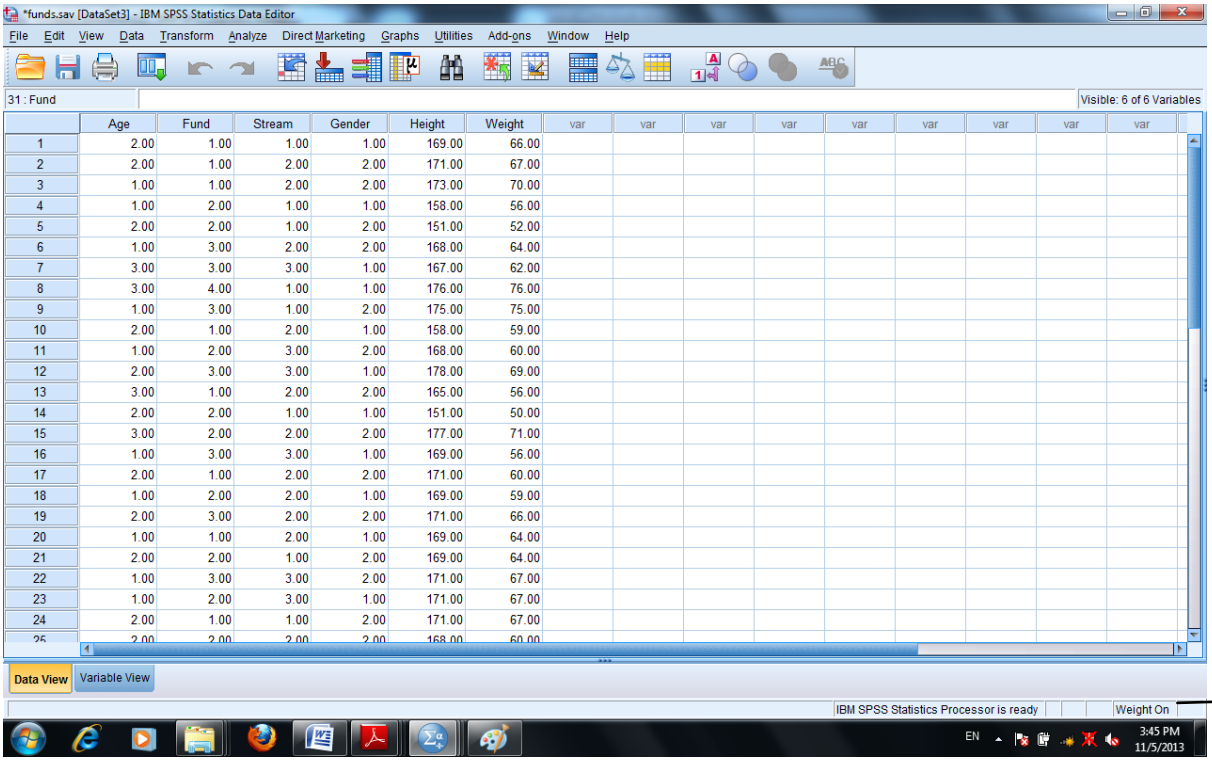
H1= There is a significant difference in the funds invested on the basis of cost.

For this first go to data and select the last item as weigh cases then a dialogue box would open.



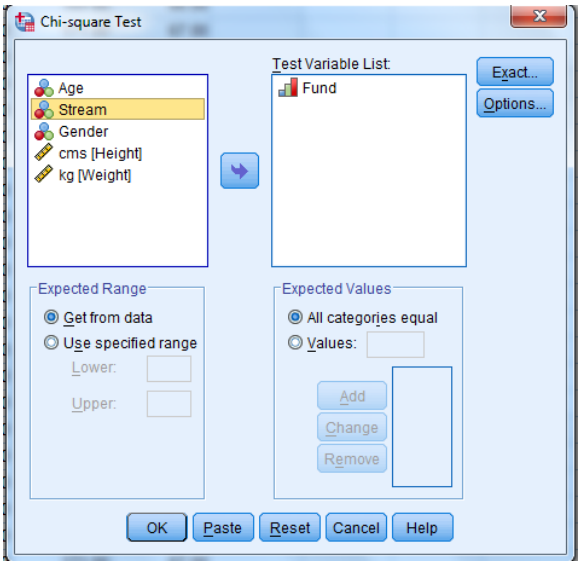


You go to Weigh Cases By and transfer the variable Funds in frequency variable box using arrow button and press OK The completed command would be displayed in the status bar.



Weight On

Now go to Non parametric test and select legacy dialogue and then Chi square.



A dialogue box would appear, then select fund variable in Test Variable List using arrow button. Now click on options and select descriptive in the options sub dialogue box. Then click continue and OK , the following output would be generated in the output window.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Stream	51	1.9804	.81216	1.00	3.00

Chi-Square Test

Frequencies

Fund

	Observed N	Expected N	Residual
Lowest Cost Fund	8	12.8	-4.8
Second Lowest Cost Fund	18	12.8	5.3
Third lowest cost fund	21	12.8	8.3
Highest Cost Fund	4	12.8	-8.8
Total	51		

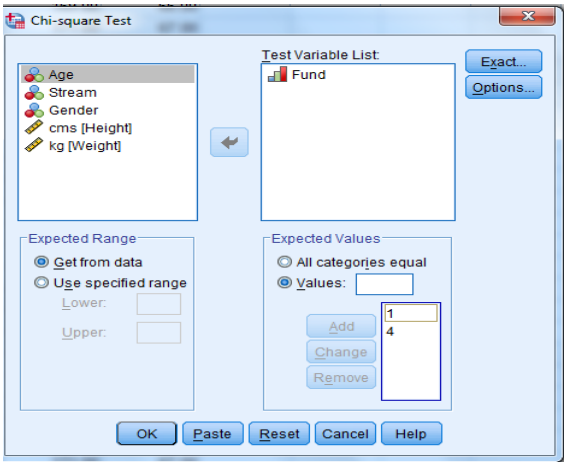
Test Statistics

	Fund
Chi-Square	15.275[a]
df	3
Asymp. Sig.	.002

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 12.8.

Thus, it explains that Chi-square value of 15.275 at 3 degrees of freedom and N being 51 , $p<0.05$ is significant at 3 degrees of freedom , showing that there is a significant difference between expected and observed frequencies. Hence, null hypothesis is rejected and accept the alternate hypothesis that there is a significant difference in the funds invested on the basis of cost.

You can also mention that the expected value for computing χ^2 in the Chi-square dialogue box in sub head Expected Values.



This test can also be used to find out the changes in the dependent variables on the whole but it fails to ascertain the changes in each variable.

2.5.1.3 Chi-square as Test of Independence

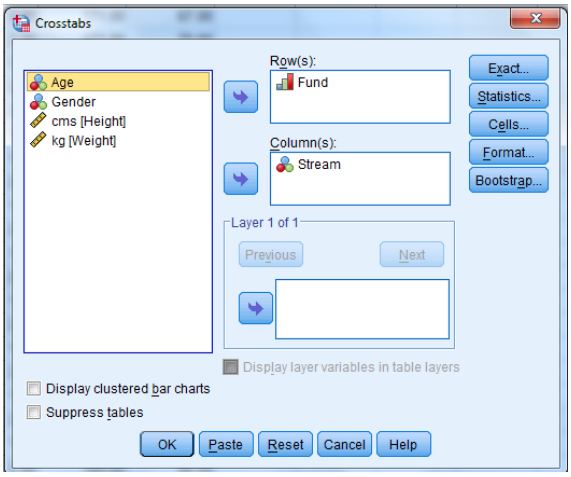
Other than goodness of fit chi-square can also be used to evaluate the relationship between two or more variables. It is termed as Chi-square as Test of Independence. This is useful when analyzing cross tabulations. This helps in identifying any significant association between the variables.

χ^2 test enables to explain whether or not two attributes are associated. Thus, χ^2 is not a measure of the degree of relationship between two attributes but it helps in judging the significance of such association between attributes. It helps in stating whether different samples come from the same Universe.

The procedure for calculating Chi-square is as follows:

Analyse → Descriptive statistics → Cross tabs

First you have to activate weigh cases as you did for finding out Chi square test for Goodness of fit (based on weigh cases). Now you select the Analyze menu and then choose descriptive statistics and then choose cross tabs. A dialogue box would appear then select a row and column variable using arrow buttons. Then click Statistics select chi-square in the box and then click Continue.



All the options shall be kept as default and then OK, the output would be generated as under;

Crosstabs

[DataSet3] C:\Users\magarwal\Downloads\funds.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Fund * Stream	51	100.0%	0	0.0%	51	100.0%

Fund * Stream Crosstabulation

Count

		Stream			
		Science	Arts	Commerce	Total
Fund	Lowest Cost Fund	2	6	0	8
	Second Lowest Cost Fund	8	6	4	18
	Third lowest cost fund	3	6	12	21
	Highest Cost Fund	4	0	0	4
Total		17	18	16	51

Chi-Square Tests

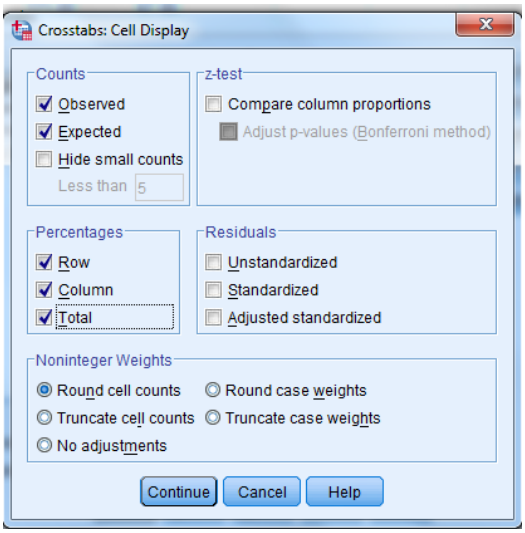
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22.417[a]	6	.001
Likelihood Ratio	24.613	6	.000
Linear-by-Linear Association	.485	1	.486
N of Valid Cases	51		

a. 6 cells (50.0%) have expected count less than 5. The minimum expected count is 1.25.

Three tables would be generated in the output, first table depicts case processing summary that explains the information about the variables. Cross tabulation of stream and funds would be given in second table. (funds.sav)

The third table conveys information about Chi-square test. The value of Pearson chi-square is 22.141 at 6 degree of freedom and N=51, associated significance value is 0.01<0.05.This shows that the there is significance of association between stream adopted and funds invested.

Further, observed and expected value can also be computed using cell display option.

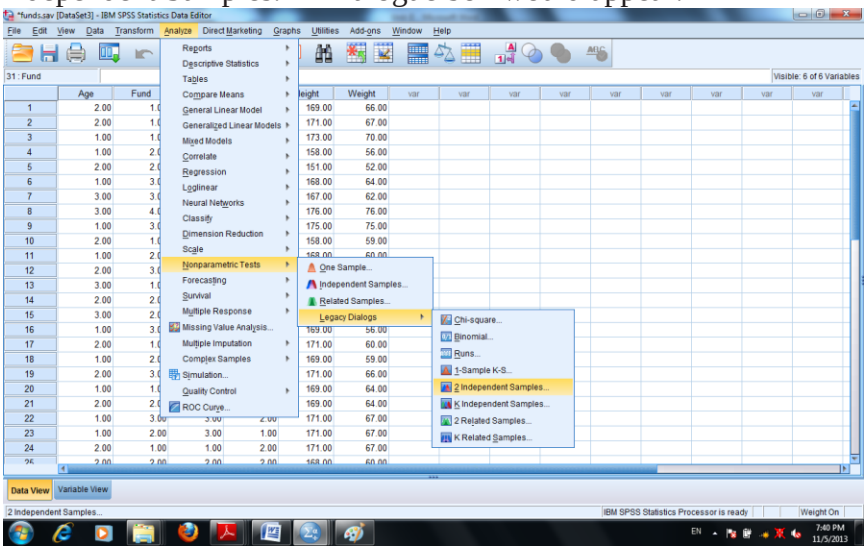


To know more about Chi- Square Test refer to the following link: http://en.wikipedia.org/wiki/Chi-squared_test

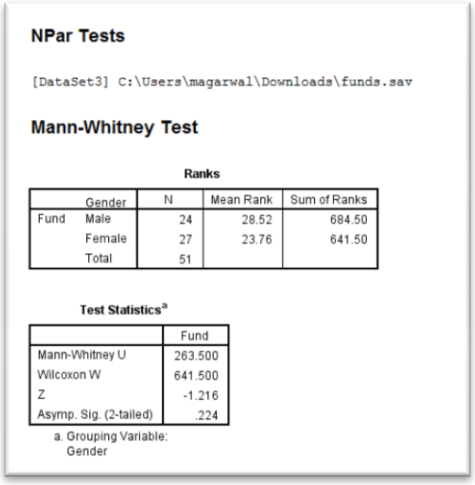
2.5.2 Mann- Whitney U Test (Wilcoxon Rank Sum W test)

Mann Whitney U test is used to check that the two populations which are being compared have identical distributions. The test is based on the joint ranking of the observations from the two samples. Hence, it checks the hypothesis that two independent samples come from populations having same distributions. This non parametric test is equivalent to parametric t-test of independent groups. If several values are tied, you assign each the average of the ranks that otherwise would have been assigned had there been no ties.
The alternative hypothesis in Mann Whitney U-Test is that the population distributions differ in location (the median).

To compute in SPSS first Click Analyze the go to Non parametric Test and then to 2 Independent Samples. A Dialogue box would appear.



Then select dependent variable in the Test variable list using Arrow button and the Gender in the grouping variable list. Then you have to define the groups in the sub dialogue box as the you have assigned in Gender and then click continue. Then select Mann Whitney U in the Test Type and then click OK
The output would be generated as under:



NPar Tests

[DataSet3] C:\Users\magarwal\Downloads\funds.sav

Mann-Whitney Test

Ranks

	Gender	N	Mean Rank	Sum of Ranks
Fund	Male	24	28.52	684.50
	Female	27	23.76	641.50
	Total	51		

Test Statistics[a]

	Fund
Mann-Whitney U	263.500
Wilcoxon W	641.500
Z	-1.216
Asymp. Sig. (2-tailed)	.224

a. Grouping Variable: Gender

The ranks table displays the details about the mean ranks and sums of ranks for each category . The sum of ranks for male respondents investing in funds is higher than the sum of ranks of female respondents.
You have to look for z-score and two tailed p-value which has been corrected for ties. The result conveys that Z= -1.218 and p is greater than 0.05 and no significant difference is noticed in male and female investment in different types of funds in terms of cost.

For exploring more about U test refer to this link:
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U

To have conceptual understanding about Non Parametric test refer to the following link:
http://en.wikipedia.org/wiki/Non-parametric_statistics

2.6 SUMMARY

This unit has introduced us how to compute measures of central tendency using SPSS. It helped in determining the single value for the entire mass of data, so that it describes the overall level of group of observations and depicts the representative value for the whole set of the data. This unit also explained that the how to compare means using t-test to infer that whether the means of the corresponding populations differ or not. We also learnt that how to compute parametric test in SPSS for knowing the degree of confidence you can accept or reject a hypothesis. Further, we also learnt to compute Non Parametric Test in SPSS when you do not assume that the outcome is approximately normally distributed.

2.7 GLOSSARY

- **Type I Error** It refers to the error when we reject a hypothesis when it may be true.
- **Type II Error:** It refers to the error when we accept a hypothesis when it may be false.
- **Hypothesis:** A hypothesis is a statement or assumption concerning a population.
- **Goodness of fit :** This is to test that how well an observed frequency distribution fits a theoretical distribution
- **Display order:** It allow us to choose the order/sequence in which the variables will be displayed.
- **Standard Error:** This is standard deviation divided by the square root. It measures stability or sampling error of the sample means.
- ***p* value** The probability of a test statistic (assuming the null hypothesis to be true). If this value is very small (e.g. 0.02763) then we reject the null hypothesis. We claim a significant effect if the *p* value is smaller than a conventional significance level (such as 0.05).

2.8 ANSWERS TO CHECK YOUR PROGRESS

Check your Progress-A

5.

a) (Null)
b) (Levene)
c) (Explore)
d) (Descriptives)

6.

a) True
b) True
c) False

2.9 REFERENCES

1. Gupta S.L. and Gupta Hitesh, SPSS 17.0 for Researchers, International Book House Pvt. Ltd., New Delhi

2. Pandya Kiran, Bulsari Smruti, Sinha Sanjay(2012) , SPSS in Simple Steps, Dreamtech press, New Delhi

3. Hooda R. P(2000), Statistics for Business and Economics, Macmillan India Ltd.

4. George Darren and Mallery Paul (2011), SPSS for windows Step by Step, Dorling Kindersley Publishing

5. Research and Communications Methodology, Self Learning Material for MBA Course CP1007, Unit 9, Vikas Publications, Noida

2.10 SUGGESTED READINGS

1. Gupta S.L. and Gupta Hitesh, SPSS 17.0 for Reserachers, International Book House Pvt. Ltd., New Delhi
2. Pandya Kiran, Bulsari Smruti, Sinha Sanjay(2012) , SPSS in Simple Steps, Dreamtech press, New Delhi
3. Hooda R. P(2000), Statistics for Business and Economics, Macmillan India Ltd.
4. George Darren and Mallery Paul (2011), SPSS for windows Step by Step, Dorling Kindersley Publishing
5. *IBM* reference guide posted on the SPSS website.
6. Landau Sabine and Everitt Brian S 'A Hand book on statistical analysis using SPSS', free downloadable

2.11 TERMINAL QUESTIONS

Q1. How do you use the function of descriptive statistics command in SPSS?

Q2. What is Chi-square test? Explain its significance in statistical analysis. How do you compute chi-square test in SPSS?

Q3. Check using SPSS whether the following samples from normal populations have the same variance.

Sample A 0 5 11 14 16 22 25 27 31

Sample B 1 6 7 25 18 3 25 26 28

Q4. Explain briefly the t-test, pinpointing its salient features. Also explain the process for its computation is SPSS.

Q5. The following is a data set from a sample of n= 20

677, 700, 835, 679, 677, 677, 720, 834, 910, 500,344,786, 896,767, 549, 493, 264,577, 989, 677

a. Compute the mean, median and mode using SPSS.
b. Compute the range, interquartile range, variance, standard deviation and coefficient of variation using SPSS.

Q6. Using academic index. sav file, compare the witting habits with the mean reading habits for the members of the faculty.

All inferential statistics are found under the **Analyze** command.