


STATISTICS WITH SPSS FOR RESEARCH

A 3D bar chart with yellow bars of varying heights, overlaid with a thick red line graph that trends upwards from left to right. The background is a vibrant pink with a pattern of 3D cubes.

TAY CHOO CHUAN
MOHD RAZALI MUHAMAD
TAM CAI LIAN
SEK YONG WEE
SITI AZIRAH ASMAI

STATISTICS WITH SPSS FOR RESEARCH

Tay Choo Chuan
Mohd Razali Muhamad
Tam Cai Lian
Sek Yong Wee
Siti Azirah Asmai

Penerbit Universiti
Universiti Teknikal Malaysia Melaka

© FIRST PUBLISHED 2011
Universiti Teknikal Malaysia Melaka

ISBN: 978-967-0257-04-04

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, electronic, mechanical photocopying, recording or otherwise, without the prior permission of the Publisher.

Perpustakaan Negara Malaysia

Cataloguing-in-Publication Data

Statistic with SPSS for research / Tay Choo Chuan ... [et al.]

Bibliography: p. 201

ISBN 978-967-0257-04-4

1. Statistic--Study and teaching (Higher). 2. SPSS for Windows.

I. Tay, Choo Chuan.

519.537

Printed and Published in Malaysia by:

Penerbit Universiti
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100, Durian Tunggal, Melaka

TABLE OF CONTENTS

CONTENTS	PAGES
Preface.....	v
Acknowledgements.....	vii
Chapter One	
Correlation.....	1
Worked Examples.....	21
Review Exercises.....	31
Chapter Two	
Regression.....	35
Worked Examples.....	47
Review Exercises.....	63
Chapter Three	
t – Test.....	67
Worked Examples.....	71
Review Exercises.....	89
Chapter Four	
ANOVA.....	93
Worked Examples.....	97
Review Exercises.....	117
Chapter Five	
Crosstabs Procedure.....	121
Worked Examples.....	141
Review Exercises.....	149
Review Exercises' Answers.....	153
References.....	189
Appendices.....	191

ACKNOWLEDGEMENT

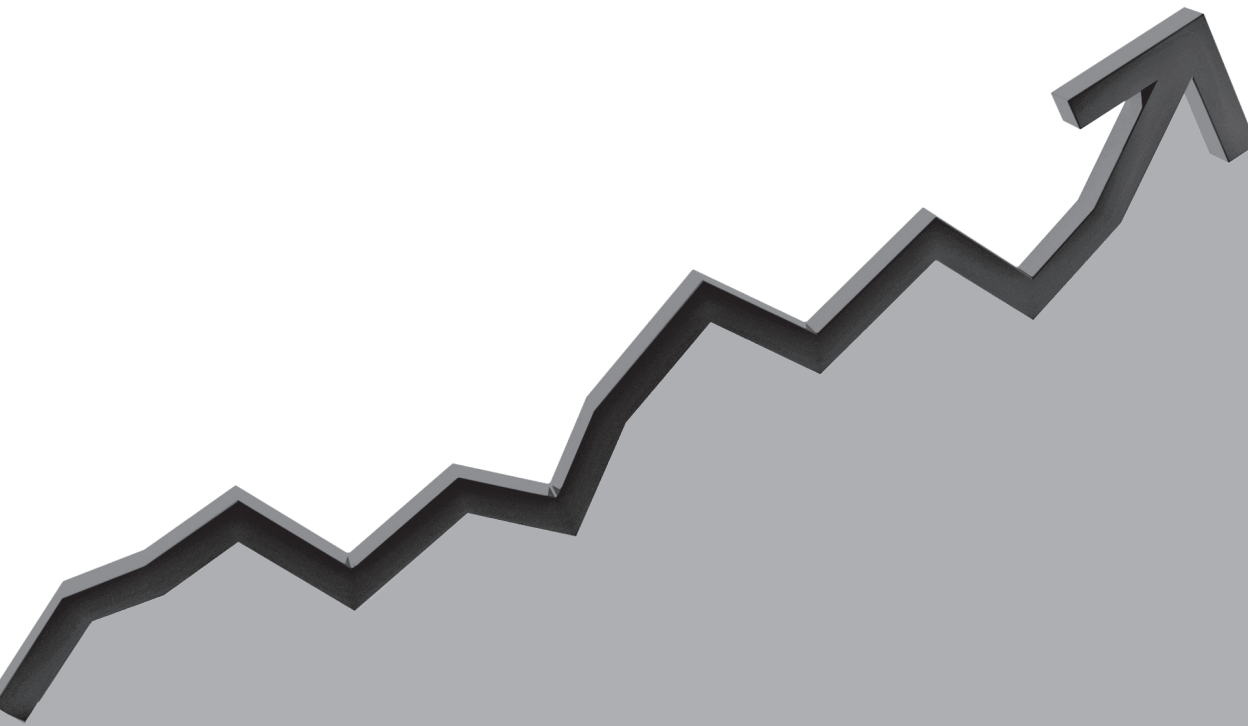
There are many individuals, who gave freely of their time, wisdom and insight. Without producing a long list, to them we are especially grateful. Their spirit of sharing, openness and generosity made it all possible. We are much indebted to them.

PREFACE

Many students are having difficulties in learning Statistics. For the purpose of helping students to master the skills of this subject, step by step easy approach has been designed and adopted by the authors of this book. This is deemed as the best approach when the design of the content is structured and well-organized. Furthermore, a large part of the syllabus involves the mastery of manual mathematical calculation and step by step, direct instruction is the best approach in order to obtain an answer. In addition, for students who are not so good in math-logic reasoning and dealing with numbers, this method may enable them a sure way of obtaining the correct answer by following a set of given guidelines. Assessment is the process of evaluating the extent to which students have developed their knowledge, understanding and abilities. It involves determining what does the student know, understand, and can do with the knowledge gained from the lessons. For the purpose of evaluating whether the students have achieved the objectives of the chapters, exercises in form of assessments based on theory was used. In order to maximize the student's learning capability, the authors took effort in providing extra exercises that cover the manual calculation and SPSS calculation for all different chapters. We earnestly hope that Step by Step Learning SPSS would work as a good guide book that enhance the level of understanding of the concepts and enable students to execute the appropriate steps in coming up with the correct interpretation.

chapter one

CORRELATION



Learning Objectives

At the completion of the chapter, students should be able to:

- Apply the technique of data collection using paper and pencil methods.
- Differentiate the direction of the relationship between two variables which include positive, negative and zero.
- Analyze the strength or magnitude of the relationship using correlation coefficients.
- Illustrate the linear relationship between variables using Scatter plot.
- Interpret the results in a manner that provide answers to research questions.
- Construct simple types of graphs and charts.
- Apply the technique of data collection using statistical software.

1.0 Introduction

The first chapter focuses on the nature of statistical data of correlation. The aim of the series of exercises is to ensure the students are able to use SPSS to explain the statistical theories and calculations. Meanwhile they are also competent in the analysis of data through manual calculation.

Correlation refers to a statistical measurement to measure and describe the relationship between two variables whereby changes in one variable will associated with be a concurrent change in the other variable.

According to Gravetter and Wallnau (2005), the two variables in correlation are observed as they naturally existed in the environment. In any of the correlation study, there shouldn't be any attempt or effort to manipulate the variables. For example, the relationship between number of hours of revision and anxiety level during exam could measure and record the number of hours spent by a group of students to do revision and then observe their anxiety levels in the exam hall, but the researcher is merely observing the occurrence naturally with no attempt to manipulate the number of hours spent and also their anxiety level in the exam hall.

1.1 Scatter Plots

A **scatter plot** is an extremely useful tool when it comes to looking at the association between two variables (Caldwell, 2007). In short, a scatter plot allows simultaneously viewing the values of two variables on a case-by-case basis. A typical example used to illustrate the utility of a scatter plot involves the association between height and weight. Table 1.1 shows a hypothetical distribution of values of those variables (height and weight) for 20 cases.

Table 1.1: Height/Weight Data for 20 Cases

Case	Height (Inches)	Weight (pounds)
1	59	92
2	61	105
3	61	100
4	62	107
5	62	114
6	63	112
7	63	120
8	63	130
9	64	132
10	64	137
11	65	132
12	65	138
13	65	120
14	66	136
15	66	132
16	67	140
17	67	143
18	68	139
19	68	134
20	69	153

A visual representation of the same data in the form of a scatter plot is shown in Figure 1. Height measurements values are shown along the horizontal or X axis of the graph; weight measurement values are shown along the vertical or Y axis of the graph. Focusing on case number 1, shown in the lower left corner of the scatter plot, interpret the point as reflecting a person (case) with a height of 59 inches (or 4' 11") and a weight of 92 pounds.

Each of the 20 points can be interpreted in the same fashion—a reflection of the values of two variables (height and weight) for a given case. Note that the scales along the X and Y axes are different. The variable of height is expressed in inches, but the variable of weight is expressed in pounds.

1.2 Linear Associations: Direction and Strength

Two variables (Variable X and Variable Y) can be associated in several ways. A scatter plot can provide a graphic and concise statement as to the general relationship or association between two variables. In short, a scatter plot tells something about the *direction* and *strength* of association.

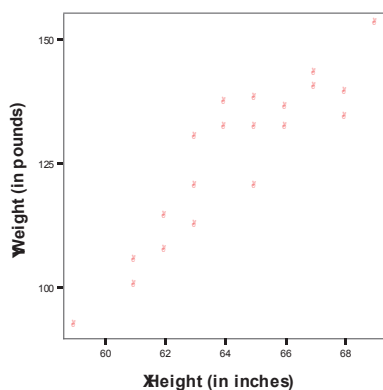


Figure 1.1: Height/Weight Data for 20 cases: Scatter Plot

To create a scatter plot, the score obtained from the observation will be recorded as X and Y in a table and a scatter plot will be used to represent the data. In scatter plot, the X values will be placed on the horizontal axis of a graph and the Y values are placed on the vertical axis of the graph. The pattern of the data will show the relationship between the X values and Y values.

To say that two variables are related or associated in a positive or direct association is to say that they track together; it means that high values on Variable X are generally associated with high values on Variable Y , and low values on Variable X are generally associated with low values of Variable Y . In a negative or inverse association, high values on Variable X are associated with low values on Variable Y , and low values on Variable X are associated with high values on Variable Y . In short, the variables track in opposite directions.

The idea of a perfect relationship is useful, because it helps to understand what is meant by strength of association. In many respects, the strength of an association is just an expression of how close one might be to being able to predict the value on one variable from knowledge of the value on another variable. Some associations are stronger than others in the sense that they come closer than others to the notion of perfect predictability.

1.3 Positive Correlation and Negative Correlation

A positive correlation occurs when one variable indicates a high score while the other also indicates a high score; or when one variable indicates a low score and the other also indicates a low score. Take for example a study to find out whether there is a relationship between going on a diet and becoming thin. If the results indicate that the less people eat, the thinner they become; then we can conclude that there is a relationship between these two variables, and in this case, it is a positive correlation. Take another example, a study to find out whether there is a relationship between those who eat lots of sweet things and getting diabetes. If the results indicate that the more people eat sweet things, the higher chances of them getting diabetes; then we can conclude that there is a relationship between these two variables, and in this case, it is also a positive correlation.

A negative correlation on the other hand occurs when one variable indicates a high score while the other indicates a low score. Take for example a study to find out whether there is a relationship between people who drink lots of milk and the occurrence of osteoporosis. If the results indicate that the more the people drink milk, the less likely they get osteoporosis; then we can conclude that there is a relationship between these two variables, and in this case, it is a negative correlation.

What is important to note in correlation is not whether it is a positive correlation or a negative correlation. What is important is the degree of relationship, also known as 'correlation coefficient' or simply ' r '. The bigger the degree of the correlation, i.e. the higher the value of r the stronger the relationship is between the two variables.

However, when the study indicates a correlation of 0 or almost 0, it means that there is no relationship between the two variables. This usually occurs when the two variables measured do not have any relevance to each other. Take for example, a study to find out whether there is any relationship between people who are intelligent and the ability to play badminton. If the results indicate a very scattered data and when calculated, it indicates a 0, we can conclude that there is no relationship between these two variables. This means that the ability to play badminton has nothing to do with the intelligence of that person.

The results can be observed either through a manual calculation process or by using the SPSS software. Firstly, the manual calculation method will be explained and then followed by the exploration of SPSS.

In general, there are 3 types of characteristics that correlation is measuring between variable X and variable Y as stated below:

- Firstly is the direction of the relationship.

A positive correlation is a direct relationship where as the amount of one variable increases, the amount of a second variable also increases.

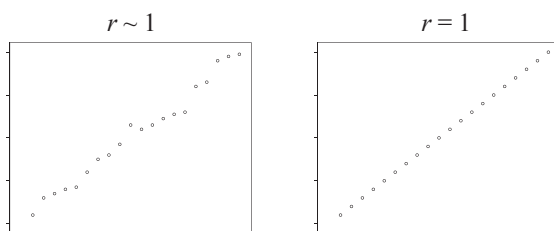


Figure 1.2a: Scatter plot of positive correlation

In a negative correlation, as the amount of one variable goes up, the levels of another variable go down.

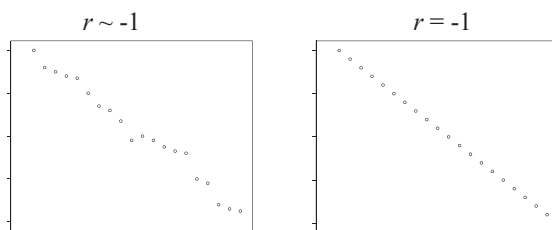


Figure 1.2b: Scatter plot of negative correlation

In both types of correlation, there is no evidence or proof that changes in one variable cause changes in the other variable. A correlation simply indicates that there is a relationship between the two variables.

In zero correlation, there is no relationship between the variables.

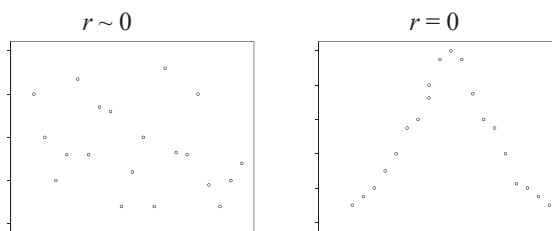


Figure 1.2c: Scatter plot of zero correlation

- The Form of the Relationship
 - The commonly seen form of correlation is the linear form which the points in the scatter plot tend to form a straight line. Correlation is also commonly used to measure the straight-line relationships.

- The Degree of the Relationship
 - A perfect correlations is always identified by a correlation of +1.00 or - 1.00 where the positive and negative sign does not reflects the face value of the data, it simply shows the direction of the data.
 - A zero correlation simply shows the data do not fit at all, it shows no relationship between the two measured variables.

1.4 Two Variables: *X* and *Y*

When speaking in terms of *Variable X* and *Variable Y*, it is proposed that some logical connection between the two. For example, the *X* variable is commonly regarded as the **independent variable**, and the *Y* variable is commonly regarded as the **dependent variable**. In the language of research, an independent variable is a variable that is presumed to influence another variable. The dependent variable, in turn, is the variable that is presumed to be influenced by another variable.

For example, it is common to assert that there is a connection between a person's level of education and level of income. Educational level would be treated as the independent variable, and the level of income would be regarded as the dependent variable. In other words, education (independent) is thought to exert an influence on income (dependent).

1.5 The Logic of Simple Correlation

In truth, correlation analysis takes many forms (such as multiple correlation or partial correlation); the one being considered here is referred to as **simple correlation**. In other words, simple correlation analysis allows measuring the association between two interval/ratio level variables (assuming that the two variables, if associated, are associated in a linear fashion).

Z transformation is used to convert the scores on different scales to a single scale based on Z scores (or points along the baseline of the normal curve).

$$Z = \frac{\text{Raw score}}{\text{Standard Deviation}}$$

For example, look back at Figure 1. Note that the values along the horizontal and vertical axes are expressed in different scales or units of measurement; inches along the horizontal axis and pounds along the vertical axis. When considering the raw scores of the points represented in the scatter plot, then, deal with two different scales. The two sets of scores will be the on the same scale, though, if they're transformed into Z scores. The same is true for any number of situations.

For example, student aptitude test scores (SAT scores) and grade point averages (GPAs). When expressed as raw scores, are based on very different underlying scales, but the raw scores can easily be transformed into Z scores to create a single scale of comparison. Data on education (expressed as the number of school years completed) and income (expressed in dollars) can share a common scale when transformed into Z scores. The list goes on. All it needs is the mean and standard deviation of each distribution. It is a simple

transformation: Subtract the mean for each raw score in the distribution, and divide the difference by the end of standard deviation.

1.6 The Formula for Pearson's r

Because the computational formula r includes the steps necessary to convert raw scores to Z scores, it has a way of appearing extremely complex. Assume knowing the basis of Pearson's r (namely, the conversion raw scores into Z scores), though in a position to rely on a more *conceptual* formula. The heart of the more conceptual approach has to do with what is referred to as the cross products of the Z scores.

Table 1.2 shows pairs of values or scores associated with 10 cases. Columns 2 and 4 show the raw score distributions for the two variables, X and Y . The means and standard deviations of the raw score distributions are given at the bottom of the table. Columns 3 and 5 show the Z scores or transformations based on the associated raw scores. (Recall that these are calculated by subtracting the mean of each raw score and dividing it by the standard deviation). Case number 1, for example, has a raw score X value of 20 (shown in Column 2) and a Z score (Z_X) value of -1.49 (shown in Column 3). The raw score Y value for case number 1 is 105 (shown in Column 4), and the Z score (Z_Y) value of -1.57 (shown in Column 5).

The cross products are obtained by multiplying each Z_X value (the entry in Column 3) by the associated Z_Y value (the entry in Column 5). The results of the cross product multiplication are shown in Column 6 ($Z_X \cdot Z_Y$).

As in Table 1.2, the symbol Z_X denotes the Z scores for the X variable, and Z_Y denotes the Z scores for the Y variable. Sum the cross products and divide the sum by the number of paired cases minus 1. The result is the calculated r .

$$r = \frac{\sum(Z_X \cdot Z_Y)}{n - 1}$$

Table 1.2: Cross Product Calculations for X and Y Variables (Positive Association)

(1) Case	(2) X	(3) Z_X	(4) Y	(5) Z_Y	(6) $Z_X \cdot Z_Y$
1	20	-1.49	105	-1.57	2.34
2	25	-1.16	126	-0.84	0.97
3	30	-0.83	122	-0.98	0.81
4	35	-0.50	130	-0.70	0.35
5	40	-0.17	155	0.17	-0.03
6	45	0.17	159	0.31	0.05
7	50	0.50	153	0.10	0.05
8	55	0.83	184	1.18	0.98
9	60	1.16	177	0.94	1.09
10	65	1.49	190	1.39	2.07
Sum of Cross Products = 8.68					
Mean of X = 42.50					
Standard Deviation of X = 15.14					
Mean of Y = 150.10					
Standard Deviation of Y = 28.65					

Note that using $n-1$ in the denominator of the formula that some presentations of the formula rely on n alone. The difference in the two approaches traces back to the manner in which the standard deviation for each distribution was calculated (recall that the standard deviation is a necessary ingredient for the calculation of a Z score). The assumption in this text is that $n-1$ was used in the calculation of the standard deviations of Variable X and Variable Y .

The formula that follows, for example, is typical of how a computational; the formula for r might be presented:

$$r = \frac{\sum(Z_X \cdot Z_Y)}{n - 1}$$

Such a formula can be very handy if when using a calculator to compute the value of r . Given the increasing use of computers and statistical software, however, the real issue is likely to be whether or not a solid understanding of what lies behind a procedure and how to interpret the results. In the case of Pearson's r , the conceptual formula, based on the cross product calculations, gives one a better understanding of what is really involved in the calculation.

For the data presented in Table 1.2, the calculation is as follows:

$$r = \frac{\sum(Z_X \cdot Z_Y)}{n - 1}$$

$$r = \frac{8.68}{9}$$

$$r = +0.96$$

1.6.1 Interpretation

The calculated r value is $+0.96$, but there is still the question of how one would interpret it. Statisticians actually use the information provided by the r value in two ways.

The value of r is referred to as the **correlation coefficient**. The sign (+ or -) in front of the r value indicates whether the association is positive (direct) or negative (inverse). The absolute value of r (the magnitude, without respect of the sign) is a measure of the strength of the relationship. The closer the value gets to 1.0 (either $+1.0$ or -1.0), the stronger the association.

1.6.2 Considerations while interpreting correlations

- Correlation is merely describing a relationship between two variables. It does not prove and explain which variable causes the other variable to change. For example, the amount of time spent on reading does not necessary cause the increase of IQ level of a person. There is a relationship between these two variables but correlation is not trying to explain or prove the causal relationship.
- Healey (2007) stated that the value of correlation must come from a representative sample size. The range of scores represented in the data will greatly affect the value of correlation.
- The outlier or the extreme data points will also influence the value of a correlation greatly (Gravetter & Wallnau, 2005).

- Correlation must always be squared, r^2 before the relationship is being judged. This value is known as Coefficient of Determination (Gravetter & Wallnau, 2005). For example, when a correlation, $r = +.6$ is calculated, it does not mean it is a moderate degree being $+.6$ is half way in between 0 and $+1.00$. Although $+1.00$ is equals to 100% perfectly predictable relationship, a correlation of $.6$ does not equals to 60% accuracy. To describe the accuracy, one must square the correlation. Thus, when $r = .6$, the accuracy of predicting the relationship is equals to $r^2 = .6^2 = .36$, or 36% accuracy.

1.6.3 Pearson Correlation (Pearson product-moment correlation)

The Pearson correlation measures the degree and direction of the linear relationship between two variables. Below are the computational formulas involved in measuring the relationship between two variables.

$$\text{Calculated } r: r_{\text{cal}} = \frac{(n \cdot \sum XY) - (\sum X \cdot \sum Y)}{\sqrt{[(n \cdot \sum X^2) - (\sum X)^2] \times [(n \cdot \sum Y^2) - (\sum Y)^2]}}$$

Degree of freedom, $df = n - 2$

One condition for correlation to exist is that, the results should indicate that the *computed 'r'* is higher than the *critical 'r'*. *Computed 'r'* refers to the value which is obtained either through manual calculation or through SPSS. *Critical 'r'* on the other hand, refers to the value which is obtained from the statistical table.

To get the critical value of r from the table, one must know the sample size (n), the probability of making error (alpha level). When the $r_{\text{cri}} > r_{\text{cal}}$, it shows that there is no significant relationship between the 2 variables and when $r_{\text{cri}} < r_{\text{cal}}$, it shows that there is a significant relationship between the two variables.

When there is no significant relationship between the two variables, it means the researcher has to accept the null hypothesis whereas when there is a significant relationship between the two variables, the researcher has to reject the research hypothesis.

1.7 Correlation for SPSS

Correlation analysis is used to describe the relationships and directions between two variables. There are different ways available from SPSS to measure correlation. In this chapter, Pearson product moment correlation coefficient is presented.

There are 2 types of correlation.

- Simple bivariate correlation: also known as zero-order correlation which means relationship between two variables.
- Partial correlation: Such correlation will allow you to explore the relationship between two variables, while at the same time controlling for another variable.

1.8 Pearson Correlation Coefficients (r)

The range of r is from -1 to +1. Positive relationship indicates a positive relationship between two variables. This means when one variable increases, another variable will also increase. For example, increasing number of hours in study is associated with better performance in the examination. Whereas negative relationship indicates a negative relationship between two variables. As one variable increases, the other decreases. For example, increasing hours of watching television are associated with lower performance in the examination.

The absolute value (ignoring the sign) indicates the strength of the relationships. For example:

- +1 or -1 means perfect correlation
- 0.9 means positive strong correlation
- 0.6 means positive average/moderate correlation
- 0.1 means positive weak correlation
- -0.9 means negative strong correlation
- -0.6 means negative average/moderate correlation
- -0.1 means negative weak correlation
- 0 means no relationship between the two variables

Table 1.3a: Strength of Positive Correlation Coefficients

Perfect											+1
Strong										+0.9	
									+0.8		
								+0.7			
Moderate						+0.6					
					+0.5						
				+0.4							
Weak			+0.3								
		+0.2									
		+0.1									
Zero	0										

Table 1.3b: Strength of Negative Correlation Coefficients

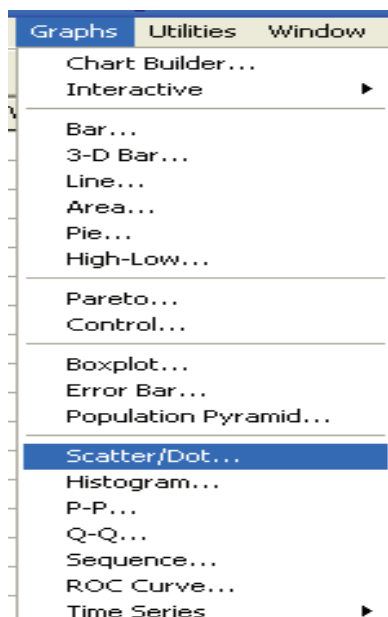
Perfect	-1										
Strong		-0.9									
			-0.8								
				-0.7							
Moderate					-0.6						
						-0.5					
							-0.4				
Weak								-0.3			
									-0.2		
										-0.1	
Zero											0

1.9 Scatterplot

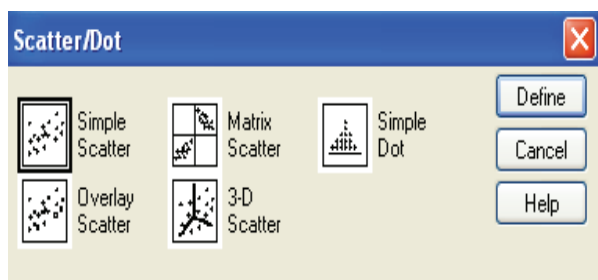
Before performing a correlation analysis, it is better to generate a scatterplot. This enables you to check for the assumption of linearity. In addition, it gives you better idea of the direction and relationship between the variables.

1.9.1 Procedure for Generating a Scatter plot

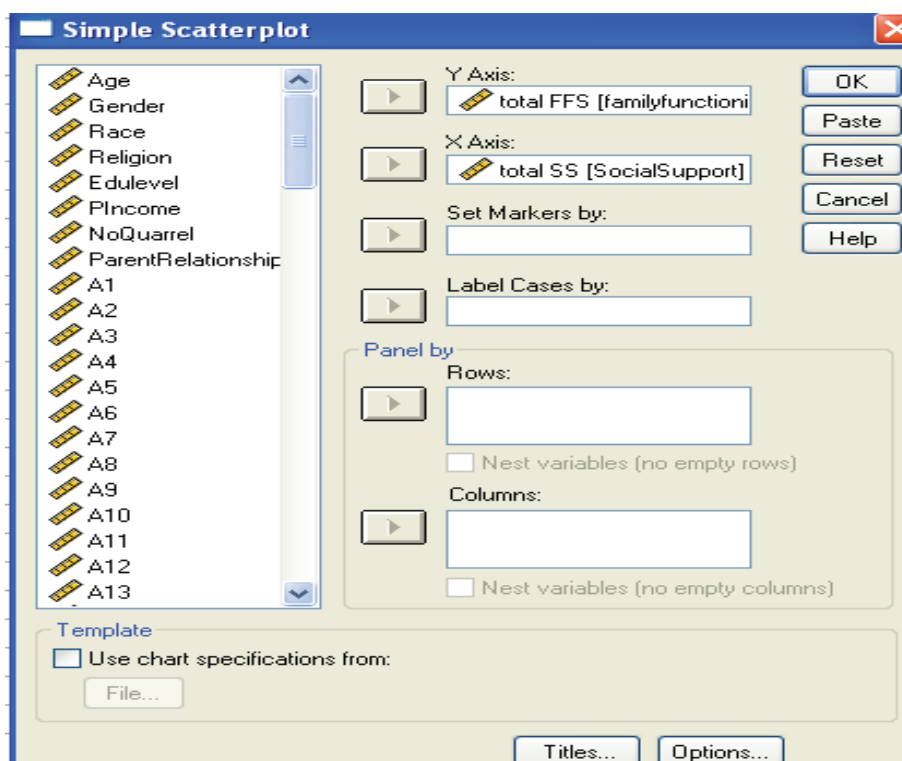
- Use Data Set1 Family Functioning
- Select **Graph** from the menu
- Click on **Scatter** as shown in the following figure



- Click on **Simple Scatter** and **define** button as shown in the following figure



- Move dependent variable to Y-axis
- Move independent variable to X-axis
- Click on continue and then **OK** as shown in the following figure



1.9.2 Details of Example: Family Functioning, Social Support and Self-Esteem

The interrelationships among some of the variables included in the family functioning data set provided in the CD given can be used to demonstrate the use of correlation. The survey was designed to find out the relationship among family functioning, perceived social support and self-esteem among college and university students. Refer Data Set1 Family Functioning.

In this survey, the aim is to examine relationships between variables. This data set contains information on:

- Age
- Gender
- Race
- Religion
- Educational Level
- Parents Income
- Number of Quarrels
- Parent Relationship
- Family Functioning
- Social Support
- Self Esteem

1.9.3 Use Pearson's r to show the relationship between Family Functioning and Social Support.

Answer:

Table 1.4a: Correlation between Social Support and Self-esteem

Family Functioning	
	.377(**)
Social Support	

** $p < .01$

The relationship between family functioning and social support was explored using Pearson's product moment correlation. Result revealed that there was a positive correlation ($r=.377, p<.01$) between family functioning and social support suggesting that students with better family functioning have higher perceived social support. See Table 7.

Overall, social support contributed 14.21% ($r = .377$) to family functioning.

Table 1.4b: Correlations with SPSS

		total FFS	total SS
total FFS	Pearson Correlation	1	.377(**)
	Sig. (2-tailed)		.000
	N	246	246
total SS	Pearson Correlation	.377(**)	1
	Sig. (2-tailed)	.000	
	N	246	246

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

1.9.4 Use Pearson's r to show the relationship between Family Functioning and Self-Esteem.

Table 1.4c: Correlations with SPSS

		total FFS	Total SE
total FFS	Pearson Correlation	1	.117
	Sig. (2-tailed)		.066
	N	246	246
total SE	Pearson Correlation	.117	1
	Sig. (2-tailed)	.066	
	N	246	246

Answer

There is no significant relationship between family functioning and self-esteem
 $r = .117, p>0.05$

1.9.5 Let us proceed further and examine how family functioning and social support predicts student's self-esteem. Use Pearson correlation matrix in the analysis.

Table 1.4d: Correlations with SPSS

		total FFS	Total SE	total SS
total FFS	Pearson Correlation	1	.117	.377(**)
	Sig. (2-tailed)		.066	.000
	N	246	246	246
total SE	Pearson Correlation	.117	1	.439(**)
	Sig. (2-tailed)	.066		.000
	N	246	246	246
total SS	Pearson Correlation	.377(**)	.439(**)	1
	Sig. (2-tailed)	.000	.000	
	N	246	246	246

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

Answer

- There is a significant relationship between family functioning social support, $r = .377$, $p < .01$. When there is more social support, there is better family functioning. $r^2 = .142$. Social support contributes 14.2% towards family functioning.
- There is a significant relationship between social support and self esteem, $r = .439$, $p < .01$. When there is more social support, the student's self esteem becomes higher. $r^2 = .193$. Social support contributes 19.3% towards self esteem.
- There is no significant relationship between family functioning and self-esteem, $r = .117$, $p > .01$.

1.9.6 Examine the following scatterplots. Try to predict:

- The strength /magnitude of the relationships (perfect, strong, moderate, weak or zero).
- The direction of the relationships (positive, negative or zero)

Answer

Positive Weak Relationship between family functioning and social support, $r = .377$, $n = 246$

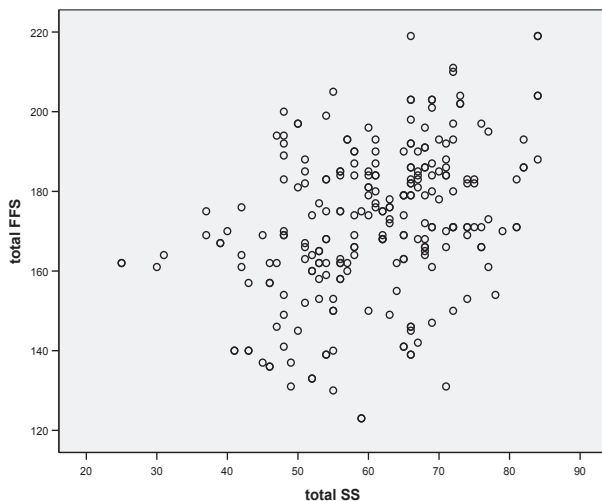


Figure 1.3a: Scatterplot of FFS and SS

Answer

Positive Moderate Relationship between social support and self-esteem, $r = .439$, $n = 246$

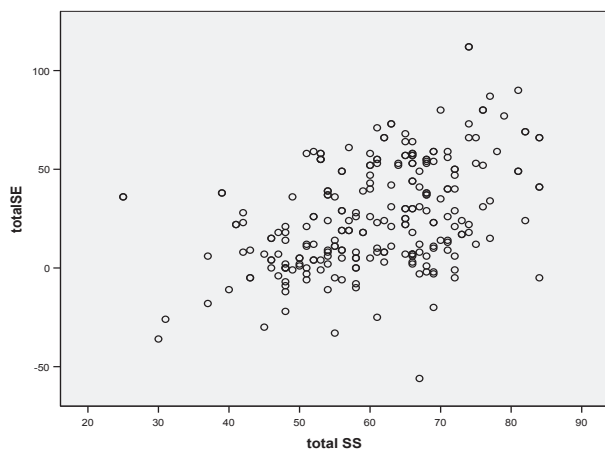


Figure 1.3b: Scatterplot of SE and SS

Answer

Zero Relationship between family functioning and self-esteem, $r = .117$, $n = 246$

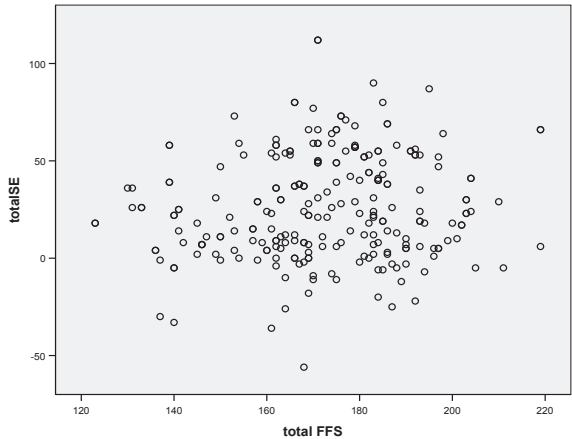


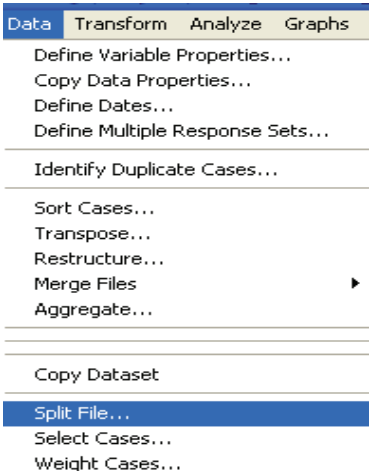
Figure 1.3c: Scatterplot of SE and FFS

1.10 Comparing the Correlation Coefficients for Two Groups

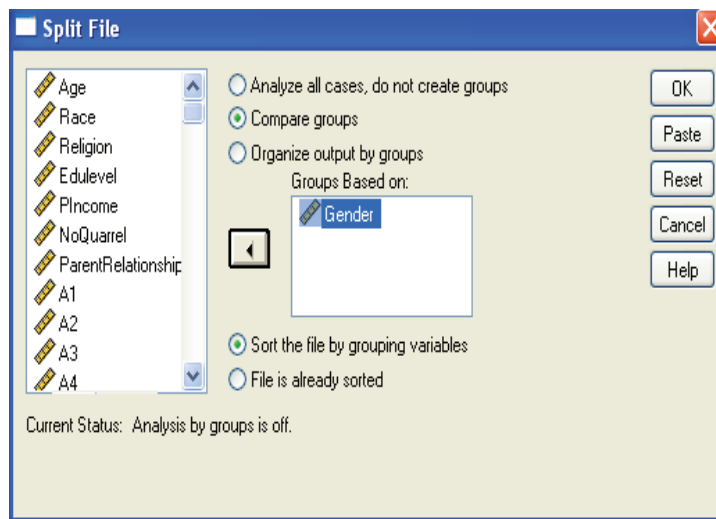
When doing correlational research, you may want to compare the strength of correlation coefficients for two separate groups. For example, you may want to look at the relationship between family functioning, social support and self-esteem for males and females separately. Below is the procedure for the comparing.

Split the sample (Use Data Set1 Family Functioning)

- Select **Data** from the main menu.
- Click **Split File** as shown in the following figure



- Click on **Compare Groups**
- Move the grouping variable (e.g gender) into the box labeled **Groups based on**.
- Click on **OK** as shown in the following figure



This will split the sample by gender and analyse the two groups separately. When you have finished analyzing males and females separately, you have to turn off the **Split File**. Then click on the button **Analyse all cases, do not create groups** and click on **OK**.

The output generated from family functioning data set is shown in Table 1.5

Table 1.5: Correlations for Male and Female Groups with SPSS

Gender			total FFS	total SS
Male	total FFS	Pearson Correlation	1	.440(**)
		Sig. (2-tailed)		.000
		N	123	123
	total SS	Pearson Correlation	.440(**)	1
		Sig. (2-tailed)	.000	
		N	123	123
Female	total FFS	Pearson Correlation	1	.282(**)
		Sig. (2-tailed)		.002
		N	123	123
	total SS	Pearson Correlation	.282(**)	1
		Sig. (2-tailed)	.002	
		N	123	123

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

1.11 Interpretation of Output from Correlation for Two Groups

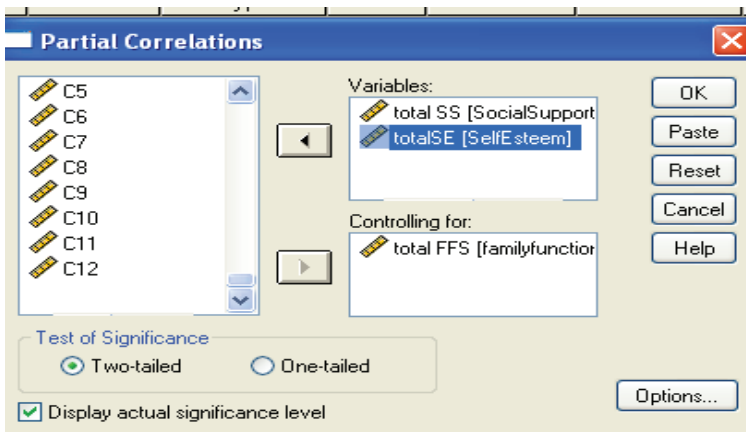
The correlation between family functioning and social support for male was .440, while for females it was slightly lower, $r=.282$. It is important to note that this process is different from testing the statistical significance of the correlation coefficients reported in the output table above. The significance levels reported for males: Sig. = .000; for females: Sig. = .000 provide a test of the null hypothesis that the correlation coefficient in the population is 0. However, assesses the probability that the difference in the correlations observed for the two groups (males and females) would occur as a function of a sampling error. In fact, there was no real difference in the strength of the relationship for males and females.

1.12 Partial Correlation

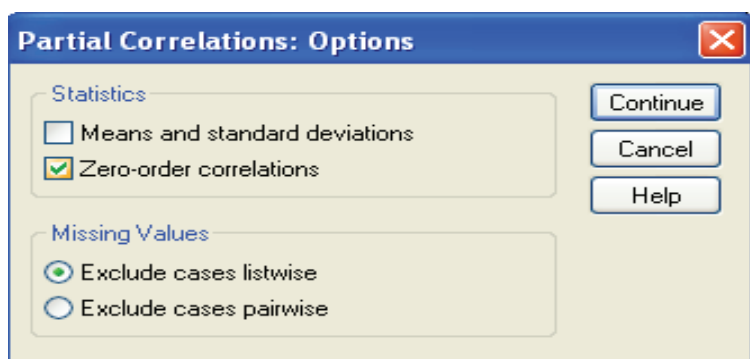
Partial correlation is similar to Pearson product-moment correlation. However, it allows you to control for an additional variable. This additional variable might be influencing your two variables by interest. By removing the confounding variable, you can get a more accurate indication of the relationship between your two variables. Use Data Set1 Family Functioning.

1.12.1 Procedure for Partial Correlation

- Select **Analyze** the menu
- Click on **Correlate**, then on **Partial**
- Move two continuous variables (e.g Social support and Self-esteem) that you want to correlate into the **variables** box.
- Move the variable that you want to control into the **Controlling for** box (e.g Family functioning)
- Click on Options as shown in the following figure



- In the **Missing values** section, click on Exclude cases pairwise
- In the **Statistics** section, click on **Zero Order Correlations**
- Click on **Continue** and then **OK** as shown in the following figure



1.12.2 Example of Partial Correlation

Use family functioning data set and demonstrate the generated output as shown in Table 1.6

Table 1.6: Partial Correlations with SPSS

Control Variables			total SS	totalSE	total FFS
-none-(a)	total SS	Correlation	1.000	.439	.377
		Significance (2-tailed)	.	.000	.000
		Df	0	244	244
	totalSE	Correlation	.439	1.000	.117
		Significance (2-tailed)	.000	.	.066
		Df	244	0	244
	total FFS	Correlation	.377	.117	1.000
		Significance (2-tailed)	.000	.066	.
		Df	244	244	0
total FFS	total SS	Correlation	1.000	.429	
		Significance (2-tailed)	.	.000	
		Df	0	243	
	totalSE	Correlation	.429	1.000	
		Significance (2-tailed)	.000	.	
		Df	243	0	

Cells contain zero-order (Pearson) correlations.

1.12.3 Interpretation of Partial Correlation

In the top half of the table, the word “none” indicates that no control variable is in operation. In this case the $r = .439$

The bottom half of the table repeats the same set of correlation analyses but controlling or taking out the effects of the control variable (e.g. family functioning). The $r = .429$.

The bottom half of the table repeats the same set of correlation analyses but controlling or taking out the effects of the control variable (e.g. family functioning). The $r = .429$.

By comparing the two sets of correlation coefficients, you will be able to see whether controlling the additional variable had any impact on the relationship between the two variables.

In this case, there was only a small decrease in the strength of the correlation (from .439 to .429). This suggests that the observed relationship between social support and self-esteem is not due merely to the influence of family functioning.

1.12.4 Presenting the Result

Partial Correlation was used to explore the relationship between social support and self-esteem while controlling for scores on the family functioning. There was a moderate, positive, partial correlation between social support and self-esteem ($r = .429, p < .01$). This indicated that high level of self-esteem was associated with an increase level of social support. An inspection of the zero order correlation ($r = .439$) suggested that controlling for family functioning had very little effect on the strength of the relationship between these two variables.

Worked Examples

Worked Example 1

The following is an example of a previous study done which indicates a correlation between two variables:-

Researchers at the Malaysian Centre for Road Safety Testing are trying to find out how the age of cars affects their braking capability. They test a group of ten cars of differing ages and find out the minimum stopping distances that the cars can achieve. The results are set out in the table below:

Table Worked Example 1(a): Car ages and stopping distances

Car	Age (months)	Minimum Stopping at 40 kph (metres)
A	9	28.4
B	15	29.3
C	24	37.6
D	30	36.2
E	38	36.5
F	46	35.3
G	53	36.2
H	60	44.1
I	64	44.8
J	76	47.2

The results show a reasonably strong positive correlation – the older the car, the longer the stopping distance.

The calculation needed for us to work out the correlation coefficient is as follows.

Note that in the Table Worked Example 1(b) below,

x = age of car

y = stopping distance

Table Worked Example 1(b): Calculation for Car ages and stopping distances

Car	X	Y	X^2	Y^2	XY
A	9	28.4	81	806.56	255.6
B	15	29.3	225	858.49	439.5
C	24	37.6	576	1413.76	902.4
D	30	36.2	900	1310.44	1086
E	38	36.5	1444	1332.25	1387
F	46	35.3	2116	1246.09	1623.8
G	53	36.2	2809	1310.44	1918.6
H	60	44.1	3600	1944.81	2646
I	64	44.8	4096	2007.04	2867.2
J	76	47.2	5776	2227.84	3587.2
Totals	415	375.6	21623	14457.72	16713.3

The formula to calculate the correlation coefficient is:

$$\text{Calculated } r: r_{cal} = \frac{(n \cdot \sum XY) - (\sum X \cdot \sum Y)}{\sqrt{[(n \cdot \sum X^2) - (\sum X)^2] \times [(n \cdot \sum Y^2) - (\sum Y)^2]}}$$

Substituting the values from the table gives:

$$r = 10 \times 16713.3 - 415 \times 375.6 / \sqrt{[(10 \times 21623 - 415^2) (10 \times 14457.72 - 375.6^2)]}$$

$$r = 11259 / \sqrt{(44005 \times 3501.84)}$$

$$r = 11259 / 124.14$$

$$r = 0.91$$

In this case of car ages and stopping distance performance, we can conclude that there is a strong correlation between the two variables.

It is also important to understand that correlation does not equal to causation. In correlation, one cannot conclude that one variable causes the other variable, simply because there is a relationship. For example, one cannot say that going on a diet is the cause of losing weight. Correlation only tells us that there is a relationship between the two variables. However, it is not the cause, whereby one variable causes the other variable.

Worked Example 2

In the study of the relationship between the grip strength and levels of extroversion, a group of participants were introduced and shook hands with a friendly research team. Members of the research team were trained to independently score grip strength using a one to nine scale with one indicates very weak grip and nine indicates very strong grip. Participants were then asked to complete a personality inventory that included an introversion or extroversion scale. In this study, level of extroversion is the independent variable whereas grip strength is the dependent variable.

The data collected for this study is shown below:

Table Worked Example 2: Relationship between Grip Strength and Levels of Extroversion

Student	Level of extroversion, X	Grip Strength, Y	X^2	Y^2	XY
Jamie	19	4	361	16	76
Kelvin	36	7	1296	49	252
Harrison	27	9	729	81	243
Bobby	11	3	121	9	33
Akane	27	6	729	36	162
Gloria	15	2	225	4	30
	$\Sigma X = 135$	$\Sigma Y = 31$	$\Sigma X^2 = 3461$	$\Sigma Y^2 = 195$	$\Sigma XY = 796$
	$\Sigma (X)^2 = 18225$	$\Sigma (Y)^2 = 961$			

$$\begin{aligned}
 \text{Calculated } r &= (6 \times 796) - (135 \times 31) / \sqrt{[(6 \times 3461) - (18225)] \times [(6 \times 195) - (961)]} \\
 &= (4776 - 4185) / \sqrt{[(20766 - 18225) \times (1170 - 961)]} \\
 &= (591) / \sqrt{(2541 \times 209)} \\
 &= (591) / \sqrt{531069} \\
 &= 591 / 728.7448 \\
 &= 0.811
 \end{aligned}$$

$$df = n - 2$$

$$= 6 - 2$$

$$= 4$$

Critical $r = 0.811$ (from the table of correlation)

Conclusion

Calculated $r =$ critical r , Null hypothesis accepted. There is no significant relationship between the levels of extroversion with the grip of strength. Therefore the research hypothesis is rejected.

Worked Example 3

A researcher hypothesizes that body height affects self-esteem of males. The information on body height and level of self-esteem are collected from 10 males. Body height is measured in centimeters. Self-esteem is measured using 1-5 rating items where higher score indicates higher level of self-esteem.

Table Worked Example 3(a): Relationship between Height and Self-esteem

Subject	Height (X)	Self-esteem (Y)
1	168	4.1
2	171	4.6
3	171	3.8
4	165	4.1
5	163	4.0
6	158	3.2
7	162	3.3
8	172	3.4
9	181	4.3
10	169	3.7

- Calculate ΣXY , ΣX , ΣY , ΣX^2 , and ΣY^2 .
- Compute the correlation coefficient and critical r .
- Write brief interpretation for the result at $p = 0.05$.

Answer:

$$(a) \quad r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[n(\Sigma X^2) - (\Sigma X)^2][n(\Sigma Y^2) - (\Sigma Y)^2]}}$$

Where:

n = sample size

ΣXY = sum of the products of X and Y

ΣX = sum of X

ΣY = sum of Y

ΣX^2 = sum of squared X

ΣY^2 = sum of squared Y

Table Worked Example 3(b): Calculation for Height and Self-esteem

Subject	Height (X)	X ²	Self-esteem (Y)	Y ²	XY
1	168	28224	4.1	16.81	688.8
2	171	29241	4.6	21.16	786.6
3	171	29241	3.8	14.44	649.8
4	165	27225	4.1	16.81	676.5
5	163	26569	4.0	16.00	652.0
6	158	24964	3.2	10.24	505.6
7	162	26244	3.3	10.89	534.6
8	172	29584	3.4	11.56	584.8
9	181	32761	4.3	18.49	778.3
10	169	28561	3.7	13.69	625.3
Sum	1680	282614	38.5	150.09	6482.3

(b)

Applying the formula,

$$r_{calc} = \frac{(10)(6482.3) - (1680)(38.5)}{\sqrt{[(10)(282614) - (1680)^2] [(10)(150.09) - (38.5)^2]}}$$

$$r_{calc} = 143 / \sqrt{(3740)(18.65)}$$

$$r_{calc} = 0.54$$

$$df = n - 2$$

$$= 10 - 2$$

$$= 8$$

At $p = .05$, $df = 8$,

$$r_{cri} = 0.632 \quad (\text{from table of correlation})$$

(c)

Based on the calculation, there is no significant correlation between body height and self-esteem of males. Since r_{calc} is smaller than r_{cri} . Thus, null hypothesis is accepted.

Worked Example 4

Refer to Data Set 2 Coffee Images for the following exercise.

Information on the different brands of coffee, frequency of drinking coffee (in a month), and the reasons why people drink coffee is shown in Table Worked Example 4(a).

- 4.1 A particular company has an interest to find out the important reasons as to why people drink coffee. You may use table of frequency to determine some of the significant reasons.

Table Worked Example 4(a): Frequency Table for Reasons of Drinking Coffee

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	I got addicted	925	19.4	19.4	19.4
	I enjoy drinking coffee	985	20.6	20.6	40.0
	The taste is nice	765	16.0	16.0	56.0
	I can't survive without coffee	500	10.5	10.5	66.5
	I am inspired after drinking coffee	226	4.7	4.7	71.2
	Coffee helps to reduce my anxiety	755	15.8	15.8	87.0
	I am more energetic with coffee	620	13.0	13.0	100.0
	Total	4776	100.0	100.0	

Answer:

From the table of frequency, the strongest reason as to drink coffee is “I enjoy drinking coffee” ($f = 985$) which contributes 20.6% of total percentage. The second reason as to why people drink coffee is “I got addicted” ($f = 925$) which contributes 19.4% of total percentage. Whereas the third reason for drinking coffee among the respondents is “the taste of coffee is nice” ($f = 765$), which equals to 16% of total contribution.

- 4.2 Name three of the popular brands of coffee among the respondents?

Table Worked Example 4(b): Frequency Table for Brands of Different Coffee

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NestCoffee	1195	25.0	25.0	25.0
	Blue Mountain	603	12.6	12.6	37.6
	Mucha	801	16.8	16.8	54.4
	Inspiration	753	15.8	15.8	70.2
	Hoollay	709	14.8	14.8	85.0
	Sweetie	715	15.0	15.0	100.0
	Total	4776	100.0	100.0	

Answer:

The most popular brand is NestCoffee, followed by Blue Mountain and Mucha.

4.3 Use pie chart to demonstrate the frequency distribution of different coffee brands.

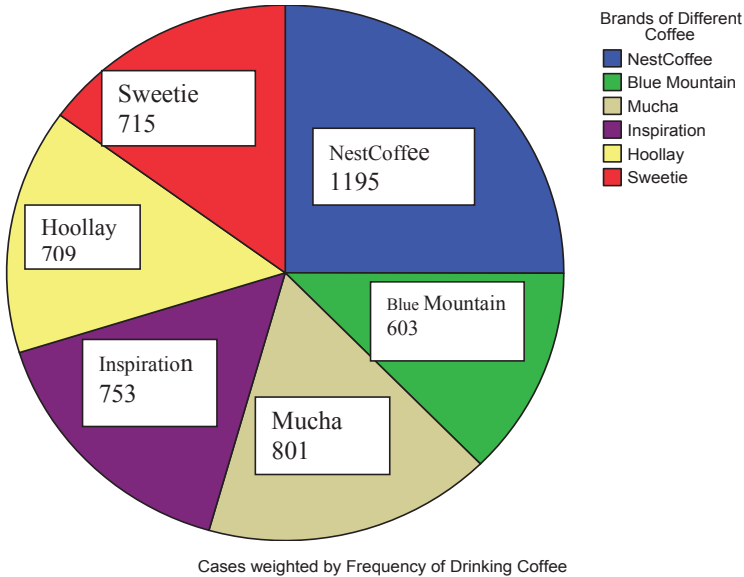


Figure 4: Pie Chart (Frequency of Drinking Coffee)

4.4 Is there a relationship between the reasons of drinking coffee and the number of hours sleep?

Table Worked Example 4(c): Correlations with SPSS

		Frequency of Drinking Coffee	Number of Hours Sleep per day
Frequency of Drinking Coffee	Pearson Correlation	1	-.537(**)
	Sig. (2-tailed)		.000
	N	4776	4776
Number of Hours Sleep per day	Pearson Correlation	-.537(**)	1
	Sig. (2-tailed)	.000	
	N	4776	4776

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

Answer:

There is a significant negative relationship between the frequency of drinking coffee and number of hours sleep ($r = -.537$, $p < .01$). Hence, when there is an increase in the frequency of coffee drank, the number of hours slept per day decreases.

Worked Example 5

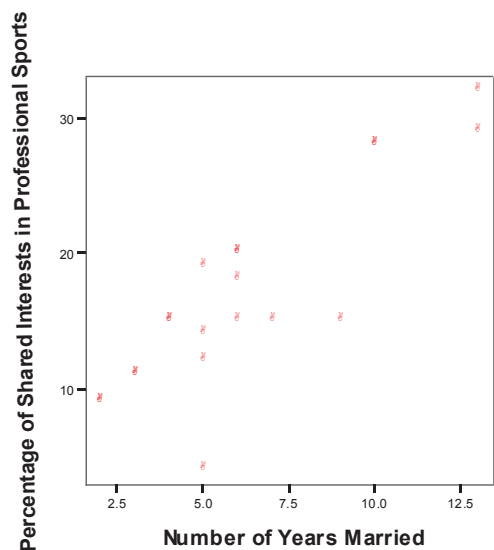
A psychologist, interested in marital relationships, hypothesized a positive relationship between the percentage of a couple's shared interests in professional sports and the number of years they are married. Use the data collected from the pilot study to answer the questions listed below.

Table Worked Example 5(a): Relationships of Number of Years Married and Percentage of Shared Interests

Participant	(X) Number of Years Married	(Y) Percentage of Shared Interests in Professional Sports
Couple 1	2	9
Couple 2	6	20
Couple 3	4	15
Couple 4	6	15
Couple 5	5	19
Couple 6	3	11
Couple 7	7	15
Couple 8	13	32
Couple 9	10	28
Couple 10	5	12
Couple 11	2	9
Couple 12	6	20
Couple 13	4	15
Couple 14	6	18
Couple 15	5	14
Couple 16	3	11
Couple 17	9	15
Couple 18	13	29
Couple 19	10	28
Couple 20	5	4

- Do data entry for the above dataset.
- Draw a scatterplot of these data.
- Briefly describe the findings. Was the researcher's hypothesis supported?
- Calculate the r value manually. Briefly interpret the result.

Answer:
b)



Scatter plot of Percentage of Shared Interests and Number of Years Married

c) Briefly describe the findings. Was the researcher's hypothesis supported?

Table Worked Example 5(b): Correlations with SPSS

		Number of Years Married	Percentage of Shared Interests in Professional Sports
Number of Years Married	Pearson Correlation	1	.859(**)
	Sig. (2-tailed)		.000
	N	20	20
Percentage of Shared Interests in Professional Sports	Pearson Correlation	.859(**)	1
	Sig. (2-tailed)	.000	
	N	20	20

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

There is a very significant strong positive relationship between the percentage of a couple's share interest in professional sports and the number of years they are married ($r = 0.859$, $p < 0.01$). Therefore, the longer the couples are married, the higher the percentage of shared interests in professional sports.

d)

Table Worked Example 5(c): Calculation for Number of Years Married and Percentage of Shared Interests

Participant	Number of Years Married (X)	Percentage of Shared Interests in Professional Sports (Y)	XY	Y^2	X^2
1	2	9	18	81	4
2	6	20	120	400	36
3	4	15	60	225	16
4	6	15	90	225	36
5	5	19	95	361	25
6	3	11	33	121	9
7	7	15	105	225	49
8	13	32	416	1024	169
9	10	28	280	784	100
10	5	12	60	144	25
11	2	9	18	81	4
12	6	20	120	400	36
13	4	15	60	225	16
14	6	18	108	324	36
15	5	14	70	196	25
16	3	11	33	121	9
17	9	15	135	225	81
18	13	29	377	841	169
19	10	28	280	784	100
20	5	4	20	16	25
	$\Sigma X = 124$	$\Sigma Y = 339$	$\Sigma XY = 2498$	$\Sigma Y^2 = 6803$	$\Sigma X^2 = 970$

$$e) r = \frac{(n \cdot \Sigma XY) - (\Sigma X \cdot \Sigma Y)}{\sqrt{[(n \cdot \Sigma X^2) - (\Sigma X)^2] \cdot [(n \cdot \Sigma Y^2) - (\Sigma Y)^2]}}$$

$$= \frac{(20 \cdot 2498) - (124 \cdot 339)}{\sqrt{[20 \cdot 970 - (124)^2] \cdot [20 \cdot 6803 - (339)^2]}}$$

$$= 0.859$$

$$df = 20 - 2 = 18$$

$$r_{cri} = 0.444$$

$r_{cal} > r_{cri}$, reject null hypothesis

There is a significant positive relationship between the percentage of a couple's shared interests in professional sports and the number of years they are married. Therefore, the longer the couples are married, the higher the percentage of shared interests in professional sports.

Chapter 1 Review Exercises

Question 1

The table below shows the incomes, X , and entertainment expenditures, Y (in hundreds of ringgits) of 10 sales engineers.

Incomes, X	28	40	45	60	35	55	32	72	30	70
Entertainment Expenses, Y	9	7	15	10	8	12	5	23	8	10

- Plot a scatter diagram.
- Find the value of the correlation coefficient for these data.
- Discuss the relationship between X and Y .

Question 2

The data for age and systolic blood pressure of eight randomly selected subjects are shown in the following table.

Subject	Age, X	Pressure, Y
A	34	110
B	45	126
C	50	125
D	58	135
E	63	145
F	72	150
G	75	152
H	38	120

- Construct a scatter plot for these data.
- Compute the value of the correlation coefficient for these data.
- Discuss the relationship between X and Y .

Question 3

A vehicle manufacturing company wants to investigate how the price of one of the motorcycle models depreciates with its age. The research department of the company took a sample of ten motorcycles of this model and collected the following information on the ages (in years) and prices (in thousands of RM) of these motorcycles.

Age	5	3	2	9	1	4	7	6	8	10
Price	2.2	2.5	3.5	1.0	4.0	2.4	1.5	1.8	1.2	0.7

- Plot a scatter diagram and determine whether the two variables have positive, negative or zero linear relationship.
- Confirm your answer for the above by computing the correlation coefficient.

Question 4

Modern warehouses employ computerized and automated guided vehicles to handle materials. As a result, the physical layout of the warehouse must be carefully designed to prevent vehicle congestion and also to optimize the response time. The data are shown in the following table. Of interest to the researcher is the relationship between congestion time, Y and number of vehicles, X .

Number of vehicles	Congestion time (s)
3	38
4	40
5	45
6	75
7	75
8	80
9	110
10	120
11	125
12	130

- Construct a scatter diagram for the data.
- Does the diagram exhibit a linear relationship between x and y ?
- Next, calculate the linear correlation coefficient for the data. Interpret your answer.

Question 5

A research was done by a lecturer to compare between the average number of hours student spent in study everyday, X and the average number of hours they sleep everyday, Y . The data was recorded as follows:

X	2	2	6.6	4	6	1	2.5	1
Y	10	9	9	12	8	7	5	6

- Plot a scatter diagram.
- Find the linear correlation coefficient for X and Y . Interpret your answer.

Question 6

Pressure, X	Compression, Y
1.5	1.5
3.0	1.5
4.5	3.0
6.0	3.0
7.5	5.0

- Draw a scatter diagram for the data above.
- Does the diagram exhibit a linear relationship between pressure and compression?
- Next, calculate the linear correlation coefficient for the data. Interpret your answer.

Question 7

The following data were collected to see whether a relationship exists between the temperature and the number of accidents in a highway from 12.00 p.m. to 5.00p.m. for a week.

Temperature °C	No. of Accident
28	1
27	0
33	0
31	4
26	3
34	2
30	1

- Plot a scatter diagram.
- Find the value of the correlation coefficient for these data.
- Discuss the relationship between X and Y .

Question 8

A study is done to see whether there is a relationship between the average number of hours the students spent for online games per week and their grade point average (GPA).

Hour (X)	30	12	19	10	35	12	30	10	8
GPA (Y)	2.2	3.5	3.0	4.0	1.8	3.2	2.4	3.8	3.7

- Plot a scatter diagram.
- Find the linear correlation coefficient for x and y . Interpret your answer.

Question 9

A researcher wants to investigate the relationship between the sizes and prices of houses in a town. The researcher collected the following information on the sizes (in thousands of square meters) of seven houses and the buying prices (in thousands of ringgits) paid by the house owner.

Sizes of houses ('000 m^2)	1.8	2.2	4.5	3.0	3.2	3.5	2.3
Buying Prices ('000 ringgits)	150	180	350	250	270	300	200

- Plot a scatter diagram.
- Find the linear correlation coefficient for the variables. Interpret your answer.

Question 10

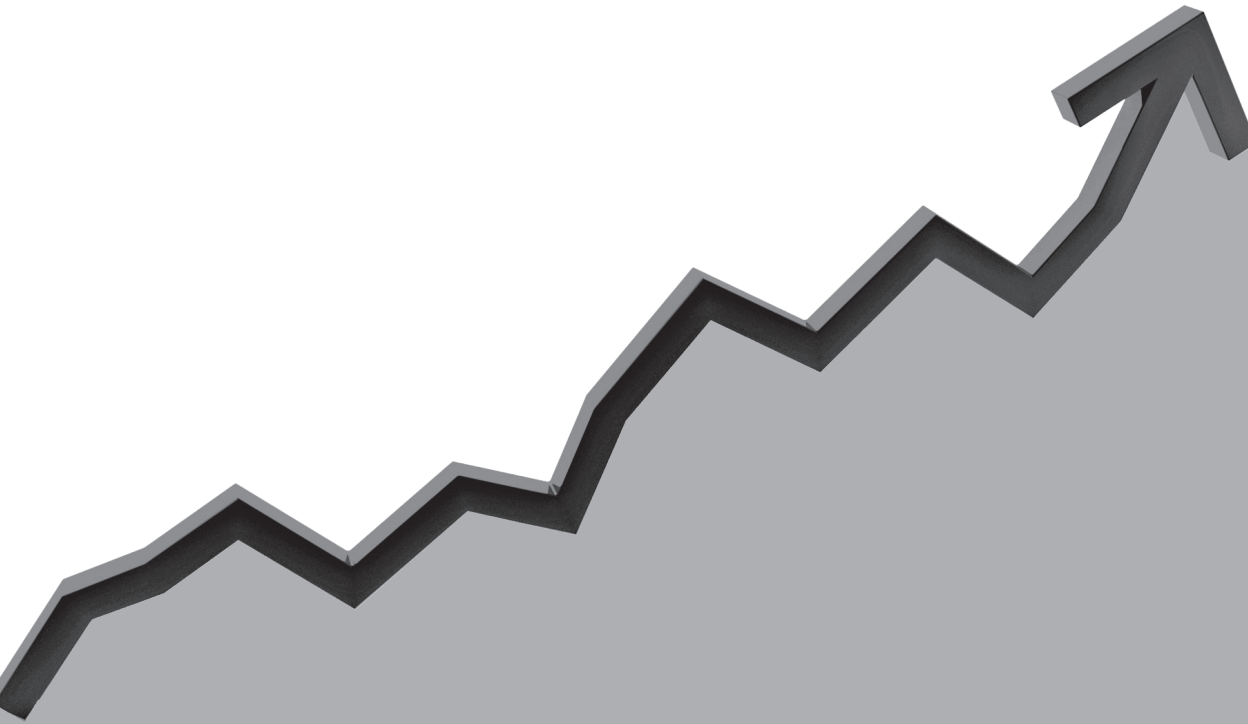
The data below were obtained in a study on the number of hours ten people exercise each week and the amount of milo (in *ml*) each person consumes per week.

Hours, x	2	0	3	6	8	5	10	3	4	7
Amount, y	450	50	330	640	100	320	550	600	500	150

- Plot a scatter diagram.
- Find the value of the correlation coefficient for these data.
- Discuss the relationship between X and Y .

chapter two

REGRESSION



Learning Objectives

At the completion of the chapter, students would be able to:

- List the characteristics of regression line.
- Find the relationship between two variables using the regression equation.
- Differentiate the characteristics of bivariate regression and multiple regression
- Explain the assumptions of multiple regressions.
- Apply the procedure for generating hierarchical multiple regression.
- Interpret the result of hierarchical multiple regression with the display of model summary.

2.0 Introduction

The second chapter, Regression analysis is an extension of correlation. The aim of the discussion of exercises is to enhance students' capability to assess the effect of one variable x on another variable y which is known as the bivariate regression. Meanwhile students will also be exposed to the discussion of multiple regressions; that is the effect of several variables.

Regression statistical techniques attempt to investigate the best way of describing the relationship between a dependent variable and one or more independent variables. The regression line represents the best fit straight line through a set of coordinates X and Y .

Generally independent variable is also known as predictor variable will be assigned as X variable, whereas dependent variable will be assigned as Y . When explaining the relationship between X and Y , it is often said as X predicting Y . The mathematical formula is as follows:

$$Y = a + bX$$

Where

Y = predicted value for dependent variable Y

a = value of Y intercept (point cut at Y axis)

b = regression coefficient (gradient of the line)

X = value for independent or predictor variable X

2.1 Regression Line

Regression is used to make predictions based on linear relationship. It is a statistical technique for finding the best-fitting straight line for a set of data. The resulting straight line is called regression line. An example of a regression line is shown in Figure 2.1 (Medcalc, 2009).

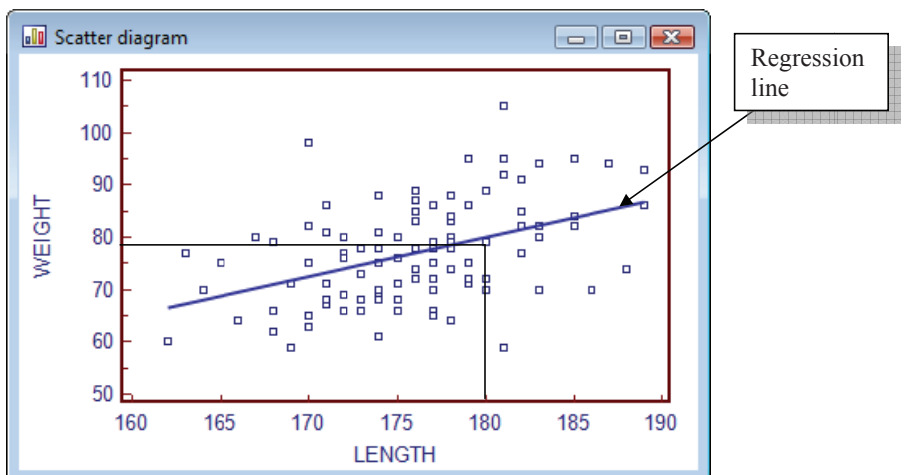


Figure 2.1: Regression Line (Medcalc, 2009)

2.2 Characteristics of Regression Line

Based on Figure 2.1, the following characteristics of the regression line can be observed:

- The line makes the relationship between length and weight easier to see.
- The line identifies the center, or central tendency, of the relationship, just as the mean describes central tendency for a set of scores. Therefore, regression line provides a simplified description of the relationship. Even when all the data points on the scatterplot were removed, the regression line will still give a general picture of the relationship between length and weight.
- The regression line can be used to make prediction. According to Gravetter (2005), the line establishes a precise, one-to-one relationship between each X value (length) and a corresponding Y value (weight). For example, in the scatterplot above, when $X = 180$, Y can be predicted as 79 with the presence of the regression line.

However, regression line does not have to be drawn on a graph; it can be presented in a simple equation $Y = bX + a$, where b and a are fixed constants (Gravetter & Wallnau, 2005).

2.3 Types of Regression

There are 2 types of regression analysis, Bivariate Regression and Multiple Regression. Bivariate Regression involves analyzing the relationship between the dependent variable and one independent variable. Multiple Regression involves the relationship between dependent variable and more than one independent variable (such as X_1, X_2, X_3 , etc).

2.3.1 Bivariate Regression

Bivariate Regression is used to examine the relationship between two variables (X) and (Y). The results indicated in the study of regression are then used to make predictions. In other words, we can use regression for a study when we have knowledge of one variable, while trying to predict the other.

For discussion purposes, let's take example of a study to find out whether the increase in sales of a product is due to the recent advertising on radio. In order to predict the result of this study, regression is an appropriate approach to use. In this situation, we do have knowledge of one variable – sales of the product have increased. However, what we do not know is the reasons behind the increase in sales. As such, we can test and predict if the increase of sales was due to the recent radio advertising. Nevertheless, the results may also show otherwise.

The two variables mentioned earlier (' X ' and ' Y '), one represents an independent variable while the other is an independent variable. An independent variable is a factor which is selected by the researcher of which he/she has control of. A dependent variable is the result of which the researcher wants to find.

As mentioned earlier, the formula for linear regression is $Y = a + bX$, whereby X is the independent variable and Y is the dependent variable. The slope of the line is b , and a is the point where X and Y intercept (the value of Y when $X = 0$).

Example:

The director of admissions of a small college administered a newly designed entrance test to 20 students selected at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the entrance test score (X). The results of the study are as follow:

Table 2.1: Entrance Test Score and GPA

Entrance Test Score (X)	GPA (Y)	XY	X^2
5.50	3.10	17.05	30.25
4.80	2.30	11.04	23.04
4.70	3.00	14.10	22.09
3.90	1.90	7.41	15.21
4.50	2.50	11.25	20.25
6.20	3.70	22.94	38.44
6.00	3.40	20.40	36.00
5.20	2.60	13.52	27.04
4.70	2.80	13.16	22.09
4.30	1.60	6.88	18.49
4.90	2.00	9.80	24.01
5.40	2.90	15.66	29.16
5.00	2.30	11.50	25.00
6.30	3.20	20.16	39.69
4.60	1.80	8.28	21.16
4.30	1.40	6.02	18.49
5.00	2.00	10.00	25.00
5.90	3.80	22.42	34.81
4.10	2.20	9.02	16.81
4.70	1.50	7.05	22.09
$\Sigma X=100.00$	$\Sigma Y=50.00$	$\Sigma XY=257.66$	$\Sigma X^2=509.12$

$$\begin{aligned}
 \text{Slope (b)} &= [N\Sigma XY - (\Sigma X)(\Sigma Y)] / [N\Sigma X^2 - (\Sigma X)^2] \\
 &= (5153.32 - 5000) / (10182.4 - 10000) \\
 &= 153.32 / 182.4 \\
 &= 0.84
 \end{aligned}$$

$$\begin{aligned}
 \text{Intercept (a)} &= (\Sigma Y - b(\Sigma X)) / N \\
 &= (50 - 0.84(100)) / 20 \\
 &= -34 / 20 \\
 &= -1.7
 \end{aligned}$$

$$\begin{aligned}
 \text{Regression Equation } Y &= a + bX \\
 Y &= -1.7 + 0.84X
 \end{aligned}$$

Find out the Predicted Grade Point Average (GPA) of a student if the entrance test score is 5.8.

$$\begin{aligned}
 Y &= -1.7 + 0.84X \\
 &= -1.7 + 0.84(5.8) \\
 &= -1.7 + 4.872 \\
 &= 3.172
 \end{aligned}$$

2.3.2 Multiple Regression

Multiple Regression is used to explore the relationship between one continuous dependent variable and a number of independent variables or predictors. It is based on correlation, but allows a more sophisticated exploration of the interrelationship among a set of variables. One shouldn't use multiple regressions as a fishing expedition. You must support your analysis with theoretical reason.

Mathematical formula for multiple regression is as follow:

$$Y = a + b_1X_1 + b_2X_2$$

Y = predicted value for dependent variable Y

a = value of Y intercept (point cut at Y axis)

b_1 = regression coefficient (gradient of the line) for the first independent variable

X_1 = value for first independent or predictor variable X_1

b_2 = regression coefficient (gradient of the line) for the second independent variable

X_2 = value for second independent or predictor variable X_2

Multiple regression can tell you how well a set of variables is able to predict a particular outcome. For example, you may be interested how well a set of subscales on predicting the academic performance among students. In addition, it will provide you information about the contribution of total subscales, and the relative contribution of each subscale. Main types of research questions that can be used:

- How well a set of variable is able to predict a particular outcome
- Which variable in a set of variables is the best predictor of an outcome

a) Assumptions of Multiple Regression

i) Sample Size

Sample size need to be big enough to be generalized to a bigger population. What should be the ideal sample size? Different authors will have different criteria. For example, Steven (1996), recommends 15 subjects per variable for social science research. Whereas Tabachnick and Fidell (2001) give a formula to calculate the size of sample, $N > 50 + 8m$ (where m = number of independent variables). If you have 3 independent variables, you will need 74 subjects in a particular study.

ii) Outliers

Outliers should be excluded in multiple regression. Extreme scores which are too high or too low should be deleted from both dependent and independent variables.

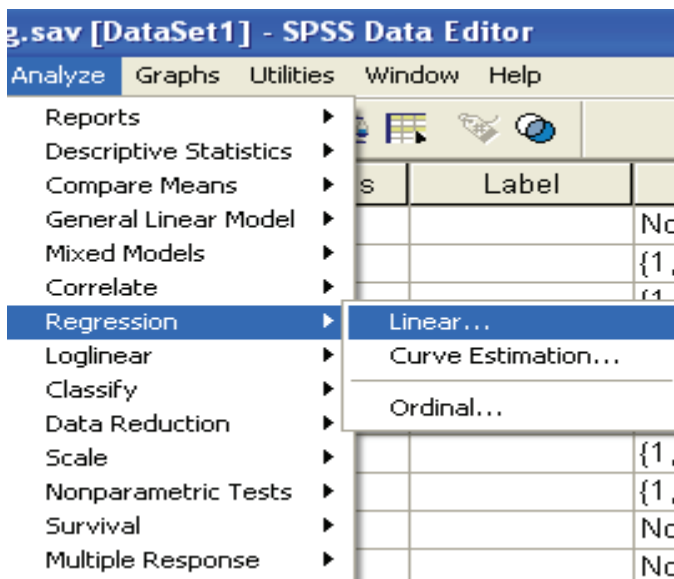
b) Multiple Regression analyses

- **Standard Multiple Regression:** All the independent (or predictor) variables are entered into the equation simultaneously. Each independent variable is evaluated in terms of its predictive power, over and above that offered by all the other independent variables.

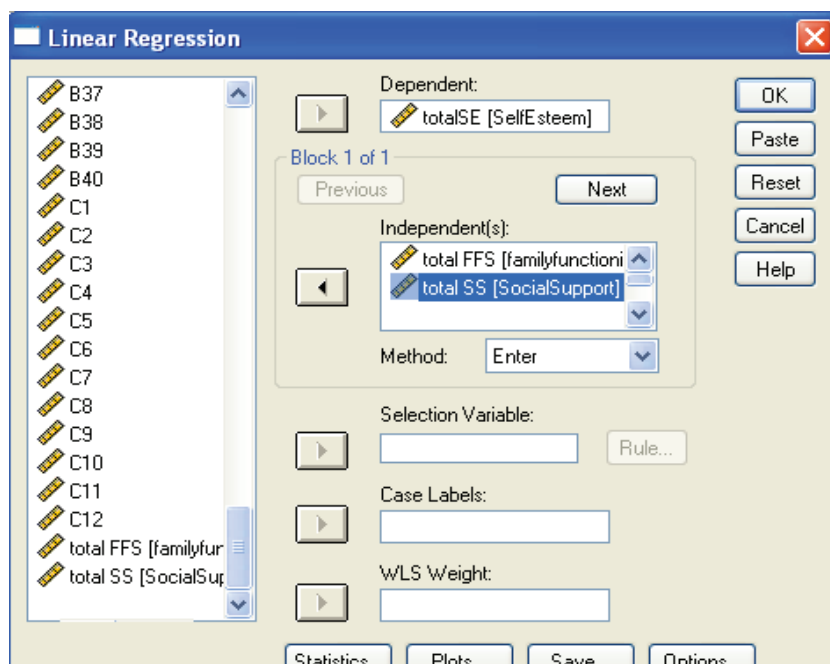
- Hierarchical multiple Regression: In hierarchical regression (also called sequential) the independent variables are entered into the equation in the order specified by the researcher based on theoretical grounds.
- Stepwise multiple regression: In stepwise regression the researcher provides SPSS with a list of independent variables and then allows the program to select which variables it will enter, and in which order they go into the equation, based on a set of statistical criteria.
- Multiple regression would be used if you had a set of variables (eg., various personality scales) and wanted to know how much variance in a dependent variable (eg., anxiety) they were able to explain as a group or block.
- Multiple regression approach would also tell you how much unique variance in the dependent variable that each of the independent variables explained.
- Hierarchical multiple regression would be used if you wanted to know how much a variable predicts another variable.
- Once all sets of variables are entered, the overall model is assessed in terms of its ability to predict the dependent measure. The relative contribution of each block of variables is also assessed.
- Stepwise multiple regression would be used when you have a large number of predictor variables.
- Stepwise multiple regression would be used when no underlying theory concerning their possible predictive power.

2.4 Procedure for Generating Multiple Regression

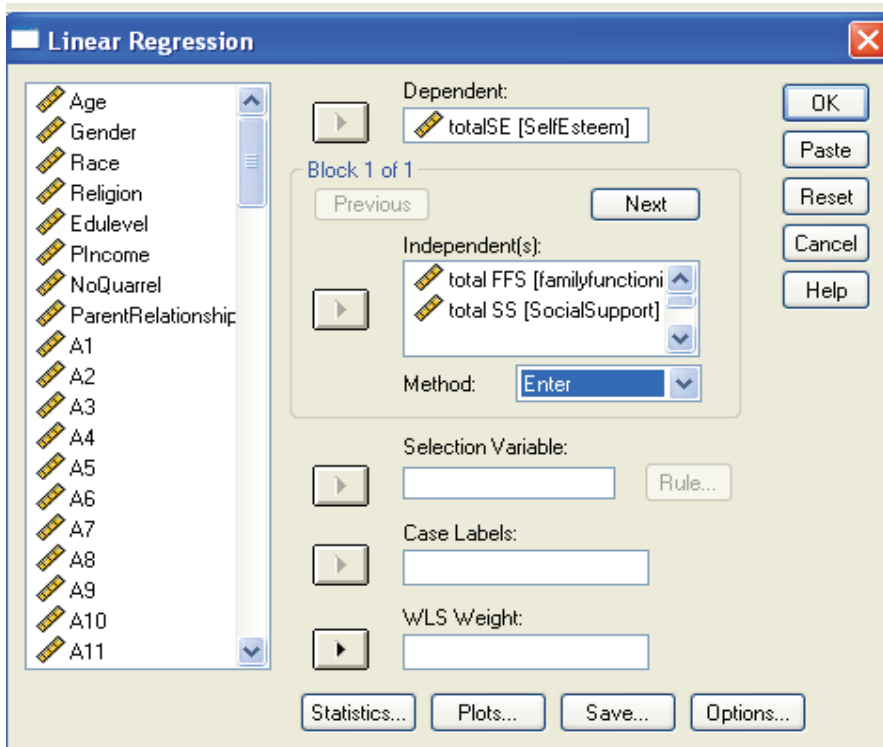
- Select **Analyze**, click on **Regression**, then on **Linear** as shown in the following figure



- Select dependent variable (e.g Self-esteem) and move it to **Dependent** box..
- Select independent variable (e.g Family Functioning and Social Support) and move them to **Independent** box as shown in the following figure



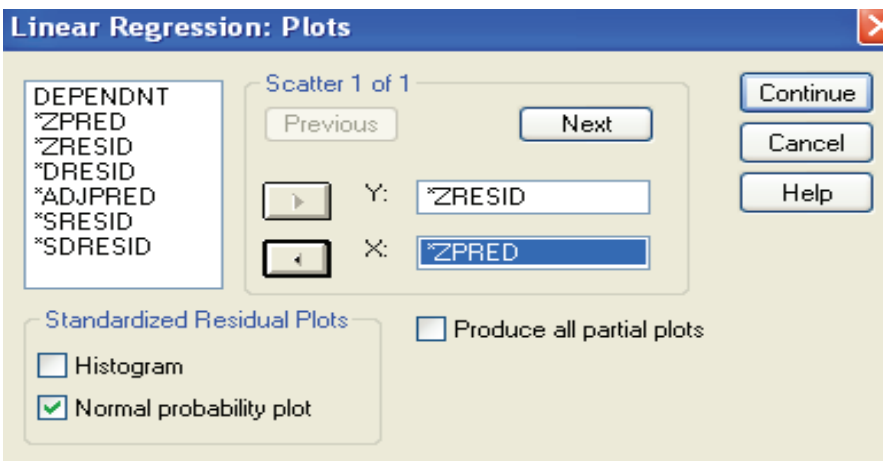
- For **Method**, select **Enter**.
- Click on the **Option** button, in the **Missing Values** section click on **Exclude cases pairwise** as shown in the following figure



The **Linear Regression** dialog box shows the following settings:

- Dependent:** totalSE [Self Esteem]
- Block 1 of 1**
 - Independent(s):** total FFS [familyfunctioni], total SS [SocialSupport]
 - Method:** Enter
- Selection Variable:** (empty)
- Case Labels:** (empty)
- WLS Weight:** (empty)
- Buttons:** Statistics..., Plots..., Save..., Options...

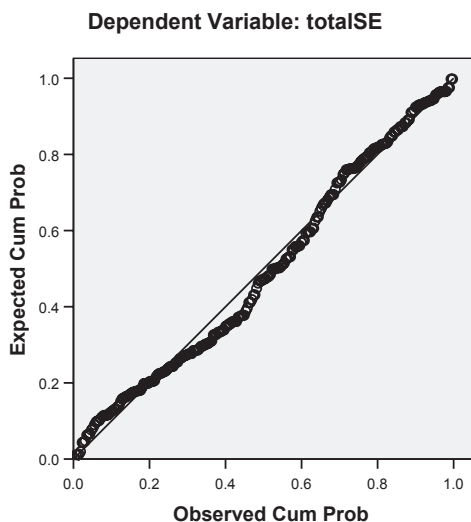
- Click on the **Plots** button.
- Click on **"ZRESID"** and move this into the **Y** box.
- Click on the **"ZPRED"** and move this into the **X** box.
- Select **Standardized Residual Plots**, tick the **Normal probability plot**.
- Click **Continue** as shown in the following figure



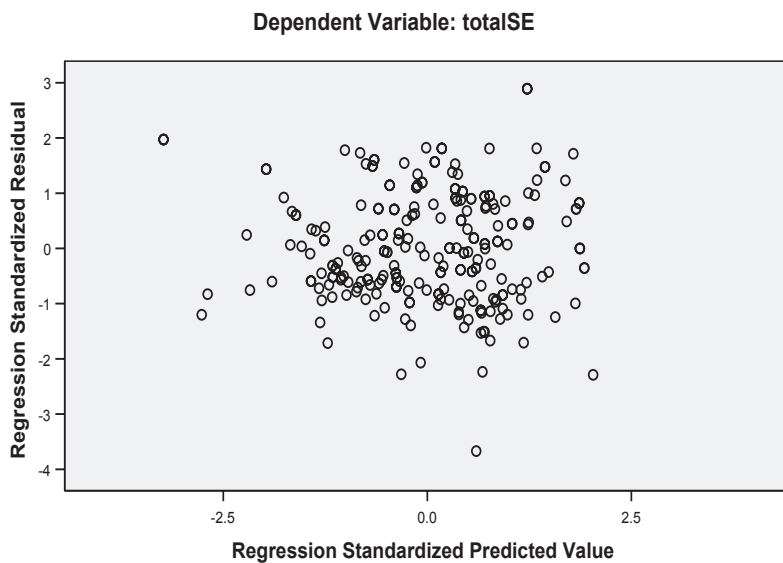
The **Linear Regression: Plots** dialog box shows the following settings:

- Scatter 1 of 1**
 - Y:** *ZRESID
 - X:** *ZPRED
- Standardized Residual Plots**
 - ☐ Histogram
 - ☒ Normal probability plot
- ☐ Produce all partial plots
- Buttons:** Continue, Cancel, Help

Normal P-P Plot of Regression Standardized Residual



Scatterplot



2.4.1 Outliers, Normality, Linearity and Independence of Residuals

- In the Normal Probability Plot you are hoping that your points will lie in a reasonably straight diagonal line from bottom left to top right.
- This would be no major deviations from normality.

- In the Scatterplot of the Standardised residuals, the residuals will be roughly rectangular distributed, with most of the scores in the centre.
- Tabachinick and Fidell (2001) define outliers as cases that have a standardized residual more than 3.3 or less than -3.3.
- If there are only a few outlying residuals in a large sample, it may not be necessary to take any action.

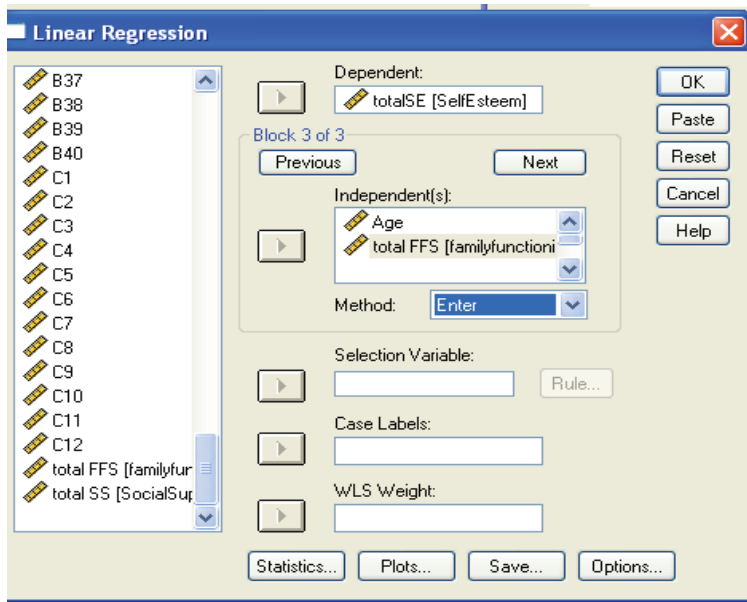
2.4.2 Hierarchical Multiple Regression

To illustrate the use of hierarchical multiple regression, we have to evaluate the effectiveness of the model. For example, after controlling for age and Total Family Functioning, will Total Perceived Social Support still able to predict a significant amount the variance in self-esteem? To answer this question, we need to use hierarchical multiple regression (also known as sequential regression). In the first block, we have to “force” age and Total Family Functioning responding into the analysis. This has the effect of statistically controlling for the variables.

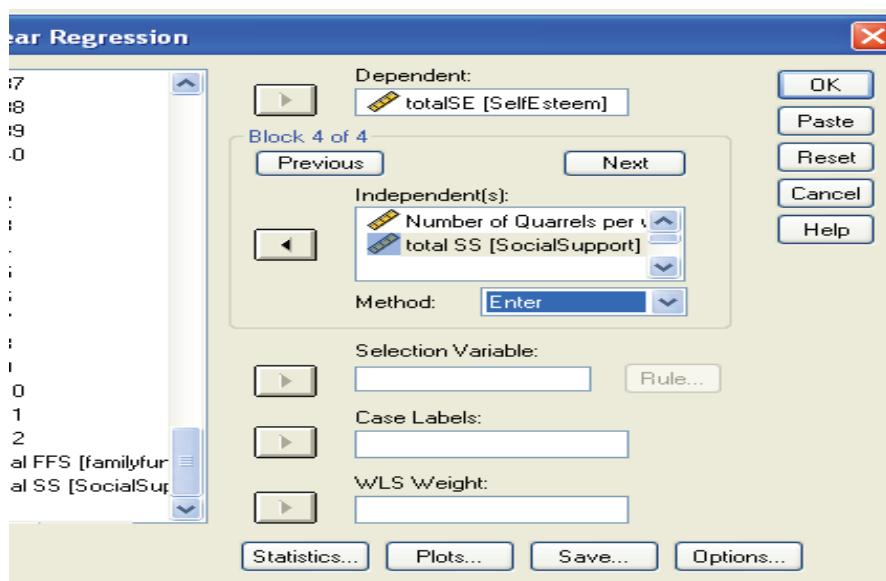
The next step we enter Total Perceived Social Support (independent) into the equation as a block where the possible effect of age and Total Family Functioning has been removed.

2.4.3 Procedure for Generating Hierarchical Multiple Regression

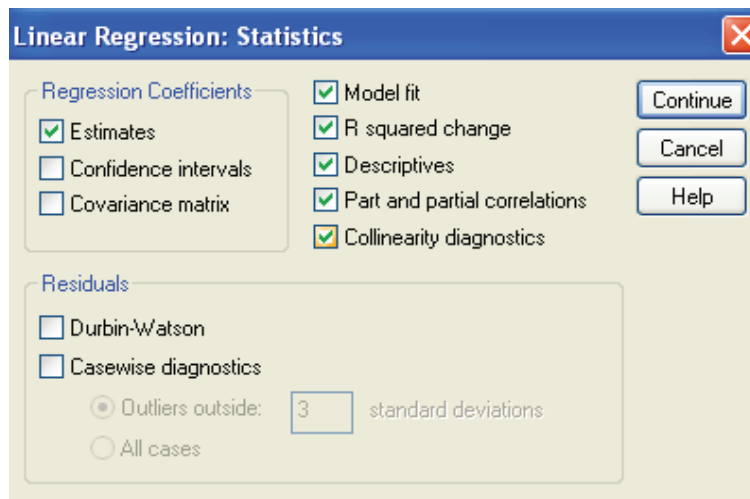
- Select **Analyze**, click on **Regression** and **Linear**.
- Select dependent variable (self-esteem) and move it into the **Dependent** box.
- Move Total Family Functioning and age (independent variable) that you wish to control into **Independent** box.
- Click on **Next** as shown in the following figure



- Select the next block of independent variable variables (Number of Quarrels and Total Perceived Social Support). Move the independent variables into the **Independent** box.
- Set default (**Enter**) in the **Method** box as shown in the following figure



- Click on **Statistic**. Tick **Estimates, Model fit, R squared change, Descriptive, Part and partial correlations and Collinearity diagnostics**.
- Click on **Continue**.
- Click on the **Options** button. Select **Missing Values** and **Exclude cases pair wise**.
- Click on **Save**.
- Select **Mahalanobis** and **Cook's**.
- Click on **Continue** and **OK** as shown in the following figure



2.4.4 Interpretation of Hierarchical Multiple Regression

Output of SPSS

Table 2.2: Model Summary

Model	R	R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	Df1	Df2	Sig F Change
1	.120(a)	.014	27.079	.014	1.786	2	243	.170
2	.456(b)	.208	24.375	.194	29.461	2	241	.000

a Predictors: (Constant), total FFS, Age

b Predictors: (Constant), total FFS, Age, Continuous data of Number of Quarrels per week, total SS

Table 2.3: ANOVA (c)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	2619.854	2	1309.927	1.786	.170(a)
	Residual	178191.105	243	733.297		
	Total	180810.959	245			
2	Regression	37627.040	4	9406.760	15.833	.000(b)
	Residual	143183.919	241	594.124		
	Total	180810.959	245			

a Predictors: (Constant), total FFS, Age

b Predictors: (Constant), total FFS, Age, Continuous data of Number of Quarrels per week, total SS

c Dependent Variable: totalSE

Table 2.4: Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	Zero-order	Partial
1	(Constant)	6.785	26.902		.252	.801
	Age	-.444	1.047	-.027	-.424	.672
	total FFS	.169	.091	.117	1.844	.066
2	(Constant)	-40.407	25.343		-1.594	.112
	Age	.809	.959	.049	.844	.400
	total FFS	-.080	.089	-.056	-.900	.369
	total SS	1.127	.151	.468	7.442	.000
	Continuous data of Number of Quarrels per week	-2.159	1.238	-.100	-1.744	.082

a Dependent Variable: totalSE

In the model summary, there are two models listed. Model **Regression**

Regression is used to make predictions based on linear relationship. It is a statistical technique for finding the best-fitting straight line for a set of data. The resulting straight line is called regression line.

Model 1 refers to the first block of independent variables (Age and Total Family Functioning), whereas Model 2 includes the second block of independent variables (Number of quarrels per week and Total Perceived Social Support).

To find out how much of this overall variance is explained by independent of interest (Number of quarrels per week and Total Perceived Social Support) after the effects of age and Total Family Functioning responding are removed, we need to look in the column labeled R Square change. R Square change value is .194. This means that Number of Quarrels per week and Total Perceived Social Support explain an additional 19.4 % of the variance in self-esteem, even when the effects of age and Total Family Functioning responding are controlled. The ANOVA table indicates that the model as a whole is significant [$F(4, 241) = 15.833, p < .01$].

Worked Examples

Worked Example 1

Find the Simple/Linear Regression of the following two variables, X and Y from Table Worked Example 1 (a).

Table Worked Example 1 (a): X and Y values

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

To find regression equation, we will first find slope, intercept and use the values to form the regression equation.

Step 1: Count the number of values.

$$N = 5$$

Step 2: Find XY , X^2 from Table Worked Example 1 (b).

Table Worked Example 1 (b): Calculation of X and Y values

X Values	Y Values	$X*Y$	$X*X$
60	3.1	$60 * 3.1 = 186$	$60 * 60 = 3600$
61	3.6	$61 * 3.6 = 219.6$	$61 * 61 = 3721$
62	3.8	$62 * 3.8 = 235.6$	$62 * 62 = 3844$
63	4	$63 * 4 = 252$	$63 * 63 = 3969$
65	4.1	$65 * 4.1 = 266.5$	$65 * 65 = 4225$

Step 3: Find ΣX , ΣY , ΣXY , ΣX^2 .

$$\Sigma X = 311$$

$$\Sigma Y = 18.6$$

$$\Sigma XY = 1159.7$$

$$\Sigma X^2 = 19359$$

Step 4: Substitute all values into the slope formula given below.

$$\begin{aligned}
 \text{Slope (b)} &= [\Sigma XY - (\Sigma X)(\Sigma Y)] / [\Sigma X^2 - (\Sigma X)^2] \\
 &= [(5)(1159.7) - (311)(18.6)] / [(5)(19359) - (311)^2] \\
 &= (5798.5 - 5784.6) / (96795 - 96721) \\
 &= 13.9 / 74 \\
 &= 0.19
 \end{aligned}$$

Step 5: Now, substitute values into the intercept formula given below.

$$\begin{aligned}
 \text{Intercept } (a) &= (\Sigma Y - b(\Sigma X)) / N \\
 &= (18.6 - 0.19(311)) / 5 \\
 &= (18.6 - 59.09) / 5 \\
 &= -40.49 / 5 \\
 &= -8.098
 \end{aligned}$$

Step 6: Then substitute the values of a and b into the regression equation formula.

$$\begin{aligned}
 \text{Regression Equation } Y &= a + bX \\
 &= -8.098 + 0.19X
 \end{aligned}$$

Suppose we want to know the approximate Y value for the variable $X = 64$. Then we can substitute the value into the equation.

$$\begin{aligned}
 \text{Regression Equation } Y &= a + bX \\
 &= -8.098 + 0.19(64) \\
 &= -8.098 + 12.16 \\
 &= 4.06
 \end{aligned}$$

This example will guide you to find the relationship between two variables by calculating the Regression Equation based on the above steps.

Worked Example 2

A financier whose specialty is investing in movie production has observed that, in general, movies with “big name” stars seem to generate more profits than those movies whose stars are less well known. To examine his beliefs he records the profits and the payment (in \$ millions) given to the two highest-paid actors in their two recently released movies. The data is as given in Table Worked Example 2.

Table Worked Example 2: Cost Paid to the Actor and Profit of the Movie

Actor	Cost Paid to the Actor	Profit of the Movie
Y	$S_y = 7.2$	65
X	$S_x = 8.0$	73
$r_{xy} = 0.48$		

- Develop the regression equation for predicting the profits of a movie acted by Y actor.
- Predict the profit of a movie that the Y actor act in if X actor act in a movie that gets a profit of 77.2.
- Develop the regression equation for predicting the profits of a movie acted by X actor.
- Predict the profit of a movie that the X actor act in if Y actor act in a movie that gets a profit of 69.4.
- Compute the standard error of the estimate for predicting the profit of a movie acted by Y actor from the profit of a movie acted by X actor.
- Compute the standard error of the estimate for predicting the profit of a movie acted by X actor from the profit of a movie acted by Y actor.

Answer:

(a)

$$\begin{aligned}
 b &= r_{xy} \cdot (S_y / S_x) \\
 &= .48 (7.2 / 8.0) \\
 &= 0.432
 \end{aligned}$$

$$\begin{aligned}
 a &= \hat{Y} - bX \\
 &= 65 - 0.432 (73) \\
 &= 65 - 31.536 \\
 &= 33.464
 \end{aligned}$$

$$\begin{aligned}
 \hat{Y} &= bX + a \\
 \hat{Y} &= 0.432X + 33.464
 \end{aligned}$$

(b)

$$\begin{aligned}
 \hat{Y} &= 0.432X + 33.464 \\
 \hat{Y} &= 0.432 (77.2) + 33.464 \\
 &= 33.3504 + 33.464 \\
 &= 66.8144
 \end{aligned}$$

(c)

$$\begin{aligned}
 b &= r_{xy} \cdot (S_x / S_y) \\
 &= .48 (8.0 / 7.2) \\
 &= 0.533
 \end{aligned}$$

$$\begin{aligned}
 a &= X - bY \\
 &= 73 - 0.533 (65) \\
 &= 73 - 34.645 \\
 &= 38.355
 \end{aligned}$$

$$\begin{aligned}
 X &= bY + a \\
 X &= 0.533Y + 38.355
 \end{aligned}$$

(d)

$$\begin{aligned}
 X &= 0.533Y + 38.355 \\
 X &= 0.533 (69.4) + 38.355 \\
 &= 36.9902 + 38.355 \\
 &= 75.3452
 \end{aligned}$$

(e)

$$\begin{aligned}
 S_{xy} &= S_y \sqrt{1 - r^2} \\
 &= 7.2 \sqrt{1 - (0.48)^2} \\
 &= 7.2 \sqrt{1 - 0.2304} \\
 &= 7.2 \sqrt{0.7696} \\
 &= \pm 6.316
 \end{aligned}$$

(f)

$$\begin{aligned}
 S_{xy} &= S_x \sqrt{1 - r^2} \\
 &= 8.0 \sqrt{1 - (0.48)^2} \\
 &= 8.0 \sqrt{1 - 0.234} \\
 &= 8.0 \sqrt{0.7696} \\
 &= \pm 7.018
 \end{aligned}$$

Worked Example 3

Presentation and assignment has become part of life for most of the undergraduates. Students always try their best to score high marks on presentation as it can contribute to their final results. Some students are able to perform well in presentation but some students cannot. Shyness is a tendency to avoid social interactions and fail to participate appropriately in social situations. Previous research had stated that majority of the population had experienced this feeling at some point in their lives. A researcher is interested in investigating whether there is a relationship between shyness, fear of negative evaluation and performance of presentation. The corresponding shyness scores, FNE scores and presentation marks for 10 participants are listed in the Table Worked Example 3(a).

Table Worked Example 3(a): The corresponding shyness scores, FNE scores and presentation marks

Number of students	Presentation Scores	Shyness	Fear of Negative Evaluation
1	85	23	76
2	68	33	57
3	78	26	70
4	86	20	77
5	60	35	50
6	76	23	67
7	65	34	57
8	74	32	70
9	67	37	62
10	58	42	50

Table Worked Example 3(b): Shyness and Presentation Scores

Number of students	Presentation Scores (X)	Shyness (Y)	X^2	Y^2	XY
1	85	23	7225	529	1955
2	68	33	4624	1089	2244
3	78	26	6084	676	2028
4	86	20	7396	400	1720
5	60	35	3600	1225	2100
6	76	23	5776	529	1748
7	65	34	4225	1156	2210
8	74	32	5476	1024	2368
9	67	37	4489	1369	2479
10	58	42	3364	1764	2436
TOTAL	$\sum X = 717$	$\sum Y = 305$	$\sum X^2 = 52259$	$\sum Y^2 = 9761$	$\sum XY = 21288$

Answer:

1. Compute the means, \bar{X} , Standard deviation, S , correlation coefficient, r , and covariance, cov .

$$\bar{X}_X = \frac{\sum X}{n} = \frac{717}{10} = 71.7$$

$$S_X = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}} = \sqrt{\frac{52259 - \frac{717^2}{10}}{9}} = \sqrt{\frac{850.1}{9}} = \sqrt{94.4556} = 9.7188$$

$$\bar{X}_Y = \frac{\sum Y}{n} = \frac{305}{10} = 30.5$$

$$S_Y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n-1}} = \sqrt{\frac{9761 - \frac{305^2}{10}}{9}} = \sqrt{\frac{458.5}{9}} = \sqrt{50.9444} = 7.1375$$

$$\text{COV}_{XY} = \frac{\sum XY - \frac{\sum X \cdot \sum Y}{n}}{n-1} = \frac{21288 - \frac{717 \cdot 305}{10}}{9} = -\frac{580.5}{9} = -64.5$$

$$r_{XY} = \frac{\text{COV}_{XY}}{S_X \cdot S_Y} = -\frac{64.5}{9.7188 \cdot 7.1375} = -\frac{64.5}{69.3679} = -0.9298$$

$$r_{XY}^2 = 0.8646$$

2. Compute the standard error of the estimate for predict (Y) from (X) and predict (X) from (Y).

$$S_{YX} = S_Y \cdot \sqrt{1-r^2} = 7.1375 \cdot \sqrt{1-0.8646} = 7.1375 \cdot 0.368 = 2.6266$$

$$S_{XY} = S_X \cdot \sqrt{1-r^2} = 9.7188 \cdot \sqrt{1-0.8646} = 9.7188 \cdot 0.368 = 3.5765$$

3. Compute the regression coefficients for the data and form the regression equation.

To predict (X)

$$b = \frac{\text{COV}_{XY}}{S_Y^2} = -\frac{64.5}{50.9444} = -1.2661$$

$$a = \bar{X} - (b \cdot \bar{Y}) = 71.7 - (-1.2661 \cdot 30.5) = 71.7 + 38.6161 = 110.3161$$

$$\hat{X} = bY + a = -1.2661Y + 110.3161$$

To predict (Y)

$$b = \frac{COV_{XY}}{S_X^2} = -\frac{64.5}{94.4556} = -0.6829$$

$$a = \bar{Y} - (b \cdot \bar{X}) = 30.5 - (-0.6829 \cdot 71.7) = 30.5 + 48.9611 = 79.4611$$

$$\hat{Y} = bX + a = -0.6829X + 79.4611$$

4. Use regression equation to predict performance for presentation of shyness score 39.

$$\hat{X} = bY + a = -1.2661Y + 110.3161 = -1.2661(39) + 110.3161 = 60.9382$$

5. Use regression equation to predict shyness score if performance of presentation score 80.

$$\hat{Y} = bX + a = -0.6829X + 79.4611 = -0.6829(80) + 79.4611 = 24.8291$$

6. Use z-score regression equations to predict performance of presentation if shyness scores 45.

$$Z_Y = \frac{Y - \bar{Y}}{S_Y} = \frac{45 - 30.5}{7.1375} = 2.0315$$

$$\hat{Z}_X = Z_Y \cdot r_{XY} = 2.0315 \cdot (-0.9298) = -1.8889$$

$$\hat{X} = \bar{X} + (\hat{Z}_X \cdot S_X) = 71.7 + (-1.8889 \cdot 9.7188) = 71.7 - 18.3579 = 53.3421$$

7. Using z-score regression equations to predict shyness if performance of presentation score 79.

$$Z_X = \frac{X - \bar{X}}{S_X} = \frac{79 - 71.7}{9.7188} = 0.7511$$

$$\hat{Z}_Y = Z_X \cdot r_{XY} = 0.7511 \cdot (-0.9298) = -0.6984$$

$$\hat{Y} = \bar{Y} + (\hat{Z}_Y \cdot S_Y) = 30.5 + (-0.6984 \cdot 7.1375) = 30.5 - 4.9846 = 25.5154$$

- Overall, shyness have a high correlation with performance of presentation which $r = 0.9298$.

Table Worked Example 3(c): Fear of negative evaluation and presentation scores

Number of students	Presentation Scores (x)	Fear of Negative Evaluation (y)	X^2	Y^2	XY
1	85	76	7225	5776	6460
2	68	57	4624	3249	3876
3	78	70	6084	4900	5460
4	86	77	7396	5929	6622
5	60	50	3600	2500	3000
6	76	67	5776	4489	5092
7	65	57	4225	3249	3705
8	74	70	5476	4900	5180
9	67	62	4489	3844	4154
10	58	50	3364	2500	2900
TOTAL	$\sum X = 717$	$\sum Y = 636$	$\sum X^2 = 52259$	$\sum Y^2 = 41336$	$\sum XY = 46449$

1. Compute the means, \bar{X} , Standard deviation, S , correlation coefficient, r , and covariance, cov.

$$\bar{X}_x = \frac{\sum X}{n} = \frac{717}{10} = 71.7$$

$$S_x = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}} = \sqrt{\frac{52259 - \frac{717^2}{10}}{9}} = \sqrt{\frac{850.1}{9}} = \sqrt{94.4556} = 9.7188$$

$$\bar{X}_y = \frac{\sum Y}{n} = \frac{636}{10} = 63.6$$

$$S_y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n-1}} = \sqrt{\frac{41336 - \frac{636^2}{10}}{9}} = \sqrt{\frac{886.4}{9}} = \sqrt{98.4889} = 9.9242$$

$$COV_{xy} = \frac{\sum XY - \frac{\sum X \cdot \sum Y}{n}}{n-1} = \frac{46449 - \frac{717 \cdot 636}{10}}{9} = \frac{847.8}{9} = 94.2$$

$$r_{xy} = \frac{COV_{xy}}{S_x \cdot S_y} = -\frac{94.2}{9.7188 \cdot 9.9242} = -\frac{64.5}{96.4513} = 0.9767$$

$$r_{xy}^2 = 0.9539$$

2. Compute the standard error of the estimate for predict (Y) from (X) and predict (X) from (Y).

$$S_{YX} = S_Y \cdot \sqrt{1 - r^2} = 9.9242 \cdot \sqrt{0.0461} = 9.9242 \cdot 0.3146 = 2.1298$$

$$S_{XY} = S_X \cdot \sqrt{1 - r^2} = 9.7188 \cdot \sqrt{0.0461} = 9.7188 \cdot 0.2146 = 2.0857$$

3. Compute the regression coefficients for the data and form the regression equation.

To predict (X)

$$b = \frac{COV_{XY}}{S_Y^2} = \frac{94.2}{98.4889} = 0.9565$$

$$a = \bar{X} - (b \cdot \bar{Y}) = 71.7 - (0.9565 \cdot 63.6) = 71.7 - 60.8334 = 10.8666$$

$$\hat{X} = bY + a = 0.9565Y + 10.8666$$

To predict (Y)

$$b = \frac{COV_{XY}}{S_X^2} = \frac{94.2}{94.4556} = 0.9973$$

$$a = \bar{Y} - (b \cdot \bar{X}) = 63.6 - (0.9973 \cdot 71.7) = 63.6 - 71.5064 = -7.9064$$

$$\hat{Y} = bX + a = 0.9973X - 7.9064$$

4. Use regression equation to predict performance for presentation of FNE score 65.

$$\hat{X} = bY + a = 0.9565Y + 10.8666 = 0.9565(65) + 10.8666 = 62.1725 + 10.8666 = 73.0391$$

5. Use regression equation to predict FNE score if performances of presentation score 90.

$$\hat{Y} = bX + a = 0.9973X - 7.9064 = 0.9973(90) - 7.9064 = 89.757 - 7.9064 = 81.8506$$

6. Using z-score regression equations to predict performance of presentation if FNE scores 65.

$$Z_Y = \frac{Y - \bar{Y}}{S_Y} = \frac{65 - 63.6}{9.9242} = 0.1411$$

$$\hat{Z}_X = Z_Y \cdot r_{XY} = 0.1411 \cdot (0.9767) = 0.1378$$

$$\hat{X} = \bar{X} + (\hat{Z}_X \cdot S_X) = 71.7 + (0.1378 \cdot 9.7188) = 71.7 + 1.3391 = 73.0391$$

7. Using z-score regression equations to predict FNE scores if performances of presentation score 90.

$$Z_X = \frac{X - \bar{X}}{S_X} = \frac{90 - 71.7}{9.7188} = 1.8829$$

$$\hat{Z}_Y = Z_X \cdot r_{XY} = 1.8829 \cdot (0.9767) = 1.8391$$

$$\hat{Y} = \bar{Y} + (\hat{Z}_Y \cdot S_Y) = 63.6 + (1.8391 \cdot 9.9242) = 63.6 + 18.2514 = 81.8514$$

- Proved that z-score regression and regression equation are able to find the same answer and FNE have a high correlation with performance of presentation which $r = 0.9767$

OUTPUT OF SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,993 ^a	,987	,983	1,277

a. Predictors: (Constant), fne, shyness

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	838,678	2	419,339	256,989	,000 ^a
	Residual	11,422	7	1,632		
	Total	850,100	9			

a. Predictors: (Constant), fne, shyness

b. Dependent Variable: scores

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	44,104	8,512		5,182	,001
	shyness	-,477	,116	-,350	-4,127	,004
	fne	,663	,083	,677	7,973	,000

a. Dependent Variable: scores

Worked Example 4

Refer to Data Set1 Family Functioning from the CD given for the exploration of Multiple Regression. Through the analysis of Matrix Correlation, we know that family functioning, perceived social support all affect the respondent's self-esteem. However, which is more important? When you use the 2 variables to predict self-esteem, how much contribution do each variable play in predicting self-esteem?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.442(a)	.195	.189	24.467

Predictors: (Constant), total SS, total FFS

ANOVA(b)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	35342.361	2	17671.180	29.519	.000(a)
	Residual	145468.598	243	598.636		
	Total	180810.959	245			

a Predictors: (Constant), total SS, total FFS

b Dependent Variable: totalSE

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-26.842	14.699		-1.826	.069
	total FFS	-.081	.089	-.056	-.904	.367
	total SS	1.107	.149	.460	7.408	.000

a Dependent Variable: totalSE

Answer:

Table: Summary of Ordinary Least Squares Multiple Regression Analysis for Social Support and Family Functioning for Predicting Self-Esteem ($N=245$)

	B	SE B	B
Self-Esteem			
Social Support	1.107	.149	-.056
Family Functioning	-.081	.089	.460**

Note: $R^2 = .195$ $F[(2, 243) = 29.519, p < .01]$. ** Significant at $p < .01$ level.

Family Functioning and social support together contribute 19.5% towards respondent self esteem. The 2 variables significantly contribute toward self esteem and the likelihood of such a result arising by sampling error is 1 in 100. [$F(2, 243) = 29.519, p < .01$]

Family Functioning

Family Functioning has a regression coefficient of $-.056$. However, $t = -.904$, $p > .05$. Thus our regression coefficient is likely to have arisen by sampling error. We can conclude that Family Functioning does not play a significant role in affecting the respondent's self esteem as compare to Social Support.

Social Support

Social Support has a regression coefficient of $.460$. Thus as social support increases by one unit, self esteem increase by $.460$. The t -value is 7.408 with an associated probability of $p < .01$. Hence our regression coefficient is unlikely to have arisen by sampling error.

Conclusion

Based on the above coefficient table, it can be concluded that between the two variables, social support plays a more important role in affecting the respondent's self esteem.

Worked Example 5

The number of students who are likely to work while pursuing their tertiary education is mushrooming. These groups of working students can be categorized into two clusters: those who primarily identify themselves as students but who work in order to pay the fees, and those who are first and foremost workers who also take some college classes. Generally, it is undoubted that, employing part-time students may have beneficial effects. For instance, important working experience that will improve future labour market prospects can be obtained. Nevertheless, working part-time also appears to replace non-productive activities, such as watching television. Research indeed supports this scenario, whereby students who work fewer than 10 hours per week seem to score slightly higher CGPAs than other similar students. In other words, working for a limited number of hours (Eg: 10 hours a week) appear to leave a positive impacts on student performance. In contrast, full-time employment (Eg: 35 hours a week) can impair students' academic performance. As full-time working among students appears to have negative effects on student academic performance, it has raised the Board of Directors of ABC University concern.

Research Questions

1. Is there a relationship between number of hours worked per week and academic performance among ABC University students?
2. Which group of students tends to perform better in terms of academic?
3. Which factor contributes the most towards academic performance among ABC university students?

Research Hypothesis

There is a positive relationship between numbers of hours worked per week and academic performance among ABC University students.

The higher the numbers of hours worked per week, the better the academic performance among ABC University students.

Null Hypothesis

There is no relationship between numbers of hours worked per week and academic performance among ABC University students.

Table Worked Example 5(a): Numbers of hours worked per week and Academic Performance

Students	Number of working hours per week. (X)	X^2	Academic Performance (in CGPA) (Y)	Y^2	Amount of Grant provided (RM)
1	18.5	342.25	3.02	9.1204	4478
1	22.9	524.41	2.97	8.8209	4578
1	29.5	870.25	2.24	5.0176	3500
1	32.5	1056.25	2.33	5.4289	3320
1	24.5	600.25	2.71	7.3441	4465
2	14	196	3.76	14.1376	5100
2	18	324	3.14	9.8596	4578
2	20	400	2.90	8.41	4986
2	14.5	210.25	3.90	15.21	4996
2	17.5	306.25	3.20	10.24	4645
1	29.5	870.25	2.44	5.9536	3517
1	14.5	210.25	3.67	13.4689	4140
1	16.5	272.25	3.40	11.56	3999
2	10.5	110.25	3.87	14.9769	5486
2	14.5	210.25	2.95	8.7025	5109
2	13.5	182.25	3.14	9.8596	5217
1	24.5	600.25	2.26	5.1076	3814
1	32	1024	2.44	5.9536	3400
2	17.5	306.25	3.49	12.1801	4319
2	15.5	240.25	3.00	9.00	5015
Total	400.4	8856.16	60.83	190.3519	88662

Indicator: 1 = Part time students

2 = Full time students

Answer for Research Question 1:

According to SPSS 16.0,

Descriptive Statistics

	Mean	Std. Deviation	N
Students CGPA - performance	3.0415	.53002	20
Number of hours worked per week	20.0200	6.64970	20

Table Worked Example 5(b): Correlation between numbers of hours worked per week and academic performance.

Correlations

		Number of hours worked per week	Students CGPA - performance
Number of hours worked per week	Pearson Correlation	1.000	-.865**
	Sig. (2-tailed)		.000
	N	20.000	20
Students CGPA - performance	Pearson Correlation	-.865**	1.000
	Sig. (2-tailed)	.000	
	N	20	20.000

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

According to manual calculation,

$$\sum X = 400.4 \quad \sum X^2 = 8856.16 \quad \sum Y = 60.83 \quad \sum Y^2 = 190.3519 \quad \sum XY = 1159.913$$

$$\begin{aligned} \text{Correlation coefficient, } r &= \frac{(N \cdot \sum XY) - (\sum X \cdot \sum Y)}{\sqrt{[(N \cdot \sum X^2) - (\sum X)^2] \cdot [(N \cdot \sum Y^2) - (\sum Y)^2]}} \\ &= \frac{(20 \times 1159.913) - (400.4 \times 60.83)}{\sqrt{[(20 \times 8856.16) - (400.4)^2] \cdot [(20 \times 190.3519) - (60.83)^2]}} \\ &= -.8647 \end{aligned}$$

$$\begin{aligned} \text{Degree of freedom, } df &= N-2 \\ &= 18 \end{aligned}$$

Results on Correlation Analysis:

Correlation analysis was conducted to identify the relationships between numbers of hours worked per week and academic performance among ABC University students. And according to the displayed results in Table 1, numbers of hours worked per week ($M=20.02$, $SD=6.6497$) was significantly correlated with the academic performance ($M=3.0415$, $SD=.53002$) at the level of 0.01 and with the correlation coefficients ($r = -.865$, $p<.01$). Consequently, when ABC university students tend to work overtime per week, turns up to obtain poorer CGPA.

Research hypothesis 1: Positively correlated – accepted

Null hypothesis: Rejected

Answer for Research Question 2:

According to SPSS 16.0,

Group Statistics					
Students		N	Mean	Std. Deviation	Std. Error Mean
Students CGPA - performance	Part-time Students	10	2.7480	.50117	.15848
	Full-time Students	10	3.3350	.38788	.12266

Table Worked Example 5(c): Independent Samples Test between part time and full time working ABC University students

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Students CGPA - performance	Equal variances assumed	.699	.414	-2.929	18	.009	-.58700	.20041	-1.00804	-.16596
	Equal variances not assumed			-2.929	16.935	.009	-.58700	.20041	-1.00994	-.16406

According to manual calculation,
Variance for part-time student,

$$\begin{aligned}
 \sigma^2 &= \frac{\sum A^2 - \frac{\sum A^2}{N}}{N-1} \\
 &= \frac{77.7756 - 27.48^2/10}{9} \\
 &= .25117
 \end{aligned}$$

Hence, standard deviation for part-time student = .50084

$$\begin{aligned}
 \text{Variance for full-time student, } \sigma^2 &= \frac{\sum B^2 - \frac{\sum (B)^2}{N}}{n-1} \\
 &= \frac{112.5763 - 33.35^2/10}{9} \\
 &= 0.15048
 \end{aligned}$$

Hence, standard deviation for full-time student = .3879

Chapter 2 Review Exercises

Question 1

The table below shows the incomes, x and entertainment expenditures, y (in hundreds of ringgits) of 10 sales engineers.

Incomes	28	40	45	60	35	55	32	72	30	70
Entertainment	9	7	15	10	8	12	5	23	8	10

- Find the simple regression line.
- Interpret the values of a and b .
- Predict the entertainment expenditure of a sales engineer whose income is RM5000.

Question 2

The data for age and systolic blood pressure of eight randomly selected subjects are shown in the following table.

Subject	Age, X	Pressure, Y
A	34	110
B	45	126
C	50	125
D	58	135
E	63	145
F	72	150
G	75	152
H	38	120

- Find the simple regression line.
- Predict the systolic pressure of a man who is 55 years old.

Question 3

A vehicle manufacturing company wants to investigate how the price of one of the motorcycle models depreciates with its age. The research department of the company took a sample of ten motorcycles of this model and collected the following information on the ages (in years) and prices (in thousands of RM) of these motorcycles.

Age	5	3	2	9	1	4	7	6	8	10
Price	2.2	2.5	3.5	1.0	4.0	2.4	1.5	1.8	1.2	0.7

- Find the simple regression line.
- Interpret the values of a and b .
- Predict the price of a 8.5-year-old motorcycle.

Question 4

Modern warehouses employ computerized and automated guided vehicles to handle materials. As a result, the physical layout of the warehouse must be carefully designed to prevent vehicle congestion and also to optimize the response time. The data are shown in the following table. Of interest to the researcher is the relationship between congestion time, y and number of vehicles, x .

Number of vehicles	Congestion time (s)
3	38
4	40
5	45
6	75
7	75
8	80
9	110
10	120
11	125
12	130

- Find the simple regression line.
- Interpret the values of a and b .
- Predict the congestion time when there are 15 vehicles in the warehouse.

Question 5

A research was done by a lecturer to compare between the average number of hours student spent in study everyday, x and the average number of hours they sleep everyday, y . The data was recorded as follows:

x	2	2	6.6	4	6	1	2.5	1
y	10	9	9	12	8	7	5	6

Find the simple regression line.

Question 6

Listed below are pressure and compression of a machine.

Pressure, x	Compression, y
1.5	1.5
3.0	1.5
4.5	3.0
6.0	3.0
7.5	5.0

- Find the simple regression line.
- Predict the compression when the pressure is 6.5.

Question 7

A survey was carried out by a marketing representative to determine the relation between monthly promotion items and sales. The following data were recorded:

Number of Promotion Items	Sales (RM, in thousands)
45	38
25	40
30	39
25	36
35	47
55	44
45	49
60	55
50	56
40	52
25	48
50	51

- Find the equation of the regression line to predict monthly sale.
- Estimate the monthly sales when there are 53 promotion items.

Question 8

A study is done to see whether there is a relationship between the average number of hours the students spent for online games per week and their grade point average (GPA).

Hour (x)	30	12	19	10	35	12	30	10	8
GPA (y)	2.2	3.5	3.0	4.0	1.8	3.2	2.4	3.8	3.7

- Find the simple regression line.
- Interpret the values of a and b .
- Predict the GPA if a student spends 20 hours per week playing online games.

Question 9

A researcher wants to investigate the relationship between the sizes and prices of houses in a town. The researcher collected the following information on the sizes (in thousands of square meters) of seven houses and the buying prices (in thousands of ringgits) paid by the house owner.

Sizes of houses ($\times 1000 \text{ m}^2$)	1.8	2.2	4.5	3.0	3.2	3.5	2.3
Buying Prices ($\times 1000$ ringgits)	150	180	350	250	270	300	200

- Find the simple regression line.
- Predict the buying price of a house which is 4000 m^2 .

Question 10

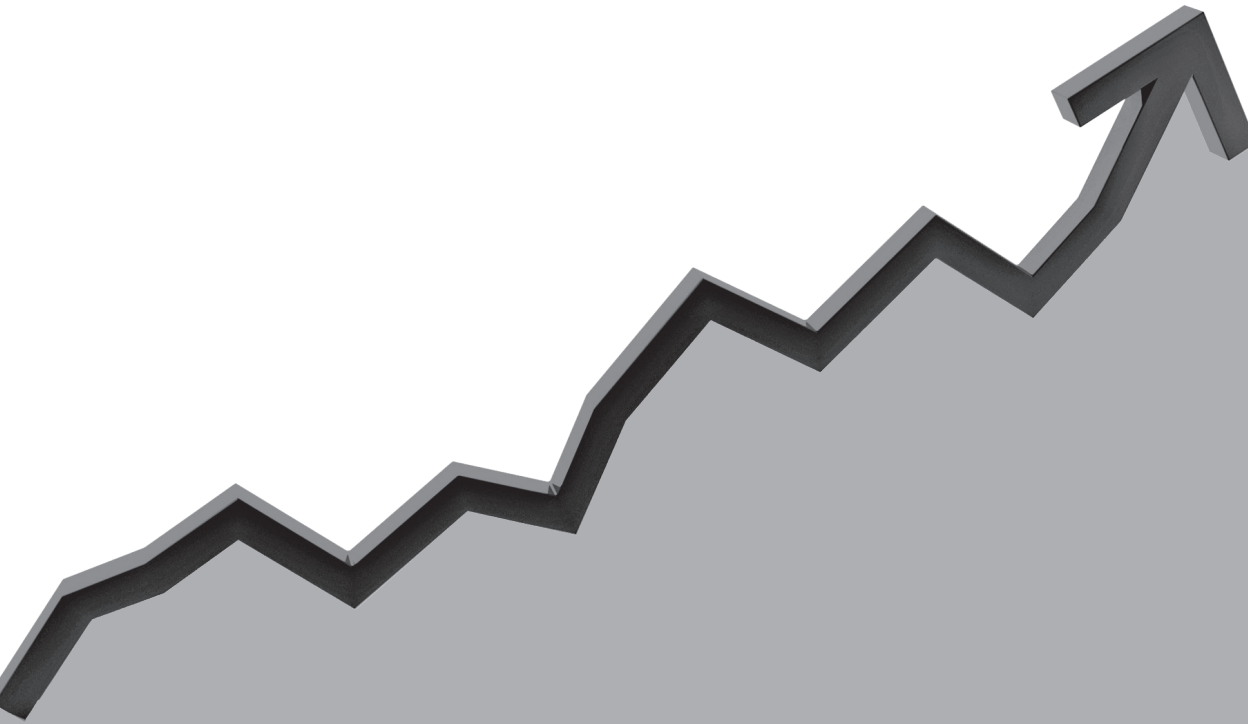
The data below were obtained in a study on the number of hours ten people exercise each week and the amount of water (in *l*) each person consumes per week.

Hours, x	2	0	3	6	8	5	10	3	4	7
Amount, y	14	13	15	16.5	16	13	15	14	13.5	17

- Find the simple regression line.
- Interpret the values of a and b .
- Predict the consumption of drinking water for a woman who exercises for 14 hours per week.

chapter three

t - TEST



Learning Objectives

Chapter 3 highlights the analysis in the differences between the two groups, specifically the mean differences. The main aim for the exercises is to enable students to understand the proper ways to do manual calculation for independent sample t-test and paired sample t-test. At the completion of the chapter, students should be able to:

- Analyze the result of the manual calculation which includes mean, standard deviation and standard error.
- Form research and null hypotheses for independent-sample t-test.
- Apply the procedure for independent-samples t-test and paired sample t-test using SPSS.
- Report the significance of the analysis using manual calculation and SPSS.

3.0 Introduction

There are many statistical techniques for comparison, such as:

- Single sample t-Test
- Independent-sample t-Test
- Paired-samples t-Test
- One-way analysis of variance
- Two-way analysis of variance
- Multiple analysis of variance
- Non-parametric techniques

t-Test is another statistical approach. It is used to compare whether there are differences between two means or if the mean of a sample is different from the mean of the population.

Nevertheless, there are a few different types of t-tests as stated which can be used, depending on the number of samples. The first type, known as '*single sample t-test*', is used when comparing the mean of a sample with the mean of the population. However, this method is not very popular and is seldom used, as it is rather difficult to obtain the mean of a population.

The second type of t-test, known as '*independent sample t-test*', is used when comparing whether there are differences between two independent sample means. For example, we can use the independent sample t-test for a study to compare the mean height among boys and girls in a class.

The third type of t-test, known as '*correlated sample t-test*' or '*paired sample t-test*' is used for comparing whether there are differences between two sample means which come from the same sample. For example, we can use this method for a study to compare the mean score of a quiz of the same 20 students before and after revising their notes. t-Test can be conducted even though the sample size is very small – even as small as ten.

In this chapter, independent-samples t-test and paired-samples t-test will be discussed.

3.1 Independent Sample t-Test

Independent Sample t-Test is used to determine whether the difference between means of two groups or conditions is due to the independent variable, or if the difference is simply due to chance. Thus, this procedure establishes the probability of the outcome of an experiment, and in doing so enables the researcher to reject or retain the null hypothesis.

The null hypothesis states that the experimental manipulation has no effect, therefore the means of the groups will be equal. In this respect, the t-test is an inferential statistic used to test hypotheses. Under ideal conditions, these types of inferential statistics allow the researcher to infer a causal relationship between the independent and dependent variable.

There are two distinct applications of the t-test. When a between-subjects design is used, the independent-samples t-test is the appropriate test. For example when comparing the mean score of statistic among male and female. Use of a within-subjects design (sometimes called a repeated measures design) or a participant-by-participant matched design requires analysis with the paired samples t-test (also known as the correlated or paired-samples t-test). For example, when comparing the mean score of statistic within

the same sample group before and after a certain activities carried out namely the difference before therapy and after therapy on a same sample.

In order to work out the calculated t-value, one must perform the following steps:

Step 1: Calculate the mean and standard deviation for both samples.

Step 2: Calculate the estimate of the standard error of the mean for both samples.

Step 3: Work out the Value of calculated t.

Step 4: Calculate the degree of freedom.

Step 5: Identify the critical value of t from the t value table.

Step 6: Compare the calculated t and the critical t value and make statistical decision.

If t calculated is greater than t critical, the null hypothesis has to be rejected and the research hypothesis will be accepted. This simply shows that there is a significant difference between the mean scores of the 2 samples.

3.2 Single-Sample t-Test Concerning The Mean

Case : “ σ^2 is unknown and $n < 30$ ”

Suppose we have a sample of size n taken from a normally distributed population whose mean is μ and variance is unknown. We would like to test whether this sample is taken from a population whose mean is μ_0 .

Steps:

- (1) $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$
- (2) $\alpha = 0.05$
- (3) Critical region :
Reject H_0 if $T > t_{\alpha/2, n-1}$ or $T < -t_{\alpha/2, n-1}$
- (4) Test statistic $T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$
- (5) Conclusion

3.3 Two Samples: t-Test Concerning The Difference Between Two Means

Case 1: σ_1^2 and σ_2^2 are unknown and $n_1, n_2 \leq 30$

Two independent samples of size n_1 and n_2 taken from approximately normally distributed population with the mean μ_1, μ_2 and unknown variances. To test whether these samples are taken from the population whose means are equal,

Case 2: $\sigma_1^2 = \sigma_2^2$ (Equal variance)

- (i) $H_0 : \mu_1 = \mu_2$, or $H_0 : \mu_1 - \mu_2 = 0$, vs $H_a : \mu_1 \neq \mu_2$

- (ii) For a particular value of α , determine the critical region for t -value:
- (iii) Rejection criteria: Reject H_o if $T > t_{\alpha/2, n_1+n_2-2}$ or $T < -t_{\alpha/2, n_1+n_2-2}$

(v) Test statistic
$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where **pooled variance**,

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

- (vi) Conclusion

Case : $\sigma_1^2 \neq \sigma_2^2$ (**Unequal variance**)

$$H_0 : \mu_1 = \mu_2 \text{ or } H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_a : \mu_1 \neq \mu_2$$

- (i) For a particular value of α , determine the critical region for t .
- (ii) Rejection criteria: Reject H_o if $T > t_{\alpha/2, v}$ or $T < -t_{\alpha/2, v}$

(v) Test statistic:
$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where

$$w_1 = \frac{s_1^2}{n_1}, \quad w_2 = \frac{s_2^2}{n_2}, \quad v = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}}$$

3.4 Paired Observation / Paired t-Test

When the samples are **not independent** and \bar{d} and s_d are the mean and standard deviation of the normally distributed difference of n random pairs of measurement, then

- (i) $H_0 : \mu_D = 0$ vs $H_a : \mu_D \neq 0$
- (ii) $\alpha = 0.05$
- (iii) Critical region : Reject H_o if $T > t_{\alpha/2, n-1}$ or $T < -t_{\alpha/2, n-1}$
- (iv) Test statistic
$$T = \frac{\bar{d} - \mu_D}{s_d / \sqrt{n}}$$

where

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} \quad \text{and} \quad s_d = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n d_i^2 - \frac{\left(\sum_{i=1}^n d_i \right)^2}{n} \right]}$$

3.5 Independent-samples t-test with SPSS

An independent-samples t-test is used when you want to compare the mean score (continuous variable), for two different groups of subjects.

3.5.1 Procedure for independent-samples t-test

- Use Data Set1 Family Functioning (from the CD given)
- Select **Analyze**, click on **Compare means**, then on **Independent Samples T-test**.
- Move the dependent (continuous variable, e.g. total self-esteem) into the **Test variable**.
- Move the independent variable (categorical) variable, e.g. sex in the **Grouping variable**.
- Click on **Define groups**. In the **Group 1** box, type 1; and in the **Group 2** box, type 2.
- Click on **Continue** and then **OK**.

3.5.2 Interpretation of SPSS Output

Group Statistics

		N	Mean	Std. Deviation	Std. Error Mean
totalSE	Male	123	27.68	28.530	2.573
	Female	123	25.52	25.801	2.326

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
		Lower	Upper	Lower	Upper	Lower	Upper	Lower
totalSE	Equal variances assumed	1.043	.308	.624	244	.534	2.163	3.468
	Equal variances not assumed			.624	241.574	.534	2.163	3.468

From the SPSS output above, there is a slight difference of mean for self-esteem among male and female respondents. From the group statistic table, it indicates that mean for male = 27.68 and mean for female = 25.52. However, when examine the independent sample t-test table, $t(244) = .624$, $p > .05$. This means that there is no significant difference for the mean score of self-esteem for male and female Hence, the slight difference of mean among male and female would have arisen by sampling errors.

Worked Examples

Worked Example 1

A researcher wanted to study the impact of keyboard designs on the number of words typed. He designed two different brands of keyboard and randomly selected 2 groups of students to determine the number of words typed per minute. Group one typed words using brand A whereas group 2 using brand B. Analyze the data below and determine whether the two keyboards are significantly different on the number of words typed.

Table Worked Example 1: Keyboard A and Keyboard B on the Words Typed per minute

Words typed per minute			
Brand A, (X)	Brand B, (Y)	(X) ²	(Y) ²
63	51	3969	2601
55	46	3025	2116
70	60	4900	3600
68	63	4624	3969
67	55	4489	3025
75	70	5625	4900
50	48	2500	2304
65	55	4225	3025
79	42	6241	1764
48	50	2304	2500
ΣX 640	ΣY 540	ΣX^2 41902	ΣY^2 29804
$\Sigma(X)^2 = 409600$	$\Sigma(Y)^2 = 291600$		

Solutions:

$$\begin{aligned}
 \text{a) } (S_1)^2 &= [41902 - (409600 / 10)] / (10-1) \\
 &= (41902 - 40960) / 9 \\
 &= 942 / 9 \\
 &= 104.6667 \\
 S_1 &= \sqrt{104.6667} \\
 &= 10.2307
 \end{aligned}$$

$$\begin{aligned}
 \text{b) } (S_2)^2 &= [29804 - (291600/10)] / (10-1) \\
 &= (41902 - 40960) / 9 \\
 &= 644 / 9 \\
 &= 71.5556 \\
 S_2 &= \sqrt{71.5556} \\
 &= 8.4591
 \end{aligned}$$

$$\begin{aligned}
 \text{c) est } \sigma_{\text{diff}} &= \sqrt{\frac{S_1^2}{n_1}} + \sqrt{\frac{S_2^2}{n_2}} \\
 &= \sqrt{[104.6667 / 10]} + \sqrt{[71.5556 / 10]} \\
 &= \sqrt{10.46667 + 7.1556} \\
 &= \sqrt{17.6223} \\
 &= 4.1979
 \end{aligned}$$

$$\begin{aligned}
 \text{d) df} &= (10-1) + (10-1) \\
 &= 9 + 9 \\
 &= 18
 \end{aligned}$$

$$\begin{aligned}
 \text{e) calculated } t &= (\bar{X}_1 - \bar{X}_2) / \sigma_{\text{diff}} \\
 &= (64 - 54) / 4.1979 \\
 &= 10 / 4.1979 \\
 &= 2.3821
 \end{aligned}$$

$$\text{f) critical } t = 2.101$$

Conclusion:

Calculated $t >$ critical t . Null hypothesis is rejected. Research hypothesis is accepted. There is a significant difference between types of keyboard with number of words typed. Keyboard A has the design that allows respondents to type more words than keyboard B. ($\bar{X}_1 = 64, > \bar{X}_2 = 54$)

Worked Example 2

One group of 6 participants was shown a movie on the evils of alcohol use and the other group of 7 participants was shown a movie about dog nutrition. Both groups were then given a test measuring their attitude toward drug use. The scores for both groups are listed in Table Worked Example 2 (high scores indicate a resistance to the use of drugs).

Table Worked Example 2: Scores for Drug Addiction

Alcohol Movie	Dog Movie
132	121
144	130
131	122
145	125
156	119
139	100
$\bar{X}_1 = 141.167$	$\bar{X}_2 = 119.5$
$S_1 = 9.326$	$S_2 = 10.291$
$n_1 = 6$	$n_2 = 6$

Given this data please see if there is a significant difference between the effect that the alcohol movie and the dog movie have on attitudes toward drug use.

Solutions:

1. Generate a null hypothesis and a research hypothesis.

Null hypothesis: The drug attitude scores for those watching the alcohol movie will not differ from those watching the dog movie.

$$H_0: \mu_1 = \mu_2$$

The research hypothesis: The drug attitude scores for those watching the alcohol movie will be higher than those watching the dog movie.

$$H_1: \mu_1 > \mu_2$$

2. Calculate the appropriate t

First compute the estimate of the standard error of the mean for each sample then use those values to compute the estimate of the standard error of the difference.

$$\begin{aligned} \text{est } \alpha_{x1} &= \frac{S_1}{\sqrt{n_1}} \\ &= \frac{9.326}{\sqrt{6}} \\ &= \frac{9.326}{2.449} \\ &= 3.808 \end{aligned}$$

$$\begin{aligned} \text{est } \alpha_{x2} &= \frac{S_2}{\sqrt{n_2}} \\ &= \frac{10.291}{\sqrt{6}} \\ &= \frac{10.875}{2.449} \\ &= 4.202 \end{aligned}$$

$$\begin{aligned} \text{est } \alpha_{\text{diff}} &= \sqrt{(\text{estimated } \alpha_{x1})^2 + (\text{estimated } \alpha_{x2})^2} \\ &= \sqrt{(3.808)^2 + (4.202)^2} \\ &= \sqrt{14.501 + 17.657} \\ &= \sqrt{32.158} \\ &= 5.671 \\ t &= \frac{(\bar{X}_1 - \bar{X}_2)}{\text{est } \alpha_{\text{diff}}} \\ &= \frac{141.167 - 119.5}{5.671} \\ &= \frac{21.667}{5.671} \\ &= 3.821 \end{aligned}$$

3. Calculate the degrees of freedom and find the critical value from the Table T

$$\begin{aligned} df &= (n_1 - 1) + (n_2 - 1) \\ &= (6 - 1) + (6 - 1) \\ &= 5 + 5 \\ &= 10 \end{aligned}$$

The critical value $t(df=10, p=.05, \text{two tailed}) = 2.228$

4. State whether to reject or fail to reject the null hypothesis

Reject the null hypothesis and accept the research hypothesis. Because the computed value for t , 3.821, is greater than the critical value, 2.228. People viewing the alcohol movie scored significantly higher (drug attitude scores) than people viewing the dog movie.

Worked Example 3

A health psychologist predicted that people who are moody would be happier when they consume 100g of milk chocolate every day. In order to verify his prediction on 10 participants, their level of moodiness was tested and recorded before and after consuming milk chocolate. The level of moodiness before and after consuming the milk chocolate was the dependent variable. The data collected are shown in Table Worked Example 3(a).

Table Worked Example 3(a): Level of Moodiness

X1 (Before consuming milk chocolate)	X2 (After consuming milk chocolate)
26.7	18.3
29.3	19.7
28.8	20.5
23.2	17.7
27.3	19.9
25.4	17.8
24.6	19.6
22.4	15.3
21.3	12.6
29.1	23.1

- Calculate sum of the difference (ΣD) and sum of the difference squared (ΣD^2) in the experiment.
- Calculate mean of the differences and estimate of the standard error of the differences.
- Find out the value of computed t .
- At $p = 0.05$, what is degree of freedom and value of critical t .
- Interpret the result.

Solutions:

Table Worked Example 3(b): Level of Moodiness

X1	X2	X1 - X2 (D)	D ²
26.7	18.3	8.4	70.56
29.3	19.7	9.6	92.16
28.8	20.5	8.3	68.89
23.2	17.7	5.5	30.25
27.3	19.9	7.4	54.76
25.4	17.8	7.6	57.76
24.6	19.6	5.0	25.00
22.4	15.3	7.1	50.41
21.3	12.6	8.7	75.69
29.1	23.1	6.0	36.00
		$\Sigma D = 73.6$	$\Sigma D^2 = 561.48$
		Mean D = $\Sigma D / N$ = $73/10$ = 7.36	

(a)

$$\Sigma D = 73.6$$

$$\Sigma D^2 = 561.48$$

(b) Mean of the differences

$$= \Sigma D / N = 73.6 / 10 = 7.36$$

Estimate of the standard error of the difference

$$= \frac{561.48 / 10 - (7.36)^2}{9}$$

$$= \frac{(56.15 - 54.17)}{9}$$

$$= 0.220$$

(c) Value of computed $t = 7.36 / 0.220 = 33.45$ (d) At $p = 0.05$, $df = 10 - 1 = 9$ Value of critical $t = 2.262$

(e) Computed $t >$ critical t at $p = 0.05$ indicated that there is significant difference for the level of moodiness before and after the consuming of milk chocolate. Therefore, the research hypothesis is accepted.

Worked Example 4

A counseling Psychologist, Dr Tam conducted a research on mathematical aptitude and creativity among 20 secondary school students. The collected data was as in Table Worked Example 4(a).

Table Worked Example 4(a): Mathematical Aptitude and Creativity

Participant	(X) Mathematical Aptitude	(Y) Creativity	Gender
1	245	165	1
2	273	173	1
3	238	155	1
4	246	163	1
5	283	169	1
6	245	165	1
7	276	143	1
8	238	165	1
9	246	168	2
10	289	167	2
11	283	169	2
12	245	161	2
13	273	173	2
14	237	155	2
15	246	162	2
16	288	164	2

Gender: 1: Male 2: Female

- Conduct the data entry.
- What is the correlation coefficient between mathematical aptitude and creativity?
- Form a null hypothesis.
- Should you reject your null hypothesis? Briefly interpret the result.
- Conduct t-test for the above data. Is there a significant difference for mathematical aptitude and creativity among males and females? (in your analysis, include t value and significant level).
- Manually calculate the r value.
- Manually calculate t value for mathematical aptitude and creativity among males and females. Briefly interpret the result.

Solutions:

- What is the correlation coefficient between mathematical aptitude and creativity?

Correlations

		Mathematical Aptitude	Creativity
Mathematical Aptitude	Pearson Correlation	1	.262
	Sig. (2-tailed)		.327
	N	16	16
Creativity	Pearson Correlation	.262	1
	Sig. (2-tailed)	.327	
	N	16	16

There is no significant relationship between mathematical aptitude and creativity. ($r = .262$, $p > 0.05$).

c) Form a null hypothesis.

There would be no relationship between mathematical aptitude and creativity

d) Should you reject your null hypothesis? Briefly interpret the result.

Accept null hypothesis. There is no significant relationship between mathematical aptitude and creativity. ($r = 0.262$, $p > 0.05$).

e) Conduct t-test for the above data. Is there a significant difference for mathematical aptitude and creativity among males and females?

Group Statistics

		N	Mean	Std. Deviation	Std. Error Mean
Mathematical Aptitude	Male	8	255.50	18.540	6.555
	Female	8	263.38	21.967	7.767

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower
Mathematical Aptitude	Equal variances assumed	1.433	.251	-.775	14	.451	-7.875	10.163	29.672	13.922
	Equal variances not assumed			-.775	13.616	.452	-7.875	10.163	29.730	13.980

Mean of mathematical aptitude in males = 255.50

Mean of mathematical aptitude in females= 263.38

$t(14) = -0.775$, $p > 0.05$

There is no significant difference for mathematical aptitude between males and females.

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Creativity	Male	8	162.25	9.316	3.294
	Female	8	164.88	5.592	1.977

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		Upper	Lower
Creativity	Equal variances assumed	.880	.364	.683	14	.506	-2.625	3.841		10.864	5.614
	Equal variances not assumed			.683	11.465	.508	-2.625	3.841		11.038	5.788

Mean of creativity in males = 162.25

Mean of creativity in females= 164.88

$t(14) = -0.683, p > 0.05$

There is no significant difference for creativity between males and females

f)

Table Worked Example 4(b): Relationships between Mathematical Aptitude and Creativity

Participants	Mathematical Aptitude (X)	Creativity (Y)	Gender	XY	Y ²	X ²
1	245	165	1	40425	27225	60025
2	273	173	1	47229	29929	74529
3	238	155	1	36890	24025	56644
4	246	163	1	40098	26569	60516
5	283	169	1	47827	28561	80089
6	245	165	1	40425	27225	60025
7	276	143	1	39468	20449	76176
8	238	165	1	39270	27225	56644
9	246	168	2	41328	28224	60516
10	289	167	2	48263	27889	83521
11	283	169	2	47827	28561	80089
12	245	161	2	39445	25921	60025
13	273	173	2	47229	29929	74529
14	237	155	2	36735	24025	56169
15	246	162	2	39852	26244	60516
16	288	164	2	47232	26896	82944
	$\Sigma X = 4151$	$\Sigma Y = 2617$		$\Sigma XY = 679543$	$\Sigma Y^2 = 428897$	$\Sigma X^2 = 1082957$

$$\begin{aligned}
 r &= \frac{(n \cdot \sum XY) - (\sum X \cdot \sum Y)}{\sqrt{[(n \cdot \sum X^2) - (\sum X)^2] \cdot [(n \cdot \sum Y^2) - (\sum Y)^2]}} \\
 &= \frac{(16 \cdot 679543) - (4151 \cdot 2617)}{\sqrt{[16 \cdot 1082957 - (4151)^2] \cdot [(16 \cdot 428897) - (2617)^2]}} \\
 &= 0.262
 \end{aligned}$$

$$df = 16 - 2 = 14$$

$$t_{cri} = 0.497$$

$t_{cri} > t_{cal}$, accept null hypothesis

There is no significant relationship between mathematical aptitude and creativity.

g.

Table Worked Example 4(c): t-Test of Mathematical Aptitude by Gender Status

Participants	MALE (X ₁)	FEMALE (X ₂)	X ₁ ²	X ₂ ²
1	245	246	60025	60516
2	273	289	74529	83521
3	238	283	56644	80089
4	246	245	60516	60025
5	283	273	80089	74529
6	245	237	60025	56169
7	276	246	76176	60516
8	238	288	56644	82944
	$\sum X_1 = 2044$ $\bar{x} = 255.5$	$\sum X_2 = 2107$ $\bar{x} = 263.375$	$\sum X_1^2 = 524648$	$\sum X_2^2 = 558309$

$$S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}}$$

$$\begin{aligned}
 S_1 &= \sqrt{\frac{524648 - \frac{4177936}{8}}{8 - 1}} \\
 &= 18.54
 \end{aligned}$$

$$\begin{aligned}
 S_2 &= \sqrt{\frac{558309 - \frac{4439449}{8}}{8 - 1}} \\
 &= 21.967
 \end{aligned}$$

ESTIMATED ERROR OF DIFFERENCE:

$$\begin{aligned} &\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ &= \sqrt{\frac{343.732}{8} + \frac{482.549}{8}} \\ &= 10.163 \\ &t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{diff}} \\ &= \frac{255.5 - 263.375}{10.163} \\ &= -0.775 \end{aligned}$$

df = (8-1) + (8-1) = 14

t_{cri} = 2.145 at p = 0.05

t_{cri} > t_{cal}, accept null hypothesis
There is no significant relationship between gender and mathematical aptitude.

Table Worked Example 4(d): t-Test of Creativity by Gender Status

Participants	MALE (Y ₁)	FEMALE (Y ₂)	Y ₁ ²	Y ₂ ²
1	165	168	27225	28224
2	173	167	29929	27889
3	155	169	24025	28561
4	163	161	26569	25921
5	169	173	28561	29929
6	165	155	27225	24025
7	143	162	20449	26244
8	165	164	27225	26896
	ΣY ₁ =1298 \bar{x} =162.25	ΣY ₂ =1319 \bar{x} =164.875	ΣY ₁ ² =211208	ΣY ₂ ² =217689

STANDARD DEVIATION:

$$S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}}$$

$$S_1 = \sqrt{\frac{211208 - \frac{1684804}{8}}{8 - 1}}$$

$$= 9.316$$

$$S_2 = \sqrt{\frac{217689 - \frac{1739761}{8}}{8 - 1}}$$

$$= 5.592$$

ESTIMATED ERROR OF DIFFERENCE:

$$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$= \sqrt{\frac{9.316^2}{8} + \frac{5.592^2}{8}}$$

$$= 3.842$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{diff}}$$

$$= \frac{162.25 - 164.875}{3.842}$$

$$= -0.683$$

$$df = (8-1) + (8-1) = 14$$

$$t_{cri} = 2.145 \text{ at } p = 0.05$$

$t_{cri} > t_{cal}$, accept null hypothesis. There is no significant relationship between gender and creativity.

Worked Example 5

Refer to Data Set 3 Graduate Salary for the following exercise (from the CD given).
The data set contains information for the different range of salary for different courses from a particular university.

Question 1:

Find the number of students who graduated in January 2006, June 2006, January 2007 and June 2007.

Question 2

- a) Find the number of males and females in this research and represent the data with a pie chart.
- b) Plot graph of mean for salary versus college (x axis-college, y axis-salary)
- c) Plot graph of salary for males and females (gender), across different colleges.
(x axis-different colleges, gender, y axis-salary)

Question 3

Find the number of males and females in each different college.

Question 4

What is the mean and range of salary earned per year for the 1100 graduates from the University ?

Question 5

Show graphically the normal distribution of the salary of the graduates.

Question 6

Use transform recode and divide the amount of salary earned into 5 different groups as follow:

1= < 10000

2= 10000-20000

3= 20001-30000

4= 30001-40000

5= >40000

Question 7

Which group of salary range has the most graduates?

Question 8

Graduate from which course of the University has the highest mean of salary per annum and which course has the lowest mean of salary per annum?

Question 9

List the number of males and females with salary > 40000.

Question 10

Use t test to examine whether there is a significant difference between genders in the amount of salary.

Suggested Answer

Refer to Data Set 3 Graduate Salary for the following exercise.

The data set contains information for the different range of salary for different courses from a particular university.

Question 1

Find the number of students who graduated in January 2006, June 2006, January 2007 and June 2007.

Graduation Date

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid January 2006	258	23.5	23.5	23.5
June 2006	395	35.9	35.9	59.4
January 2007	224	20.4	20.4	79.7
June 2007	223	20.3	20.3	100.0
Total	1100	100.0	100.0	

Number of students who graduated in January 2006 = 258

Number of students who graduates in June 2006= 395

Number of students who graduate in January 2007= 224

Number of students who graduate in June 2007= 223

Question 2

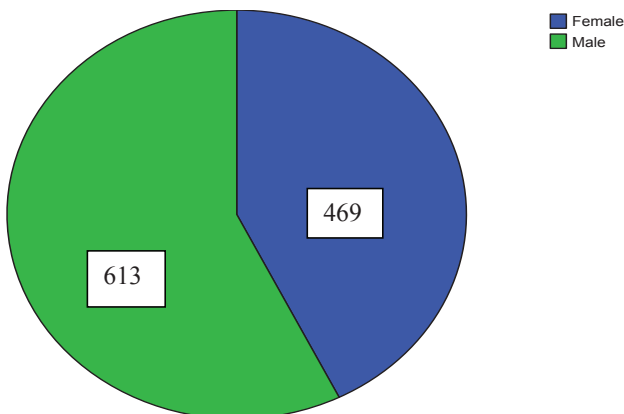
a) Find the number of males and females in this research and represent the data with a pie chart.

Gender

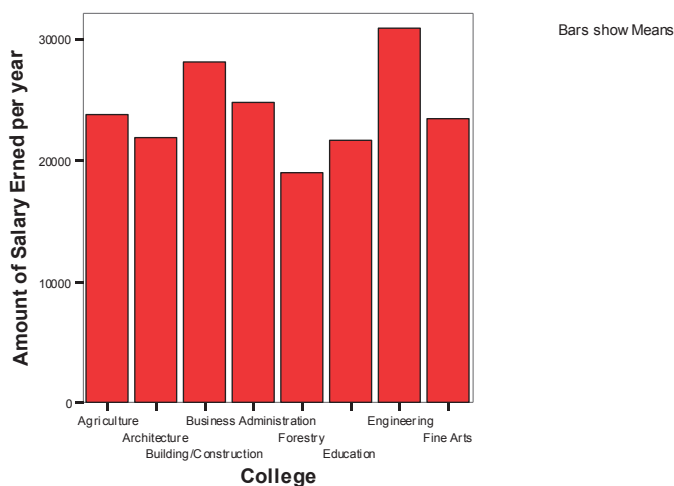
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Female	469	42.6	42.6	42.6
Male	631	57.4	57.4	100.0
Total	1100	100.0	100.0	

- Number of females= 469
- Number of males= 631

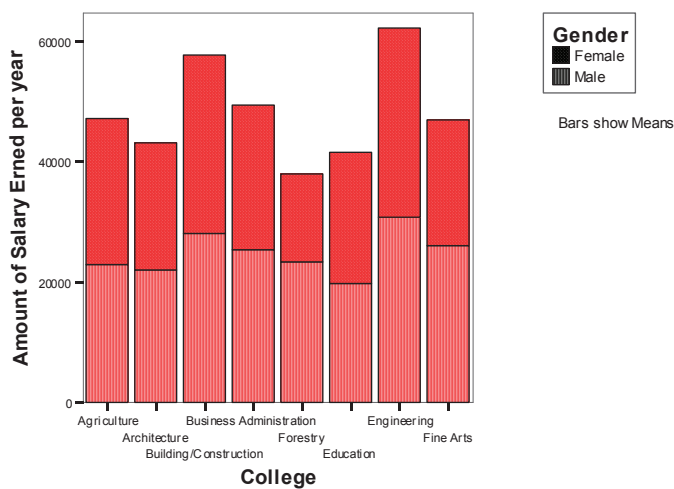
Gender



b) Plot graph of mean for salary versus college (x axis-college, y axis-salary)



c) Plot graph of salary for males and females (gender), across different colleges.
(x axis-different colleges, y axis-salary)



Question 3

Find the number of males and females in each different college.

College * Gender Crosstabulation

Count		Gender		
		Female	Male	Total
College	Agriculture	271	144	415
	Architecture	2	8	10
	Building/Construction	4	51	55
	Business Administration	133	189	322
	Forestry	1	1	2
	Education	12	1	13
	Engineering	45	236	281
	Fine Arts	1	1	2
Total		469	631	1100

Agriculture

Male= 144

Female= 271

Architecture

Male= 8

Female= 2

Building/Construction

Male= 51

Female= 4

Question 4

What is the mean and range of salary earned per year for the 1100 graduates from the University ?

Statistics

Starting Salary

N	Valid	1100
	Missing	0
Mean		26064.20
Range		58300
Minimum		7200
Maximum		65500

Mean for the amount of salary earned = 26064.20

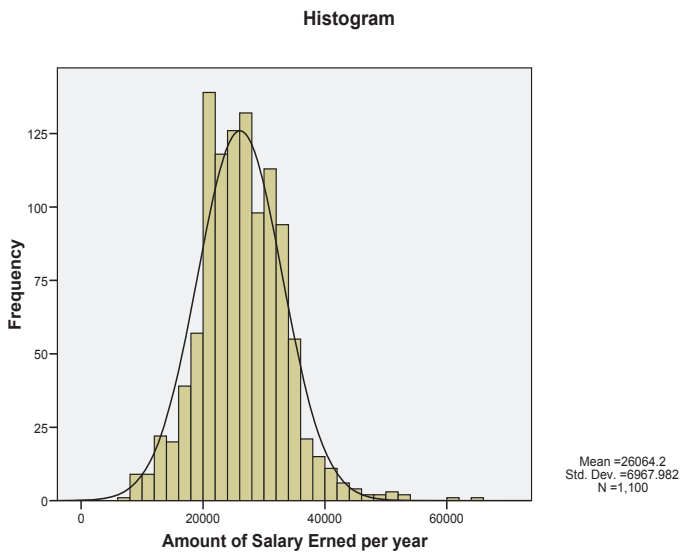
Range of salary= 58300

Question 5

Show graphically the normal distribution of the salary of the graduates.

Question 5

Show graphically the normal distribution of the salary of the graduates.



Question 6

Use transform recode and divide the amount of salary earned into 5 different groups as follow:

- 1= < 10000
- 2= 10000-20000
- 3= 20001-30000
- 4= 30001-40000
- 5= >40000

Question 7

Which group of salary range has the most graduates?

Salary Recode					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 10000	13	1.2	1.2	1.2
	10000-20000	223	20.3	20.3	21.5
	20001-30000	581	52.8	52.8	74.3
	30001-40000	258	23.5	23.5	97.7
	> 40000	25	2.3	2.3	100.0
Total		1100	100.0	100.0	

Most of the graduates have the amount of salary that fell in the range of 20001-30000 per year

Question 8

Graduate from which course of the University has the highest mean of salary per annum and which course has the lowest mean of salary per annum?

Starting Salary				
	N	Mean	Std. Deviation	Std. Error
Agriculture	415	23780.00	7678.715	376.933
Architecture	10	21920.00	5900.810	1866.000
Building/Construction	55	28163.64	3923.168	529.000
Business Administration	322	24814.05	5553.360	309.477
Forestry	2	19000.00	6363.961	4500.000
Education	13	21715.38	5485.108	1521.295
Engineering	281	30876.87	5189.219	309.563
Fine Arts	2	23450.00	3606.245	2550.000
Total	1100	26064.20	6967.982	210.093

Course with the highest mean of salary per annum is engineering.

Course with the lowest mean of salary per annum is forestry.

Question 9

List the number of males and females with salary > 40000.

Gender * salaryRecode Crosstabulation

Count		salaryRecode					Total
		< 10000	10000-20000	20001-30000	30001-40000	> 40000	
Gender	Female	9	120	253	79	8	469
	Male	4	103	328	179	17	631
Total		13	223	581	258	25	1100

Number male = 17

Number of female = 8

Question 10

Use t test to examine whether there is a significant difference between genders in the amount of salary.

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Amount of Salary	Female	469	24769.51	6895.765	318.417
	Male	631	27026.51	6870.097	273.494

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Amount of Salary	Equal variances assumed	.034	.854	-5.380	1098	.000	-2256.996	419.517	-3080.142	-1433.850
	Equal variances not assumed			-5.377	1006.360	.000	-2256.996	419.748	-3080.678	-1433.314

The mean of amount salary for males is 27026.51 and the mean for amount salary for females is 24769.51. An independent t-test revealed that, $t(1098) = -5.380$, $p < .01$. Therefore there is a significant difference between males and females in the amount of salary earned per annum.

Chapter 3 Review Exercises

Question 1

A sample of 600 students about whether they usually buy store brand or name brand products are recorded in the following table.

	Usually buy	
	Store Brand	Name Brand
Male	135	110
Female	212	143

Using the 1% significance level, can you reject the null hypothesis that the two attributes, gender and store or name products, are independent?

Question 2

A survey is being conducted to determine whether the age of a person is related to the type of TV programme he or she watches. A random sample of 700 people gives the data shown here. At $\alpha=0.025$, is the type of TV programme watched related to a person's age?

Age	TV programme			
	Cartoon	Documentary	Comedy	Mystery
1-10	54	8	8	5
11-20	11	33	29	26
21-30	13	27	34	27
31-40	18	25	41	48
41-50	14	24	42	25
51-60	26	26	40	23
61 and above	11	18	34	10

Question 3

A researcher wishes to know whether the way people obtain information is related to their educational background. A random sample of 550 people yielded the following data. At $\alpha=0.05$, test the claim that the way people obtain information is independent of their educational background.

	Internet	TV	Newspaper	Others
Primary	25	46	52	5
Secondary	50	40	63	17
Tertiary	91	61	70	30

Question 4

The table below shows classification of 450 randomly selected workers based on their status as smoker or nonsmoker and on the number of visits they made to the doctor last year.

	Number of visits to the doctor		
	0-2	3-5	> 5
Smoker	30	65	105
Nonsmoker	115	90	45

Test at the 1% significance level if there is any relation between smoking and the number of visits to the doctor.

Question 5

A researcher wishes to know whether the age of the house purchaser is related to the price of the house purchased. A sample of 240 house owners shows the following data. At $\alpha=0.05$, is the price of the house independent of the age of the owner?

	Price		
Age	Below RM150 000	RM150 001–RM300 000	RM300 001 and above
21-30	18	28	2
31-40	46	28	14
41-50	33	18	17
51 and above	11	14	11

Question 6

Olympic Electronics Company buys a type of integrated circuit chips (IC) from two subcontractors, P and Q. The quality control department at this company wanted to check if the distribution of good and defective IC is the same from both subcontractors. The quality control engineer selected a random sample of 250 IC chips from subcontractor P and 350 IC chips from subcontractor Q. These IC chips were checked for being good or defective and recorded in the following table.

	Subcontractor P	Subcontractor Q
Good	232	325
Defective	18	25

Using the 10% significance level, test the null hypothesis that the distribution of good and defective IC chips are the same for both subcontractors.

Question 7

A survey is conducted to see if the instructor's degree is related to the students' opinion of teaching quality. A random sample of 120 students' evaluations of various instructors is shown below. At $\alpha=0.10$, can we conclude that the degree of the instructor is related to students' opinions about the teaching quality?

	Degree		
Rating	Bachelor	Master	Doctorate
Excellent	19	17	9
Average	18	8	13
Poor	7	14	15

Question 8

A survey had been conducted and found out Malaysia people (different age of groups) had different tastes on beverage. The result is shown below :

	Age Groups		
Types of beverage	< 30	30 – 55	>55
Milo	200	140	50
Nescafe	60	56	130
Horlick	40	84	153

At $\alpha=0.10$, can we conclude that there is a relationship between age group and beverage preference?

Question 9

The following data shows the performance of 350 staff in a training program and their job performance.

		Performance in training program		
		Poor	Average	Excellent
Job Performance	Poor	18	55	24
	Average	22	74	53
	Excellent	3	44	57

At $\alpha=0.01$, test the null hypothesis that performance in the training program and job performance are independent.

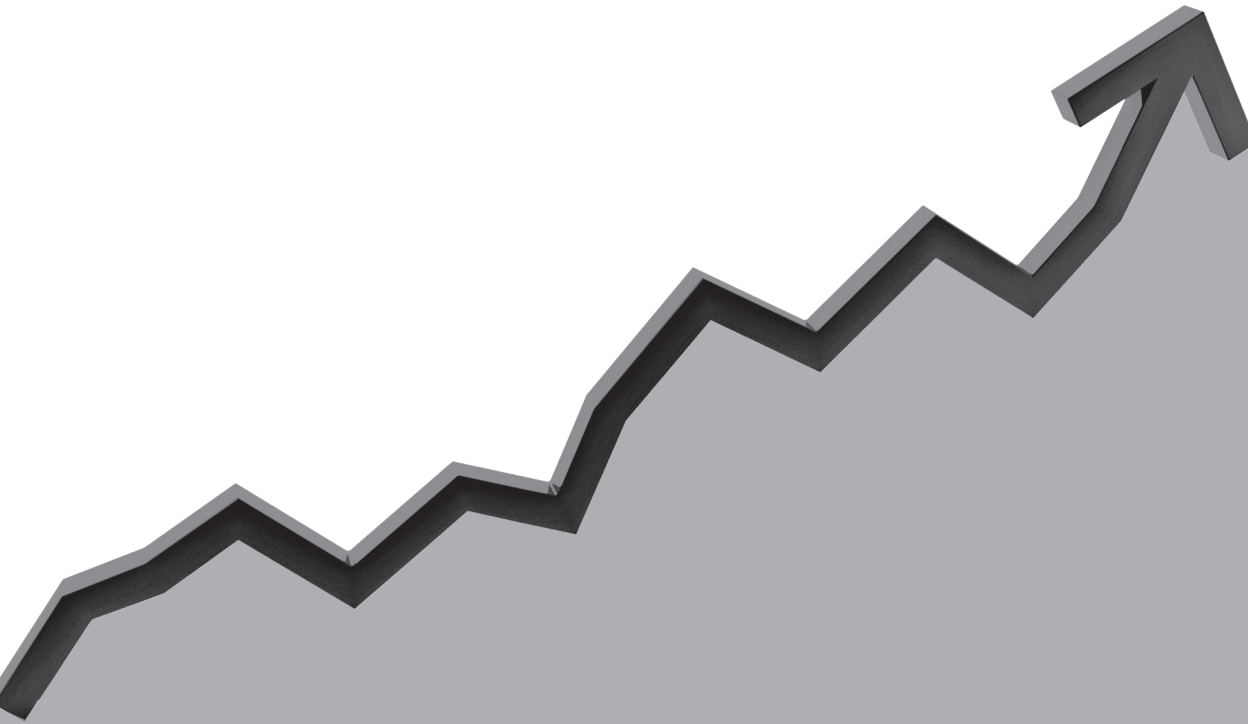
Question 10

An officer in a recreational club is interested to know if there is a relation between the facilities and gender. The table below shows the data collected from 760 club members. What can she conclude by using the data at 2.5% significance level?

Gender	Facilities			
	Tennis court	Swimming pool	Gymnasium	Track
Male	92	138	110	46
Female	88	151	98	37

chapter four

ANOVA



Learning Objectives

At the completion of this chapter, students should be able to:

- Differentiate between-groups and within-groups variance.
- Collect data using paper and pencil methods, and usage of statistical software.
- Analyse differences between three or more conditions using formula and SPSS.
- Interpret the results in a manner that provides answers to research questions.
- Apply the procedure for one-way between-groups ANOVA with post-hoc tests.

4.0 Introduction

Chapter 4 focuses on the nature of statistical data, ordering and manipulation of data, concepts of statistical inference, and hypothesis testing. The aim of the exercises is to ensure that students are able to use the manual calculation and SPSS programme to analyze the statistical theories and calculations. Throughout the discussion, students would be exposed to various data sets, and required to analyze types of data.

4.1 Analysis Of Variance: ANOVA

According to Caldwell (2007), many statisticians think of Analysis of Variance (commonly abbreviated as ANOVA) as an extension of the difference of means test because it is based, in part, on a comparison of sample means. At the same time, however, the procedure involves a comparison of different estimates of population variance—hence the name analysis of variance. Because ANOVA is appropriate for research involving three or more samples, it has wide applicability.

In the field of experimental psychology, for example, results from three or more samples, are often referred to as treatment groups. Imagine an educational psychologist wanting to know if students exposed to three different treatments conditions and learning environments (positive sanction, negative sanction and sanction neutral) exhibit different test scores. Assuming the test scores are based on an interval/ratio scale of measurement, ANOVA would be an appropriate approach to the problem. In short, ANOVA allows the comparison of multiple samples in a single application.

4.2 The Logic of ANOVA

Imagine for a moment if scores in an aptitude test actually vary for students in different types of schooling environment—home schooling, public schooling, and private schooling. The research question will involve a comparison of more than two groups. Assume that the test scores are measured at the interval/ratio level, the situation will be tailor-made for an application of ANOVA. The study could easily be one that asks whether or not aptitude test scores vary on the basis of school environment.

Another way to look at the question is whether or not the type of school environment is a legitimate classification scheme when it comes to the matter of aptitude test scores. If aptitude test scores really do vary on the basis of school environment—if there is a significant difference between the scores in the three environments—then it is probably legitimate to speak in terms of school environments when looking at test scores. If there is no significant difference between the scores, then question the legitimacy of a classification scheme.

The purpose of ANOVA is to measure whether there is more variation between groups than within groups. It examines the legitimacy of a classification scheme.

4.3 From Curves to Data Distributions

In essence, ANOVA allows calculating a ratio of the variation between groups to the variation within groups. This ratio is referred to as the *F* ratio (named after its developer Sir Ronald Fisher). Assume that in search of significant results in a hypothesis-testing situation, the more variation between the means of several groups, the more relative to

the variation within the groups. In short, there will be more variation between than within. Because the F ratio is an expression of the between-to-within ratio, look for a large F value because all factors being equal, the larger the F ratio, the greater the probability that rejects the null hypothesis. Look for more variations between the samples than within the samples to achieve significant results.

4.4 The Different Means

The two types of mean which comes into play in ANOVA is the grand mean and the individual sample means. Each sample mean, or group mean, is a function in part, of the number of cases in the sample.

4.5 From Different Means to Different Types of Variation

The concept of variation typically involves the extent to which various scores in a distribution deviate from the mean of the distribution. The variation between-groups are an expression of the amount of deviation of sample means from the grand mean. The variation within-groups is an expression of the amount of deviation of sample scores from sample means.

4.6 The Null Hypothesis

The null hypothesis simply states that the means of the regions are equal. It can be stated as follows:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

In terms of the F ratio, recall that there has to be more variation between the regions than within the regions for the F ratio to be significant. It all goes back to the notion that the F ratio is an expression of the ratio of the variation between groups to the variation of within groups; the larger the F ratio, the more likely it is to be significant. If all the sample means are equal, there would not be any between-groups variation. Calculate the F ratio (test statistic) as a test statistic (the F ratio) meets or exceeds the critical value; reject the null hypothesis (with a known probability) of having committed a Type 1 error). When the null hypothesis would be rejected, the result showed that there are significance differences between groups. However, it does not indicate where the difference lies. In order to find out where the difference lies, a 'post-hoc test' should be conducted'.

4.7 Post Hoc Test

In order to determine which groups differ from each other, a post-hoc test can be conducted after the initial ANOVA test is completed. However please note that the post-hoc test should only be carried out if only the initial ANOVA problem is significant (rejected the null). If the hypothesis fails to reject the null, then there are no differences to find.

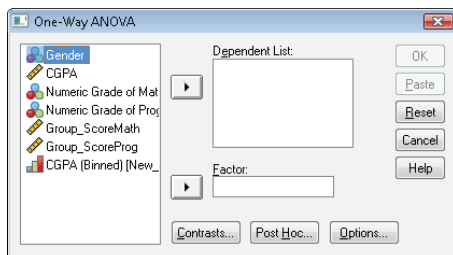
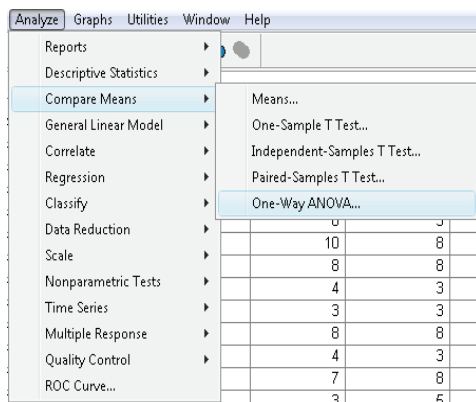
For the Tukey's post-hoc test, the differences between the means of all of our groups will be determined. This difference score will be compared to a critical value to see if the difference is significant. The critical value in this case is the HSD (honestly significant difference) and it must be computed. It is the point when a mean difference becomes honestly significantly different.

$$HSD = q \sqrt{\frac{MS_{wg}}{n}}$$

Note that “ q ” is a table value, and n is the number of values we are dealing with in each group (not total n). MS_{wg} is the mean square error from the overall F-test that from the ANOVA you already computed. The Turkey test is usually recommended in the post hoc test because studies show it has greater power than the other tests under most circumstances and it is readily available in computer packages such as SPSS.

4.8 Procedure for one-way between-groups ANOVA with post-hoc tests using SPSS

- Select **Analyze** and click on **Compare Means**, then on **One-way ANOVA**.
- Click on **dependent variable** (continuous) and move into the box of **Dependent List**.
- Click on your **independent variable** (categorical) and move into the box of **Factor**.
- Click on **Options**, select **Descriptive**, **Homogeneity of Variance test**, **Brown-Forsythe**, **Welsh** and **Means Plot**.
- Click on **Post Hoc** and select **Tukey**
- Click on **Continue** and then **OK** as shown in the following figure



Worked Examples

Worked Example 1

A researcher Dr Tam is interested in seeing if the amount of protein in a poodle's feed influences the rate at which the poodles react to receive the feed as a reward. She makes up three different types of feed: Type A, which has a large amount of protein; Type B, which has a moderate amount of protein; and Type C which has very little protein. The researcher hypothesizes that poodles will react faster for the low protein feed because they need more of it to meet their nutritional requirements.

Table Worked Example 1(a): Different Amount of Protein in Poodle's feed

Type A	Type B	Type C
20	30	20
25	35	35
25	35	40
20	35	40
30	35	45
$\Sigma X^1 = 120$	$\Sigma X^2 = 170$	$\Sigma X^3 = 180$

- a) State the null and research hypothesis.

H_0 = Null hypothesis; $\mu_1 - \mu_2 - \mu_3 = 0$

Null hypothesis: There is no significant difference in the amount of protein in the poodle's feed with the rate of the poodle react to receive the fed as reward.

H_r = Research hypothesis; $\mu_1 - \mu_2 - \mu_3 \neq 0$

Research hypothesis: There is significance difference in the amount of protein in the poodle's feed with the rate of poodle react to receive the feed as reward.

- b) Compute the appropriate statistical test

Table Worked Example 1(b): Different Amount of Protein in Poodle's Feed

Type A, (X_1)	Type A, (X_1) ²	Type B, (X_2)	Type B, (X_2) ²	Type C, (X_3)	Type C, (X_3) ²
20	400	30	900	20	400
25	625	35	1225	35	1225
25	625	35	1225	40	1600
20	400	35	1225	40	1600
30	900	35	1225	45	2025
$\Sigma X_1 = 120$	$\Sigma (X_1)^2 = 2950$	$\Sigma X_2 = 170$	$\Sigma (X_2)^2 = 5800$	$\Sigma X_3 = 180$	$\Sigma (X_3)^2 = 6850$

i. Sum of Square between group SS_{bg}

$$\begin{aligned} SS_{bg} &= \left[\frac{(120)^2}{5} + \frac{(170)^2}{5} + \frac{(180)^2}{5} \right] - \left[\frac{(120+170+180)^2}{5} \right] \\ &= \left[\frac{14400}{5} + \frac{28900}{5} + \frac{32400}{5} \right] - \left[\frac{220900}{5} \right] \\ &= (2880 + 5780 + 6480) - 14726.667 \\ &= 15140 - 14726.667 \\ &= 413.333 \end{aligned}$$

ii. Sum of Square within group SS_{wg}

$$\begin{aligned} SS_{wg} &= [2950 + 5800 + 6850] - \left[\frac{(120)^2}{5} + \frac{(170)^2}{5} + \frac{(180)^2}{5} \right] \\ &= 15600 - \left[\frac{14400}{5} + \frac{28900}{5} + \frac{32400}{5} \right] \\ &= 15600 - (2880 + 5780 + 6480) \\ &= 15600 - 15140 \\ &= 460.00 \end{aligned}$$

iii. Degree of freedom between group df_{bg}

$$\begin{aligned} df_{bg} &= 3-1 \\ &= 2 \end{aligned}$$

iv. Degree of freedom within group df_{wg}

$$\begin{aligned} df_{wg} &= (5-1) + (5-1) + (5-1) \\ &= 4+4+4 \\ &= 12 \end{aligned}$$

v. Mean square between group MS_{bg}

$$\begin{aligned} MS_{bg} &= \frac{413.333}{2} \\ &= 206.667 \end{aligned}$$

vi. Mean square within group MS_{wg}

$$\begin{aligned} MS_{wg} &= \frac{460.00}{12} \\ &= 38.333 \end{aligned}$$

vii. F ratio

$$\begin{aligned} \text{Calculated F} &= \frac{206.667}{38.333} \\ &= 5.391 \end{aligned}$$

As a result a source table can be generated as shown in table below:

Table Worked Example 1(c): Source table

Source	Sum of squares	df	Mean of square	F	P
Between	413.333	2	206.667	5.391	<.01
Within	460.00	12	38.333		
Total	873.331	14			

- c) Refer to Table F^1 , find the critical value of the test statistic for the appropriate degrees of freedom.
Critical F = 3.88
- d) State the statistical conclusion
Calculated F > critical F
Thus, reject null hypothesis, accept H_r
- e) Calculate the post hoc test in order to determine which groups differ each other.

$$\begin{aligned}
 \text{HSD} &= q \times \sqrt{\frac{MS_{wg}}{n}} \\
 &= 3.77 \times \sqrt{\frac{38.333}{5}} \\
 &= 10.439
 \end{aligned}$$

Min differences: $\bar{A} = 24$, $\bar{B} = 34$, $\bar{C} = 36$

$\bar{A} - \bar{C} = 12$ (greater than HSD (10.439), therefore very significant min differences between type A and type C)

$$\bar{A} - \bar{B} = 10$$

$$\bar{B} - \bar{C} = 2$$

- f) Briefly describe the findings

Results indicated a significant difference in the rate of reaction to receive feed depending on the amount of protein in feed A. Tukey HSD post-hoc analysis revealed that the rate of reactions among poodles were significantly faster for type C feed with the least protein compared to type A with the most protein. The rate of reactions for type B feed with the moderate amount of protein did not differ from either Type A or Type C feed.

Worked Example 2

A research study was conducted to examine the clinical efficacy of a new antidepressant. Depressed patients were randomly assigned to one of the three groups: a placebo group, a group that received a low dose of the drug, and a group that received a moderate dose of the drug. After four weeks of treatment, the patients completed the Beck Depression Inventory. The higher the score, the more depressed the patient. The data are presented below. Use this information to solve the following questions:

Table Worked Example 2(a): The Depression Scores of Different Groups

Placebo Dose	Low Dose	Moderate Dose
38	22	14
47	19	26
39	8	11
25	23	18
42	31	5
$\Sigma=191$	$\Sigma=103$	$\Sigma=74$

- a) In the above example, what is a null hypothesis and a research hypothesis can be generated?

Solutions:

Null hypothesis: There will be no difference in depression levels among the three groups.

Research hypothesis: There will be a difference in depression levels among the three levels of drug groups.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3$$

- b) Perform the appropriate statistical test and generate the source table that includes:

- i. Sum of Squares

Solutions:

$$\begin{aligned}
 SS_{bg} &= \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} \right] - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3)^2}{n_{total}} \\
 &= \left[\frac{(191)^2}{5} + \frac{(103)^2}{5} + \frac{(74)^2}{5} \right] - \frac{(191 + 103 + 74)^2}{15} \\
 &= \left[\frac{36481}{5} + \frac{10609}{5} + \frac{5476}{5} \right] - \left[\frac{(368)^2}{15} \right]
 \end{aligned}$$

$$\begin{aligned}
 &= (7296.2 + 2121.8 + 1095.2) - 9028.27 \\
 &= 10513.2 + 9028.27 \\
 &= 1484.93
 \end{aligned}$$

$$\begin{aligned}
 SS_{wg} &= (\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2) - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} \right] \\
 &= (7563 + 2399 + 1342) - \left[\frac{191^2}{5} + \frac{103^2}{5} + \frac{74^2}{5} \right] \\
 &= 11304 + 10513.2 \\
 &= 790.8
 \end{aligned}$$

$$\begin{aligned}
 SS_{total} &= SS_{bg} + SS_{wg} \\
 &= 1484.93 + 790.8 \\
 &= 2275.73
 \end{aligned}$$

ii. Degree of freedoms

Solutions:

$$\begin{aligned}
 df_{bg} &= k - 1 \\
 &= 3 - 1 \\
 &= 2 \\
 df_{wg} &= (n_1 - 1) + (n_2 - 1) + (n_3 - 1) \\
 &= (5 - 1) + (5 - 1) + (5 - 1) \\
 &= 12
 \end{aligned}$$

$$df_{total} = N - 1 = 15 - 1 = 14$$

iii. Mean Squares

Solutions:

$$\begin{aligned}
 MS_{bg} &= \frac{SS_{bg}}{df_{bg}} \\
 &= \frac{1484.93}{2} \\
 &= 742.47
 \end{aligned}$$

$$\begin{aligned}
 MS_{wg} &= \frac{SS_{wg}}{df_{wg}} \\
 &= \frac{790.8}{12} \\
 &= 65.9
 \end{aligned}$$

iv. F ratio

Solutions:

$$\begin{aligned}
 F &= \frac{MS_{bg}}{MS_{wg}} \\
 &= \frac{742.47}{65.9} \\
 &= 11.27
 \end{aligned}$$

Thus, the source table can be generated as shown in table 4.5.

Table Worked Example 2(b): Source Table

Source	Sum of Square	df	Mean square	F	p
Between	1484.93	2	742.47	11.27	<.05
Within	790.8	12	65.9		
Total	2275.73	14			

- c) Refer to Table F, find the critical value for the appropriate degrees of freedom and determine whether the f ratio is significant.

Solutions:

The critical value $F(2, 12, \alpha = 0.05) = 3.88$. Calculated $F(2, 12, \alpha = .05)$
 $11.27 > 3.88$. Therefore reject the null hypothesis

- d) If the F ratio is significantly large for you to reject the null hypothesis, then perform the necessary post-hoc analysis.

Solutions:

By using Tukey's post-hoc test, the HSD can be given as:

$$K = 3, \quad df_{wg} = 3.77$$

$$HSD = q \cdot \sqrt{\frac{MS_{wg}}{n}}$$

$$\begin{aligned}
 HSD &= q \cdot \sqrt{\frac{65.9}{5}} \\
 &= 3.77 (3.63) \\
 &= 13.685
 \end{aligned}$$

The difference between means must be greater than 13.685 to be statistically.

$$\bar{X}_1 - \bar{X}_2 = 38.2 - 20.6 = 17.6^* ; \bar{X}_1 - \bar{X}_3 = 38.2 - 14.8 = 23.4^*$$

$$\bar{X}_2 - \bar{X}_3 = 20.6 - 14.8 = 5.8$$

- e) Briefly interpret the results of this study relative to the research question. Be sure to include the means and standard deviations for each condition in the interpretation.

The means for placebo group, low dose group and moderate dose group are 17.6, 23.4 and 5.8 respectively. The difference between the means for placebo group and low dose group is 17.6 which is greater than the value of HSD (13.685). The difference between the means for placebo group and moderate dose group is 23.4 which is also greater than the HSD of 13.685. The mean difference between the low dose group and the moderate dose group is 5.8 which is lower than the value HSD of 13.685.

Worked Example 3

Dr Smith is a psychologist who works with patients with Down's syndrome. He has designed a study to determine types of rewards which are effective for training his patients. Dr Smith selected four different groups of patients and recorded the number of days he took to teach them the same task, with each group received one of the four types of rewards: Reward A, Reward B, Reward C, and Reward D. The numbers of days for task training were given in the Table Worked Example 3(a).

Table Worked Example 3(a): Types of Rewards with Task Training

Reward A(X_1)	Reward B(X_2)	Reward C(X_3)	Reward D(X_4)
4	7	10	13
4	6	10	13
6	8	14	15
3	8	13	17
2	10	11	16
1	7	12	14

- Generate null and research hypothesis
- Create a source table that includes
 - Sum of Squares between groups and Sum of Squares within groups
 - Degrees of Freedom
 - Mean Square between group and Mean Square within group
 - F ratio
- Do you reject or accept the null hypothesis? Explain briefly.
- Conduct Post Hoc test to determine which levels of the independent variables are significantly different from one another.

Solutions:

Table Worked Example 3(b): Types of Rewards with Task Training

X1	X1 ²	X2	X2 ²	X3	X3 ²	X4	X4 ²
4	16	7	49	10	100	13	169
4	16	6	36	10	100	13	169
6	36	8	64	14	196	15	225
3	9	8	64	13	169	17	289
2	4	10	100	11	121	16	256
1	1	7	49	12	144	14	196
$\sum X1 = 20$	$\sum X1^2 = 82$	$\sum X2 = 46$	$\sum X2^2 = 362$	$\sum X3 = 70$	$\sum X3^2 = 830$	$\sum X4 = 88$	$\sum X4^2 = 1304$

- a) Generate a null and research hypothesis

Solution:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

There are no significant differences between the numbers of days it takes to teach four different groups of patients the same task, with each group receiving one of the four types of rewards.

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

There are significant differences between the numbers of days it takes to teach four different groups of patients the same task, with each group receiving one of the four types of rewards.

- b) Create a source table that includes:

Solution:

- i. Sum of Squares

$$\begin{aligned} SS_{bg} &= \left[\frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} + \frac{(\sum X_4)^2}{n_4} \right] - \left[\frac{(\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4)^2}{N_{total}} \right] \\ &= \left[\frac{(20)^2}{6} + \frac{(46)^2}{6} + \frac{(70)^2}{6} + \frac{(88)^2}{6} \right] - \left[\frac{(20 + 46 + 70 + 88)^2}{24} \right] \\ &= [66.667 + 352.667 + 816.667 + 1290.667] - [2090.667] \\ &= 436.001 \end{aligned}$$

$$\begin{aligned} SS_{wg} &= \left[\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 \right] - \left[\frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} + \frac{(\sum X_4)^2}{n_4} \right] \\ &= [82 + 362 + 830 + 1304] - \left[\frac{(20)^2}{6} + \frac{(46)^2}{6} + \frac{(70)^2}{6} + \frac{(88)^2}{6} \right] \\ &= [2578] - [66.667 + 352.667 + 816.667 + 1290.667] \\ &= 51.332 \end{aligned}$$

$$\begin{aligned} SS_{total} &= SS_{bg} + SS_{wg} \\ &= 436.001 + 51.332 \\ &= 487.333 \end{aligned}$$

- ii. Degrees of Freedom

Solution:

$$df_{bg} = k - 1$$

$$df_{wg} = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1)$$

$$= 4 - 1$$

$$= 3$$

$$= (6-1) + (6-1) + (6-1) + (6-1)$$

$$= 20$$

$$df_{total} = N_{total} - 1$$

$$= 24 - 1$$

$$= 23$$

iii. Mean Square

Solution:

$$MS_{wg} = \frac{SS_{wg}}{df_{wg}}$$

$$MS_{bg} = \frac{SS_{bg}}{df_{bg}}$$

$$= \frac{436.001}{3}$$

$$= 145.334$$

$$= \frac{51.332}{20}$$

$$= 2.567$$

iv. F ratio

Solution:

$$F = \frac{MS_{bg}}{MS_{wg}}$$

$$= \frac{145.334}{2.567}$$

$$= 56.616$$

Thus, the source table can be generated as shown in Table Worked Example 3(c).

Table Worked Example 3(c): Source Table

Source	Sums of Squares	df	Mean Square	F	p
Between	436.001	3	145.334	56.616	< .05
Within	51.332	20	2.567		
Total	487.333	23			

c) P-value and indicate whether to reject or fail to reject the null hypothesis.

Solution:

$$F = \frac{MS_{bg}}{MS_{wg}}$$

From Table F, the critical value

Hence, Computed $F_{(3,20,\alpha=.05)} = 56.616 > 3.10$

Therefore, reject the null hypothesis.

$$= \frac{145.334}{2.567}$$

$$= 56.616$$

$$F_{(3,20,\alpha=.05)} = 3.10$$

- d) If the F ratio is significantly large enough to reject the null hypothesis, then compute an HSD to determine which levels of the independent variable are significantly different from one another.

Solution:

From Table H², value $q_{(4,20)} = 3.96$

$$\begin{aligned} HSD &= q \cdot \sqrt{\frac{MS_{wg}}{n}} \\ &= 3.96 \cdot \sqrt{\frac{2.567}{6}} \\ &= 2.590 \end{aligned}$$

$$\begin{aligned} \bar{X}_1 &= \frac{\sum X_1}{n_1} \\ &= \frac{20}{6} \\ &= 3.333 \end{aligned}$$

$$\begin{aligned} S_1 &= \sqrt{\frac{\sum X_1^2 - \frac{(\sum X_1)^2}{n_1}}{n_1 - 1}} \\ &= \sqrt{\frac{82 - \frac{(20)^2}{6}}{5}} \\ &= 1.751 \end{aligned}$$

$$\begin{aligned} \bar{X}_2 &= \frac{\sum X_2}{n_2} \\ &= \frac{46}{6} \\ &= 7.667 \end{aligned}$$

$$\begin{aligned} S_2 &= \sqrt{\frac{\sum X_2^2 - \frac{(\sum X_2)^2}{n_2}}{n_2 - 1}} \\ &= \sqrt{\frac{362 - \frac{(46)^2}{6}}{5}} \\ &= 1.366 \end{aligned}$$

$$\begin{aligned}\bar{X}_3 &= \frac{\sum X_3}{n_3} \\ &= \frac{70}{6} \\ &= 11.667\end{aligned}$$

$$\begin{aligned}S_3 &= \sqrt{\frac{\sum X_3^2 - \frac{(\sum X_3)^2}{n_3}}{n_3 - 1}} \\ &= \sqrt{\frac{830 - \frac{(70)^2}{6}}{5}} \\ &= 1.633\end{aligned}$$

$$\begin{aligned}\bar{X}_4 &= \frac{\sum X_4}{n_4} \\ &= \frac{88}{6} \\ &= 14.667\end{aligned}$$

$$\begin{aligned}S_4 &= \sqrt{\frac{\sum X_4^2 - \frac{(\sum X_4)^2}{n_4}}{n_4 - 1}} \\ &= \sqrt{\frac{1304 - \frac{(88)^2}{6}}{5}} \\ &= 1.633\end{aligned}$$

*The difference between means must be greater than 2.590 to be statistically different.

$$\bar{X}_1 - \bar{X}_2 = 3.333 - 7.667 = 4.334^*$$

$$\bar{X}_1 - \bar{X}_3 = 3.333 - 11.667 = 8.334^*$$

$$\bar{X}_1 - \bar{X}_4 = 3.333 - 14.667 = 11.334^*$$

$$\bar{X}_2 - \bar{X}_3 = 7.667 - 11.667 = 4.00^*$$

$$\bar{X}_2 - \bar{X}_4 = 7.667 - 14.667 = 7.00^*$$

$$\bar{X}_3 - \bar{X}_4 = 11.667 - 14.667 = 3.00^*$$

Worked Example 4

Dr Siva is a researcher of a beauty and slimming center. He wanted to find out whether there were differences for 3 types of pills for reducing body weight. He gave these pills to 15 participants and then measured the changes for their body weight after one month. The 1st to 5th participants took pill S, 6th to 10th participants took pills H, and participants 11th to 15th took pills E. The weight lost in one month for each participant was recorded as below:

Table Worked Example 4(a): Types of Pills and Weight Lost

Number of pills taken	Weight lost per month (KG)		
	Pills S	Pills H	Pills E
1	2.2	3.0	4.0
2	3.3	4.6	4.7
3	1.0	3.1	5.2
4	4.1	3.1	5.1
5	2.3	3.5	5.7

- Form null and research hypotheses.
- Calculate mean for pill S, H and E.
- Calculate standard deviation for pill S, H and E.
- Find SS_{bg} and SS_{wg}
- Find MS_{bg} and MS_{wg}
- What is the Computed F value?
- What is the critical F value?
- Analyze the results.
- Match your answer with using SPSS.

Answers:

- Null Hypothesis:
There is no significant difference among 3 different types of pills taken by participants towards weight lost in one month.

$$H_0 = \mu_1 = \mu_2 = \mu_3$$

Research Hypothesis:

There is a significant difference among 3 different types of pills taken by participants towards weight lost in one month.

$$H_1 = \mu_1 \neq \mu_2 \neq \mu_3$$

Table Worked Example 4(b): Types of Pills and Weight Lost

Number of pills taken	Weight lost per month (KG)					
	Pills S	Pills H	Pills E	X_1^2	X_2^2	X_3^2
1	2.2	3.0	4.0	4.84	9.00	16.00
2	3.3	4.6	4.7	10.89	21.16	22.09
3	1.0	3.1	5.2	1.00	9.61	27.04
4	4.1	3.1	5.1	16.81	9.61	26.01
5	2.3	3.5	5.7	5.29	12.25	32.49
TOTAL	12.9	17.3	24.7	38.83	61.63	123.63

b)

$$\text{Mean of pills S, } \bar{X}_1 = \frac{12.9}{5} = 2.58,$$

$$\text{Mean of pills H, } \bar{X}_2 = \frac{17.3}{5} = 3.46,$$

$$\text{Mean of pills E, } \bar{X}_3 = \frac{24.7}{5} = 4.94$$

c)

Standard Deviation of pills S,

$$S_1 = \sqrt{\frac{\sum X_1^2 - \frac{(\sum X_1)^2}{n_1}}{n_1 - 1}} = \sqrt{\frac{38.83 - \frac{12.9^2}{5}}{4}} = \sqrt{1.387} = 1.1777$$

Standard Deviation of pills H,

$$S_2 = \sqrt{\frac{\sum X_2^2 - \frac{(\sum X_2)^2}{n_2}}{n_2 - 1}} = \sqrt{\frac{61.63 - \frac{17.3^2}{5}}{4}} = \sqrt{0.443} = 0.6656$$

Standard Deviation of pills E,

$$S_3 = \sqrt{\frac{\sum X_3^2 - \frac{(\sum X_3)^2}{n_3}}{n_3 - 1}} = \sqrt{\frac{123.63 - \frac{24.7^2}{5}}{4}} = \sqrt{0.403} = 0.6348$$

d)

$$\begin{aligned}
 SS_{bg} &= \left[\frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} \right] - \left[\frac{(\sum X_1 + \sum X_2 + \sum X_3)^2}{N_{total}} \right] \\
 &= \left[\frac{12.9^2}{5} + \frac{17.3^2}{5} + \frac{24.7^2}{5} \right] - \left[\frac{(12.9 + 17.3 + 24.7)^2}{15} \right] \\
 &= [33.282 + 59.858 + 122.018] - 200.934 \\
 &= 14.224
 \end{aligned}$$

$$\begin{aligned}
 SS_{wg} &= \left[\sum X_1^2 + \sum X_2^2 + \sum X_3^2 \right] - \left[\frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} \right] \\
 &= [38.83 + 61.63 + 123.63] - \left[\frac{12.9^2}{5} + \frac{17.3^2}{5} + \frac{24.7^2}{5} \right] \\
 &= 224.09 - (33.282 + 59.858 + 122.018) \\
 &= 224.09 - 215.158 \\
 &= 8.932
 \end{aligned}$$

$$\begin{aligned}
 SS_{total} &= SS_{bg} + SS_{wg} \\
 &= 14.224 + 8.932 \\
 &= 23.156
 \end{aligned}$$

e) In order to obtain Mean square, degree of freedom must be acquired initially as shown below:

Degree of freedom, $df_{bg} = k - 1 = 3 - 1 = 2$,

$$df_{wg} = (n_1 - 1) + (n_2 - 1) + (n_3 - 1)$$

Degree of freedom, $= 4 + 4 + 4$
 $= 12$

Degree of freedom, $df_{total} = N - 1 = 15 - 1 = 14$

$$\text{Mean square, } MS_{bg} = \frac{SS_{bg}}{df_{bg}} = \frac{14.224}{2} = 7.112$$

$$\text{Mean square, } MS_{wg} = \frac{SS_{wg}}{df_{wg}} = \frac{8.932}{12} = 0.7443$$

f)

$$F = \frac{MS_{bg}}{MS_{wg}} = \frac{7.112}{0.7443}$$

$$= 9.555$$

Thus, the source table can be generated as shown in Table Worked Example 4(c).

Table Worked Example 4(c): The source table

Source	Sums of Square	Df	Mean Square	F	p
Between	14.224	2	7.112	9.555	.05
Within	8.932	12	0.744		
Total	23.156	14			

g) From table F, the critical $F_{(2,12,\alpha=.05)} = 3.88$,

h) Calculated $F_{(2,12,\alpha=.05)} 9.555 > 3.88$, hence, null hypothesis being rejected and continue with Tukey HSD post-hoc analysis.

$$HSD = q \cdot \sqrt{\frac{MS_{wg}}{n}}$$

$$= 3.77 \cdot \sqrt{\frac{0.7443}{5}}$$

$$= 3.77 \cdot \sqrt{0.14886}$$

$$= 3.77 \cdot 0.3858$$

$$= 1.4545$$

From table H, $q_{(3,12)} = 3.77$

The difference between means must be greater than 1.4545 to be statistically different.

- $\bar{X}_1 - \bar{X}_2 = 2.58 - 3.46 = -0.88$
- $\bar{X}_1 - \bar{X}_3 = 2.58 - 4.94 = -2.36^*$
- $\bar{X}_2 - \bar{X}_3 = 3.46 - 4.94 = -1.48^*$

Conclusion:

There was a significant differences in weight lost in a month by taking 3 types of diet pills by participants ($F_{(2,12)} = 9.555, p < .05$). Tukey HSD post-hoc analysis (HSD= 1.4545) revealed that participants significantly lost more weight by taking pills $E_{(\bar{X}_3=4.94, SD_3=0.6348)}$ and pills $H_{(\bar{X}_2=3.46, SD_2=0.6656)}$ compared to pills $S_{(\bar{X}_1=2.58, SD_1=1.1777)}$. There was no significant difference in the weight lost in one month by taking pills H and pills E.

Output generated in SPSS :

ANOVA

decrease weight in one month

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	14,224	2	7,112	9,555	,003
Within Groups	8,932	12	,744		
Total	23,156	14			

Multiple Comparisons

decrease weight in one month

Tukey HSD

(I) diet pills	(J) diet pills	Mean Difference (I- J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
mS	mH	-,88000	,54565	,278	-2,3357	,5757
	mE	-2,36000*	,54565	,003	-3,8157	-,9043
mH	mS	,88000	,54565	,278	-,5757	2,3357
	mE	-1,48000*	,54565	,046	-2,9357	-,0243
mE	mS	2,36000*	,54565	,003	,9043	3,8157
	mH	1,48000*	,54565	,046	,0243	2,9357

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

decrease weight in one month

Tukey HSD

diet pills	N	Subset for alpha = 0.05	
		1	2
mS	5	2,5800	
mH	5	3,4600	
mE	5		4,9400
Sig.		,278	1,000

Means for groups in homogeneous subsets are displayed.

Worked Example 5

A company invented some new fertilizers and the scientists want to test the effects of the different fertilizers on growth of tomato plants after the fertilizers are applied to the tomato plants. The fertilizers were applied to the plants once a week. The heights of tomato plants are measured after 4 months of applying the fertilizers. The following table shows that the results of the growth of tomato plants.

Table Worked Example 5: Types of Fertilizers and Height of Plants

Types of fertilizers	Initial height of plant (inch)	Final height of plant (inch)
Fertilizer X	3	74
Fertilizer X	4	68
Fertilizer Y	2	76
Fertilizer Y	4	80
Fertilizer Z	3	87
Fertilizer Z	7	91

- What is the average and range of the height of the tomato plants after 4 months?
- Is the fertilizer effective in the growth of the tomato plants? (t-test)
- Carry out a one way ANOVA to find out whether the type of fertilizers influences the height of the tomato plant after 4 months. Give mean and standard deviation for each types of fertilizers.
- Which types of fertilizer will you choose if you were to make sure you have the highest yield of tomatoes?

Solutions:

a)

Descriptive Statistics**Statistics**

Final Height

N	Valid	6
	Missing	0
Mean		79.33
Range		23
Minimum		68
Maximum		91

The average height of the tomato plants is 79.33.

The range of the height of the tomato plant = $91 - 68 = 23$

b)

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 Final Height	79.33	6	8.524	3.480
Initial Height	3.83	6	1.722	.703

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Final Height - Initial Height	75.500	7.740	3.160	67.378	83.622	23.895	5	.000

The mean of the initial height of the plant is 3.83 and the final height of the plant is 79.33.

Pair-sample t-test revealed that there is a significant difference between the initial and final height of the plant. $F(5)=23.893$, $p<.01$ ($p<.00$ means that means there is 0% chance of t-test result occurs by sampling error by chance.)

c)

Descriptive Table

Final Height

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
fertilizer X	2	71.00	4.243	3.000	32.88	109.12	68	74
fertilizer Y	2	78.00	2.828	2.000	52.59	103.41	76	80
fertilizer Z	2	89.00	2.828	2.000	63.59	114.41	87	91
Total	6	79.33	8.524	3.480	70.39	88.28	68	91

ANOVA

Final Height

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	329.333	2	164.667	14.529	.029
Within Groups	34.000	3	11.333		
Total	363.333	5			

Presenting the results from one-way repeated measures ANOVA

Descriptive Statistics for Height of Tomato with Fertilizer X, Y and Z

Type of Fertilizer	N	Mean	Standard Deviation
Fertilizer X	2	71.00	4.243
Fertilizer Y	2	78.00	2.828
Fertilizer Z	2	89.00	2.828

The mean height and standard deviation of tomato plants after using fertilizer X is 71.00 and 4.243 .

The mean height and standard deviation of tomato plants after using fertilizer Y is 78.00 and 2.828.

The mean height and standard deviation of tomato plants after using fertilizer Z is 89.00 and 2.828.

The result of ANOVA indicates that there is a significant difference between the types of fertilizers used and the final height of the tomato plants. $F(2,3)=14.529$, $p<.05$.

Multiple Comparisons

Dependent Variable: Final Height
Tukey HSD

(I) Fertilizer	(J) Fertilizer	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
fertilizer X	fertilizer Y	-7.000	3.367	.241	-21.07	7.07
	fertilizer Z	-18.000(*)	3.367	.026	-32.07	-3.93
fertilizer Y	fertilizer X	7.000	3.367	.241	-7.07	21.07
	fertilizer Z	-11.000	3.367	.092	-25.07	3.07
fertilizer Z	fertilizer X	18.000(*)	3.367	.026	3.93	32.07
	fertilizer Y	11.000	3.367	.092	-3.07	25.07

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

Post Hoc comparison using the Tukey honestly significant difference test indicates that there is a significant difference (in terms of mean of the plants' height and types of fertilizer used) between fertilizer X and Y, mean difference = 18.

d)

Descriptive statistics shows that fertilizer Z has most effective effect on the growth of tomato plants. Mean =89.00.

Question 1

Eighteen students were randomly assigned to three groups to experiment with three different methods of teaching mathematics. Their test scores are shown in the table below. At 1% significance level, can we reject the null hypothesis that the mean mathematics score of all students taught by each of these three methods is the same? Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Method I	Method II	Method III
46	90	86
74	67	64
49	72	98
61	83	76
72	60	67
85	57	77

Question 2

Three brands of laptops are selected, and the number of defects in each is as recorded below. At $\alpha=0.01$, can one conclude that there is a difference in the means of the number of defects for the three groups? Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Brand P	Brand Q	Brand R
2	1	0
0	0	0
0	3	0
0	5	1
2	1	4
3	2	3
0	0	2
1	4	0
2	1	0
4	0	1
0	6	1

Question 3

A research was conducted to compare the time taken by graduates with four different majors to find their first job after graduation. The table below lists the time (in days) the graduates in the year 2009 taken to find their first job. At the 5% significance level, can you conclude that the mean time taken to find their job for all 2009 graduates in these fields is the same? Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Law	Business	IT	Engineering
25	25	90	28
33	100	83	39
40	70	30	60
52	89	55	45
60	68	60	35
74	77	68	75
35	91	72	30
50	120	50	52
29	65	78	80

Question 4

A researcher lives in a house equipped with a solar electric system. He collected voltages readings from the meter connected to the system and recorded in the table as shown below. Test the claim at $\alpha=0.05$, can we conclude that sunny days result in greater amounts of electrical energy? Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Rainy Days	Cloudy Days	Sunny Days
12.0	12.5	13.7
12.3	12.8	13.2
12.5	12.7	13.0
11.8	13.2	14.2
12.6	13.0	13.8
11.3	12.9	14.1
12.2	13.1	13.9
11.6	12.4	12.9

Question 5

A researcher wants to see whether there is any difference in the weight gains of kids following one of three special diets. Kids are randomly assigned to three groups and placed on the diet for 8 weeks. The weight gains (in kg) are shown in the table below. At $\alpha=0.05$, can the researcher conclude that there is a difference in the diets? Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Diet X	2	5	8	6	1	4
Diet Y	9	11	10	13	7	5
Diet Z	9	4	3	4	6	2

Question 6

Operators are randomly assigned to four machines on a production line. The number of defective parts produced by each operator for one week is recorded as shown in the table below. At $\alpha=0.05$, test the claim that the mean number of defective parts produced by the operators is the same. Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Machine A	Machine B	Machine C	Machine D
1	2	9	8
3	3	8	16
1	4	10	0
4	7	12	3
6	7	15	1
5	9	11	0
3	1	18	2

Question 7

Three different vacation packages are given to randomly selected patients in an effort to reduce their stress levels. A special instrument has been designed to measure the percentage of stress reduction in each person as shown in the table below. At $\alpha=0.05$, can one conclude that there is a difference in the means of the percentages? Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Package A	Package B	Package C
5	14	15
12	14	15
7	19	16
3	11	13
15	16	21
5	7	19
4	12	14
6	10	9

Question 8

A researcher wishes to try three different techniques to lower the cholesterol level of patient diagnosed with high cholesterol level. The subjects are randomly assigned to three groups as shown in the table below. After two months the reduction in each person's cholesterol level is recorded. At $\alpha=0.05$, test the claim that there is no difference among the means. Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Medication	Exercise	Diet
1	0.6	0.5
1.2	0.8	0.9
0.9	0.3	1.2
1.5	0	0.8
1.3	0.2	0.4

Question 9

The research department of a bank is established to observe various employees for their work productivity. A recent research is conducted to check whether the four tellers at the bank serve, on average, the same number of customers per hour. The data are recorded as follows. At the 5% significance level, test the null hypothesis that the mean number of customers served per hour by each of these four tellers is the same. Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Teller A	Teller B	Teller C	Teller D
12	12	9	18
15	10	11	20
13	13	8	16
11	14	10	19
14	17	12	13
16	17	15	11
15	9	11	9
13	10	18	12

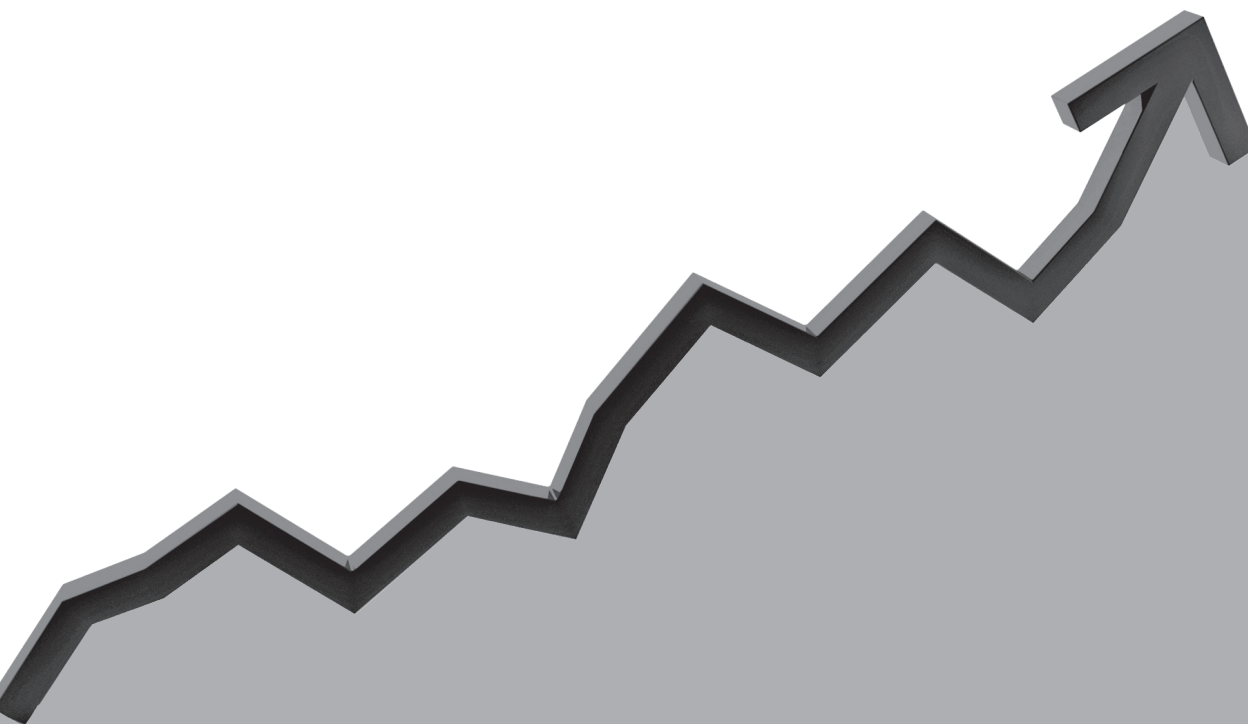
Question 10

The GPA of students joining co-curricular society in a university are compared. The data are shown here. At $\alpha=0.10$, can one conclude that there is a difference in the mean GPA of the three groups? Assume that all the assumptions required to apply the one-way ANOVA procedure hold true.

Chess club	Engineering Society	History Club
1.9	3.4	2.5
2.5	3.9	2.0
2.5	2.7	1.7
2.7	3.2	2.4
3.3	3.8	3.4
3.5	2.5	2.8

chapter five

CROSSTABS PROCEDURE



Learning Objectives

At the completion of the chapter, student should be able to:

- Use the procedure of Crosstabs to construct a cross tabulation table for describing relationships between combinations of variable.
- Request the percentage into the table.
- Analyze the strength of the relationship from the results of chi-squares test of independence.
- Interpret the results in a manner that provide answers to research questions.

5.0 Introduction

This chapter focuses on how to compare groups when the outcome is categorical (nominal or ordinal) by using SPSS. The aim of the series of exercise is to ensure the student can summarize and interpret at the relationship between the variables.

Crosstabs procedure is used as a statistical measurement to describe and examine the relationship between two categorical variables via a cross tabulation table. A cross tabulation table which also known as contingency table, is a co-frequency table of counts, where each row or column is frequency value of one variable for observation falling within in the same category of the other variable. In survey work, a cross tabulation table can serve in two main purposes; descriptive and relationship inferences. Descriptive is normally used to provide some useful information about the survey data such as demographic information on courses of a faculty or student grades of a subject.

Crosstabs procedure is also used to conclude the relationships among the variables. In order to make such inference of relationship, a set of hypothesis must be constructed first and statistical test using chi-square test of independence is applied into the procedure. A lot of surveys require the relationship inferences in their analysis. Because of that, crosstabs procedure is used for both of purposes.

In this chapter, we first need to understand about the type of data. In SPSS we have to identify the data whether it is scale or in categorical form to ensure that we choose the right procedure to analyze. Next, we describe the use of Chi-square through a manual calculation. Later, by using a student data set, we perform the crosstabs procedure for testing the relationship among the variables.

5.1 Understand the categorical data

In general, there are two main types of variables for statistical analysis; scale and categorical. Most of the scale variable in survey work is in interval measurement which means a unit increase in numeric value represents the same changes in quantity. Statistical technique such as regression and analysis of variance assume that the dependent (or outcome) measure is measured on an interval scale. For example income in ringgit Malaysia(RM) or age in years.

Categorical variable can be easily determined when each data represents an identified group. In SPSS, categorical variable can be defined either nominal or ordinal. Nominal is where each numeric value represents a category or group identifier only. The categories have no underlying any numeric value or ranking value. For example, gender can be coded as 1 (Female) and 0 (Male) or vice versa. Ordinal is the data value that represent ranking or ordering information. An example would be specifying how satisfy you are with your internet service provider, coded 1(Not Satisfy), 2(Satisfy), and 3(Very Satisfy).

5.2 Chi-square (χ^2) test of independence: Direction and Strength

Chi-square (χ^2) test of independence is used for deciding whether the hypothesis of independence between difference variable is tenable (Hamburg and Young 1994). The test compares the observed frequencies in different categories with the expected frequencies from the hypothesis.

Statistically χ^2 is defined as:

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t}$$

where

f_o = is an observed frequency

f_t = is a expected frequency

and χ^2 test is used to provide an answer under the hypothesis of independence which may be stated in general as follows:

H_0 : There is independency between observed variables

H_1 : There is not independency between observed variables

The general nature of the test is best explained with the further example. However before continue the tests there are few assumptions need to be concerned:

1. Sample is randomly selected from the population.
2. All observations are independent.
3. Limited to nominal data
4. No more than 20% of the cells have an expected frequency less than 5

5.3 A Chi-Square example

Let consider the example below:

Do men or women differ in what they are willing to give up in order to keep the Internet hook-up?

The results are shown in Table 5.1. This type of table refers as a cross tabulation table or contingent table. Based on the number of rows and columns, the table is called as a two-by-three (often written 2 x 3) cross tabulation table. In general, in a $r \times c$ contingency table, where r denotes the number of rows and c denotes the number of columns, there are $r \times c$ cells.

The χ^2 test consists of calculating observed and expected frequencies under the hypothesis of independence. Thus, the hypotheses being tested in this problem may be stated as follows:

H_0 : Gender is independent on what they are willing to give up in order to keep internet hook-up

H_1 : Gender is not independent on what they are willing to give up in order to keep internet hook-up

Table 5.1: Cross Tabulation Table

	Morning Coffee	Cable TV	Newspaper	Total
Men	87	73	66	226
Women	113	77	84	274
Total	200	150	150	500

Based on the χ^2 test formula, we are interested in determining f_o and f_t . In this case the value of f_o and f_t initially can be illustrated as shown in table 5.2.

Table 5.2 : Cross Tabulation Table

	Morning Coffee	Cable TV	Newspaper	Total
Men	$f_o = 87$ $f_t =$	$f_o = 73$ $f_t =$	$f_o = 66$ $f_t =$	226
Women	$f_o = 113$ $f_t =$	$f_o = 77$ $f_t =$	$f_o = 84$ $f_t =$	274
Total	200	150	150	500

The value of f_t can be calculated as follows:

$$f_t = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}$$

The number of degree of freedom in the cross tabulation table must be determined in order to apply the χ^2 test. The number of degree of freedoms can be calculated as follows:

$$df = (\text{number of rows} - 1) \cdot (\text{number of columns} - 1)$$

Chi-Square Solution

From above formulas, we can construct the solution as follows:

Table 5.3: Cross Tabulation Table

	Morning Coffee	Cable TV	Newspaper	Total
Men	$f_o = 87$ $f_t = 90.4$	$f_o = 73$ $f_t = 67.80$	$f_o = 66$ $f_t = 67.80$	226
Women	$f_o = 113$ $f_t = 109.60$	$f_o = 77$ $f_t = 82.20$	$f_o = 84$ $f_t = 82.20$	274
Total	200	150	150	500

$$\text{Calculated } \chi^2 = .1279 + .3988 + .0478 + .1055 + .329 + .0394 = 1.0484$$

$$\begin{aligned} \text{Calculated } \chi^2 &= 1.048. \\ \text{degrees of freedom } df &= k - 1 = 2 \end{aligned}$$

We now need to find the critical value χ^2 for comparing with the calculated χ^2 . By refer χ^2 distribution for $df = 2$, we can find a critical value at the 0.05 level of significance as follows:

$$\text{Critical Value } \chi^2 (2, .05) = 5.99$$

Then we can state the decision rule for this problem as (Hamburg and Young 1994):

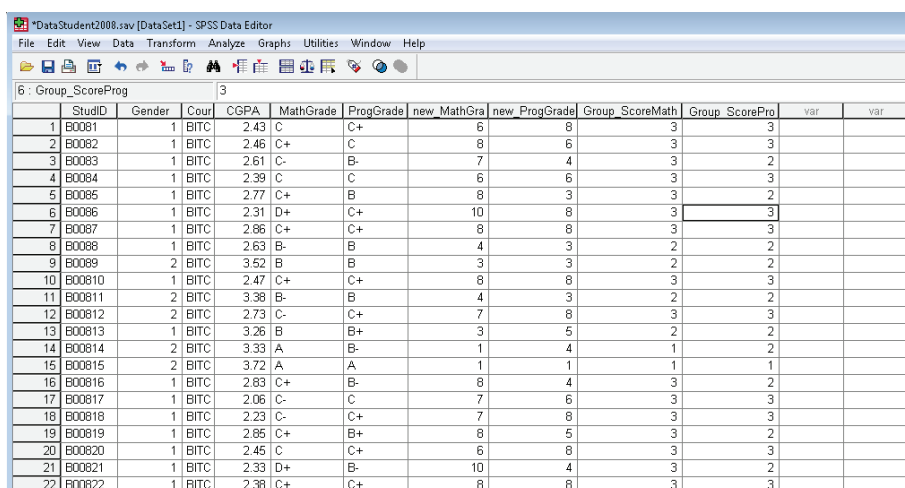
$$\begin{aligned} \text{If } \chi^2 > 5.99, & \text{ reject } H_0 \\ \text{If } \chi^2 < 5.99, & \text{ do not reject } H_0 \end{aligned}$$

In this problem, calculated $\chi^2 < \text{critical value } \chi^2$ ($1.0484 < 5.99$). Thus, it failed to reject the null hypothesis (accept the null hypothesis). So we can conclude that there were no significant differences between men and women in their willingness to give up morning coffee, cable television, or the newspaper to keep the Internet.

5.4 Crosstab procedure using SPSS

In SPSS, crosstabs is one of analysis under descriptive statistics. To understand how we can use crosstabs procedure in SPSS, let us consider the problem below:

A lecturer wishes to determine whether the number of students in each course was influenced of the gender of student. The lecturer also wants to determine the effect of gender on the CGPA result. She obtained 215 student data in a year. The data can be found in *DataStudent2008.sav* as shown in Figure 5.1.



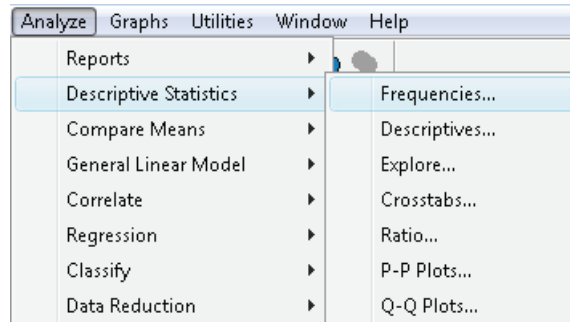
	StudID	Gender	Cour	CGPA	MathGrade	ProgGrade	new_MathGra	new_ProgGrade	Group_ScoreMath	Group_ScorePro	Var1	Var2
1	B0081	1	BITC	2.43	C	C+	6	8	3	3		
2	B0082	1	BITC	2.46	C+	C	8	6	3	3		
3	B0083	1	BITC	2.61	C-	B-	7	4	3	2		
4	B0084	1	BITC	2.39	C	C	6	6	3	3		
5	B0085	1	BITC	2.77	C+	B	8	3	3	2		
6	B0086	1	BITC	2.31	D+	C+	10	8	3	3		
7	B0087	1	BITC	2.86	C+	C+	8	8	3	3		
8	B0088	1	BITC	2.63	B-	B	4	3	2	2		
9	B0089	2	BITC	3.52	B	B	3	3	2	2		
10	B0090	1	BITC	2.47	C+	C+	8	8	3	3		
11	B0091	2	BITC	3.38	B-	B	4	3	2	2		
12	B0092	2	BITC	2.73	C-	C+	7	8	3	3		
13	B0093	1	BITC	3.26	B	B+	3	5	2	2		
14	B0094	2	BITC	3.33	A	B-	1	4	1	2		
15	B0095	2	BITC	3.72	A	A	1	1	1	1		
16	B0096	1	BITC	2.83	C+	B-	8	4	3	2		
17	B0097	1	BITC	2.06	C-	C	7	6	3	3		
18	B0098	1	BITC	2.23	C-	C+	7	8	3	3		
19	B0099	1	BITC	2.85	C+	B+	8	5	3	2		
20	B0100	1	BITC	2.45	C	C+	6	8	3	3		
21	B0101	1	BITC	2.33	D+	B-	10	4	3	2		
22	B0102	1	BITC	2.38	C+	C+	8	8	3	3		

Figure 5.1: *DataStudent2008.sav* dataset

To start analyzing, first it is recommended to explore the number of occurrences or frequencies cases in each variable using **Descriptive Frequencies**.

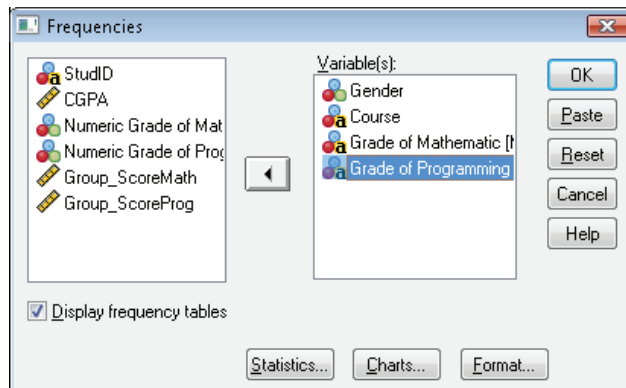
Procedure for conducting **Descriptive Frequencies** :

- From the Analyze menu click on **Descriptive Statistics**
- Select **Frequencies** as shown in the following figure



In the **Frequencies** dialog box :

- Select the variable(s) you require and move the variable into the **Variable(s):** box
- Then click OK as shown in the following figure



Frequencies Output

We now examine the frequencies output in the **Output Viewer**. The outputs begin with frequencies table for **Gender**.

Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MALE	108	50.2	50.2	50.2
	FEMALE	107	49.8	49.8	100.0
Total		215	100.0	100.0	

By default, value labels appear in the first column of the table. The **Frequency** column contains counts or the number of occurrences of each variable. The **Percent** column shows the percentage of cases in each group relative to the number cases in the entire data set including those with missing values. The **Valid Percent** column contains the percentage of cases in each group without missing (non-missing). The **Cumulative Percent** contains cumulative percentages without containing a missing value. In our case, the variable gender contains non-missing value which means, the Percent and Valid Percent are similar. By examine the frequencies of gender variable, we might be understand that MALE and FEMALE have slightly equal in sample size.

Since we also select other variables in this frequencies analysis, we can also examine the frequencies of *Course*, the student grade in mathematics and programming subjects as shown in the outputs below :

Course					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	BITC	47	21.9	21.9	21.9
	BITD	48	22.3	22.3	44.2
	BITI	38	17.7	17.7	61.9
	BITM	41	19.1	19.1	80.9
	BITS	41	19.1	19.1	100.0
	Total	215	100.0	100.0	

Grade of Mathematic

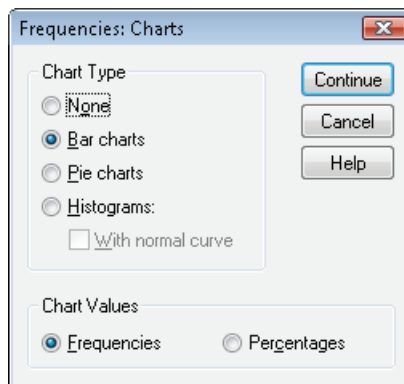
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A-	13	6.0	6.0	6.0
	A	11	5.1	5.1	11.2
	B-	34	15.8	15.8	27.0
	B	29	13.5	13.5	40.5
	B+	18	8.4	8.4	48.8
	C-	16	7.4	7.4	56.3
	C	26	12.1	12.1	68.4
	C+	39	18.1	18.1	86.5
	D	11	5.1	5.1	91.6
	D+	11	5.1	5.1	96.7
	E	7	3.3	3.3	100.0
	Total	215	100.0	100.0	

Grade of Programming

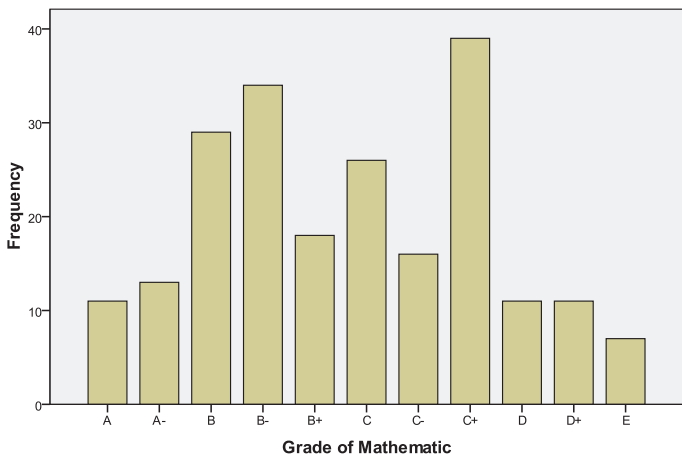
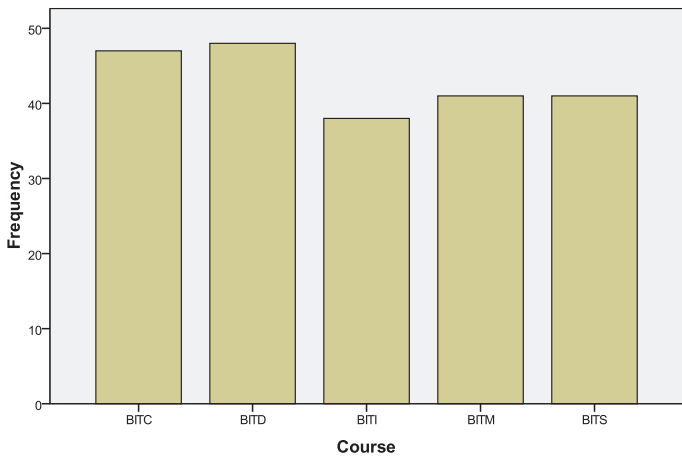
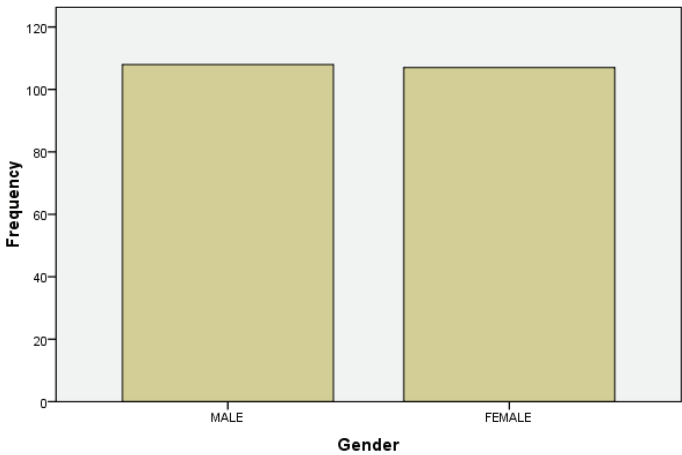
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A-	15	7.0	7.0	7.0
	A	18	8.4	8.4	15.3
	B-	46	21.4	21.4	36.7
	B	34	15.8	15.8	52.6
	B+	25	11.6	11.6	64.2
	C-	11	5.1	5.1	69.3
	C	27	12.6	12.6	81.9
	C+	35	16.3	16.3	98.1
	D	1	.5	.5	98.6
	D+	3	1.4	1.4	100.0
	Total	215	100.0	100.0	

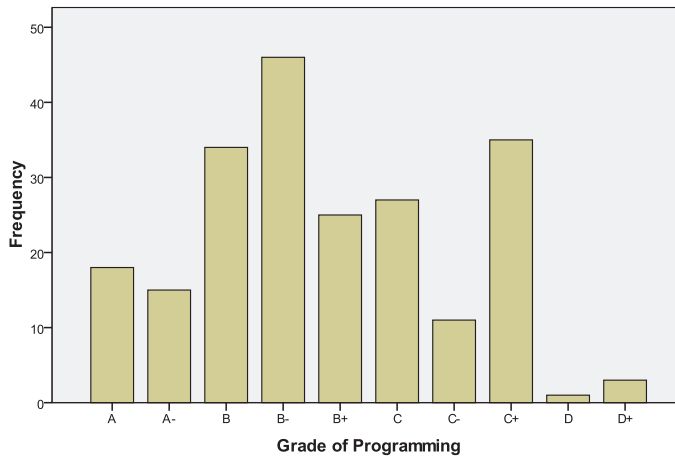
The frequency analysis also can be presented in graphical form. For categorical variables, we can use bar charts.

- Reselect the **Descriptive Statistics**....Click on **Frequencies** procedure
- In Frequencies dialog box, click on **Charts**...
- In **Frequencies: Charts** sub-dialogbox, select the Bar chart radio button.
- Click on **Continue** and then OK as shown in the following figure



Then, the output viewer shows the bar charts as follow:



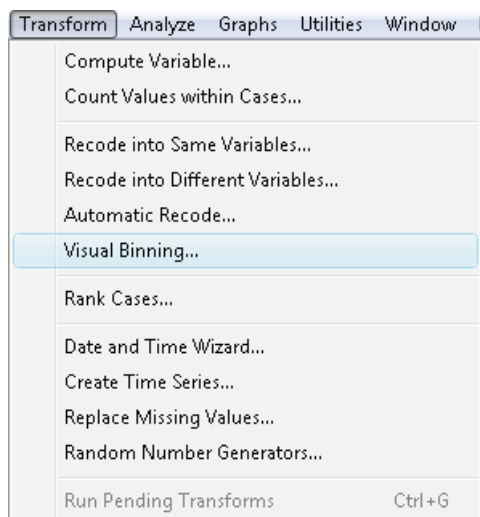


5.5 Modifying data from scale to categorical

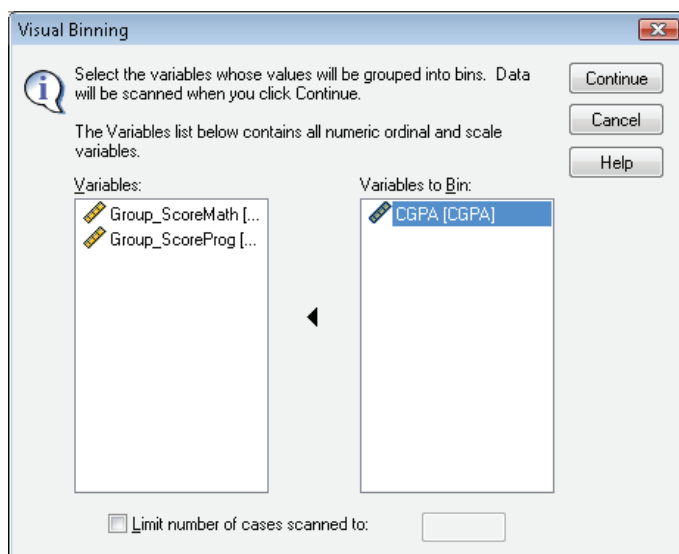
Crosstabs procedure requires data be in categorical form. Therefore you may need to transform your data value in scale form into categorical form. For example using *DataStudent2008.sav*, we wish to study the independency between gender and result of CGPA in the first semester study. In SPSS, the scale variable can be transformed to categorical variables by using **Visual Binning**.

Visual Binning is a useful feature that allows to create a new category variable from a scale variable.

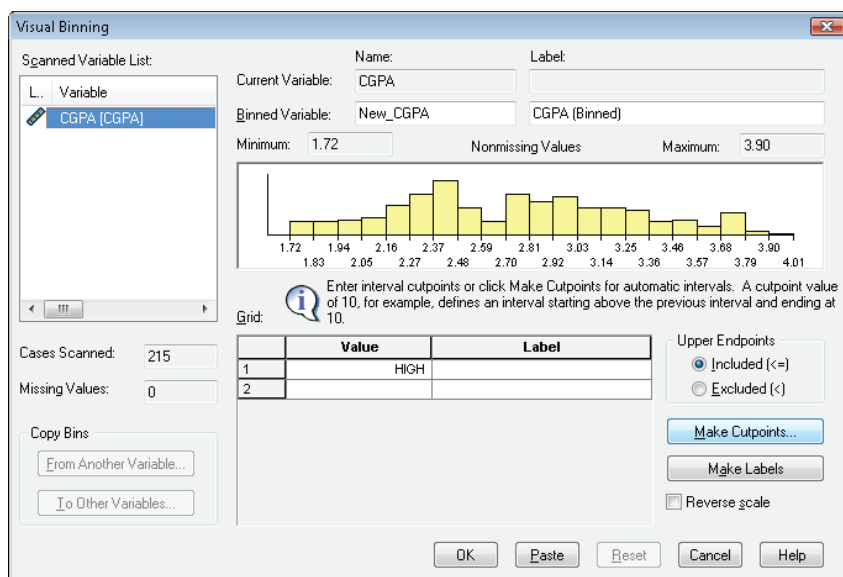
- From Transform...
- Click on Visual Binning as shown in the following figure



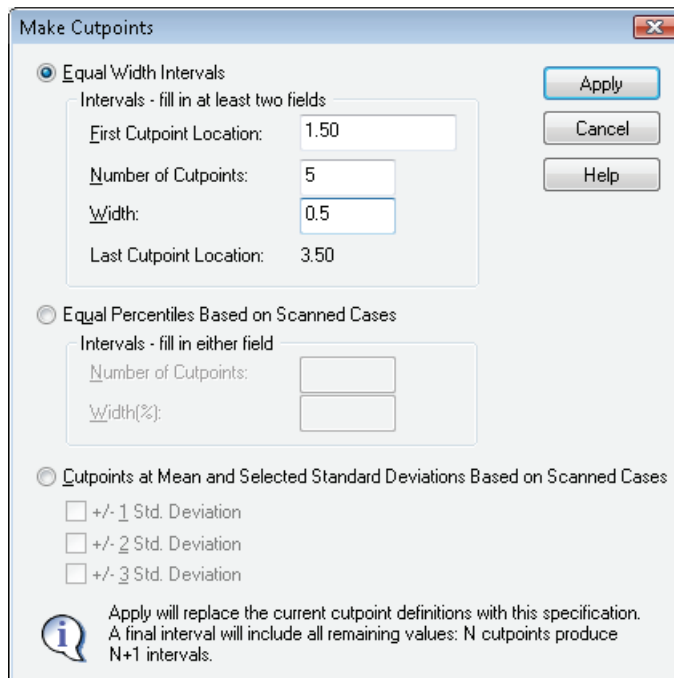
- In the **Visual Binning** dialog box, select the scale variables that needed to be transformed to categorical. In our case, we select CGPA variable and move to variable to Bin: box.
- Click on **Continue** as shown in the following figure



- In the next **Visual Binning** dialog box, select again the CGPA from the scanned variable list
- Type *new_CGPA* in the Binned variable text box
- Click on **Make Cutpoints...** as shown in the following figure



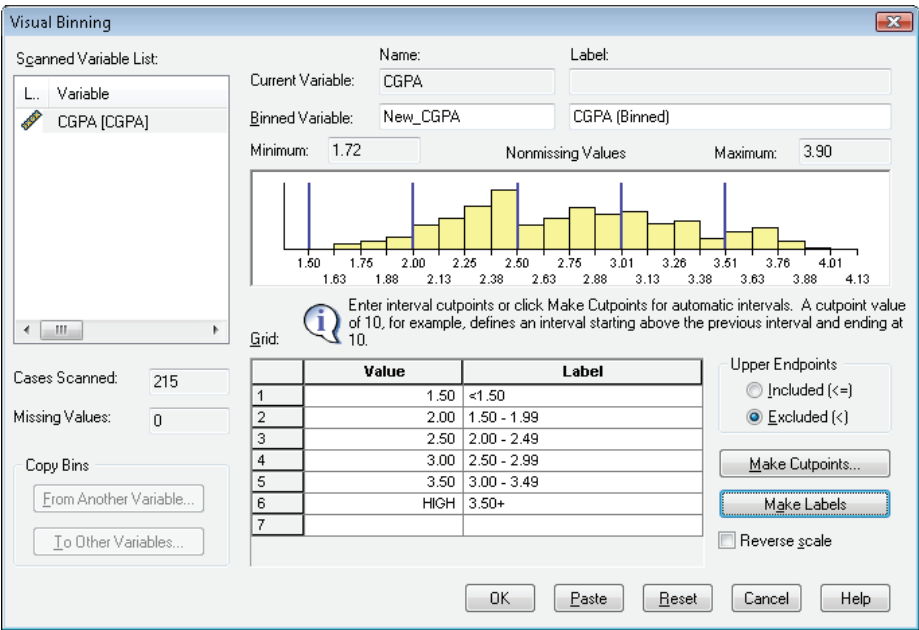
- From the visual distribution in the Visual Binning, it can give you some idea on how you to create the new categories of your scale variable.
- In our case, we can consider that the first category containing the CGPA less than 1.5 and we want 5 numbers of cutpoints. After all, SPSS is able to calculate automatically the width and the last cutpoint location. You may adjust these detail according to your analysis.
- If you agree with these cutpoints, click on the **Apply** button as shown in the following figure



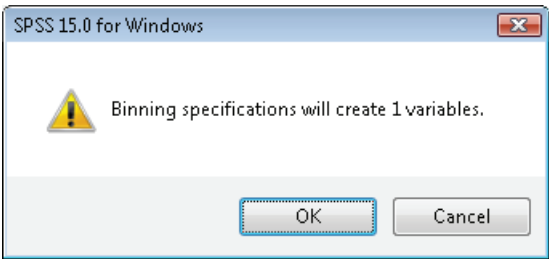
- Select **Excluded** (<) in order to have the Upper Endpoint category
- click Make Labels

This action generates automatically the new value labels for each category in the **Grid**: box.

- Click OK as shown in the following figure



SPSS gives a pop-up message to inform that 1 new binned variable will be created.



Finally, a new variable name *new_CGPA* was created at the end of list as shown in figure 5.2

SPSS Data Editor window showing the dataset "DataStudent2008.sav". The variable "Group_ScoreProg" is selected, and its values are displayed in the adjacent column.

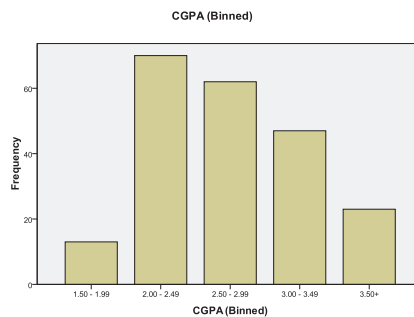
StudID	Gender	Cour	CGPA	MathGrade	ProgGrade	new_MathGra	new_ProgGrade	Group_ScoreMath	Group_ScorePro	New_CGPA	var
1 B0081	1	BITC	2.43	C	C+	6	8	3	3	3	
2 B0082	1	BITC	2.46	C+	C	8	6	3	3	3	
3 B0083	1	BITC	2.61	C-	B-	7	4	3	2	4	
4 B0084	1	BITC	2.39	C	C	6	6	3	3	3	
5 B0085	1	BITC	2.77	C+	B	8	3	3	2	4	
6 B0086	1	BITC	2.31	D+	C+	10	8	3	3	3	
7 B0087	1	BITC	2.86	C+	C+	8	8	3	3	4	
8 B0088	1	BITC	2.63	B-	B	4	3	2	2	4	
9 B0089	2	BITC	3.52	B	B	3	3	2	2	6	
10 B00810	1	BITC	2.47	C+	C+	8	8	3	3	3	
11 B00811	2	BITC	3.38	B-	B	4	3	2	2	5	
12 B00812	2	BITC	2.73	C-	C+	7	8	3	3	4	
13 B00813	1	BITC	3.26	B	B+	3	5	2	2	5	
14 B00814	2	BITC	3.33	A	B-	1	4	1	2	5	
15 B00815	2	BITC	3.72	A	A	1	1	1	1	6	
16 B00816	1	BITC	2.83	C+	B-	8	4	3	2	4	
17 B00817	1	BITC	2.06	C-	C	7	6	3	3	3	
18 B00818	1	BITC	2.23	C-	C+	7	8	3	3	3	
19 B00819	1	BITC	2.85	C+	B+	8	5	3	2	4	
20 B00820	1	BITC	2.45	C	C+	6	8	3	3	3	
21 B00821	1	BITC	2.33	D+	B-	10	4	3	2	3	
22 B00822	1	BITC	2.38	C+	C+	8	8	3	3	3	

Figure 5.2 DataStudent2008.sav with a binned variable

- Run **Frequencies** Analysis from **Descriptive Statistics** to see the frequency of CGPA in the categorical form.

CGPA (Binned)

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1.50 - 1.99	13	6.0	6.0	6.0
2.00 - 2.49	70	32.6	32.6	38.6
2.50 - 2.99	62	28.8	28.8	67.4
3.00 - 3.49	47	21.9	21.9	89.3
3.50+	23	10.7	10.7	100.0
Total	215	100.0	100.0	



By looking at the **Valid Percent** in the Frequencies table and the bar chart, we can see that the majority of students are in the CGPA value from 2.00-2.49 and less than 20 students got the CGPA below than 2.00.

At this point, we only explore the frequencies for each variable isolated through frequency table or by visual charts. We not yet compare the frequencies between the variables and see their relationship. In order to do that, we continue by using **Crosstabs** procedure.

5.6 Conducting a Crosstabs procedure

Objective 1: A lecturer wishes to determine whether the number of student in each course was influenced by of the gender of student.

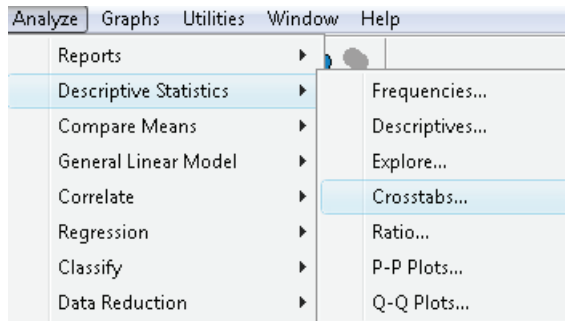
So we do understand that we want to test the relationship between gender and course. Before we start to do analysis, it is better to state the hypothesis so that, we can easily answer the test question. The hypotheses may be stated as follows:

H_0 : There is no relationship between gender and the number of student in each course

H_1 : There is relationship between gender and the number of student in each course

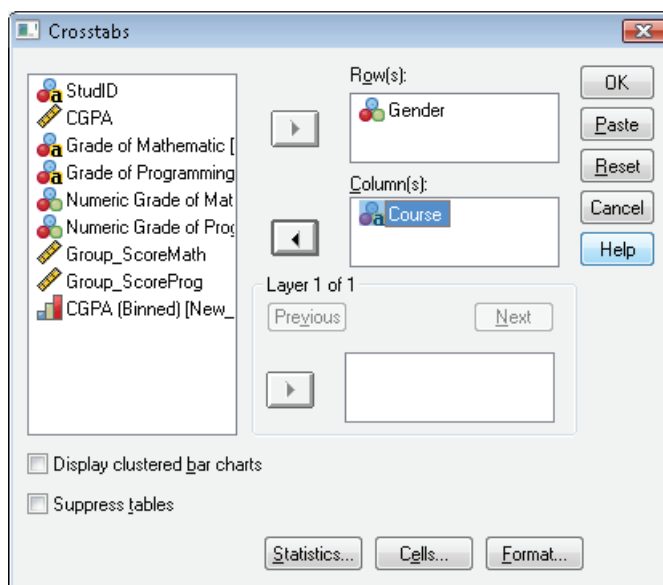
Now, you may start to analyze the data.

- From the **Analyze** menu, select **Descriptive Statistics**
- Click on **Crosstabs...** as shown in the following figure



In Crosstab dialog box :

- Select the Gender variable and move the variable into the **Row(s) :** box
- Select Course variable and move the variable **Column(s) :** box
- Click on OK as shown in the following figure



In the output viewer, the first table is **Case Processing Summary**. It shows the number of cases that are valid on both variables which is 215. Since the dataset do not contain missing cases. Thus, the total cases on both variables remain same.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * Course	215	100.0%	0	.0%	215	100.0%

The second part of the output gives the crosstabulation table of observed frequencies for each possible combination of the two variables. In this example, we can see the frequencies of female and male for the each course.

Gender * Course Crosstabulation

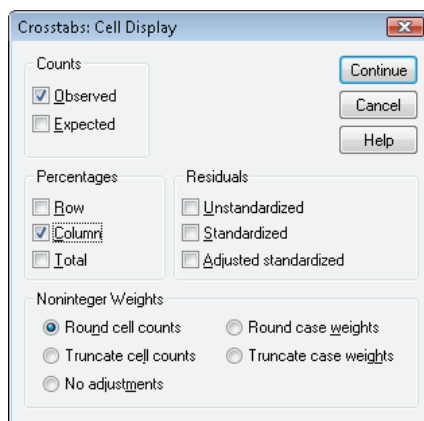
Count		Course					Total
		BITC	BITD	BITI	BITM	BITS	
	MALE	27	15	24	19	23	108
	FEMALE	20	33	14	22	18	107
	Total	47	48	38	41	41	215

As can be seen from crosstabulation table above, by default the Crosstabulation procedure will display only counts in the each cell of the table. In the total column for *Gender*, there were 108 male students and 107 female students. Thus we can understand that female and male only give very slight difference (difference in one case). In the total for *Course*, *BITD* and *BITC* have more number of students compared other courses. *BITI*

is the course that has fewer students compared other courses. It is often difficult, to interpret the frequency between variable in the each cell if we use only observed count. Therefore you may request percentages for easy interpretation.

5.7 Procedure request percentages in Crosstabs:

- Re- select **Descriptive Statistics**
- Click on **Crosstabs...**
- In Crosstabs dialog box, click on the **Cell...**
- Then it open the **Crosstabs: Cell Display** dialog box
- In the percentages box, click on the Column check box as shown in the following figure



Since *Course* variable is our column variable, column percentages allow immediate comparison of the percentage of male and female in each course. You can request also row percentages, but for more convenience and to keep the table simple, it is suggested to request only one type percentages in one time.

Gender * Course Crosstabulation

			Course					Total
			BITC	BITD	BITI	BITM	BITS	
Gender	MALE	Count	27	15	24	19	23	108
		% within Course	57.4%	31.3%	63.2%	46.3%	56.1%	50.2%
	FEMALE	Count	20	33	14	22	18	107
		% within Course	42.6%	68.8%	36.8%	53.7%	43.9%	49.8%
Total		Count	47	48	38	41	41	215
		% within Course	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

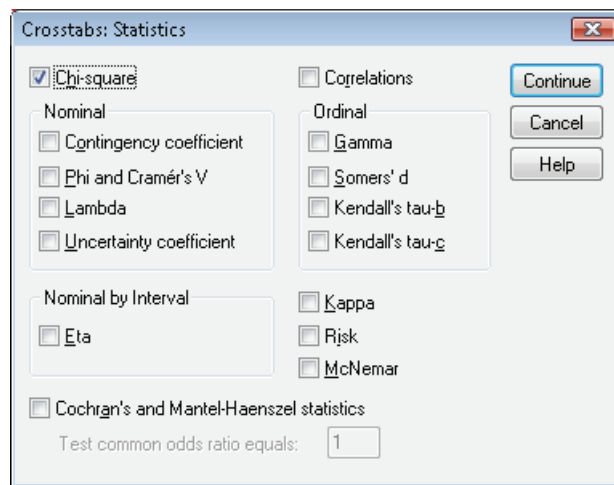
Now, there are two numbers in each cell; counts and column percentages. By using percentage, we can see that *BITD* has more female student at 69% and only 31% of male students. On the other hand, *BITI* has more male students at 63% and only 37 % of female students. Even though in the other courses also have differences frequencies

between male and female, but we can say that the differences female and male student do not have much difference compare to BITD and BITI course.

Through the crosstabulation table, we can clearly see the differences of frequencies between two variables through the observed count or percentages. However the difference of frequencies analysis is not able to conclude the relationship or test the independence of between variables. So we will use the Chi square test in the analysis

5.8 Procedure for requesting Chi-Square Test of Independence

- Reselect Descriptive Statistics from Analyze Menu
- Click on Crosstabs... to reopen the crosstabs dialog box
- Make sure in Row(s): box is Gender and Column(s) box is Course
- Click on the Statistics...
- In Crosstabs: Statistics dialog box
- Click on the Chi-Square check box
- Click on Continue.
- Then OK as shown in the following figure



To interpret the Chi-square, we need to look at the Pearson statistics. It is similar with the chi-square that we have been calculated in the manual calculation. In the SPSS output table, it is referred to as Pearson Chi square.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11.249 ^a	4	.024
Likelihood Ratio	11.453	4	.022
N of Valid Cases	215		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 18.91.

Pearson Chi-square is able to test the hypothesis that row and column variables are independent. In manual calculation, the Chi-square value is assessed based on comparison of the calculated value and the theoretical chi-square distribution. In SPSS, it can be easily accessed through an interpreted probability is returned by referring at column labeled Asymp. Sig. (2-sided) or also called as p-value (Coakes and Steed 2006).

In our example, the pearson Chi square is 11.249, has degrees of freedom (from the df column) and has p-value of 0.024 which is less than alpha value 0.05(normal cut off for significant level). So that allows you to reject the null hypothesis. This means that gender and course is related. We would claim that student female or student male influences which course that you are primarily interested in.

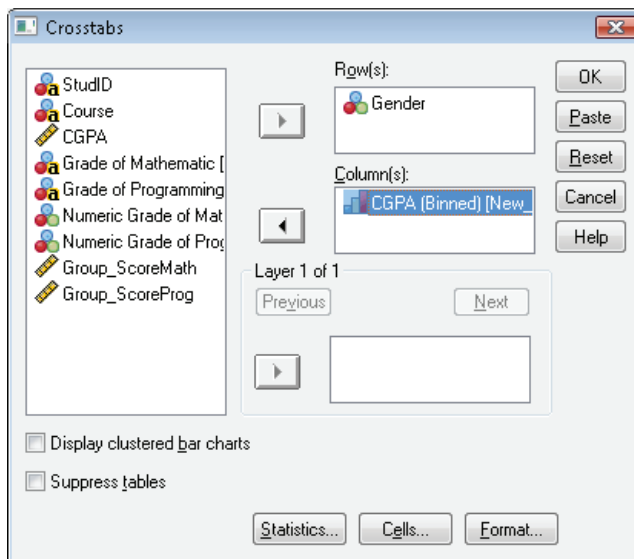
Objective 2: let try another test for independences by using same dataset *DataStudent 2008.sav*. In the test, the lecturer wished to test whether there is relationship between gender and the CGPA result student (CGPA_Binned). Again, it is suggested that to state the hypothesis in order to easy answer the test question. The hypotheses may be stated as follow:

H_0 : There is no relationship between gender and CGPA result

H_1 : There is relationship between gender and CGPA result

Now, you may start to analyze the data.

- Select **Descriptive Statistics** from the **Analyze** menu,
- click on **Crosstabs...**
- In Crosstabs dialog box
- Move *Gender* into **Row(s):** box and move *CGPA(Binned)* into **Column(s):** box



- Click on the **Statistics** button...
- In the **Crosstabs: Statistics** dialog box
- Click on the Chi-square check box

- Click on **Continue**
- Click on the **Cells...**
- In **Crosstabs: Cell Display** dialog box
- In the percentages, click on the column check boxes.
- Click on **Continue** and then **OK**.

The SPSS output viewer appears as follows:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * CGPA (Binned)	215	100.0%	0	.0%	215	100.0%

Gender * CGPA (Binned) Crosstabulation

			CGPA (Binned)					Total
			1.50 - 1.99	2.00 - 2.49	2.50 - 2.99	3.00 - 3.49	3.50+	
Gender	MALE	Count	9	41	28	18	12	108
		% within CGPA (Binned)	69.2%	58.6%	45.2%	38.3%	52.2%	50.2%
	FEMALE	Count	4	29	34	29	11	107
		% within CGPA (Binned)	30.8%	41.4%	54.8%	61.7%	47.8%	49.8%
Total	Count	13	70	62	47	23	215	
	% within CGPA (Binned)	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7.174 ^a	4	.127
Likelihood Ratio	7.260	4	.123
Linear-by-Linear Association	3.662	1	.056
N of Valid Cases	215		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.47.

In Gender * CGPA(Binned), we can see that the highest frequency student is in category CGPA 2.00-2.49 . On the first inspection on percentage, this seems that female students tend to obtain good CGPA result. As we can see from the table, it shows that from CGPA 2.50-2.99 category until CGPA more than 3.5 category, female students have more percentage compare with male students. But let us check the Chi-square test.

In the Chi-square Tests, Pearson Chi-square is 7174 , has degree of freedom 4 and has p-value 0.127 which means it more than alpha level 0.05 (p-value>0.05). Refer back to our hypotheses, thus it fail to reject H_0 (accept H_0). It means that gender of student is not influence in the CGPA result.

Worked Examples

Worked Example 1

A researcher deals with the need to overcome the job-related anxiety. A group of workers were interviewed to determine whether the dreadful job and wonderful job are able to cause on their level of anxiousness.

	Anxious	Not anxious	Total
Dreadful job	210	150	360
Wonderful job	62	52	114
Total	272	202	

- Prepare a table and calculate the expected frequency
- Calculate the computed χ^2 and critical χ^2 .
- Briefly analyze the result.

Answer :

a)

The calculation needed for us to work out the crosstab procedure is as follows:

	Anxious	Not anxious	Total
Dreadful job	$f_{\bullet} = 210$ $f_z =$	$f_{\bullet} = 150$ $f_z =$	360
Wonderful job	$f_{\bullet} = 62$ $f_z =$	$f_{\bullet} = 52$ $f_z =$	114
Total	272	202	

$$f_z = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}$$

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t}$$

Calculated χ^2 = ?

Degrees of freedom = ?

Critical Value χ^2 (1, .05) = ?

	Anxious	Not anxious	Total
Dreadful job	$f_o = 210$ $f_t = 206.58$	$f_o = 150$ $f_t = 153.42$	360
Wonderful job	$f_o = 62$ $f_t = 65.42$	$f_o = 52$ $f_t = 48.58$	114
Total	272	202	474

b)

$$f_t = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}$$

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t}$$

c)

$$\text{Calculated } \chi^2 = 0.0566 + 0.1788 + 0.0762 + 0.2408 = 0.5524$$

$$\text{Calculated } \chi^2 = 0.5524.$$

$$\text{degrees of freedom} = 2 - 1 = 1$$

$$\text{Critical Value } \chi^2 (1, .05) = 3.84$$

Calculated $\chi^2 < \text{Critical Value } \chi^2$ (0.5524 < 3.84) Fail to reject the null hypothesis (accept the null hypothesis). There were no significant differences between dreadful job and wonderful job in their level of anxiousness (anxious or not anxious). This means there is no significant association between the two categorical variables.

Worked Example 2

A study was to explore a possible relation between the human feeling and the choice of drinking tea or coffee. The data collected for this study are as below:

	Drink tea	Drink coffee
Feel terrific	70	50
Feel lousy	30	80

Solutions:

	Drink tea	Drink coffee	Total
Feel terrific	$f_o = 70$ $f_t = 52.17$	$f_o = 50$ $f_t = 67.83$	120
Feel lousy	$f_o = 30$ $f_t = 47.83$	$f_o = 80$ $f_t = 62.17$	110
Total	100	130	230

$$f_t = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}$$

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t}$$

$$\text{Calculated } \chi^2 = 6.094 + 6.647 + 4.687 + 5.114 = 22.54$$

$$\text{Calculated } \chi^2 = 22.54$$

$$\text{degrees of freedom} = 2 - 1 = 1$$

$$\text{Critical Value } \chi^2 (1, .05) = 3.84$$

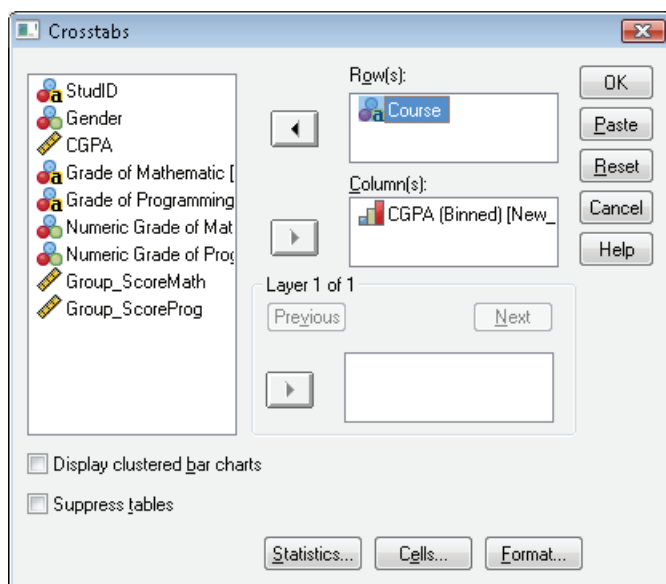
$$\text{Calculated } \chi^2 > \text{Critical Value } \chi^2 \text{ (22.54} > 3.84)$$

Reject the null hypothesis. There were significant differences between respondents feeling who drink tea and drink coffee. Thus the choice of drinking tea or coffee significantly is influenced by the human feeling.

Worked Example 3

Refer to Data Set *DataStudent 2008.sav* from the CD given for the following exercise.

A lecturer wants to test whether there is difference in the CGPA result based upon the type of Course. To do so, you may select *Course* and new *CGPA(Binned)* as shown in figure below :



You may request column percentage for easy interpretation and also Chi-square test for the significance relationship between the variables.

The outputs may appear as follows:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Course * CGPA (Binned)	215	100.0%	0	.0%	215	100.0%

Course * CGPA (Binned) Crosstabulation								
			CGPA (Binned)					Total
			1.50 - 1.99	2.00 - 2.49	2.50 - 2.99	3.00 - 3.49	3.50+	
Course	BITC	Count	0	13	14	14	6	47
		% within CGPA (Binned)	.0%	18.6%	22.6%	29.8%	26.1%	21.9%
	BITD	Count	3	12	22	7	4	48
		% within CGPA (Binned)	23.1%	17.1%	35.5%	14.9%	17.4%	22.3%
	BITI	Count	0	10	10	14	4	38
		% within CGPA (Binned)	.0%	14.3%	16.1%	29.8%	17.4%	17.7%
	BITM	Count	9	21	3	3	5	41
		% within CGPA (Binned)	69.2%	30.0%	4.8%	6.4%	21.7%	19.1%
	BITS	Count	1	14	13	9	4	41
		% within CGPA (Binned)	7.7%	20.0%	21.0%	19.1%	17.4%	19.1%
	Total	Count	13	70	62	47	23	215
		% within CGPA (Binned)	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	51.789 ^a	16	.000
Likelihood Ratio	52.507	16	.000
N of Valid Cases	215		

a. 8 cells (32.0%) have expected count less than 5. The minimum expected count is 2.30.

Answer :

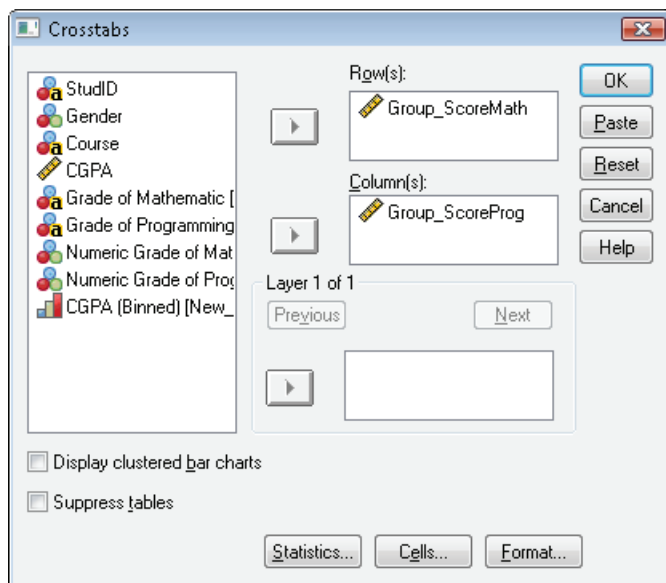
- The result shows that Pearson chi-square has a value of 51.789 with a very low significance of .000. This significance value is well below the alpha level of 0.05. Thus, initially it shows that there is a significant relationship between course and the Cumulative Grade Point Average (CGPA) result.
- However in this example, there is an important warning at the bottom of the Chi-Square output. The warning tells us that 32.2 % of the cells have expected frequencies less than 5. Thus, one of the assumptions of chi-square has been violated and the results may not be meaningful.
- Because of that, we can conclude that there is insufficient evidence to conclude that whether course influences the Cumulative Grade Point Average (CGPA) result.

Worked Example 4

Refer to Data Set *DataStudent 2008.sav* from the CD given for the following exercise.

A lecturer wishes to compare the result of student in mathematic with the result of student in programming. Can you conclude that there is relationship between the performances of student in both subjects?

To answer this question, you may need to use *Group_ScoreMath* and *Group_ScoreProg* as shown in figure below:



The outputs may appear as follows:

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Group_ScoreMath * Group_ScoreProg	215	100.0%	0	.0%	215	100.0%

Group_ScoreMath * Group_ScoreProg Crosstabulation

			Group_ScoreProg			Total
			1	2	3	
Group_ScoreMath	1	Count	16	7	1	24
		% within Group_ScoreProg	48.5%	6.7%	1.3%	11.2%
	2	Count	15	47	19	81
		% within Group_ScoreProg	45.5%	44.8%	24.7%	37.7%
	3	Count	2	51	57	110
		% within Group_ScoreProg	6.1%	48.6%	74.0%	51.2%
Total		Count	33	105	77	215
		% within Group_ScoreProg	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	76.301 ^a	4	.000
Likelihood Ratio	69.510	4	.000
Linear-by-Linear Association	58.414	1	.000
N of Valid Cases	215		

a. 1 cells (11.1%) have expected count less than 5. The minimum expected count is 3.68.

Answer:

Since the chi-square test has less significant value (.000) than alpha value 0.05 and the number of cells have less than 5 frequencies is not more than 20%, so you may conclude there is a statistically significant relationship between mathematic score and programming score, students have a high score in mathematic score are more likely to have high score in programming subject.

Chapter 5 Review Exercises

Question 1

A sample of 600 students who are asked whether they usually buy store brand or name brand products are recorded in the following table.

	Usually buy	
	Store Brand	Name Brand
Male	135	110
Female	212	143

Using the 1% significance level, can you reject the null hypothesis that the two attributes, gender and store or name products, are independent?

Question 2

A survey is being conducted to determine whether the age of a person is related to the type of TV programme he or she watches. A random sample of 700 people gives the data shown here. At $\alpha=0.025$, is the type of TV programme watched related to a person's age?

Age	TV programme			
	Cartoon	Documentary	Comedy	Mystery
1 -10	54	8	8	5
11-20	11	33	29	26
21-30	13	27	34	27
31-40	18	25	41	48
41-50	14	24	42	25
51-60	26	26	40	23
61 and above	11	18	34	10

Question 3

A researcher wishes to know whether the way people obtain information is related to their educational background. A random sample of 550 people yielded the following data. At $\alpha=0.05$, test the claim that the way people obtain information is independent of their educational background.

	Internet	TV	Newspaper	Others
Primary	25	46	52	5
Secondary	50	40	63	17
Tertiary	91	61	70	30

Question 4

The table below shows classification of 450 randomly selected workers based on their status as smoker or nonsmoker and on the number of visits they made to the doctor last year.

	Number of visits to the doctor		
	0-2	3-5	> 5
Smoker	30	65	105
Nonsmoker	115	90	45

Test at the 1% significance level if there is any relation between smoking and the number of visits to the doctor.

Question 5

A researcher wishes to know whether the age of the house purchaser is related to the price of the house purchased. A sample of 240 house owners shows the following data. At $\alpha=0.05$, is the price of the house independent of the age of the owner?

Age	Price		
	Below RM150 000	RM150 001–RM300 000	RM300 001 and above
21-30	18	28	2
31-40	46	28	14
41-50	33	18	17
51 and above	11	14	11

Question 6

Olympic Electronics Company buys a type of integrated circuit chips (IC) from two subcontractors, P and Q. The quality control department at this company wanted to check if the distribution of good and defective IC is the same from both subcontractors. The quality control engineer selected a random sample of 250 IC chips from subcontractor P and 350 IC chips from subcontractor Q. These IC chips were checked for being good or defective and recorded in the following table.

	Subcontractor P	Subcontractor Q
Good	232	325
Defective	18	25

Using the 10% significance level, test the null hypothesis that the distribution of good and defective IC chips are the same for both subcontractors.

Question 7

A survey is conducted to see if the instructor's degree is related to the students' opinion of teaching quality. A random sample of 120 students' evaluations of various instructors is shown below. At $\alpha=0.10$, can we conclude that the degree of the instructor is related to students' opinions about the teaching quality?

	Degree		
Rating	Bachelor	Master	Doctorate
Excellent	19	17	9
Average	18	8	13
Poor	7	14	15

Question 8

A survey had been conducted and found out Malaysia people (different age of groups) had different tastes on beverage. The result is shown below :

	Age Groups		
Types of beverage	< 30	30 – 55	>55
Milo	200	140	50
Nescafe	60	56	130
Horlick	40	84	153

At $\alpha=0.10$, can we conclude that there is a relationship between age group and beverage preference?

Question 9

The following data shows the performance of 350 staff in a training program and their job performance.

		Performance in training program		
		Poor	Average	Excellent
Job Performance	Poor	18	55	24
	Average	22	74	53
	Excellent	3	44	57

At $\alpha=0.01$, test the null hypothesis that performance in the training program and job performance are independent.

Question 10

An officer in a recreational club is interested to know if there is a relation between the facilities and gender. The table below shows the data collected from 760 club members. What can she conclude by using the data at 2.5% significance level?

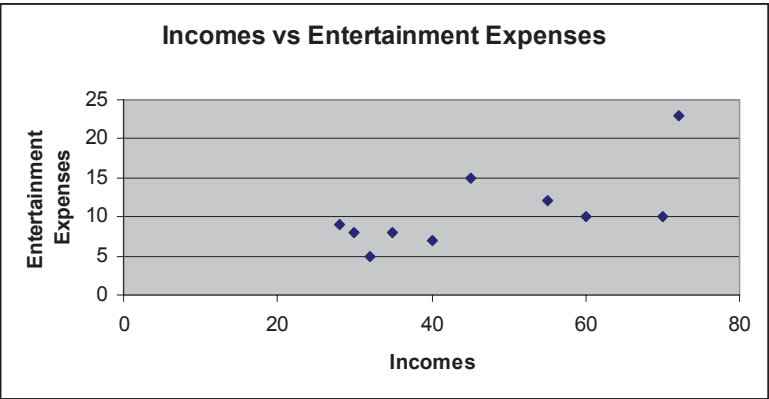
	Facilities			
Gender	Tennis court	Swimming pool	Gymnasium	Track
Male	92	138	110	46
Female	88	151	98	37



REVIEW EXERCISES' ANSWERS

Question 1

1.a)

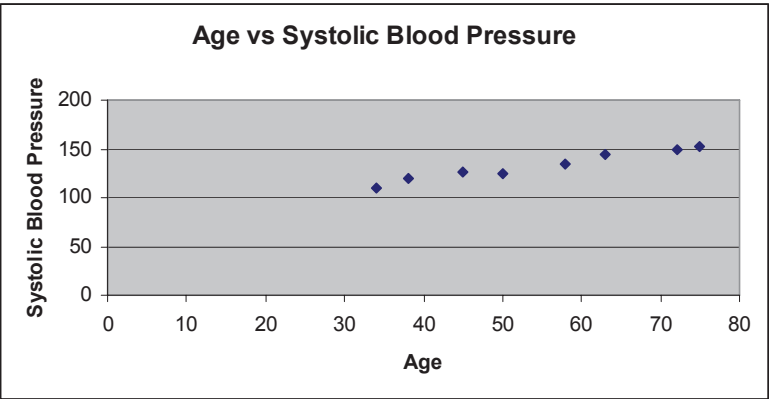


1.b) 0.66

1.c) Strong positive linear correlation.

Question 2

2.a)

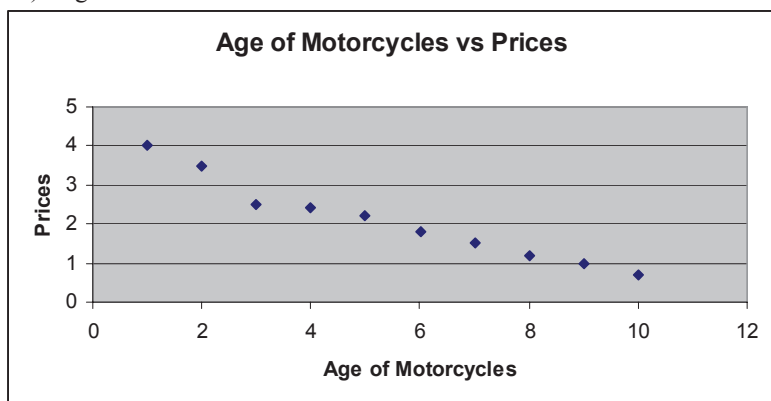


2.b) 0.98

2.c) Very strong positive linear correlation.

Question 3

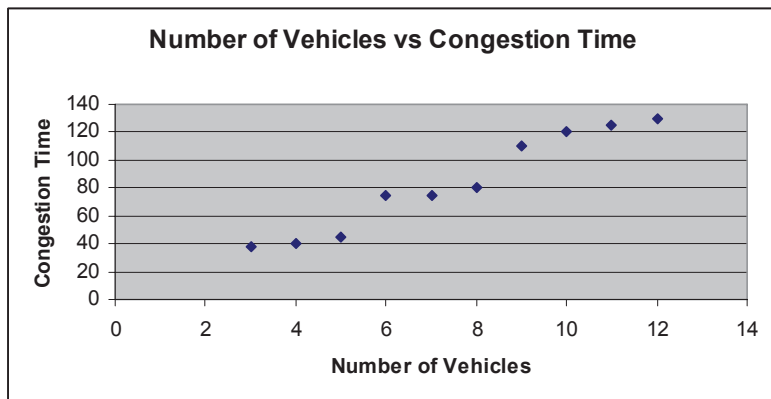
3.a) Negative linear correlation.



3.b) -0.98

Question 4

4.a)

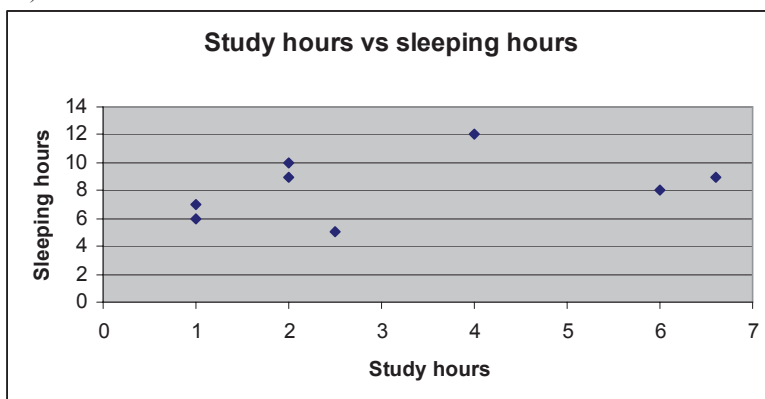


4.b) Yes

4.c) 0.97. Very strong positive linear correlation.

Question 5

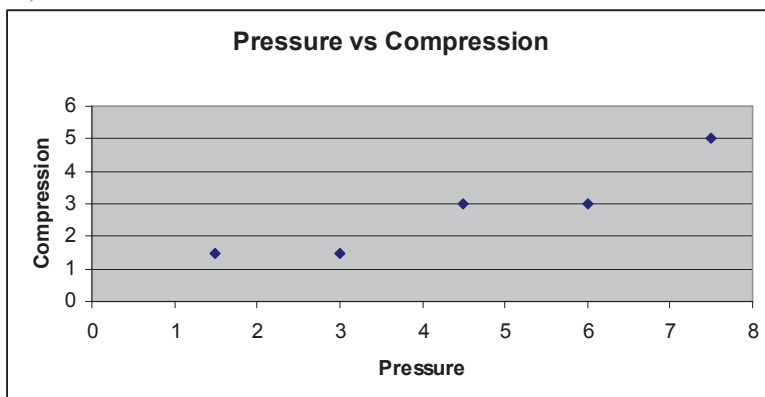
5.a)



5.b) 0.35. Low positive linear correlation.

Question 6

6.a)

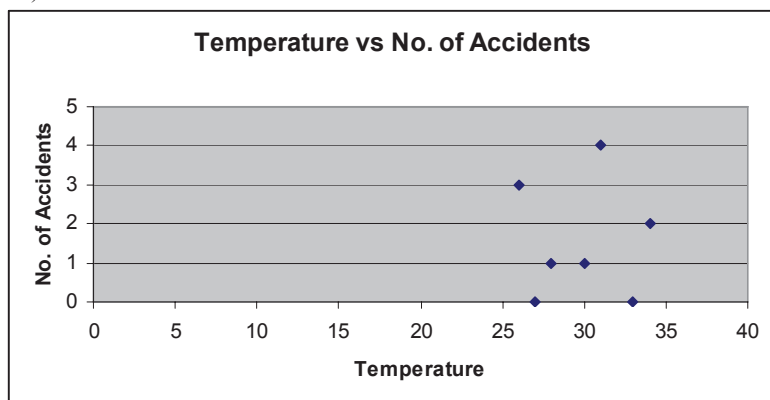


6.b) Yes.

6.c) 0.93. Very strong positive linear correlation.

Question 7

7.a)

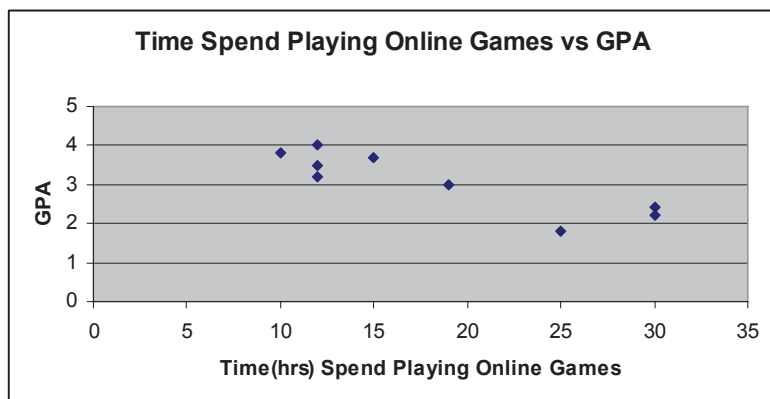


7.b) -0.02

7.c) Very weak negative linear correlation.

Question 8

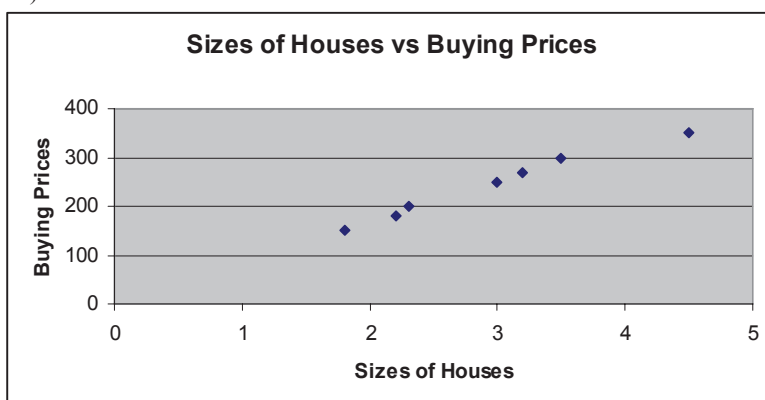
8.a)



8.b) -0.88 Very strong negative linear correlation.

Question 9

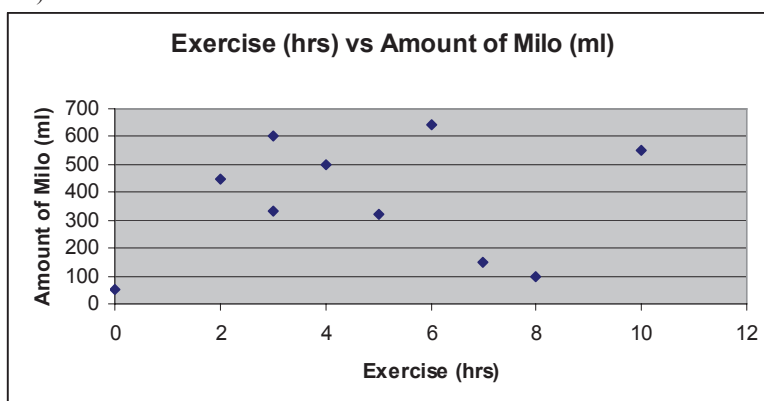
9.a)



9.b) 0.99 Very strong positive linear correlation.

Question 10

10.a)



10.b) 0.13

10.c) Very weak positive linear correlation.

Chapter 2 Review Exercises' Answers

Question 1

1.a) $\hat{y} = 1.0849 + 0.2059x$

1.b) The value of $a = 1.0849$ is the entertainment expenditure when there is no income. On average, for every RM100 increase in income will result in RM108.49 increase in entertainment expenditure.

1.c) RM1138

Question 2

2.a) $\hat{y} = 79.49 + 0.98x$

2.b) 133.4

Question 3

3.a) $\hat{y} = 3.97333 - 0.34424x$

3.b) The value of $a = 3.97333$ is the price of a new motorcycle. On average, for every increase of 1 year on the age of motorcycle will result in the depreciation of RM344.24.

3.c) RM1047.29

Question 4

4.a) $\hat{y} = -2.93 + 11.56x$

4.b) The value of $a = -2.93$ is the congestion time when there is no vehicle. The value of $b = 11.56$ means on average, for every 1 unit increase in the number of vehicle will result 11.56 seconds increase in congestion time.

4.c) 170.5 s

Question 5

5. $\hat{y} = 7.1277 + 0.3577x$

Question 6

6.a) $\hat{y} = 0.25 + 0.57x$

6.b) 4.0

Question 7

7.a) $\hat{y} = 32.99253 + 0.32802x$

7.b) RM50377.59

Question 8

8.a) $\hat{y} = 4.63 - 0.085x$

8.b) The value of $a = 4.63$ is GPA for a student who does not play any online games. The value of $b = -0.085$ means on average, for every 1 hour increase in playing online games will result 0.085 point decrease in GPA.

8.c) 2.93

Question 9

9.a) $\hat{y} = 19.83390 + 76.15428x$

9.b) RM324 451.02

Question 10

10.a) $\hat{y} = 13.26 + 0.299x$

10.b) The value of $a = 13.26$ is the consumption of drinking water (in l) without doing any exercise.

The value of $b = 0.299$ means on average, for every 1 hour increase in doing exercise will result 0.299 l of drinking water consumption.

10.c) 17.45 l

Chapter 3 Review Exercises' Answers

Question 1

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies			
	Usually buy		
Gender	Store Brand	Name Brand	Total
Male	135	110	245
Female	212	143	355
Total	347	253	600

Calculations

fo-fe

-6.69167 6.691667

6.691667 -6.69167

Expected Frequencies			
	Usually buy		
Gender	Store Brand	Name Brand	Total
Male	141.6916667	103.3083333	245
Female	205.3083333	149.6916667	355
Total	347	253	600

(fo-fe)²/fe

0.316027 0.433444

0.218103 0.299138

Data	
Level of Significance	0.01
Number of Rows	2
Number of Columns	2
Degrees of Freedom	1

Results	
Critical Value	6.634896712
Chi-Square Test Statistic	1.266712092
p-Value	0.260384586
Do not reject the null hypothesis	

Null Hypothesis : the two attributes, gender and store or name products, are independent**Conclusion : the gender and store product are independent.**

Question 2

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies					
	TV Programme				
Age	Cartoon	Documentary	Comedy	Mystery	Total
1-10	54.00	8.00	8.00	5.00	75.00
11-20	11.00	33.00	29.00	26.00	99.00
21-30	13.00	27.00	34.00	27.00	101.00
31-40	18.00	25.00	41.00	48.00	132.00
41-50	14.00	24.00	42.00	25.00	105.00
51-60	26.00	26.00	40.00	23.00	115.00
61 and above	11.00	18.00	34.00	10.00	73.00
Total	147.00	161.00	228.00	164.00	700.00

Calculations

fo-fe			
38.25	-9.25	-16.43	-12.57
-9.79	10.23	-3.25	2.81
-8.21	3.77	1.10	3.34
-9.72	-5.36	-1.99	17.07
-8.05	-0.15	7.80	0.40
1.85	-0.45	2.54	-3.94
-4.33	1.21	10.22	-7.10

Expected Frequencies					
	TV Programme				
Age	Cartoon	Documentary	Comedy	Mystery	Total
1-10	15.75	17.25	24.43	17.57	75.00
11-20	20.79	22.77	32.25	23.19	99.00
21-30	21.21	23.23	32.90	23.66	101.00
31-40	27.72	30.36	42.99	30.93	132.00
41-50	22.05	24.15	34.20	24.60	105.00
51-60	24.15	26.45	37.46	26.94	115.00
61 and above	15.33	16.79	23.78	17.10	73.00
Total	147.00	161.00	228.00	164.00	700.00

(fo-fe) ² /fe			
92.89	4.96	11.05	8.99
4.61	4.60	0.33	0.34
3.18	0.61	0.04	0.47
3.41	0.95	0.09	9.43
2.94	0.00	1.78	0.01
0.14	0.01	0.17	0.58
1.22	0.09	4.40	2.95

Data	
Level of Significance	0.025
Number of Rows	7.00
Number of Columns	4.00
Degrees of Freedom	18.00

Results	
Critical Value	31.53
Chi-Square Test Statistic	160.22
p-Value	0.00000
Reject the null hypothesis	

Null Hypothesis : type of TV programme watched is not related to a person's age

Conclusion : the type of TV programme watched is related to a person's age.

Question 3

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies					
	Way obtain information				
Educational background	Internet	TV	Newspaper	Others	Total
Primary	25	46	52	5	128
Secondary	50	40	63	17	170
Tertiary	91	61	70	30	252
Total	166	147	185	52	550

Calculations

fo-fe			
-13.63	11.79	8.95	-7.10
-1.31	-5.44	5.82	0.93
14.94	-6.35	-14.76	6.17

Expected Frequencies					
	Way obtain information				
Educational background	Internet	TV	Newspaper	Others	Total
Primary	38.63	34.21	43.05	12.10	128
Secondary	51.31	45.44	57.18	16.07	170
Tertiary	76.06	67.35	84.76	23.83	252
Total	166	147	185	52	550

(fo-fe) ² /fe			
4.81	4.06	1.86	4.17
0.03	0.65	0.59	0.05
2.94	0.60	2.57	1.60

Data	
Level of Significance	0.05
Number of Rows	3
Number of Columns	4
Degrees of Freedom	6

Results	
Critical Value	12.5915872
Chi-Square Test Statistic	23.9349791
p-Value	0.00053684
Reject the null hypothesis	

Null Hypothesis : the way people obtain information is independent of their educational background.

Conclusion : The way of people obtain information is dependent of their educational background.

Question 4

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Number of visits to the doctor			
status	0-2	3-5	> 5	Total
Smoker	30	65	105	200
Nonsmoker	115	90	45	250
Total	145	155	150	450

Calculations

fo-fe		
-34.44	-3.89	38.33
34.44	3.89	-38.33

Expected Frequencies				
	Number of visits to the doctor			
status	0-2	3-5	> 5	Total
Smoker	64.44	68.89	66.67	200
Nonsmoker	80.56	86.11	83.33	250
Total	145	155	150	450

(fo-fe) ² /fe		
18.41	0.22	22.04
14.73	0.18	17.63

Data	
Level of Significance	0.01
Number of Rows	2
Number of Columns	3
Degrees of Freedom	2

Results	
Critical Value	9.21034
Chi-Square Test Statistic	73.20809
p-Value	1.27E-16
Reject the null hypothesis	

Null Hypothesis : there is no relation between smoking and the number of visits to the doctor.

Conclusion : there is relation between smoking and the number of visits to the doctor

Question 5

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Price			
Age	Below RM150 000	RM150 001–RM300 000	RM300 001 and above	Total
21-30	18	28	2	48
31-40	46	28	14	88
41-50	33	18	17	68
51 and above	11	14	11	36
Total	108	88	44	240

Calculations

fo-fe		
-3.60	10.40	-6.80
6.40	-4.27	-2.13
2.40	-6.93	4.53
-5.20	0.80	4.40

Expected Frequencies				
	Price			
Age	Below RM150 000	RM150 001–RM300 000	RM300 001 and above	Total
21-30	21.6	17.60	8.80	48
31-40	39.6	32.27	16.13	88
41-50	30.6	24.93	12.47	68
51 and above	16.2	13.20	6.60	36
Total	108	88	44	240

(fo-fe) ² /fe		
0.60	6.15	5.25
1.03	0.56	0.28
0.19	1.93	1.65
1.67	0.05	2.93

Data	
Level of Significance	0.05
Number of Rows	4
Number of Columns	3
Degrees of Freedom	6

Results	
Critical Value	12.59159
Chi-Square Test Statistic	22.29628
p-Value	0.00107
Reject the null hypothesis	

Null Hypothesis : the price of the house independent of the age of the owner.**Conclusion** : the price of the house dependent of the age of the owner

Question 6

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies			
	Type of contractor		
status of IC	Subcontractor P	Subcontractor Q	Total
Good	232	325	557
Defective	18	25	43
Total	250	350	600

Calculations

fo-fe	
-0.08	0.08
0.08	-0.08

Expected Frequencies			
	Type of contractor		
status of IC	Subcontractor P	Subcontractor Q	Total
Good	232.08	324.92	557
Defective	17.92	25.08	43
Total	250	350	600

(fo-fe) ² /fe	
0.00	0.00
0.00	0.00

Data	
Level of Significance	0.1
Number of Rows	2
Number of Columns	2
Degrees of Freedom	1

Results	
Critical Value	2.705543971
Chi-Square Test Statistic	0.000715747
p-Value	0.978656382
Do not reject the null hypothesis	

Null Hypothesis : the distribution of good and defective IC chips are the same for both subcontractors

Conclusion : There is no relation between the status of IC and type of contractor

Question 7

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Degree			
Rating	Bachelor	Master	Doctorate	Total
Excellent	19	17	9	45
Average	18	8	13	39
Poor	7	14	15	36
Total	44	39	37	120

Calculations

fo-fe		
2.50	2.38	-4.88
3.70	-4.68	0.98
-6.20	2.30	3.90

Expected Frequencies				
	Degree			
Rating	Bachelor	Master	Doctorate	Total
Excellent	16.50	14.63	13.88	45
Average	14.30	12.68	12.03	39
Poor	13.20	11.70	11.10	36
Total	44	39	37	120

(fo-fe) ² /fe		
0.38	0.39	1.71
0.96	1.72	0.08
2.91	0.45	1.37

Data	
Level of Significance	0.1
Number of Rows	3
Number of Columns	3
Degrees of Freedom	4

Results	
Critical Value	7.77944
Chi-Square Test Statistic	9.972544
p-Value	0.040893
Reject the null hypothesis	

Null Hypothesis : the degree of the instructor is not related to students' opinions about the teaching quality

Conclusion : the degree of the instructor is related to students' opinions about the teaching quality

Question 8

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Age Group			
Type of beverages	< 30	30 – 55	>55	Total
Milo	200	140	50	390
Nescafe	60	56	130	246
Horlick	40	84	153	277
Total	300	280	333	913

Calculations

fo-fe		
71.85	20.39	-92.25
-20.83	-19.44	40.28
-51.02	-0.95	51.97

Expected Frequencies				
	Age Group			
Type of beverages	< 30	30 – 55	>55	Total
Milo	128.15	119.61	142.25	390
Nescafe	80.83	75.44	89.72	246
Horlick	91.02	84.95	101.03	277
Total	300	280	333	913

(fo-fe) ² /fe		
40.29	3.48	59.82
5.37	5.01	18.08
28.60	0.01	26.73

Data	
Level of Significance	0.1
Number of Rows	3
Number of Columns	3
Degrees of Freedom	4

Results	
Critical Value	7.77944
Chi-Square Test Statistic	187.384
p-Value	1.93E-39
Reject the null hypothesis	

Null Hypothesis : there is no relationship between age group and beverage preference

Conclusion : there is a relationship between age group and beverage preference

Question 9

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Performance in training program			
Job Performance	Poor	Average	Excellent	Total
Poor	18	55	24	97
Average	22	74	53	149
Excellent	3	44	57	104
Total	43	173	134	350

Expected Frequencies				
	Performance in training program			
Job Performance	Poor	Average	Excellent	Total
Poor	11.92	47.95	37.14	97
Average	18.31	73.65	57.05	149
Excellent	12.78	51.41	39.82	104
Total	43	173	134	350

Calculations

fo-fe		
6.08	7.05	-13.14
3.69	0.35	-4.05
-9.78	-7.41	17.18

(fo-fe) ² /fe		
3.10	1.04	4.65
0.75	0.00	0.29
7.48	1.07	7.42

Data	
Level of Significance	0.01
Number of Rows	3
Number of Columns	3
Degrees of Freedom	4

Results	
Critical Value	13.2767
Chi-Square Test Statistic	25.78772
p-Value	3.49E-05
Reject the null hypothesis	

Null Hypothesis : that performance in the training program and job performance are independent.

Conclusion : that performance in the training program and job performance are dependent.

Question 10

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies					
	Column variable				
Gender	Tennis court	Swimming pool	Gymnasium	Track	Total
Male	92	138	110	46	386
Female	88	151	98	37	374
Total	180	289	208	83	760

Calculations

fo-fe				
0.58	-8.78	4.36	3.84	
-0.58	8.78	-4.36	-3.84	

Expected Frequencies					
	Column variable				
Gender	Tennis court	Swimming pool	Gymnasium	Track	Total
Male	91.42	146.78	105.64	42.16	386
Female	88.58	142.22	102.36	40.84	374
Total	180	289	208	83	760

(fo-fe)^2/fe				
0.00	0.53	0.18	0.35	
0.00	0.54	0.19	0.36	

Data	
Level of Significance	0.025
Number of Rows	2
Number of Columns	4
Degrees of Freedom	3

Results	
Critical Value	9.348404
Chi-Square Test Statistic	2.152938
p-Value	0.541277
Do not reject the null hypothesis	

Null Hypothesis : there is no relationship between the facilities and gender

Conclusion : there is no relationship between the facilities and gender

Chapter 4 Review Exercises' Answers

Question 1

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	6	387	64.5	232.3
Column 2	6	429	71.5	167.5
Column 3	6	468	78	157.2

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	547	2	273.5	1.47307	0.260556	6.358873
Within Groups	2785	15	185.6667			
Total	3332	17				

Since $F < F_{crit}$, do not reject null hypothesis

Question 2

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	11	14	1.272727	2.018182
Column 2	11	23	2.090909	4.490909
Column 3	11	12	1.090909	1.890909

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	6.242424	2	3.121212	1.114719	0.341215	5.390346
Within Groups	84	30	2.8			
Total	90.24242	32				

Since $F < F_{crit}$, do not reject null hypothesis

Question 3

Anova: Single Factor

SUMMARY				
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Column 1	9	398	44.22222	257.4444
Column 2	9	705	78.33333	710
Column 3	9	586	65.11111	343.8611
Column 4	9	444	49.33333	360

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	6504.306	3	2168.102	5.189002	0.004917	2.90112
Within Groups	13370.44	32	417.8264			
Total	19874.75	35				

Since $F > F_{crit}$, reject the null hypothesis**Question 4**

Anova: Single Factor

SUMMARY				
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Column 1	8	96.3	12.0375	0.202679
Column 2	8	102.6	12.825	0.079286
Column 3	8	108.8	13.6	0.251429

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	9.765833	2	4.882917	27.46334	1.38E-06	3.4668
Within Groups	3.73375	21	0.177798			
Total	13.49958	23				

Since $F > F_{crit}$, reject the null hypothesis

Question 5

Anova: Single Factor

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Row 1	6	26	4.333333	6.666667
Row 2	6	55	9.166667	8.166667
Row 3	6	28	4.666667	6.266667

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	87.44444	2	43.72222	6.21643	0.010806	3.68232
Within Groups	105.5	15	7.033333			
Total	192.9444	17				

Since $F > F_{crit}$, reject the null hypothesis**Question 6**

Answer :

Anova: Single Factor

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Column 1	7	23	3.285714	3.571429
Column 2	7	33	4.714286	8.904762
Column 3	7	83	11.85714	12.47619
Column 4	7	30	4.285714	34.2381

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	323.8214	3	107.9405	7.294449	0.001215	3.008787
Within Groups	355.1429	24	14.79762			
Total	678.9643	27				

Since $F > F_{crit}$, reject the null hypothesis

Question 7

Anova: Single Factor

SUMMARY				
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Column 1	8	57	7.125	17.55357
Column 2	8	103	12.875	13.83929
Column 3	8	122	15.25	13.35714

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	279.25	2	139.625	9.360335	0.001241	3.4668
Within Groups	313.25	21	14.91667			
Total	592.5	23				

Since $F > F_{crit}$, reject the null hypothesis**Question 8**

Anova: Single Factor

SUMMARY				
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Column 1	5	5.9	1.18	0.057
Column 2	5	1.9	0.38	0.102
Column 3	5	3.8	0.76	0.103

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	1.601333	2	0.800667	9.167939	0.003831	3.885294
Within Groups	1.048	12	0.087333			
Total	2.649333	14				

Since $F > F_{crit}$, reject the null hypothesis

Question 9

Anova: Single Factor

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Column 1	8	109	13.625	2.839286
Column 2	8	102	12.75	9.642857
Column 3	8	94	11.75	10.78571
Column 4	8	118	14.75	16.5

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	39.09375	3	13.03125	1.310732	0.290525	2.946685
Within Groups	278.375	28	9.941964			
Total	317.4688	31				

Since $F < F_{crit}$, do not reject null hypothesis**Question 10**

Anova: Single Factor

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Column 1	6	16.4	2.733333	0.342667
Column 2	6	19.5	3.25	0.323
Column 3	6	14.8	2.466667	0.358667

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	1.903333	2	0.951667	2.787179	0.093484	2.695173
Within Groups	5.121667	15	0.341444			
Total	7.025	17				

Since $F > F_{crit}$, reject the null hypothesis

Chapter 5 Review Exercises' Answers

Question 1

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies			
	Usually buy		
Gender	Store Brand	Name Brand	Total
Male	135	110	245
Female	212	143	355
Total	347	253	600

Calculations

fo-fe

-6.69167 6.691667

6.691667 -6.69167

Expected Frequencies			
	Usually buy		
Gender	Store Brand	Name Brand	Total
Male	141.6916667	103.3083333	245
Female	205.3083333	149.6916667	355
Total	347	253	600

(fo-fe)²/fe

0.316027 0.433444

0.218103 0.299138

Data	
Level of Significance	0.01
Number of Rows	2
Number of Columns	2
Degrees of Freedom	1

Results	
Critical Value	6.634896712
Chi-Square Test Statistic	1.266712092
p-Value	0.260384586
Do not reject the null hypothesis	

Null Hypothesis : the two attributes, gender and store or name products, are independent**Conclusion** : the gender and store product are independent.

Question 2

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies					
	TV Programme				
Age	Cartoon	Documentary	Comedy	Mystery	Total
1-10	54.00	8.00	8.00	5.00	75.00
11-20	11.00	33.00	29.00	26.00	99.00
21-30	13.00	27.00	34.00	27.00	101.00
31-40	18.00	25.00	41.00	48.00	132.00
41-50	14.00	24.00	42.00	25.00	105.00
51-60	26.00	26.00	40.00	23.00	115.00
61 and above	11.00	18.00	34.00	10.00	73.00
Total	147.00	161.00	228.00	164.00	700.00

Calculations

fo-fe				
38.25	-9.25	-16.43	-12.57	
-9.79	10.23	-3.25	2.81	
-8.21	3.77	1.10	3.34	
-9.72	-5.36	-1.99	17.07	
-8.05	-0.15	7.80	0.40	
1.85	-0.45	2.54	-3.94	
-4.33	1.21	10.22	-7.10	

Expected Frequencies					
	TV Programme				
Age	Cartoon	Documentary	Comedy	Mystery	Total
1-10	15.75	17.25	24.43	17.57	75.00
11-20	20.79	22.77	32.25	23.19	99.00
21-30	21.21	23.23	32.90	23.66	101.00
31-40	27.72	30.36	42.99	30.93	132.00
41-50	22.05	24.15	34.20	24.60	105.00
51-60	24.15	26.45	37.46	26.94	115.00
61 and above	15.33	16.79	23.78	17.10	73.00
Total	147.00	161.00	228.00	164.00	700.00

(fo-fe) ² /fe				
92.89	4.96	11.05	8.99	
4.61	4.60	0.33	0.34	
3.18	0.61	0.04	0.47	
3.41	0.95	0.09	9.43	
2.94	0.00	1.78	0.01	
0.14	0.01	0.17	0.58	
1.22	0.09	4.40	2.95	

Data	
Level of Significance	0.025
Number of Rows	7.00
Number of Columns	4.00
Degrees of Freedom	18.00

Results	
Critical Value	31.53
Chi-Square Test Statistic	160.22
p-Value	0.00000
Reject the null hypothesis	

Null Hypothesis : type of TV programme watched is not related to a person's age

Conclusion : the type of TV programme watched is related to a person's age.

Question 3

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies					
	Way obtain information				
Educational background	Internet	TV	Newspaper	Others	Total
Primary	25	46	52	5	128
Secondary	50	40	63	17	170
Tertiary	91	61	70	30	252
Total	166	147	185	52	550

Calculations

fo-fe			
-13.63	11.79	8.95	-7.10
-1.31	-5.44	5.82	0.93
14.94	-6.35	-14.76	6.17

Expected Frequencies					
	Way obtain information				
Educational background	Internet	TV	Newspaper	Others	Total
Primary	38.63	34.21	43.05	12.10	128
Secondary	51.31	45.44	57.18	16.07	170
Tertiary	76.06	67.35	84.76	23.83	252
Total	166	147	185	52	550

(fo-fe) ² /fe			
4.81	4.06	1.86	4.17
0.03	0.65	0.59	0.05
2.94	0.60	2.57	1.60

Data	
Level of Significance	0.05
Number of Rows	3
Number of Columns	4
Degrees of Freedom	6

Results	
Critical Value	12.5915872
Chi-Square Test Statistic	23.9349791
p-Value	0.00053684
Reject the null hypothesis	

Null Hypothesis : the way people obtain information is independent of their educational background.

Conclusion : The way of people obtain information is dependent of their educational background.

Question 4

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Number of visits to the doctor			
status	0-2	3-5	> 5	Total
Smoker	30	65	105	200
Nonsmoker	115	90	45	250
Total	145	155	150	450

Calculations

fo-fe		
-34.44	-3.89	38.33
34.44	3.89	-38.33

Expected Frequencies				
	Number of visits to the doctor			
status	0-2	3-5	> 5	Total
Smoker	64.44	68.89	66.67	200
Nonsmoker	80.56	86.11	83.33	250
Total	145	155	150	450

(fo-fe) ² /fe		
18.41	0.22	22.04
14.73	0.18	17.63

Data	
Level of Significance	0.01
Number of Rows	2
Number of Columns	3
Degrees of Freedom	2

Results	
Critical Value	9.21034
Chi-Square Test Statistic	73.20809
p-Value	1.27E-16
Reject the null hypothesis	

Null Hypothesis : there is no relation between smoking and the number of visits to the doctor.

Conclusion : there is relation between smoking and the number of visits to the doctor

Question 5

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Price			
Age	Below RM150 000	RM150 001–RM300 000	RM300 001 and above	Total
21-30	18	28	2	48
31-40	46	28	14	88
41-50	33	18	17	68
51 and above	11	14	11	36
Total	108	88	44	240

Calculations

fo-fe		
-3.60	10.40	-6.80
6.40	-4.27	-2.13
2.40	-6.93	4.53
-5.20	0.80	4.40

Expected Frequencies				
	Price			
Age	Below RM150 000	RM150 001–RM300 000	RM300 001 and above	Total
21-30	21.6	17.60	8.80	48
31-40	39.6	32.27	16.13	88
41-50	30.6	24.93	12.47	68
51 and above	16.2	13.20	6.60	36
Total	108	88	44	240

(fo-fe) ² /fe		
0.60	6.15	5.25
1.03	0.56	0.28
0.19	1.93	1.65
1.67	0.05	2.93

Data	
Level of Significance	0.05
Number of Rows	4
Number of Columns	3
Degrees of Freedom	6

Results	
Critical Value	12.59159
Chi-Square Test Statistic	22.29628
p-Value	0.00107
Reject the null hypothesis	

Null Hypothesis : the price of the house independent of the age of the owner.**Conclusion** : the price of the house dependent of the age of the owner

Question 6

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies			
	Type of contractor		
status of IC	Subcontractor P	Subcontractor Q	Total
Good	232	325	557
Defective	18	25	43
Total	250	350	600

Calculations

fo-fe	
-0.08	0.08
0.08	-0.08

Expected Frequencies			
	Type of contractor		
status of IC	Subcontractor P	Subcontractor Q	Total
Good	232.08	324.92	557
Defective	17.92	25.08	43
Total	250	350	600

(fo-fe) ² /fe	
0.00	0.00
0.00	0.00

Data	
Level of Significance	0.1
Number of Rows	2
Number of Columns	2
Degrees of Freedom	1

Results	
Critical Value	2.705543971
Chi-Square Test Statistic	0.000715747
p-Value	0.978656382
Do not reject the null hypothesis	

Null Hypothesis : the distribution of good and defective IC chips are the same for both subcontractors

Conclusion : There is no relation between the status of IC and type of contractor

Question 7

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Degree			
Rating	Bachelor	Master	Doctorate	Total
Excellent	19	17	9	45
Average	18	8	13	39
Poor	7	14	15	36
Total	44	39	37	120

Calculations

fo-fe		
2.50	2.38	-4.88
3.70	-4.68	0.98
-6.20	2.30	3.90

Expected Frequencies				
	Degree			
Rating	Bachelor	Master	Doctorate	Total
Excellent	16.50	14.63	13.88	45
Average	14.30	12.68	12.03	39
Poor	13.20	11.70	11.10	36
Total	44	39	37	120

(fo-fe) ² /fe		
0.38	0.39	1.71
0.96	1.72	0.08
2.91	0.45	1.37

Data	
Level of Significance	0.1
Number of Rows	3
Number of Columns	3
Degrees of Freedom	4

Results	
Critical Value	7.77944
Chi-Square Test Statistic	9.972544
p-Value	0.040893
Reject the null hypothesis	

Null Hypothesis : the degree of the instructor is not related to students' opinions about the teaching quality

Conclusion : the degree of the instructor is related to students' opinions about the teaching quality

Question 8

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Age Group			
Type of beverages	< 30	30 – 55	>55	Total
Milo	200	140	50	390
Nescafe	60	56	130	246
Horlick	40	84	153	277
Total	300	280	333	913

Calculations

fo-fe		
71.85	20.39	-92.25
-20.83	-19.44	40.28
-51.02	-0.95	51.97

Expected Frequencies				
	Age Group			
Type of beverages	< 30	30 – 55	>55	Total
Milo	128.15	119.61	142.25	390
Nescafe	80.83	75.44	89.72	246
Horlick	91.02	84.95	101.03	277
Total	300	280	333	913

(fo-fe)^2/fe		
40.29	3.48	59.82
5.37	5.01	18.08
28.60	0.01	26.73

Data	
Level of Significance	0.1
Number of Rows	3
Number of Columns	3
Degrees of Freedom	4

Results	
Critical Value	7.77944
Chi-Square Test Statistic	187.384
p-Value	1.93E-39
Reject the null hypothesis	

Null Hypothesis : there is no relationship between age group and beverage preference

Conclusion : there is a relationship between age group and beverage preference

Question 9

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies				
	Performance in training program			
Job Performance	Poor	Average	Excellent	Total
Poor	18	55	24	97
Average	22	74	53	149
Excellent	3	44	57	104
Total	43	173	134	350

Calculations

fo-fe		
6.08	7.05	-13.14
3.69	0.35	-4.05
-9.78	-7.41	17.18

Expected Frequencies				
	Performance in training program			
Job Performance	Poor	Average	Excellent	Total
Poor	11.92	47.95	37.14	97
Average	18.31	73.65	57.05	149
Excellent	12.78	51.41	39.82	104
Total	43	173	134	350

(fo-fe) ² /fe		
3.10	1.04	4.65
0.75	0.00	0.29
7.48	1.07	7.42

Data	
Level of Significance	0.01
Number of Rows	3
Number of Columns	3
Degrees of Freedom	4

Results	
Critical Value	13.2767
Chi-Square Test Statistic	25.78772
p-Value	3.49E-05
Reject the null hypothesis	

Null Hypothesis : that performance in the training program and job performance are independent.

Conclusion : that performance in the training program and job performance are dependent.

Question 10

fo = Observed Frequencies

fe = Expected Frequencies

Chi-Square Test

Observed Frequencies					
	Column variable				
Gender	Tennis court	Swimming pool	Gymnasium	Track	Total
Male	92	138	110	46	386
Female	88	151	98	37	374
Total	180	289	208	83	760

Calculations

fo-fe			
0.58	-8.78	4.36	3.84
-0.58	8.78	-4.36	-3.84

Expected Frequencies					
	Column variable				
Gender	Tennis court	Swimming pool	Gymnasium	Track	Total
Male	91.42	146.78	105.64	42.16	386
Female	88.58	142.22	102.36	40.84	374
Total	180	289	208	83	760

(fo-fe)^2/fe			
0.00	0.53	0.18	0.35
0.00	0.54	0.19	0.36

Data	
Level of Significance	0.025
Number of Rows	2
Number of Columns	4
Degrees of Freedom	3

Results	
Critical Value	9.348404
Chi-Square Test Statistic	2.152938
p-Value	0.541277
Do not reject the null hypothesis	

Null Hypothesis : there is no relationship between the facilities and gender

Conclusion : there is no relationship between the facilities and gender



REFERENCES

REFERENCES

- Caldwell, S. (2007). *Statistic Unplugged* (2nd ed). Belmont, CA: Thomson Wadsworth.
- Coakes, S. and L. Steed (2006). SPSS: Analysis without anguish: Version 13.0 for windows, John Wiley & Sons Australia.
- Gravetter, F. J. & Wallnau, L. B. (5th). (2005). *Essential of statistics for behavioral sciences*. Belmont: Thomson Learning.
- Hamburg, M. and P. Young (1994). Statistical analysis for decision making, Harcourt Brace Jovanovich New York.
- Healey, J. F. (2007). *The essentials of statistics: a tool for social research*. Belmont: Thomson Learning.
- Medcalc. (2009). Electronic references. Retrieved April 9, 2009 from http://www.medcalc.be/manual/scatter_diagram_regression_line.php
- Statistical Analysis Using SPSS handbook (2004), SPSS, Inc United States of America.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). Needham Heights, MA: Allyn and Bacon.



APPENDICES

Appendices

Appendix 1: Computational Formulas: Descriptive Statistics, Correlation, Regression and *t* Tests

Mean	$\bar{X} = \frac{\sum X}{n} \quad \mu = \frac{\sum X}{N}$
Median	$Median_{\text{odd number of scores}} = \left[\frac{n+1}{2} \right]^{th} \text{score}$ $Median_{\text{even number of scores}} = \frac{\left[\frac{n+2}{2} \right]^{th} \text{score} + \left[\frac{n}{2} \right]^{th} \text{score}}{2}$
Variance	$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$
Standard deviation	$S = \sqrt{S^2}$
z score	$z = \frac{X - \bar{X}}{S}$
Covariance	$cov_{XY} = \frac{\sum (X - \bar{X}) \cdot (Y - \bar{Y})}{n-1}$ $cov_{XY} = \frac{\sum XY - \frac{\sum X \cdot \sum Y}{n}}{n-1}$
Correlation	$r = \frac{cov_{XY}}{S_X \cdot S_Y}$ $r = \frac{(n \cdot \sum XY) - (\sum X \cdot \sum Y)}{\sqrt{[(n \cdot \sum X^2) - (\sum X)^2] \cdot [(n \cdot \sum Y^2) - (\sum Y)^2]}}$ $df = n - 2$
Coefficient of determination	r^2
Linear regression	$\hat{Y} = bX + a$ $b = \frac{cov_{XY}}{S_X^2} = r_{XY} = \frac{S_Y}{S_X} \quad \text{or} \quad b = \frac{cov_{XY}}{S_Y^2} = r_{XY} = \frac{S_X}{S_Y}$ $a = \bar{Y} - (b \cdot \bar{X}) \quad \text{or} \quad a = \bar{X} - (b \cdot \bar{Y})$
Standard error of the estimate	$S_{XY} = S_X \cdot \sqrt{1-r^2} \quad \text{or} \quad S_{YX} = S_Y \cdot \sqrt{1-r^2}$

Standard error of the mean	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
z test	$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$
Estimated standard error of the mean	$Est\sigma_{\bar{X}} = \frac{S}{\sqrt{n}}$
t Tests	$t = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad df = n - 1$ $t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{diff}} \quad est\sigma_{diff} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad df = (n_1 - 1) + (n_2 - 1)$ $est\sigma_{diff} = \sqrt{(estimated\sigma_{\bar{X}_1})^2 + (estimated\sigma_{\bar{X}_2})^2 - (2 \cdot cov)}$ $\sigma_{diff} = \sqrt{(estimated\sigma_{\bar{X}_1})^2 + (estimated\sigma_{\bar{X}_2})^2 - (2 \cdot r \cdot estimated\sigma_{\bar{X}_1} \cdot estimated\sigma_{\bar{X}_2})}$ $estimated\sigma_{diff} = \sqrt{\frac{\sum D^2}{n} - \bar{D}^2} \quad t = \frac{\bar{D}}{estimated\sigma_{diff}} \quad df = \text{number of pairs} - 1$

Appendix 2: Computational Formulas: One-Way Analysis of Variance

$$MS_{wg} = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2 + \sum (X_3 - \bar{X}_3)^2 + \cdots + \sum (X_k - \bar{X}_k)^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + \cdots + (n_k - 1) +}$$

$$MS_{bg} = \frac{n_1(\bar{X}_1 - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + n_3(\bar{X}_3 - \bar{\bar{X}})^2 + \cdots + n_k(\bar{X}_k - \bar{\bar{X}})^2}{k - 1}$$

$$\bar{\bar{X}} = \frac{\sum \bar{X}}{k} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \cdots + \bar{X}_k}{k}$$

$$\bar{\bar{X}} = \frac{\sum \sum X}{N_{total}} = \frac{\sum X_1 + \sum X_2 + \sum X_3 + \cdots + \sum X_k}{N_{total}}$$

$$F = \frac{MS_{bg}}{MS_{wg}} \quad df_{bg} = k - 1 \quad df_{wg} = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) + \cdots + (n_k - 1)$$

$$MS_{bg} = \frac{SS_{bg}}{df_{bg}} \quad MS_{wg} = \frac{SS_{wg}}{df_{wg}}$$

$$SS_{bg} = \left[\frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \cdots + \frac{(\sum X_k)^2}{n_k} \right] - \left[\frac{(\sum X_1 + \sum X_2 + \cdots + \sum X_k)^2}{N_{total}} \right]$$

$$SS_{wg} = \left[\sum X_1^2 + \sum X_2^2 + \cdots + \sum X_k^2 \right] - \left[\frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \cdots + \frac{(\sum X_k)^2}{n_k} \right]$$

$$SS_{total} = \left[\sum X_1^2 + \sum X_2^2 + \cdots + \sum X_k^2 \right] - \left[\frac{(\sum X_1 + \sum X_2 + \cdots + \sum X_k)^2}{N_{total}} \right]$$

$$SS_{total} = SS_{bg} + SS_{wg} \quad HSD = q \cdot \sqrt{\frac{MS_{wg}}{n}}$$

Appendix 3: Computational Formulas: Two-Way Analysis of Variance

$$SS_{total} = \sum \sum X^2 - \frac{(\sum \sum X)^2}{N_{total}}$$

$$SS_{wg} = \sum \sum X^2 - \frac{\sum (\sum X_{cell})^2}{n_{cell}} \quad SS_r = \frac{\sum (\sum X_{row})^2}{n_{row}} - \frac{(\sum \sum X)^2}{N_{total}}$$

$$SS_c = \frac{\sum (\sum X_{col})^2}{n_{col}} - \frac{(\sum \sum X)^2}{N_{total}} \quad SS_{rxc} = SS_{Total} - (SS_{wg} + SS_r + SS_c)$$

$$df_r = \text{number of rows} - 1 \quad df_c = \text{number of columns} - 1$$

$$df_{rxc} = df_r \cdot df_c = (\text{number of rows} - 1) \cdot (\text{number of columns} - 1)$$

$$df_{wg} = (n_{cell_1} - 1) + (n_{cell_2} - 1) + \dots + (n_{cell_k} - 1)$$

$$df_{wg} = N_{Total} - \text{number of cells} \quad df_{Total} = N_{Total} - 1$$

$$MS_r = \frac{SS_r}{df_r} \quad MS_c = \frac{SS_c}{df_c} \quad MS_{rxc} = \frac{SS_{rxc}}{df_{rxc}} \quad MS_{wg} = \frac{SS_{wg}}{df_{wg}}$$

$$F_r = \frac{MS_r}{MS_{wg}} \quad F_c = \frac{MS_c}{MS_{wg}} \quad F_{rxc} = \frac{MS_{rxc}}{MS_{wg}}$$

Appendix 4: Computational Formulas: Nonparametric Statistics

Chi-square	$X^2 = \sum \frac{f_o - f_e}{f_e}$ $f_e = \frac{\text{row total} \cdot \text{column total}}{\text{grand total}}$ $df = (\text{number of rows} - 1) \cdot (\text{number of columns} - 1)$
Mann Whitney U	$U_1 = (n_1 \cdot n_2) + \frac{n_1(n_1 + 1)}{2} - \sum R_1$ $U_2 = (n_2 \cdot n_1) + \frac{n_2(n_2 + 1)}{2} - \sum R_2$ $U_1 + U_2 = n_1 \cdot n_2$
Kruskal Wallis	$H = \left[\frac{12}{N_{total} \cdot N_{total} + 1} \right] \cdot \left[\frac{\sum R_1^2}{n_1} + \frac{\sum R_2^2}{n_2} + \dots + \frac{\sum R_k^2}{n_k} \right] - 3 \cdot N_{total} + 1$

Authors' Profiles



Tay Choo Chuan is a graduate with Bachelor of Science (Hons) degree in Mathematics, Master of Science degree (Quality and Productivity Improvement) and PhD in Mathematics from University Kebangsaan Malaysia (UKM). The author has over 20 years of experience teaching in two level of education: secondary and tertiary. He is currently attached to University Teknikal Malaysia Melaka (UTeM) as a senior lecturer in the Faculty of Electrical Engineering. He is also the author of several Mathematics books.



Mohd Razali Muhamad is a Professor in the Faculty of Manufacturing Engineering and Dean of the Centre for Graduate Studies, Universiti Teknikal Malaysia Melaka (UTeM). Prior to joining UTeM, he was a lecturer and an Associate Professor in the School of Mechanical Engineering and the School of Materials and Mineral Resources Engineering, Universiti Sains Malaysia. His research interests include, concurrent engineering, hard coating of materials, design of manufacturing systems, and management of technology. He has published journal and conference papers, as well as supervising master and PhD students on these topics. He has used SPSS extensively while completing his doctoral degree at the University of Liverpool, UK.



Tam Cai Lian is the lecturer from the School of Medicine and Health Sciences, Monash University Malaysia. She was the Assistant professor for Universiti Tunku Abdul Rahman (2005-2006). Dr. Tam is actively involved in writing commentaries on current issues for Dewan Masyarakat and local newspaper.



Sek Yong Wee holds a Bachelor of Science degree in Statistics from University Kebangsaan Malaysia. He also obtained MSc in Information Technology from University Putra Malaysia. He has served as a lecturer since 2002. He is an experienced lecturer in teaching mathematics, probability and statistics.



Siti Azirah Asmai is a graduate with Bachelor of Science (Hons) degree in Computer Science from Universiti Teknologi Malaysia(UTM), and Master of Science in IT for Engineers from Coventry University, United Kingdom. She is currently working toward the PhD Degree in IT at Universiti Teknikal Malaysia Melaka. Her research interests are in data analysis, artificial intelligent, prognostic modeling, decision support technology and simulation.



**Penerbit Universiti
Universiti Teknikal Malaysia Melaka**

ISBN 978-967-0257-04-4



9 789670 257044