SPSS Tutorial

Which Statistical test?

Introduction

Irrespective of the statistical package that you are using, deciding on the right statistical test to use can be a daunting exercise. In this document, I will try to provide guidance to help you select the appropriate test from among the many variety of statistical tests available. In order to select the right test, the following must be considered:

1. The question you want to address.
2. The level of measurement of your data.
3. The design of your research.

Statistical Analysis (Test)

After considering the above three factors, it should also be very clear in your mind what you want to achieve.

If you are interested in the degree of relationship among variables, then the following statistical analyses or tests should be use:

Correlation

> This measures the association between two variables.

Regression

> Simple regression - This predicts one variable from the knowledge of another.
> Multiple regression - This predicts one variable from the knowledge of several others.

Crosstabs

> This procedure forms two-way and multi-way tables and provides measure of association for the two-way tables.

Loglinear Analysis

> When data are in the form of counts in the cells of a multi-way contingency table, loglinear analysis provides a means of constructing the model that gives the best approximation of the values of the cell frequencies. Suitable for nominal data.

Nonparametric Tests

> Use nonparametric test if your sample does not satisfy the assumptions underlying the use of most statistical tests. Most statistical tests assumed that your sample is drawn from a population with normal distribution and equal variance.

If you are interested in the significance of differences in level between / among variables, then the following statistical analyses or tests should be use:

- T-Test
- One-way ANOVA
- ANOVA
- Nonparametric Tests

If you are interested in the prediction of group membership then you should use **Discriminant Analysis**.

If you are interested in finding latent variables then you should use **Factor Analysis**. If your data contains many variables, you can use Factor Analysis to reduce the number of variables. Factor analysis group variables with similar characteristics together.

If you are interested in identifying a relatively homogeneous groups of cases based on some selected characteristics then you should use **Cluster Analysis**. The procedure use an algorithm that starts with each case in a separate cluster (group) and combines clusters until only one is left.

Conclusion

Although the above is not exhaustive, it covers the most common statistical problems that you are likely to encounter.

Some Common Statistical Terms

Introduction

In order to use any statistical package (such as SPSS, Minitab, SAS, etc.) successfully, there are some common statistical terms that you should know. This document introduces the most commonly used statistical terms. These terms serve as a useful conceptual interface between methodology and any statistical data analysis technique. Irrespective of the statistical package that you are using, it is important that you understand the meaning of the following terms.

Variables

Most statistical data analysis involves the investigation of some supposed relationship among variables. A variable is therefore a feature or characteristic of a person, a place, an object or a situation which the experimenter wants to investigate. A variable comprises different values or categories and there are different types of variables.

Quantitative variables

Quantitative variables are possessed in degree. Some common examples of these types of variables are height, weight and temperature.

Qualitative variables

Qualitative variables are possessed in kind. Some common examples of these types of variables are sex, blood group, and nationality.

Hypotheses

Often, most statistical data analysis wants to test some sort of hypothesis. A hypothesis is therefore a provisional supposition among variables. It may be hypothesized, for example, that tall mothers give birth to tall children. The investigator will have to collect data to test the hypothesis. The collected data can confirmed or disproved the hypothesis.

Independent and dependent variables

The independent variable has a causal effect upon another, the dependent variable. In the example hypothesized above, the height of mothers is the independent variable while the height of children is the dependent variable. This is so because children heights are suppose to depend on the heights of their mothers.

Kinds of data

There are basically three kinds of data:

Interval data

These are data taken from an independent scale with units. Examples include height, weight and temperature.

Ordinal data

These are data collected from ranking variables on a given scale. For example, you may ask respondents to rank some variable based on their perceived level of importance of the variables.

Nominal data

Merely statements of qualitative category of membership. Example include sex (male or female), race (black or white), nationality (British, American, African, etc.).

It should be appreciated that both Interval and Ordinal data relate to quantitative variables while Nominal data refers to qualitative variables.

Some cautions of using statistical packages

The availability of powerful statistical packages such as SPSS, Minitab, and SAS has made statistical data analysis very simple. It is easy and straightforward to subject a data set to all manner of statistical analysis and tests of significance. It is, however, not advisable to proceed to formal statistical analysis without first exploring your data for transcription errors and the presence of outliers (extreme values). The importance of thorough preliminary examination of your data set before formal statistical analysis can not be overemphasized.

The Golden Rule of Data Analysis

Know exactly how you are going to analyse your data before you even begin to think of how to collect it. Ignoring this advice could lead to difficulties in your project.

How to Perform and Interpret Regression Analysis

Introduction

Regression is a technique use to predict the value of a **dependent** variable using one or more **independent** variables. For example, you can predict a salesperson's total yearly sales (the dependent variable) from his age, education, and years of experience (the independent variables). There are two types of regression analysis namely **Simple** and **Multiple** regressions. Simple regression involves two variables, the dependent variable and one independent variable. Multiple regression involves many variables, one dependent variable and many independent variables.

Mathematically, the simple regression equation is as shown below:

$y^1 = b_0 + b_1x$

Mathematically, the multiple regression equation is as shown below:

$y^1 = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_nx_n$

where y^1 is the estimated value for y (the dependent variable), b_1, b_2, b_3,... are the partial regression coefficients, x, x_1, x_2, x_3,... are the independent variables and b_0 is the regression constant. These coefficients will be generated automatically after running the simple regression procedure.

Residuals

It is important to understand the concept of **Residuals**. It does not only help you to understand the analysis, they form the basis for measuring the accuracy of the estimates and the extent to which the regression model gives a good account of the collected data. The residual is simply the difference between the actual and the predicted values (i.e. y-y^1). A simple correlation analysis between y and y^1 gives an indication of the accuracy of the model.

Simple Regression

The data shown on Table 1 was collected through a questionnaire survey. Thirty sales people were approached and their ages and total sales values in the preceding year solicited. We want to use the data to illustrate the procedure of simple regression analysis.

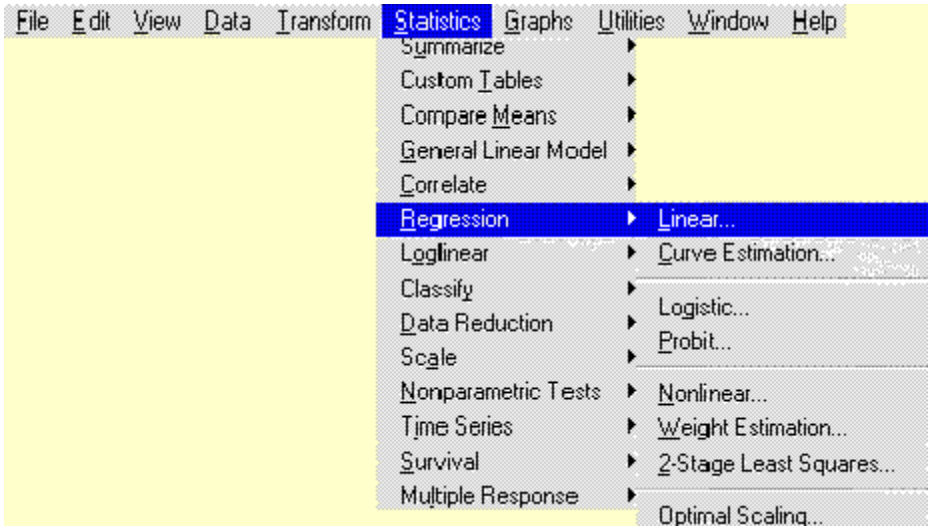
Table 1: Ages and sales total

Age	Sales in £000	Age	Sales in £000	Age	Sales in £000
29	195	42	169	38	164
35	145	36	142	32	140
26	114	21	114	29	112
23	105	28	103	27	100
29	95	21	94	25	101
20	78	27	76	24	90
24	65	23	61	20	91
41	50	20	50	19	74
25	126	35	45	19	49
27	50	33	25	18	38

Before we can conduct any statistical procedure the data has to be entered correctly into a suitable statistical package such as SPSS. Using the techniques described in *Getting Started with SPSS for Windows*, define the variables *age* and *sales*, using the labelling procedure to provide more informative names as *Age for salesperson* and *Total sales*. Type the data into columns and save under a suitable name such as *simreg*. Note that all SPSS data set files have the extension *.sav. You* can leave out the thousand when entering the sales values, but remember to multiply by a thousand when calculating the total sales of a salesperson.

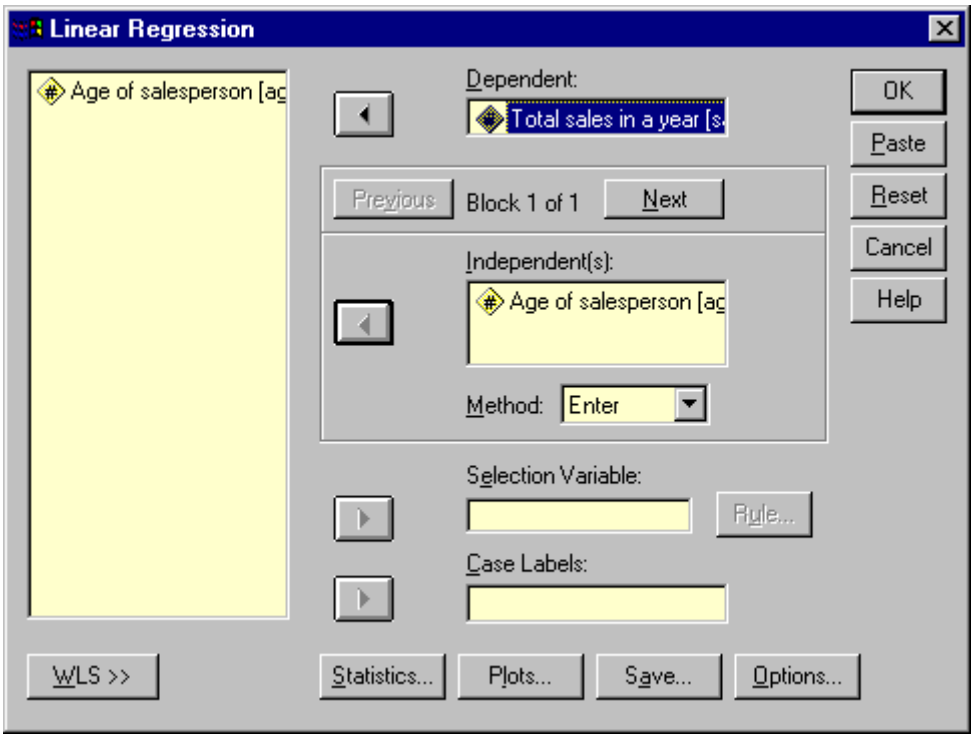
The Simple Regression Procedure

From the menus choose:
Statistics
Regression
Linear...
The Linear regression dialog box will be loaded on the screen as shown below.



Finding the Linear Regression procedure

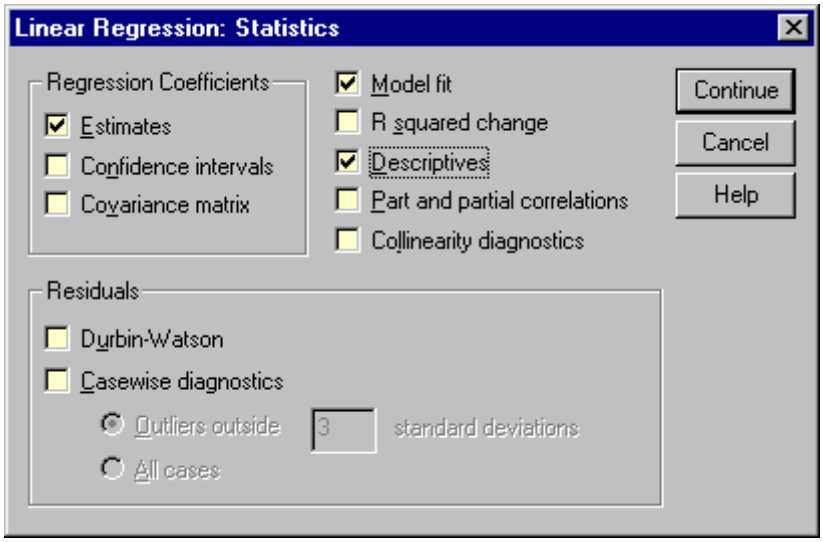
The Linear Regression dialog box



The two variables names *age* and *sales* will appear on the left-hand box. Transfer the dependent variable *sales* to the **Dependent** text box by clicking on the variable name and then on the arrow >. Transfer the independent variable *age* to the **Independent** text box.

To obtain additional descriptive statistics and residuals analysis click on the **Statistics** button. The **Linear Regression: Statistics** dialog box will be loaded on the screen as shown below. Click on the **Descriptives** check box and then on **Continue** to return to the **Linear Regression** dialog box.

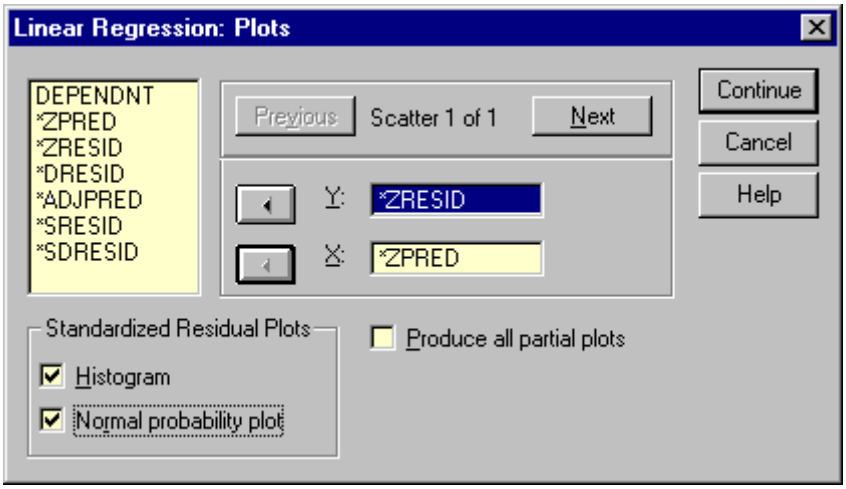
The Linear Regression: Statistics dialog box



Residuals analysis can be obtained by clicking on the **Plots** button. The **Linear Regression: Plots** dialog box will loaded on the screen as shown below. Click to check the boxes for **Histogram** and **Normal probability plots**.

We recommend you plot the residuals against the predicted values. The correct ones for this plots are **zpred* and **zresid*. Click on **zresid* and then on the arrow > to transfer it to the left of the **Y:** text box. Transfer **zpred* to the left of the **X:** text box. The completed box is as shown below. Click on **Continue** and then **OK** to run the regression. Now let's look at the output after running the procedure.

The Linear Regression: Plots dialog box



Output listing for Simple Regression

You will be surprise by the amount of output that the simple regression procedure will generate. We will attempt to explain and interpret the output for you. You should be able to interpret the output of any statistical procedure that you generate.

The descriptive statistics and correlation coefficient are shown on the tables below. The mean total sales in a year for all the 30 salespersons is £95370 (i.e. 95.37x1000). The mean age is 27.20 and N stand for the sample size. In the correlation table, the 0.393 gives the correlation between total sales value and age and it is significant at 5% level (0.016 < 0.05).

Descriptive Statistics

	Mean	Std. Deviation	N
Total sales in a year	95.37	42.02	30
Age of salesperson	27.20	6.72	30

Correlations

		Total sales in a year	Age of salesperson
Pearson Correlation	Total sales in a year	1.000	.393
	Age of salesperson	.393	1.000
Sig. (1-tailed)	Total sales in a year	.	.016
	Age of salesperson	.016	.
N	Total sales in a year	30	30
	Age of salesperson	30	30

The table below shows which variables has been entered or removed from the analysis. It is more relevant to multiple regression.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Age of salesperson	.	Enter

a. All requested variables entered.

b. Dependent Variable: Total sales in a year

The next table below gives a summary of the model. The R value stand for the correlation coefficient which is the same as r. R is use mainly to refer to multiple regression while r refers to simple regression. There is also an ANOVA table, which test if the two variables have a linear relationship. In this example, the F value of 5.109 is

highly significant indicating a linear relationship between the two variables. Only an examination of the scatter plot of the variables can ensure that the relationship is genuinely linear.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.393[a]	.154	.124	39.32

a. Predictors: (Constant), Age of salesperson

b. Dependent Variable: Total sales in a year

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7899.174	1	7899.174	5.109	.032[a]
	Residual	43293.793	28	1546.207		
	Total	51192.967	29			

a. Predictors: (Constant), Age of salesperson

b. Dependent Variable: Total sales in a year

The table below is the main aim of a regression analysis, because it contains the regression equation. The values of the **regression coefficient** and **constant** are given in column **B** of the table. Don't forget to multiply the constant and coefficient by a thousand. The equation is, therefore,

Total sales value = 28595 + 2455 x (age)

Thus a salesperson who is 24 years old would be predicted to generate yearly sales total of

28595 + 2455 x 24 = £87515

Notice from the data that the 24 old sales person actually generate £90000 worth of sales. The residual is £90000 - £87515 = £2485.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	28.595	30.401		.941	.355
	Age of salesperson	2.455	1.086	.393	2.260	.032

a. Dependent Variable: Total sales in a year

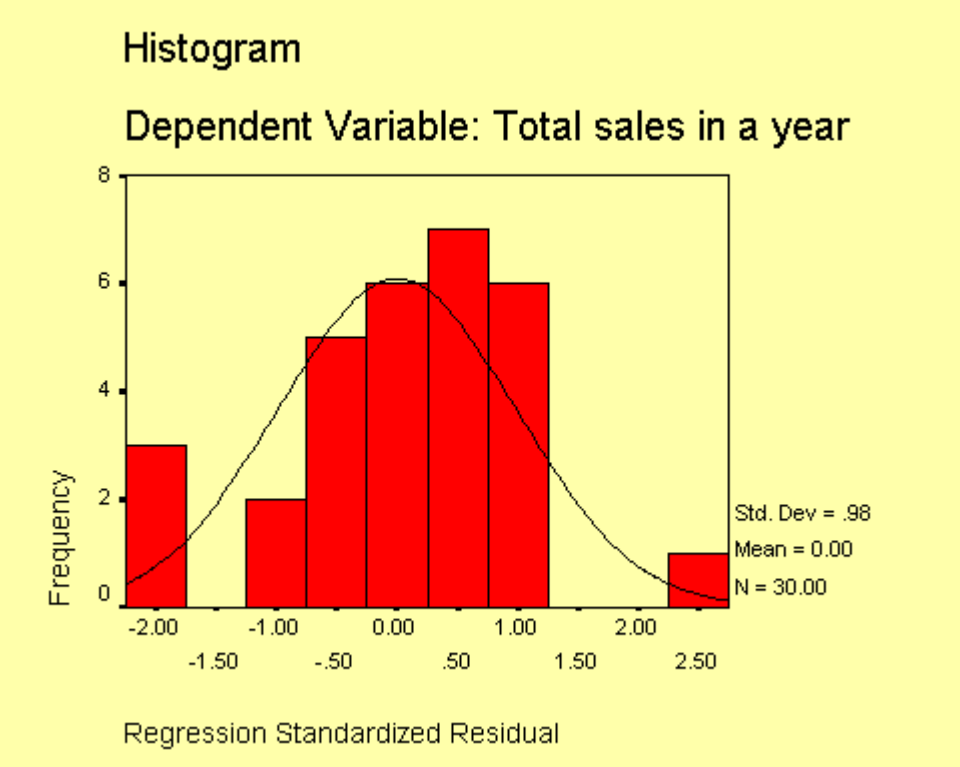
The remaining output listing relate to the residuals analysis. The table below contains the residuals statistics. It comprises the unstandardized predicted and residuals values. It also contains the standardized (std.) predicted and residuals values. Standardized means that the values have been scale such that they have a mean of 0 and a standard deviation of 1.

Residuals Statistics[a]

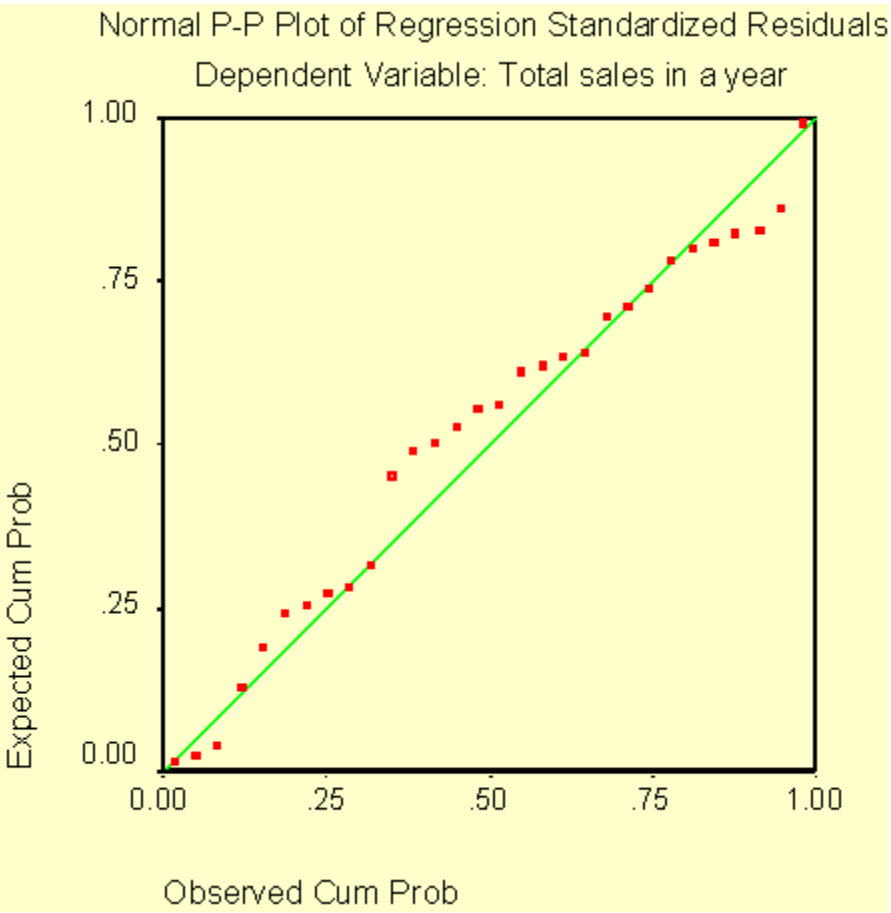
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	72.78	131.70	95.37	16.50	30
Residual	-84.60	95.21	4.26E-15	38.64	30
Std. Predicted Value	-1.368	2.201	.000	1.000	30
Std. Residual	-2.152	2.421	.000	.983	30

a. Dependent Variable: Total sales in a year

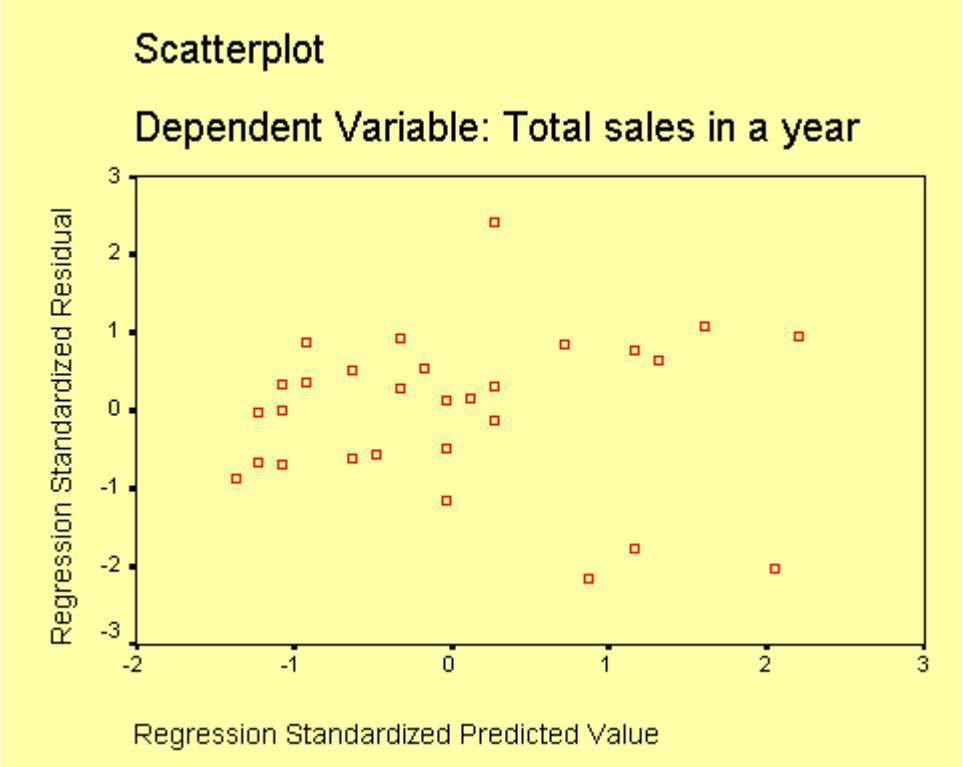
The histogram of the standardized residual is shown below. The bars shows the frequencies while the superimposed curve represent the ideal normal distribution for the residuals.



The next plot shown below is a cumulative probability plot of standardized residuals. If all the points lies on the diagonal, it means the residual are normally distributed.



The last plot of the output listing (shown below) is a scatter plot of the predicted scores against residuals. No pattern is indicated, confirming the linearity of the relationship.



So far, we have looked at how to generate and interpret a simple regression analysis. Now let us look at how to generate and interpret a multiple regression analysis.

Multiple Regression

It has already been mentioned that multiple regression involves two or more independent variables and one dependent variable. The data use for the simple regression above, will be extended and use to illustrate multiple regression. Two extra variables, the salesperson's *education* (*educ*) and *years of experience* (*years*) have been added. See Table 2 below. The salespersons education were assess by their scores obtained on a relevant academic project.

In discussing the output listing from the multiple regression procedure, there are two main questions that we need to address:

1. How does the addition of more independent variables affect the accurate prediction of total sales?
2. How can we determine the relative importance of the new variables?

Data Entry

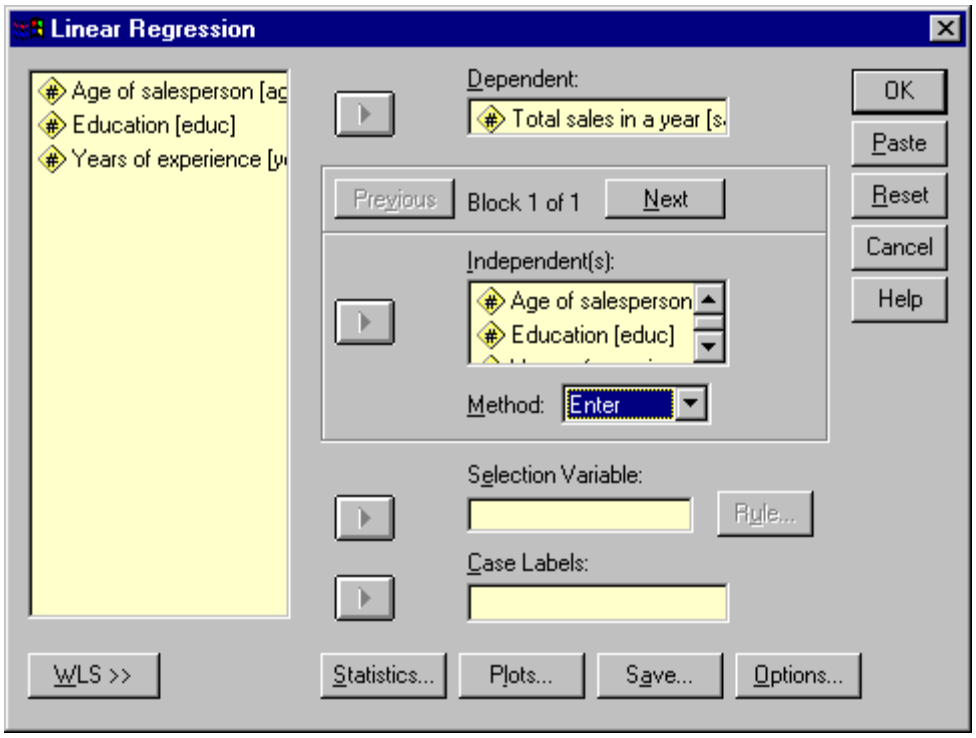
Restore the file named *simreg* into the **Data Editor** window. Define and label the two new variables. Type in the new data. Save the file under a new name such as *mulreg*.

Table 2: Extension of Table 1

sales	age	educ	years	sales	age		
195	29	65	10	76	27	75	8
145	35	84	14	61	23	65	4
114	26	76	7	50	20	70	3
105	23	60	5	45	35	68	15
95	29	84	11	25	33	78	13
78	20	79	3	164	38	64	17
65	24	77	5	140	32	69	10
50	41	70	15	112	29	60	9
126	25	74	6	100	27	68	8
50	27	72	7	101	25	61	6
169	42	60	16	90	24	65	5
142	36	68	14	91	20	82	3
114	21	50	7	74	19	60	2
103	28	69	8	49	19	54	3
94	21	72	3	38	18	75	2

After the data has been entered successfully into the Data Editor, it is time to conduct some analysis. The multiple regression procedure is the same as the simple regression procedure except that the **Linear Regression** dialog box is filled out as shown below. Transfer the variables names *age*, *educ* and *years* into the **Independent** text box, by highlighting them and clicking on the appropriate arrow (>) button. The **Dependent** variable text box must contain the variable *sales*. For **Method**, select **Enter** and click **OK** to run the procedure. The **Enter Method** enter the variables in the block in a single step. Other entry methods include **Backward**, **Forward**, and **Stepwise**.

The Linear regression dialog box fill out for multiple regression



Output listing for multiple regression

The first table from the multiple regression procedure is shown below. It shows what method was selected to enter the variables. It also shows all the variables entered.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Years of experience, Education, Age of salesperson	.	Enter

a. All requested variables entered.

b. Dependent Variable: Total sales in a year

The next table shown below gives a summary of the regression model. The multiple regression coefficient (R) is 0.447. Recalling that for the simple regression case R was 0.393, we see that the answer to the question whether adding more independent variables improves the predictive power of the regression equation is 'yes'.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.447[a]	.200	.107	39.69

a. Predictors: (Constant), Years of experience, Education, Age of salesperson

The next table from the multiple regression output listing is the ANOVA shown below.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10228.351	3	3409.450	2.164	.116[a]
	Residual	40964.616	26	1575.562		
	Total	51192.967	29			

a. Predictors: (Constant), Years of experience, Education, Age of salesperson

b. Dependent Variable: Total sales in a year

The final table is the coefficients of the variables. From column B on the table, we can write the regression equation of *total sales* upon *age*, *educ*, *years* as:

Total sales = 124906 + 312x(age) + 3343x(years) - 935x(educ)

Note the coefficients have been multiplied by a thousand.

This equation tells us nothing about the relative important of each variable. The values for the coefficients reflect the original units in which the variables were measured. Therefore, we can not conclude that *years of experience* with a larger coefficient

How to Perform and Interpret Discriminant Analysis (DA)

Introduction

Discriminant analysis is a technique use to build a predictive model of group membership based on observed characteristics of each case. For example, it is possible to group children into two main groups of *Very Clever* or *Just Clever* children based on their performance on the three core subjects English, Mathematics, and Science. Discriminant analysis generate functions from a sample of cases for which group membership is known; the functions can then be applied to new cases with measurements for the predictor variables but unknown group membership. That is, knowing a child's score on three subjects, we can use the discriminant function to determine whether the child belongs to the *Very Clever* group or the *Just Clever* group.

When there are two groups, only one discriminant function is generated. When there are more than two groups, several functions will be generated. Usually, only the first three of these functions will be useful.

Types of Discriminant Analysis

There are basically three types of DA: **direct**, **hierarchical** and **stepwise**. In **direct** DA, all the variables enter at once; in **hierarchical** DA, the order of variable entry is determine by the researcher; and in **stepwise** DA, statistical criteria alone determine the order of entry. This document concentrate on **stepwise** DA.

Preparing the SPSS data set

Data was collected on two groups of students. One group is considered to be *Very Clever* while the other is considered to be *Just Clever*. The scores of the students on the following subjects *English*, *Mathematics* and *Science* were noted. The maximum score for each subject is 100. The two groups are the dependent variables while the subjects are the independent variables. The collected data is as shown on Table 1 below. Enter the data into SPSS **Data Editor** window. The data should fit 4 columns and 30 rows. Define the independent variables. Define the coding variable, comprising two values *1 = Very Clever*, *2 = Just Clever*. After the data has been entered successfully, we are ready to perform some analysis.

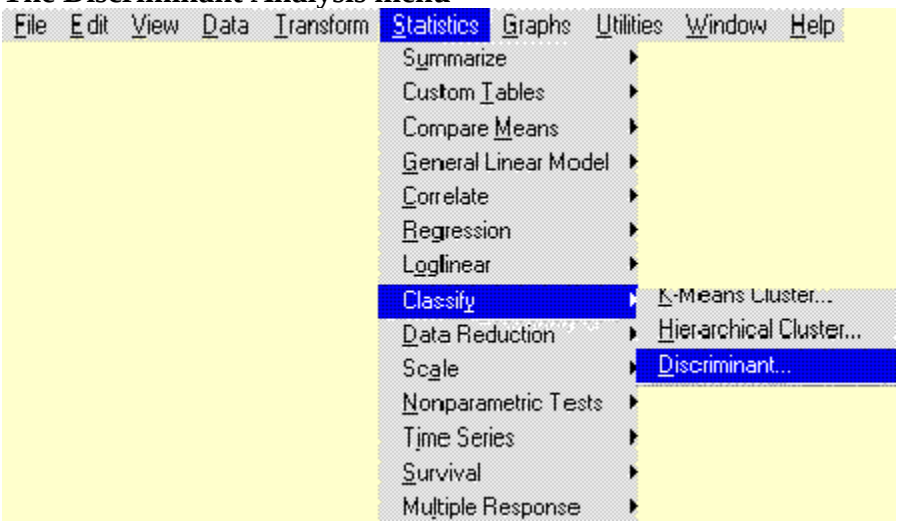
Table 1: Collected Data

English	Maths	Science	Group	English	Maths	Science	Group
44	44	28	1	40	54	40	1
61	29	25	1	29	53	19	2
19	68	77	1	28	66	71	2
48	58	45	1	27	67	17	2
38	41	30	1	45	66	79	2
25	55	77	1	55	43	51	2
39	30	50	1	68	45	58	2
33	59	44	1	52	56	51	2
30	65	49	1	74	47	33	2
17	60	21	1	70	51	29	2
42	49	30	1	49	67	74	2
47	44	43	1	80	53	40	2
13	76	52	1	50	61	13	2
63	31	54	1	48	71	71	2
54	47	8	1	65	60	39	2

Finding and running discriminant analysis

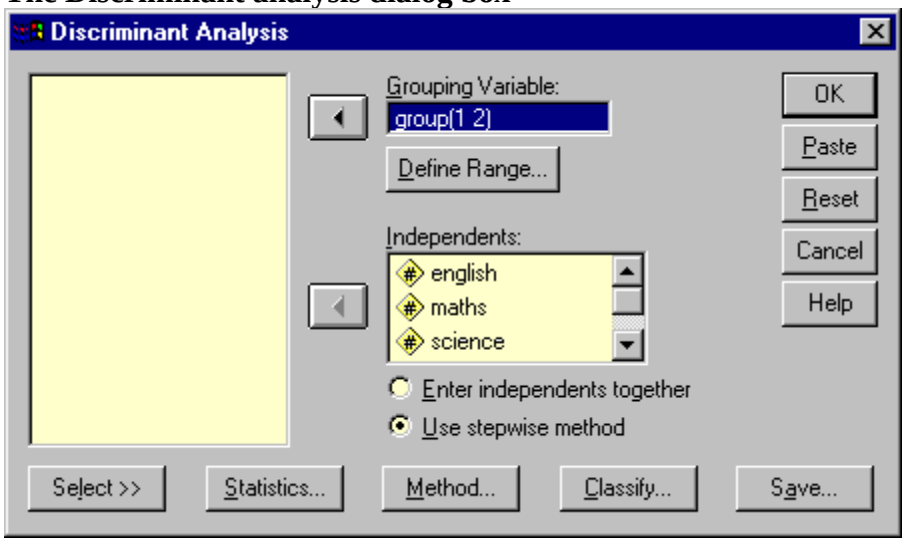
From the menus choose:
Statistics
Classify
Discriminant... (See diagram below)

The Discriminant Analysis menu



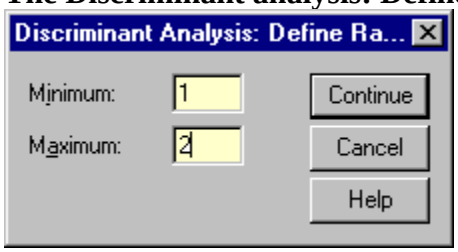
The **Discriminant analysis** dialog box will be loaded on the screen as shown below. Click to select the dependent variable *group* and click on the arrow (>) to transfer it into the **Grouping Variable** text box.

The Discriminant analysis dialog box



Now click on **Define Range** to load the **Discriminant analysis: Define Range** dialog box on to the screen (see diagram below). Type *1* into the **Minimum** text box and *2* into the **Maximum** text box. Click on **Continue** to return to the **Discriminant analysis** dialog box.

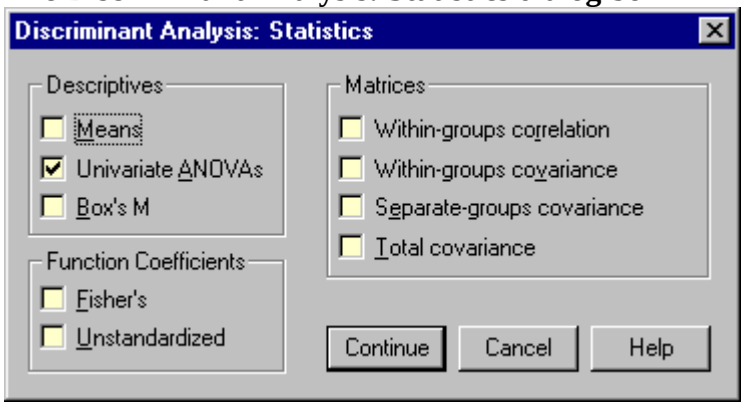
The Discriminant analysis: Define Range dialog box



Now drag the cursor over the rest of the variables (i.e. *english*, *maths* and *science*) to highlight them, and click on the arrow (>) to tranfer them into the Independent text box. Click on **Use stepwise method**. The completed dialog box is as shown above.

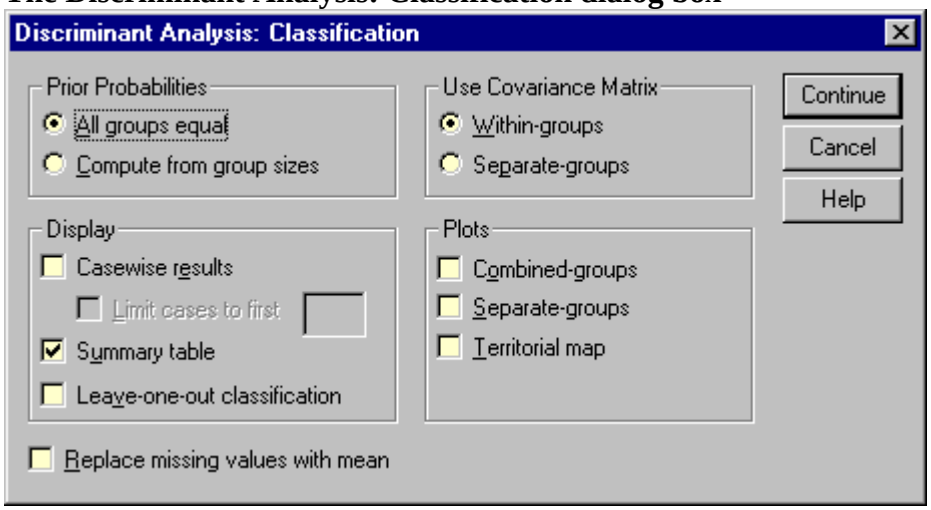
Click on **Statistics** and the **Discriminant analysis: Statistics** dialog box will be loaded on the screen (see diagram below). Within the **Descriptives** box select **Univariate ANOVAs**. Click on **Continue**.

The Discriminant Analysis: Statistics dialog box



To obtain the success/failure table, click on **Classify** and the **Discriminant analysis: Classification** dialog box will be loaded on the screen (see diagram below). Within the **Display** box, select **Summary table**. Click on **Continue** and then on **OK** to run the procedure.

The Discriminant Analysis: Classification dialog box



Now let us examine the output and try to offer some interpretation.

Output listing for discriminant analysis

The first two tables from the output listing shown below gives information about the data and the number of cases in each category of dependent variable.

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		30	100.0
Excluded	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	0	.0
Total		30	100.0

Group Statistics

GROUP		Valid N (listwise) Unweighted	Weighted
Very Clever Children	ENGLISH	16	16.000
	MATHS	16	16.000
	SCIENCE	16	16.000
Just Clever Children	ENGLISH	14	14.000
	MATHS	14	14.000
	SCIENCE	14	14.000
Total	ENGLISH	30	30.000
	MATHS	30	30.000
	SCIENCE	30	30.000

The table shown below was generated by the selected **Univariate ANOVAs**. This indicates whether there is a statistically significant difference among the dependent variable means (*group*) for each independent variable. Only *English* is statistically significant. The **Wilks' Lambda** is a statistical criteria that is used to add or remove variables from the analysis. Several other criteria are available.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
ENGLISH	.819	6.181	1	28	.019
MATHS	.917	2.549	1	28	.122
SCIENCE	.990	.287	1	28	.596

The table below shows which variables have entered the analysis. The variables are *English* and *Maths* with Wilks' Lambda of 0.819 and 0.513 respectively. Note that, at each step the variable that minimizes the overall Wilks' Lambda is entered. The table also gives more statistical information about the two variables that have entered the analysis. The F statistic and their significant is shown on the table. Note the information provided at the bottom of the table.

Step	Entered	Wilks' Lambda							
		Statistics	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	ENGLISH	.819	1	1	28.000	6.181	1	28.000	.019
2	MATHS	.513	2	1	28.000	12.833	2	27.000	.000

At each step, the variable that minimizes the overall Wilks' Lamda is entered.

- Maximum number of steps is 6
- Minimum partial F to enter is 3.84
- Maximum partial F to remove is 2.71
- F level, tolerance, or VIN insufficient for further computation

The next table shown below gives a summary of the variables in the analysis. The step at which they were entered is also shown along with other useful statistics.

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	ENGLISH	1.000	6.181	
2	ENGLISH	.532	21.271	.917
	MATHS	.532	16.142	.819

The table below shows variables not in the analysis at each step. Note, at step 0, none of the variable was yet in the analysis. At step 1, *English* was entered and at step 2 *Maths* was entered.

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	ENGLISH	1.000	1.000	6.181	.819
	MATHS	1.000	1.000	2.549	.917
	SCIENCE	1.000	1.000	.287	.990
1	MATHS	.532	.532	16.142	.513
	SCIENCE	.954	.954	.948	.791
2	SCIENCE	.903	.504	.002	.513

The next two tables shown below gives the percentage of the variance accounted for by the one discriminant function generated. The significant of the function is also shown. Becuase there are two groups only one discriminant function was generated.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.951[a]	100.0	100.0	.698

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.513	18.040	2	.000

The standardised conical discriminant function coefficients for the two variables in the analysis are shown on the table below.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
ENGLISH	1.304
MATHS	1.201

The pooled within groups correlations between the discriminanting variables and the function is shown on the table below. It is clear from this output that the association between the variable *Science* and the discriminant function is very small.

Structure Matrix

	Function
	1
ENGLISH	.482
MATHS	.309
SCIENCE[a]	.093

a. This variable not used in the analysis.

The next table below shows the group centroids for each group. The group centroids are quite different for the two groups.

Functions at Group Centroids

	Function
GROUP	1
Very Clever Children	-.881
Just Clever Children	1.007

Unstandardized canonical discriminant functions evaluated at group means

The last three tables from the output listing was generated from the optional selection of **Summary table** from the **Classify options** in the **Discriminant Analysis** dialog box. The last of the table provide an indication of the success rate for prediction of membership of the grouping variable's categories using the discriminant function developed from the analysis.

The last table shows that the *Very Clever* chidren are the more accurately classified with 93.8% of the cases correct. For the *Just Clever* children 71.4% of cases were correctly classified. Overall, 83.3% of the original cases was correctly classified.

Classification Processing Summary

Processed		30
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		30

Prior Probabilities for Groups

GROUP	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Very Clever Children	.500	16	16.000
Just Clever Children	.500	14	14.000
Total	1.000	30	30.000

Classification Results[a]

		GROUP	Predicted Group Membership		Total
			Very Clever Children	Just Clever Children	
Original	Count	Very Clever Children	15	1	16
		Just Clever Children	4	10	14
	%	Very Clever Children	93.8	6.3	100.0
		Just Clever Children	28.6	71.4	100.0

a. 83.3% of original grouped cases correctly classified.

How to Perform and Interpret T Tests

Introduction

There are basically three types of t tests. We are going to look at each one in turn, that is, how to perform and interpret the output. The three types are:

Independent-samples t test (two-sample t test):

This is used to compare the means of one variable for two groups of cases. As an example, a practical application would be to find out the effect of a new drug on blood pressure. Patients with high blood pressure would be randomly assigned into two groups, a *placebo* group and a *treatment* group. The *placebo* group would receive conventional treatment while the *treatment* group would receive a new drug that is expected to lower blood pressure. After treatment for a couple of months, the two-sample t test is used to compare the average blood pressure of the two groups. Note that each patient is measured once and belongs to one group.

Paired-samples t test (dependent t test):

This is used to compare the means of two variables for a single group. The procedure computes the differences between values of the two variables for each case and tests whether the average differs from zero. For example, you may be interested to evaluate the effectiveness of a mnemonic method on memory recall. Subjects are given a passage from a book to read, a few days later, they are asked to reproduce the passage and the number of words noted. Subjects are then sent to a mnemonic training session. They are then asked to read and reproduce the passage again and the number of words noted. Thus each subject has two measures, often called, *before* and *after* measures.

An alternative design for which this test is used is a matched-pairs or case-control study. To illustrate an example in this situation, consider treatment patients. In a blood pressure study, patients and control might be matched by age, that is, a 64-year-old patient with a 64-year-old control group member. Each record in the data file will contain response from the patient and also for his matched control subject.

One-sample t test:

This is used to compare the mean of one variable with a known or hypothesized value. In other words, the One-sample t test procedure tests whether the mean of a single variable differs from a specified constant. For instance, you might be interested to test whether the average IQ of some 50 students differs from 125.

Assumptions underlying the use of t test

Before we look at the details of how to perform and interpret a t test, it is good idea for you to understand the assumption underlying the use of t test. It is assumed that the data has been derived from a population with normal distribution and equal variance. With moderate violation of the assumption, you can still proceed to use the t test provided the following is adhere to:

1. The samples are not too small.
2. The samples do not contain outliers.
3. The samples are of equal or nearly equal size.

However, if the sample seriously violates the assumption then **Nonparametric Tests** should be used. Nonparametric tests do not carry specific assumptions about population distributions and variance.

The p-value

In the interpretation of the ***t statistics***, we will be looking at its p-value. Generally, there are three situations where you will need to interpret the p-value:

1. If the p-value is greater than 0.05, the **null hypothesis** is accepted and the result is **not significant**.
2. If the p-value is less than 0.05 but greater than 0.01, the **null hypothesis** is rejected and the result is **significant beyond the 5 per cent level**.
3. If the p-value is smaller than 0.01, the **null hypothesis** is rejected and the result is **significant beyond the 1 percent level**.

After understanding the background of a t test, let us consider some real examples.

The Independent-Samples T Test

To illustrate this procedure, consider the data shown on Table 1 below. Twenty patients suffering from high blood pressure were randomly selected and assigned to two separate groups. One group called the *placebo* group were given conventional treatment and the other group called *newdrug* were given a new drug. The aim was to investigate whether the new drug will reduced blood pressure.

Table 1: Patients with high blood pressure

Group	Blood pressure
1	90
1	95
1	67
1	120
1	89
1	92
1	100
1	82
1	79
1	85
2	71
2	79
2	69
2	98
2	91

2	85
2	89
2	75
2	78
2	80

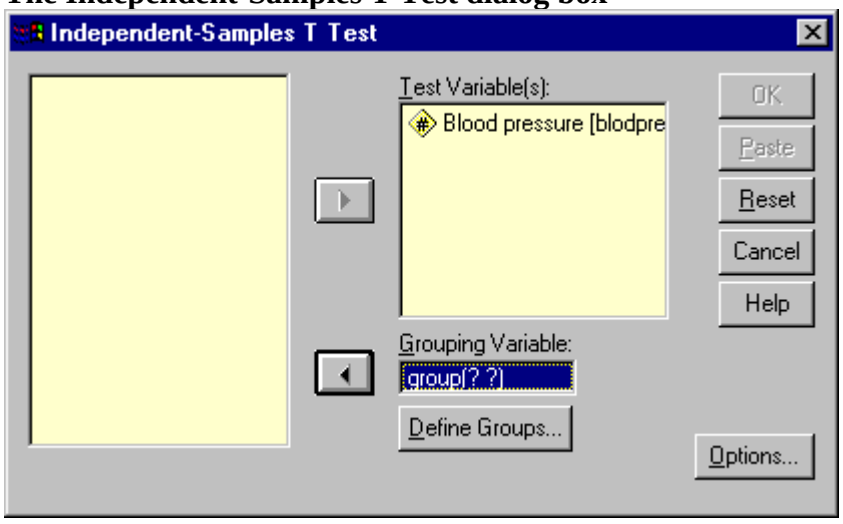
Note: 1 is a code for the *placebo* group and 2 is a code for the *newdrug* group

Using the techniques described in *Getting Started for SPSS* define the grouping (independent) variable as *group* and the dependent variable as *blodpres*. Fullers names (e.g. *Treatment Groups* and *Blood pressure*) and value labels (e.g. *placebo* and *newdrug*) can be assigned by using the **Define Labels** procedure. Type in the data and save in a suitable file. After the data has correctly been entered into the **Data Editor** of SPSS, we are now ready to perform some analysis. To carry out the t test procedure follow these instructions:

From the menus choose:
Statistics
Compare means
Independent-samples T Test.

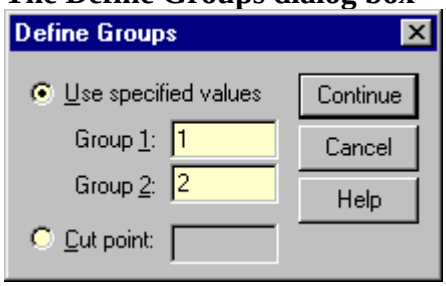
The **Independent-Samples T Test** dialog box will be loaded on the screen as shown below. Highlight the variable *blodpres* and click on the top arrow (>) to transfer it to the **Test Variable(s)** text box. Highlight the grouping variable *group* and click on the bottom arrow (>) to transfer it to the **Grouping Variable** text box.

The Independent-Samples T Test dialog box



At this point, the **Grouping Variable** text box will appear with ??. Now click on the **Define Groups** button and the **Define Groups** dialog box will be loaded on the screen. Type the value *1* into the **Group 1** text box and the value *2* into the **Group 2** text box. The **Define Groups** dialog box should be completed as shown below. Click on **Continue** and then on **OK** to run the t test.

The Define Groups dialog box



Now let us look at the output listing.

T test output listing for Independent Samples

The output listing starts with a table of statistics for the two groups followed by another table showing the mean difference between the two groups and some other statistics. One of the assumption underlying the use of t test is the equality of variance, the **Levene test** for homogeneity (equality) of variance is included in the table. Provided the F value is **not significant** ($p > 0.05$), the variances can be assumed to be homogeneous and the **Equal Variance** line values for the t test be used. If $p < 0.05$, then the equality of variance assumption has been violated and the t test based on the separate variance estimates (**Unequal Variances**) should be used.

In this case, the **Levene test** is not significant, so the t value calculated with the pooled variance estimate (Equal Variance) is appropriate. With a 2-Tail Sig (i.e. p-value) of 0.130 (i.e. 13%), the difference between means is not significant.

Group Statistics

	Treatment Group	N	Mean	Std. Deviation	Std. Error Mean
Blood pressure	placebo group	10	89.90	14.02	4.43
	newdrug group	10	81.50	9.19	2.91

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. E Differe
Blood pressure	Equal variances assumed	.368	.552	1.584	18	.130	8.40	5
	Equal variances not assumed			1.584	15.531	.133	8.40	5

The Paired-Samples T Test

As mentioned above, paired-samples t test is used to compare the means of two variables for a single group. To illustrate this procedure, consider the data shown on Table 2 below. Subjects were given a passage to read and ask to reproduce it on a later date. Subjects were then sent to a mnemonic training session and after the training, subjects were given the same passage and asked to reproduce it on a later date. The table show the number of words recalled by subjects *before* and *after* the mnemonic training session.

Table 2: Number of words recalled

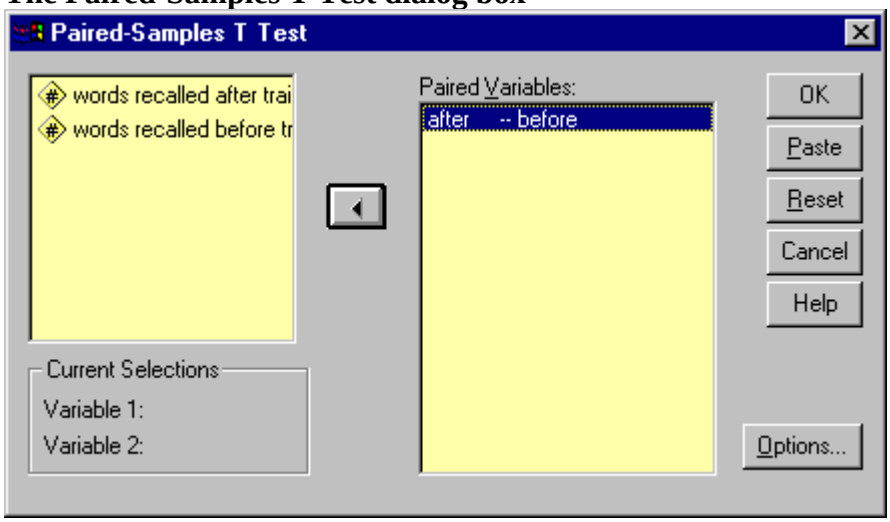
Before **mnemonic training**	***After*** **mnemonic training**
204	223
393	412
391	402
265	285
326	353
220	243
423	443
342	340
480	582
464	490

Define the variables names as *before* and *after* and use the **Define Labels** procedure to provide fuller names such as *Words recalled before training* and *Words recalled after training*. Type the data into two columns and save under a suitable name. After the data has correctly been entered into the **Data Editor** of SPSS, we are now ready to perform some analysis. To carry out the t test procedure follow these instructions:

From the menus choose:
Statistics
Compare means
Paired-Samples T Test.

The **Paired-Samples T Test** dialog box will be loaded on the screen as shown below. Highlight the two variables names from the left hand box and transfer them into the **Paired Variables** text box. Now click on **OK** to run the procedure. Let us look at the output listing.

The Paired-Samples T Test dialog box



T test output listing for Paired-Samples

The output listing starts with a table of statistics for the two variables (see below).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	words recalled after training	377.30	10	112.09	35.45
	words recalled before training	350.80	10	97.09	30.70

The next table from the output listing gives the correlation between the two variables which is 0.975.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	words recalled after training & words recalled before training	10	.975	.000

The last table from the output listing contains the t-value (3.013) and the 2-tail p-value (0.015). The 95% confidence interval of 6.60 to 46.40 is also shown on the table. Since the p-value of 0.015 is less than 0.05 the difference between the means is significant. In other words, sending subjects to mnemonic training session improves their memory recall.

One-Sample T Test

Paired Samples Test

		Paired Differences					t	df	S (2-ta
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	words recalled after training - words recalled before training	26.50	27.81	8.80	6.60	46.40	3.013	9	

The One-Sample t test procedure tests whether the mean of a single variable differs from a specified constant. For example, you might want to test whether the IQs of 10 students in your class differs from 125. Data for the ten students is shown on Table 3 below.

Table 3: IQs of students

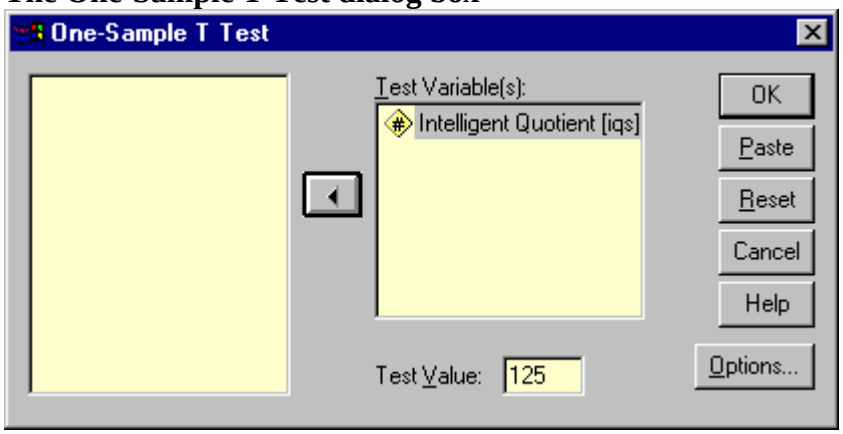
IQs
128
134
134
131
134
126
140
133
127
131

Define the name of the variables as *iqs* and use the **Define Labels** procedure to provide a fuller name such as *Intelligence Quotient*. Type the data into a single column and save under a suitable name. After the data has correctly been entered into the **Data Editor** of SPSS, we are now ready to perform some analysis. To carry out the t test procedure follow these instructions:

From the menus choose:
Statistics
Compare means
One-Sample T Test

The **One-Sample T Test** dialog box will be loaded on the screen as shown below. Highlight the variable *iqs* and transfer it to the **Test Variable(s)** text box. Enter 125 into the **Test Value** text box. Now click on **OK** to run the procedure. Let us look at the output listing.

The One-Sample T Test dialog box



T test output listing for One-Sample

The output listing starts with a table of statistics for the variable. These include the mean, standard deviation and standard error (see table below).

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Intelligent Quotient	10	131.80	4.16	1.31

The next table contains information about the t test. The t-value is 5.172 and the p-value is 0.001. Since the p-value is less than 0.05, the difference between the mean (131.80) and the test value (125) is statistically significant. The 95% confidence interval of the difference is 3.83 to 9.77.

One-Sample Test

	Test Value = 125					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Intelligent Quotient	5.172	9	.001	6.80	3.83	9.77

Correlational Analysis

Introduction

In most statistical packages, correlational analysis is a technique use to measure the association between two variables. A **correlation coefficient** (**r**) is a statistic used for measuring the strength of a supposed linear association between two variables. The most common correlation coefficient is the **Pearson** correlation coefficient. Other types of correlation coefficients are available. Generally, the correlation coefficient varies from -1 to +1.

Learning Outcomes

After studying this document you should be able to do the following:

1. Conduct and interpret a correlation analysis using interval data.
2. Conduct and interpret a correlation analysis using ordinal data.
3. Conduct and interpret a correlation analysis using categorical data.

Scatterplot

The existence of a statistical association between two variables is most apparent in the appearance of a diagram called a scatterplot. A scatterplot is simply a cloud of points of the two variables under invistigation. The diagrams below shows the scatterplots of sets of data with varying degrees of linear association.

Scatterplots of sets of data with varying degrees of linear association

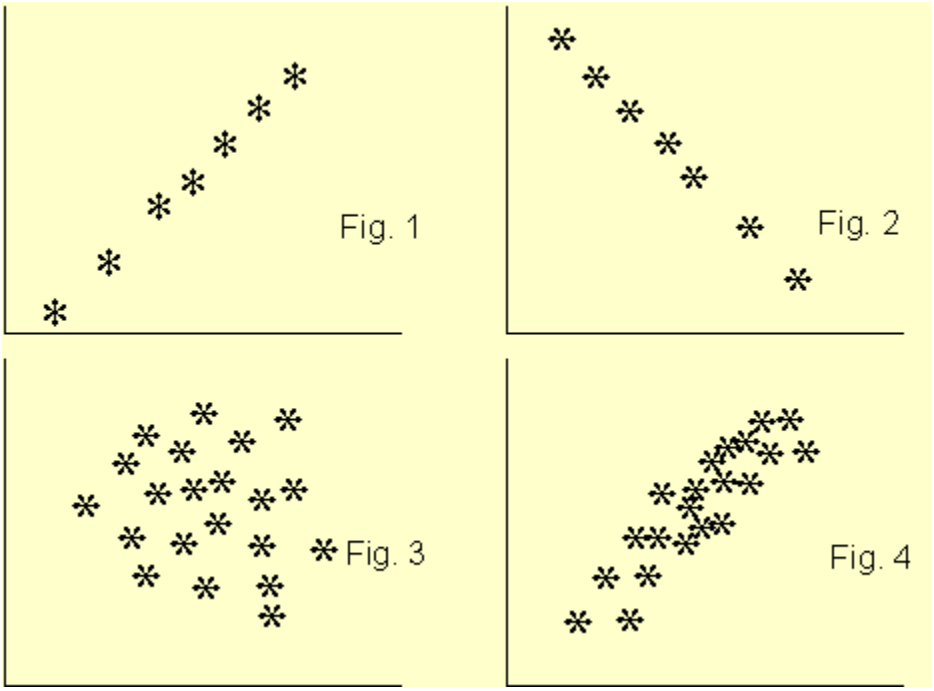


Figure 1 clearly shows a linear association between the two variables and the coefficent of correlation r is +1. For Figure 2, r is -1. In Figure 3, the two variables do not show any degree of linear association at all, r = 0. The scatterplot of Figure 4 shows some degree of association between the two variables and r is about +0.65. From the

scatterplot, we can see very clearly whether there is a linear association between the two variables and guess accurately the value of the correlation coefficient. After looking a the scatterplot, we then go ahead and confirm the association by conducting a correlation analysis. However, from the correlation coefficient alone, we can not say much about the linear association between the two variables.

How to conduct and interpret a correlation analysis using interval data

Suppose you are interested in finding whether there is an association between people monthly expenditure and income. To investigate this, you collected data from ten subjects as shown on Table 1 below.

Table 1: Set of paired data

Income / month (£)	Expenditure / month (£)
4000	4000
4000	5000
5000	6000
2000	2000
9000	6000
4000	2000
7000	5000
8000	6000
9000	9000
5000	3000

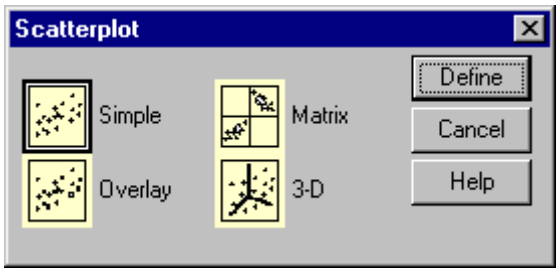
Preparing the Data set

Start SPSS, define the variables names *income* and *expend* and use the **Define Labels** procedure to provide fuller names such as *Income / month* and *Expenditure / month*. Type in the data and save under a suitable name. To conduct the correlation analysis, it is advisable to produce a scatterplot of the two variables first.

To produce the scatterplot choose:
Graphs
Scatter

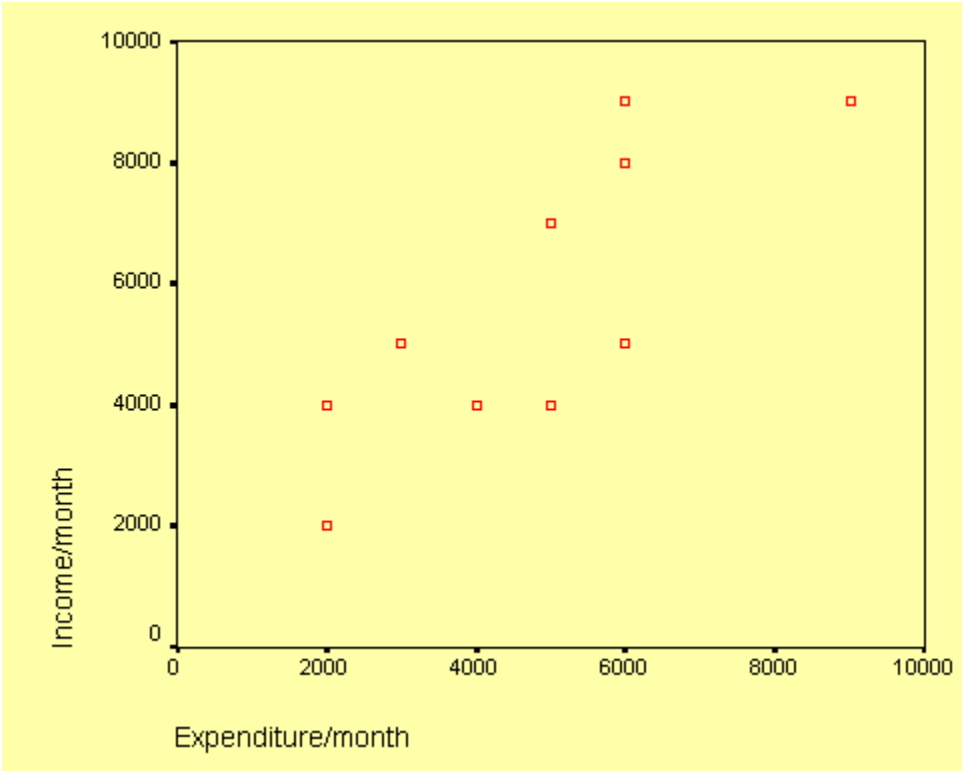
The **Scatterplot** selection box will be loaded to the screen as shown below, with **Simple** scatterplot selected by default. Click on **Define** to specify the axes of the plot. Enter the variables names *income* and *expend* into the **y-axis** and the **x-axis** box, respectively. Click on **OK**.

The Scatterplot selection box



The scatterplot is shown below and it seems to indicate a linear association between the two variables.

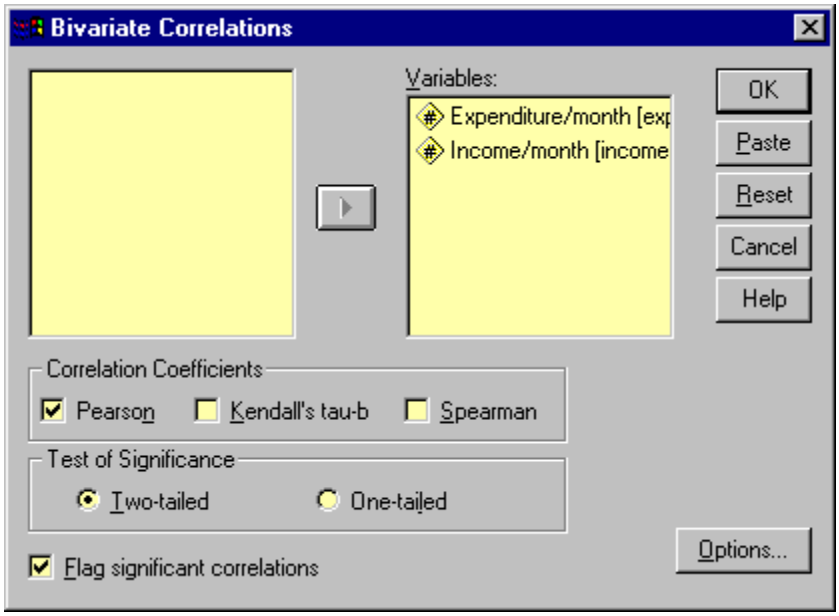
Scatterplot Income/month against Expenditure/month



To produce the correlation analysis choose:
Statistics
Correlate
Bivariate

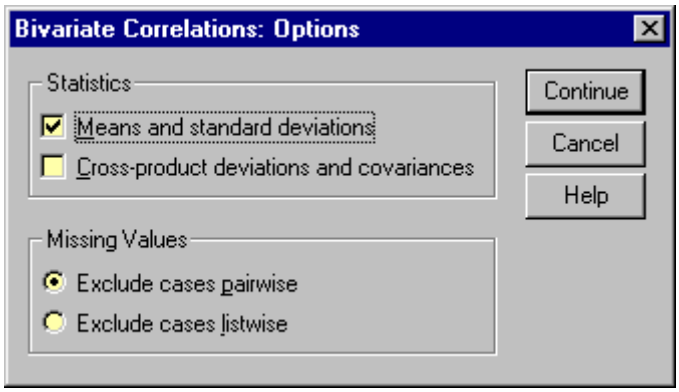
This will open the **Bivariate Correlation** dialog box as shown below. Transfer the two variables to the Variables text box.

The Bivariate Correlation dialog box



Click on **Options** and the **Bivariate Correlation: Options** dialog box will be loaded on the screen as shown below. Click on the **Means and Standard Deviations** check box. Click on Continue and then **OK** to run the procedure.

The Bivariate Correlation: Options dialog box



Let us now look at the output listing.

Output Listing of Pearson Correlation Analysis

The output listing starts with the means and standard deviation of the two variables as requested under the **Options** dialog box. This result is shown on the table below.

Descriptive Statistics

	Mean	Std. Deviation	N
Expenditure/month	4800.00	2149.94	10
Income/month	5700.00	2406.01	10

The next table from the output listing shown below gives the actual value of the correlation coefficient along with its p-value. The correlation coefficient is 0.803 and the p-value is 0.005. From these values, it can be concluded that the correlation coefficient is significant beyond the 1 per cent level. In order words, people with high monthly income are also likely to have a high monthly expenditure budget.

Correlations

		Expenditure/month	Income/month
Expenditure/month	Pearson Correlation	1.000	.803**
	Sig. (2-tailed)	.	.005
	N	10	10
Income/month	Pearson Correlation	.803**	1.000
	Sig. (2-tailed)	.005	.
	N	10	10

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

How to conduct and interpret a correlation analysis using ordinal data

The **Pearson** correlation analysis as demonstrated above is only suitable for interval data. With other types of data such as ordinal or nominal data other methods of measuring association between variables must be used. Ordinal data are either ranks or ordered category membership and nominal data are records of qualitative category membership. A brief introduction of types of data can be found under *Some Common Statistical Terms* which is located under *Documentation* found on the Content page on the left.

Suppose you are a psychology student. Twelve books dealing with the same psychological topic have just been published by 12 different authors. You and a friend were asked to rank the books in order depending on how well the authors covered the topic. The ranking is show on Table 2 below. Is there any association of the ranking by the two students?

Table 2: Ranks assigned by two students to each of twelve books

Books	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Student 1	1	2	3	4	5	6	7	8	9	10	11	12
Student 2	1	3	2	4	6	5	8	7	10	9	12	11

Preparing the Data set

In the **Data Editor** grid of SPSS, define the two variables, *student1* and *student2*. Enter the data from Table 2 into the respective column.

To obtain the correlation coefficient follow these instructions:

Choose
Statistics
Correlate
Bivariate

This will open the **Bivariate Correlation** dialog box. See daigram above. Select the **Kendall's tau-b** and the **Spearman** check boxes. Notice that by default the **Pearson** box is selected. Click on **OK** to run the procedure.

Output Listing of Spearman and Kendall rank correlation

The two tables from the output listing are shown below. Notice that both the **Pearson** and the **Spearman** correlationn coefficient are exactly the same 0.965 and significant beyond the 1 per cent level. The **Kendall** correlation coefficient is 0.848 and also significant beyond the 1 per cent level. The different between the **Spearman** and the **Kendall** coeffiecients is due to the fact that they have different theoretical background. You should not worry about the difference.

The association between the two ranks is significant indicating that the two students ranked the twelve books in a similar way. In fact, close examination of the data on Table 2 shows that, at most, the ranks assigned by the students differ by a single rank.

Correlations

		STUDENT1	STUDENT2
STUDENT1	Pearson Correlation	1.000	.965**
	Sig. (2-tailed)	.	.000
	N	12	12
STUDENT2	Pearson Correlation	.965**	1.000
	Sig. (2-tailed)	.000	.
	N	12	12

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

Correlations

			STUDENT1	STUDENT2
Kendall's tau_b	STUDENT1	Correlation Coefficient	1.000	.848**
		Sig. (2-tailed)	.	.000
		N	12	12
	STUDENT2	Correlation Coefficient	.848**	1.000
		Sig. (2-tailed)	.000	.
		N	12	12
Spearman's rho	STUDENT1	Correlation Coefficient	1.000	.965**
		Sig. (2-tailed)	.	.000
		N	12	12
	STUDENT2	Correlation Coefficient	.965**	1.000
		Sig. (2-tailed)	.000	.
		N	12	12

**. Correlation is significant at the .01 level (2-tailed).

How to conduct and interpret a correlation analysis using categorical data

Soppose that 150 students (75 boys and 75 girls) starting at a university are asked to show their preference of study by indicating whether they prefer art or science degrees. We can hypothesised that boys should prefer science degree and girls art. There are two nominal variables here *group* (boys or girls); and *student's choice* (art or science). The null hypothesis is that there is no association between the two variables. The table below shows the student's choices.

Table 3: A contingency table

		STUDENT'S CHOICE		Total
		Art degree	Science degree	
GROUP	Boys	25	50	75
	Girls	55	20	75
	Total	80	70	150

Close examination of Table 3 indicate that there is an association between the two variables. The majority of the boys chose science degree while the majority of the girls chose art degree.

Preparing the Data set

You need to define three variables here, two coding variables for *group* and *choice*. The third variable is simply the frequency *count* for the choice of degree. Note that no individual can fall into more than one combination of categories. Define the three variables *group*, *choice* and *count*. In the *group* variable, use the code numbers *1* and *2* to represent boys and girls respectively. Similarly, in the *choice* variable use the values *1* and *2* to represent art and science degrees respectively. Type the data into the three columns as shown below.

Showing coding of data in Data Editor

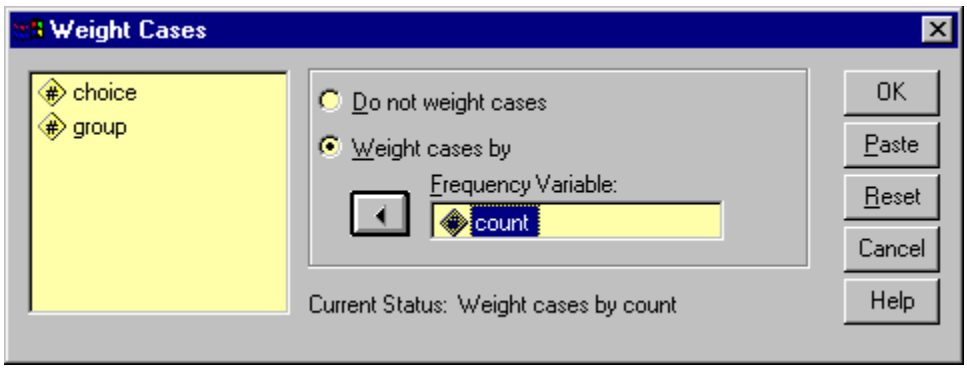
	group	choice	count
1	1	1	25
2	1	2	50
3	2	1	55
4	2	2	20

Before we can proceed, we need to tell SPSS that the data in the count column represent cell frequencies of a variable and not actual values. To do this, follow this instructions.

Choose
Data
Weight Cases

The **Weight Cases** dialog box will be loaded on the screen as shown below. Select the item **Weight cases by**. Click on the variable *count* and on the arrow (>) to transfer it into the **Frequency Variable** text box. Click on **OK**.

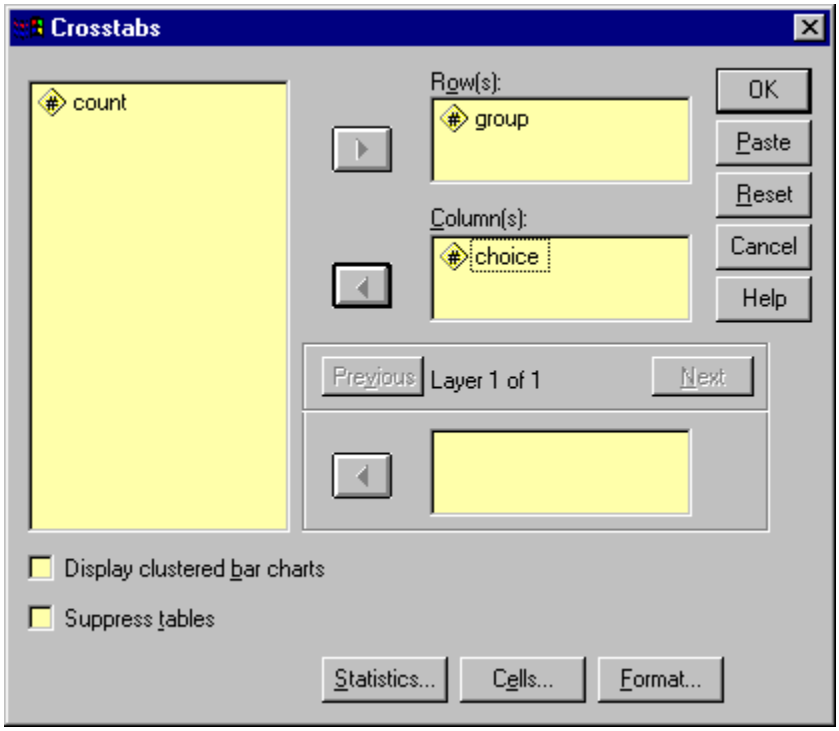
The Weight Cases dialog box



To analyse the contingency table data, choose
Statistics
Summarize
Crosstabs

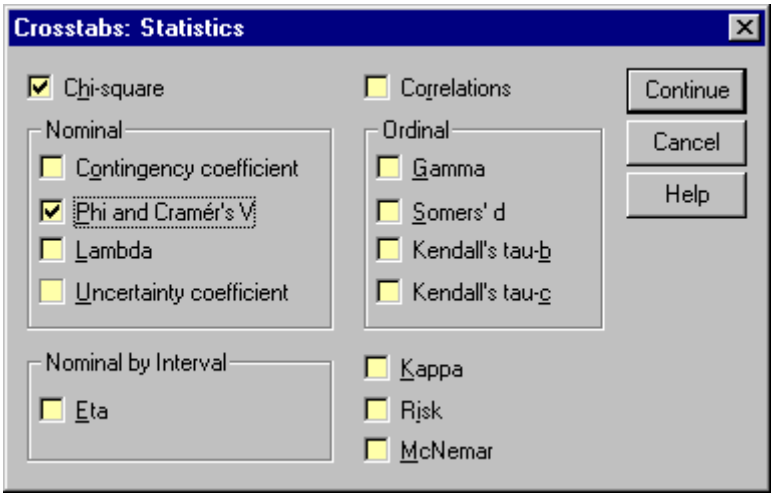
The **Crosstabs** dialog box will be loaded on the screen as shown below. Click on the variable *group* and on the top arrow (>) to transfer *group* into the **Row(s)** text box. Click on variable *choice* and then on the middle arrow (>) to transfer *choice* into the **Column(s)** text box.

The completed Crosstabs dialog box



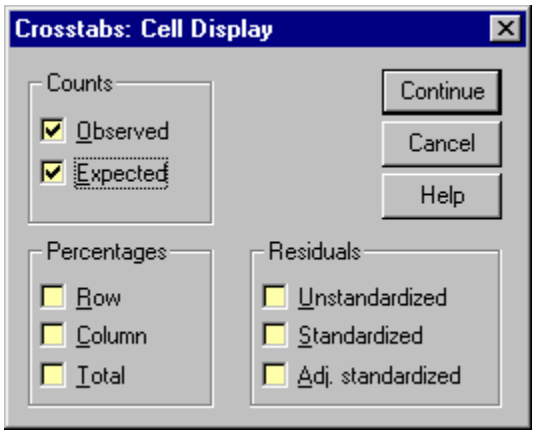
Click on **Statistics** to open the **Crosstabs: Statistics** dialog box. See diagram below. Select the **Chi-square** and **Phi and Cramer's V** check boxes. Click on **Continue** to return to the **Crosstabs** dialog box.

The completed Crosstabs: Statistics dialog box



Click on **Cells** at the foot of the **Crosstabs** dialog box to open the **Crosstabs: Cell Display** dialog box. See diagram below. Select the **Expected** check box. Click on **Continue** and then **OK** to run the procedure. We have computed the cell frequencies to ensure that the prescribed minimum requirements for the valid use of chi-square have been fulfilled, i.e. a cell frequency should not be less than 5.

The Crosstabs: Cell Display dialog box



Output Listing for Crosstabulation

The first table from the output listing shown below gives a summary of variables and the number of cases.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GROUP * CHOICE	150	100.0%	0	.0%	150	100.0%

The table below shows the observed and expected frequencies as requested in the **Crosstabs: Cell Display** dialog box. Notice that none of the expected frequencies is less than 5.

GROUP * CHOICE Crosstabulation

			CHOICE		
			art degree	science degree	Total
GROUP	boys	Count	25	50	75
		Expected Count	40.0	35.0	75.0
	girls	Count	55	20	75
		Expected Count	40.0	35.0	75.0
Total		Count	80	70	150
		Expected Count	80.0	70.0	150.0

The table below gives the Chi-square statistics for the contingency table. It can be concluded that there is a significant association between the variables *group* and *choice*, as shown by the p-value (less than 0.01).

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	24.107[b]	1	.000		
Continuity Correction[a]	22.527	1	.000		
Likelihood Ratio	24.813	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	23.946	1	.000		
N of Valid Cases	150				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 35.00.

The **Phi** and **Cramer's V** coefficients (shown on the table below) of 0.401 gives the strength of the association between the two variables.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-.401	.000
	Cramer's V	.401	.000
N of Valid Cases		150	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Conclusion

You should now be able to perform and interpret the results of correlational analysis using SPSS for interval, ordinal and categorical data.

How to Perform and Interpret Factor Analysis using SPSS

Introduction

Factor analysis is used to find latent variables or factors among observed variables. In other words, if your data contains many variables, you can use factor analysis to reduce the number of variables. Factor analysis groups

variables with similar characteristics together. With factor analysis you can produce a small number of factors from a large number of variables which is capable of explaining the observed variance in the larger number of variables. The reduced factors can also be used for further analysis.

There are three stages in factor analysis:

1. First, a correlation matrix is generated for all the variables. A correlation matrix is a rectangular array of the correlation coefficients of the variables with each other.
2. Second, factors are extracted from the correlation matrix based on the correlation coefficients of the variables.
3. Third, the factors are rotated in order to maximize the relationship between the variables and some of the factors.

Example

You may be interested to investigate the reasons why customers buy a product such as a particular brand of soft drink (e.g. coca cola). Several variables were identified which influence customer to buy coca cola. Some of the variables identified as being influential include *cost of product*, *quality of product*, *availability of product*, *quantity of product*, *respectability of product*, *prestige attached to product*, *experience with product*, and *popularity of product*. From this, you designed a questionnaire to solicit customers' view on a seven point scale, where 1 = not important and 7 = very important. The results from your questionnaire are show on the table below. Only the first twelve respondents (cases) are used in this example.

Table 1: Customer survey

cost	quality	avabity	quantity	respect	prestige	experie	popula
1	3	4	6	7	2	4	5
2	3	4	3	4	6	7	6
4	5	6	7	7	2	3	4
3	4	5	6	7	3	5	4
2	5	5	5	6	2	4	5
3	4	6	7	7	4	3	5
2	3	6	4	5	4	4	4
1	3	4	5	6	3	3	4
3	3	5	6	6	4	4	3
4	4	5	6	7	4	3	4
2	3	6	7	5	4	4	4
2	3	5	7	6	3	3	3

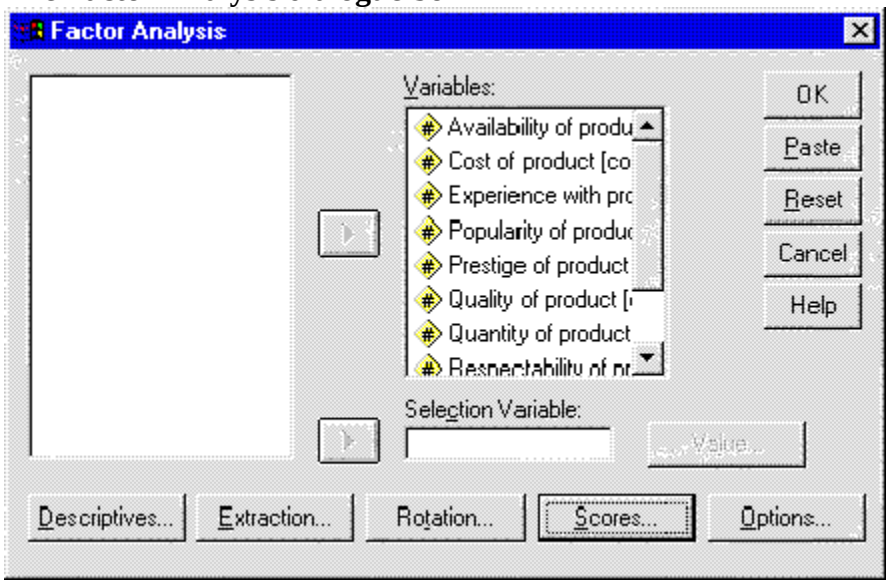
Preparing the Data set

Prepare and enter the data into SPSS **Data Editor** window. If you do not know how to create SPSS data set see *Getting Started with SPSS for Windows*. Define the eight variables *cost*, *quality*, *avabity*, *quantity*, *respect*, *prestige*, *experie*, *popula*, and use the **Variable Labels** procedure to provide fuller labels *cost of product*, *quality of product*, *availability of product*, *respectability of product*, and so on to the variables names. The completed data set look like the one shown above in Table 1.

Running the Factor Analysis Procedure

From the menu bar select **Statistics** and choose **Data Reduction** and then click on **Factor**. The **Factor Analysis** dialogue box will be loaded on the screen. Click on the first variables on the list and drag down to highlight all the variables. Click on the arrow (>) to transfer them to the **Variables** box. The completed dialogue box should look like the one shown below.

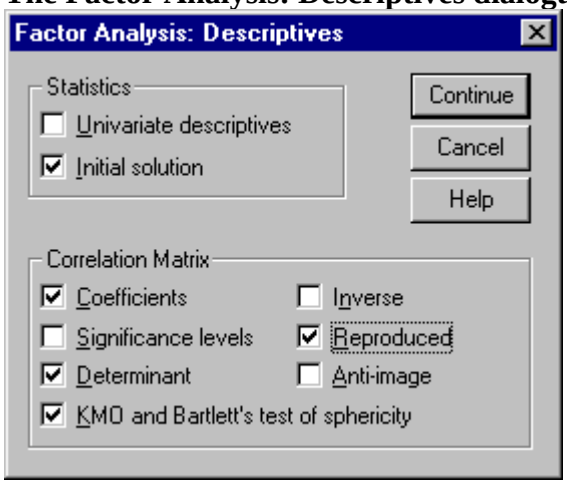
The Factor Analysis dialogue box



All we need to do now is to select some options and run the procedure.

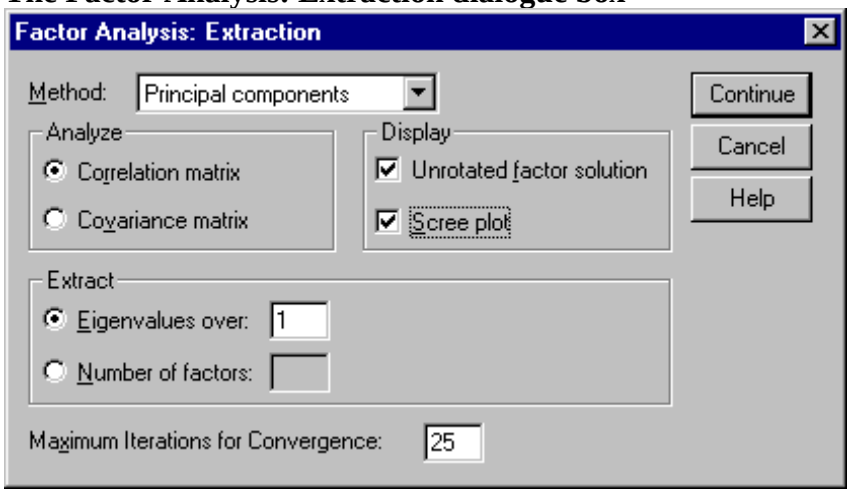
Click on the **Descriptives** button and its dialogue box will be loaded on the screen. Within this dialogue box select the following check boxes **Coefficients**, **Determinant**, **KMO and Bartlett's test of sphericity**, and **Reproduced**. Click on **Continue** to return to the **Factor Analysis** dialogue box. The **Factor Analysis: Descriptives** dialogue box should be completed as shown below.

The Factor Analysis: Descriptives dialogue box



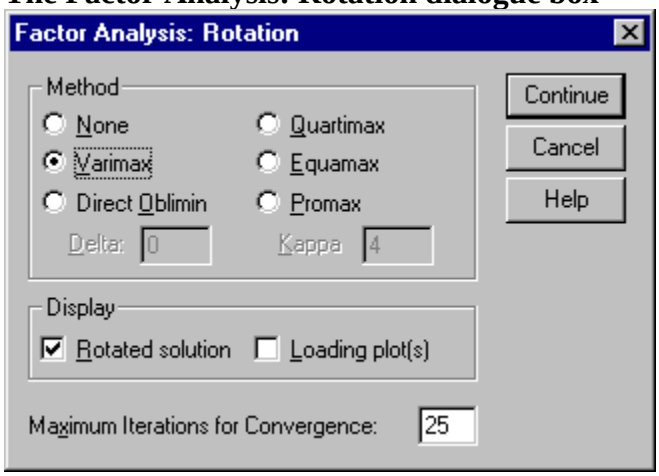
From the **Factor Analysis** dialogue box click on the **Extraction** button and its dialogue box will be loaded on the screen. Select the check box for **Scree Plot**. Click on **Continue** to return to the **Factor Analysis** dialogue box. The **Factor Analysis: Extraction** dialogue box should be completed as shown below.

The Factor Analysis: Extraction dialogue box



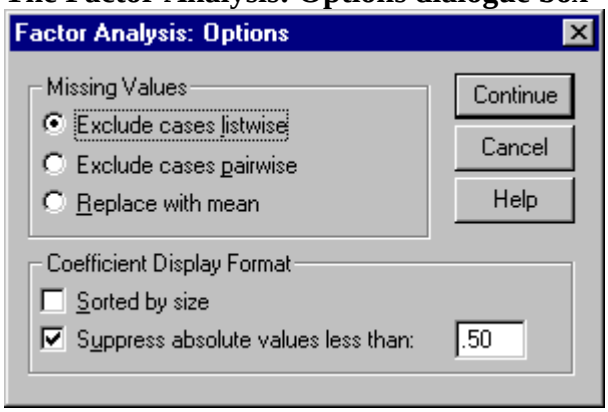
From the **Factor Analysis** dialogue box click on the **Rotation** button and its dialogue box will be loaded on the screen. Click on the radio button next to **Varimax** to select it. Click on **Continue** to return to the **Factor Analysis** dialogue box. The **Factor Analysis: Rotation** dialogue box should be completed as shown below.

The Factor Analysis: Rotation dialogue box



From the **Factor Analysis** dialogue box click on the **Options** button and its dialogue box will be loaded on the screen. Click on the check box of **Suppress absolute values less than** to select it. Type **0.50** in the text box. Click on **Continue** to return to the **Factor Analysis** dialogue box. Click on **OK** to run the procedure. The **Factor Analysis: Options** dialogue box should be completed as shown below.

The Factor Analysis: Options dialogue box



Interpretation of the Output

Descriptive Statistics

The first output from the analysis is a table of descriptive statistics for all the variables under investigation. Typically, the *mean*, *standard deviation* and *number of respondents* (N) who participated in the survey are given. Looking at the *mean*, one can conclude that *respectability of product* is the most important variable that influence customers to buy the product. It has the highest *mean* of 6.08.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Availability of product	5.08	.79	12
Cost of product	2.42	1.00	12
Experience with product	3.92	1.16	12
Popularity of product	4.25	.87	12
Prestige of product	3.42	1.16	12
Quality of product	3.58	.79	12
Quantity of product	5.75	1.29	12
Respectability of product	6.08	1.00	12

The Correlation matrix

The next output from the analysis is the correlation coefficient. A correlation matrix is simply a rectangular array of numbers which gives the correlation coefficients between a single variable and every other variables in the investigation. The correlation coefficient between a variable and itself is always 1, hence the principal diagonal of the correlation matrix contains 1s. The correlation coefficients above and below the principal diagonal are the same. The determinant of the correlation matrix is shown at the foot of the table below.

Correlation Matrix

		Availability of product	Cost of product	Experience with product	Popularity of product	Prestige of product	Quality of product	Quantit produ
Correlation	Availability of product	1.000	.527	-.386	-.298	-.041	.349	.
	Cost of product	.527	1.000	-.202	-.237	.072	.585	.
	Experience with product	-.386	-.202	1.000	.563	.564	-.238	-.
	Popularity of product	-.298	-.237	.563	1.000	.248	.165	-.
	Prestige of product	-.041	.072	.564	.248	1.000	-.484	-.
	Quality of product	.349	.585	-.238	.165	-.484	1.000	.
	Quantity of product	.467	.372	-.682	-.509	-.470	.245	1.
	Respectability of product	.105	.420	-.620	-.237	-.660	.508	.

a. Determinant = 1.731E-03

Kaiser-Meyer-Olkin (KMO) and Bartlett's Test

The next item from the output is the Kaiser-Meyer-Olkin (KMO) and Bartlett's test. The KMO measures the sampling adequacy which should be greater than 0.5 for a satisfactory factor analysis to proceed. Looking at the table below, the KMO measure is 0.417. From the same table, we can see that the Bartlett's test of sphericity is significant. That is, its associated probability is less than 0.05. In fact, it is actually 0.012. This means that the correlation matrix is not an identity matrix.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.417
Bartlett's Test of Sphericity	Approx. Chi-Square	47.694
	df	28
	Sig.	.012

Communalities

The next item from the output is a table of communalities which shows how much of the variance in the variables has been accounted for by the extracted factors. For instance over 90% of the variance in *quality of product* is accounted for while 73.5% of the variance in *availability of product* is accounted for.

Communalities

	Initial	Extraction
Availability of product	1.000	.735
Cost of product	1.000	.844
Experience with product	1.000	.800
Popularity of product	1.000	.804
Prestige of product	1.000	.865
Quality of product	1.000	.918
Quantity of product	1.000	.768
Respectability of product	1.000	.814

Extraction Method: Principal Component Analysis.

Total Variance Explained

The next item shows all the factors extractable from the analysis along with their eigenvalues, the percent of variance attributable to each factor, and the cumulative variance of the factor and the previous factors. Notice that the first factor accounts for 46.367% of the variance, the second 18.471% and the third 17.013%. All the remaining factors are not significant.

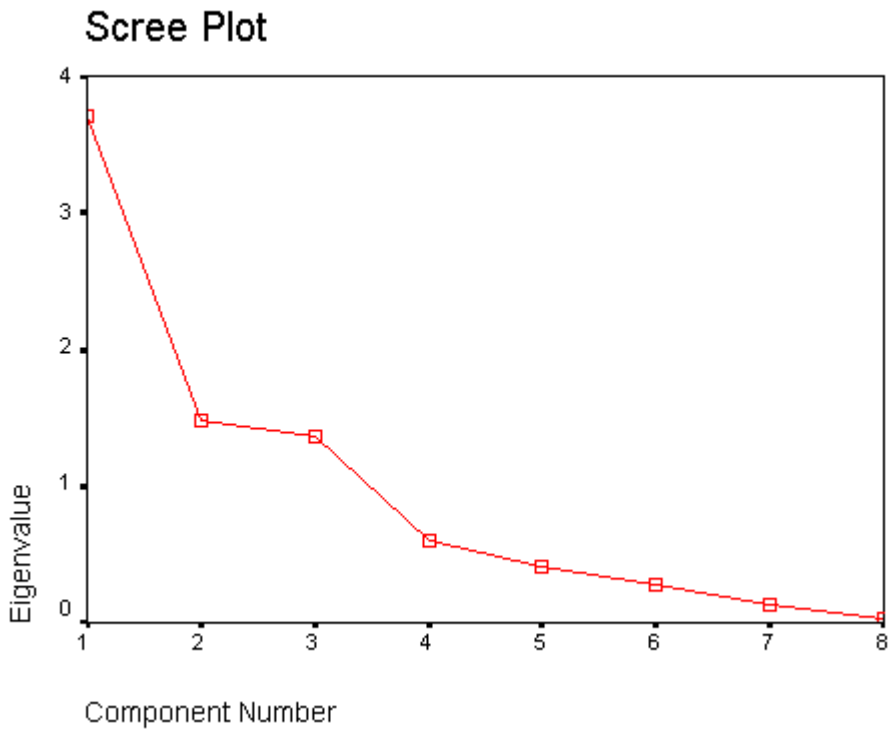
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loading		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulati %
1	3.709	46.367	46.367	3.709	46.367	46.367	2.501	31.260	31.26
2	1.478	18.471	64.838	1.478	18.471	64.838	2.045	25.566	56.82
3	1.361	17.013	81.850	1.361	17.013	81.850	2.002	25.024	81.85
4	.600	7.499	89.349						
5	.417	5.214	94.563						
6	.281	3.508	98.071						
7	.129	1.608	99.679						
8	2.569E-02	.321	100.000						

Extraction Method: Principal Component Analysis.

Scree Plot

The scree plot is a graph of the eigenvalues against all the factors. The graph is useful for determining how many factors to retain. The point of interest is where the curve starts to flatten. It can be seen that the curve begins to flatten between factors 3 and 4. Note also that factor 4 has an eigenvalue of less than 1, so only three factors have been retained.



Component (Factor) Matrix

The table below shows the loadings of the eight variables on the three factors extracted. The higher the absolute value of the loading, the more the factor contributes to the variable. The gap on the table represent loadings that are less than 0.5, this makes reading the table easier. We suppressed all loadings less than 0.5.

Component Matrix[a]

	Component		
	1	2	3
Availability of product	.546		
Cost of product	.563	.697	
Experience with product	-.817		
Popularity of product	-.534		.598
Prestige of product	-.660		-.526
Quality of product	.578	.570	.508
Quantity of product	.841		
Respectability of product	.815		

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Rotated Component (Factor) Matrix

The idea of rotation is to reduce the number factors on which the variables under investigation have high loadings. Rotation does not actually change anything but makes the interpretation of the analysis easier. Looking at the table below, we can see that *availability of product*, and *cost of product* are substantially loaded on Factor (Component) 3 while *experience with product*, *popularity of product*, and *quantity of product* are substantially loaded on Factor 2. All the remaining variables are substantially loaded on Factor 1. These factors can be used as variables for further analysis.

Rotated Component Matrix[a]

	Component		
	1	2	3
Availability of product			.760
Cost of product			.908
Experience with product	-.560	.684	
Popularity of product		.893	
Prestige of product	-.901		
Quality of product	.637		.634
Quantity of product		-.652	
Respectability of product	.845		

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Conclusion

You should now be able to perform a factor analysis and interpret the output. Many other items are produce in the output, for the purpose of this illustration they have been ignored. Note that the correlation matrix can used as input to factor analysis. In this case you have to use SPSS command syntax which is outside the scope of this document.