

# **Advanced Techniques: ANOVA (SPSS 10.0)**

*SPSS Inc.  
233 S Wacker Drive, 11th Floor  
Chicago, Illinois 60606  
312.651.3300*

*Training Department  
800.543.2185*

*v10.0 Revised 1/17/00 hc/ss*

SPSS Neural Connection, SPSS QI Analyst, SPSS for Windows, SPSS Data Entry II, SPSS-X, SCSS, SPSS/PC, SPSS/PC+, SPSS Categories, SPSS Graphics, SPSS Professional Models, SPSS Advanced Models, SPSS Tables, SPSS Trends and SPSS Exact Tests are the trademarks of SPSS Inc. for its proprietary computer software. CHAID for Windows is the trademark of SPSS Inc. and Statistical Innovations Inc. for its proprietary computer software. Excel for Windows and Word for Windows are trademarks of Microsoft; dBase is a trademark of Borland; Lotus 1-2-3 is a trademark of Lotus Development Corp. No material describing such software may be produced or distributed without the written permission of the owners of the trademark and license rights in the software and the copyrights in the published materials.

General notice: Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

Copyright(c) 2000 by SPSS Inc.

All rights reserved.

Printed in the United States of America.

No part of this publication may be reproduced or distributed in any form or by any means, or stored on a database or retrieval system, without the prior written permission of the publisher, except as permitted under the United States Copyright Act of 1976.

# ADVANCED TECHNIQUES: ANOVA (SPSS 10.0) TABLE OF CONTENTS

<b>Chapter 1</b>	<b>Introduction</b>	
	Why do Analysis of Variance	1-1
	Visualizing Analysis of Variance	1-1
	What is Analysis of Variance?	1-3
	Variance of Means	1-4
	Basic Principle of ANOVA	1-6
	A Formal Statement of ANOVA Assumptions	1-8
<b>Chapter 2</b>	<b>Examining Data and Testing Assumptions</b>	
	Why Examine the Data?	2-2
	Exploratory Data Analysis	2-3
	A Look at the Variable Cost	2-5
	A Look at the Subgroups	2-9
	Normality	2-11
	Comparing the Groups	2-17
	Homogeneity of Variance	2-17
	Effects of Violations of Assumptions in ANOVA	2-19
<b>Chapter 3</b>	<b>One-Factor ANOVA</b>	
	Logic of Testing for Mean Differences	3-2
	Factors	3-2
	Running One-Factor ANOVA	3-3
	One Factor ANOVA Results	3-5
	Post Hoc Testing	3-7
	Why So Many Tests?	3-8
	Planned Comparisons	3-16
	How Planned Comparisons are Done	3-17
	Graphic the Results	3-19
	Appendix: Group Differences on Ranks	3-20
<b>Chapter 4</b>	<b>Multi-Way Univariate ANOVA</b>	
	The Logic of Testing, and Assumptions	4-2
	How Many Factors?	4-2
	Interactions	4-3
	Exploring the Data	4-5
	Two-Factor ANOVA	4-13
	The ANOVA Table	4-18
	Predicted Means	4-19
	Ecological Significance	4-20
	Residual Analysis	4-21
	Post Hoc Tests of ANOVA Results	4-22

	Unequal Samples and Unbalanced Designs	4-24
	Sums of Squares	4-25
	Equivalence and Recommendations	4-26
	Empty Cells and Nested Designs	4-26
<b>Chapter 5</b>	<b>Multivariate Analysis of Variance</b>	
	Why Perform MANOVA?	5-2
	How MANOVA Differs from ANOVA	5-3
	Assumptions of MANOVA	5-3
	What to Look for in MANOVA	5-4
	Significance Testing	5-4
	Checking the Assumptions	5-5
	The Multivariate Analysis	5-11
	Examining Results	5-17
	What if Homogeneity Failed	5-19
	Multivariate Tests	5-19
	Checking the Residuals	5-23
	Conclusion	5-25
	Post Hoc Tests	5-26
<b>Chapter 6</b>	<b>Within-Subject Designs: Repeated Measures</b>	
	Why Do a Repeated Measures Study?	6-2
	The Logic of Repeated Measures	6-2
	Assumptions	6-5
	Proposed Analysis	6-7
	Key Concept	6-7
	Comparing the Grade Levels	6-13
	Examining Results	6-19
	Planned Comparisons	6-26
<b>Chapter 7</b>	<b>Between and Within-Subject ANOVA: (Split-Plot)</b>	
	Assumptions of Mixed Model ANOVA	7-2
	Proposed Analysis	7-2
	A Look at the Data	7-2
	Summary of Explore	7-8
	Split-Plot Analysis	7-8
	Examining Results	7-12
	Tests of Assumptions	7-13
	Sphericity	7-14
	Multivariate Tests Involving Time	7-15
	Tests of Between-Subject Factors	7-15
	Averaged F Tests Involving Time	7-16
	Additional Within-Subject Factors and Sphericity	7-18
	Exploring the Interaction - Simple Effects	7-18
	Graphing the Interaction	7-25

<b>Chapter 8</b>	<b>More Split-Plot Design</b>	
	Introduction: Ad Viewing with Pre-Post Brand Ratings	8-1
	Setting Up the Analysis	8-2
	Examining Results	8-7
	Tests of Assumptions	8-8
	ANOVA Results	8-11
	Profile Plots	8-13
	Summary of Results	8-15
<b>Chapter 9</b>	<b>Analysis of Covariance</b>	
	How is Analysis of Covariance Done?	9-2
	Assumptions of ANCOVA	9-2
	Checking the Assumptions	9-3
	Baseline ANOVA	9-3
	ANCOVA - Homogeneity of Slopes	9-5
	Standard ANCOVA	9-7
	Describing the Relationship	9-8
	Fitting Non-Parallel Slopes	9-9
	Repeated Measures ANCOVA with a Single Covariate	9-11
	Repeated Measures ANCOVA with a Varying Covariate	9-16
	Further Variations	9-18
<b>Chapter 10</b>	<b>Special Topics</b>	
	Latin Square Designs	10-2
	An Example	10-2
	Complex Designs	10-6
	Random Effects Models	10-6
<b>References</b>	<b>References</b>	
	References	R-1
<b>Exercises</b>	<b>Exercises</b>	
	Exercises	E-1



# Chapter 1 Introduction

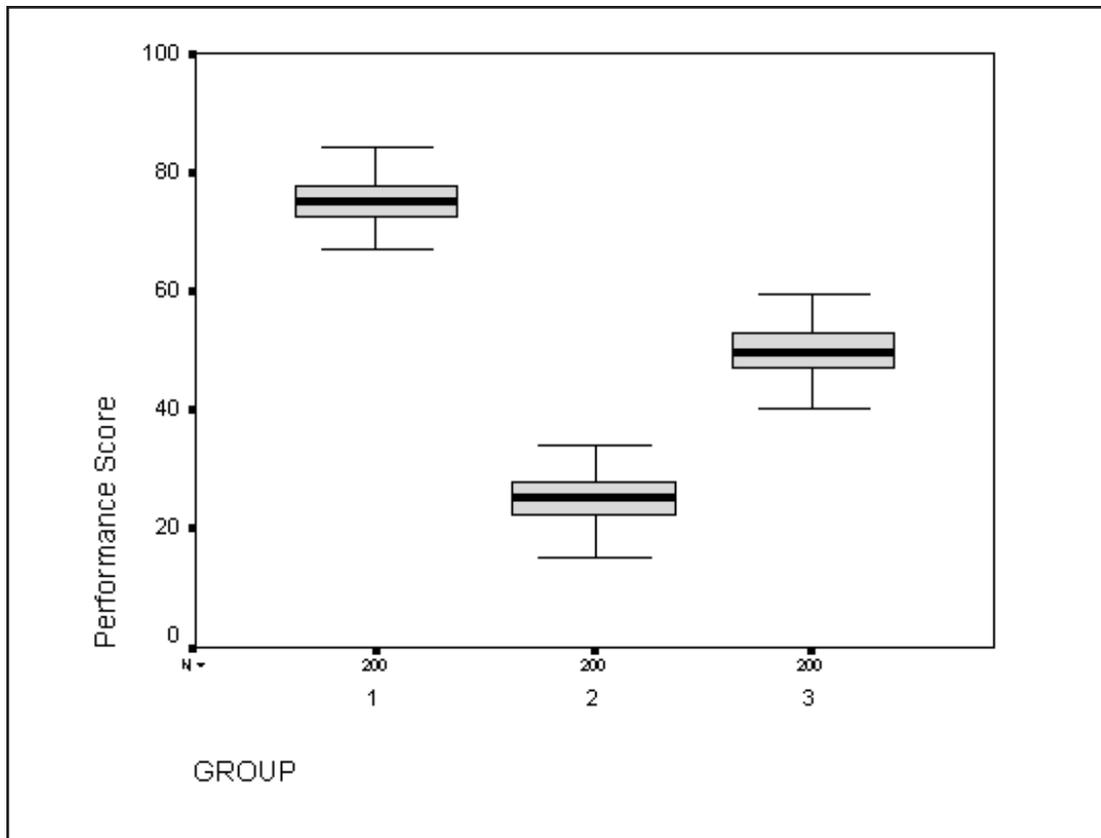
## WHY DO ANALYSIS OF VARIANCE?

**A**nalysis of variance is performed in order to determine whether there are differences in the means between groups or across different conditions. From a simple two-group experiment, to a complex study involving many factors and covariates, the same core principle applies. Why this technique is called analysis of variance (ANOVA) and not analysis of means, has to do with the methodology used to determine if the means are far enough apart to be considered “significantly” different.

## VISUALIZING ANALYSIS OF VARIANCE

To examine the basic principle of ANOVA, image a simple experiment in which subjects are randomly assigned to one of three treatment groups, the treatments are applied, then subjects are tested on some performance measure. One possible outcome appears below. Performance scores are plotted along the vertical axis and each box represents the distribution of scores within a treatment group.

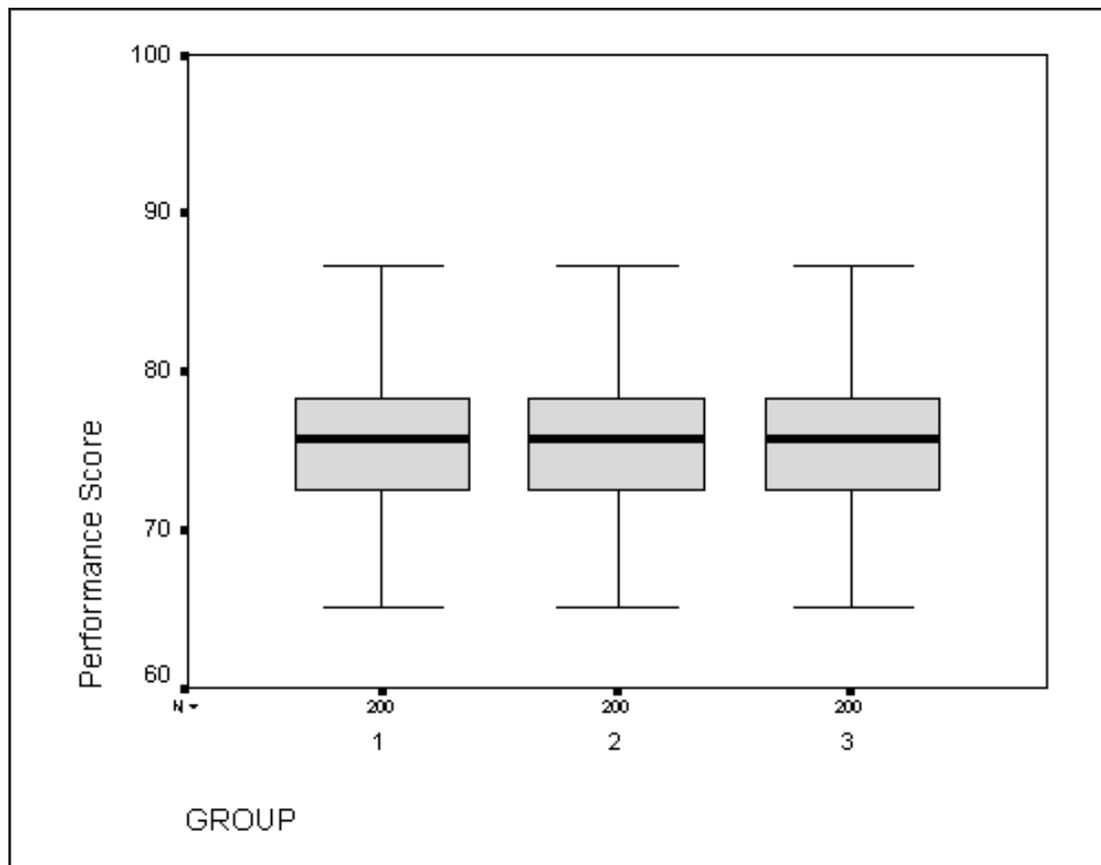
Figure 1.1 Performance Scores: Distinct Populations



Here a formal testing of the differences is almost unnecessary. The groups show no overlap in performance scores and the group means (medians are the dark bar at the center of each box) are well spaced relative to the standard deviation of each group. Think of the variation, or distances going from group mean to group mean, and compare this to the variation of the individual scores within each group.

Let us take another example. Suppose the same experiment described above results in the performance scores having little or no difference. We picture this below.

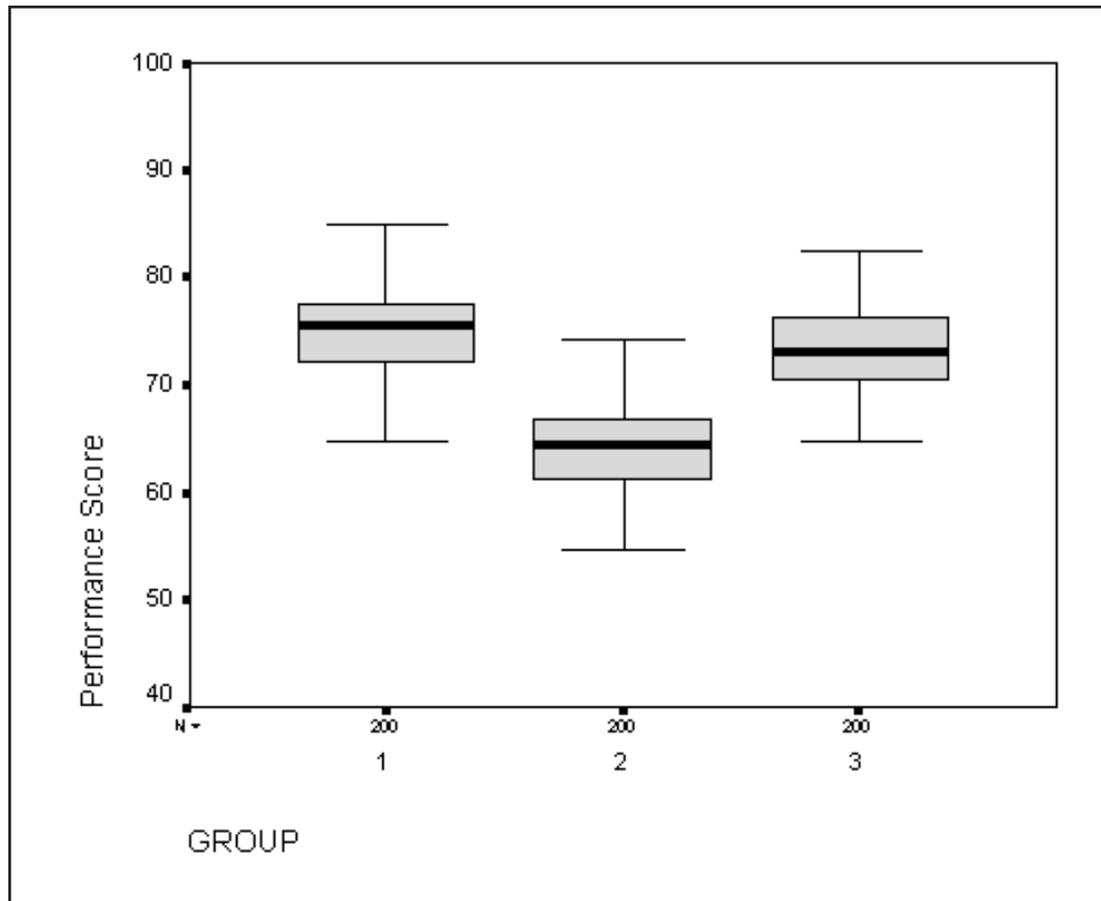
**Figure 1.2 Performance Scores: Identical Populations**



Here the group means are all but identical, so there is little variation or distance going from group mean to group mean compared to the variation of performance scores within the groups. A formal ANOVA analysis would merely confirm this.

A more realistic example involves groups with overlapping scores and group means that differ. This is shown in the plot below.

Figure 1.3 Performance Scores: Overlapping Groups



The formal ANOVA analysis needs to be done to determine if the group means do indeed differ in the population, that is, with what confidence can we claim that the group means are not the same. Once again, think of the variation of the group means (distances) between pairs of groups, or variation of the group means around the grand mean) relates to the variation of the performance scores within each group.

**WHAT IS ANALYSIS OF VARIANCE?**

Stripped of technical adjustments and distributional assumptions, you are comparing the variation of group means to the variation of individual scores within the groups constitute the basis for analysis of variance. To the extent that the differences or variation between groups is large relative to the variation of individual scores within the groups, we speak of the groups showing significant differences. Another way of reasoning about the experiment we described is to say that if the treatments applied to the three groups had no effect (no group differences), then the variation in group means should be due to the same sources and be of the same magnitude (after technical adjustments) as the variation among individuals within the groups.

### VARIANCE OF MEANS

The technical adjustment just mentioned is required when comparing variation in means scores to variation in individual scores. This is because the variance of means will be less than the variance of the individual scores on which the mean is based. The basic mathematical relation is that the variance of the means based on a sample size of " $n$ " will be equal to the variance of the individual scores in the sample divided by " $n$ ". The standard deviation of the mean is called the standard error or the standard error of the mean. We will illustrate this law with a little under 10,000 observations produced by a pseudo-random number generator in SPSS, based on a normal distribution with a mean of zero and a standard deviation of one. The results appear in Figure 1.4.

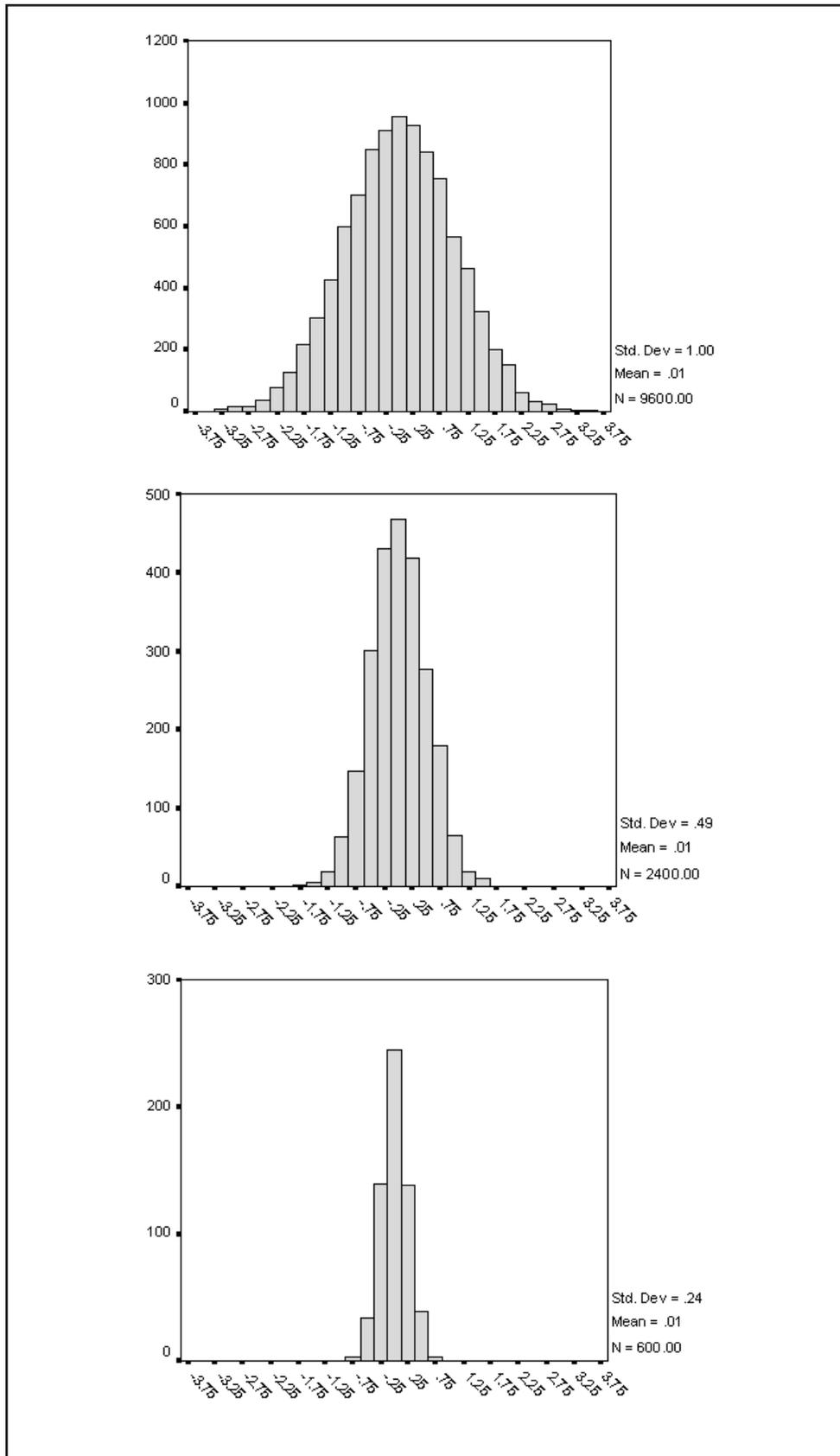
The first histogram shows the distribution of the original 9,600 data points. Notice almost all of the points fall between the values of  $-3$  and  $+3$ .

The second histogram contains the mean scores of samples of size 4 drawn from the original 9,600 data points. Each point is a mean score for a sample of size 4 for a total of 2,400 data points. The distribution of means is narrower than that of the first histogram; almost all the points fall between  $-1.5$  and  $+1.5$ .

In the final histogram each point is a mean of 16 observations from the original sample. The variation of these 600 points is less than that of the previous histograms with most points between  $-.9$  and  $+.9$ . Despite the decrease in variance, the means (or centers of the distributions) remain at zero.

This relation is relevant to analysis of variance. In ANOVA, when comparing the variation between group mean scores to variation of individuals within groups, the sample sizes upon which the means are based are explicitly taken into account.

Figure 1.4 Variation in Means as a Function of Sample Size



### **BASIC PRINCIPLE OF ANOVA**

While we will give a formal statement of the assumptions of ANOVA and proceed with complex variations, this basic principle comparing the variation of group or treatment means to the variation of individuals within groups (or some other grouping) will be the underlying theme.

### **A FORMAL STATEMENT OF ANOVA ASSUMPTIONS**

The term “factor” denotes a categorical predictor variable. “Dependent Variables” are interval level outcome variables, and “covariates” are interval level predictor variables. ANOVA is considered a form of the general linear model and most of the assumptions follow from that and are listed below:

- All variables must exhibit independent variance. In other words, a variable must vary, and it must not be a one-to-one function of any other variable. Though it is the dream of any data analyst to have a dependent variable that is perfectly predicted, if such were the case, the “F-ratio” for an analysis of variance could not be formed (Note: as a practical matter, if you find such a perfect prediction, lack of an “F-ratio” should not result in any lost sleep).
- Dependent variables and covariates must be measured in interval or ratio scale. Factors may be nominal or categorized from ordinal or interval variables. However, ordinal hypothesis can only be tested in a pairwise fashion. Imposing the desired metric through the appropriate set of contrasts can test interval hypothesis.
- For fixed effect models, all levels of predictor variables that are of interest must be included in the analysis.
- The linear model specified is the correct one; it includes all the relevant sources of variation, excludes all irrelevant ones, and is correct in its functional form (Note: in the words of the Sgt. in Hill Street Blues “so, be careful out there”).
- Errors of measurement must be unbiased (have a zero mean).
- Errors must be independent of each other and of the predictor variables.
- Error variances must be homogeneous.
- Errors must be normally distributed. This final assumption is not required for estimation, but must be met in order for an “F-ratio” to be accurately referred to as an “F-distribution” (Note: that is, it is required for testing, which is why you are doing the analysis).

We will examine some of these assumptions in the data sets used in the rest of this course.

We use an “analysis of variance” to test for differences between means for the following formal reason:

The formulation of the analysis of variance approach as a test of equality of means follows a deductive format. We can show that if it is true that two (or more) means are equal, then certain properties must hold for other functions of the data, such as between group and within group variation. The idea

behind the formulation of the familiar “F-ratio” is that if the means being compared are equal, then the numerator and denominator of the “F-ratio” represent independent estimates of the same quantity (error variance) and their ratio must then follow a known distribution. This allows us to place a distinct probability on the occurrence of sample means as different as those observed under the hypothesis of zero difference among population means.

### **SUMMARY**

In this chapter we discussed the basic principle of analysis of variance and gave a formal statement of the assumptions of the model. We turn next to examining these assumptions and the implications if the assumptions are not met (Note: life as it really is).



# Chapter 2 Examining Data and Testing Assumptions

## DESCRIPTION OF THE DATA

The data set comes from Cox and Snell (1981). They obtained it from a report (Mooz, 1978) and reproduced it with the permission of the Rand Corporation. Only a subset of the original variables is used in the data set we will use.

The data set we will be using contains information for 32 light water nuclear power plants. Four variables are included: the capacity and cost of the plant; time to completion from start of construction; and experience of the architect-engineer who built the plant. These variables are described in more detail below.

We will use only a subset of all the variables that were in the original data set, and have created categories from the variables capacity and experience in order to use them as factors in an analysis of variance.

In order of the variables in the data file, they are:

Capacity	Generating capacity
1	Less than 800 MW's (Mega Watts)
2	800-1000
3	Greater than 1000

Experience	Experience of the architect-engineer in building power plants
1	1-3 plants
2	4-9 plants
3	10 or more plants

Time time in months between issuing of construction permit and issuing of operating license.

Cost cost in millions of dollars adjusted to a 1976 base (In 1976 dollars).

## Note About the Analyses That Follow

The analyst should choose the analysis that best conforms to the type of information collected in the data and the research or analysis question(s) you wish to answer. We feel that in a short course there is an advantage in describing the various types of analyses that can be done. However, in practice you would run only the most appropriate analysis. In other words if there were two factors in your study, you would run a two-factor analysis and not begin with one factor analysis as we do here.

### **Research Question(s)**

The researcher should state the research questions clearly and concisely, and refer to these questions regularly as the design and implementation of the study progresses. Without this statement of questions, it is easy to deviate from them when engrossed in the details of planning or to make decisions that are at variance with the questions when involved in a complex study. Translating study objectives into questions serves as a check on whether the study has met the objectives.

The next task is to analyze the researchable question(s). In doing this one must

- Identify and define key terms
- Identify sub questions, which must also be answered
- Identify the scope and time frame imposed by the researchable question

### **Data to be Collected**

One of the most important decisions that should not be overlooked is to set down in terms of utmost clarity exactly what information is needed. It is usually good procedure to verify that all the data are relevant to the purposes of the study and that no essential data are omitted. Unless this is specified, the reporting forms may yield information that is quite different from what is needed, since there is a tendency to request too much data, some of which is subsequently never analyzed.

### **Know the Data**

It is critical that the researcher be familiar with the data being analyzed, whether it is primary (data you collected) or secondary (someone else collected it) data. Not only is knowing your data important to defining your population, but it can (1) help to spot trends on which to focus, and (2) provide assurance that you are measuring what you want to measure.

### **Scan the Data**

Visually review the data for several cases (or the entire data set if it is relatively small). Be familiar with the meaning of every variable and with the codes associated with the variables of interest.

### **WHY EXAMINE THE DATA?**

Before applying formal tests (ANOVA for example in this course) to your data, it is important to first examine and check the data. This is done for several reasons:

- To identify data errors
- To identify unusual points – outliers
- To become aware of unexpected or interesting patterns
- To check on or test the assumptions of the planned analysis
- For ANOVA:
  - Homogeneity of variance
  - Normality of error

### EXPLORATORY DATA ANALYSIS

Bar charts and histograms, as well as such summaries as means and standard deviations have been used in statistical work for many years. Sometimes such summaries are ends in their own right; other times they constitute a preliminary look at the data before proceeding with more formal methods. Seeing limitations in this standard set of procedures, John Tukey, a statistician at Princeton and Bell Labs, devised a collection of statistics and plots designed to reveal data features that might not be readily apparent from standard statistical summaries. In his book describing these methods, entitled Exploratory Data Analysis (1977), Tukey described the work of a data analyst to be similar to that of a detective, the goal being to discover surprising, interesting, and unusual things about the data. To further this effort Tukey developed both plots and data summaries. These methods, called exploratory data analysis and abbreviated EDA, have become very popular in applied statistics and data analysis. Exploratory data analysis can be viewed either as an analysis in its own right, or as a set of data checks and investigations performed before applying inferential testing procedures.

These methods are best applied to variables that have at least ordinal (more commonly interval) scale properties and can take on many different values. The plots and summaries would be less helpful for a variable that takes on only a few values (for example, on five point rating scales)

#### Plan of Analysis

We will use the SPSS **EXPLORE** procedure to examine the data and test some of the ANOVA assumptions. In windows we first open the file.

#### Note on Course Data Files

All files for this class are located in the c:\Train\Anova folder on your training machine. If you are not working in an SPSS Training center, the training files can be copied from the floppy disk that accompanies this course guide. If you are running SPSS Server (click File..Switch Server to check), then you should copy these files to the server or a machine that can be accessed (mapped from) the computer running SPSS Server.

#### A Note About Variable Names and Labels in Dialog Boxes

SPSS can display either variable names or variable labels in dialog boxes. In this course we display the variable names in alphabetical order. In order to match the dialog boxes shown here:

Click **Edit..Options**

Within the **General** tab of the Options dialog:

Click the **Display names** and **Alphabetical** option buttons in the Display Variables area

Click **OK**.

Click **File..Open..Data** (move to the c:\Train\Anova directory)

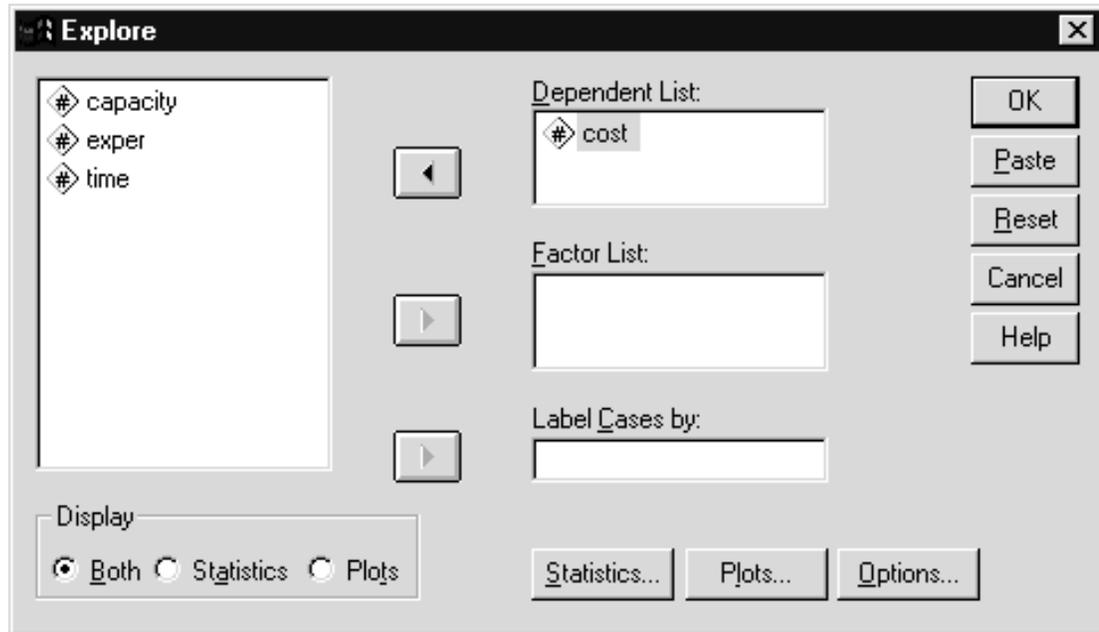
Select **SPSS Portable file (.por)** from **Files of Type** list

Double-click on **Plant.por** to open the file.

Click on **Analyze..Descriptive Statistics..Explore**

Move the **cost** variable into the **Dependent List** box

Figure 2.1 Explore Dialog Box



The syntax for running the Explore procedure is given below:

```
EXAMINE  
  VARIABLES=cost  
  /PLOT BOXPLOT STEMLEAF  
  /COMPARE GROUP  
  /STATISTICS DESCRIPTIVES  
  /CINTERVAL 95  
  /MISSING LISTWISE  
  /NOTOTAL.
```

The variable to be summarized (here cost) appears in the Dependent List box. The Factor list box can contain one or more categorical (for example, in our data set capacity) variables, and if used would cause the procedure to present summaries for each subgroup based on the factor variable(s). We will use this feature later in this chapter when we want to see differences between the groups. By default, both plots and statistical summaries will appear. We can request specific statistical summaries and plots using the Statistics and Plots pushbuttons. While not discussed here, the Explore procedure can print robust mean estimates (M-estimators) and lists of extreme values, as well as normal probability and homogeneity plots.

Click **OK** to run the Explore procedure.

**A LOOK AT THE VARIABLE COST**

The Explore procedure provides for us in this first run a summary of the variable cost for all 32 plants.

**Figure 2.2 Descriptives for the Variable Cost**

			Statistic	Std. Error
COST	Mean		461.5603	30.0734
	95% Confidence Interval for Mean	Lower Bound	400.2253	
		Upper Bound	522.8954	
	5% Trimmed Mean		455.6733	
	Median		448.1050	
	Variance		28941.042	
	Std. Deviation		170.1207	
	Minimum		207.51	
	Maximum		881.24	
	Range		673.73	
	Interquartile Range		321.7400	
	Skewness		.500	.414
	Kurtosis		-.456	.809

Explore first displays information about missing data. The Case Process Summary pivot table (not shown) displays the number of valid and missing observations; this information appears at the beginning of the statistical summary. Here we have data for the variable cost for all 32 observations. (Typically an analyst does not have all the data.)

**Measures of Central Tendency**

Next several measures of central tendency appear. Such statistics attempt to describe, with a single number, where the data values are typically found, or the center of the distribution. The mean is the arithmetic average. The median is the value at the center of the distribution when it is ordered (either lowest to highest or highest to lowest), that is, half the data values are greater than, and half the data values are less than, the median. Medians are resistant to extreme scores, and so are considered to be a robust measure of central tendency. The 5% trimmed mean is the mean calculated after the extreme upper 5% and the extreme lower 5% of the data values are dropped from the calculation. Such a measure would be resistant to small numbers of extreme or wild scores. In this case the three measures of central tendency are similar (461.56, 448.11, and 455.67), and we can say that the typical plant costs about \$450 million. If the mean were considerably above or below the median and the trimmed mean, it would suggest a

skewed or asymmetric distribution. A perfectly symmetric distribution, for example, the normal, would produce identical expected means, medians, and trimmed means.

### Variability Measures

Explore provides several measures of the amount of variation across the plants. They indicate to what degree observations tend to cluster near the center of the distribution. Both the standard deviation and variance (standard deviation squared) appear. For example, if all the observations were located at the mean then the standard deviation would be zero. In this case the standard deviation is \$170.12 (million). Another way to express the variability is that the standard deviation is 36.86% of the mean, which indicates that the data is moderately variable. The standard error is an estimate of the standard deviation of the mean if repeated samples of the same size were taken from the same population (\$30.07). It is used in calculating the 95% confidence interval for the sample mean discussed below. Also appearing is the interquartile range, which is essentially the range between the 25<sup>th</sup> and 75<sup>th</sup> percentile values. Thus the interquartile range represents the range including the middle 50 percent of the sample (321.74). It is a variability measure more resistant to extreme scores than the standard deviation. We also see the minimum and maximum dollar amounts and the range. It is useful to check the minimum and maximum to make sure no impossible data values are recorded (here a cost at zero or below).

### Confidence Interval for Mean

The 95% confidence interval has a technical definition: if we were to repeatedly perform the study and computed the confidence intervals for each sample drawn, on average, 95 out of each 100 such confidence intervals would contain the true population mean. It is useful in that it combines measures of both central tendency (mean) and variation (standard error) to provide information about where we should expect the population mean to fall. Here, we can say that we estimate the cost of the light water nuclear power plants to be \$461.56 and we are 95-percent confident that the true but unknown cost would be between \$400.23 and \$522.90.

The 95% confidence interval for the mean can be easily obtained from the sample mean, standard deviation, and sample size. The confidence interval is based on the sample mean, plus or minus 1.96 times the standard error of the mean. (1.96 is used because 95% of the area under a normal curve is within 1.96 standard deviation of the mean [when doing in my head I cheat and use 2 since it is easier to multiply by]). Since the sample standard error of the mean is simply the sample standard deviation divided by the square root of the sample size, the 95% confidence interval is equal to the sample mean plus or minus 1.96 times (sample standard deviation divided by {square root of the sample size}). Thus if you have the sample mean, sample standard deviation, and the sample size, you can easily compute the 95-percent confidence interval.

### Shape of the Distribution

Skewness and Kurtosis provide numeric summaries about the shape of the distribution of the data. While many analysts are content to view histograms in order to make judgments regarding the distribution of a variable, these measures quantify the shape. Skewness is a measure of the symmetry of a distribution. It is normed so that a symmetric distribution has zero skewness. Positive skewness indicates bunching of the data on the left and a longer tail on the right (for example, income distribution in the U.S.); negative skewness follows the reverse pattern (long tail on the left and bunching of the data on the right). The standard error of skewness also appears, and we can use it to determine if the data are significantly skewed. In our case, the skewness is .5 with a standard error of .414. Thus, using the formula above the 95-percent confidence interval for skewness is between  $-0.311$  and  $+1.311$ . Since the interval contains zero the data is not significantly skewed. (As a quick and dirty rule of thumb, however, if the skewness is over 3 in either direction you might want to consider a different approach in your study.)

Kurtosis also has to do with the shape of a distribution and is a measure of how peaked the distribution is. It is normed to the normal curve (kurtosis is zero). A curve that is more peaked than the normal has a positive value and one that is flatter than the normal has negative kurtosis. Again our data is not significantly peaked. (Again the same rule of thumb can be applied although some say that the value should be larger). The shape of the distribution can be of interest in its own right. Also, assumptions are made about the shape of the data distribution within each group when performing significance tests on mean differences between groups. (As a quick rule of thumb, however, if the kurtosis is over 3 in either direction you might want to consider a different approach in your study.)

### Stem & Leaf Plot

The stem & leaf plot is modeled after the histogram, but is designed to provide more information. Instead of using a standard symbol (for example, an asterisk "\*" or block character) to display a case or group of cases, the stem & leaf plot uses data values as the plot symbols. Thus the shape of the distribution is shown and the plot can be read to obtain specific data values. The stem & leaf plot for the cost appears below:

Figure 2.3 Stem & Leaf Plot for Cost

COST Stem-and-Leaf Plot		
Frequency	Stem	Leaf
8.00	2 .	01778888
4.00	3 .	1459
10.00	4 .	0124556799
1.00	5 .	6
7.00	6 .	0245699
1.00	7 .	1
1.00	8 .	8
Stem width:	100.00	
Each leaf:	1 case(s)	

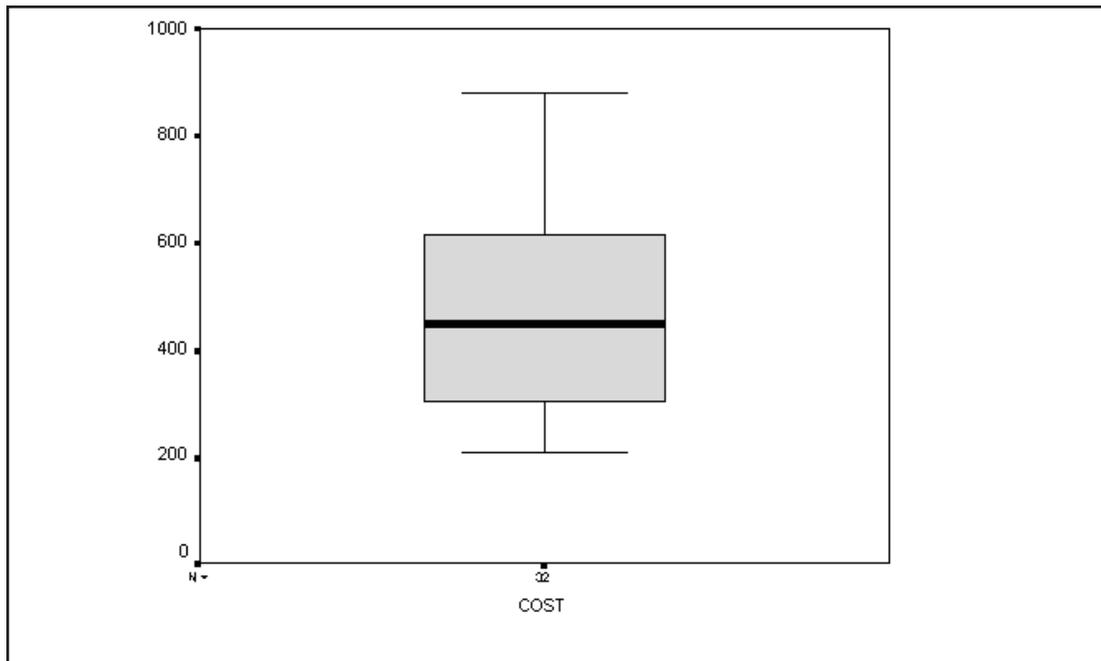
In a stem & leaf plot the stem is the vertical axis and the leaves branch horizontally from the stem (Tukey devised the stem & leaf). The stem width indicates how to interpret the units in the stem; in this case a stem unit represents one hundred dollars in the cost scale. The actual numbers in the chart (leaves) provide an extra decimal place of information about the data values. For example the stem of 5 and a leaf of 6 would indicate a cost of \$560 to \$569. Thus besides viewing the shape of the distribution we can pick out individual scores. Below the diagram a note indicates that each leaf represents one case. For large samples a leaf may represent two or more cases and in such situations an ampersand (&) represents two or more cases that have different data values.

The last line identifies outliers. These are data points far enough from the center of the distribution (defined more exactly under Box & Whisker plots below) that they might merit more careful checking – extreme points might be data errors or possibly represent a separate subgroup. If the stem & leaf plot were extended to include these outliers the skewness would be apparent.

**Box & Whisker Plot**

The stem & leaf plot attempts to describe data by showing every observation. In comparison, displaying only a few summaries, the box & whisker plot will identify outliers (data values far from the center of the distribution). Below we see the box & whisker plot (also called a box plot) for cost.

**Figure 2.4 Box & Whisker Plot for Cost**



The vertical axis is the cost of the plants. In the plot, the solid line inside the box represents the median. The “hinges” provide the top and

bottom borders to the box; they correspond to the 75<sup>th</sup> and 25<sup>th</sup> percentile values of cost, and thus define the interquartile range (IQR). In other words, the middle 50% of the data values fall within the box. The “whiskers” are the last data values that lie within 1.5 box lengths (or IQRs) of the respective hinge (edge of box). Tukey considers data points more than 1.5 box lengths from the hinges to be far enough from the center to be noted as outliers. Such points are marked with a circle. Points more than 3 box lengths from the hinges are viewed by Tukey to be “far out” points and are marked with an asterisk type symbol. This plot has no outliers or far-out points. If a single outlier appears at a given data value, the case sequence number prints out beside it (an id variable can be substituted), which aids data checking.

If the distribution were symmetric, then the median would be centered within the hinges and the whiskers. In the plot above, the different lengths of the whiskers show the skewness. Such plots are also useful when comparing several groups, as we will see shortly.

### A LOOK AT THE SUBGROUPS

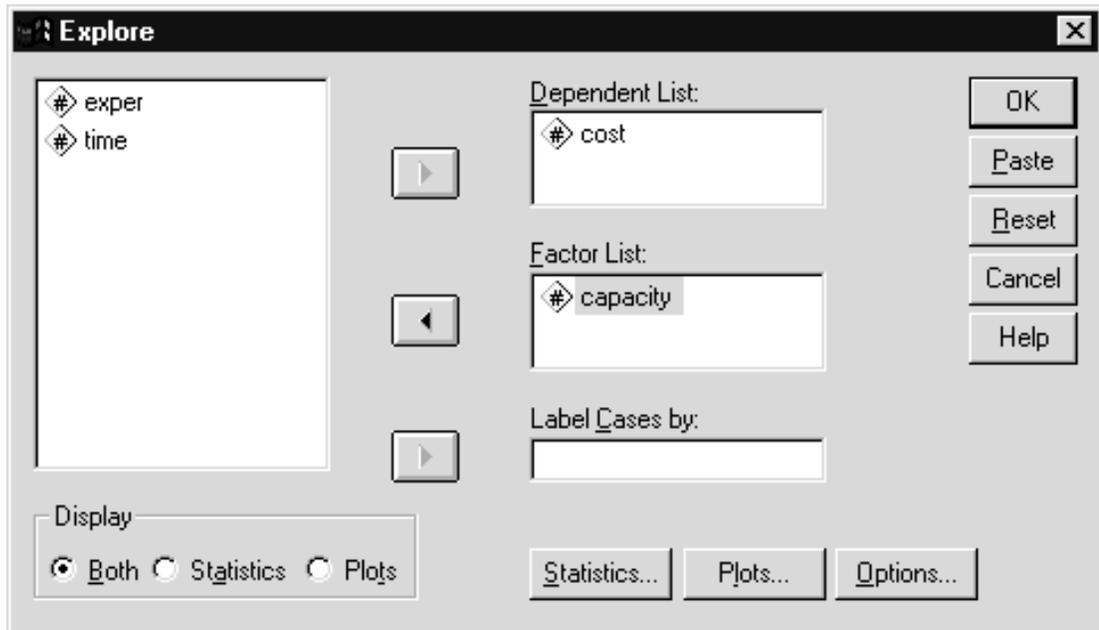
We now produce the same summaries and plots for each subgroup (here based on plant capacity).

Click on the **Dialog Recall** tool  on the toolbar.

Click on the **Explore** procedure

When the dialog box opens move the variable **capacity** to the **Factors List** box.

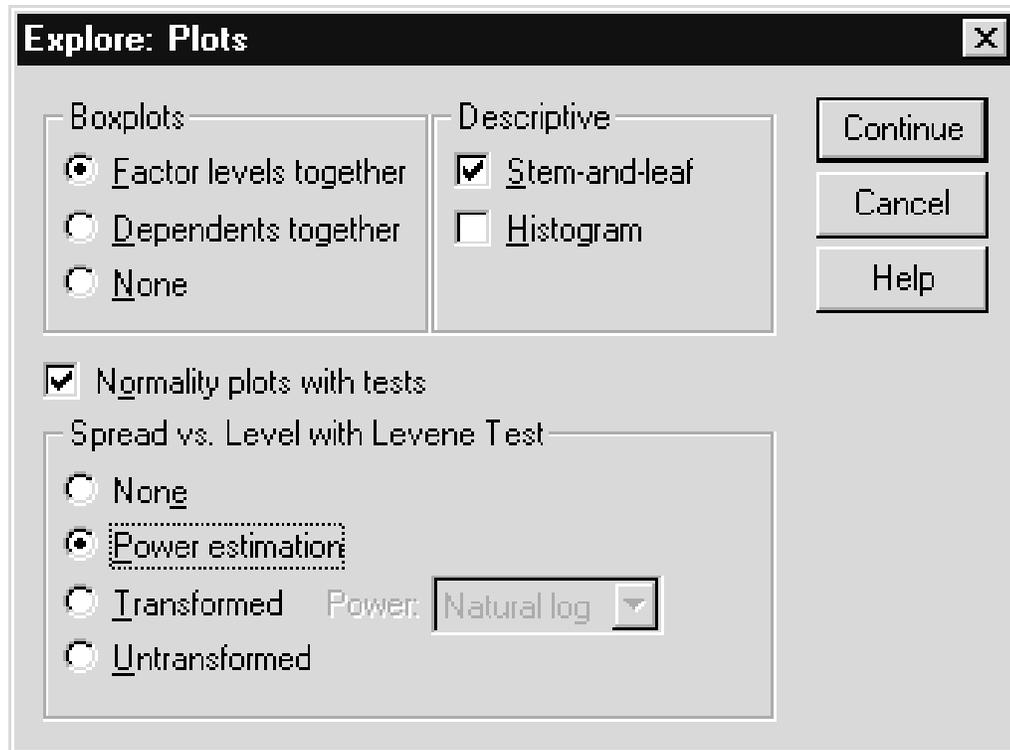
Figure 2.5 Explore Dialog Box



We also request normality plots and homogeneity tests.

Click **Plots** pushbutton  
Click **Normality plots with tests** check box  
Click **Power estimation** option button

Figure 2.6 Plots Sub-Dialog Box



Click **Continue**  
Click **OK**

The command below will run the analysis

```
EXAMINE  
  VARIABLES=cost BY capacity  
  /PLOT BOXPLOT STEMLEAF NPLOT SPREADLEVEL  
  /COMPARE GROUP  
  /STATISTICS DESCRIPTIVES  
  /CINTERVAL 95  
  /MISSING LISTWISE /NOTOTAL.
```

The Npplot keyword on the /Plot subcommand requests the normal probability plots, while the Spreadlevel keyword will produce the spread & level plots and the homogeneity of variance tests.

Below we see the statistics and the stem & leaf plot for the first capacity group (under 800 MW). Notice that relative to the group (not the entire set of plants as in the previous plots) there is an extreme score.

Figure 2.7 Descriptives for the First Group

CAPACITY			Statistic	Std. Error
COST < 800 MWe	Mean		400.8615	39.0447
	95% Confidence Interval for Mean	Lower Bound	315.7905	
		Upper Bound	485.9325	
	5% Trimmed Mean		395.5295	
	Median		402.5900	
	Variance		19818.303	
	Std. Deviation		140.7775	
	Minimum		207.51	
	Maximum		690.19	
	Range		482.68	
	Interquartile Range		163.4100	
	Skewness		.702	.616
	Kurtosis		.343	1.191

Figure 2.8 Stem & leaf Plot for the First Group

```

COST Stem-and-Leaf Plot for
CAPACITY= < 800 MWe

  Frequency      Stem & Leaf

    3.00         2 . 018
    3.00         3 . 145
    5.00         4 . 01267
    .00          5 .
    1.00         6 . 2
    1.00 Extremes      (>=690)

Stem width:      100.00
Each leaf:       1 case(s)
    
```

**NORMALITY**

The next pair of plots provides some specific information about the normality of data points within the group. This is equivalent to examining the normality of the residuals in ANOVA and is one of the assumptions made when the “F” tests of significance are made.

Figure 2.9 Q-Q Plot of the First Group

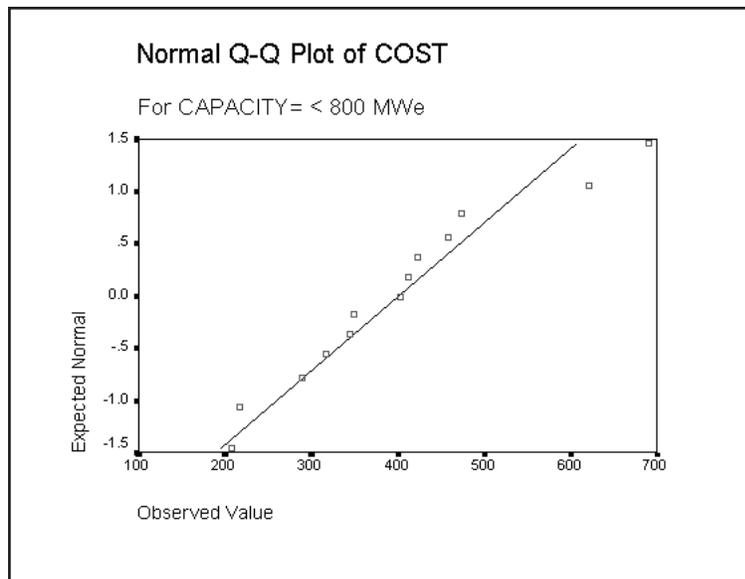
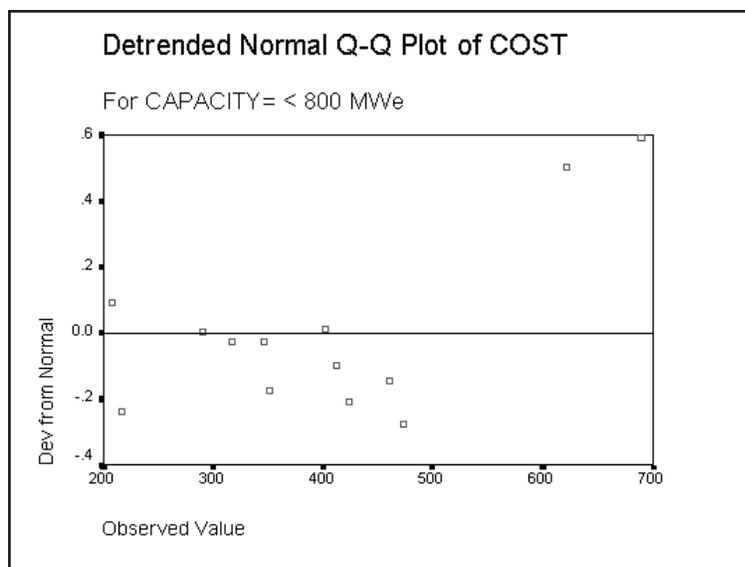


Figure 2.10 Detrended Q-Q Plot of the First Group



The first plot is called a normal probability plot. Each point is plotted with its actual value on the horizontal axis and its expected normal deviate value (based on the point's rank-order within the group). If the data follow a normal distribution, the points form a straight line.

The second plot is a detrended normal plot. Here the deviations of each point from a straight line (normal distribution) in the previous plot are plotted against the actual values. Ideally, they would distribute randomly around zero.

Next we look at the second group.

Figure 2.11 Descriptives for the Second Group

CAPACITY			Statistic	Std. Error	
COST	800-1000 MWe	Mean	436.6482	52.3574	
		95% Confidence Interval for Mean	Lower Bound	319.9886	
			Upper Bound	553.3078	
		5% Trimmed Mean	430.5546		
		Median	394.3600		
		Variance	30154.305		
		Std. Deviation	173.6500		
		Minimum	270.71		
		Maximum	712.27		
		Range	441.56		
		Interquartile Range	328.4400		
		Skewness	.484	.661	
		Kurtosis	-1.584	1.279	

Figure 2.12 Stem & Leaf Plot for the Second Group

```

COST Stem-and-Leaf Plot for
CAPACITY= 800-1000 MWe

  Frequency      Stem & Leaf

    5.00         2 . 77888
    1.00         3 . 9
    1.00         4 . 5
    1.00         5 . 6
    2.00         6 . 06
    1.00         7 . 1

Stem width:      100.00
Each leaf:       1 case(s)
    
```

For the second group the stem & leaf plot shows a concentration of costs at the low end.

Figure 2.13 Q-Q Plot for the Second Group

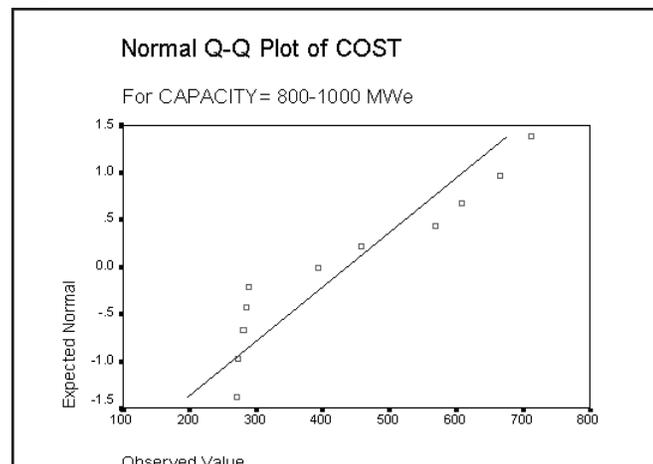
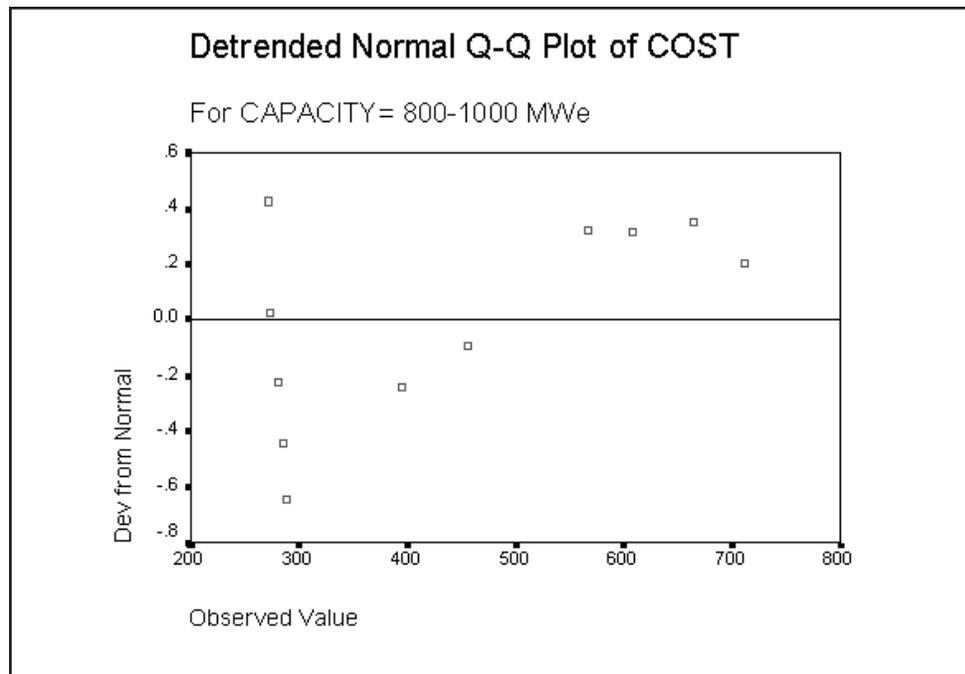


Figure 2.14 Detrended Q-Q Plot for the Second Group



The pattern from the stem & leaf plot carries over to the normal probability plot where the cluster of low cost values show in the lower left corner of the plot.

Let us examine the results for the third group.

Figure 2.15 Descriptives for the Third Group

CAPACITY			Statistic	Std. Error
COST	> 1000 MWe	Mean	594.4500	53.7536
		95% Confidence Interval for Mean	Lower Bound	467.3430
		Upper Bound	721.5570	
		5% Trimmed Mean	586.9189	
		Median	568.9050	
		Variance	23115.582	
		Std. Deviation	152.0381	
		Minimum	443.22	
		Maximum	881.24	
		Range	438.02	
		Interquartile Range	223.4725	
		Skewness	.900	.752
		Kurtosis	.250	1.481

Figure 2.16 Stem & Leaf Plot for the Third Group

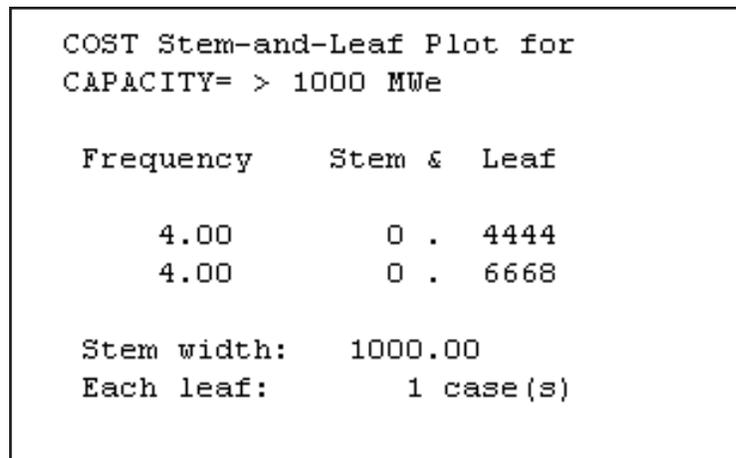


Figure 2.17 Q-Q Plot for the Third Group

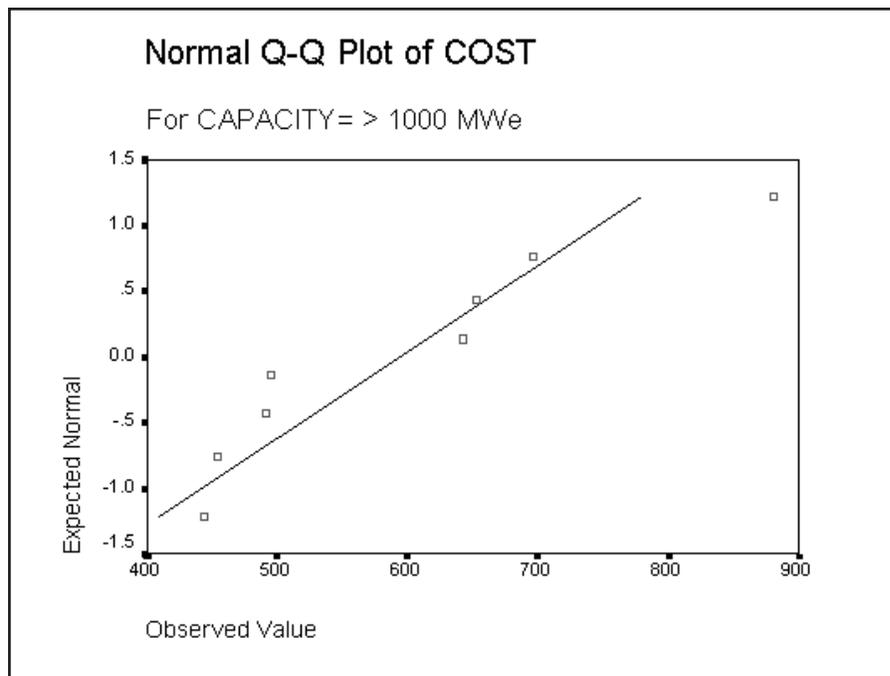
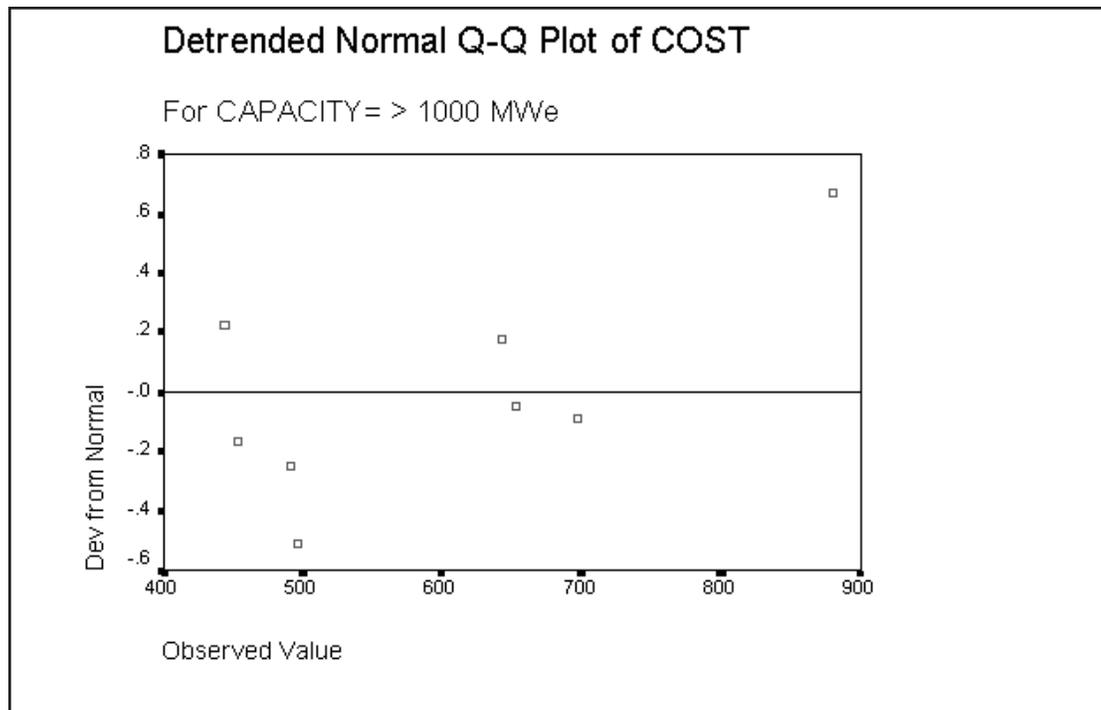


Figure 2.18 Detrended Q-Q Plot for the Third Group



In addition to a visual inspection, two tests of normality of the data are provided. The test labeled Kolmogorov-Smirnov is a modification of it using the Lilliefors Significance Correction (in which means and variances must be estimated from the data) comparing the distribution of the data values within the group to the normal distribution. The Shapiro-Wilks test also compares the observed data to the normal distribution and has been found to have good power in many situations when compared to other tests of normality (see Conover, 1980). For the first group there seem to be no problems regarding normality, nor any strikingly odd data values. Notice also that for the second group the tests of normality reject the null hypothesis that the data comes from a normal distribution, while the third group the null hypothesis is not rejected.

Figure 2.19 Tests of Normality

		Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
CAPACITY		Statistic	df	Sig.	Statistic	df	Sig.
COST	< 800 MWe	.149	13	.200*	.943	13	.489
	800-1000 MWe	.258	11	.040	.849	11	.049
	> 1000 MWe	.242	8	.185	.887	8	.279

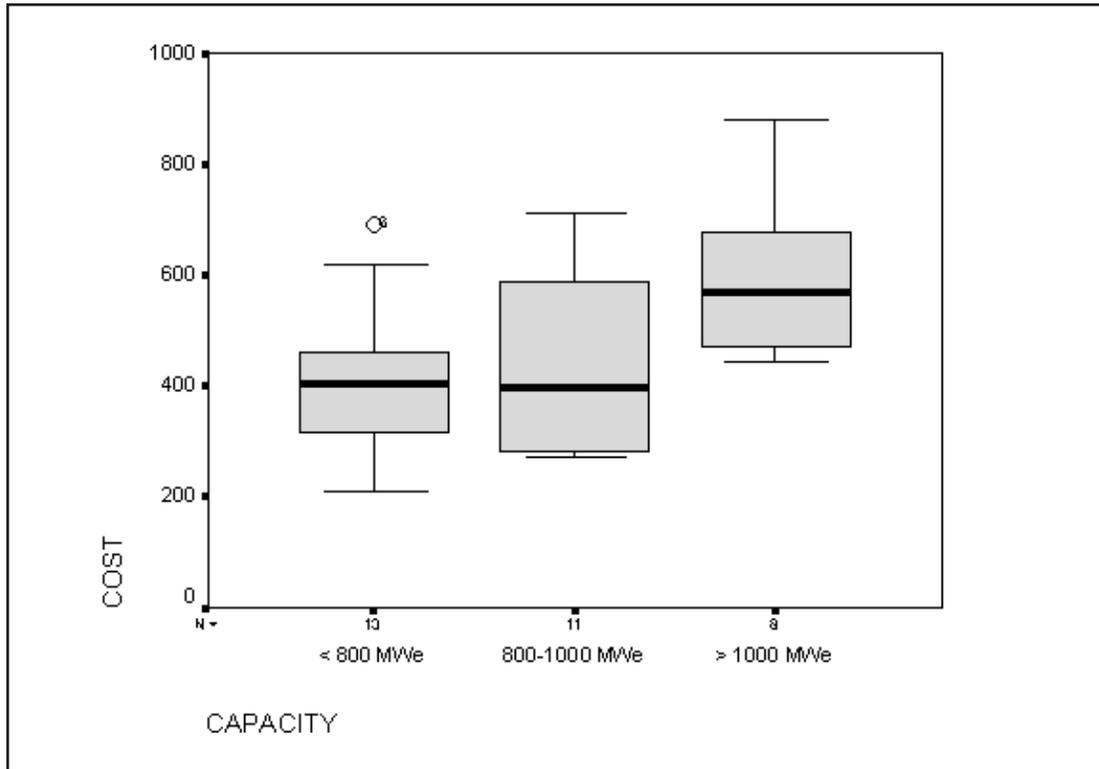
\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

**COMPARING THE GROUPS**

The box and whisker allows visual comparison of the groups.

Figure 2.20 Box and Whiskers Plot



The third group appears to contain higher cost plants than the first and second groups. The variation within each group as gauged by the whiskers seems fairly uniform. Notice the outlier in group one is identified by its case sequence number. There does not seem to be any increase in variation or spread as the median cost rises from the first to third group.

**HOMOGENEITY OF VARIANCE**

Homogeneity of variance within each population group is one of the assumptions in ANOVA. This can be tested by any of several statistics and if the variance is systematically related to the level of the group (mean, median) data transformations can be performed to relieve this (we will say more on this later in this chapter). The spread and level plot below provides a display of this by plotting the natural log of the spread (interquartile range) of the group against the natural log of the group median. If you can overcome a seemingly inborn aversion to logs and view the plot, we desire relatively little variation in the log spread going across the groups – which would suggest that the variances are stable across groups. The reason for taking logs is technical. If there is a systematic relation between the spread and the level (or variances and means), the slope of the best fitting line indicates what data transformation (within the class of power transformations) will best stabilize the variances across the different groups. We will say more about such transformations later.

Figure 2.21 Spread and Level Plot

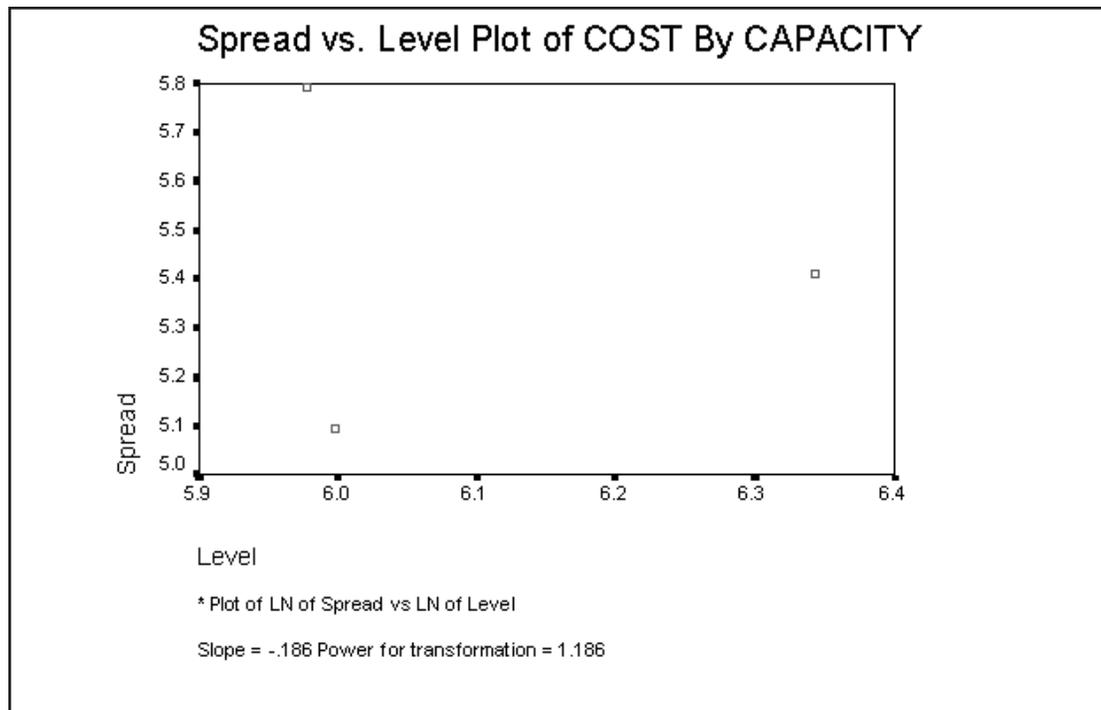


Figure 2.22 Test of Homogeneity of Variance

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
COST	Based on Mean	.995	2	29	.382
	Based on Median	.700	2	29	.505
	Based on Median and with adjusted df	.700	2	28.684	.505
	Based on trimmed mean	.945	2	29	.400

A number of tests are available for testing homogeneity of variance, such as the Bartlett-Box and Cochran's C tests of homogeneity of variance. However, these are sensitive to departures from normality as well. The Levene tests appearing above are less sensitive to departures from normality and might be preferred for that reason. Some statisticians consider the former tests too powerful in general; that is, they tend to reject the homogeneity of variance assumption when the differences are too small to influence the analysis. Above, the Levene test suggests no problem with the homogeneity assumption.

### Summary of the Plant Data

Overall the data fared fairly well in terms of the ANOVA assumptions. The only problem was normality of group 2. If inequality of variances was a problem and a data transformation applied, that might relieve the difficulty but no such transformation is called for. Since two of the three groups seem fine we will proceed with the analysis.

### EFFECTS OF VIOLATIONS OF ASSUMPTIONS IN ANOVA

Below we state in more detailed and formal terms the implications of violations of the assumptions and general conditions under which they constitute a serious problem.

### Normality of Errors in the Population

In the fixed effects model this assumption is equivalent to assuming that the dependent variable is normally distributed in the population, since all other terms in the model are to be considered fixed effects. “F” and “t” tests used to test for differences among means in the analysis of variance are unaffected by non-normality in large sample (this has led to the common practice of referring to the analysis of variance as robust with respect to violations of the normality assumption). Less is known about small sample behavior, but the current belief among most statisticians is that normality violations are generally not a cause for concern in fixed effect models.

While inferences about means are generally not heavily affected by non-normality, inferences about variances and about ratios of variances are quite dependent on the normality assumption. Thus random effects models are vulnerable to violations of normality where fixed effects models are not. More important in the general case, since most analyses of variance involve fixed effects models, is the fact that many standard tests of the homogeneity of error variance depend on inferences about variances, and are therefore vulnerable to violations of the normality assumption.

Tests of the homogeneity of variance assumption such as the Bartlett-Box F, Cochran's C and the F-max criterion all assume normality and are inaccurate in the presence of nonzero population kurtosis. If the population kurtosis is positive (signifying a peaked or leptokurtic distribution), these tests will tend to reject the homogeneity assumption too often, while a negative population kurtosis (indicative of a flat or platykurtic distribution) will lead to too many failures to recognize violations of the homogeneity assumption. For this reason the Levene test for homogeneity of variance (included in the Explore procedure) is strongly recommended, as it is robust to violations of the normality assumption.

### Homogeneity of Population Error Variances Among Groups

Violations of the homogeneity of variance assumption are in general more troublesome than violations of the normality assumption. In general, the smaller the sample sizes of the groups and the more dissimilar the sizes of the groups, the more problematic violations of this assumption become. Thus in a large sample with equal group sizes, even

moderate to severe departures from homogeneity may not have large effects on inferences, while in small samples with unequal group sizes, even slight to moderate departures can be troublesome. This is one reason that statisticians recommend large samples and equal group sizes whenever possible.

The magnitude of effects on actual Type I error level of violations of the homogeneity assumption depends on how dissimilar the variances are, how large is the sample, and how dissimilar are the group sizes, as mentioned above. The direction of the distortion of actual Type I error level depends on the relationship between variances and group sizes. Smaller sample from populations with larger variances lead to inflation of the actual Type I error level, while smaller samples from populations with smaller variances result in actual Type I error levels smaller than the nominal test alpha levels.

### **Population Errors Uncorrelated with Predictors and with Each Other**

Violations of the independence assumption can be serious even with large samples and equal group sizes. Methods such as generalized least squares should be used with autocorrelated data.

Two further points should be considered here. First, our discussion has centered on the impact of violations of assumptions on the actual Type I (alpha) error level. When considerations such as the power of a particular test are introduced, the situation can quickly become much more complicated. In addition, most of the work on the effects of assumption violations has considered each assumption in isolation. The effects of violations of two or more assumptions simultaneously are less well known. For more detailed discussions of these topics, see Scheffe (1959) or Kirk (1982). Also, see Wilcox (1996, 1997) for who discusses the effects of ANOVA assumption violation and presents robust alternatives.

### **A Note on Transformations**

Many researchers deal with violations of normality or homogeneity of variance assumptions by transforming their dependent variable in a nonlinear manner. Such transformations include natural logarithms, square roots, etc. These types of transformations are also employed to achieve additivity of effects in factorial designs with non-crossover interactions. There are, however, serious potential problems with such an approach.

While statistical procedures such as those employed by SPSS are not concerned with the sources of the numbers they are used to analyze, and will produce valid probabilities assuming only that distributional assumptions are met. The interpretation of analyses of transformed data can be quite problematic if the transformation employed is nonlinear.

If data are originally measured on an interval scale, which the calculation of means assumes, then nonlinearly transforming the dependent variable and running a standard analysis results in a very different set of questions being asked than with the dependent variable in

the original metric. Aside from the fact that a nonlinear transformation of an interval scale destroys the interval properties assumed in the calculation of means, the test of equality of a set of means of nonlinearly transformed data does not test the hypothesis that the means of the original data are equal, and there is no one to one relationship between the two tests. Attempts to back-transform parameter estimates by applying the inverse of the original transformation in order to apply the results to the original research hypothesis do not work. The bias introduced is a complicated one that actually increases with increasing sample size. For further information on this bias, see Kendall & Stuart (1968).

The practical implications of this point are that studies should be designed such that the variables which are of interest are measured, care should be taken to see that they meet the assumptions required to make the computation of basic descriptive statistics meaningful, and that commonly applied transformations in cases where ANOVA model assumptions are violated may cause more trouble than they avert. Accurate probabilities attached to significance tests of the equality of meaningless quantities are of even less use than distorted probabilities attached to tests concerning meaningful variables, especially when the direction and magnitude of distortions are of some degree estimable and can be taken into account when interpreting research results.

### **SUMMARY**

In this chapter we discussed the implications of violation of some of the assumptions of ANOVA: homogeneity of variance, and normality of error. We used exploratory data analysis techniques on the data set prior to formal analysis in order to view the data and check on the assumptions. In the next chapter we will proceed with the actual one-factor ANOVA analysis and consider planned and post-hoc comparisons.



## Chapter 3 One-Factor ANOVA

- Objective** Apply the principles of testing for population mean differences to situations involving more than two comparison groups. Understand the concept behind and the practical use of post-hoc tests applied to a set of sample means.
- Method** We will run a one-factor (Oneway procedure) analysis of variance comparing the different capacity groups on the cost of building a nuclear power plant. Then, we will rerun the analysis requesting multiple comparison (post hoc) tests to see specifically which population groups differ. We will then plot the results using an error bar chart. The appendix contains a nonparametric analysis of the same data.
- Data** We use the light water nuclear power plant data used in the last chapter.
- Scenario** We wish to investigate the relationship between the level of capacity of these plants and the cost associated with building the plants. One way to approach this is to group the plants according to their generating capacity and compare these groups on their average cost. In our data set we have the plants grouped into three capacity categories. Assuming we retain these categories we might first ask if there are any population differences in cost among these groups. If there are significant mean differences overall, we next want to know specifically which groups differ from which others.

### INTRODUCTION

**A**nalysis of variance (ANOVA) is a general method of drawing conclusions regarding differences in population means when two or more comparison groups are involved. The independent-groups t test applies only to the simplest instance (two groups), while ANOVA can accommodate more complex situations. It is worth mentioning that the t test can be viewed as a special case of ANOVA and they yield the same result in the two-group situation (same significance value, and the t statistic squared is equal to the ANOVA's F statistic).

We will compare three groups of plants based on their capacity and determine whether the populations they represent differ in the cost of being built.

### LOGIC OF TESTING FOR MEAN DIFFERENCES

The basic logic of significance testing is that we will assume that the population groups have the same mean (null hypothesis), then determine the probability of obtaining a sample with group mean differences as large (or larger) as what we find in our data. To make this assessment the amount of variation among the group means (between-group variation) is compared to the amount of variation among the observations within each group (within-group variation). Assuming that in the population the group means are equal (null hypothesis), the only source of variation among the sample means would be the fact that the groups are composed of different individual observations. Thus the ratio of the two sources of variation (between-group/within-group) should be about one when there are no population differences. When the distribution of the individual observations within each group follows the normal curve, the statistical distribution of this ratio is known (F distribution) and we can make a probability statement about the consistency of our data with the null hypothesis. The final result is the probability of obtaining sample differences as large (or larger) as what we found, if there were no population differences. If this probability is sufficiently small (usually less than .05, i.e., less than 5 chances in 100) we conclude the population groups differ.

### FACTORS

When performing a t test comparing two groups there is only one comparison that can be made: group one versus group two. For this reason the groups are constructed so their members systematically vary in only one aspect: for example, males versus females, or drug A versus drug B. If the two groups differed on more than one characteristic (for example, males given drug A versus females given drug B) it would be impossible to differentiate between the two effects (gender and drug).

Why couldn't a series of t tests be used to make comparisons among three groups? Couldn't we simply use t tests to compare group one versus group two, group one versus group three, and group two versus group three? One problem with this approach is that when multiple comparisons are made among a set of group means, the probability of at least one test showing significance even when the **null hypothesis is true** is higher than the significance level at which each test is performed (usually 0.05 or 0.01). In fact, if there is a large array of group means, the probability of at least one test showing significance is close to one (certainty)! It is sometimes asserted that an unplanned multiple comparison procedure can only be carried out if the ANOVA F test has shown significance. This is not necessarily true as it depends on what the research question(s) are.

There remains a problem, however. If the null hypothesis is that all the means are equal, the alternative hypothesis is that at least one of the means is different. If the ANOVA F test gives significance, we know there is a difference somewhere among the means, but that does not justify us in saying that any particular comparison is significant. The ANOVA F test, in fact, is an **omnibus test**, and further analysis is necessary to localize whatever differences there may be among the individual group means.

The question of exactly how one should proceed with further analysis after making the omnibus F test in ANOVA is not a simple one. It is important to distinguish between those comparisons that were **planned** before the **data** were actually gathered, and those that are made as part of the inevitable process of unplanned data-snooping that takes place after the results have been obtained. Planned comparisons are often known as *a-priori* comparisons. Unplanned comparisons should be termed *a-posteriori* comparisons, but unfortunately the misnomer **post hoc** is more often used.

When the data can be partitioned into more than two groups, additional comparisons can be made. This might involve one aspect or dimension, for example four groups each representing a region of the country. Or the groups might vary along several dimensions, for example eight groups each composed of a gender (two categories) by region (four categories) combination. In this latter case, we can ask additional questions: (1) is there a gender difference? (2) is there a region difference? (3) do gender and region interact? Each aspect or dimension the groups differ on is called a factor. Thus one might discuss a study or experiment involving one, two, even three or more factors. A factor is represented in the data set as a categorical variable and would be considered an independent variable. SPSS allows analysis of multiple factors, and has different procedures available based on how many factors are involved and their degree of complexity. If only one factor is to be studied use the Oneway (or One Factor ANOVA) procedure. When two or more factors are involved simply shift to the general factorial procedure (General Linear Model..General Factorial). In this chapter we consider a one-factor study (capacity relating to the cost of the plants), but we will discuss multiple factor ANOVA in later chapters.

### RUNNING ONE-FACTOR ANOVA

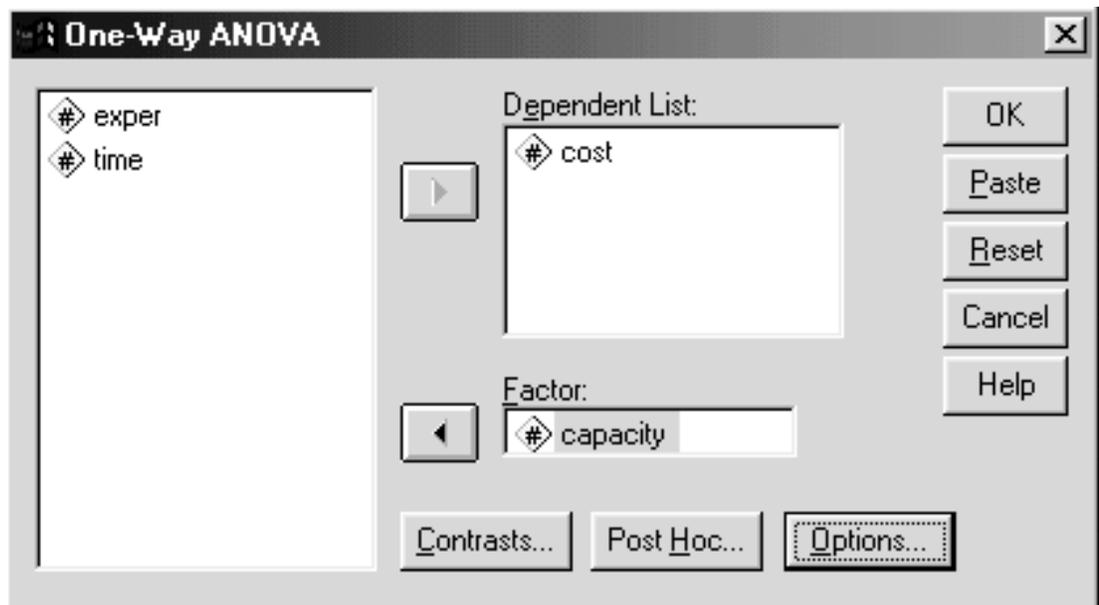
First we need to open our data set.

Click **File..Open..Data** (move to the **c:\Train\Anova** directory)  
Select **SPSS Portable (.por)** from the Files of Type drop-down list  
Double-click on **plant.por** to open the file.

To run the analysis using SPSS for Windows:

Click **Analyze..Compare Means ..One-Way ANOVA**.  
Move **cost** into the **Dependent List** box  
Move **capacity** into the **Factor** list box.

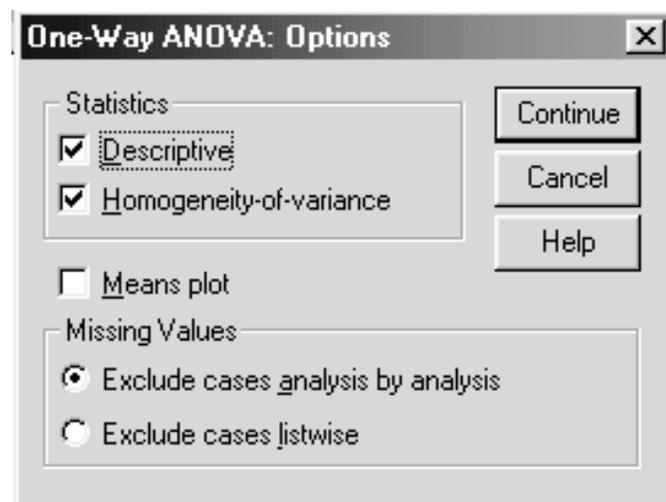
Figure 3.1 One-Way ANOVA Dialog Box



Enough information has been provided to run the basic analysis. The Contrasts pushbutton allows users to request statistical tests for planned group comparisons of interest to them. The Post Hoc pushbutton will produce multiple comparison tests that can test each group mean against every other one. Such tests facilitate determination of just which groups differ from which others and are usually performed after the overall analysis establishes that some significant differences exist. Finally, the Options pushbutton controls such features as missing value inclusion and whether descriptive statistics and homogeneity tests are desired.

- Click on the **Options** pushbutton
- Click to select both the **Descriptive** and **Homogeneity-of-variance**
- Click the **Exclude cases analysis by analysis** option button

Figure 3.2 One-way ANOVA Options Dialog Box



Click **Continue**  
Click **OK**

The missing value choices deal with how missing data are to be handled when several dependent variables are given. By default cases with missing values on a particular dependent variable are dropped only for the specific analysis involving that variable. Since we are looking at a single dependent variable, the choice has no relevance to our analysis.

The following syntax will run the analysis:

```
ONEWAY
  cost BY capacity
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /MISSING ANALYSIS .
```

The ONEWAY procedure performs a one-factor analysis of variance. Cost is the dependent measure and the keyword BY separates the dependent variable from the factor variable. We request descriptive statistics and a homogeneity of variance test. We also told SPSS to exclude cases with missing data on an analysis by analysis basis.

**ONE-FACTOR ANOVA RESULTS**

**Descriptive Statistics**

Information about the groups appears in the figure below. We see that costs increase with the increase in capacity. 95-percent confidence intervals for the capacity groups are presented in the table. One should note that the standard deviations for the three groups appear to be fairly close.

**Figure 3.3 Descriptive Statistics**

**Descriptives**

COST

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
< 800 MWe	13	400.8615	140.7775	39.045	315.79	485.93	207.51	690.19
800-1000 MWe	11	436.6482	173.6500	52.357	319.99	553.31	270.71	712.27
> 1000 MWe	8	594.4500	152.0381	53.754	467.34	721.56	443.22	881.24
Total	32	461.5603	170.1207	30.073	400.23	522.90	207.51	881.24

**Homogeneity of Variance**

We also requested the Levene test of homogeneity of variance.

**Figure 3.4 Levene Test of Homogeneity of Variance**

**Test of Homogeneity of Variances**

COST

Levene Statistic	df1	df2	Sig.
.995	2	29	.382

This assumption of equality of variance for all groups was tested in Chapter 2 using the **EXPLORE** (Examine) procedure. The Levene test also shows that with this particular data set the assumption of homogeneity of variance is met, indicating that the variances do not differ across groups.

What do we do if the assumption of equal variances is not met? If the sample sizes are close to the same size and sufficiently large we could count on the robustness of the assumption to allow the process to continue. However, there is no general adjustment for the F test in the case of unequal variances, as there was for the t test. A statistically sophisticated analyst might attempt to apply transformations to the dependent variable in order to stabilize the within-group variances (variance stabilizing transforms). These are beyond the scope of this course. Interested readers might turn to Emerson's chapter in Hoaglin, Mosteller, and Tukey (1991) for a discussion from the perspective of exploratory data analysis, and note that the spread & level plot in **EXPLORE** will suggest a variance stabilizing transform. A second and conservative approach would be to perform the analysis using a statistical method that does not assume homogeneity of variance. A one-factor analysis of group differences assuming that the dependent variable is only an ordinal (rank) variable is available as a nonparametric procedure within SPSS. This analysis is provided in the appendix to this chapter. However, one should note that corresponding nonparametric tests are not available for all analysis of variance models.

**The ANOVA Table**

**Figure 3.5 ANOVA Summary Table**

**ANOVA**

COST

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	196000.544	2	98000.272	4.053	.028
Within Groups	701171.764	29	24178.337		
Total	897172.309	31			

The output includes the analysis of variance summary table and the probability value we will use to judge statistical significance.

Most of the information in the ANOVA table is technical in nature and is not directly interpreted. Rather the summaries are used to obtain the F statistic and, more importantly, the probability value we use in evaluating the population differences. Notice that in the first column there is a row for the between-group and a row for the within-group variation. The df column contains information about the degrees of freedom, related to the number of groups and the number of individual observations within each group. The degrees of freedom are not interpreted directly, but are used in estimating the between-group and within-group variation (variances). Similarly, the sums of squares are intermediate summary numbers used in calculating the between and within-group variances. Technically they represent the sum of the squared deviations of the individual group means around the total grand mean (between) and the sum of the squared deviations of the individual observations around their respective sample group mean (within). These numbers are never interpreted and are reported because it is traditional to do so. The mean squares are measures of between and within group variances. Recall in our discussion of the logic of testing that under the null hypothesis both variances should have the same source and the ratio of between to within would be about one. This ratio, the sample F statistic, is 4.05 and we need to decide if it is far enough from one to say that the group means are not equal. The significance (Sig.) column indicates that under the null hypothesis of no group differences, the probability of getting mean costs this far (or more) apart by chance is under three percent (.028). If we were testing at the .05 level, we would conclude the capacity groups differ in average cost. In the language of statistical testing, the null hypothesis that power plants of these different capacities do not differ in cost is rejected at the 5% level.

**Conclusion** From this analysis we conclude that the capacity groups differ in terms of cost. In addition, we would like to know which groups differ from which others (Are they all different? Does the high capacity group differ from each of the other two?). This secondary examination of pairwise differences is done via procedures called multiple comparison testing (also called post hoc testing and multiple range testing). We turn to this issue next.

**POST-HOC TESTING** The purpose of post hoc testing is to determine exactly which groups differ from which others in terms of mean differences. This is usually done after the original ANOVA F test indicates that all groups are not identical. Special methods are employed because of concern with excessive Type I error.

In statistical testing, a Type I error is made if one falsely concludes that differences exist when in fact the null hypothesis of no differences is correct (sometimes called a false positive). When we test at a given level of significance say 5% (.05), we implicitly accept a five percent chance of a Type I error occurring. The more tests we perform, the greater the overall chances of one or more Type I errors cropping up.

This is of particular concern in our examination of which groups differ from which others since the more groups we have the more tests we make. If we consider pairwise tests (all pairings of groups, the number of tests for  $K$  groups is  $\{(K)(K-1)/2\}$ . Thus for three groups, three tests are made, but for 10 groups, 45 tests would apply. The purpose of the post hoc methodology is to allow such testing since we have interest in knowing which groups differ, yet apply some degree of control over the Type I error.

There are different schemes of controlling for Type I error in post hoc testing. SPSS makes many of them available. We will briefly discuss the different post hoc tests, and then apply some of them to the nuclear plant data. We will apply several post hoc methods for comparison purposes, in practice, usually only one would be run.

### WHY SO MANY TESTS?

The ideal post hoc test would demonstrate tight control of Type I error, have good statistical power (probability of detecting true population differences), and be robust over assumption violations (failure of homogeneity of variance, nonnormal error distributions). Unfortunately, there are implicit tradeoffs involving some of these desired features (Type I error and power) and no one current post hoc procedure is best in all areas. Couple to this the facts that there are different statistical distributions on which pairwise tests can be based (t, F, studentized range, and others) and that there are different levels at which Type I error can be controlled (per individual test, per family of tests, variations in between), and you have a huge collection of post hoc tests.

We will briefly compare post hoc tests from the perspective of being liberal or conservative regarding the control of the false positive rate and apply several to our data. There is a full literature (including several books) devoted to the study of post hoc (also called multiple comparison or multiple range tests, although there is a technical distinction between the two) tests. More recent books (Toothaker, 1991) summarize simulation studies that compare post hoc tests on their power (probability of detecting true population differences) as well as performance under different scenarios of patterns of group means, and assumption violations (homogeneity of variance).

The existence of numerous post hoc tests suggests that there is no single approach that statisticians agree will be optimal in all situations. In some research areas, publication reviewers require a particular post hoc method, which simplifies the researcher's decision.

Below we present some tests roughly ordered from the most liberal (greater statistical power and greater false positive rate) to the most conservative (smaller false positive rate, less statistical power), and mention some designed to adjust for the lack of homogeneity of variance.

<b>LSD</b>	The LSD or least significant difference method simply applies the standard t tests to all possible pairs of group means. No adjustment is made based on the number of tests performed. The argument is that since an overall difference in group means has already been established at the selected criterion level (say .05), no additional control is necessary. This is the most liberal of the post hoc tests.
<b>SNK, REGWF, REGWQ, and Duncan</b>	The SNK (Student-Newman-Keuls), REGWF (Ryan-Einot-Gabriel-Walsh F), REGWQ (Ryan-Einot-Gabriel-Walsh Q [based on studentized range statistic]), and Duncan methods involve sequential testing. After ordering the group means from lowest to highest, the two most extreme means are tested for a significant difference using a critical value adjusted for the fact that these are extremes from a larger set of means. If these means are found not to be significantly different, the testing stops; if they are different then the testing continues with the next most extreme pairs, and so on. All are more conservative than the LSD. REGWF and REGWQ improve on the traditionally used SNK in that they adjust for the slightly elevated false positive rate (Type I error) that SNK has when the set of means tested is much smaller than the full set.
<b>Bonferroni &amp; Sidak</b>	The Bonferroni (also called the Dunn procedure) and Sidak (also called Dunn-Sidak) perform each test at a stringent significance level to ensure that the overall (experiment wide) false positive rate does not exceed the specified value. They are based on inequalities relating the probability of one or more false positives for a set of independent tests. For example, the Bonferroni is based on an additive inequality, so the criterion level for each pairwise test is obtained by dividing the original criterion level (say .05) by the number of pairwise comparisons made. Thus with three means and therefore 3 pairwise comparisons, each Bonferroni test will be performed at the $.05/3$ or $.016667$ level.
<b>Tukey(b)</b>	The Tukey(b) test is a compromise test, combining the Tukey (see below) and the SNK criterion producing a test that falls between the two.
<b>Tukey</b>	Tukey (also called Tukey HSD, WSD, or Tukey(a) test): Tukey's HSD (Honestly Significant Difference) controls the false positive rate experiment wide. This means if you are testing at the .05 level, that when performing all pairwise comparisons, the probability of obtaining one or more false positives is .05. It is more conservative than the Duncan and SNK. If all pairwise comparisons are of interest, which is usually the case, Tukey's test is more powerful than the Bonferroni and Sidak.
<b>Scheffe</b>	Scheffe's method also controls the overall (or experiment wide) error rate. It adjusts not only for the pairwise comparisons, but for any possible comparison the researcher might ask. As such it is the most conservative of the available methods (false positive rate is least), but has less statistical power.

### Specialized Post Hoc Unequal Ns:

#### Hochberg's GT2 & Gabriel

Most post hoc procedures mentioned earlier (excepting LSD, Bonferroni, and Sidak) were derived assuming equal sample sizes in addition to homogeneity of variance and normality of error. When subgroup sample sizes are unequal, SPSS substitutes a compromise value (the harmonic mean) for the sample sizes. Hochberg's GT2 and Gabriel's post hoc test explicitly allow for unequal sample sizes.

#### Waller-Duncan

The Waller-Duncan takes an interesting approach (Bayesian) that adjusts the criterion value based on the size of the overall F statistic in order to be sensitive to the types of group differences associated with the F (for example, large or small). Also, you can specify the ratio of Type I (false positive) to Type II (false negative) error in the test. This feature allows for adjustments if there are differential costs to the two types of error.

#### Unequal Variances and Unequal Ns:

#### Tamhane T2, Dunnett's T3, Games-Howell, Dunnett's C

Each of these post hoc tests adjusts for unequal variances and sample sizes in the groups. Simulation studies suggest that although Games-Howell can be too liberal when the group variances are equal and sample sizes are unequal, it is more powerful than the others.

An approach some analysts take is to run both a liberal (say LSD) and a conservative (Scheffe or Tukey HSD) post hoc test. Group differences that show up under both criteria are considered solid findings, while those found different only under the liberal criterion are viewed as tentative results.

To illustrate the differences among the post hoc tests we will request six different post hoc tests: (1) LSD, (2) Duncan, (3) SNK, (4) Tukey's HSD, (5) Bonferroni, and (6) Scheffe.

Click **Dialog Recall** button 

Select **One-Way ANOVA**

Within the One-Way ANOVA Dialog Box

Click on the **Post Hoc** pushbutton.

Select the following types of post hoc tests: **LSD, Duncan, SNK, Tukey, Bonferroni, and Scheffe.**

Figure 3.6 Post Hoc Dialog Box

**One-Way ANOVA: Post Hoc Multiple Comparisons**

Equal Variances Assumed

LSD       S-N-K       Waller-Duncan  
 Bonferroni       Tukey      Type I/Type II Error Ratio: 100  
 Sidak       Tukey's-b       Dunnett  
 Scheffe       Duncan      Control Category: Last  
 R-E-G-W F       Hochberg's GT2      - Test:  
 R-E-G-W Q       Gabriel       2-sided     < Control     > Control

Equal Variances Not Assumed

Tamhane's T2     Dunnett's T3     Games-Howell     Dunnett's C

Significance level: .05

Continue    Cancel    Help

By default, statistical tests will be done at the .05 level. For some tests you may supply your preferred criterion level. The command to run the post hoc analysis appears below.

```
ONEWAY
  cost BY capacity
  /MISSING ANALYSIS
  /POSTHOC = SNK TUKEY DUNCAN SCHEFFE LSD
  BONFERRONI ALPHA(.05).
```

Post hoc tests are requested using the **POSTHOC** subcommand. The **STATISTICS** subcommand need not be included here since we have already viewed the means and discussed the homogeneity test.

Click **Continue**  
 Click **OK**

The beginning part of the output contains the ANOVA table, descriptive statistics, and the homogeneity test, which we have already reviewed. We will move directly to the post hoc test results.

Figure 3.7 LSD Post Hoc Results

Post Hoc Tests						
Multiple Comparisons						
Dependent Variable: COST						
LSD						
(I) CAPACITY	(J) CAPACITY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
< 800 MWe	800-1000 MWe	-35.7866	63.702	.579	-166.0712	94.4979
	> 1000 MWe	-193.5885*	69.872	.010	-336.4937	-50.6832
800-1000 MWe	< 800 MWe	35.7866	63.702	.579	-94.4979	166.0712
	> 1000 MWe	-157.8018*	72.252	.037	-305.5733	-10.0303
> 1000 MWe	< 800 MWe	193.5885*	69.872	.010	50.6832	336.4937
	800-1000 MWe	157.8018*	72.252	.037	10.0303	305.5733

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

**Note** All tests appear in one table. However, the Post Hoc Tests and Homogeneous Subsets pivot tables were edited in the Pivot Table Editor so that each test can be viewed and discussed separately. (To do so, double-click on the pivot table to invoke the Pivot Table Editor, then click Pivot..Pivot Trays so that the Pivot Trays option is checked and the Pivot Trays window is visible. Next click and drag the pivot tray icon for Test (to see an icon's label, just click on the icon) from the Row dimension tray into the Layer dimension tray. Now test results for any single post hoc test can be viewed by selecting the desired test from the Test drop-down list located just above the table.)

The rows are constructed from every possible pairing of groups. For example, the less than 800 Mwe group is paired against the other two groups, then the 800-1000 Mwe group is paired against the other two groups, etc. The column label "Mean Difference (I-J)" contains the mean difference between each pairing of groups. We see that the <800 group has a mean cost difference of -\$35.7866 with the 800-1000 group and a difference of -\$193.5885 with the >1000 group. If a difference is statistically significant at the specified level after applying any post hoc adjustments (none for LSD), then an asterisk (\*) appears beside the mean difference. Notice the actual significance value for the test appears in the column labeled "Sig."

The first LSD block indicates that in the population those plants having less than 800 Mwe's differ significantly in cost from the plants having a capacity of greater than 1000 Mwe's. In addition, the standard errors and 95% confidence intervals for each mean difference are displayed. These provide information of the precision with which we have estimated the mean differences. Note that, as you expect, if a mean difference is not significant, the confidence interval contains zero. Using LSD, the high capacity group differs from each of the other two, but the lower capacity groups do not differ from each other.

Figure 3.8 Duncan Results

**COST**

Duncan<sup>a,b</sup>

CAPACITY	N	Subset for alpha = .05	
		1	2
< 800 MWe	13	400.8615	
800-1000 MWe	11	436.6482	
> 1000 MWe	8		594.4500
Sig.		.606	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 10.245.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

SPSS does not present the Duncan results in the same format as we saw for the LSD. This is because for some of the post hoc test methods standard errors and 95-percent confidence intervals are not defined (for multiple-range tests, recall testing stops once the remaining most extreme means are not found different). Rather than display results with empty columns in such situations, a different format, homogeneous subsets, is used. A homogeneous subset is a set of groups for which no pair of group means differs significantly. Depending on the post hoc test requested SPSS will display a multiple comparison table, a homogeneous subset table, or both. In this data set, it shows that the two lower capacity groups do not differ in cost, but differ from the highest capacity group.

Figure 3.9 SNK Results

**Homogeneous Subsets**

**COST**

Student-Newman-Keuls<sup>a,b</sup>

CAPACITY	N	Subset for alpha = .05	
		1	2
< 800 MWe	13	400.8615	
800-1000 MWe	11	436.6482	
> 1000 MWe	8		594.4500
Sig.		.606	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 10.245.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

The SNK results display the same pattern as the Duncan tests.

Figure 3.10 Tukey Results for Multiple Comparisons

**Post Hoc Tests**

**Multiple Comparisons**

Dependent Variable: COST  
Tukey HSD

(I) CAPACITY	(J) CAPACITY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
< 800 MWe	800-1000 MWe	-35.7866	63.702	.841	-193.1079	121.5346
	> 1000 MWe	-193.5885*	69.872	.025	-366.1495	-21.0275
800-1000 MWe	< 800 MWe	35.7866	63.702	.841	-121.5346	193.1079
	> 1000 MWe	-157.8018	72.252	.091	-336.2389	20.6353
> 1000 MWe	< 800 MWe	193.5885*	69.872	.025	21.0275	366.1495
	800-1000 MWe	157.8018	72.252	.091	-20.6353	336.2389

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

The Tukey multiple comparison tests show that the less than 800 Mwe group is significantly different from the greater than 1000 Mwe group, but this is the only pairwise difference.

Figure 3.11 Tukey Results for Homogeneous Subsets

**Homogeneous Subsets**

**COST**

Tukey HSD <sup>a,b</sup>

CAPACITY	N	Subset for alpha = .05	
		1	2
< 800 MWe	13	400.8615	
800-1000 MWe	11	436.6482	436.6482
> 1000 MWe	8		594.4500
Sig.		.862	.072

Means for groups in homogeneous subsets are displayed.

- Uses Harmonic Mean Sample Size = 10.245.
- The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

The Tukey homogeneous subset table is consistent with the multiple comparison table. The first homogeneous subset contains the two lower

capacity plants (they do not differ). The second homogeneous subset is made up of the second and third groups (they do not differ). Thus the only difference is between the less than 800 Mwe group and the greater than 1000 Mwe group. It should be pointed out that the second and third groups are barely not significant (.07) and had the sample sizes been larger their difference might have been significant.

Figure 3.12 Bonferroni Results

**Post Hoc Tests**

**Multiple Comparisons**

Dependent Variable: COST  
Bonferroni

(I) CAPACITY	(J) CAPACITY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
< 800 MWe	800-1000 MWe	-35.7866	63.702	1.000	-197.6467	126.0734
	> 1000 MWe	-193.5885*	69.872	.029	-371.1280	-16.0489
800-1000 MWe	< 800 MWe	35.7866	63.702	1.000	-126.0734	197.6467
	> 1000 MWe	-157.8018	72.252	.112	-341.3870	25.7834
> 1000 MWe	< 800 MWe	193.5885*	69.872	.029	16.0489	371.1280
	800-1000 MWe	157.8018	72.252	.112	-25.7834	341.3870

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

The test shows that the less than 800 Mwe's group has a significantly different cost than the greater than 1000 Mwe's group.

Figure 3.13 Scheffe Results for Multiple Comparisons

**Post Hoc Tests**

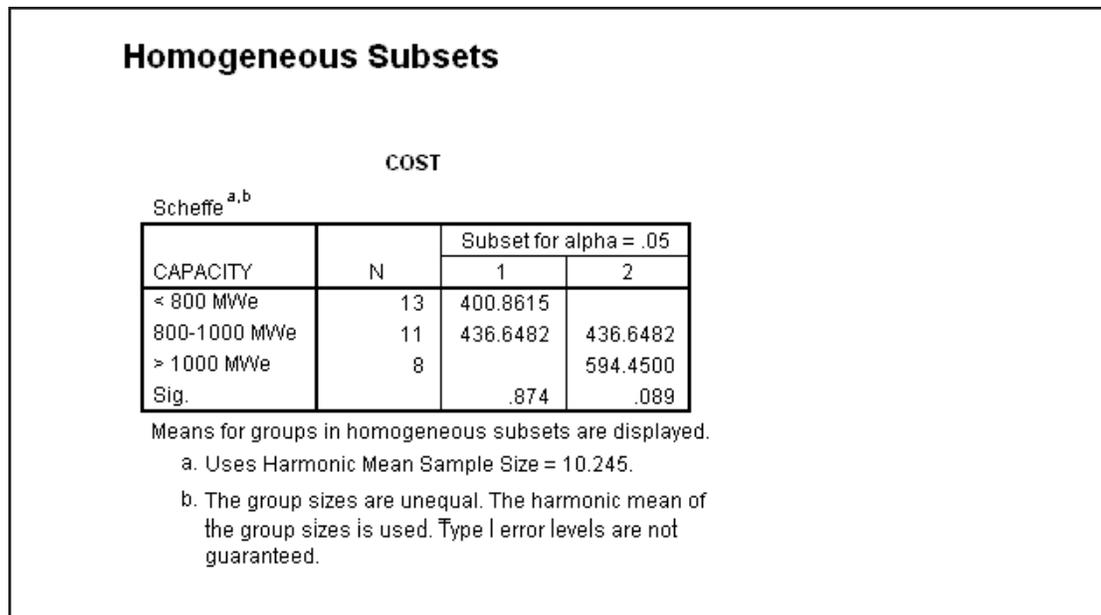
**Multiple Comparisons**

Dependent Variable: COST  
Scheffe

(I) CAPACITY	(J) CAPACITY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
< 800 MWe	800-1000 MWe	-35.7866	63.702	.855	-200.1235	128.5502
	> 1000 MWe	-193.5885*	69.872	.033	-373.8447	-13.3322
800-1000 MWe	< 800 MWe	35.7866	63.702	.855	-128.5502	200.1235
	> 1000 MWe	-157.8018	72.252	.110	-344.1962	28.5926
> 1000 MWe	< 800 MWe	193.5885*	69.872	.033	13.3322	373.8447
	800-1000 MWe	157.8018	72.252	.110	-28.5926	344.1962

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

Figure 3.14 Scheffe Results for Homogeneous Subsets



From these results we can see that similar to the Bonferroni test, only the high and low capacity groups differ.

**Conclusion**

As discussed before, the different post hoc procedures offer different trade-offs between Type I error (falsely claiming a significant difference) and power (ability to detect a real difference). Your choice in the matter depends on how you want to balance the two. In this analysis it appears that the high and low capacity groups do differ in cost, while the low and middle groups do not. The middle to high capacity difference might be usefully considered as a tentative finding.

**PLANNED COMPARISONS**

Post hoc tests compare all pairs of groups and most of the methods discussed apply a penalty function (adjusting the critical value) because so many tests are being made. In some experiments and studies, the researcher has in mind some specific comparisons to be made between group means. Compared to post hoc tests, planned comparisons are fewer in number and are to be formulated before viewing the data. Because they are limited in number (based on between-group degrees of freedom (the number of groups minus one)) and specified beforehand, the adjustments made for post hoc tests are not required.

A broad variety of planned comparisons (sometimes called *a priori* comparisons) can be requested: all treatment groups might be compared to a control group; a linear trend line could be fit; step comparisons could be made to detect a threshold.

To demonstrate this method, let us suppose that there is interest in making some specific comparisons between capacity groups. The idea is that at some point the change in capacity would result in a large change in cost. To see if and where this occurs, we can compare the low to middle capacity plants, then the middle to high capacity plants. If either of these

comparisons is significant, we have an idea of where the big cost increase will occur.

### HOW PLANNED COMPARISONS ARE DONE

Planned comparisons between groups are done by applying a set of coefficients to the group means and testing whether the result is zero. For example, to compare the low and middle plant groups, multiply the mean of the low plants by one, the mean of the middle plants by negative one, the mean of the large plants by zero, and sum the result. Thus, we compare the means, and if this difference is significantly different from zero, then the low capacity plants differ from the middle capacity plants. In **ONEWAY** you can request planned comparisons by providing sets of coefficients.

To request tests of low versus middle, and the middle versus high capacity groups we use the Contrasts pushbutton and apply the necessary coefficients.

Click **Dialog Recall** tool 

Select **One-Way ANOVA**

Click the **Contrasts** pushbutton

Type **1** in the **Coefficients** text box and click **Add** pushbutton

Type **-1** in the **Coefficients** text box and click **Add** pushbutton

Type **0** in the **Coefficients** text box and click **Add** pushbutton

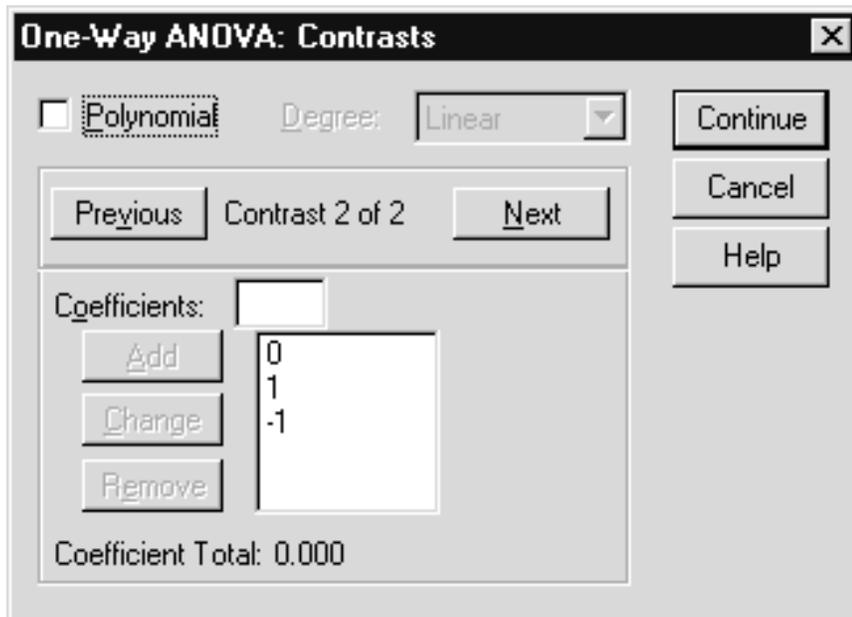
Click **Next** pushbutton

Type **0** in the **Coefficients** text box and click **Add** pushbutton

Type **1** in the **Coefficients** text box and click **Add** pushbutton

Type **-1** in the **Coefficients** text box and click **Add** pushbutton

Figure 3.15 Contrasts Dialog Box



Each set of contrast coefficients is assigned a number (1,2, ...) and appears as a column in the Coefficients list box.

Click **Continue** to process the Contrasts  
 Click **OK** to run the analysis

This leads to the syntax below (note the PostHoc subcommand is not included although our previous post hoc requests are still stored in the Post Hoc dialog.

```
ONEWAY
cost BY capacity
/CONTRAST= 1 -1 0 /CONTRAST = 0 1 -1
/MISSING ANALYSIS.
```

The first contrast requests the difference between the low and middle groups; the second compares the middle and high groups. We are limited to two planned comparisons because with three groups we have but two between-group degrees of freedom.

Scroll to the **Contrast Coefficients Pivot Table** in the Viewer window.

Figure 3.16 Contrast Coefficients

**Contrast Coefficients**

Contrast	CAPACITY		
	< 800 MWe	800-1000 MWe	> 1000 MWe
1	1	-1	0
2	0	1	-1

The requested comparisons are first reproduced along with the group labels. We verify that the first compares the low to middle group, and the second compares the middle to high group.

Figure 3.17 Contrast Results

**Contrast Tests**

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
COST	Assume equal variances	1	-35.7866	63.7017	-.562	29	.579
		2	-157.8018	72.2518	-2.184	29	.037
	Does not assume equal variances	1	-35.7866	65.3130	-.548	19.253	.590
		2	-157.8018	75.0383	-2.103	16.308	.051

Notice that there are two sets of results, one labeled “assume equal variances” and the other “does not assume equal variances”. Results labeled “does not assume equal variances” are adjusted results that can be used if the homogeneity of variance assumption is not met. We

previously determined the variances are equal and will use the “assume equal variances” statistics.

The column labeled "Value of Contrast" contains the values of the contrast coefficients applied to the sample means, which here represent the mean difference between pairs of groups. This can be verified by checking the group means appearing earlier. The first comparison (between the low and middle groups) is not significant, but the second one (comparing the middle and high capacity groups) is. This suggests that the big cost increase comes when shifting from the middle to high capacity plants. A t test is used since each comparison involves one degree of freedom; it is equivalent to using an F test (the t statistic squared would equal the F).

Thus a limited number of planned comparisons between group means can be specified as part of the general analysis. Performing planned comparisons does not preclude running post hoc analyses later.

## GRAPHING THE RESULTS

For presentations it is useful to display the sample group means along with their 95-percent confidence intervals. In SPSS for Windows

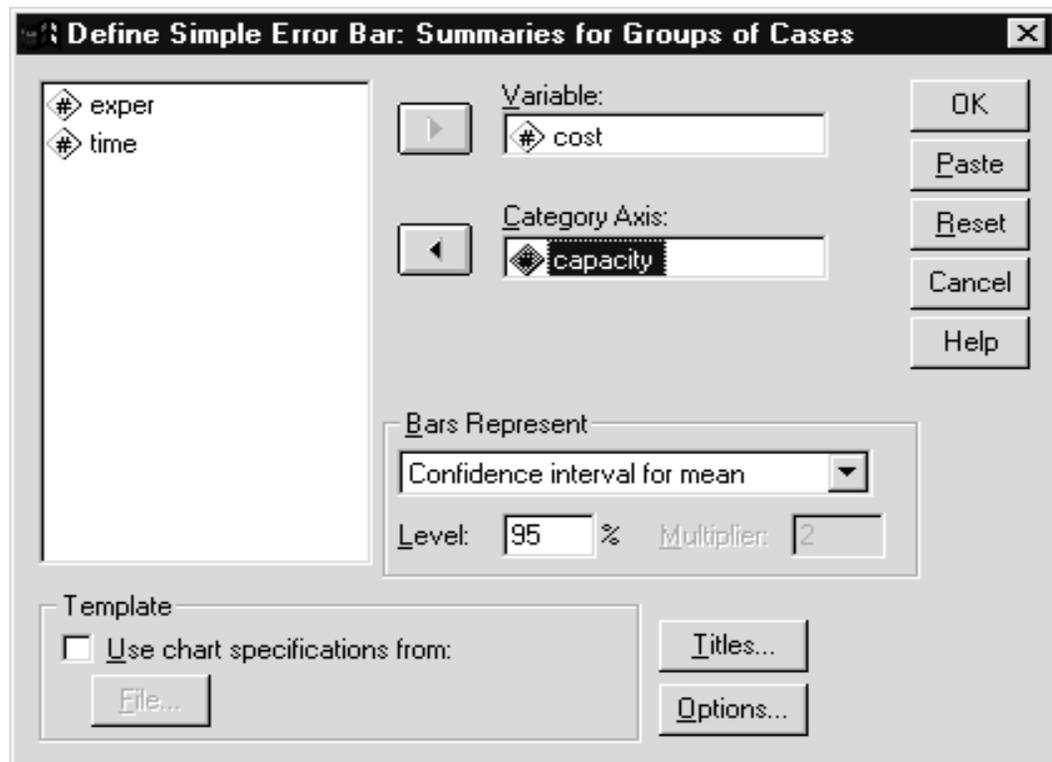
Click on **Graphs..Error Bar**

Verify that **Simple** is selected, then click **Define** pushbutton

Move **cost** into the **Variable** list box

Move **capacity** into the **Category Axis** list box.

Figure 3.18 Error Bar Dialog Box

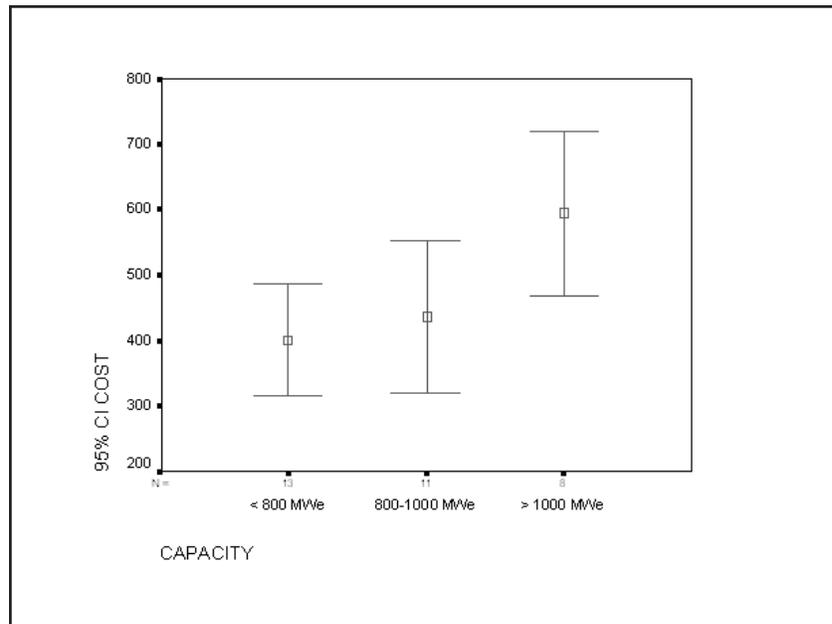


Click on **OK**

The command below will produce the error bar chart using a standard graph (there is also Interactive graph that produces an error bar chart).

```
GRAPH  
/ERRORBAR( CI 95 )=cost BY capacity  
/MISSING=REPORT.
```

**Figure 3.19 Error Bar Chart of Cost by Capacity Group**



The chart provides a visual sense of how far the groups are separated. The confidence bands are determined for each group separately (thus inspection of the confidence band overlap is not formally equivalent to testing for group differences) and no adjustment is made based on the number of groups that are compared. However, from the graph we have a clearer sense of the relation between capacity and cost.

## SUMMARY

In this chapter we tested for population mean differences with more than two groups when these groups constitute a single factor. We examined the data to check for assumption violations, discussed alternatives, and interpreted the ANOVA results. Having found significant differences we performed post hoc tests to determine which specific groups differed from which others, and summarized the analysis with an error bar graph. The appendix contains a nonparametric analysis of the same data.

## APPENDIX: GROUP DIFFERENCES ON RANKS

Analysis of variance assumes that the dependent measure is interval scale, that its distribution within each group follows a normal curve, and that the within-group variation is homogeneous across groups. If any of these assumptions fail in a gross way, one may be able to apply techniques that make fewer assumptions about the data. Such tests fall under the class of nonparametric statistics (they do not assume specific data distributions described by the parameters such as the mean and

standard deviation). Since these methods make few if any distributional assumptions, they can often be applied when the usual assumptions are not met. If you are tempted to think that something is obtained for nothing, the downside of such methods is that if the stronger data assumptions hold, the nonparametric tests are generally less powerful (probability of finding true differences) than the appropriate parametric method. Also, there are some parametric statistical analyses that currently have no corresponding nonparametric method. It is fair to say that the boundaries concerning when to use parametric versus nonparametric methods are in practice somewhat vague, and statisticians can and often do disagree about which approach is optimal in a specific situation.

For the purposes of this appendix let us assume that we needed to run the test using nonparametric methods. We will perform a nonparametric procedure that only assumes that the dependent measure has ordinal properties. The basic logic behind this test, the Kruskal-Wallis test, is as follows. If we rank order the dependent measure throughout the entire sample, we would expect under the null hypothesis (of no population differences) that the average rank (technically the sum of the ranks adjusted for sample size) should be about the same for each group. The Kruskal-Wallis test calculates the ranks, each sample group's mean rank, and the probability of obtaining group average ranks (weighted summed ranks) as far apart (or further) as what is observed in the sample, if the population groups were identical.

To run the Kruskal-Wallis test in SPSS we would

Click **Analyze..Nonparametric Tests..K Independent Samples**

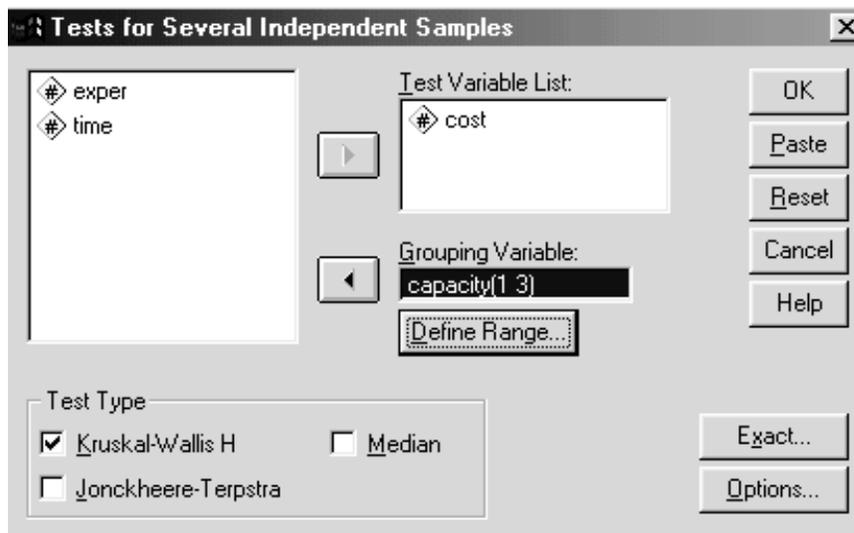
Move **cost** into the Test Variable List

Move **capacity** into the Grouping Variable box

Click the **Define Range** button and enter a **Minimum** of 1 and **Maximum** of 3.

Click **Continue**

Figure 3.20 Analysis of Ranks Dialog Box



By default, the Kruskal-Wallis test will be performed. The organization of this dialog box closely resembles that of the **One-Way ANOVA**. The command to run this analysis using SPSS follows.

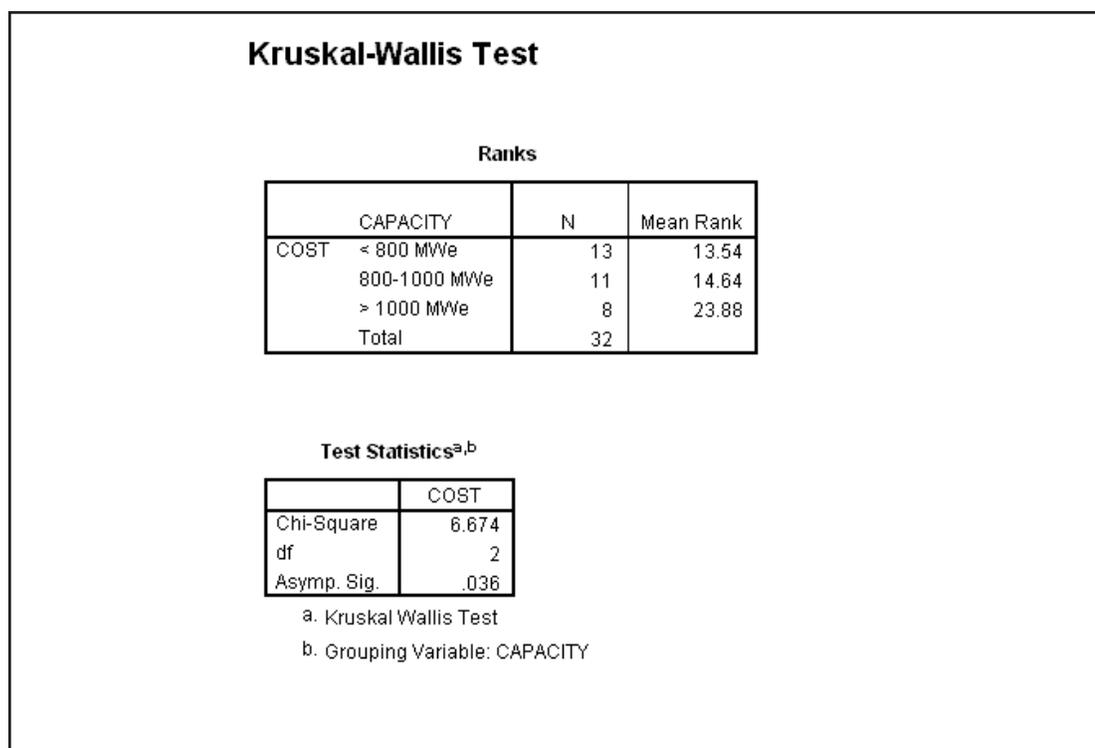
```

NPAR TESTS
  /K-W=cost BY capacity(1 3)
    
```

The K-W subcommand instructs the nonparametric testing routine to perform the Kruskal-Wallis analysis of variance of ranks on the dependent variable cost with capacity as the independent or grouping variable.

Click **OK** to run the analysis

Figure 3.21 Results of Kruskal-Wallis Nonparametric Analysis



We see the pattern of mean ranks (remember smaller ranks imply lower cost) follows that of the original means of cost, increasing as the capacity increases. The chi-square statistic is used in the Kruskal-Wallis indicates that it is very unlikely (fewer than 4 chances in 100) to obtain samples with average ranks so far apart if the null hypothesis (no cost differences between groups) were true. This is consistent with our conclusion from the initial one-way ANOVA analysis.

# Chapter 4 Multi-Way Univariate ANOVA

**Objective** We will apply the principles of testing for differences in population means to situations involving more than one factor. Also we will show how the two-factor ANOVA is a generalization of the one-factor design that we covered in the last chapter. We will develop some understanding of the new features of the analysis. We will then discuss the implications of unequal sample sizes and empty cells.

**Method** We wish to test whether there are any differences in the cost of the nuclear power plants based on capacity or the experience of the architect/engineer. First we use the EXPLORE procedure to explore the subgroups involved in the analysis. Next, we make use of the Univariate procedure to run the two-factor ANOVA, specifying cost as the dependent variable with capacity and experience as factors. We display the results using an error bar chart. In the appendix we perform post hoc tests based on the results of our analysis.

**Data** We continue to use the nuclear plant data. The data set is an SPSS portable file (plant.por) containing information about 32 light water nuclear power plants. Four variables are included: the capacity and cost of the plant; time to completion; and experience of the architect-engineer who built the plant.

## INTRODUCTION

**A**nalysis of variance (ANOVA) is a general method for drawing conclusions about differences in population means when two or more comparison groups are involved. In an introductory statistics class you have seen how a “t” test is used to contrast two groups, and in the last chapter we saw how one-way ANOVA compares more than two groups which differ along a single factor. In this chapter, we expand our consideration of ANOVA to allow multiple factors in a single analysis. Such an approach is efficient in that several questions are addressed within one study. The assumptions and issues considered in the last chapter (normality of the dependent variable within each group, homogeneity of variance, and the importance of both) apply to general ANOVA and will not be repeated here.

We will investigate whether there are differences in the average cost of a plant for the different plant capacities and levels of experience of the designer/engineer. Since two factors, capacity and experience, are under consideration, we can ask three different questions: (1) Are there cost differences based on capacity? (2) Are there differences based on experience? (3) Do capacity and experience interact?

A multi-factor analysis involves the same approach and principles, as did a one-way ANOVA. The between-groups variation can now be partitioned into pieces attributable to main effects and interaction components, but the method is much the same. Some complications arise with unequal cell sizes and empty cells that were not a problem when we tested a single factor. We will discuss these issues and illustrate the analysis.

As in earlier chapters, we begin by running an exploratory data analysis, then proceed with more formal testing.

### LOGIC OF TESTING, AND ASSUMPTIONS

As before, we wish to draw conclusions about the populations from which we sample. The main difference in moving from a one-way ANOVA to the general ANOVA is that more questions can be asked about the populations. However, the results will be stated in the same terms: how likely is it that we would obtain means as far apart as what we observe in our sample, if there were no mean differences in the populations. Comparisons are again framed as a ratio of the variation among the group means (between-group variation) to the variation among observations within each group (within-group variation). When statistical tests are performed, homogeneity of variance and normality of the dependent variable within each group are assumed. Comments made earlier regarding robustness of the means analysis when these assumptions are violated apply directly.

### HOW MANY FACTORS?

The new aspect we consider is how to include several factors, or ask several different questions of the data, within a single analysis of variance. We will test whether there are differences in cost based on capacity, whether there are differences based on the experience of the engineer, and finally, whether capacity and experience interact concerning the cost of the plants. The interpretation of an interaction is discussed in the next section.

Although our example involves only two factors (capacity and experience), ANOVA can accommodate more. Usually, the number of factors is limited by either the interests of the researcher, who might wish to examine a few specific issues, or by sample size considerations. Sample size plays a role in that the greater the number of factors, the greater the number of cell means that must be computed, and the smaller the sample for each mean. For example, suppose we have a sample of 800 plants and wish to look at cost differences due to whether the plant was light water or heavy water (2 levels), capacity (3 levels), experience (3 levels), region of the country (9 levels), and age of the plant (4 levels). There are  $2 \times 3 \times 3 \times 9 \times 4$  or 648 subgroup means involved. If the data were distributed evenly across the levels, each subgroup mean would be based on approximately two observations, and this would not produce a very powerful analysis. Such analyses can be performed, and technically, questions involving single effects like capacity or experience would be based on means involving fairly large samples. Also, some planned experiments permit many subgroups to be dropped (for example, incomplete designs). Yet the fact remains that with smaller samples, there are practical limitations in the number of questions you can ask of the data.

### INTERACTIONS

When moving beyond one-factor ANOVA, the distinction between main effects and interactions becomes relevant. A main effect is an effect (or group difference) attributable to a single factor (independent variable). For example, when we study cost differences across capacity groups and experience groups, the effect of capacity alone, and the effect of experience alone, would be considered main effects. The two-way interaction would test whether the effect of one factor is the same at each level of the other factor.

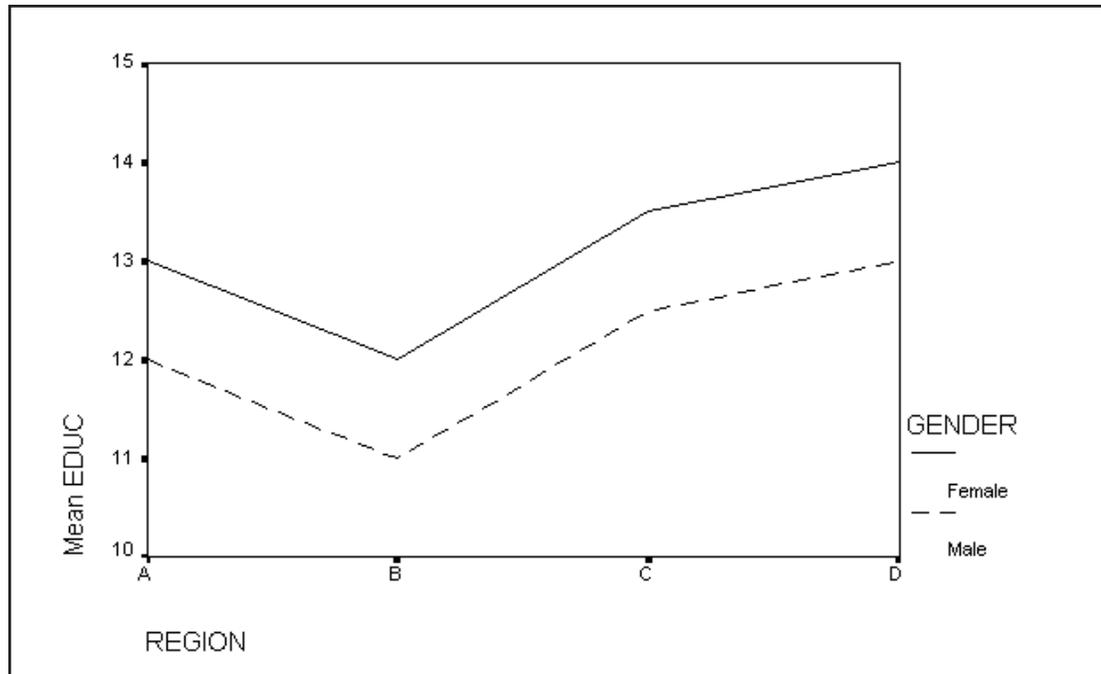
In our example, this can be phrased in either of two ways. We could say the interaction tests whether the cost difference due to capacity (which could be zero) is the same for each level of experience. Alternatively, we can say that the two-way interaction tests whether the experience difference in cost is the same for each capacity group. While these two phrasings are mathematically equivalent, it can sometimes be simpler (based on the number of levels in each factor) for you to present the information from one perspective instead of the other. The presence of a two-way interaction is important to report, since it qualifies our interpretation of a main effect. For example, a capacity by experience interaction implies that the magnitude of the capacity difference varies across levels of experience. In fact, there may be no difference or a reversal in the pattern of the capacity means for some experience levels. Thus statements about capacity differences must be qualified by experience information.

Since we are studying two factors, there can be only one interaction. If we expand our analysis to three factors (say capacity, experience, and age of plant) we can ask both two-way (capacity by experience, capacity by age, experience by age) and three-way (capacity by experience by age) interaction questions. As the number of factors increases, so does the possible complexity of the interactions. In practice, significant high-order (three, four, five-way, etc.) interactions are relatively rare compared to the number of significant main effects.

Interpretation of an interaction can be done directly from a table of relevant subgroup means, but it is more convenient and common to view a multiple-line chart of the means. We illustrate this below under several scenarios.

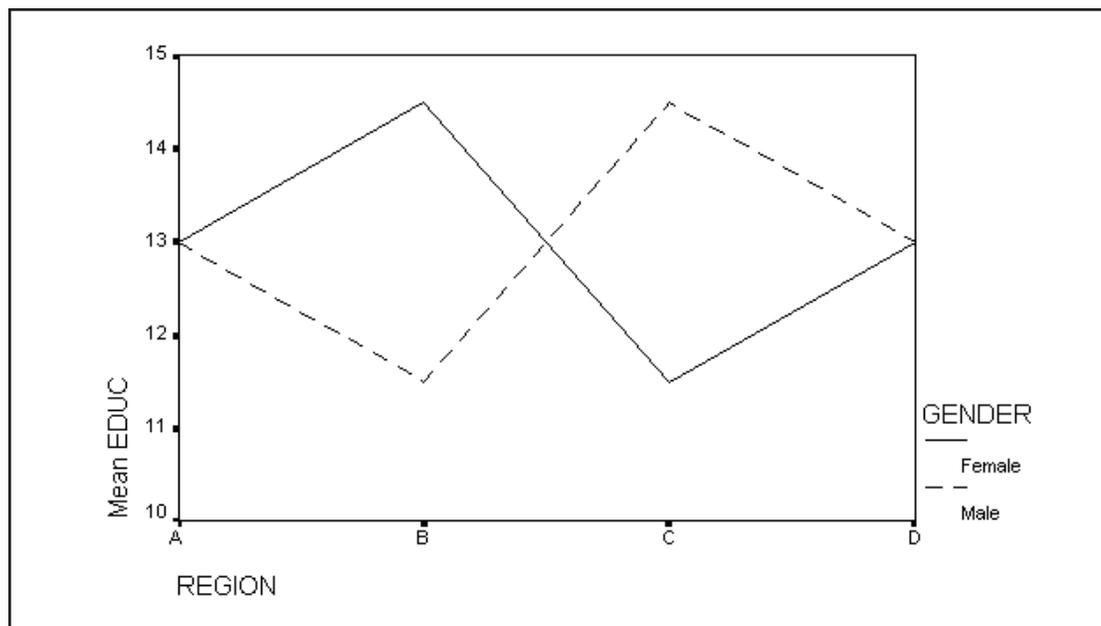
Suppose that we have four levels of one independent variable (say location) and two levels for the second independent variable (say gender). In our scenario, suppose that women are more highly educated than men, there are regional differences in education, and that the gender differences are the same across regions. The line chart below plots a set of means consistent with this pattern.

Illustration 4.1 Main Effects, No Interaction



In the illustration we see that the mean line for women is above that of the men. In addition, there are differences among the four locations. However, note that the gender differences are nearly identical across the four locations. This equal distance between the lines (parallelism of lines) indicates that there is no interaction present.

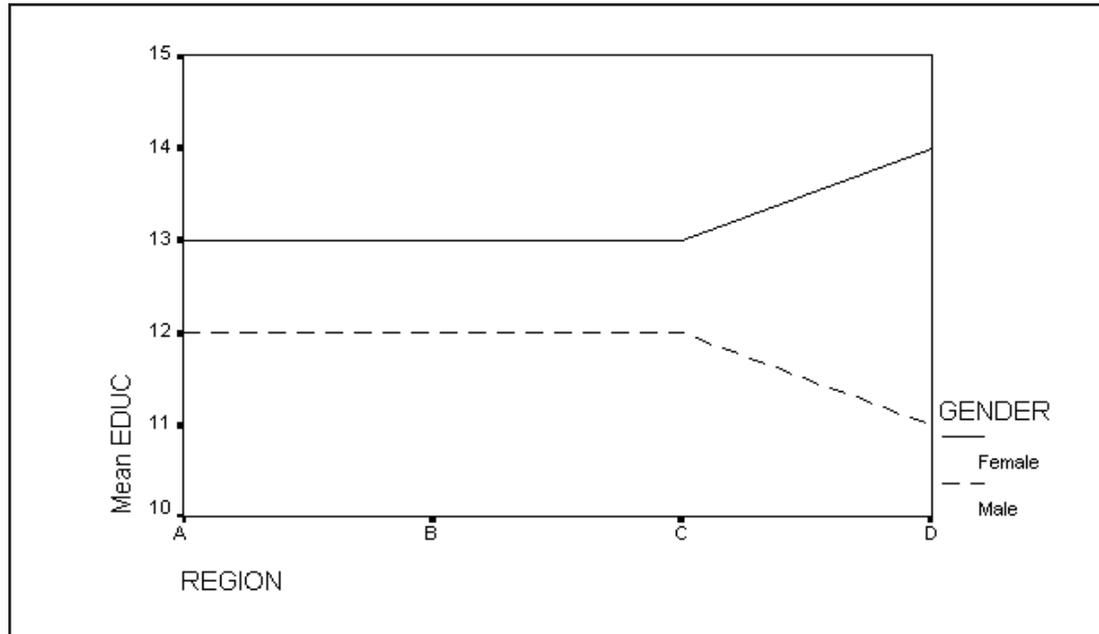
Illustration 4.2 No Main Effects, Strong Interaction



Here the overall means for men and women are about the same, as are the means for each location (pooling the two gender groups).

However, the gender differences vary dramatically across the different locations: in location B women have higher education, in locations A and D there is no gender difference, and in location C males have higher education. We cannot make a statement about gender differences without qualifying it with location information, nor can we make location claims without mentioning gender. Strong interactions are marked by this crossover pattern in a multiple-line chart.

**Illustration 4.3 One Main Effect, Weak Interaction**



We see a gender difference for each of the four locations, but the magnitude of this difference varies across locations (substantially greater for location D). This difference in magnitude of the gender effect would constitute an interaction between gender and location. It would be termed a weak interaction because there is no crossover of the mean lines.

Additional scenarios can be charted, and we have not mentioned three-way and higher interactions. Such topics are discussed in introductory statistics and analysis of variance books (see the reference page for suggestions). We will now proceed to analyze our data set.

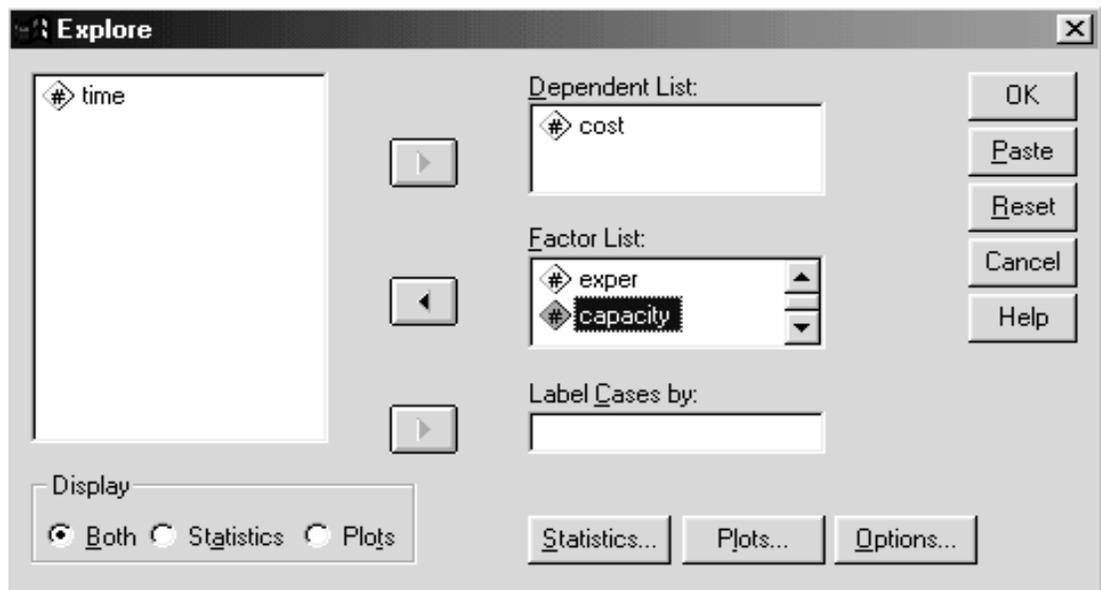
### EXPLORING THE DATA

We begin by applying exploratory data analysis to the cost of the plants within subgroups defined by combinations of capacity and experience. In practice, you would check each group's summaries, look for patterns in the data, and note any unusual points. Also, we will request that the Explore procedure perform a homogeneity of variance test.

Click on **File..Open..Data** (move to the **c:\Train\Anova** folder)  
Select **SPSS Portable (\*.por)** from the Files of Type **drop-down** list  
Double-click on **plant.por**

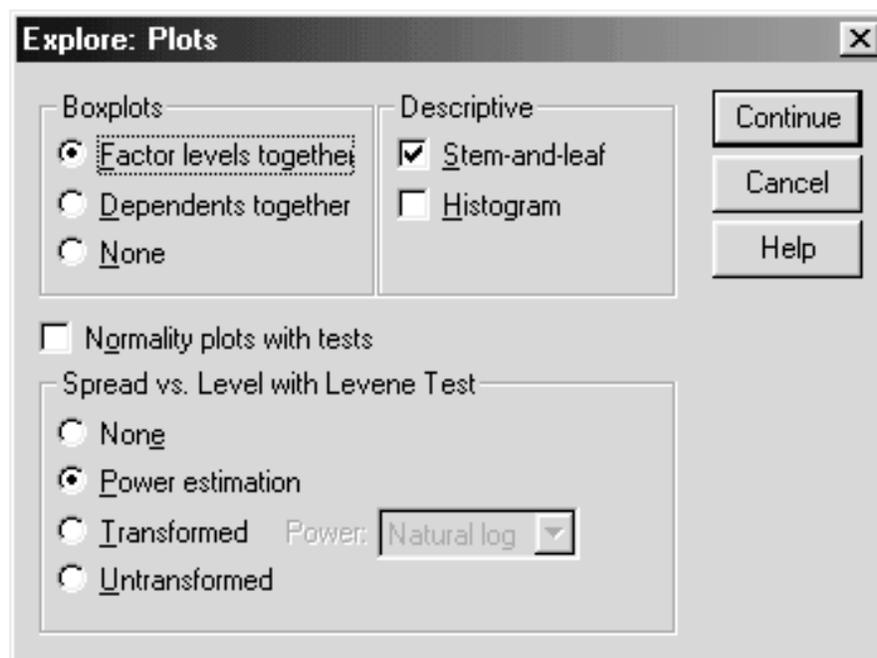
Click on **Analyze..Descriptive Statistics..Explore**  
Move **cost** into the **Dependent** List box  
Move the **exper** and **capacity** into the **Factor** List box.

Figure 4.1 Explore Dialog Box



Click on the **Plots** pushbutton  
Click the **Power estimation** option button in the “Spread vs. Level with Levene Test” area

Figure 4.2 Plots Dialog Box



By default, no homogeneity test is performed (“None” option button).  
Each of the remaining choices will lead to homogeneity being tested. The

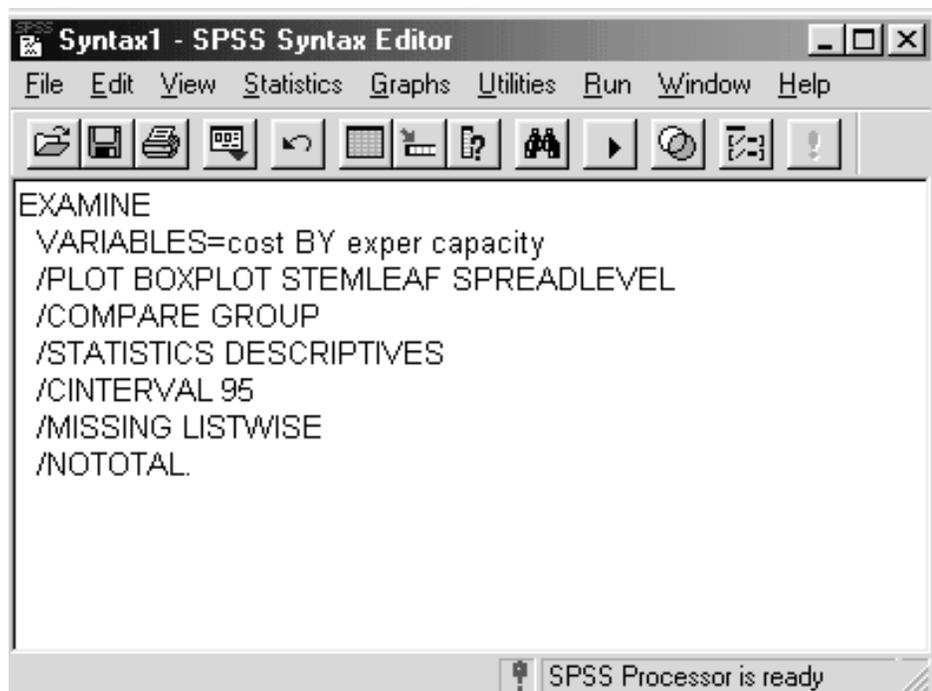
second (Power estimation) and third (Transformed) choices are used by more technical analysts to investigate power transformations of the dependent measure that would yield greater homogeneity of variance. These issues are of interest to serious practitioners of ANOVA, but are beyond the scope of this course (see Emerson in Hoaglin, Mosteller, and Tukey (1991), also a brief discussion in Box, Hunter, and Hunter (1978), and the original (technical) paper by Box and Cox (1964)). The Untransformed choice builds a plot without transforming the scale of the dependent measure.

Click on the **Continue** button to return to the **Explore** dialog box

Since we are comparing capacity by experience subgroups, we designate **exper** (experience) and **capacity** as the factors (or nominal independent variables). However, if we were to run this analysis, it would produce a set of summaries for each capacity group, then a set for each experience group. In other words, each of the two factors would be treated separately, instead of being combined, which we desire. To instruct SPSS to treat each capacity by experience combination as a subgroup we must use SPSS syntax. The easiest way to accomplish this would be to click the Paste pushbutton that opens a syntax window and builds an Examine command that will perform an analysis for each factor.

Click on the **Paste** pushbutton to paste the syntax into a Syntax window

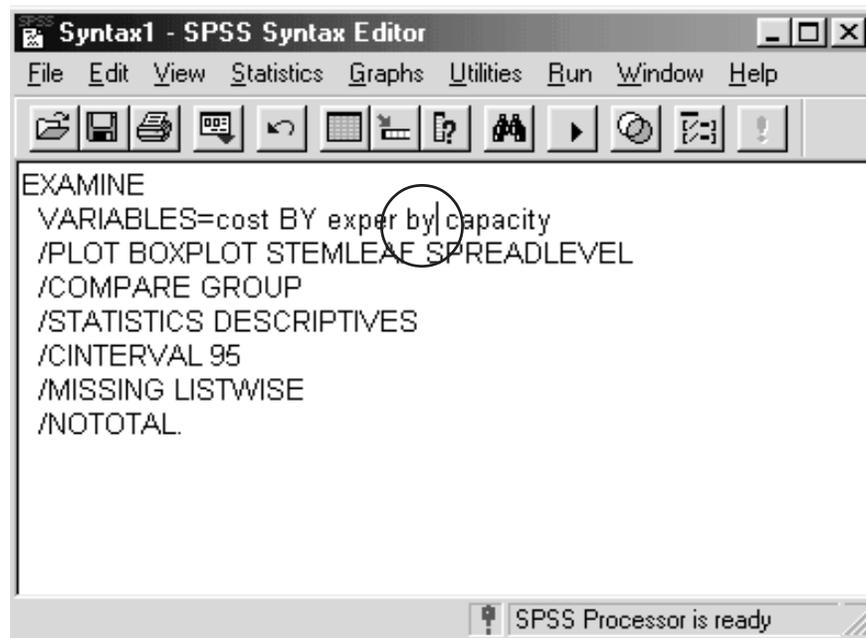
Figure 4.3 Examine Command in Syntax Window



The **Examine** command requires only the **Variables** subcommand in order to run. We also include the **Plot** subcommand since we desire the homogeneity test (controlled by the **SPREADLEVEL** keyword). The

other subcommands specify default values and appear in order to make it simple for the analyst to modify the command when necessary. Note the keyword **BY** separates the dependent variable **cost** from the **exper** and **capacity** factors. Currently, both **exper** and **capacity** follow the **BY** keyword and thus have the same status: an analysis will be run for each separately. To indicate we wish a joint analysis, we insert an additional **BY** keyword between **exper** and **capacity** on the **VARIABLES** subcommand.

Figure 4.4 Examine Command Requesting Subgroup Analysis



SPSS now interprets the factor groupings to be based on each capacity by experience combination (**exper BY capacity**).

Click **Run..Current** or click the Run button  .

Looking at the descriptives for each of our subgroups we see the following information.

**Note** All descriptive statistics appear in a single pivot table; the figures present separate sections of the table

Figure 4.5 Descriptives for 1-3 Plants and <800 MWe

EXPER	CAPACITY		Statistic	Std. Error		
COST	1-3 PLANTS	< 800 MWe	Mean	404.0829	50.9982	
			95% Confidence Interval for Mean	Lower Bound	279.2947	
			Upper Bound	528.8711		
			5% Trimmed Mean		394.5448	
			Median		350.6300	
			Variance		18205.745	
			Std. Deviation		134.9287	
			Minimum		289.66	
			Maximum		690.19	
			Range		400.53	
			Interquartile Range		106.1100	
			Skewness		1.987	.794
			Kurtosis		4.384	1.587

We find in this subgroup that the mean cost is 404.0829 and other descriptive information is available to us.

Figure 4.6 Descriptives for 1-3 Plants and >1000 MWe

EXPER	CAPACITY		Statistic	Std. Error		
COST	1-3 PLANTS	> 1000 MWe	Mean	592.4833	144.4059	
			95% Confidence Interval for Mean	Lower Bound	-28.8450	
			Upper Bound	1213.8117		
			5% Trimmed Mean		.	
			Median		452.9900	
			Variance		62559.173	
			Std. Deviation		250.1183	
			Minimum		443.22	
			Maximum		881.24	
			Range		438.02	
			Interquartile Range		.	
			Skewness		1.729	1.225
			Kurtosis		.	.

Figure 4.7 Descriptives for 4-9 Plants and <800 MWe

EXPER	CAPACITY		Statistic	Std. Error		
COST	4-9 PLANTS	< 800 MWe	Mean	275.8267	63.4457	
			95% Confidence Interval for Mean	Lower Bound	2.8420	
			Upper Bound	548.8114		
			5% Trimmed Mean		.	
			Median		217.3800	
			Variance		12076.061	
			Std. Deviation		109.8911	
			Minimum		207.51	
			Maximum		402.59	
			Range		195.08	
			Interquartile Range		.	
			Skewness		1.716	1.225
			Kurtosis		.	.

Figure 4.8 Descriptives for 4-9 Plants and 800-1000 MWe

EXPER	CAPACITY		Statistic	Std. Error		
COST	4-9 PLANTS	800-1000 MWe	Mean	339.3233	59.0816	
			95% Confidence Interval for Mean	85.1155		
			Lower Bound	593.5312		
			Upper Bound			
			5% Trimmed Mean	.		
			Median	288.4800		
			Variance	10471.924		
			Std. Deviation	102.3324		
			Minimum	272.37		
			Maximum	457.12		
			Range	184.75		
			Interquartile Range	.		
			Skewness	1.684		1.225
			Kurtosis			

Figure 4.9 Descriptives for 4-9 Plants and > 1000 MWe

EXPER	CAPACITY		Statistic	Std. Error		
COST	4-9 PLANTS	> 1000 MWe	Mean	493.2300	2.3500	
			95% Confidence Interval for Mean	463.3704		
			Lower Bound	523.0896		
			Upper Bound			
			5% Trimmed Mean	.		
			Median	493.2300		
			Variance	11.045		
			Std. Deviation	3.3234		
			Minimum	490.88		
			Maximum	495.58		
			Range	4.70		
			Interquartile Range	.		
			Skewness	.		.
			Kurtosis	.		.

Figure 4.10 Descriptives for 10 or more Plants and <800 MWe

EXPER	CAPACITY		Statistic	Std. Error		
COST	10 OR MORE PLANTS	< 800 MWe	Mean	518.3800	51.6841	
			95% Confidence Interval for Mean	296.0012		
			Lower Bound	740.7588		
			Upper Bound			
			5% Trimmed Mean	.		
			Median	473.6400		
			Variance	8013.741		
			Std. Deviation	89.5195		
			Minimum	460.05		
			Maximum	621.45		
			Range	161.40		
			Interquartile Range	.		
			Skewness	1.687		1.225
			Kurtosis			

Figure 4.11 Descriptives for 10 or more Plants and 800-1000 MWe

EXPER		CAPACITY		Statistic	Std. Error		
COST	10 OR MORE PLANTS	800-1000 MWe	Mean		453.7657	72.4561	
			95% Confidence Interval for Mean	Lower Bound	276.4719		
				Upper Bound	631.0595		
			5% Trimmed Mean		449.5741		
			Median		394.3600		
			Variance		36749.252		
			Std. Deviation		191.7009		
			Minimum		270.71		
			Maximum		712.27		
			Range		441.56		
			Interquartile Range		385.6300		
			Skewness		.395		.794
			Kurtosis		-2.112		1.587

Figure 4.12 Descriptives for 10 or more Plants and >1000 MWe

EXPER		CAPACITY		Statistic	Std. Error		
COST	10 OR MORE PLANTS	> 1000 MWe	Mean		663.8967	16.8749	
			95% Confidence Interval for Mean	Lower Bound	591.2896		
				Upper Bound	736.5037		
			5% Trimmed Mean		.		
			Median		652.3200		
			Variance		854.291		
			Std. Deviation		29.2283		
			Minimum		642.23		
			Maximum		697.14		
			Range		54.91		
			Interquartile Range		.		
			Skewness		1.503		1.225
			Kurtosis		.		.

a. COST is constant when EXPER = 1-3 PLANTS, CAPACITY = 800-1000 MWe. It has been omitted.

In the Viewer window we find a warning message concerning the spread and level plot and the test for homogeneity of variance. This message tells us that because we had a small number of cases in some of our subgroups that the median and/or the interquartile range was not defined. Thus the test and plot are not produced.

Figure 4.13 Test of Homogeneity of Variance

<p><b>Test of Homogeneity of Variance<sup>a</sup></b></p> <hr style="width: 20%; margin: auto;"/> <p>a. COST is constant when EXPER = 1-3 PLANTS, CAPACITY = 800-1000 MWe. It has been omitted.</p>
---

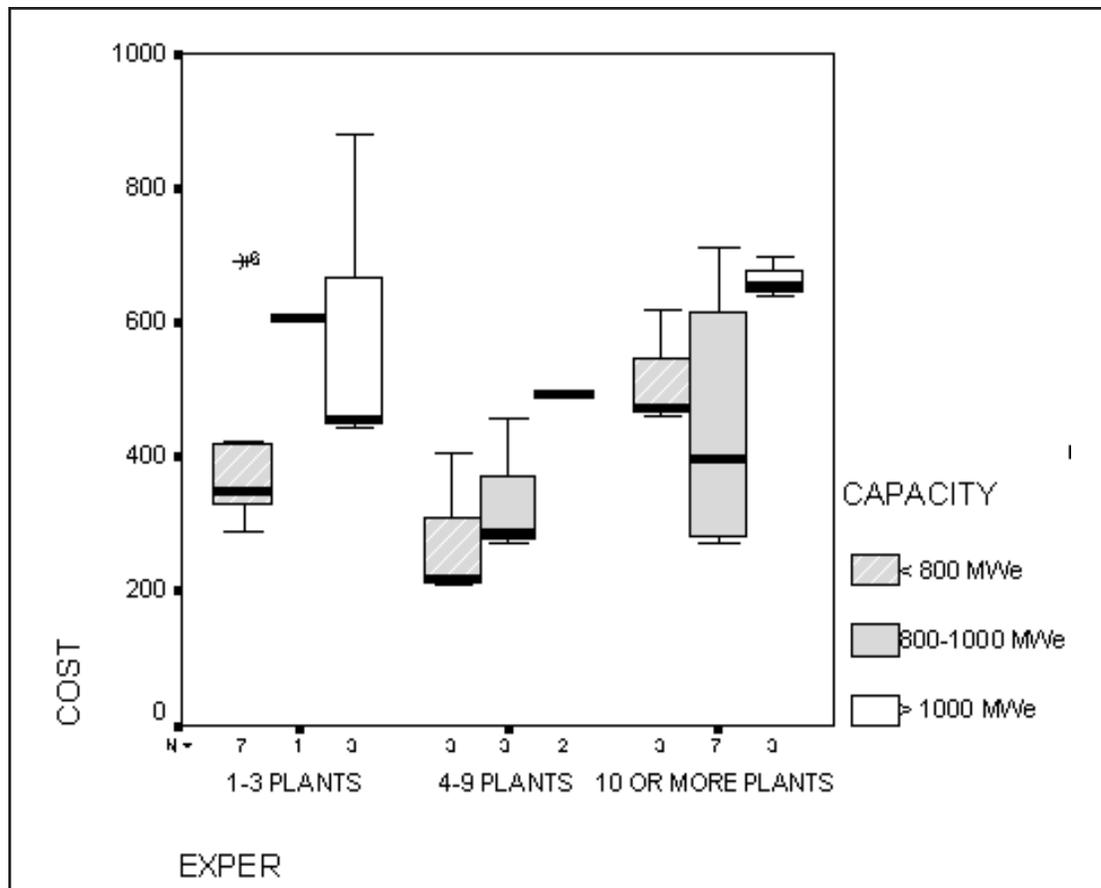
Figure 4.14 Warning Messages

**Warnings**

COST is constant when EXPER = 1-3 PLANTS, CAPACITY = 800-1000 MWe. It will be included in any boxplots produced but other output will be omitted.  
 A spread vs. level plot is not provided because the median and/ or interquartile range is not defined, using the HAVERAGE method, for COST when EXPER = 1-3 PLANTS, CAPACITY = > 1000 MWe.

Now we move to view the Box and Whiskers Plot.

Figure 4.15 Box and Whisker Plot of Cost



We see variation in the lengths of the boxes that suggests that the variation of cost within the subgroups is not homogeneous. We can see that there are differences in the median cost across our subgroups, but with the small sample sizes are they different enough to be statistically significant? An outlier is visible at the high end. Does it seem so extreme as to suggest a data error?

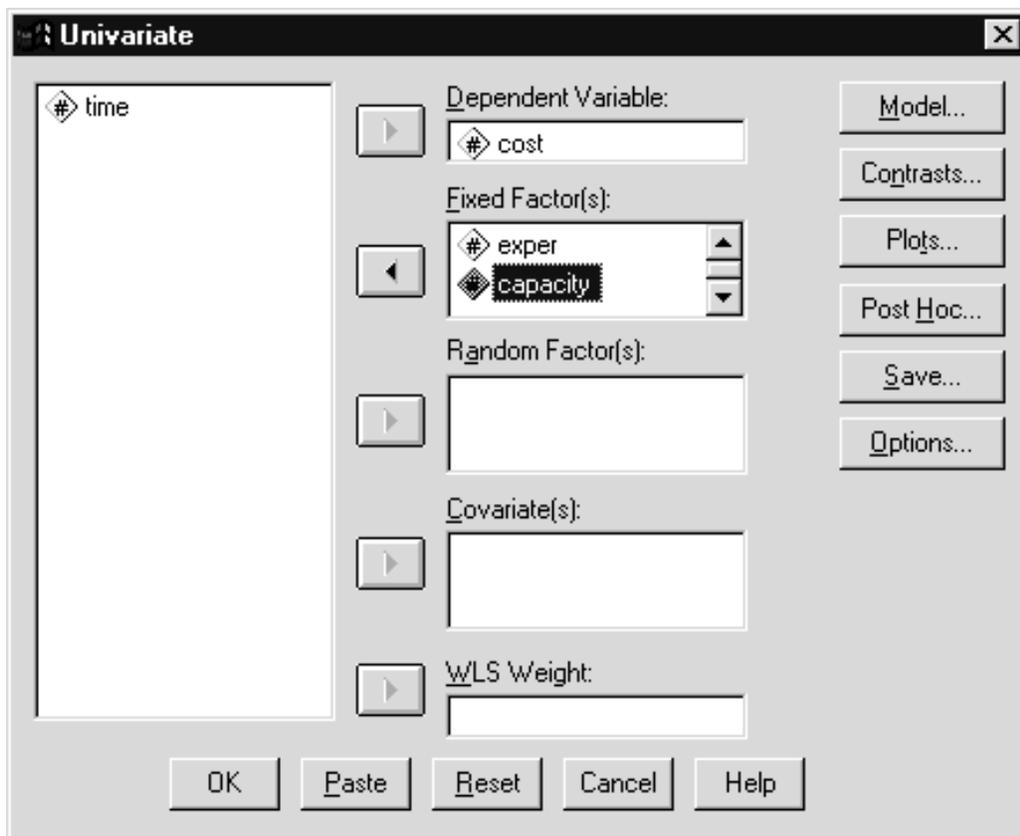
**TWO-FACTOR ANOVA**

To run the analysis in SPSS we choose **Analyze..General Linear Model** menu. Please note that the General Linear Model menu choices will vary depending on your version of SPSS and whether you have the SPSS Advanced Models option installed. We will use the Univariate procedure.

The Univariate choice permits the analyst to handle designs from the simple to the more complex (incomplete block, Latin square, etc.) and also provides the user with control over various aspects of the analyses. The Multivariate menu choice performs multivariate (multiple dependent measures) analysis of variance, while the Repeated Measures menu choice is used for studies in which an observation contributes to several factor levels (these are commonly called split-plot or repeated measure designs). Finally, the Variance Components menu choice performs an analysis that estimates the variation in the dependent variable attributable to each random effect in a model (see discussion of random and fixed effects below). Thus it assesses the relative influence of each random effect in a model containing multiple random effects.

Click on **Analyze..General Linear Model..Univariate**  
 Move **cost** and to the **Dependent Variable** list box  
 Move **exper** and **capacity** to the **Fixed Factor(s)** list box.

Figure 4.16 Univariate Dialog Box



Our analysis does not include random factors (other than the plant to plant variation that is already accounted for as the within-group variation. Briefly, fixed factors have a limited (finite) number of levels and we wish to draw population conclusions about only those levels.

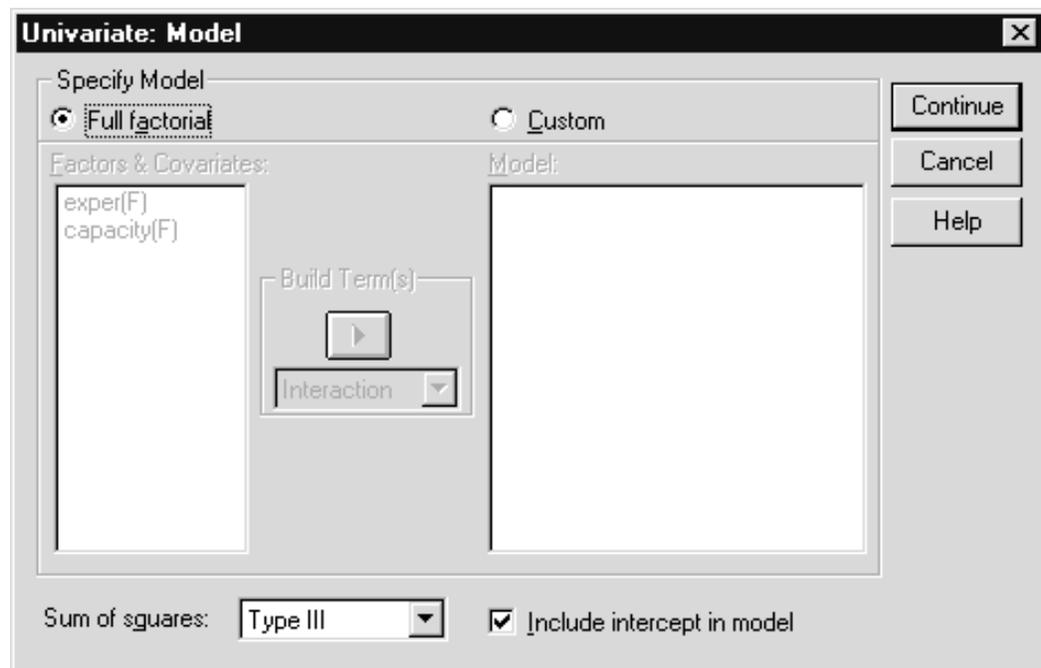
Random factors are those in which a random sample of a few levels from all possible ones are included in the study, but population conclusions are to be applied to all levels. For example, an institutional researcher might randomly select schools from a large school district to be included in a study investigating sex differences in learning mathematics. Here sex is a fixed factor while school is a random factor. It is important to distinguish between fixed and random factors since error terms differ.

Our analysis also does not include covariates. They are interval scale independent variables, whose relationships with the dependent measure you wish to statistically control, before performing the ANOVA itself.

The OK button is active, so we can run the analysis. However, we will request some additional information.

Click on the **Model** pushbutton.

Figure 4.17 Model dialog box



Within the Model dialog you can specify the model you want applied to the data. By default a model containing all main effects and interactions is run. Analysts who analyze data based on incomplete designs (some combinations of factors are not evaluated in order to reduce the sample size requirements) would use this dialog to indicate which effects should be evaluated. Also, if your sample sizes are unequal across subgroups you can choose among several sums of squares adjustments. This issue is discussed later in the chapter.

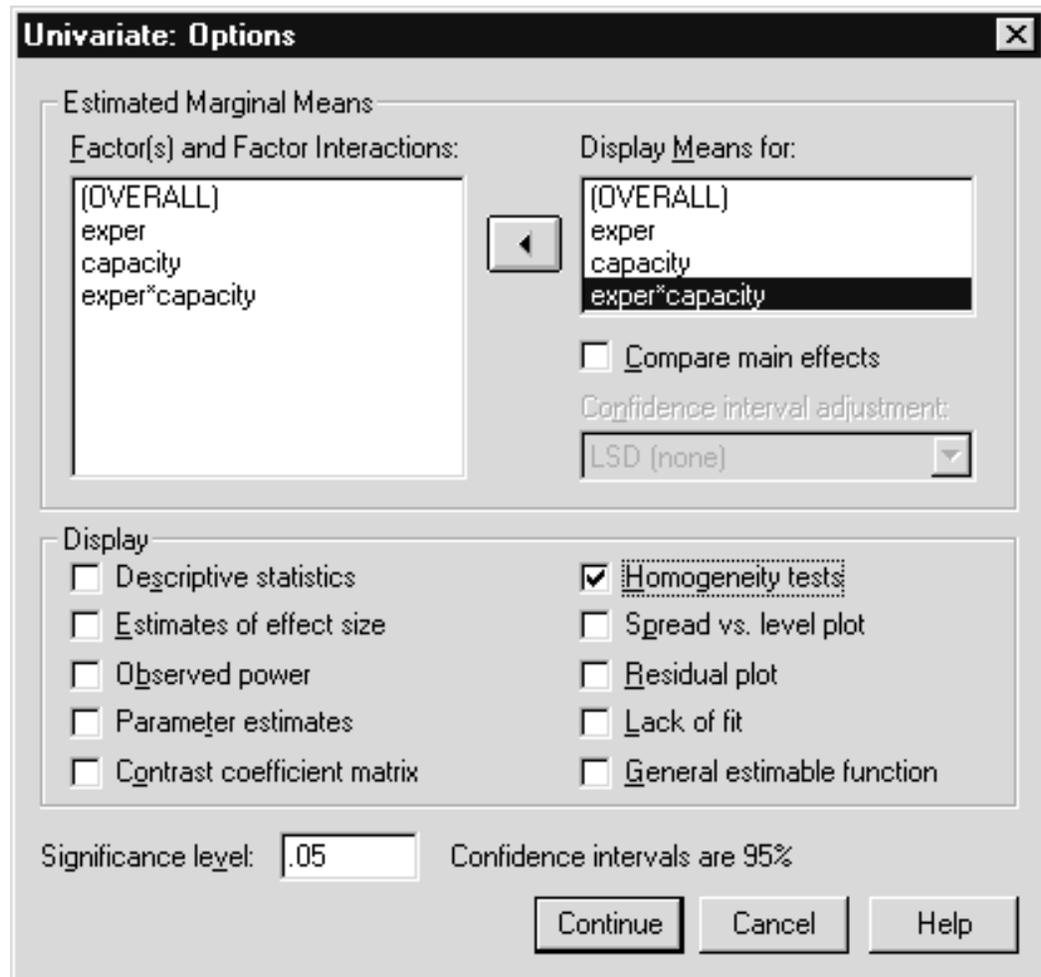
Click the **Cancel** button.

The next pushbutton we will look at is the Options button. We ask Univariate to provide us with means for the main effects and the two-way

interaction. Also we will request a test of homogeneity of variance.

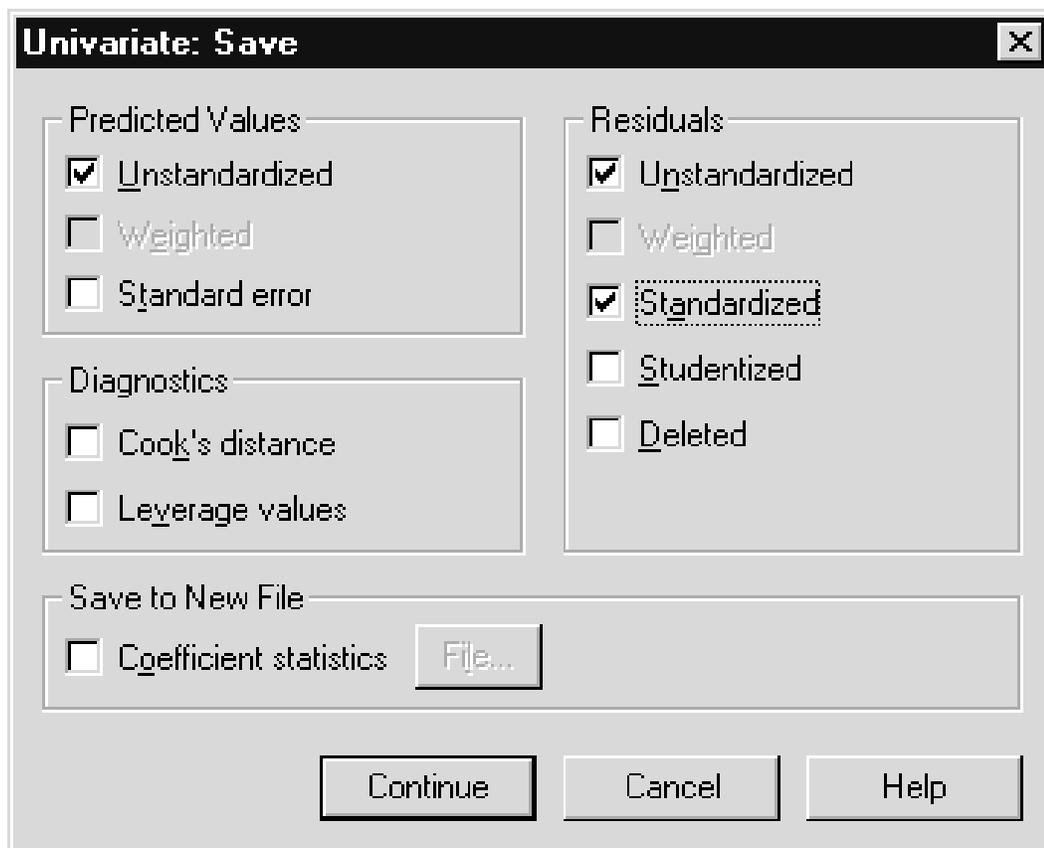
- Click the **Options** pushbutton
- Move **(Overall)**, **exper**, **capacity**, and **exper \*capacity** into the **Display Means for** list box
- Click the **Homogeneity tests** check box

Figure 4.18 Univariate: Options Dialog Box



- Click on **Continue**
- Click the **Save** pushbutton
- Click the check boxes for **Unstandardized Predicted Values** and both the **Unstandardized** and **Standardized Residuals**.

Figure 4.19 Univariate: Save Dialog Box



Click on **Continue**  
Click on **OK**.

The following syntax will also run the analysis:

```
UNIANOVA
  cost BY exper capacity
  /METHOD = SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED RESID ZRESID
  /EMMEANS=TABLES(OVERALL)
  /EMMEANS=TABLES(exper)
  /EMMEANS=TABLES(capacity)
  /EMMEANS=TABLES(exper*capacity)
  /PRINT=HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=exper capacity exper*capacity.
```

Now we will look at the output from our analysis. The first result is a listing of the Between-Subjects Factors.

Figure 4.20 Between Subjects Factors

**Univariate Analysis of Variance**

**Between-Subjects Factors**

		Value Label	N
EXPER	1.00	1-3 PLANTS	11
	2.00	4-9 PLANTS	8
	3.00	10 OR MORE PLANTS	13
CAPACITY	1.00	< 800 MWe	13
	2.00	800-1000 MWe	11
	3.00	> 1000 MWe	8

Next we see the result of the Levene's test of equality of error variances (homogeneity of variance test).

Figure 4.21 Levene's Test of Homogeneity of Variance

**Levene's Test of Equality of Error Variances<sup>a</sup>**

Dependent Variable: COST

F	df1	df2	Sig.
3.184	8	23	.014

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

a. Design: Intercept+EXPER+CAPACITY+EXPER \* CAPACITY

The significance level is .014 which means that if the error variances were equal in the population, we would get an "F" statistic this large only 14 times in one thousand. Thus the homogeneity of variance assumption does not hold. A technical analyst might move to the spread and level plots to see if the dependent variable can be transformed such that homogeneity of variance holds. A nonparametric analysis could be done, although SPSS currently does not contain a two-factor nonparametric Anova procedure). We will proceed with the analysis, realizing that the test results may not be completely accurate.

**THE ANOVA TABLE**

The ANOVA table contains the information, much of it technical, necessary to evaluate whether there are significant differences in cost across capacity groups, across experience groups, and whether the two factors interact.

**Figure 4.22 The ANOVA Table**

Tests of Between-Subjects Effects					
Dependent Variable: COST					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	379480.902 <sup>a</sup>	8	47435.113	2.107	.077
Intercept	5480467.2	1	5480467.2	243.486	.000
EXPER	158495.497	2	79247.749	3.521	.046
CAPACITY	154061.360	2	77030.680	3.422	.050
EXPER * CAPACITY	49479.036	4	12369.759	.550	.701
Error	517691.407	23	22508.322		
Total	7714385.8	32			
Corrected Total	897172.309	31			

a. R Squared = .423 (Adjusted R Squared = .222)

The first column lists the different sources of variation. We are interested in the capacity and experience main effects, as well as the capacity by experience interaction. The source labeled “Error” contains summaries of the within-group variation (or Residual term) which will be used when calculating the “F” ratios (ratios of between-group to within-group variation). The remaining sources in the list are simply totals involving the sources already described, and as such are generally not of interest. The Sum of Squares column contains a technical summary (sum of the squared deviations of group means around the overall mean, or of individual observations around their group mean) that is not interpreted directly, but is used in calculating the later column values. The “df” (degrees of freedom) column contains values that are functions of the number of levels of the factors (for capacity, experience, and capacity by experience) or the number of observations (for residual). Although this is a gross oversimplification, you might think of degrees of freedom as measuring the number of independent values (whether means or observations) that contribute to the sum of squares in the previous column. As with sums of squares, degrees of freedom are technical measures, not interpreted themselves, but used in later calculations.

Mean Square values are variance measures attributable to the various effects (capacity, experience, capacity by experience) and to the variation of individuals within groups (error). The ratio of an effect mean square to the mean square of the error provides the between-group to within-group variance ratio, or “F” statistic. If there were no group differences in the population, then the ratio of the between-group

variation to the within-group variation should be about one. The column “Sig” contains the most interpretable numbers in the table: the probabilities that one can obtain “F” ratios as large or larger (or group means as far or farther apart) as what we find in our sample, if there were no mean differences in the population.

The ANOVA table summarizes the statistical testing. Both experience and capacity are marginally significant at the .046 and .050 respectively. The result for capacity is similar but not identical to its result in the one factor ANOVA for several reasons. First, the within-groups error term is now based on nine cells and not only three as before. Also, since the sample sizes are neither equal nor proportional, the effects of capacity and experience are not independent of each other and the test for capacity adjusts for the experience factor. In the one factor analysis the second factor was ignored. The interaction is not significant indicating that the capacity differences do not change across different levels of experience.

**Conclusion** Average cost shows significant differences across levels of plant capacity and levels of building experience. The two factors do not seem to interact.

**PREDICTED MEANS** In the Options dialog box we asked for the means to be displayed for each main effect and the interaction. The following figures provide those requested means.

Figure 4.23 Grand Mean and Means for Experience Levels

Estimated Marginal Means				
<b>1. Grand Mean</b>				
Dependent Variable: COST				
Mean	Std. Error	95% Confidence Interval		
		Lower Bound	Upper Bound	
483.310	30.973	419.237	547.383	
<b>2. EXPER</b>				
Dependent Variable: COST				
EXPER	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1-3 PLANTS	535.122	60.761	409.429	660.815
4-9 PLANTS	369.460	54.016	257.719	481.201
10 OR MORE PLANTS	545.347	44.995	452.268	638.427

Note that surprisingly, the mean cost is lowest for the middle level of experience.

Figure 4.24 Means for Capacity Levels

3. CAPACITY				
Dependent Variable: COST				
CAPACITY	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
< 800 MWe	399.430	44.995	306.350	492.509
800-1000 MWe	467.296	60.761	341.604	592.989
> 1000 MWe	583.203	54.016	471.462	694.944

Figure 4.25 Means for Capacity\*Experience Levels

4. EXPER * CAPACITY					
Dependent Variable: COST					
EXPER	CAPACITY	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1-3 PLANTS	< 800 MWe	404.083	56.705	286.779	521.386
	800-1000 MWe	608.800	150.028	298.444	919.156
	> 1000 MWe	592.483	86.619	413.299	771.667
4-9 PLANTS	< 800 MWe	275.827	86.619	96.643	455.011
	800-1000 MWe	339.323	86.619	160.139	518.507
	> 1000 MWe	493.230	106.086	273.775	712.685
10 OR MORE PLANTS	< 800 MWe	518.380	86.619	339.196	697.564
	800-1000 MWe	453.766	56.705	336.462	571.069
	> 1000 MWe	663.897	86.619	484.713	843.081

**ECOLOGICAL SIGNIFICANCE**

We have found that both main effects are statistically significant although the assumption of homogeneity of the variances is not met and may compromise the results. Also the analyst must ask him or herself if any differences are significant in a practical sense. It is again important to recall that a statistically significant mean difference implies that the population difference is not zero. Differences can be small yet statistically significant when the sample size is large. This effect of large samples is certainly not a problem in this study.

**RESIDUAL ANALYSIS**

To view the predicted values and residuals we turn to the case summary procedure, although we could simply examine them in the Data Editor window.

Click **Analyze..Reports..Case Summaries**

Move **cost, pre\_1, res\_1, and zre\_1** to the Variables list box

Click on **OK**

The following syntax will also produce the case summary report.

```
SUMMARIZE
/TABLES=cost pre_1 res_1 zre_1
/FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=100
/TITLE='Case Summaries' /FOOTNOTE ''
/MISSING=VARIABLE
/CELLS=COUNT.
```

**Figure 4.26 Case Summary Report**

**Case Summaries<sup>a</sup>**

	COST	Predicted Value for COST	Residual for COST	Standardized Residual for COST
1	345.390	404.083	-58.693	-.391
2	317.210	404.083	-86.873	-.579
3	423.320	404.083	19.237	.128
4	289.660	404.083	-114.423	-.763
5	412.180	404.083	8.097	.054
6	690.190	404.083	286.107	1.907
7	350.630	404.083	-53.453	-.356
8	402.590	275.827	126.763	.845
9	217.380	275.827	-58.447	-.390
10	207.510	275.827	-68.317	-.455
11	473.640	518.380	-44.740	-.298
12	621.450	518.380	103.070	.687
13	460.050	518.380	-58.330	-.389
14	608.800	608.800	.000	.000
15	288.480	339.323	-50.843	-.339
16	272.370	339.323	-66.953	-.446
17	457.120	339.323	117.797	.785
18	394.360	453.766	-59.406	-.396
19	712.270	453.766	258.504	1.723
20	567.790	453.766	114.024	.760
21	665.990	453.766	212.224	1.415
22	284.880	453.766	-168.886	-1.126
23	280.360	453.766	-173.406	-1.156
24	270.710	453.766	-183.056	-1.220

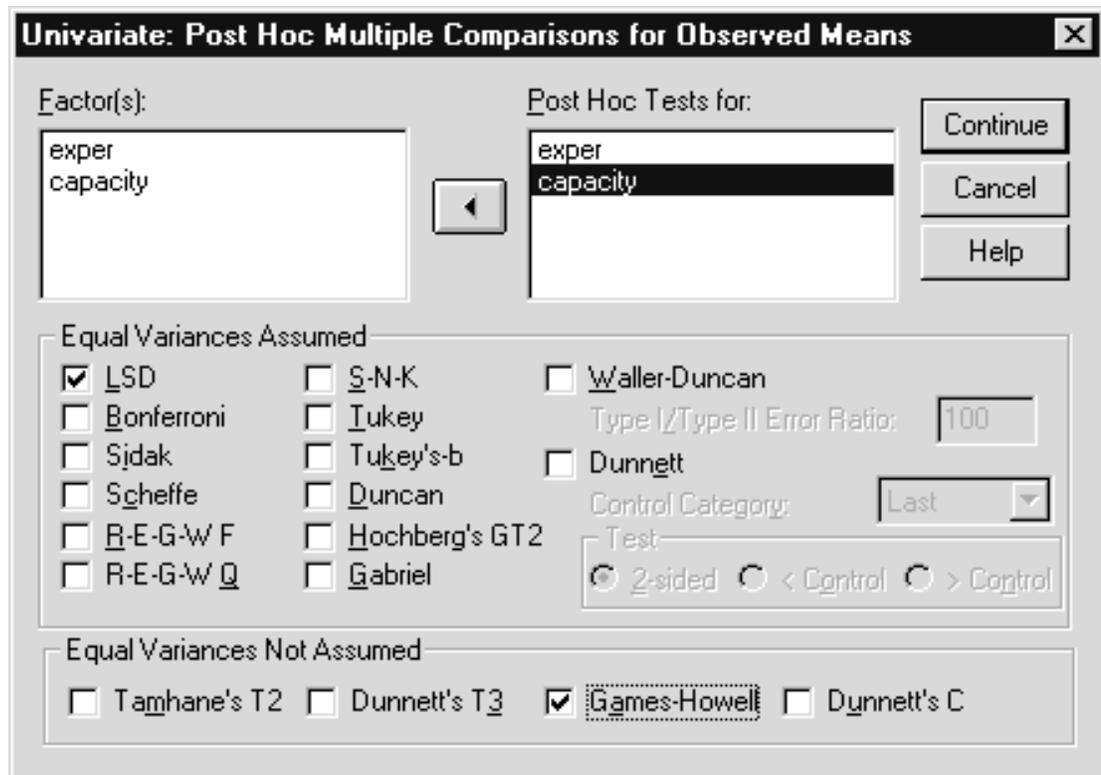
Notice the predicted values are identical for all cases in the same cell, that is, group membership determines the predicted value. The standardized residuals are in standard deviation units; do you see any surprisingly large residuals?

**POST HOC TESTS OF ANOVA RESULTS**

To run post hoc tests on our results we will need to re-open the Univariate dialog box.

- Click the Dialog Recall Tool , then click **Univariate**
- Click the **Post Hoc** pushbutton
- Move **exper** and **capacity** into the **Post Hoc Tests for** box
- Select the **LSD**, **Games-Howell**, and **Scheffe** post hoc tests (click their check boxes)

Figure 4.27 Post Hoc Test Dialog Box



- Click **Continue**
- Click **OK**.

As shown below, the Posthoc subcommand requests the post hoc tests.

UNIANOVA

```

cost BY exper capacity
/METHOD = SSTYPE(3)
/INTERCEPT=INCLUDE
/SAVE=PRED RESID ZRESID
/POSTHOC = capacity exper ( SCHEFFE LSD GH )
/EMMEANS=TABLES(OVERALL)
/EMMEANS=TABLES(exper)
/EMMEANS=TABLES(capacity)
/EMMEANS=TABLES(exper*capacity)
/PRINT=HOMOGENEITY
/CRITERIA=ALPHA(.05)
/DESIGN=exper capacity exper*capacity.
    
```

We have selected the LSD, Games-Howell (because of the failure of the homogeneity of variance assumption), and Scheffe tests. We will examine the post hoc tests only for capacity (since this chapter is lengthy as it is). In practice you would examine the results for both capacity and experience if they were found to be significant. If time permits, review the post hoc test results for experience. What do you find?

Figure 4.28 Post Hoc Tests For Capacity

Multiple Comparisons

Dependent Variable: COST

	(I) CAPACITY	(J) CAPACITY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	< 800 MWe	800-1000 MWe	-35.7866	61.462	.845	-196.5816	125.0084
		> 1000 MWe	-193.5885*	67.416	.029	-369.9598	-17.2172
	800-1000 MWe	< 800 MWe	35.7866	61.462	.845	-125.0084	196.5816
		> 1000 MWe	-157.8018	69.712	.099	-340.1790	24.5753
	> 1000 MWe	< 800 MWe	193.5885*	67.416	.029	17.2172	369.9598
		800-1000 MWe	157.8018	69.712	.099	-24.5753	340.1790
LSD	< 800 MWe	800-1000 MWe	-35.7866	61.462	.566	-162.9312	91.3579
		> 1000 MWe	-193.5885*	67.416	.009	-333.0496	-54.1273
	800-1000 MWe	< 800 MWe	35.7866	61.462	.566	-91.3579	162.9312
		> 1000 MWe	-157.8018*	69.712	.033	-302.0119	-13.5917
	> 1000 MWe	< 800 MWe	193.5885*	67.416	.009	54.1273	333.0496
		800-1000 MWe	157.8018*	69.712	.033	13.5917	302.0119
Games-Howell	< 800 MWe	800-1000 MWe	-35.7866	61.462	.849	-201.5316	129.9584
		> 1000 MWe	-193.5885*	67.416	.029	-367.3995	-19.7774
	800-1000 MWe	< 800 MWe	35.7866	61.462	.849	-129.9584	201.5316
		> 1000 MWe	-157.8018	69.712	.120	-351.0644	35.4608
	> 1000 MWe	< 800 MWe	193.5885*	67.416	.029	19.7774	367.3995
		800-1000 MWe	157.8018	69.712	.120	-35.4608	351.0644

Based on observed means.

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

We can see from the post hoc results with the LSD testing that both the less than 800 MWe and 800-1000 MWe plants were different from the over 1000 MWe plants. This however is the most liberal test. The two

other tests find only that the less than 800MWe plants are different from the over 1000MWe plants.

Figure 4.29 Homogenous Subsets for Capacity

COST			
CAPACITY	N	Subset	
		1	2
Scheffe <sup>a,1</sup> < 800 MWe	13	400.8615	
800-1000 MWe	11	436.6482	436.6482
> 1000 MWe	8		594.4500
Sig.		.865	.079

Means for groups in homogeneous subsets are displayed.  
 Based on Type III Sum of Squares  
 The error term is Mean Square(Error) = 22508.322.

- Uses Harmonic Mean Sample Size = 10.245.
- The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.
- Alpha = .05.

As we would expect given the post hoc results, the homogeneous subsets produced by the Scheffe test confirms that only the lowest and highest capacity groups differ.

**UNEQUAL  
 SAMPLES AND  
 UNBALANCED  
 DESIGNS**

Up to now we have not discussed the implications of unequal sample sizes. The basic problem arises when the sample sizes are not equal across groups (or not proportional if you are mainly interested in main effects). When this occurs, or if cells are missing entirely, the effects in the analysis become correlated, that is, they overlap. As the cell size imbalance increases, it becomes increasingly difficult to speak of independent effects. For example, if almost all high-capacity plants were built by people with experience building 10 or more plants, how can we speak of separate effects? The same problem, high correlation among predictor variables, is frequently discussed in the literature on regression. There are different methods for adjusting for such overlap of effects and we discuss some of these approaches and their implications below.

### SUMS OF SQUARES

From the Model dialog box you can choose a type of sums of squares. Type III is the most commonly used and is the default. Each type adjusts for unequal sample sizes in a different way. When all subgroup sample sizes are the same, the various sums of squares' calculations yield the identical result.

- Type I. This method is also known as hierarchical decomposition of the sum-of-squares. Each term is adjusted for only the terms that precedes it in the model. Type I sums of squares are commonly used in situations in which the researcher has a prior ordering of effects in mind. For example, if previous research has always found a factor to be significant there might be interest in determining if a second factors makes a substantial contribution. In this situation, the known factor might be entered first in the model (not adjusting for the second factor), while the new factor follows in the model (so it is tested after adjusting for the first factor).
- Type II. This method calculates the sums of squares of an effect in the model adjusted for all other "appropriate" effects. An appropriate effect is one that corresponds to all effects that do not contain the effect being examined. Thus a main effect would adjust for all other main effects but interactions. A two-way interaction would adjust for all main effects and other two-way interactions, but ignore three-way and higher effects.
- Type III. This is the default. This method calculates the sums of squares of an effect in the design as the sums of squares adjusted for any other effects that do not contain it and orthogonal to any effects (if any) that contain it. Essentially, each effect is adjusted for all other effects (main effects, same order interactions, higher order interactions) in the model. Thus you can speak of an effect independent of all other effects. The Type III sums of squares have a major advantage in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant. In practice, this means that Type III sums of squares is often considered useful for an unbalanced model with no missing cells. In a factorial design with no missing cells, this method is equivalent to the Yates' weighted-squares-of-means technique. Type III is recommended on the strength of the fact that a statistical test for an effect adjusts for all other effects in the model. However, if there are missing data cells, Type IV is preferred.
- Type IV. This method is designed for situations in which there are missing cells. The technical description of Type IV sums of squares follows. For any effect F in the design, if F is not contained in any other effect, the Type IV = Type III = Type II. When F is contained in other effects, Type IV distributes the contrasts being made among the parameters in F to all higher-level effects equitably. To give a practical example, suppose we were testing salary differences due to

two factors: experience (in three categories) programming in a computer language, and the computer language itself (two categories). If there were no programmers with the highest experience level for one of the languages (say Java), then that experience category would not be used when evaluating the computer language main effect. Thus the computer language effect would be evaluated from only those experience categories containing programmers of both languages. This is the source of the equity mentioned above. The Type IV sum-of-squares method is commonly used for an unbalanced model with empty cells.

### **EQUIVALENCE AND RECOMMENDATIONS**

All of these types of sums of squares are equivalent when there is only one effect to be tested. Thus the one-way ANOVA procedure does not offer any options in terms of sums of squares. Also as mentioned above, they give identical results for a balanced design. In practice today, the Type III sums of squares method is usually used if there are no missing cells. If cells are missing the Type IV method is generally chosen. When a researcher wants to test effects after adjusting for certain effects, but ignoring others, then the Type I or Type II methods are employed.

### **EMPTY CELLS AND NESTED DESIGNS**

Any time that the between-subject portion of an analysis cannot be laid out in a full factorial setup with all cells filled (having at least one observation), matters can become quite complicated, and knowledge of the theory of estimable functions is required in order to determine just what hypotheses can be tested. The best advice that can be given here is to consult a statistician knowledgeable in experimental design in order to determine the appropriate, testable hypotheses in a particular case

Virtually any testable hypothesis can be tested using the General Linear Model - Univariate procedure with its flexible DESIGN subcommand, but determining the appropriate hypothesis to test when there are missing cells can be extremely difficult. It should be noted in particular that simply applying standard sets of commands to such data can produce results that are uninterpretable, since the particular hypotheses tested have not been identified.

For further information on the analysis of such data, see Searle (1987) or Milliken and Johnson (1984). Of the two, Searle's book is more complete but rather technical. Milliken and Johnson's book is more accessible.

### **SUMMARY**

In this chapter we generalized ANOVA to the case with two or more factors and discussed post hoc comparisons in the context. In addition, the effects of unequal sample size and missing cells were presented. We turn next to another generalization: ANOVA with multiple dependent measures – multivariate analysis of variance.

# Chapter 5 Multivariate Analysis Of Variance

**Objective** The purpose of this chapter is to understand the properties of multivariate analysis of variance, drawing on our previous discussions of univariate ANOVA.

**Method** We run the EXPLORE procedure to check on some of the assumptions of the multivariate ANOVA. We will use the General Linear Model-Multivariate procedure to run a two-factor multivariate analysis of variance with two dependent variables.

**Data** We use the same data set as in the prior chapters, i.e., the nuclear power plant data set (plant.por).

**Design** A two-factor two dependent variable multivariate analysis of variance – experience and plant capacity are the two fixed factors (3 levels each), cost and time (time before plant was licensed) are the dependent measures.

## INTRODUCTION

**M**ultivariate analysis of variance (MANOVA) is a generalization of analysis of variance that permits testing for mean differences on several dependent measures simultaneously. In this chapter we will explore the rationale and assumptions of multivariate analysis of variance, review the key summaries to examine in the results, and then step through an analysis looking at group differences on two measures in our data set.

Multivariate analysis of variance is used when there is an interest in testing for mean differences between groups on several dependent variables simultaneously. ANOVA will test whether the mean of a single variable (scalar) differs across groups. MANOVA covers the broader case of testing for mean differences in several variables (vector) across groups.

### WHY PERFORM MANOVA?

Multivariate analysis of variance (MANOVA) tests for population group differences on several dependent measures simultaneously. Instead of examining differences for a single outcome variable (as analysis of variance does), MANOVA tests for differences on a set or vector of means. The outcome measures (dependent variables) are typically related; for example, a set of ratings of employee performance, multiple physiological measures of stress, several scales assessing an attitude, a collection of fitness measures, multiple scales measuring a product's appearance, several measures of the fiscal health of a company.

MANOVA is typically performed for two reasons: statistical power and control of false positive results (also known as Type I error).

First, there can be greater statistical power, that is, the ability to detect true differences, in a multivariate analysis. The argument is that if you have several imperfect measures of an outcome, for example, several physiological measures of stress, the joint analysis will be more likely to show a true difference in stress than any individual analysis. A multivariate analysis compares mean differences across several variables and takes formal account of their intercorrelations. In this way a small difference appearing in several related outcome variables may result in a significant multivariate test, although no single outcome measure shows a significant difference. This is not to say there is a power advantage in throwing 20 or so unrelated variables into a multivariate analysis of variance, since a true difference in a single outcome measure can be diluted in a joint test involving many variables that display no effect. However, if you are interested in studying group differences in outcomes for which various measures exist (this occurs in marketing, social science, medical, ecological, and engineering studies), then MANOVA probably carries greater statistical power.

The second argument for running MANOVA in place of separate univariate (single outcome variable) analyses concerns controlling the false positive rate when multiple tests are done. If a separate ANOVA is run for every outcome variable, each tested at the 0.05 level, then the overall (or experiment-wise) false positive rate (chance of obtaining one or more false positive test results) is well above 5 in 100 (0.05 or 5%) because of the multiple tests. A MANOVA applied to a study with seven outcome measures would result in a single test performed at the 0.05 level. Although there are certainly alternative methods for controlling the false positive rate when multiple tests are performed (for example, Bonferroni adjustments), using a multivariate test accomplishes this as well. Some researchers follow the procedure of first performing a multivariate test and only if the results are significant would they examine the individual univariate test results. This provides some control over the false positive rate. It is not a perfect solution (it is similar to the argument for the LSD multiple comparison procedure) and has received some criticism (see Huberty (1989)).

### HOW MANOVA DIFFERS FROM ANOVA

First, the good news, MANOVA is similar to ANOVA in that variation between group means is compared to variation of individuals within groups. Since this variation is measured on several variables, MANOVA computes a matrix containing the variation and covariation (there are several variables!) of the vector of group means and a second matrix containing within-group variances and covariances. When testing in MANOVA, a ratio is taken not of the two variances (two numbers), but of two matrices. Instead of the usual “F” test, the multivariate form – called a generalized “F” is used. The summary table will contain some unfamiliar statistics, but in the end will report the probability of obtaining means as far (or farther) apart as you did by chance alone, just as ANOVA did.

In short, while the required matrix notation used while deriving or describing MANOVA is a bit intimidating, the same principles that have guided us so far in analysis using ANOVA – variation between group means compared to variation within groups – still holds true in MANOVA. The statistics change because we are now talking about vectors of means (a set of means) being tested jointly.

### ASSUMPTIONS OF MANOVA

The assumptions made when performing multivariate analysis of variance are largely extensions of those made under ordinary analysis of variance. In addition to the usual assumptions for a linear model (additivity, independence between the error and model effects, independence of the errors), MANOVA testing assumes that the residual errors follow a multivariate normal distribution in the population; this is a generalization of the normality assumption made in ANOVA. In SPSS you can examine and test individual variables for normality within each group. This is not equivalent to testing for multivariate normality, but is still quite useful in evaluating the assumption. In addition, homogeneity of variance, familiar from ANOVA, has a multivariate extension concerning homogeneity of the within-group variance-covariance matrices. A multivariate test of homogeneity of variance (Box’s M test) is available to check this assumption.

For large samples, we expect departures from normality to make little difference. This is due to the central limit theorem argument combined with the fact that in MANOVA we are generally testing simple functions of group means. If the samples are small and multivariate normality is violated, the results of the analysis may be effected. Data transformations (for example, logs) on the dependent measure(s) may alleviate the problem, but have potential problems of their own (interpretation, incomplete equivalence between tests in the transformed and untransformed scales). Unfortunately, a general class of multivariate nonparametric tests is not currently available; developments in this area would help provide a solution.

Concerning homogeneity of variance, in practice if the sample size is similar across groups then moderate departures from homogeneity of the within-group variance-covariance matrices do not effect the analysis. If homogeneity does not hold and the sample size varies substantially across groups, then test results can be effected. In the simplest scenarios,

the direction of the effect depends on which sized group has the larger variances, but specific situations can be far more complex, in which case little is known.

### **WHAT TO LOOK FOR IN MANOVA**

After investigating whether the assumptions are met, primary interest would be in the multivariate statistical tests. If significant effects are found you might then examine univariate results. Additionally, you can perform post hoc comparisons to discover just where the differences reside.

### **SIGNIFICANCE TESTING**

When testing for mean differences between groups on a single dependent variable the test comes down to a ratio of between-group to within-group variation – a single number. In multivariate analysis, we are left with two matrices containing the between and within-group variation and covariation. There are different test statistics that can apply. Most of them involve computing the latent roots of some function of the ratio of the two matrices. In depth discussion of these test measures is beyond the scope of this course, but some comments about their characteristics will be made when we review the results.

### **Proposed Analysis**

We will perform MANOVA with experience and plant capacity as the factors with time to completion (licensing) and cost as dependent variables. This pairing of dependent variables makes sense if you believe the adage that time is money. We expect that plants that require more time to build should also be more costly. By using both variables we hope to tap a more general measure of cost.

### **CHECKING THE ASSUMPTIONS**

The analysis we conducted in Chapter 4 investigated the properties of the cost measure. There was some evidence for heterogeneity of variance, although the tests were not in complete agreement. The normal plot of the errors suggested some skewness. We will now proceed to look at some of the information from the EXPLORE procedure for the time variable. One caution, these plots look at each variable separately while the assumptions for MANOVA involve the joint distribution of the variables. As a practical matter, if the assumptions are met for the variables singly, things look good for the multivariate assumptions, but if the assumptions fail for the single variables, they should fail for the multivariate situation as well.

Click **File..Open..Data**

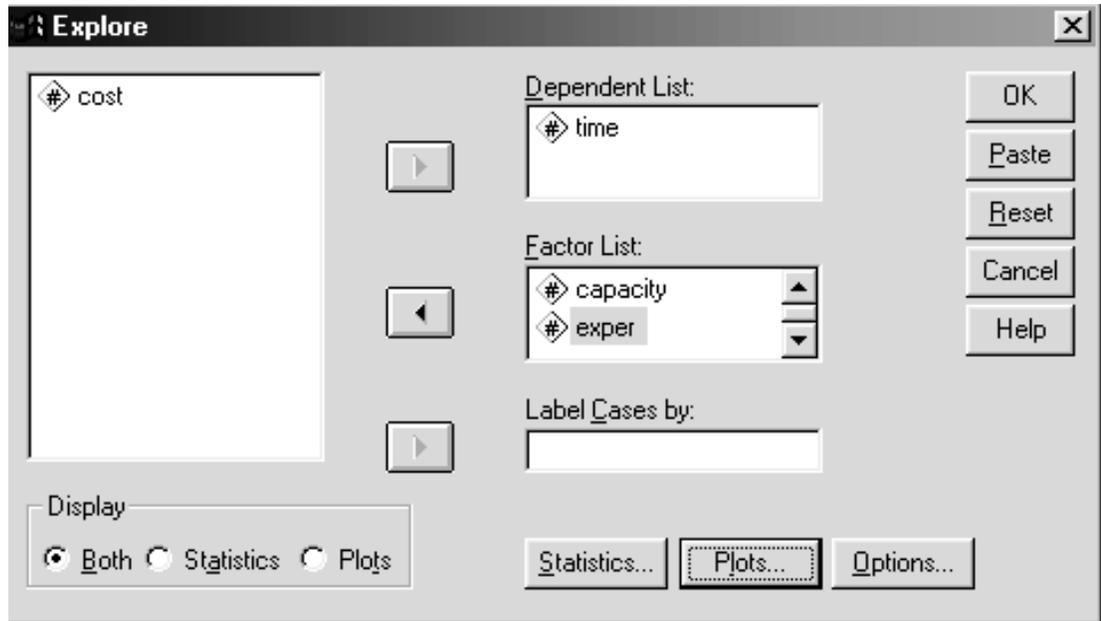
Move to the **c:\Train\Anova** directory (if necessary)

Select **SPSS Portable (\*.por)** from the Files of Type drop-down list

Double-click on **plant.por**

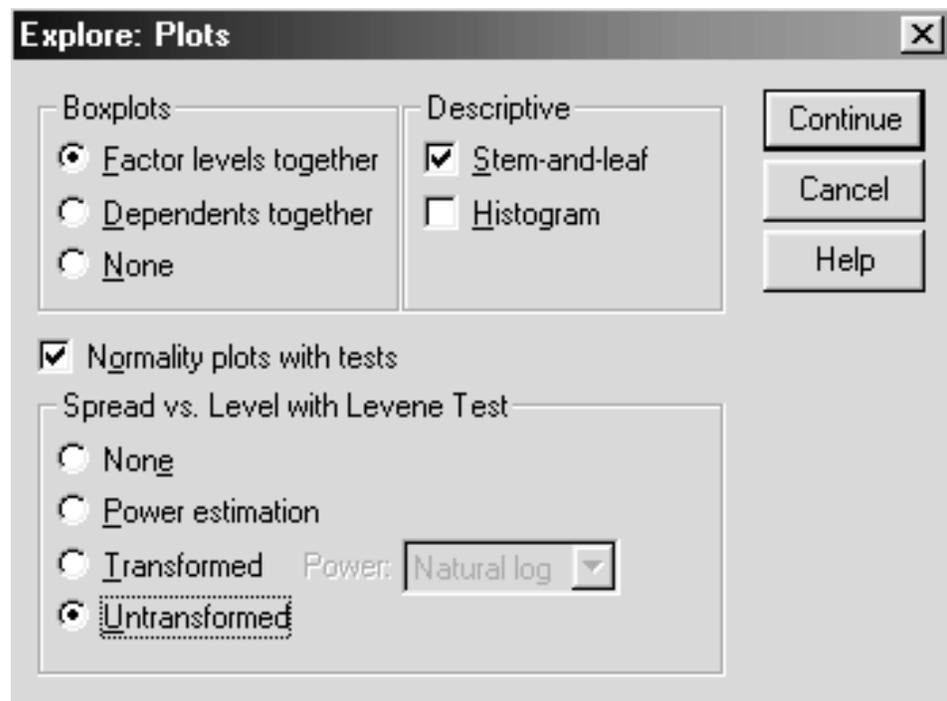
Click on **Analyze..Descriptive Statistics..Explore**  
Move **time** into the **Dependent List** box  
Move **capacity** and **exper** into the **Factor List** box

Figure 5.1 Explore Dialog Box



Click the **Plots** pushbutton  
Click the **Normality plots with tests** checkbox  
Click the **Untransformed** option button in the **Spread vs. Level with Levene Test** area

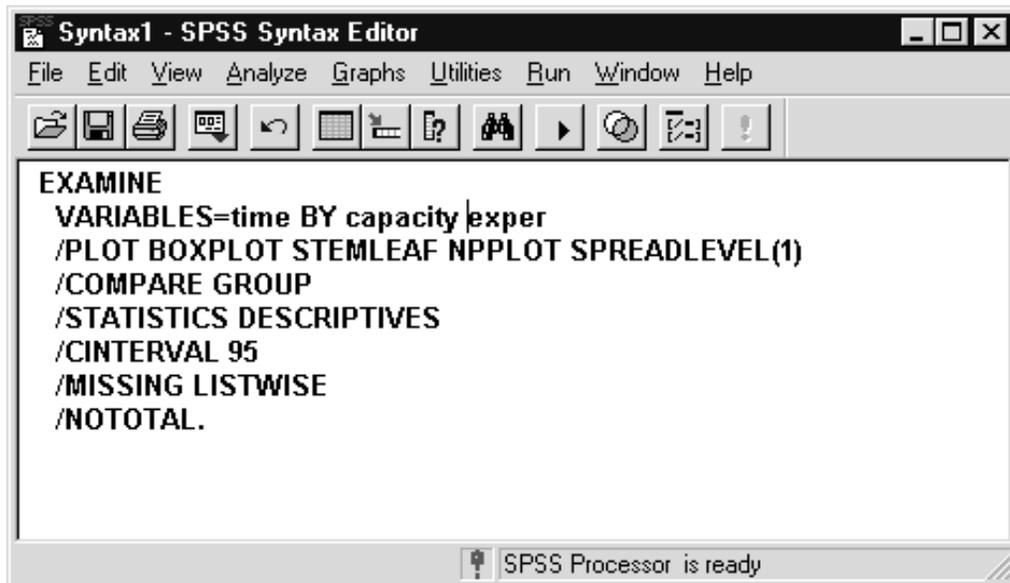
Figure 5.2 Plots Dialog Box



Click **Continue**.

Click **Paste** to paste the syntax into a Syntax Editor window

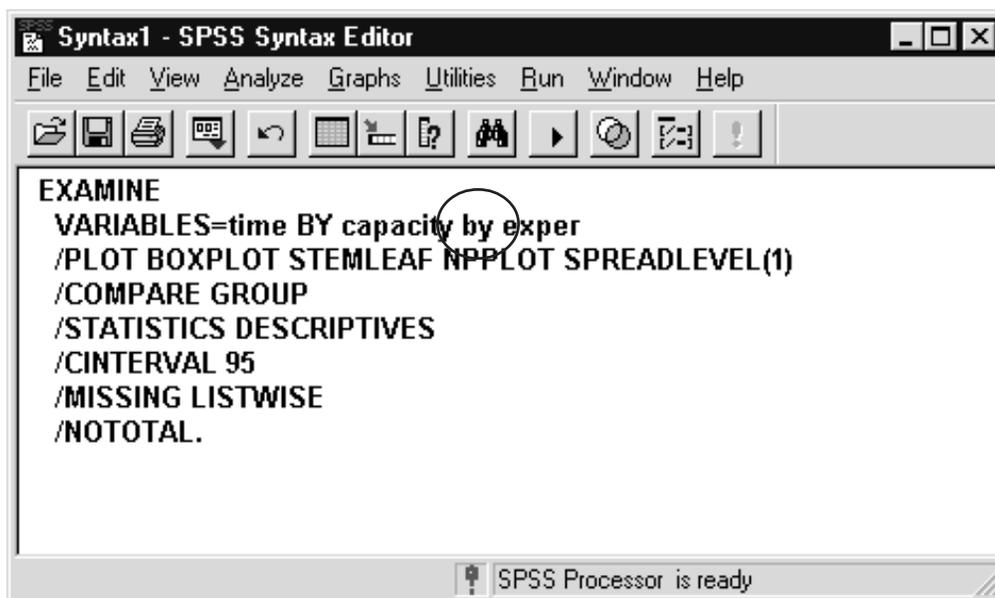
Figure 5.3 Syntax Editor Before Change



Since we want to analyze the results for time in all the combinations of capacity and experience we must insert the keyword **BY** between capacity and experience in the Examine syntax command.

Type **BY** between capacity and exper in the Examine syntax command

Figure 5.4 Syntax Editor After Change



Click **Run..Current** to run the Examine command

We show the normal plots for the time variable below, noting that in most groups there are too few observations to perform a normality test, but in the few cases that tests of normality could be made, the data was consistent with it.

**Figure 5.4A Tests of Normality**

			Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
CAPACITY	EXPER		Statistic	df	Sig.	Statistic	df	Sig.
TIME < 800 MWe	1-3 PLANTS		.214	7	.200*	.930	7	.523
	4-9 PLANTS		.240	3	.			
	10 OR MORE PLANTS		.314	3	.			
800-1000 MWe	4-9 PLANTS		.227	3	.			
	10 OR MORE PLANTS		.186	7	.200*	.955	7	.749
> 1000 MWe	1-3 PLANTS		.361	3	.			
	4-9 PLANTS		.260	2	.			
	10 OR MORE PLANTS		.358	3	.			

\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

b. TIME is constant when CAPACITY = 800-1000 MWe, EXPER = 1-3 PLANTS. It has been omitted.

**Figure 5.5 Q-Q Plot of <800 MWe and 1-3 Plants**

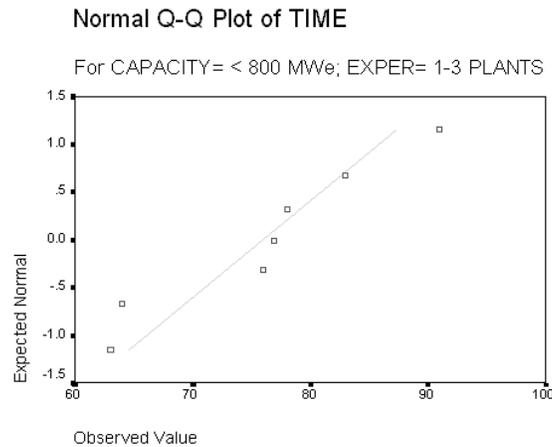


Figure 5.6 Q-Q Plot of <800 MWe and 4-9 Plants

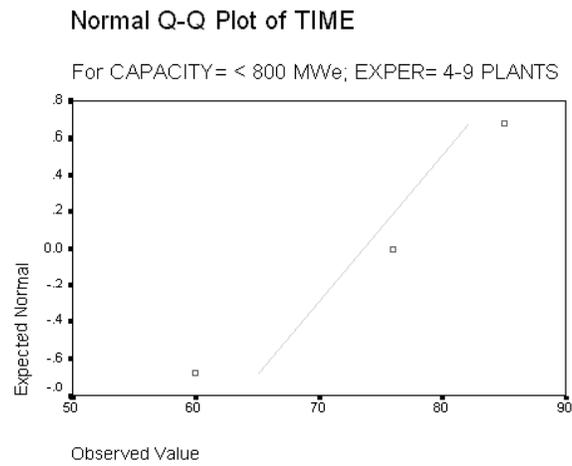


Figure 5.7 Q-Q Plot of <800 MWe and 10 or more Plants

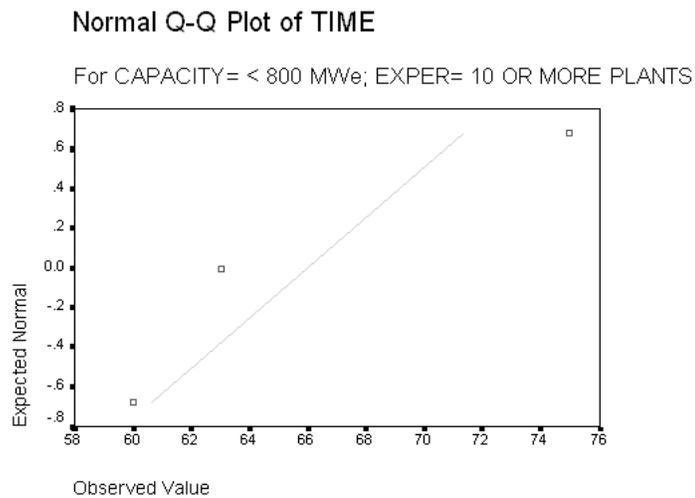
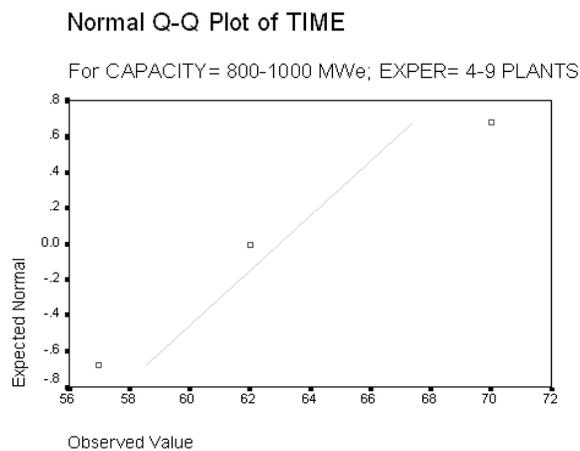


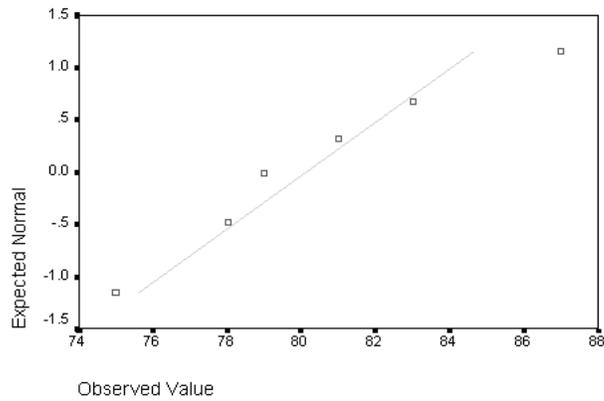
Figure 5.8 Q-Q Plot of 800-1000 MWe and 4-9 Plants



**Figure 5.9 Q-Q Plot of 800-1000 MWe and 10 or more Plants**

Normal Q-Q Plot of TIME

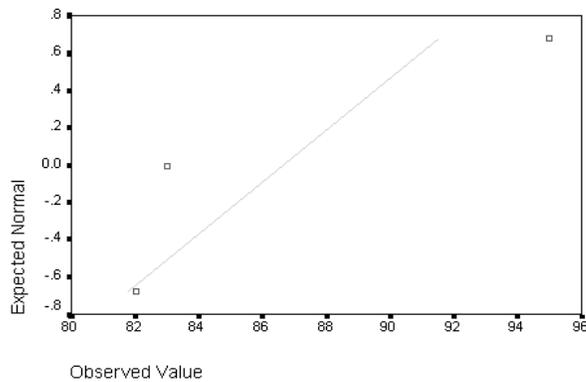
For CAPACITY= 800-1000 MWe; EXPER= 10 OR MORE PLANTS



**Figure 5.10 Q-Q Plot of >1000 MWe and 1-3 Plants**

Normal Q-Q Plot of TIME

For CAPACITY= > 1000 MWe; EXPER= 1-3 PLANTS



**Figure 5.11 Q-Q Plot of >1000 MWe and 4-9 Plants**

Normal Q-Q Plot of TIME

For CAPACITY= > 1000 MWe; EXPER= 4-9 PLANTS

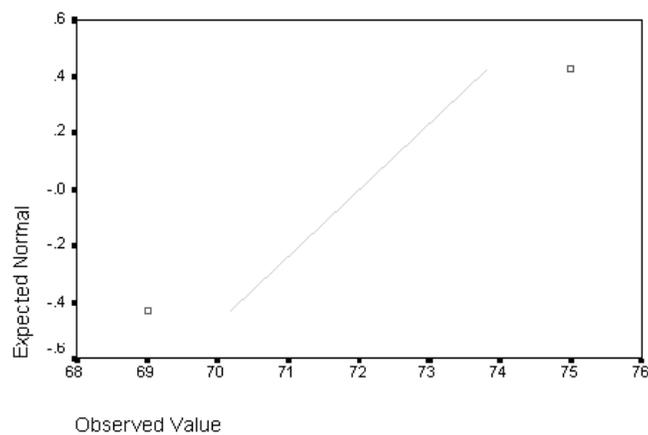
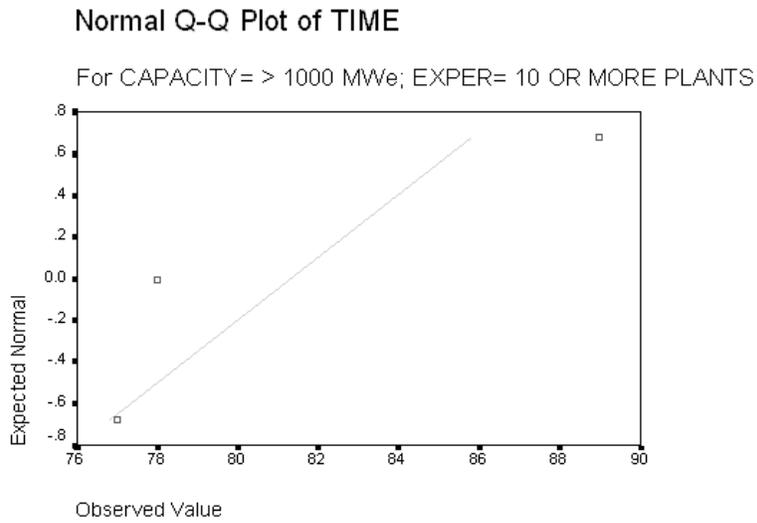
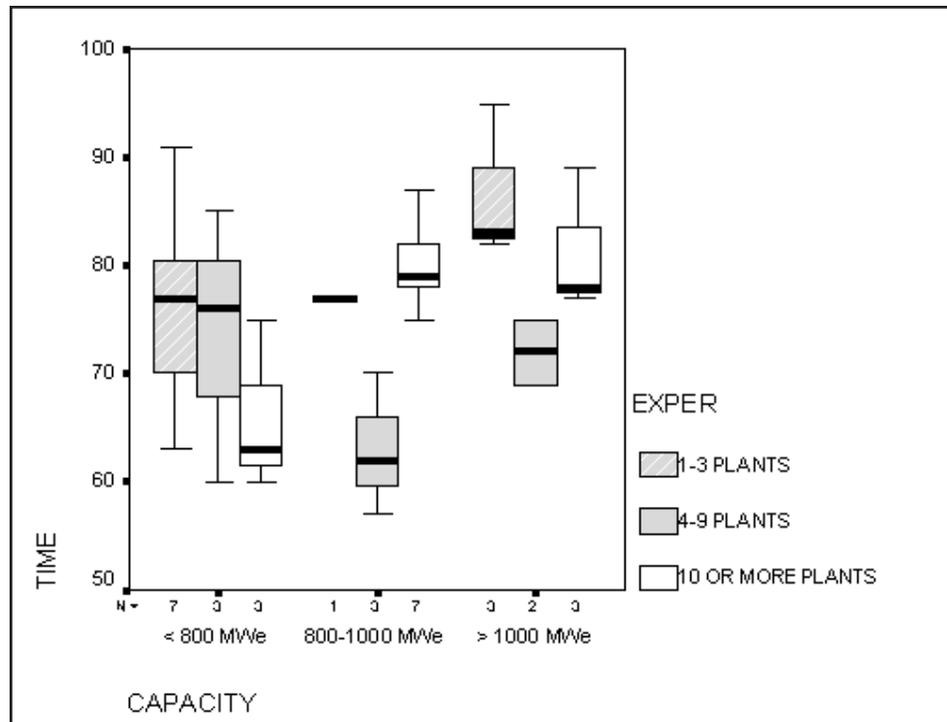


Figure 5.12 Q-Q Plot of >1000 MWe and 10 or more Plants



Next we see the Box and Whisker plot for time.

Figure 5.13 Box and Whisker Plot for Time



There seems to be a fair amount of variation among the groups in time taken to license the plant. It looks as if there is more spread for groups with less experience than there is for those with more experience.

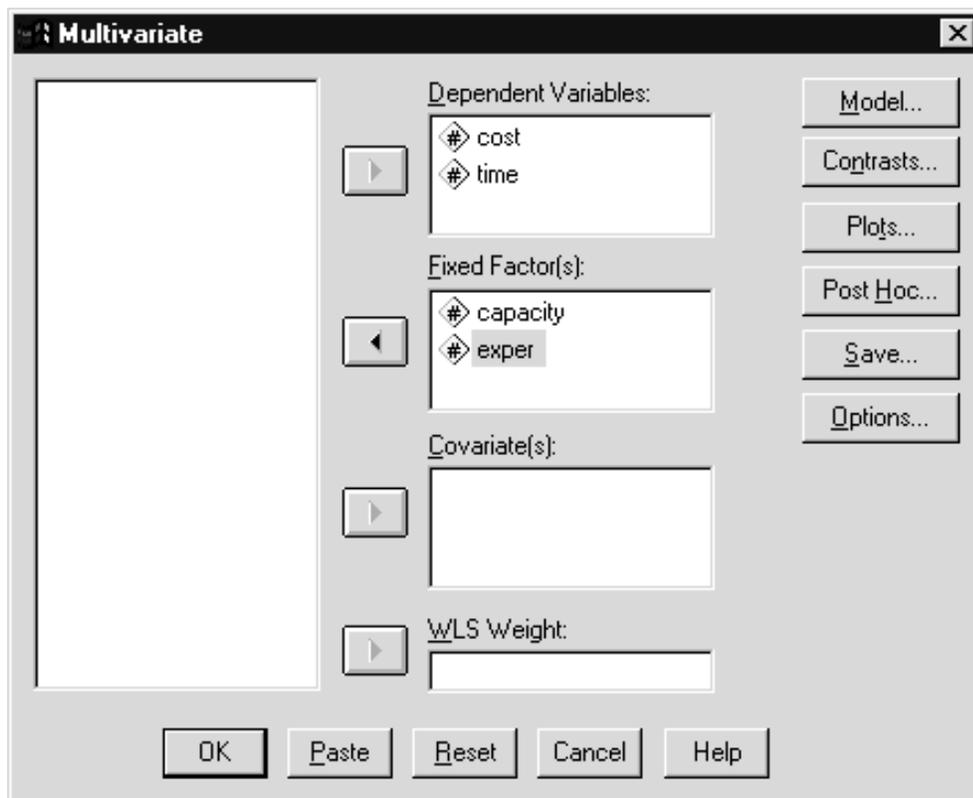
**THE  
MULTIVARIATE  
ANALYSIS**

The Advanced Models module within SPSS adds several General Linear Model (GLM) procedures (multivariate (GLM) and repeated measures (GLM)) to Univariate within the SPSS Base system. These procedures have several desirable features from the perspective of MANOVA: 1) Post hoc tests on marginal means (univariate only), 2) Type I through Type IV sums of squares available (greater flexibility in handling unbalanced designs/ missing cells), 3) Multiple Random Effect models can be easily specified, and 4) Residuals, predicted values and influence measures can be saved as new variables to the data set. However, the MANOVA procedure (which was the original procedure within SPSS performing MANOVA, and it is still available through syntax) contains several useful advanced functions. Within the MANOVA procedure are: 1) Roy-Bargmann step-down tests (testing for mean differences on a single dependent measure while controlling for the other dependent measures), and 2) Dimension reduction analysis and discriminant coefficients. These latter functions provide information as to how the dependent variables interrelate within the context of group differences (for a single main-effect analysis, this is equivalent to a discriminant analysis).

In short, while we expect the SPSS General Linear Model procedure will be your first choice for multivariate analysis of variance, the MANOVA procedure can contribute additional information. (Please note, MANOVA can only be run from syntax.)

Click **Analyze..General Linear Model..Multivariate**  
Move **cost** and **time** into the **Dependent Variables** list box  
Move **capacity** and **exper** into the **Fixed Factors** list box

Figure 5.14 Multivariate Dialog Box



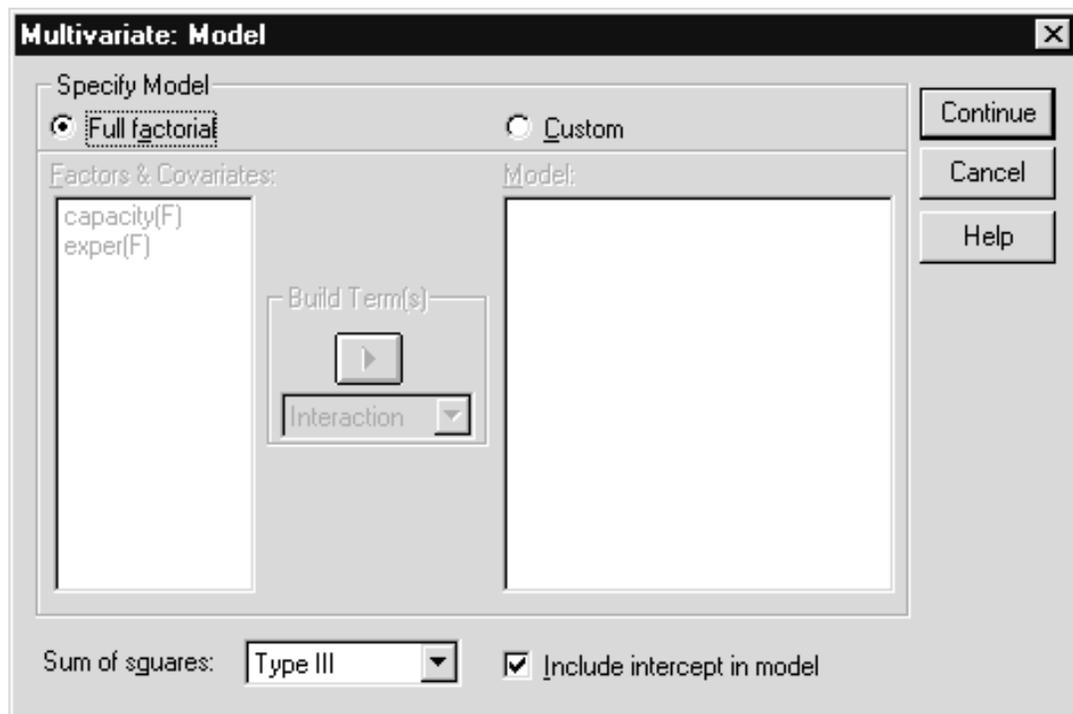
We must specify the dependent measure(s) and at least one factor. The dialog box for Multivariate contains list boxes for the dependent variables, factors and covariates. The term “Fixed Factor(s)” in the Multivariate dialog box reminds us that the factors are assumed to be fixed, that is, levels of the factor(s) used in the analysis were chosen by the researcher (not randomly sampled) and cover the range to which population conclusions will be drawn. The Multivariate dialog box also permits a weight variable to be incorporated in the analysis (performs weighted least squares). Although rarely used in multivariate analyses (when used it is typically for univariate analysis), it adjusts the analysis based on different levels of precision (or heterogeneity of variance) for different individuals or groups.

The Multivariate dialog box contains several pushbuttons. The Plots pushbutton produces for each dependent measure a profile plot displaying group means. The Post Hoc pushbutton performs post hoc tests on the marginal means (for multivariate analyses, each dependent variable is analyzed separately). The Contrasts pushbutton performs any planned contrasts that the researcher wants to conduct; while the Options pushbutton controls many options for the analysis. Finally the Save pushbutton permits you to save predicted values, residuals, and influence measures for later examination.

We could run the analysis at this point, but will examine the dialog boxes within Multivariate and request some additional options.

Click **Model** pushbutton

Figure 5.15 Multivariate: Model Dialog Box

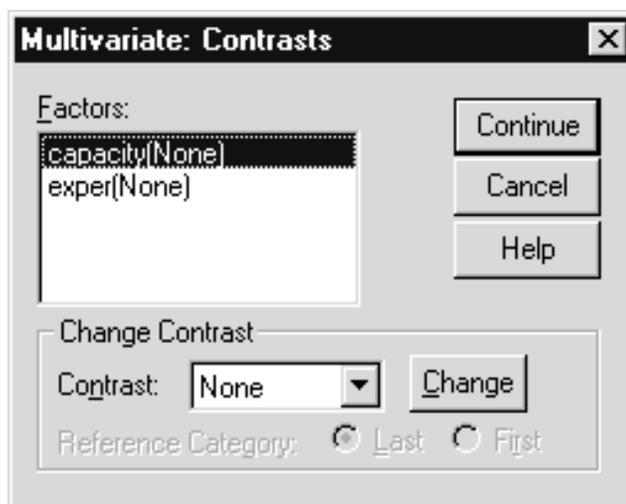


For most analyses the Model dialog box is not used. This is because by default a full factorial model (all main effects, interactions, covariates)

is fit and the various effects tested using Type III sums of squares (each effect is tested after statistically adjusting for all other effects in the model). If there are any missing cells in your analysis, you might switch to Type IV sums of squares, which better adjusts for missing cells. If you are running specialized factorial designs that are incomplete (by plan every possible combination of factor levels is not present), or in which there are no replicates (interaction effects are used as error terms), you would click the Custom option button in the Specify Model area and indicate which main effects and interactions to be included in the model. A custom model is sometimes used if there is no interest in testing high order interaction effects. Since we are interested in both main effects and the one interaction there is no need to modify this dialog box.

Click **Cancel** to exit the Model dialog box  
Click **Contrasts** pushbutton

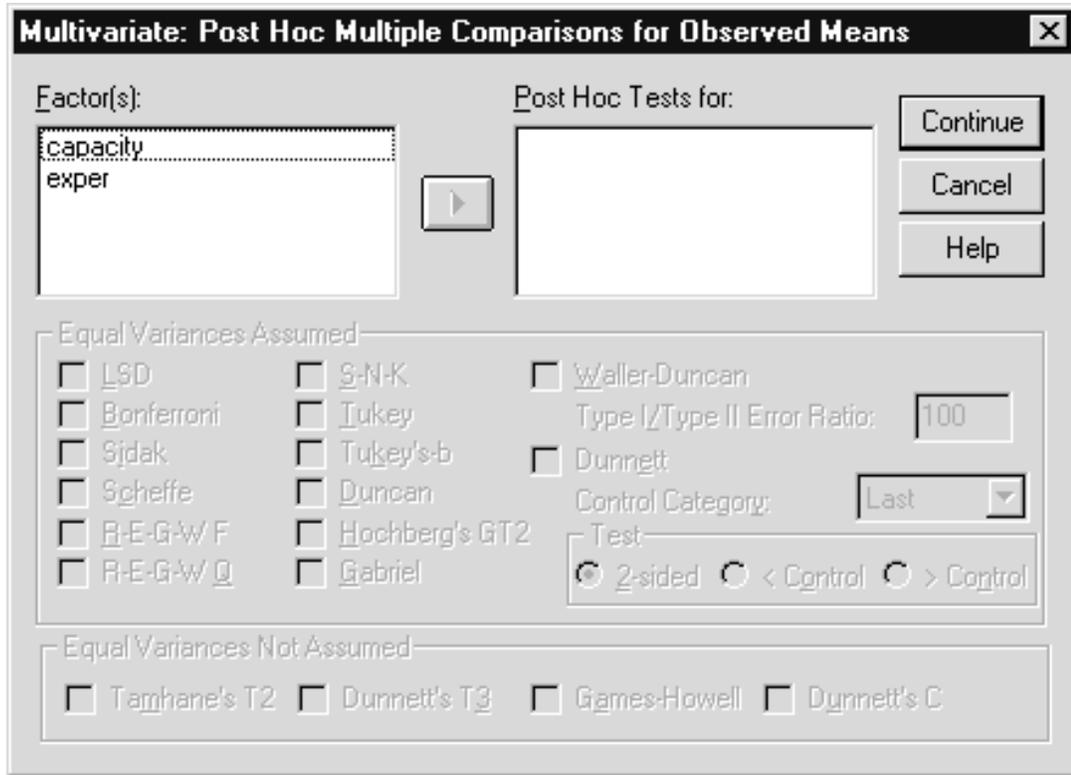
Figure 5.16 Multivariate: Contrasts Dialog Box



The Contrasts dialog box is identical for multivariate and univariate analyses. You would use it to specify main effect group comparisons of interest, for which parameter estimates can be displayed and tests performed. In statistical literature, these contrasts are sometimes called planned comparisons. For example, in an experiment in which there are three treatment groups and a control group there is a very specific interest in testing each experimental group against the control. One of the contrast choices (Simple) allows this. Several types of contrasts are available within the dialog box and using syntax you can specify your own (Special). To request a set of contrasts, select the factor from the Factor(s) list box, select the desired contrast from the Contrast drop-down list, and click the Change pushbutton. Since we have no specific planned contrasts that we wished to apply to the main effects, we will exit the Contrast dialog box.

Click **Cancel** to exit the Contrasts dialog box  
Click **Post Hoc** pushbutton

Figure 5.17 Multivariate: Post Hoc Dialog Box

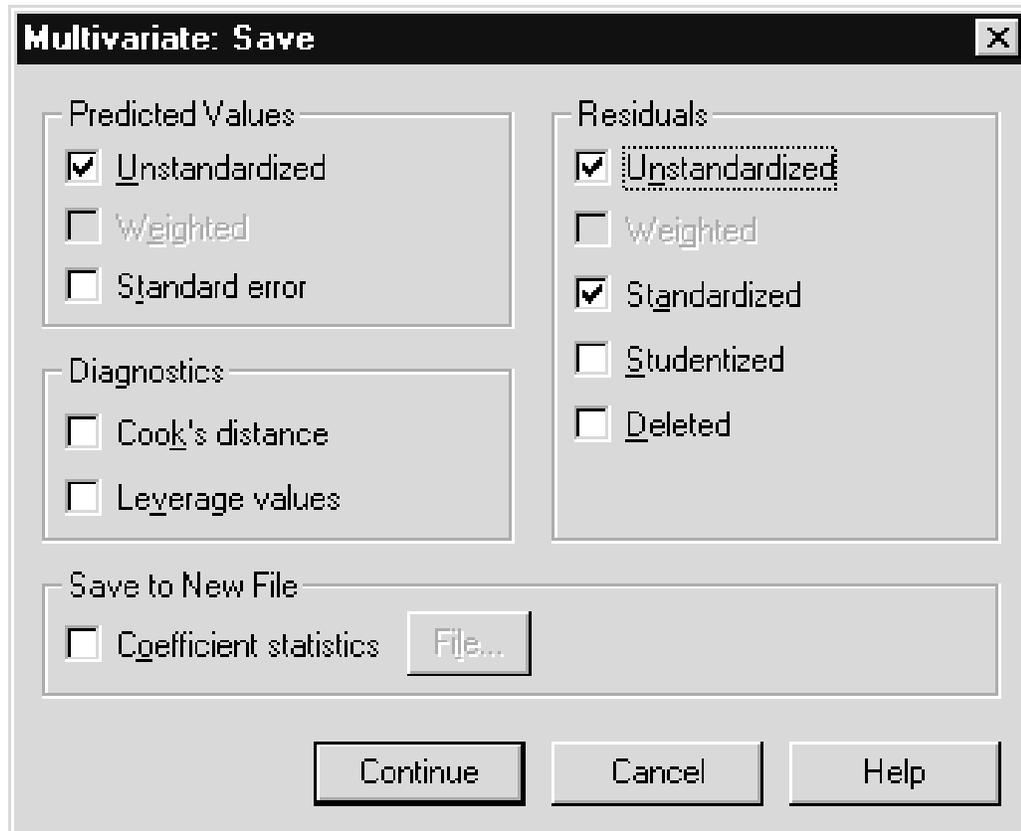


The Post Hoc dialog box is used to request post hoc comparisons on the observed subgroup means. Post hocs test for significant differences between every possible pairing of levels of a factor. Since many tests may be involved, most post hocs adjust the significance criterion based on the number of tests in order to control the false positive error rate (Type I error). Usually post hocs are performed after a significant main effect is found (in the initial analysis), and we will visit this dialog box later in this chapter.

Click **Cancel** pushbutton to exit the Post Hoc dialog box  
 Click **Save** pushbutton

Click the **Unstandardized Predicted Values**,  
**Unstandardized Residual**, and **Standardized Residual**  
 check boxes

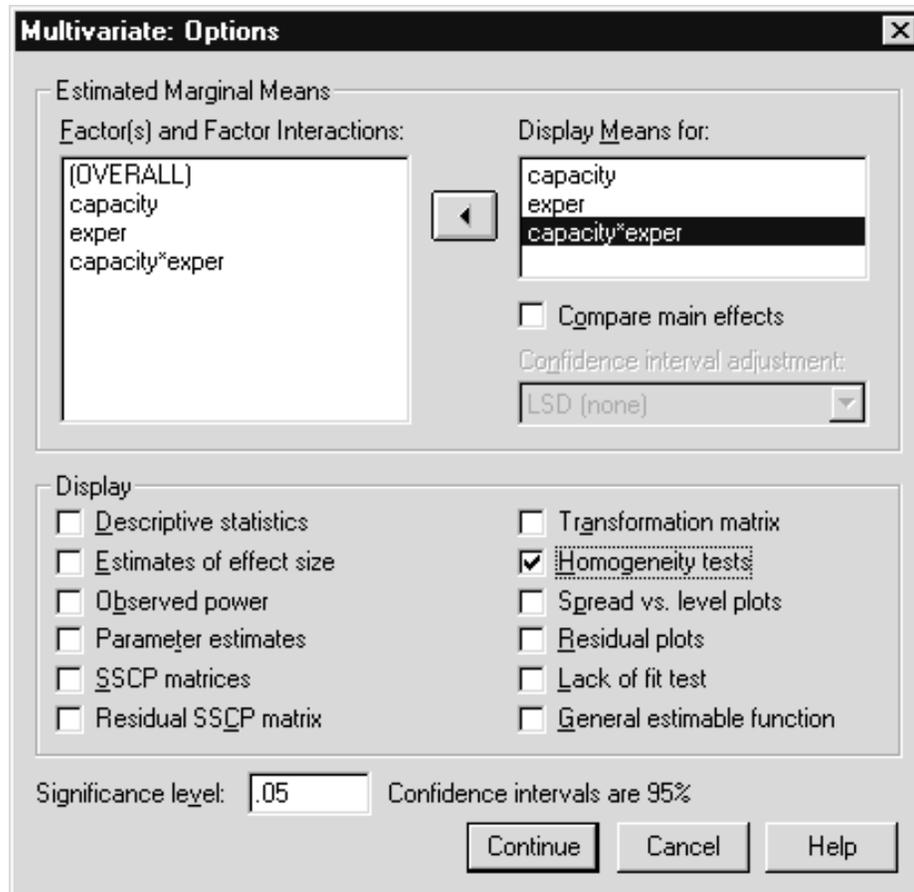
Figure 5.18 Multivariate: Save Dialog Box



The Save dialog box allows you to save predicted values, and various types of residuals and influence measures as new variables in the data file. Examining them might identify outliers and influential data points (data points whose exclusion substantially effects the analysis). Such analyses are standard for serious practitioners of regression and can be applied in this context. In addition, the coefficient statistics (coefficient estimates, standard errors, etc.) can be saved to an SPSS data file (in matrix format) and manipulated later (for example, apply the coefficients to generate predictions for future cases). Although we strongly recommend an examination of the residuals, with the limited amount of time available in this class, we will skip this step.

- Click **Continue** to process the residual requests
- Click **Options** pushbutton
- Select **capacity**, **exper**, and the **capacity\*exper** interaction, and move them into the **Display Means for** list box
- Click the **Homogeneity tests** check box in the Display area.

Figure 5.19 Multivariate: Options Dialog Box



Click **Continue** to process our option requests.  
 Click **OK** to run the analysis.

Moving these factor variables and their interaction term into the Display Means for list box will result in estimated means, predicted from the chosen model, appearing for the subgroups. These means can differ from the observed means if covariates are specified or if an incomplete model (one not containing all main effects and interactions) is used. If no covariates are included (our situation), then post hoc analyses can be applied to the observed marginal means using the Post Hoc pushbutton. The Compare main effects' checkbox can be used to have SPSS test for significant differences between every pair of estimated marginal means for each of the main effects in the Display Means for list box. Note that by default, a significance level of .05 (see Significance level text box) is applied to each test. Also notice the confidence intervals for the mean differences have no adjustment (LSD (none)) based on the number of tests made, although Bonferroni and Sidak adjustments can be requested.

In the Display area, we requested that homogeneity of variance tests be performed. The Display choices allow you to view supplemental information. Checking Descriptive Statistics will display means, standard

deviations, and counts for each cell (subgroup) in the analysis. If effect size is checked, then partial eta-square values will be presented for each effect (main effects, interactions). Eta-square is equivalent to the r-square in regression; the partial eta-square measures the proportion of variation in the dependent measure that can be attributed to each effect in the model after adjusting for the other effects. Parameter estimates are the estimates for the coefficients in the model. Typically, they would be requested if you wanted to construct a prediction equation. The various sums of square matrices are computational summaries and not interpreted directly.

The Significance level text box allows you to specify the significance level used to test for differences in the estimated marginal means (default .05), and the confidence intervals around parameter estimates (default .95).

We are now ready to proceed. The SPSS command below will run the analysis.

```
GLM
  cost time BY capacity exper
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /SAVE = PRED RESID ZRESID
  /EMMEANS = TABLES(capacity)
  /EMMEANS = TABLES(exper)
  EMMEANS = TABLES(capacity*exper)
  /PRINT = HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /DESIGN = capacity exper capacity*exper.
```

In the GLM command the dependent variables (cost, time) precede the BY keyword while the factor variables (capacity, exper) follow it. Type III (each effect is evaluated after adjusting for all other effects) sums of squares is requested (the default). The Emmeans subcommand will print a table of estimated marginal means for the factor variables. Homogeneity tests are obtained from the print subcommand and the alpha value (used for confidence intervals and significance tests of the estimated marginal means) is set to .05 (default). The Design subcommand is used to specify the model to be applied to the data; if nothing were specified, a full factorial model would be fit.

## EXAMINING THE RESULTS

The first piece of Multivariate output describes the factors involved in the analysis. They are labeled between-subject factors; this is appropriate because the three capacity groups and the three experience groups were composed of different plants. We will see within-subject analysis of variance (repeated measures) in a later chapter.

Figure 5.20 Between-Subject Factor Summary

Between-Subjects Factors			
		Value Label	N
CAPACITY	1.00	< 800 MWe	13
	2.00	800-1000 MWe	11
	3.00	> 1000 MWe	8
EXPER	1.00	1-3 PLANTS	11
	2.00	4-9 PLANTS	8
	3.00	10 OR MORE PLANTS	13

The next two pivot tables provide information about the homogeneity of variance assumption. Box's M tests for equality of covariance matrices (since there is more than a single dependent measure) across the different subgroups. Levene's homogeneity test is a univariate test and is applied separately to each dependent variable.

Figure 5.21 Box's Test of Equality of Covariance Matrices

Box's Test of Equality of Covariance Matrices <sup>a</sup>	
Box's M	33.460
F	.931
df1	18
df2	416
Sig.	.541

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

a. Design: Intercept+CAPACITY+EXPER+CAPACITY \* EXPER

Figure 5.22 Levene's Test of Equality of Error Variances

	F	df1	df2	Sig.
COST	3.184	8	23	.014
TIME	1.054	8	23	.427

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

a. Design: Intercept+CAPACITY+EXPER+CAPACITY \* EXPER

As mentioned above, Box's M statistic can be used to test for equality of variance-covariance matrices in the population. This generalizes the homogeneity of variance test to a multivariate situation, testing for equality of group variances (as univariate homogeneity tests would) and covariances (which univariate tests cannot) for all dependent measures in one test. Box's test is not significant (.541) indicating no group differences in the covariance matrices made up of the dependent measures. As a univariate statistic, Levene's test is applied to each dependent measure. The time measure is consistent with homogeneity assumption (sig. = .427), while cost measure does show group differences in variance (sig. = .014). Given that the Box's test is not significant, we will proceed to view the multivariate test results.

**WHAT IF  
HOMOGENEITY  
FAILED?**

As with ANOVA, MANOVA is robust under failure of homogeneity if the sample sizes in the cells are large and roughly equal. If the sample sizes are unequal, and larger variances are associated with larger cells, the MANOVA tests are conservative so you can be confident of significant findings. If smaller cells have larger variances, the MANOVA tests are liberal so the Type I error is greater than it should be (see Hakstian, Roed, and Lind (1979)). If variance is related to the mean level of the group, a variance stabilizing transform is a possibility.

**MULTIVARIATE  
TESTS**

There are four multivariate test statistics commonly applied: Pillai's criterion, Hotelling's Trace criterion, Wilk's Lambda, and Roy's largest root. The first three give identical results in a two-group analysis, and then can differ. They all test the null hypothesis of no group mean differences in the population. Results of Monte Carlo simulations focussing on robustness and statistical power, suggest that under general circumstances Pillai's test is preferred. However, there are specific situations, for example when the dependent measures are highly related (forming a strong core), that one of the others is the most powerful test. As a general rule, if different multivariate tests give you markedly different results, it suggests something about the dimensionality and type of group differences. For an accessible discussion of this see Olsen (1976).

The distribution of the first three multivariate statistics follows the generalized “F” distribution. While more complicated than the simple “F”, and having 3 sets of degrees of freedom it yields a probability value just as the regular “F” does. This generalized “F” test assumes a multivariate normal distribution of the errors.

Figure 5.23 Multivariate Analysis of Variance Table

**Multivariate Tests<sup>c</sup>**

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.991	1179.436 <sup>a</sup>	2.000	22.000	.000
	Wilks' Lambda	.009	1179.436 <sup>a</sup>	2.000	22.000	.000
	Hotelling's Trace	107.221	1179.436 <sup>a</sup>	2.000	22.000	.000
	Roy's Largest Root	107.221	1179.436 <sup>a</sup>	2.000	22.000	.000
CAPACITY	Pillai's Trace	.357	2.499	4.000	46.000	.055
	Wilks' Lambda	.644	2.703 <sup>a</sup>	4.000	44.000	.043
	Hotelling's Trace	.550	2.885	4.000	42.000	.034
	Roy's Largest Root	.546	6.275 <sup>b</sup>	2.000	23.000	.007
EXPER	Pillai's Trace	.394	2.824	4.000	46.000	.035
	Wilks' Lambda	.616	3.011 <sup>a</sup>	4.000	44.000	.028
	Hotelling's Trace	.605	3.177	4.000	42.000	.023
	Roy's Largest Root	.575	6.614 <sup>b</sup>	2.000	23.000	.005
CAPACITY * EXPER	Pillai's Trace	.351	1.224	8.000	46.000	.307
	Wilks' Lambda	.664	1.249 <sup>a</sup>	8.000	44.000	.294
	Hotelling's Trace	.483	1.268	8.000	42.000	.286
	Roy's Largest Root	.431	2.476 <sup>b</sup>	4.000	23.000	.073

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+CAPACITY+EXPER+CAPACITY \* EXPER

The upper part of the table tests whether the overall mean (Intercept) differs from zero in the population. It is not interesting since all it shows is that overall, cost and time were not both zero.

The Value column displays the sample value of each of the four multivariate test statistics. They are converted to “F” statistics (“F” column) and the associated hypothesis (Hypothesis df) and error (error df) degrees of freedom follow. These four columns are technical summaries; we are primarily interested in the significance values that appear under the “Sig.” Heading. Here we see that for the capacity factor, three of the four tests show that there are differences in the dependent measures (significance values of .055, .043, .034, .007). Notice the tests are not in agreement if you test at the .05 level. While Pillai’s is often the recommended test, it would be safe to conclude at least there is a marginal effect, perhaps something worth looking at with a larger sample. We also see that for the experience factor, all four tests show that there are group differences in the means (significance values of .035, .028, .023, and .005). The test of an interaction between capacity and experience was not significant (significance values of .307, .294, .286, and .073). Given these findings, we are next interested in looking at whether both cost and time show differences (univariate tests), and knowing which groups differ from which others (post hocs).

Two additional columns can appear in the multivariate (or univariate) analysis of variance table, but do not do so by default. The noncentrality parameter is a technical summary that describes the magnitude of the mean group differences in the form of a parameter for the “F” distribution. It can be used to calculate the appropriate sample size (statistical power analysis) if this study were to be repeated while expecting to find the same group differences. The Observed Power indicates how likely you are to obtain a significant group difference (testing at the .05 level) if the population group means matched the means in the sample. This can be useful in conducting postmortems of your analysis, that is, exploring why you failed to find significant differences.

We now examine the test results for each dependent measure.

Figure 5.24 Univariate Test Results

**Tests of Between-Subjects Effects**

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	COST	379480.902 <sup>a</sup>	8	47435.113	2.107	.077
	TIME	1402.018 <sup>b</sup>	8	175.252	2.821	.024
Intercept	COST	5480467.2	1	5480467.2	243.486	.000
	TIME	132290.879	1	132290.879	2129.457	.000
CAPACITY	COST	154061.360	2	77030.680	3.422	.050
	TIME	315.056	2	157.528	2.536	.101
EXPER	COST	158495.497	2	79247.749	3.521	.046
	TIME	381.596	2	190.798	3.071	.066
CAPACITY * EXPER	COST	49479.036	4	12369.759	.550	.701
	TIME	572.529	4	143.132	2.304	.089
Error	COST	517691.407	23	22508.322		
	TIME	1428.857	23	62.124		
Total	COST	7714385.8	32			
	TIME	186752.000	32			
Corrected Total	COST	897172.309	31			
	TIME	2830.875	31			

a. R Squared = .423 (Adjusted R Squared = .222)

b. R Squared = .495 (Adjusted R Squared = .320)

Although both dependent measures appear in this table the results are calculated independently, and are identical to what you would obtain if separate analyses were run on each dependent measure (univariate ANOVA). Thus we find whether both of the dependent measures showed significant group differences. The sums of squares, df (degrees of freedom), mean square, and “F” columns are what we would expect in an ordinary table. We described and disregarded the Intercept information in the multivariate summary. Moving to the capacity section, we find cost is right on the border of significance (.05) while time is not significant (.101). From the experience summary we find that again cost is significant (.046) while time is not (.066). In the interaction area we find that neither cost nor time are significant (.701 and .089). The Error

section summarizes the within-group variation. The Corrected Model summary pools together all model effects (excluding the intercept), and is equal to the Corrected Total minus the Error Total. Some analysts turn to this overall test first to see if any effects are significant, and then proceed to examine individual effects. However, most researchers move directly to the tests of specific main effects and interactions. The Total summary pools together everything in the analysis (including the error). It should be noted that if the sample sizes are not equal when multiple factors are included in the analysis, then under Type III sums of squares (the default), the sums of squares for the totals will not generally be equal to the sums of their component sums of squares.

Finally, r-square values (based on the corrected model) for each variable appear as footnotes. Notice that the adjusted r-square for time (.320) is higher than that of cost (.222). This is consistent with time having a higher “F” statistic in the corrected model section.

**Figure 5.25 Estimated Marginal Means for Capacity**

**1. CAPACITY**

Dependent Variable	CAPACITY	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
COST	< 800 MWe	399.430	44.995	306.350	492.509
	800-1000 MWe	467.296	60.761	341.604	592.989
	> 1000 MWe	583.203	54.016	471.462	694.944
TIME	< 800 MWe	71.889	2.364	66.999	76.779
	800-1000 MWe	73.381	3.192	66.778	79.984
	> 1000 MWe	80.000	2.838	74.130	85.870

**Figure 5.26 Estimated Marginal Means for Experience**

**2. EXPER**

Dependent Variable	EXPER	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
COST	1-3 PLANTS	535.122	60.761	409.429	660.815
	4-9 PLANTS	369.460	54.016	257.719	481.201
	10 OR MORE PLANTS	545.347	44.995	452.268	638.427
TIME	1-3 PLANTS	79.889	3.192	73.285	86.492
	4-9 PLANTS	69.556	2.838	63.685	75.426
	10 OR MORE PLANTS	75.825	2.364	70.935	80.715

Figure 5.27 Estimated Marginal Means for Capacity by Experience Subgroups

3. CAPACITY \* EXPER

Dependent Variable	CAPACITY	EXPER	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
COST	< 800 MWe	1-3 PLANTS	404.083	56.705	286.779	521.386
		4-9 PLANTS	275.827	86.619	96.643	455.011
		10 OR MORE PLANTS	518.380	86.619	339.196	697.564
	800-1000 MWe	1-3 PLANTS	608.800	150.028	298.444	919.156
		4-9 PLANTS	339.323	86.619	160.139	518.507
		10 OR MORE PLANTS	453.766	56.705	336.462	571.069
	> 1000 MWe	1-3 PLANTS	592.483	86.619	413.299	771.667
		4-9 PLANTS	493.230	106.086	273.775	712.685
		10 OR MORE PLANTS	663.897	86.619	484.713	843.081
TIME	< 800 MWe	1-3 PLANTS	76.000	2.979	69.837	82.163
		4-9 PLANTS	73.667	4.551	64.253	83.080
		10 OR MORE PLANTS	66.000	4.551	56.586	75.414
	800-1000 MWe	1-3 PLANTS	77.000	7.882	60.695	93.305
		4-9 PLANTS	63.000	4.551	53.586	72.414
		10 OR MORE PLANTS	80.143	2.979	73.980	86.306
	> 1000 MWe	1-3 PLANTS	86.667	4.551	77.253	96.080
		4-9 PLANTS	72.000	5.573	60.471	83.529
		10 OR MORE PLANTS	81.333	4.551	71.920	90.747

Estimated marginal means are means estimated for each level of a factor averaging across all levels of other factors (marginals), based on the specified model (estimated). By default, SPSS fits a complete model (all main-effects and interactions), and in such cases these estimated means are identical to the (unweighted) observed means. However, if a partial model were fit (for example, if all main effects were included but higher order interactions were not) then the estimated means will differ from the (unweighted) observed means. We see in the tables above that the average time and cost increase with the plant capacity. Interestingly, regarding experience, time and cost have their lowest means in the middle experience group.

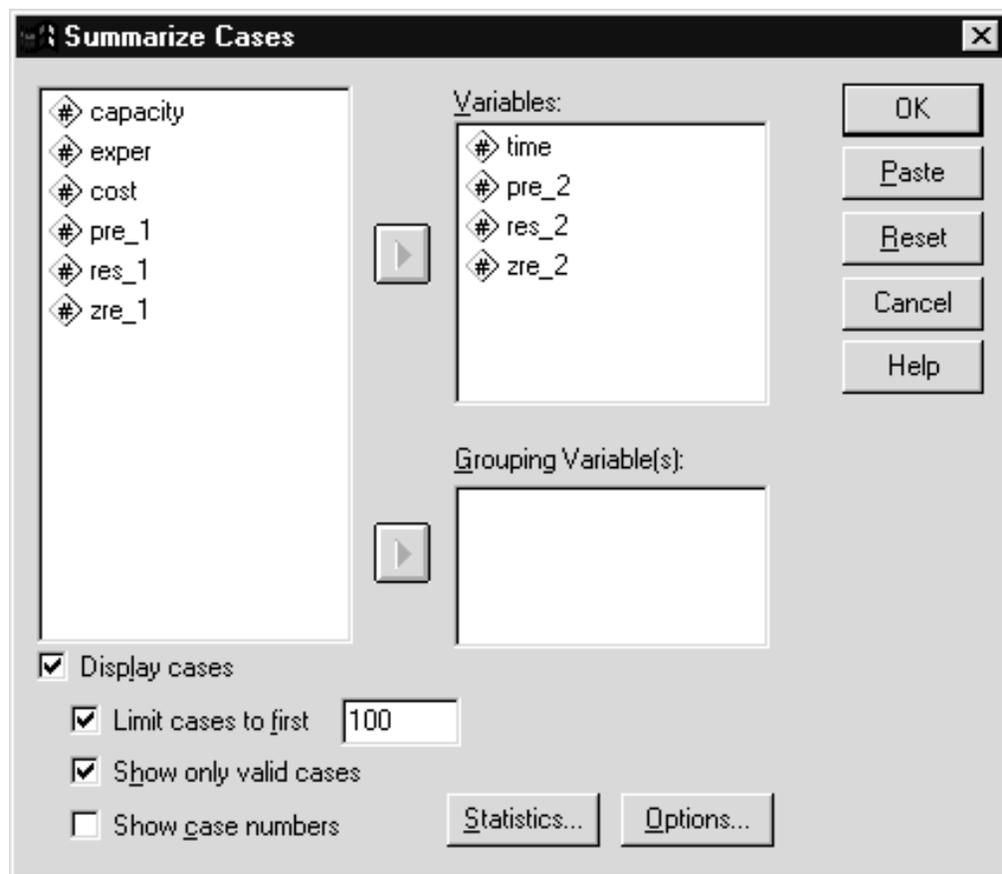
**CHECKING THE RESIDUALS**

We first view the casewise listing of residuals for time. We will skip the listing for cost since it is identical to that seen in Chapter 4, when we ran the same model on cost alone.

Click **Analyze..Reports..Case Summary**

Move **time**, **pre\_2**, **res\_2**, and **zre\_2** into the **Variables** list box.

Figure 5.28 Case Summary Dialog Box



Click on **OK**

The following syntax will also produce the case summary table.

```
SUMMARIZE
  /TABLE=time pre_2 res_2 zre_2
  /FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=100
  /TITLE='Case Summaries' /FOOTNOTE ''
  /MISSING=VARIABLE
  /CELLS=COUNT.
```

Figure 5.29 Casewise Listing of Residuals

**Case Summaries<sup>a</sup>**

	TIME	Predicted Value for TIME	Residual for TIME	Standardized Residual for TIME
1	64.00	76.00	-12.00	-1.52
2	63.00	76.00	-13.00	-1.65
3	78.00	76.00	2.00	.25
4	91.00	76.00	15.00	1.90
5	77.00	76.00	1.00	.13
6	83.00	76.00	7.00	.89
7	76.00	76.00	.00	.00
8	60.00	73.67	-13.67	-1.73
9	85.00	73.67	11.33	1.44
10	76.00	73.67	2.33	.30
11	63.00	66.00	-3.00	-.38
12	75.00	66.00	9.00	1.14
13	60.00	66.00	-6.00	-.76
14	77.00	77.00	.00	.00
15	57.00	63.00	-6.00	-.76
16	62.00	63.00	-1.00	-.13
17	70.00	63.00	7.00	.89
18	78.00	80.14	-2.14	-.27
19	78.00	80.14	-2.14	-.27
20	81.00	80.14	.86	.11
21	79.00	80.14	-1.14	-.14
22	75.00	80.14	-5.14	-.65
23	83.00	80.14	2.86	.36
24	87.00	80.14	6.86	.87

There seem to be no especially large standardized residuals. Once again the predicted values are identical for all members of the same group.

**CONCLUSION**

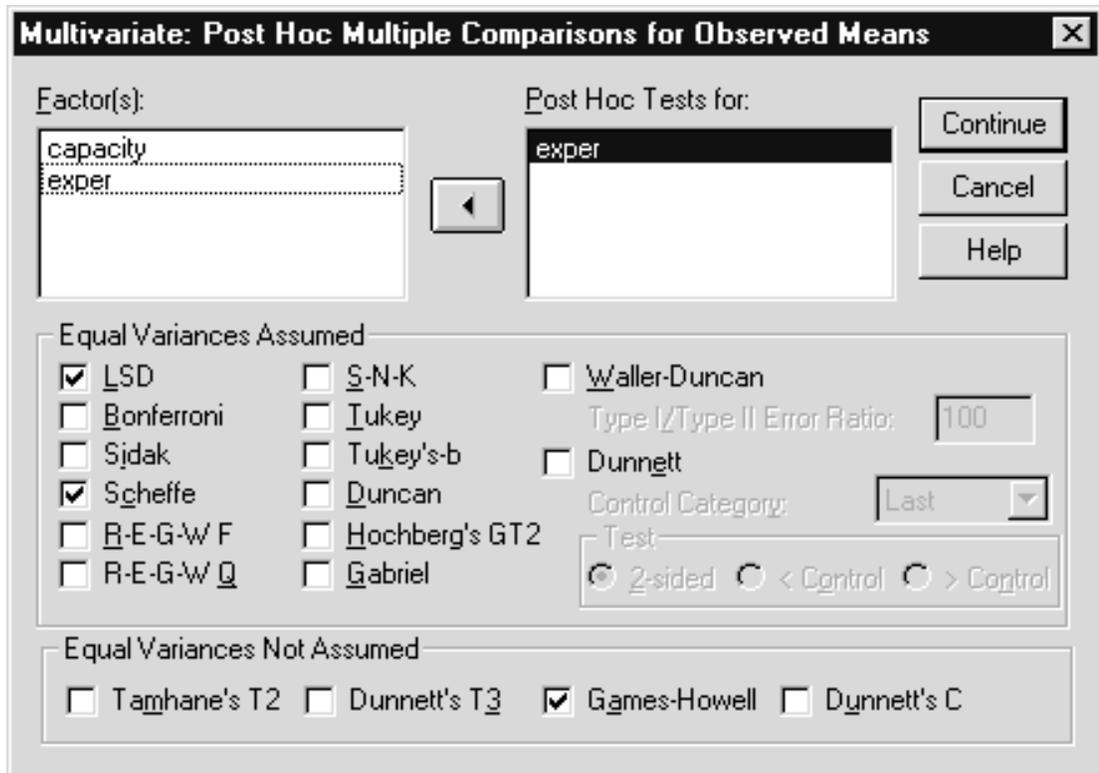
From the multivariate analysis of variance we conclude that the dependent variables show significant mean differences across experience groups, although not in a strictly increasing fashion. There is a modest effect across capacity groups and no sign of an interaction. Of the two measures, cost seems more sensitive to the group differences. What might qualify the result? You could argue that the groupings of experience and capacity levels are arbitrary and different groupings could yield different results. Also, with only 32 observations over a nine-cell design with two dependent measures, we expect very little power to detect differences.

**POST HOC TESTS**

At this point of the analysis it is natural to ask just which groups differ from which others. The GLM procedure in SPSS will perform separate post hoc tests on each dependent variable in order to determine this. Post hoc tests are usually performed to investigate which pairs of levels within a factor differ after an overall (main effect) difference has been established. SPSS offers many post hoc tests and characteristics of them were reviewed in Chapter 3. Recall the basic idea behind post hoc testing is that some adjustment of the Type I (false positive or alpha) error rate must be made due to the number of pairwise comparisons made. In our example, only three tests need to be performed within each factor (group 1 vs. 2, 1 vs. 3, and 2 vs. 3). However, if there were ten levels of experience (or capacity or both), then there would be  $[10*9]/2$  or 45 pairwise tests, and the probability of one or more false positive results would be quite substantial. We asked for the following types of post hoc tests to be performed: LSD (the most liberal), Scheffe (the most conservative), and the Games-Howell (does not assume equal variances—recall the Levene test indicated there might be a homogeneity of variance problem with cost). Although both experience and capacity were found significant, below we request post hocs only for experience. In practice you would view post hoc results for each significant main effect.

- Click the Dialog Recall tool  , then select **Multivariate**
- Click **Post Hoc** pushbutton
- Move **exper** into the **Post Hoc Tests for** list box
- Click **LSD**, **Scheffe**, and **Games-Howell** checkboxes

Figure 5.30 Post Hoc Dialog Box



Click **Continue** to process the post hoc requests  
 Click **OK** to run

The SPSS syntax below will produce the post hoc analysis.

```
GLM
cost time BY capacity exper
/METHOD = SSTYPE(3)
/INTERCEPT = INCLUDE
/SAVE = PRED RESID ZRESID
/POSTHOC = exper ( SCHEFFE LSD GH)
/EMMEANS = TABLES(capacity)
/EMMEANS = TABLES(exper)
EMMEANS = TABLES(capacity*exper)
/PRINT = HOMOGENEITY
/CRITERIA = ALPHA(.05)
/DESIGN = capacity exper capacity*exper.
```

The Posthoc subcommand instructs GLM to apply Scheffe, LSD and Games-Howell (GH) multiple comparison tests to the experience (exper) factor.

Although both multiple comparison and homogeneous subset tables will be produced, we present only the former. Also note that for ease of reading, the post hoc results, which appear in a single pivot table, are displayed below as three figures (within the pivot table editor, the pivot tray window was opened and the post hoc test (Test) icon was moved into the layer dimension).

Figure 5.31 LSD Post Hoc Test for Experience

multiple comparisons

Test		LSD					
Dependent Variable	(I) EXPER	(J) EXPER	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
COST	1-3 PLANTS	4-9 PLANTS	120.0867	69.712	.098	-24.1234	264.2968
		10 OR MORE PLANTS	-43.0930	61.462	.490	-170.2376	84.0516
	4-9 PLANTS	1-3 PLANTS	-120.0867	69.712	.098	-264.2968	24.1234
		10 OR MORE PLANTS	-183.1797*	67.416	.024	-302.6408	-23.7186
	10 OR MORE PLANTS	1-3 PLANTS	43.0930	61.462	.490	-84.0516	170.2376
		4-9 PLANTS	183.1797*	67.416	.024	23.7186	302.6408
TIME	1-3 PLANTS	4-9 PLANTS	9.7500*	3.662	.014	2.1737	17.3263
		10 OR MORE PLANTS	1.8462	3.229	.573	-4.8335	8.5258
	4-9 PLANTS	1-3 PLANTS	-9.7500*	3.662	.014	-17.3263	-2.1737
		10 OR MORE PLANTS	-7.9038*	3.542	.036	-15.2306	-.5771
	10 OR MORE PLANTS	1-3 PLANTS	-1.8462	3.229	.573	-8.5258	4.8335
		4-9 PLANTS	7.9038*	3.542	.036	.5771	15.2306

Based on observed means.

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

**Note** The multiple comparison table was edited in the Pivot Table Editor and the tests (TEST icon) were placed in the layer dimension (see Chapter 4 for instructions) so we can separately view the results from each post hoc.

Figure 5.32 Scheffe Post Hoc Test for Experience

**Multiple Comparisons**

Test		Scheffe					
Dependent Variable	(I) EXPER	(J) EXPER	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
COST	1-3 PLANTS	4-9 PLANTS	120.0867	69.712	.248	-62.2904	302.4639
		10 OR MORE PLANTS	-43.0930	61.462	.784	-203.8880	117.7020
	4-9 PLANTS	1-3 PLANTS	-120.0867	69.712	.248	-302.4639	62.2904
		10 OR MORE PLANTS	-163.1797	67.416	.074	-339.5510	13.1916
	10 OR MORE PLANTS	1-3 PLANTS	43.0930	61.462	.784	-117.7020	203.8880
		4-9 PLANTS	163.1797	67.416	.074	-13.1916	339.5510
TIME	1-3 PLANTS	4-9 PLANTS	9.7500*	3.662	.046	-.1686	19.3314
		10 OR MORE PLANTS	1.8462	3.229	.850	-6.6014	10.2937
	4-9 PLANTS	1-3 PLANTS	-9.7500*	3.662	.046	-19.3314	-.1686
		10 OR MORE PLANTS	-7.9038	3.542	.105	-17.1697	1.3620
	10 OR MORE PLANTS	1-3 PLANTS	-1.8462	3.229	.850	-10.2937	6.6014
		4-9 PLANTS	7.9038	3.542	.105	-1.3620	17.1697

Based on observed means.

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

Figure 5.33 Games-Howell Post Hoc Test for Experience

**Multiple Comparisons**

Test		Games-Howell					
Dependent Variable	(I) EXPER	(J) EXPER	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
COST	1-3 PLANTS	4-9 PLANTS	120.0867	69.712	.224	-58.2010	298.3744
		10 OR MORE PLANTS	-43.0930	61.462	.820	-223.5640	137.3780
	4-9 PLANTS	1-3 PLANTS	-120.0867	69.712	.224	-298.3744	58.2010
		10 OR MORE PLANTS	-163.1797*	67.416	.046	-323.5328	-2.8267
	10 OR MORE PLANTS	1-3 PLANTS	43.0930	61.462	.820	-137.3780	223.5640
		4-9 PLANTS	163.1797*	67.416	.046	2.8267	323.5328
TIME	1-3 PLANTS	4-9 PLANTS	9.7500	3.662	.101	-1.6887	21.1887
		10 OR MORE PLANTS	1.8462	3.229	.872	-7.5124	11.2047
	4-9 PLANTS	1-3 PLANTS	-9.7500	3.662	.101	-21.1887	1.6887
		10 OR MORE PLANTS	-7.9038	3.542	.158	-18.4635	2.6558
	10 OR MORE PLANTS	1-3 PLANTS	-1.8462	3.229	.872	-11.2047	7.5124
		4-9 PLANTS	7.9038	3.542	.158	-2.6558	18.4635

Based on observed means.

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

We can see that for both cost and time, every possible group pairing appears for the factor. The “Mean Difference” column contains the difference in sample means between the two groups, and the “Standard Error” column contains the standard error of the difference between the means. The “Sig” column contains the significance value when the particular test is applied to the group differences. These post hoc test results provide detail concerning significant main effects.

Not surprisingly, the Scheffe results show fewer significant group differences than LSD. Notice that there are no group differences on time using the Games-Howell tests, although both Scheffe and LSD show differences. This is probably due to the Games-Howell being less powerful

when homogeneity of variance holds, as it does for time.

### **SUMMARY**

In this chapter we discussed multivariate analysis of variance and applied it to the plant data. We examined residuals from the analysis and performed post hoc tests.



# Chapter 6 Within-Subject Designs: Repeated Measures

- Objectives** The objective of this chapter is to understand the distinguishing characteristics, assumptions, and methods of approaching within-subject (repeated measures) ANOVA, and to see how SPSS implements such analyses. We will discuss the univariate and multivariate approaches to the repeated measures analysis.
- Method** We discuss the logic and assumptions of repeated measures and use the Explore procedure to examine the data. We then use the GLM Repeated Measures procedure to run a repeated-measures ANOVA with a single within-subject factor. Pairwise comparisons are run and planned comparisons are set up.
- Data** The data set contains vocabulary test scores obtained from the same children over four years (grades 8 through 11). The sex of each child is also recorded, but not used in this analysis.

## INTRODUCTION

In this chapter, we discuss yet another species of ANOVA, the special case where each subject (or unit of analysis) appears in several conditions. We will see that this repeated measurement feature requires some additional assumptions and a more complicated approach to computing error terms. The variation within each group, our constant companion to this point, must undergo some revision to accommodate the fact the same subject is tested in multiple conditions. We will discuss the general features and assumptions of within-subject ANOVA, then anchor the discussion with an actual analysis.

### WHY DO A REPEATED MEASURES STUDY?

Repeated measures (also called within-subject) studies are used for several reasons. First, by using a subject as her own control a more powerful (greater likelihood of finding a real difference) analysis is possible. For example, consider testing me under two drug conditions compared to testing two individuals, each under a single condition. By testing me twice instead of different people each time, the variability due to person-to-person differences is reduced when comparing the two means, which should provide a more sensitive analysis. A second reason in practice is cost reduction; recruitment costs are less if an individual can contribute data to multiple conditions.

However, repeated measures analyses have potential problems. Since an individual appears in multiple conditions there may be practice, fatigue, or carryover effects. Counterbalancing the order of conditions addresses the carryover problem, and the different trials or conditions are often well spaced to reduce the practice and fatigue issues.

#### Examples of Repeated Measures Analysis:

1. Marketing – Compare customer's rating on four different brands, or different products, for example four different perfume fragrances.
2. Medicine – Compare test results before, immediately after, and six months after a procedure.
3. Education – Compare performance test scores before and after an intervention program.
4. Engineering – Compare output from different machines after running 1 hour, 8 hours, 16 hours, and 24 hours.
5. Agriculture – The original research area for which these methods were developed. Different chemical treatments are applied to different areas within a plot of land (split plots).
6. Human Factors – Compare performance (reaction time, accuracy) under different environmental conditions. For example, examine pilot accuracy in reading different types of dials under varying lighting conditions.

For an accessible introduction to repeated measures with a number of worked examples, see Hand and Taylor (1987). For more technical and broad (beyond ANOVA) discussions of repeated measures analysis see Lindsey (1993) or Crowder and Hand (1990).

### THE LOGIC OF REPEATED MEASURES

In the simplest case of repeated measures analysis two values are compared for each subject. For example, suppose that for each individual we record a physiological measure under two conditions. We can obtain sample means for each drug and want to determine whether there are significant differences between the drugs in the larger population. One direct way to approach this would be to compute a difference or change score for each individual, obtained by subtracting the two drug measures, and testing whether the mean difference score is different from zero. We illustrate this in the spreadsheet below.

Table 6.1 Difference Scores with Two Conditions

<b>Subject</b>	<b>Drug 1</b>	<b>Drug 2</b>	<b>Difference</b>
1	30	28	2
2	14	18	-4
3	24	20	4
4	38	34	4
5	26	28	-2
<b>Means</b>	26.40	25.60	0.80
<b>S.D.</b>			3.63

We see a difference score is calculated for every individual and these scores are averaged together. If there were no drug differences then we would expect the average difference score to be about zero. To determine if the population mean difference score is different from zero, we need some measure of the variability of sample mean difference scores. We can obtain such a variability measure by calculating the variation of individual difference scores around the sample mean difference score. If the sample mean difference score is far enough from zero that it cannot be accounted for by the variation of individual difference scores, we say there is a significant population difference. This is what a paired t test does.

The analysis becomes a bit more complex when each subject (unit of analysis) appears in more than two levels (conditions) of a repeated measure factor. Now no single difference score can summarize the differences. We illustrate this below.

Table 6.2 Difference Scores with Four Conditions

<b>Subject</b>	<b>Drug 1</b>	<b>Drug 2</b>	<b>Drug 3</b>	<b>Drug 4</b>	<b>Diff 1</b>	<b>Diff 2</b>	<b>Diff 3</b>
					<b>1 Versus 2</b>	<b>2 Versus 3</b>	<b>3 Versus 4</b>
1	30	28	16	34	2	12	-18
2	14	18	10	22	-4	8	-12
3	24	20	18	30	4	2	-12
4	38	34	20	44	4	14	-24
5	26	28	14	30	-2	14	-16
<b>Mean</b>					0.80	10.00	-16.40
<b>S.D.</b>					3.63	5.10	4.98

Although no one difference score can summarize all drug differences here, we can compute additional difference scores, and thus account for drug effects. As you would imagine the number of these differences, or contrasts, is equal to the degrees of freedom available (one less than the number of levels in the factor). For two conditions, only one contrast is possible; for four conditions, there are three; for  $k$  conditions,  $k-1$  contrasts are required. If the assumptions of repeated measures ANOVA are met then these differences, or contrasts between conditions, can be pooled together to provide a significance test for an overall effect.

We used simple differences to compare the drug conditions (drug 1 minus drug 2, etc.) There are many other contrasts that could be applied. For example, we could have calculated drug 1 minus the mean of drugs 2, 3, and 4; then drug 2 versus the mean of drugs 3 and 4; and finally drug 3 versus drug 4. As long as the assumptions of repeated measures are met, the specific choice of contrasts doesn't matter when the overall test is calculated. However, if you have planned comparisons you want tested, then you would request those.

In each of the two above examples, we wound up with one fewer difference variable than the original number of conditions. There is another variable that is calculated in repeated measures, which represents the mean across all conditions. It is used when testing effects of between-group factors, having averaged across all levels of the repeated measure factor(s). This mean effect is shown in the illustration below:

Table 6.3 Mean and Difference Scores with Four Conditions

Subject	Drug 1	Drug 2	Drug 3	Drug 4	Mean Across Four Drugs	Diff 1 1 Versus 2	Diff 2 2 Versus 3	Diff 3 3 Versus 4
1	30	28	16	34	27	2	12	-18
2	14	18	10	22	16	-4	8	-12
3	24	20	18	30	23	4	2	-12
4	38	34	20	44	34	4	14	-24
5	26	28	14	30	24.5	-2	14	-16
<b>Mean</b>					24.90	0.80	10.00	-16.40
<b>S. D.</b>					6.52	3.63	5.10	4.98

The mean score across drug conditions for each subject is recorded in the mean column. As mentioned above any tests involving only between-group factors (for example, sex, age group) would use this variable.

This idea of computing difference scores or contrasts across conditions for each subject, then using the means and subject to subject variation as the basis of testing whether the average contrast value is different from zero in the population, is the core concept of repeated measures ANOVA. Once you become comfortable with it, the rest falls into place. SPSS performs repeated measures ANOVA by computing contrasts across the

repeated measures factor levels for each subject, and then testing whether the means of the contrasts are significantly different from zero. A matrix of coefficients detailing these contrasts can be displayed and is called the transformation matrix.

**ASSUMPTIONS**

A repeated measure ANOVA has several assumptions common to all ANOVA. First, that the model is correctly specified and additive. Secondly, that the errors follow a normal distribution and are independent of the effects in the model. This latter assumption implies homogeneity of variance when more than a single group is involved. As with general ANOVA, moderate departures from normality do not have a substantial effect on the analysis, especially if the sample sizes are large and the shape of the distribution is similar from group to group (if multiple groups are involved). In multi-group studies, failure of homogeneity of variance is a problem unless the sample sizes are about equal.

In addition to standard ANOVA assumptions, there is one specific to repeated measures when there are more than two levels to a repeated measures factor. If a repeated measures factor contains only two levels, there is only one difference variable that can be calculated, and you need not be concerned about the assumption. However, if a repeated measures factor has more than two levels, you generally want an overall test of differences (main effect). Pooling the results of the contrasts (described above) between conditions creates the test statistic (F). The assumption called sphericity deals with when such pooling is appropriate. The basic idea is that if the results of two or more contrasts (the sums of squares) are to be pooled, then they should be equally weighted and uncorrelated. To illustrate why this is important, view the spreadsheet below:

**Table 6.4 Scale Differences and Redundancies in Contrasts**

Subject	Drug 1	Drug 2	Drug 3	Drug 4	Diff 1 1 Versus 2	Diff 2 100* (2-3)	Diff 3 1 Versus 2
1	30	28	16	34	2	1200	2
2	14	18	10	22	-4	800	-4
3	24	20	18	30	4	200	4
4	38	34	20	44	4	1400	4
5	26	28	14	30	-2	1400	-2
<b>Mean</b>					0.80	1000.00	0.80
<b>S.D.</b>					3.63	509.90	3.63

The first contrast variable represents the difference between drug 1 and drug 2 (Drug 1 – Drug 2). However, the second is 100 times the difference between Drug 2 and Drug 3. It is clear from the mean and standard deviation values of the second difference variable that this variable would dominate the other difference variables if the results were pooled. In order to protect against this, normalization is applied to the coefficients used in creating the contrasts (each coefficient is divided by the square root of the sum of the squared coefficients).

Also, notice that the third contrast is a duplicate of the first. Admittedly, this is an extreme example, but it serves to make the point that since the results from each contrast are pooled (summed), then any correlation among the contrast variables will yield incorrect test statistics. In order to provide the best chance of uncorrelated contrast variables, the contrasts or transformations are forced to be orthogonal (uncorrelated) before applying them to the data.

This combination of normalization and forcing the original contrasts to be orthogonal (uncorrelated) is called orthonormalization. Again, when actually applied to the data, these properties may not hold, and that is where the test of sphericity plays an important role.

This combination of assumptions, equal variances of the contrast variables and zero correlation among them, is called the sphericity assumption. It is called sphericity because a sphere in multidimensional space would be defined by an equal radius value along each perpendicular (uncorrelated) axis. Although contrasts are chosen so that sphericity will be maintained, when applied to a particular data set, sphericity may be violated. The variance-covariance matrix of a group of contrast variables that maintain sphericity would exhibit the pattern shown below.

**Table 6.5 Covariance Matrix of Contrast Variables when Sphericity Holds**

	<b>Dif 1</b>	<b>Dif 2</b>	<b>Dif 3</b>
<b>Dif 1</b>	V	0	0
<b>Dif 2</b>	0	V	0
<b>Dif 3</b>	0	0	V

The diagonal elements represent the variance of each contrast when applied to the data and the off-diagonal elements are the covariances. If the sphericity assumption holds in the population, the variances will have the same value (represented by the V) and the covariances will be zero.

A test of the sphericity assumption is available. If the sphericity assumption is met then the usual “F” test (pooling the results from each contrast) is the most powerful test. When sphericity does not hold, there are several choices available. Technical corrections (Greenhouse-Geisser, Huynh-Feldt) can be made to the “F” tests (adjusting the number of the degrees of freedom) that modify the results based on the degree of sphericity violation. Another alternative is to take a multivariate approach in which contrasts are tested simultaneously while taking explicit account of the correlation and variance differences. The difficulty in choosing between these approaches is that no single method has been

found (in Monte Carlo studies) to be best under all conditions examined. Also, the test for sphericity itself is not all that sensitive. For a summary of the various approaches and a suggested strategy for testing, see Looney and Stanley (1989).

### **Data Set**

The data are reported in Bock (1975, p.454) and consist of vocabulary scores obtained from a cohort of pupils at the eighth through eleventh grade level. Alternative forms of the vocabulary section of the Cooperative Reading Tests were administered and rescaled to an arbitrary origin. Interest is in the growth rate of vocabulary at a time when physical growth is slowing. Sixty-four subjects were studied.

### **PROPOSED ANALYSIS**

We will perform a repeated measures analysis on the vocabulary growth data. There is specific interest in the trend over time – is it linear? The data will be examined then repeated measures ANOVA applied with attention paid to the assumptions mentioned above.

### **KEY CONCEPT**

The key concept to repeated measures analysis is that the contrasts (which are data transformations) will be applied across conditions of the within-subject factors, and if we conclude the contrasts are non-zero in the population, there are significant differences between the conditions.

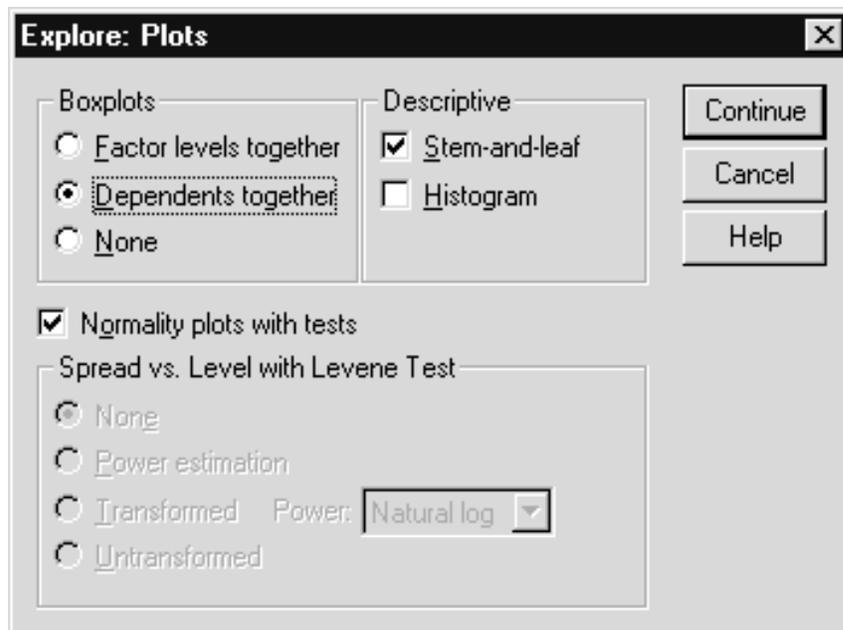
Click **File..Open..Data** (move to the **c:\Train\Anova** directory)  
Select **SPSS Portable (.por)** on the Files of Type drop-down list  
Double-Click on **Vocab**  
Click on **Analyze..Descriptive Statistics..Explore**  
Select the variables **Grade8, Grade9, Grade 10, and Grade11**  
and move them to the **Dependent List** box.

Figure 6.1 Explore Dialog Box



Click on the **Plots** pushbutton  
 Click **Dependents together** option button in the Boxplots area  
 Click the **Normality tests with plots** check box

Figure 6.2 Plots Dialog Box



Placing the dependent variables together in a single boxplot, instead of separate plots (Factor levels together), permits direct comparison of the variables. Normal probability plots and tests are also requested.

Click **Continue**  
 Click **OK**

The command below will run the analysis.

```
EXAMINE
  VARIABLES=grade8 grade9 grade10 grade11
  /PLOT BOXPLOT STEMLEAF NPLOT
  /COMPARE VARIABLES
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

We request summaries of the four variables, a normal probability plot with normality test will appear (/Plot Npplot) for each variable. Also, the four variables will appear in a single boxplot (/Compare Variables).

**Figure 6.3 Descriptives for Grade 8**

		Statistic	Std. Error
GRADE8	Mean	1.1372	.2361
	95% Confidence Interval for Mean		
	Lower Bound	.6653	
	Upper Bound	1.6090	
	5% Trimmed Mean	1.0465	
	Median	1.2300	
	Variance	3.568	
	Std. Deviation	1.8890	
	Minimum	-2.19	
	Maximum	8.26	
	Range	10.45	
	Interquartile Range	2.2650	
	Skewness	.802	.299
	Kurtosis	2.309	.590

Figure 6.4 Descriptives for Grade 9

GRADE9	Mean		2.5417	.2606
	95% Confidence Interval for Mean	Lower Bound	2.0209	
		Upper Bound	3.0625	
	5% Trimmed Mean		2.4533	
	Median		2.4550	
	Variance		4.347	
	Std. Deviation		2.0849	
	Minimum		-1.31	
	Maximum		9.55	
	Range		10.86	
	Interquartile Range		2.3775	
	Skewness		.790	.299
	Kurtosis		1.131	.590

Figure 6.5 Descriptives for Grade 10

GRADE10	Mean		2.9883	.2711
	95% Confidence Interval for Mean	Lower Bound	2.4465	
		Upper Bound	3.5300	
	5% Trimmed Mean		2.8510	
	Median		2.7150	
	Variance		4.704	
	Std. Deviation		2.1688	
	Minimum		-.66	
	Maximum		10.24	
	Range		10.90	
	Interquartile Range		2.8575	
	Skewness		.925	.299
	Kurtosis		1.516	.590

Figure 6.6 Descriptives for Grade 11

GRADE11	Mean		3.4716	.2407
	95% Confidence Interval for Mean	Lower Bound	2.9906	
		Upper Bound	3.9525	
	5% Trimmed Mean		3.4017	
	Median		3.2700	
	Variance		3.708	
	Std. Deviation		1.9255	
	Minimum		-2.22	
	Maximum		10.58	
	Range		12.80	
	Interquartile Range		2.3025	
	Skewness		.686	.299
	Kurtosis		2.813	.590

As we can see from the descriptive statistics, the mean score for the reading tests is going up in each year (from 1.1372 in Grade 8 to 3.4716 in Grade 11), but the variances are fairly constant across the years (from 3.568 to 4.704).

**Figure 6.7 Normality Tests**

**Tests of Normality**

	Kolmogorov-Smirnov <sup>a</sup>		
	Statistic	df	Sig.
GRADE8	.079	64	.200*
GRADE9	.118	64	.026
GRADE10	.069	64	.200*
GRADE11	.111	64	.048

\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

The normality tests show that there is no problem with the assumption of normality for grades 8 and 10. However, grades 9 and 11 show that there is some deviation from normality in those grade results. Although the assumption of normality is violated, the sample size is large enough that we can probably ignore that violation.

**Figure 6.8 Q-Q Plot for Grade 8**

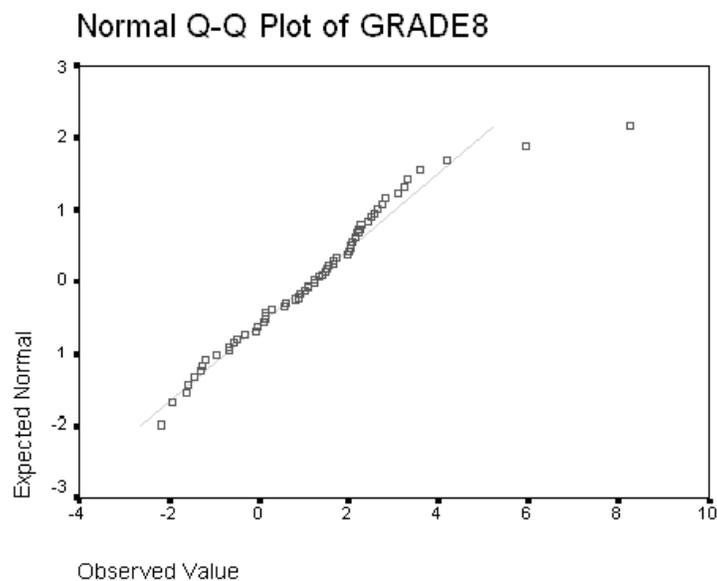


Figure 6.9 Q-Q Plot for Grade 9

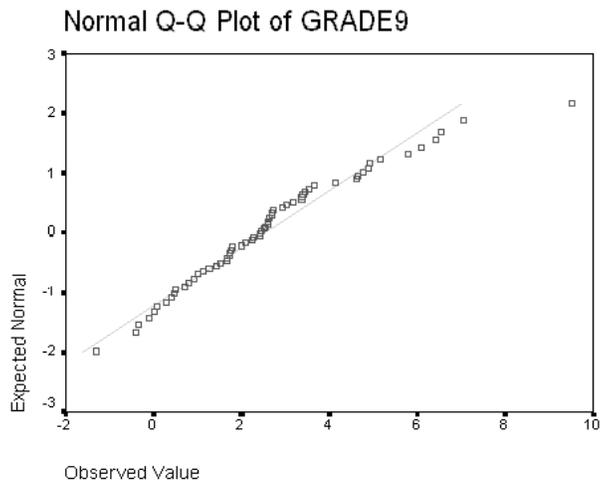


Figure 6.10 Q-Q Plot for Grade 10

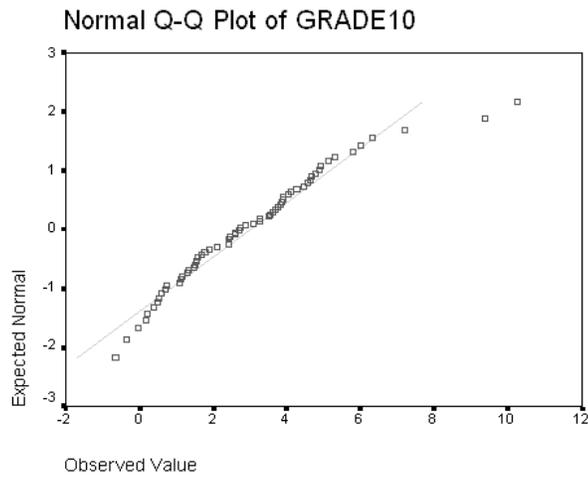
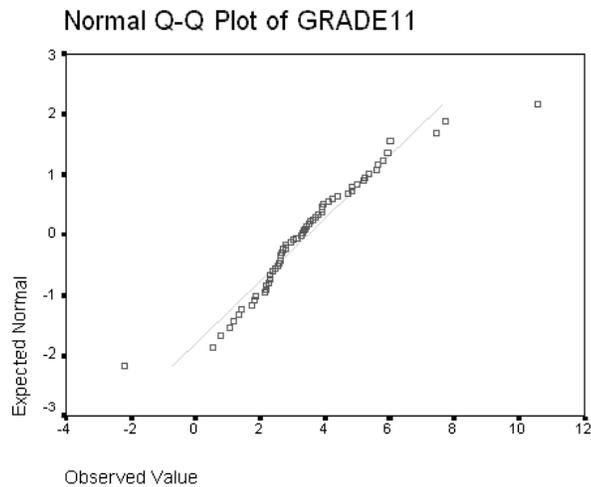
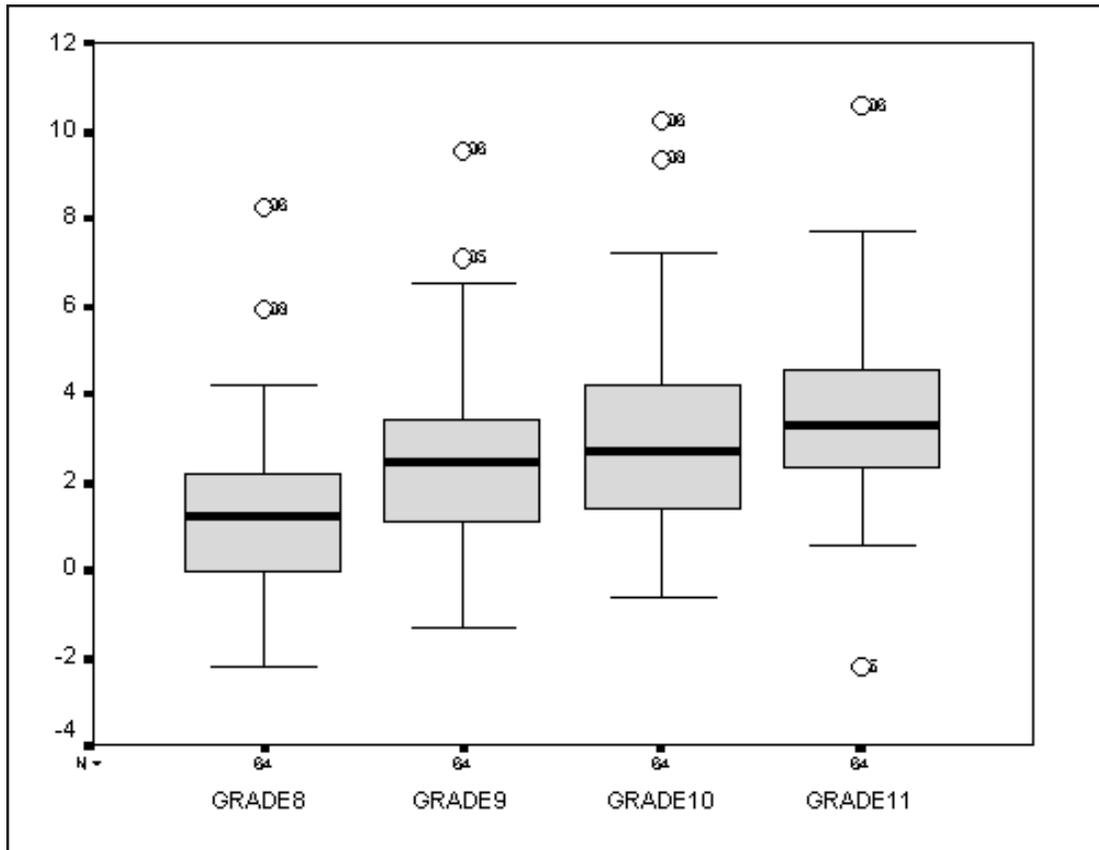


Figure 6.11 Q-Q Plot for Grade 11



These Q-Q plots also give us some indication of the degree to which the normality assumption is violated. Although the normality tests showed that grades 9 and 11 had some deviation from the normal, the Q-Q plots are similar for all the grades. Again, we should note that the sample size is somewhat large and that we can probably not worry about these violations of normality.

Figure 6.12 Box and Whiskers Plot



**COMPARING THE GRADE LEVELS**

The Box plot indicates that the variation of scores within a test year is fairly constant. There are a few outliers; the case id information indicates that for the most part, the same few individuals stand out. From the medians we see that vocabulary scores grow over the several year period, and this growth seems to be slowing.

We have only one group of subjects. Each subject has a vocabulary score under the four grade levels. Notice all four of the vocabulary scores are attached to a single case (examine data in Data Editor window – not shown). If the four measures for a subject were spread throughout the file, the analysis can still be run within SPSS, but only by using the General Linear Model Univariate dialog box.

Click **Analyze..General Linear Model..Repeated Measures**

Here we provide names for any repeated measures factors and indicate the number of levels for each. Unlike a between-group factor which would be a variable (for example, region), a repeated measures factor is expressed as a set of variables.

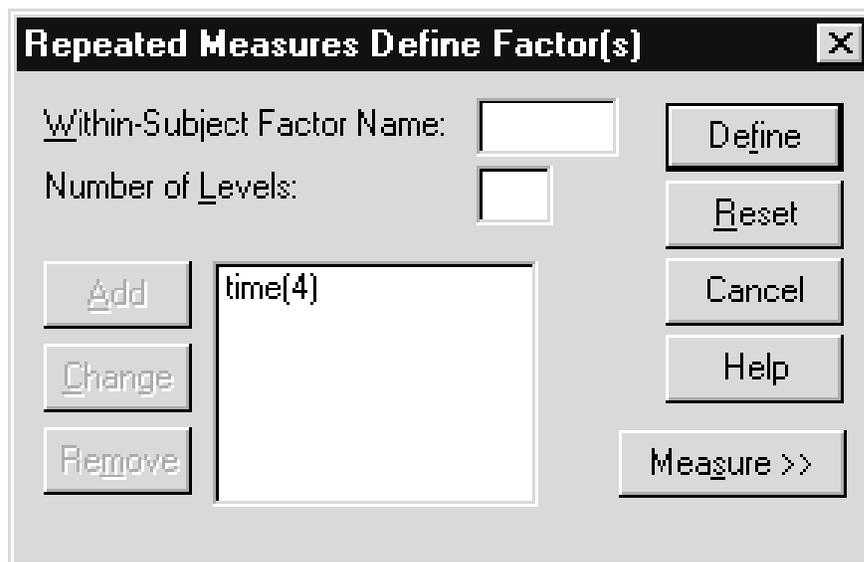
Replace **factor1** with **Time** in the Within-Subject Factor Name text box

Press **Tab** key to move to the Number of Levels text box

Type **4**

Click **Add** pushbutton

Figure 6.13 Define Repeated Measures (Within-Subject) Factor



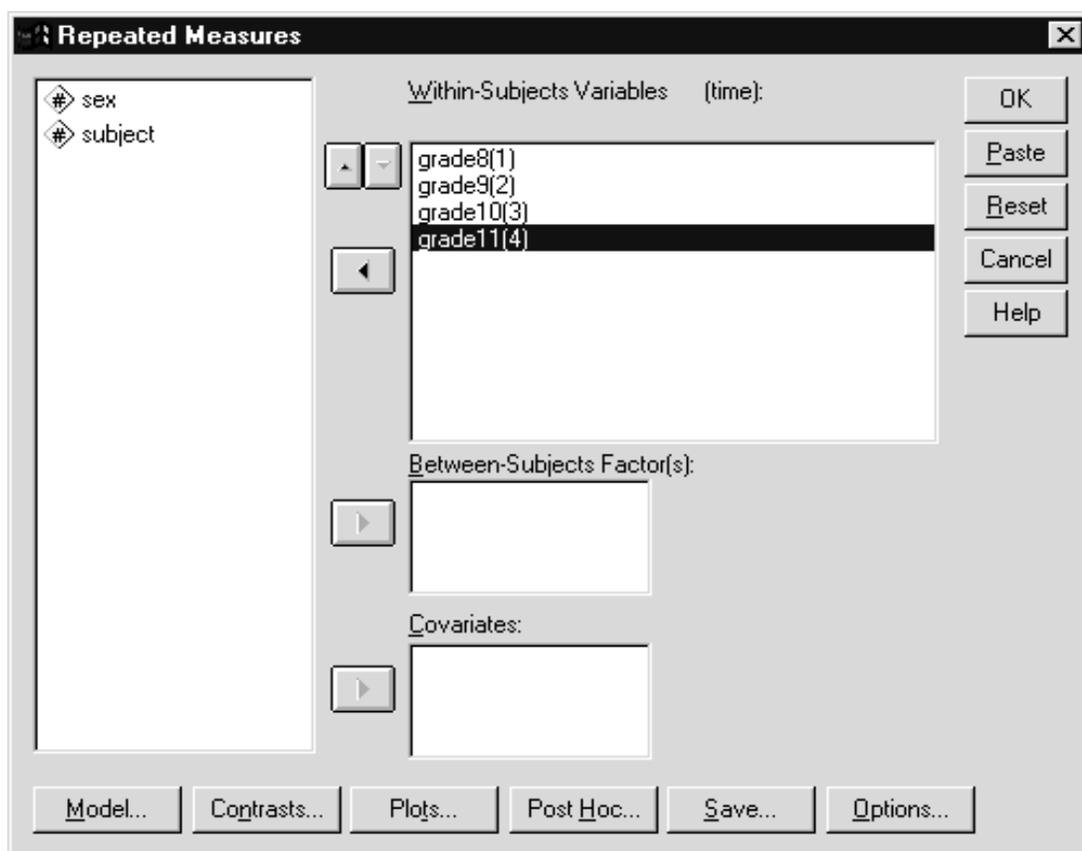
We have defined one factor with four levels. In a more complex study (we will see one later in this chapter) additional repeated measures can be added. The Measure pushbutton is used to provide two pieces of information. First if there are multiple dependent measures involved in the analysis (for example, suppose we also took four measures of mathematical skills for each of our 64 subjects), this is declared in the measure area. Secondly, you can use the Measure area to provide a label for the dependent measure in the results. Recall we named our four variables Time1 to Time4 so there would be no ambiguity about which factor level each represented. However, this choice on names does not indicate that these variables all measure vocabulary scores. You can supply such labeling information in the Measures area.

Click **Define** pushbutton

In this dialog box we link the repeated measures factor levels to variable names, and declare any between-subject factors and covariates. Notice the Within-Subjects Variables box lists Time as the factor and provides four lines labeled with level numbers 1 through 4. We must match the proper variable to each of the factor levels. This step should be done very carefully since incorrect matching of names and levels will generally produce an incorrect analysis (especially if more than one repeated measure factor is involved). We can move the variables one by one, but since they are in the correct order we will move them as a group.

Move **grade8**, **grade9**, **grade10**, and **grade11** into the Within-Subjects Variables box (maintain this variable order)  
Click **grade11** to select it.

Figure 6.14 Main Repeated Measures Dialog Box

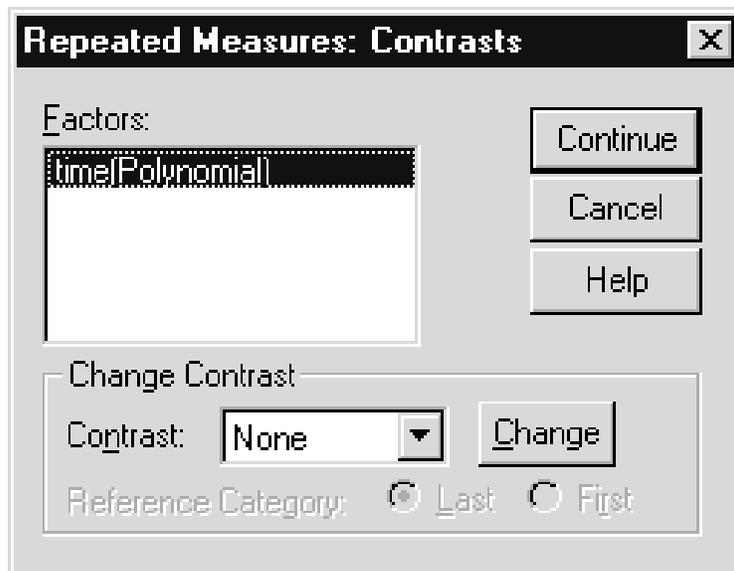


The variable corresponding to each grade level is matched with the proper time level. Since grade11 is selected, the up arrow button is active. These up and down buttons will move variables up and down the list, so you can easily make changes if the original ordering is incorrect. We have neither between-subject factors nor covariates and can proceed with the analysis, but first let us examine some of the available features.

In the model dialog box (not shown) by default a complete model (all factors and interactions) will be fit. As with procedures we saw earlier in the course, a customized model can be fit for either between or within-subject factors. This is usually done when specialty designs (Latin squares, incomplete designs) are run. The Contrasts pushbutton is used to request that particular contrasts be applied to a factor (recall our discussion of difference or contrast variables earlier).

Click **Contrasts** pushbutton

Figure 6.15 Contrasts Dialog Box

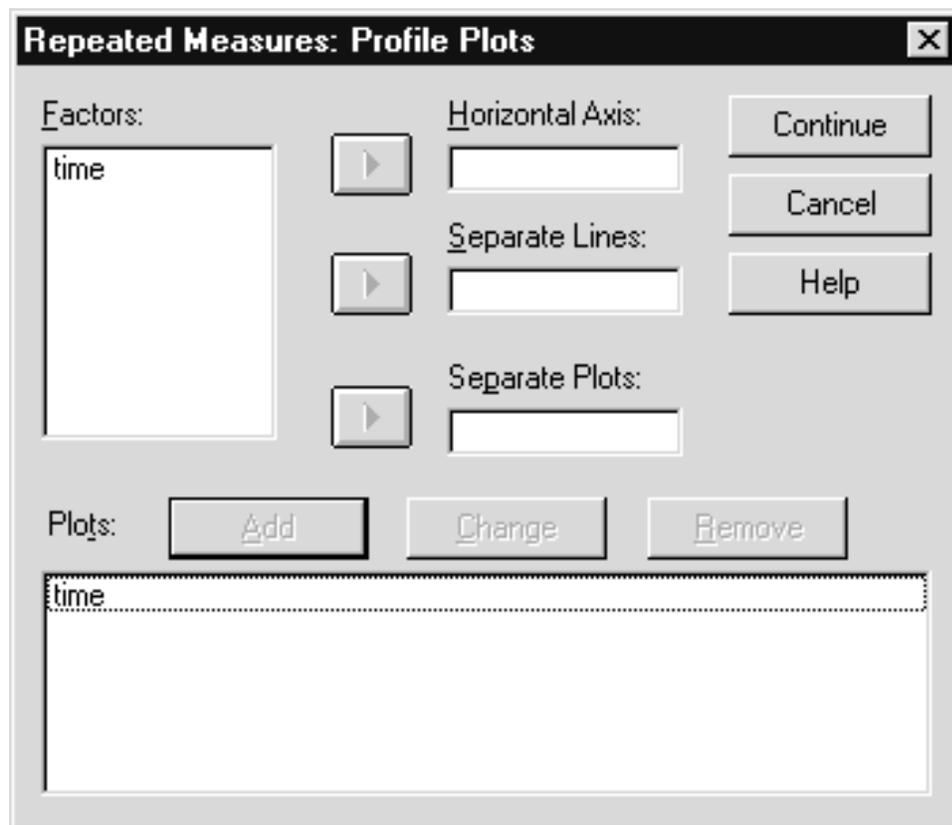


Check to see which contrast is selected  
If it is not polynomial then change it to **Polynomial**  
Click **Continue**  
Click **Plots** pushbutton

The Plots pushbutton generates profile plots that graph means at factor level combinations for up to three factors at a time. Such plots are powerful tools in understanding interaction effects. We will only request a plot for time, our repeated measure factor.

Click on **Time** and move it to the **Horizontal Axis** list box  
Click **Add**

Figure 6.16 Plots Dialog Box

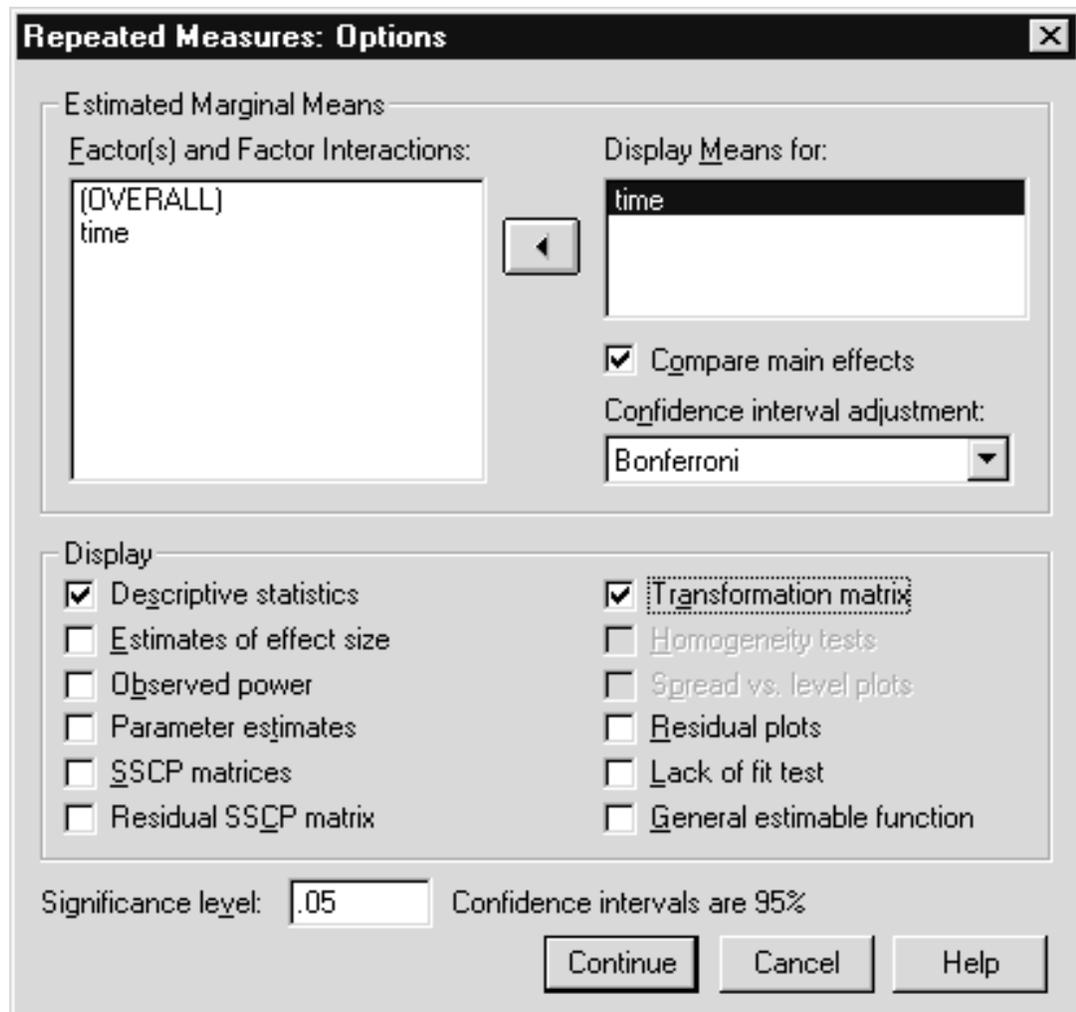


Click **Continue**

The Post Hoc dialog box was discussed earlier in the class; it performs post hoc tests of means for between-subject factors. We will look at the available tests for repeated measures factors shortly. The Save dialog box allows you to save predicted values, various residuals and influential point measures. Also, you can save the estimated coefficients to a file for later manipulation (perhaps in a prediction model, or to compare results from different data sets). We will look at the Option dialog box more closely.

- Click **Options** pushbutton
- Move **Time** into the **Display Means for** list box
- Click to check the **Compare Main Effects** checkbox
- Select **Bonferroni** on the **Confidence interval adjustment** drop-down list
- Click on **Descriptive statistics** check box
- Click **Transformation Matrix** check box

Figure 6.17 Options Dialog Box



We request descriptive statistics. Estimated marginal means can be produced for any factors in the model (here time). Since we are fitting a complete model, the estimated marginal means are identical to the estimated means. We request pairwise comparisons for the time factor using Bonferroni adjustments (the available adjustments for repeated measure factors are LSD, Bonferroni and Sidak). In addition, we have asked to see the transformation matrix. The transformation matrix contains the contrast coefficients that are applied to the repeated measures factor(s) to create the difference or contrast variables used in the analysis. Here we display it only to reinforce our earlier discussion of this topic. Diagnostic residual plots are available and there is a control to modify the confidence limits (default is 95%). The SSCP (sums of squares and cross products) matrices are not ordinarily viewed. However, they do contain the sums of squares for each of the contrast variables. By viewing them you can see that the overall test simply sums up the individual contrast sums of squares, which is why sphericity is necessary.

Click **Continue** to process the option requests  
Click **OK** to run the analysis

The SPSS syntax below will run the repeated measures analysis.

```
GLM
  grade8 grade9 grade10 grade11
  /WSFACTOR = time 4 Polynomial
  /METHOD = SSTYPE(3)
  /PLOT = PROFILE( time )
  /EMMEANS = TABLES(time) COMPARE ADJ(BONFERRONI)
  /PRINT = DESCRIPTIVE TEST(MMATRIX)
  /CRITERIA = ALPHA(.05)
  /WSDESIGN = time .
```

First the variables that constitute the repeated measures factor are listed. The WSFACTOR (within-subject factor) subcommand declares time to be a within-subject factor with four levels. In addition, polynomial contrasts will be applied when creating the contrast variables. Polynomial contrasts will perform linear, quadratic, and cubic contrasts on the time factor. If there are significant changes in vocabulary over time, as we expect, these contrasts will allow us to examine its specific form. The Print TEST (MMATRIX) specification will have the transformation (called the M Matrix) display. Method declares the sums of squares type.

### EXAMINING RESULTS

The first summary displays information about the factors in the model.

Figure 6.18 Factor Summary

Within-Subjects Factors	
Measure: MEASURE_1	
TIME	Dependent Variable
1	GRADE8
2	GRADE9
3	GRADE10
4	GRADE11

There is only a single within-subject (repeated measures) factor and no between-subject factors.

Figure 6.19 Descriptive Statistics

	Mean	Std. Deviation	N
GRADE8	1.1372	1.8890	64
GRADE9	2.5417	2.0849	64
GRADE10	2.9883	2.1688	64
GRADE11	3.4716	1.9255	64

Means, standard deviations, and sample sizes appear for each factor level. If you were unsure of your matching the variable names to factor levels in the Define Repeated Measures Factors dialog box, you can compare these means to those you would obtain from the Descriptives, Means, or Explore procedures to insure the proper variables are matched with the proper factor levels. Again, we see the increase in the mean scores as grade level increases.

Multivariate test results appear next. Since they would typically be used only if the sphericity assumption fails, we will skip these results for now and examine the sphericity test.

Figure 6.20 Mauchly's Sphericity Test

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon <sup>a</sup>		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TIME	.903	6.315	5	.277	.942	.991	.333

We see from the Significance (Sig.) information that the data are consistent with the sphericity assumption. The significance value is above .05 (.277), indicating that the covariance matrix of orthonormalized transformation variables is consistent with sphericity (diagonal elements identical and off-diagonal elements zero in the population). Since sphericity has been maintained we can use the standard (pooled) ANOVA

results, and need not resort to alternative (multivariate) or adjusted (degree of freedom adjustment) tests. The Epsilon section of the pivot table provides the degree of freedom modification factor that should be applied if the sphericity result were significant. Let us take a brief look at the multivariate results.

Figure 6.21 Multivariate Tests

Multivariate Tests <sup>b</sup>						
Effect		Value	F	Hypothesis df	Error df	Sig.
TIME	Pillai's Trace	.826	96.446 <sup>a</sup>	3.000	61.000	.000
	Wilks' Lambda	.174	96.446 <sup>a</sup>	3.000	61.000	.000
	Hotelling's Trace	4.743	96.446 <sup>a</sup>	3.000	61.000	.000
	Roy's Largest Root	4.743	96.446 <sup>a</sup>	3.000	61.000	.000

a. Exact statistic

b.  
Design: Intercept  
Within Subjects Design: TIME

Remember these results need not be viewed since sphericity has been maintained. Here the test is whether all of the contrast variables (representing vocabulary score differences) are zero in the population, while explicitly taking into account any correlation and variance differences in the contrast variables. So if sphericity were violated these results could be used. Explanations about the various multivariate tests were given in Chapter 5. The multivariate tests indicate there are vocabulary score differences by grade.

Figure 6.22 Within-Subject Effects

Tests of Within-Subjects Effects						
Measure: MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	194.338	3	64.779	79.019	.000
	Greenhouse-Geisser	194.338	2.827	68.743	79.019	.000
	Huynh-Feldt	194.338	2.974	65.355	79.019	.000
	Lower-bound	194.338	1.000	194.338	79.019	.000
Error(TIME)	Sphericity Assumed	154.942	189	.820		
	Greenhouse-Geisser	154.942	178.102	.870		
	Huynh-Feldt	154.942	187.336	.827		
	Lower-bound	154.942	63.000	2.459		

This table contains the standard repeated measures output based on summing the results from each contrast, as well as sphericity corrected results. It shows the results for (1) sphericity assumed, and then (2)

Greenhouse-Geisser, (3) Huynh-Feldt, and (4) Lower Bound adjustments. The test result (sphericity assumed) is highly significant, more so than the multivariate test, which is what we expect if sphericity holds: the pooled test is more powerful. Thus, we conclude there are significant differences in vocabulary across grade levels.

Figure 6.23 Test of Contrasts

**Tests of Within-Subjects Contrasts**

Measure: MEASURE\_1

Source	TIME	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Linear	177.593	1	177.593	221.883	.000
	Quadratic	13.579	1	13.579	19.465	.000
	Cubic	3.166	1	3.166	3.293	.074
Error(TIME)	Linear	50.425	63	.800		
	Quadratic	43.951	63	.698		
	Cubic	60.567	63	.961		

Significant tests will be performed on each of the contrast variables used to construct a repeated measure factor. Recall that by default, polynomial contrasts are used. Since the repeated measure factor is time, these contrasts test whether there are significant linear, quadratic and cubic trends in vocabulary growth over time. Note that linear and quadratic trends are significant (the linear contrast has a very large F value), while cubic is not. This is consistent with the earlier comment that vocabulary scores increase over time, but the growth seemed to be slowing down.

Figure 6.24 Test of Between-Subjects Effects

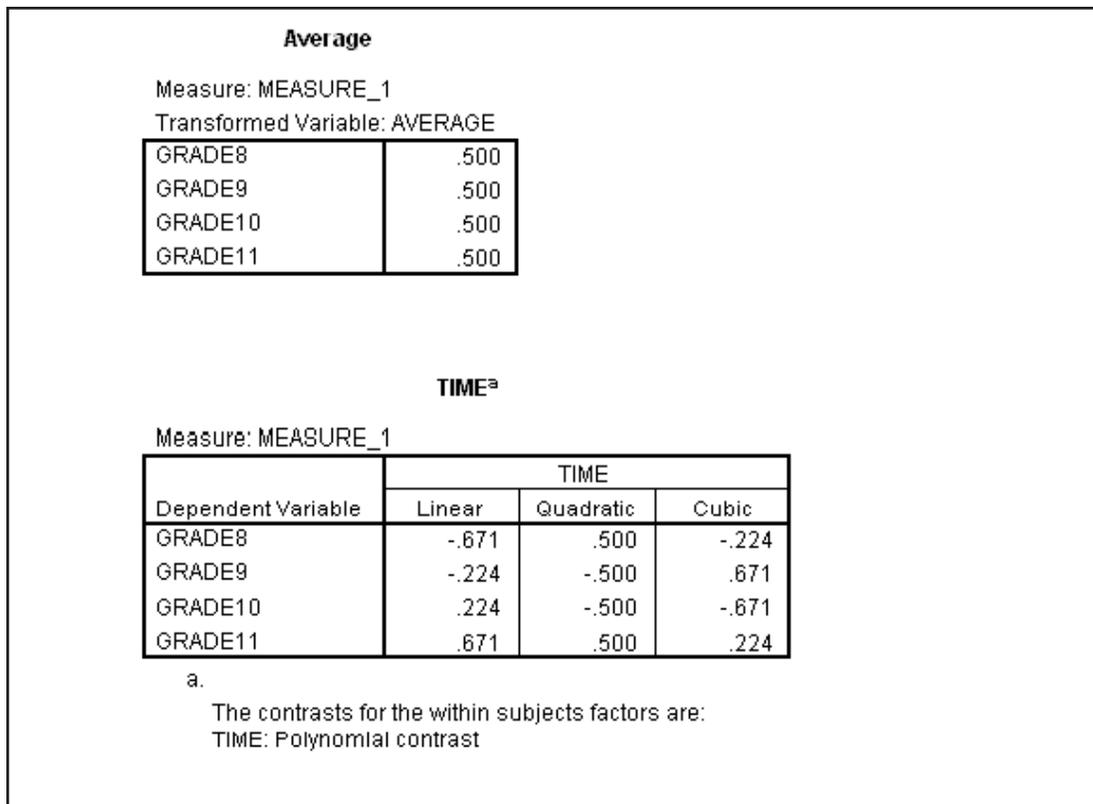
**Tests of Between-Subjects Effects**

Measure: MEASURE\_1  
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1644.708	1	1644.708	118.608	.000
Error	873.603	63	13.867		

There were no between-subject factors in this study. If there were, the test results for them would appear in this section. There is a test of the intercept, or grand mean; this simply tests whether the average of all vocabulary scores is equal to zero in the population – not an interesting hypothesis to test.

Figure 6.25 Transformation Matrix



The transformation variables are split into two groups: one corresponding to the average across the repeated measures factor, the others defining the repeated measures factor. The coefficients for the Average variable are all .5, meaning each variable is weighted equally in creating the Average transformation variable. If you wonder why the weights are not .25, recall that normalization requires the sum of the squared weights to equal one. Turning to the transformed variables that represent the time effect, the three sets of coefficients are orthogonal polynomials corresponding to linear, quadratic, and cubic terms. Looking at the first we see that there is a constant increase (of about .447) in the value of the coefficients across the four grade levels. In a similar way, the second transformation variable has two sign changes (negative to positive, then positive to negative) over the grade levels; this constitutes a quadratic effect. The *SPSS Advanced Models* manual has additional information about the commonly used transformations.

Recall that the transformations are orthogonal; you can verify this for any pair by multiplying their coefficients at each level of the factor and summing these products. The sum should be zero. For linear and quadratic we can calculate  $(-.671*.5 - .224*.5 + .224*.5 + .671*.5)$ , which is indeed zero.

Figure 6.26 Transformation Matrix (M Matrix)

**Transformation Coefficients (M Matrix)**

Measure: MEASURE\_1

Dependent Variable	TIME			
	1	2	3	4
GRADE8	1	0	0	0
GRADE9	0	1	0	0
GRADE10	0	0	1	0
GRADE11	0	0	0	1

Since we ask for analyses to compare main effects in the Options dialog box, a new transformation matrix is used to create four variables equivalent to the four grade levels: the identity transformation. Notice this is a separate analysis after the others have been completed using the original transformation matrix.

Also, note the estimated marginal means match the observed means (this pivot table is not shown).

Figure 6.27 Pairwise Test Results

**Pairwise Comparisons**

Measure: MEASURE\_1

(I) TIME	(J) TIME	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	95% Confidence Interval for Difference <sup>a</sup>	
					Lower Bound	Upper Bound
1	2	-1.405*	.155	.000	-1.826	-.983
	3	-1.851*	.135	.000	-2.219	-1.483
	4	-2.334*	.157	.000	-2.761	-1.908
2	1	1.405*	.155	.000	.983	1.826
	3	-.447	.175	.078	-.923	2.951E-02
	4	-.930*	.176	.000	-1.408	-.451
3	1	1.851*	.135	.000	1.483	2.219
	2	.447	.175	.078	-2.951E-02	.923
	4	-.483*	.160	.022	-.919	-4.803E-02
4	1	2.334*	.157	.000	1.908	2.761
	2	.930*	.176	.000	.451	1.408
	3	.483*	.160	.022	4.803E-02	.919

Based on estimated marginal means

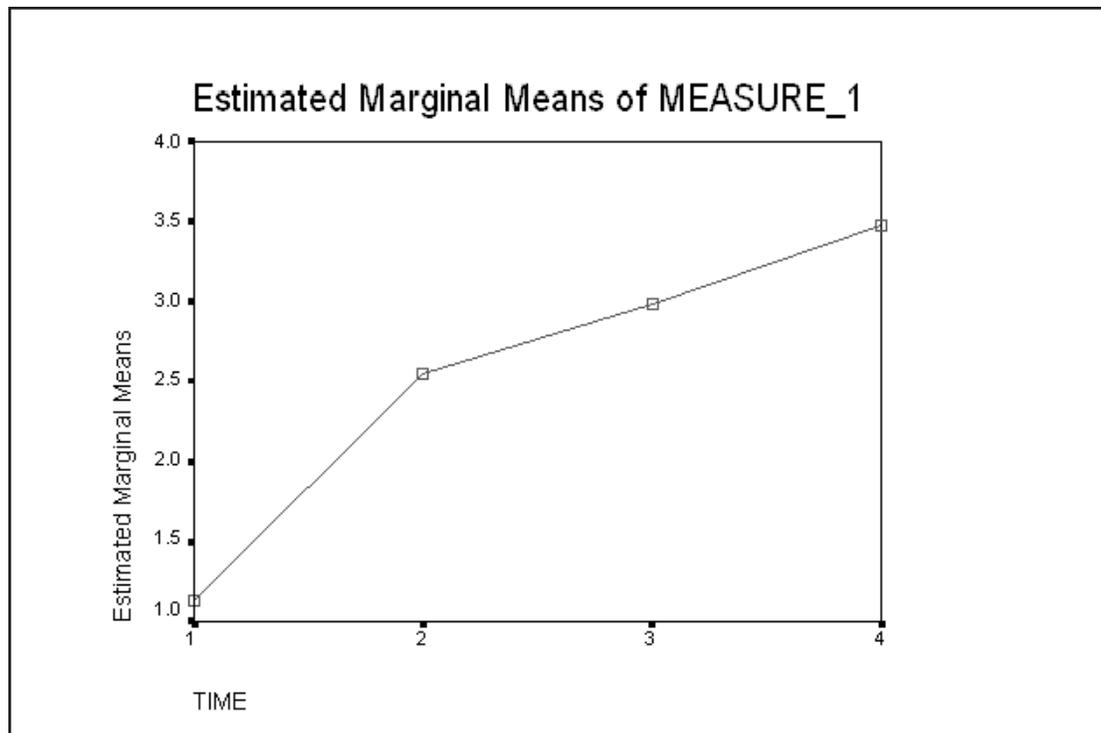
\*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

Each grade level mean is tested against every other; essentially we are performing all pairwise tests with a Bonferroni correction. The footnotes indicate that each test is performed at an adjusted level of significance using a Bonferroni correction. Thus the probability of obtaining one or more false positive results is .05. We find in our study that the grade 8 (time 1) scores are significantly different from all the others; grade 9 (time 2) is different from grade 8 and grade 11 (time 4); grade 10 (time 3) is different from grade 8 and 11; while grade 11 is significantly different from grades 8, 9, and 10. Substituting Bonferroni corrected paired t tests for post hoc comparisons provides a means to investigate differences within a repeated measures factor.

The program will also run a multivariate ANOVA attempting to test the pairwise comparisons simultaneously; this is of no interest to us.

**Figure 6.28 Profile Plot of Means**



This plot (not really necessary since with one factor there can be no interaction) shows us how the mean of the vocabulary scores is increasing with grade level.

### PLANNED COMPARISON

Suppose we had some specific hypothesis about the grade levels that we wished to test. For example, if we thought that the grade to grade promotion made a difference in the student's vocabulary score, we might want to test grade 8 versus grade 9; grade 9 versus grade 10; and grade 10 versus grade 11. The Contrast pushbutton provides a variety of planned comparisons and customized contrasts can be input using syntax.

Click the Dialog Recall tool , then click **Repeated**

#### Measures

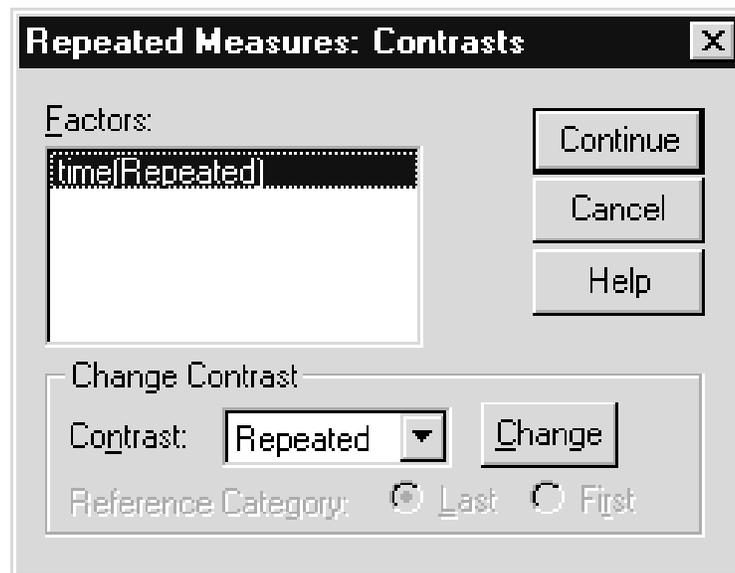
Click **Define** pushbutton

Click **Contrasts** pushbutton

Click **Contrast** drop-down arrow and select **Repeated**

Click **Change** pushbutton

Figure 6.29 Requesting Planned Comparisons



Repeated contrasts will compare each category to the one adjacent. Right click on any contrast on the list to obtain a brief description of it. *The SPSS Advanced Models* manual contains more details. Also be aware that you can provide custom contrasts using the Special keyword in syntax.

Click **Continue** to process the contrasts

Click **OK** to run the analysis

The command below will run the analysis.

```

GLM
  grade8 grade9 grade10 grade11
  /WSFACTOR = time 4 Repeated
  /METHOD = SSTYPE(3)
  /PLOT = PROFILE( time )
  /EMMEANS = TABLES(time) COMPARE ADJ(BONFERRONI)
  /PRINT = DESCRIPTIVE TEST(MMATRIX)
  /CRITERIA = ALPHA(.05)
  /WSDESIGN = time .
  
```

The Wsfactor subcommand now requests that repeated contrasts be used in place of the default polynomials.

Again, most of the output is identical to the previous runs; we focus on the contrast tests and the transformation matrix.

**Figure 6.30 Tests of Contrasts**

**Tests of Within-Subjects Contrasts**

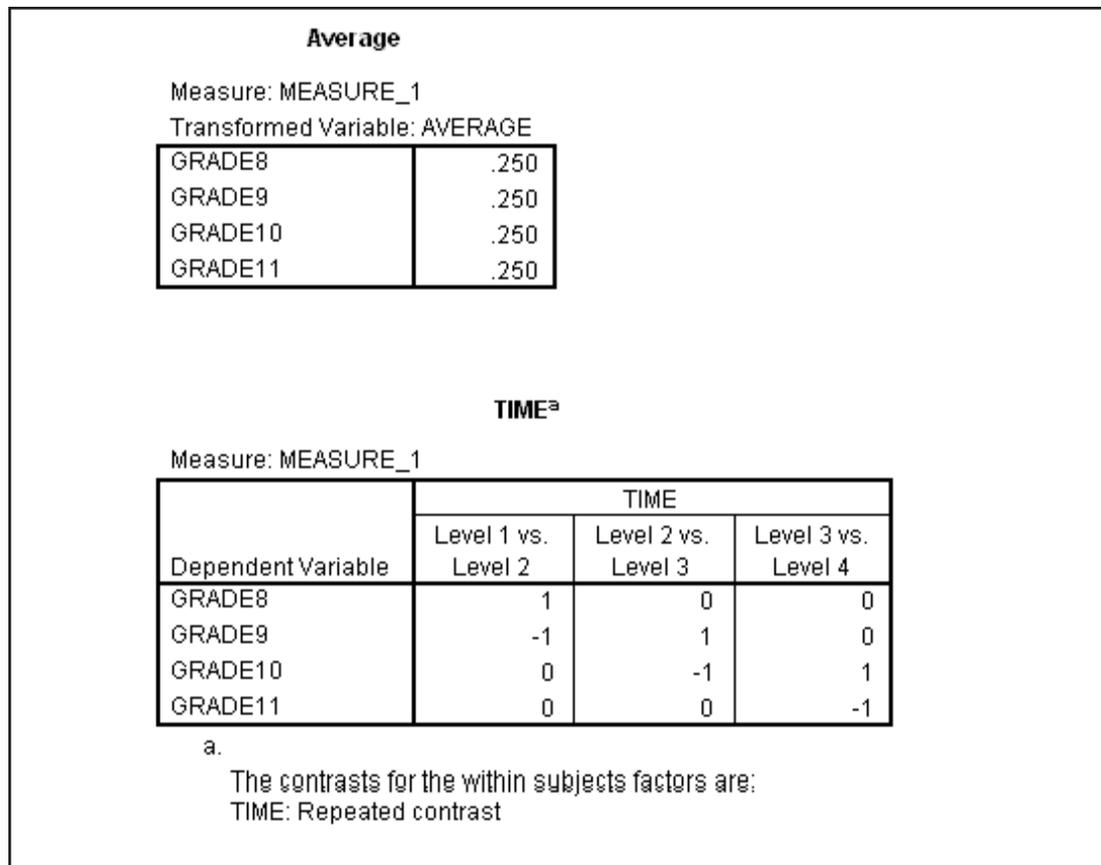
Measure: MEASURE\_1

Source	TIME	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Level 1 vs. Level 2	126.253	1	126.253	82.284	.000
	Level 2 vs. Level 3	12.763	1	12.763	6.530	.013
	Level 3 vs. Level 4	14.948	1	14.948	9.149	.004
Error(TIME)	Level 1 vs. Level 2	96.664	63	1.534		
	Level 2 vs. Level 3	123.142	63	1.955		
	Level 3 vs. Level 4	102.931	63	1.634		

We see that all three contrasts are significant at the .05 level. The first contrast, comparing grade 8 to grade 9 has by far the greatest F value. The second compares grade 9 to grade 10, and the third compares grade 10 to grade 11. These seem inconsistent with the pairwise tests we just ran in which the grade 9 scores were not different from the grade 10 scores. However, recall that we performed Bonferroni corrections on those tests and the second contrast (here with significance level of .013) would not be significant we testing at the adjusted Bonferroni level (about .008). If you return to the pairwise analysis you will see the results are quite close.

To confirm our understanding of the contrasts, we view the transformation matrix.

Figure 6.31 Transformation Matrix



We see the transformed variables do compare each grade level to the adjacent one. The transformed matrix is very useful in understanding and verifying which contrasts are being performed. These contrasts are not orthogonal, and would not be used without modification (orthonormalization) in the sphericity and pooled significance tests appearing earlier.

**SUMMARY**

In this chapter we reviewed how repeated measures ANOVA differs from between-group ANOVA and why it is used. Assumptions were discussed and an analysis was run based on student vocabulary scores measured over time. A second analysis applied planned comparisons (a priori contrasts) to a repeated measure analysis.

# Chapter 7 Between and Within-Subject ANOVA: (Split-Plot)

**Objective** In this chapter we will expand upon the last chapter to include both between and within-subject factors in one analysis. We will discuss the assumptions of this design and show an example. We will also explore the interactions using simple effects.

**Method** We will first use the Explore command to examine the data and then run the repeated measures ANOVA to do the basic mixed-model analysis (split-plot) and look at information regarding the assumptions.

**Data** The data set we will use is the same data as we used in Chapter 6, containing vocabulary test scores obtained from the same children over four years (grades 8 through 11). However, in this analysis we will use the sex of the subject as a between-subject factor.

**Technical Note** The term “mixed model” technically refers to ANOVA models containing fixed and random factors. The designs we discuss, where subject is a random effect, are a special case of the mixed model. The common usage of “mixed model” refers to designs with between and within-subject factors.

## INTRODUCTION

Many studies, especially experimental work, incorporate both between and within-subject factors. Within-subject factors will hopefully lead to a more sensitive analysis, while between-subject factors are necessary if any demographic characteristics are included or if there is reason to believe there would be strong carry-over effects. Mixed model refers to a mixture of between and within factors and is a direct generalization of the within-subjects analysis. These designs are also called split-plot designs, the term taken from agricultural experiments in which a given plot of land would receive single level of one treatment factor, but would be split into subplots that would receive all treatment levels of a second factor. This would yield between-plot and within-plot factors equivalent to the between and within-subject effects we have covered. We will discuss the features and assumptions of such analysis and run an example.

**ASSUMPTIONS  
OF MIXED  
MODEL ANOVA**

If we take the assumptions of within-subject analyses as a starting point, normality of the variables and sphericity when there are more than two levels of a within-subject factor, mixed model analyses involve little more. Since there are multiple groups, the normality of the variables now applies to the variation within each group. Also, homogeneity of covariance matrices is assumed (this can be applied to the original variables or the transformed variable – homogeneity of one implies homogeneity of the other). This combination of assumptions, homogeneity and sphericity, is sometimes called compound symmetry.

**PROPOSED  
ANALYSIS**

We will fit a model with one between-subject factor (Sex) and one within-subject factor (Time) with four levels. Thus for this analysis the sphericity issue is relevant and will be approached just as it was in Chapter 6. While we deal with a single between and a single within-subject factor, no additional assumptions are required to expand the analysis to handle multiple factors of each type.

**A LOOK AT THE  
DATA**

As before, we will use the Explore procedure to examine the distribution of vocabulary scores across grades and sex groups. Since we know from Chapter 6 that there are changes in vocabulary scores over time (grades), we will focus on the comparison of the two sex groups. This will provide some indication of normality and homogeneity of the vocabulary scores.

Click **File..Open..Data**

Move to the **c:\Train\Anova** directory

Select **SPSS Portable (.por)** from the Files of Type drop-down list

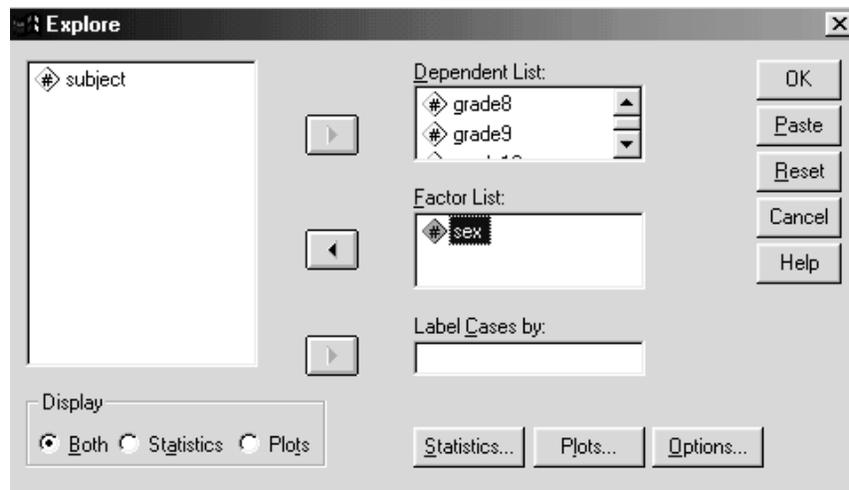
Double-click on **vocab**

**Figure 7.1 Data from Vocabulary Study**

	subject	grade8	grade9	grade10	grade11	sex	var
1	1.00	1.75	2.60	3.76	3.68	1	
2	2.00	.90	2.47	2.44	3.43	1	
3	3.00	.80	.93	.40	2.27	1	
4	4.00	2.42	4.15	4.56	4.21	1	
5	5.00	-1.31	-1.31	-.66	-2.22	1	
6	6.00	-1.56	1.67	.18	2.33	1	
7	7.00	1.09	1.50	.52	2.33	1	
8	8.00	-1.92	1.03	.50	3.04	1	
9	9.00	-1.61	.29	.73	3.24	1	
10	10.00	2.47	3.64	2.87	5.38	1	
11	11.00	-.95	.41	.21	1.82	1	

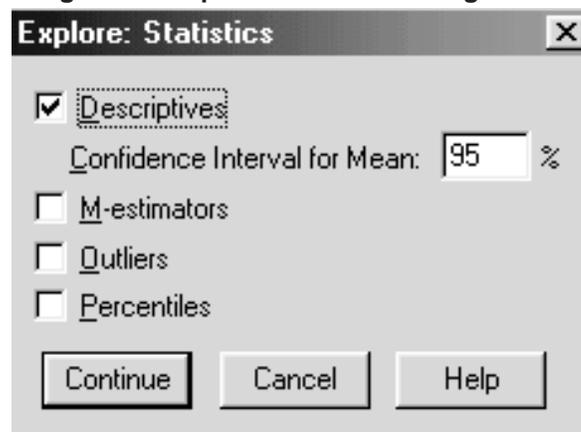
Click **Analyze..Descriptive Statistics..Explore**  
Move **Grade8, Grade9, Grade10, and Grade 11** into the  
Dependent list box.  
Move **Sex** into the Factors box

Figure 7.2 Explore Dialog Box



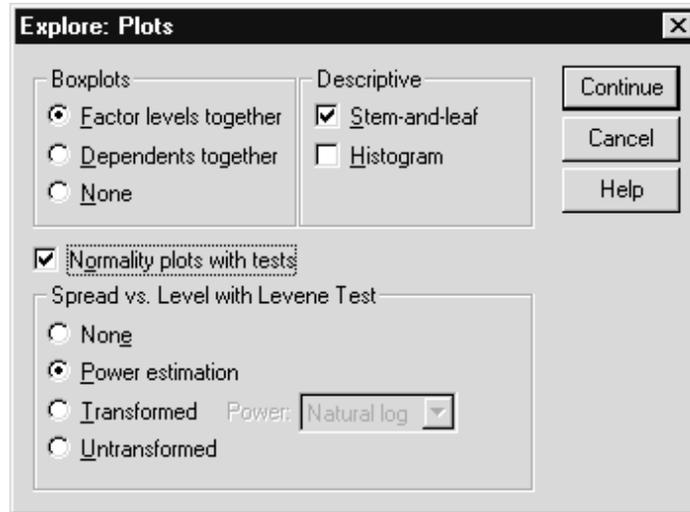
Click on the **Statistics** pushbutton  
Make sure that the **Descriptives** checkbox is the only one  
selected

Figure 7.3 Explore Statistics Dialog Box



Click on **Continue** to process the Statistics choices  
Click on the **Plots** pushbutton  
Verify that the **Factor levels together** option is selected.  
Verify that the **Stem-and-leaf** checkbox is checked  
Click **Normality plots with tests** checkbox  
Select **Power Estimation** option button in Spread vs. Level with  
Levene Test area

Figure 7.4 Explore Plots Dialog Box



Click on **Continue** to process the Plots choices  
Click on **OK** to run the EXPLORE procedure.

The syntax command below will run the analysis.

```
EXAMINE
  VARIABLES=grade8 grade9 grade10 grade11 BY sex
  /PLOT BOXPLOT STEMLEAF NPLOT SPREADLEVEL
  /COMPARE GROUP
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Normality tests and plots are generated by the Npplot keyword and homogeneity tests are due to the Spreadlevel keyword on the Plot subcommand.

Figure 7.5 Descriptives for Grade 8 Males

**Descriptives**

SEX		Statistic	Std. Error
GRADE8	Male	Mean	.8400
		95% Confidence Lower Bound	.2951
		Interval for Mean Upper Bound	1.3849
		5% Trimmed Mean	.8617
		Median	1.2850
		Variance	2.284
		Std. Deviation	1.5114
		Minimum	-1.92
		Maximum	3.30
		Range	5.22
		Interquartile Range	2.6850
		Skewness	-.454
		Kurtosis	-1.110
			.414
			.809

Figure 7.6 Stem and Leaf for Grade 8 Males

```

GRADE8 Stem-and-Leaf Plot for
SEX= Male

  Frequency      Stem & Leaf

    6.00      -1 . 134569
    3.00      -0 . 569
    5.00       0 . 11289
    8.00       1 . 02346679
    9.00       2 . 000012447
    1.00       3 . 3

Stem width:      1.00
Each leaf:       1 case(s)
    
```

Figure 7.7 Descriptives for Grade 8 Females

**Descriptives**

SEX			Statistic	Std. Error
GRADE8	Female	Mean	1.4344	.3867
		95% Confidence Interval for Mean		
		Lower Bound	.6457	
		Upper Bound	2.2230	
		5% Trimmed Mean	1.3049	
		Median	1.1700	
		Variance	4.785	
		Std. Deviation	2.1874	
		Minimum	-2.19	
		Maximum	8.26	
		Range	10.45	
		Interquartile Range	2.6475	
		Skewness	1.007	.414
		Kurtosis	2.075	.809

Figure 7.8 Stem and Leaf for Grade 8 Females

```

GRADE8 Stem-and-Leaf Plot for
SEX= Female

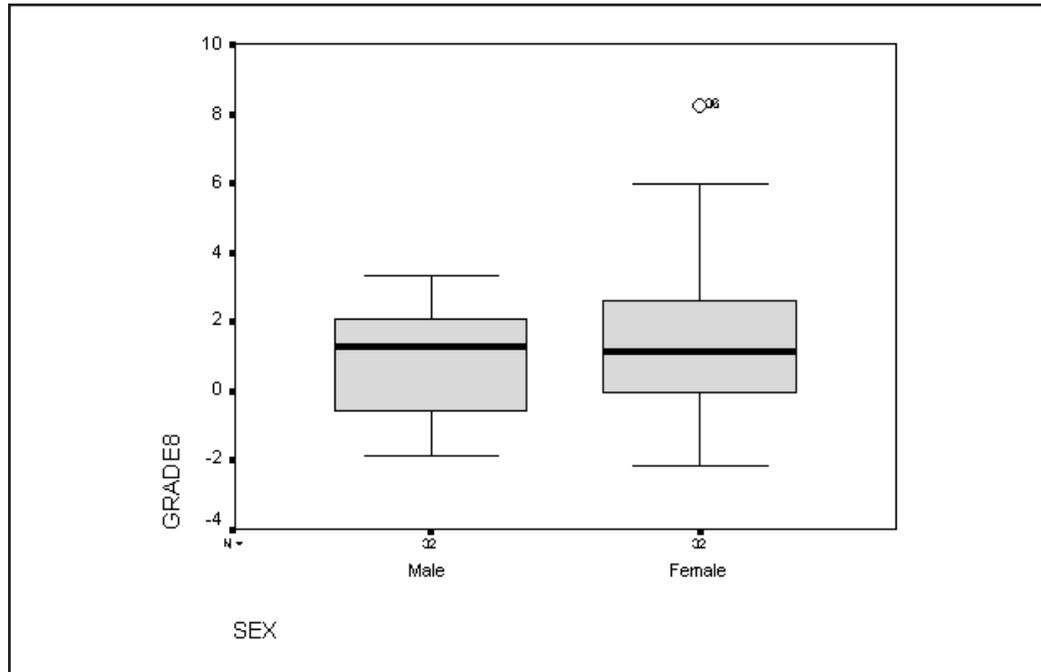
  Frequency      Stem & Leaf

    2.00      -2 . 11
    1.00      -1 . 2
    6.00      -0 . 000246
    5.00       0 . 11568
    6.00       1 . 012455
    6.00       2 . 112568
    3.00       3 . 125
    1.00       4 . 2
    1.00       5 . 9
    1.00 Extremes (>=8.3)

Stem width:      1.00
Each leaf:       1 case(s)
    
```

Notice that the range for 8<sup>th</sup> grade males is about half the range for females, but the interquartile ranges are about the same. This is due in part to an outlier among the females. While not shown, the vocabulary scores of the females were consistent with the normal distribution using the Shapiro-Wilks criterion, while the males showed a significant departure from normality.

Figure 7.9 Box Plots for Grade 8 Scores



The medians for the sex groups are very similar and the variation in the female group seems greater. Despite appearances in the plot, the 8<sup>th</sup> grade sex groups do not show significant differences in variation of test scores as evidenced by the Levene homogeneity test (not shown).

Figure 7.10 Descriptives for Grade 11 Males

Descriptives

SEX				Statistic	Std. Error
GRADE11	Male	95% Confidence Interval for Mean	Lower Bound	2.6451	
			Upper Bound	3.9581	
		5% Trimmed Mean		3.3396	
		Median		3.1900	
		Variance		3.316	
		Std. Deviation		1.8209	
		Minimum		-2.22	
		Maximum		7.46	
		Range		9.68	
		Interquartile Range		2.0675	
		Skewness		-.343	.414
		Kurtosis		1.900	.809

Figure 7.11 Stem and Leaf for Grade 11 Males

```

GRADE11 Stem-and-Leaf Plot for
SEX= Male

  Frequency    Stem & Leaf

    1.00 Extremes    (<=-2.2)
    4.00      1 .  0348
    9.00      2 .  112334677
    8.00      3 .  01234668
    4.00      4 .  0247
    4.00      5 .  2369
    1.00      6 .  0
    1.00 Extremes    (>=7.5)

Stem width:      1.00
Each leaf:      1 case(s)
    
```

Figure 7.12 Descriptives for Grade 11 Females

Descriptives

SEX		Statistic	Std. Error
GRADE11	Female	Mean	3.6416
		95% Confidence Interval for Mean	.3605
		Lower Bound	2.9063
		Upper Bound	4.3769
		5% Trimmed Mean	3.4838
		Median	3.3450
		Variance	4.159
		Std. Deviation	2.0394
		Minimum	.53
		Maximum	10.58
		Range	10.05
		Interquartile Range	2.3925
		Skewness	1.403
		Kurtosis	.414
			.809

Figure 7.13 Stem and Leaf for Grade 11 Females

```

GRADE11 Stem-and-Leaf Plot for
SEX= Female

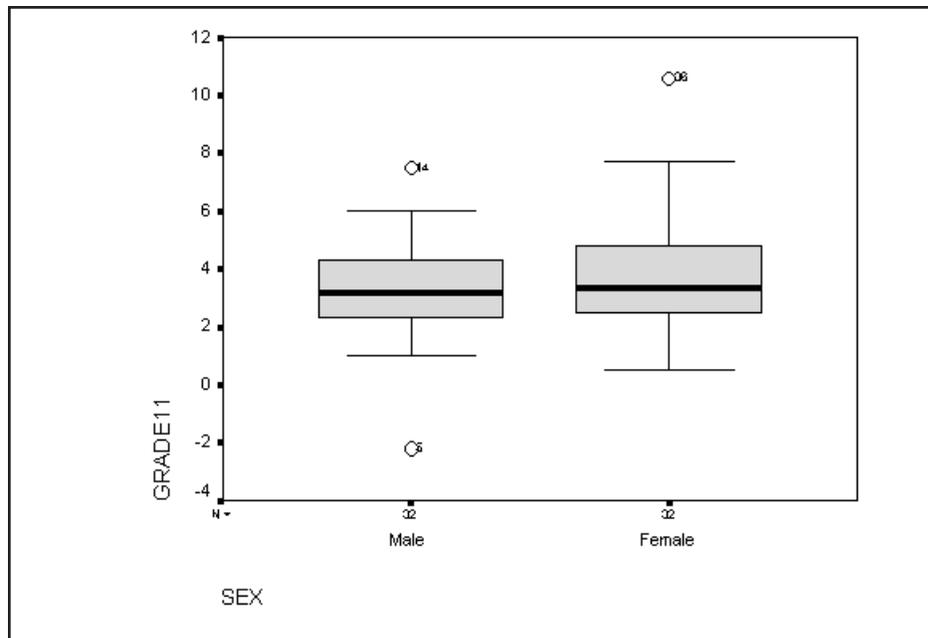
  Frequency    Stem & Leaf

    2.00      0 .  57
    3.00      1 .  178
   10.00      2 .  1345666689
    8.00      3 .  33557899
    3.00      4 .  889
    4.00      5 .  1689
    .00      6 .
    1.00      7 .  7
    1.00 Extremes    (>=10.6)

Stem width:      1.00
Each leaf:      1 case(s)
    
```

Most of the summary statistics are similar for males and females in the grade 11<sup>th</sup> grade. The normality tests (not shown) indicate that the scores for males, but not females, are consistent with normal distribution.

Figure 7.14 Box Plots for Grade 11



Once again, the medians are very close and there are a few outliers. The Levene test (not shown) indicated that the sex populations do not differ in variance on 11<sup>th</sup> grade vocabulary scores.

## 9<sup>th</sup> and 10<sup>th</sup> Grades

The distribution of vocabulary scores within sex group was consistent with the normal for 9<sup>th</sup> and 10<sup>th</sup> grade, the only exception being 10<sup>th</sup> grade females. Neither grade departed from homogeneity of variance between sex groups. (These results are not shown.)

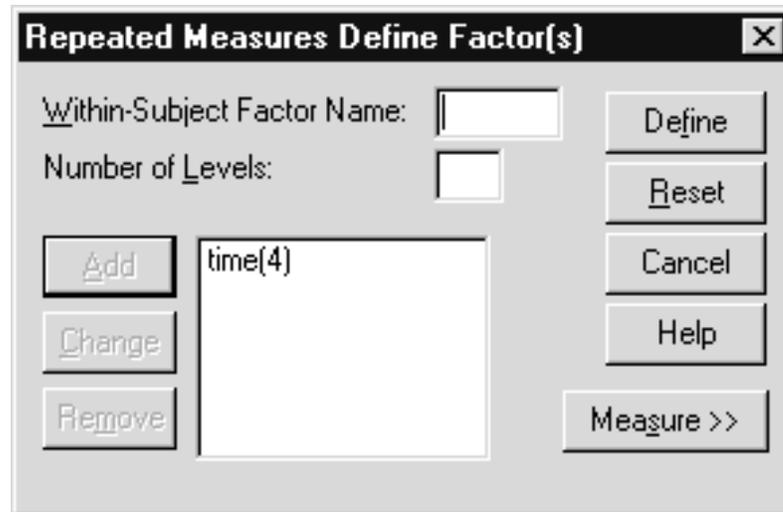
## SUMMARY OF EXPLORE

Overall, the data look good as far as homogeneity is concerned, and the departures from normality are not dramatic. If we had access to the original test sheets, we might want to check the accuracy of the scores for the outliers. We will proceed with the mixed-model ANOVA.

## SPLIT-PLOT ANALYSIS

Click **Analyze..General Linear Model..Repeated Measures**  
Replace **factor1** with **Time** in the **Within-Subject Factor Name** text box  
Press **Tab** and type **4** in the **Number of Levels** text box  
Click **Add** pushbutton

Figure 7.15 Define Factors Dialog Box

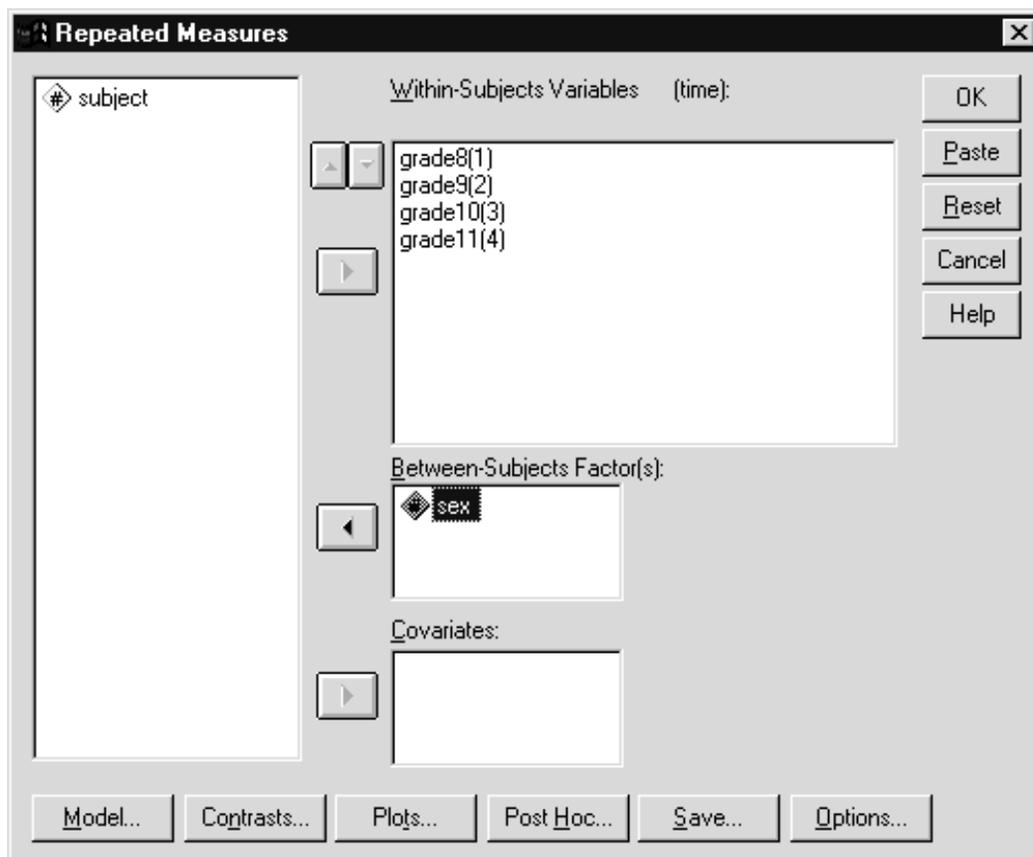


Click **Define** pushbutton

In the Repeated Measures dialog box, click and drag **Grade8**, **Grade9**, **Grade10**, and **Grade 11** to the **Within-Subject Variables** list box.

Move **Sex** into the **Between-Subjects Factors** list box

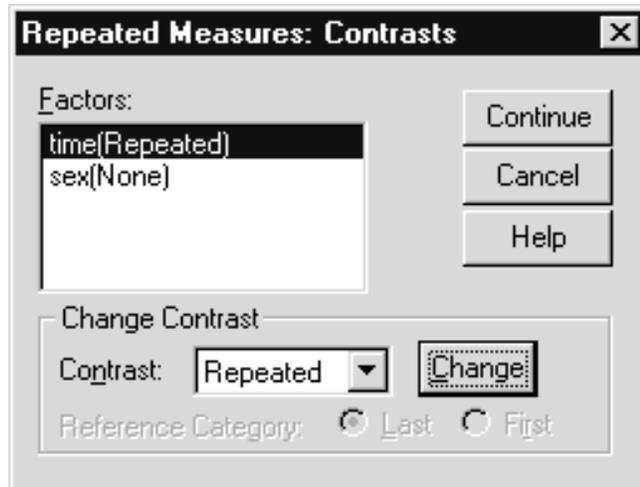
Figure 7.16 Between and Within-Subject Factors Defined



Click **Contrasts** pushbutton

- Select **time** in the Factors: list box
- Select **Repeated** from the **Contrast** drop-down list
- Click **Change** button
- Verify **Sex** is set to **none** for the contrast

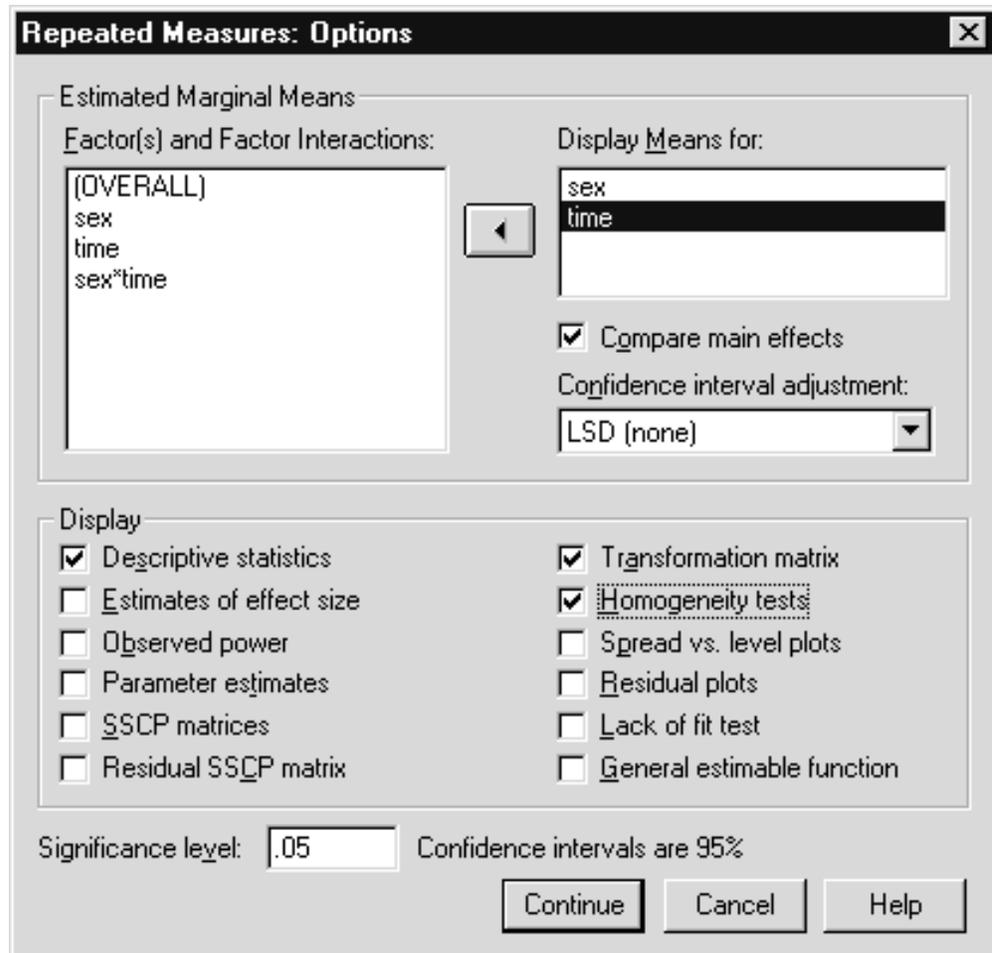
Figure 7.17 Contrasts Dialog Box



Repeated contrasts will compare each factor level with the one following it. Thus with four time level (8, 9, 10 and 11), the three repeated contrasts compare 8<sup>th</sup> to 9<sup>th</sup>, 9<sup>th</sup> to 10<sup>th</sup>, and 10<sup>th</sup> to 11<sup>th</sup> grades, respectively.

- Click **Continue** to process the Contrast changes
- Click **Options** pushbutton
- Individually move **Sex** and **Time** into the **Display Means for** list box
- Click the **Compare Main Effects** checkbox
- Click **Descriptive statistics**, **Transformation matrix**, and **Homogeneity tests** option buttons

Figure 7.18 Options Dialog Box



Click on **Continue** to process the Options requests  
 Click on **OK** to run the analysis

Besides descriptive statistics, we request estimated marginal means (which equal the observed means since we are fitting a full model) for each of the factors. Since there are several groups involved in the analysis, we ask for homogeneity of variance tests. We also request pairwise comparisons for sex and time with no adjustment (LSD (none)). Since sex has only two levels, pairwise tests are not needed.

We will proceed with the analysis. The GLM command shown below will produce this analysis (obtained by clicking the Dialog Recall tool



, then Repeated Measures, and the Paste pushbutton)

Figure 7.19 Syntax for This Analysis

```

GLM
  grade8 grade9 grade10 grade11 BY sex
  /WSFACTOR = time 4 Repeated
  /METHOD = SSTYPE(3)
  /EMMEANS = TABLES(sex) COMPARE ADJ(LSD)
  /EMMEANS = TABLES(time) COMPARE ADJ(LSD)
  /PRINT = DESCRIPTIVE TEST(MMATRIX) HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /WSDESIGN = time
  /DESIGN = sex .
  
```

The four vocabulary variables form the basis of the time factor. Estimated marginal means will be computed for the sex and time main effects. The Print subcommand requests that descriptive statistics, the transformation matrix (TEST(MMATRIX)) and homogeneity test summaries appear. The Wsdesign subcommand declares time as the only repeated measure factor in the model; similarly sex (see Design subcommand) is the only between-subject factor.

**EXAMINING RESULTS**

Figure 7.20 Factors in the Analysis

Within-Subjects Factors		
Measure: MEASURE_1		
TIME	Dependent Variable	
1	GRADE8	
2	GRADE9	
3	GRADE10	
4	GRADE11	

Between-Subjects Factors		
	Value Label	N
SEX 1	Male	32
2	Female	32

The factors in the analysis are listed along with the sample sizes for the between-subject factor.

Figure 7.21 Descriptive Statistics

**Descriptive Statistics**

	SEX	Mean	Std. Deviation	N
GRADE8	Male	.8400	1.5114	32
	Female	1.4344	2.1874	32
	Total	1.1372	1.8890	64
GRADE9	Male	1.9225	1.7449	32
	Female	3.1609	2.2356	32
	Total	2.5417	2.0849	64
GRADE10	Male	2.5472	2.0328	32
	Female	3.4294	2.2416	32
	Total	2.9883	2.1688	64
GRADE11	Male	3.3016	1.8209	32
	Female	3.6416	2.0394	32
	Total	3.4716	1.9255	64

Subgroup means appear separately for each of the repeated measure variables.

**TESTS OF ASSUMPTIONS**

Although they do not appear together in the output, we will first examine results pertaining to the assumptions of the analysis. Concerning homogeneity of variance, the program provides Box's M statistic and Levene's test. Box's M is a multivariate statistic testing whether the variance-covariance matrices composed of the four repeated measures variables are equal across the between-subject factor subgroup populations (multivariate homogeneity). Levene's test is univariate and tests homogeneity of variance for each of the four repeated measure variables separately (univariate homogeneity).

Figure 7.22 Box's M Test of Homogeneity

Box's Test of Equality of Covariance Matrices <sup>a</sup>	
Box's M	15.413
F	1.434
df1	10
df2	18378
Sig.	.158

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

a.

Design: Intercept+SEX  
Within Subjects Design: TIME

Box's M is not significant (significance value is .158), indicating that the data are consistent with the hypothesis of homogeneity of covariance matrices (based on the four repeated measures variables) across the population subgroups.

Figure 7.23 Levene's Test of Homogeneity

	F	df1	df2	Sig.
GRADE8	1.301	1	62	.258
GRADE9	1.105	1	62	.297
GRADE10	.030	1	62	.863
GRADE11	.183	1	62	.670

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

a.  
Design: Intercept+SEX  
Within Subjects Design: TIME

Not surprisingly, the results of Levene's test are consistent with Box's M. Box's M test has the advantage of being a single multivariate test. However, Box's M test is sensitive to both homogeneity and normality violations, while Levene's is relatively insensitive to lack of normality. Since homogeneity of variance violations are generally more problematic for ANOVA, Levene's test is useful.

## SPHERICITY

Since the within-subject factor (Time) has more than two levels, we will test for the sphericity assumption. As discussed in Chapter 6, if the assumption is met the usual averaged F tests are correct and are the test of choice. If sphericity conditions are not met, several choices are available: multivariate tests may be used, corrections to the averaged F test can be made (Greenhouse-Geisser, Huynh-Feldt, etc.), or more complicated decision rules may be applied (Looney & Stanley, 1989). We now view the sphericity test results.

Figure 7.24 Mauchly's Sphericity Test

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon <sup>a</sup>		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TIME	.900	6.367	5	.272	.942	1.000	.333

The Mauchly test shows no evidence of sphericity violations and the Greenhouse-Geisser and Huynh-Feldt degree of freedom adjustments are close to or equal to one. This result indicates we can proceed directly to

the averaged F tests for effects involving Time. However for comparison purposes, we will also view the multivariate tests.

**MULTIVARIATE TESTS INVOLVING TIME**

Figure 7.25 Multivariate Tests

Multivariate Tests <sup>b</sup>						
Effect		Value	F	Hypothesis df	Error df	Sig.
TIME	Pillai's Trace	.826	95.139 <sup>a</sup>	3.000	60.000	.000
	Wilks' Lambda	.174	95.139 <sup>a</sup>	3.000	60.000	.000
	Hotelling's Trace	4.757	95.139 <sup>a</sup>	3.000	60.000	.000
	Roy's Largest Root	4.757	95.139 <sup>a</sup>	3.000	60.000	.000
TIME * SEX	Pillai's Trace	.135	3.109 <sup>a</sup>	3.000	60.000	.033
	Wilks' Lambda	.865	3.109 <sup>a</sup>	3.000	60.000	.033
	Hotelling's Trace	.155	3.109 <sup>a</sup>	3.000	60.000	.033
	Roy's Largest Root	.155	3.109 <sup>a</sup>	3.000	60.000	.033

a. Exact statistic  
 b. Design: Intercept+SEX  
 Within Subjects Design: TIME

As expected from the analysis in Chapter 6, there are significant differences in vocabulary scores over time. In addition, there is a significant interaction between Sex and Time. This can be phrased in two ways; the population sex difference is not uniform across grades, or the trend over time is not identical for the two sex populations.

**TESTS OF BETWEEN-SUBJECT FACTORS**

Figure 7.26 Between-Subjects Tests

Tests of Between-Subjects Effects					
Measure: MEASURE_1					
Transformed Variable: Average					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	411.177	1	411.177	121.936	.000
SEX	9.333	1	9.333	2.768	.101
Error	209.068	62	3.372		

The Repeated Measures procedure also presents the tests for the between-subject factors, in this case Sex. There is no significant difference in overall vocabulary score between the females and males (significance value is .101).

**AVERAGED F TESTS INVOLVING TIME**

**Figure 7.27 F Tests**

**Tests of Within-Subjects Effects**

Measure: MEASURE\_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	194.338	3	64.779	81.533	.000
	Greenhouse-Geisser	194.338	2.825	68.781	81.533	.000
	Huynh-Feldt	194.338	3.000	64.779	81.533	.000
	Lower-bound	194.338	1.000	194.338	81.533	.000
TIME * SEX	Sphericity Assumed	7.162	3	2.387	3.005	.032
	Greenhouse-Geisser	7.162	2.825	2.535	3.005	.035
	Huynh-Feldt	7.162	3.000	2.387	3.005	.032
	Lower-bound	7.162	1.000	7.162	3.005	.088
Error(TIME)	Sphericity Assumed	147.780	186	.795		
	Greenhouse-Geisser	147.780	175.179	.844		
	Huynh-Feldt	147.780	186.000	.795		
	Lower-bound	147.780	62.000	2.384		

The averaged F tests indicate a significant effect of time and a sex by time interaction. Here we view only the test results labeled “sphericity assumed” since the sphericity assumption was met.

**Figure 7.28 Repeated Measures Contrasts**

**Tests of Within-Subjects Contrasts**

Measure: MEASURE\_1

Source	TIME	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Level 1 vs. Level 2	126.253	1	126.253	86.948	.000
	Level 2 vs. Level 3	12.763	1	12.763	6.534	.013
	Level 3 vs. Level 4	14.948	1	14.948	9.435	.003
TIME * SEX	Level 1 vs. Level 2	6.637	1	6.637	4.571	.036
	Level 2 vs. Level 3	2.031	1	2.031	1.040	.312
	Level 3 vs. Level 4	4.703	1	4.703	2.969	.090
Error(TIME)	Level 1 vs. Level 2	90.027	62	1.452		
	Level 2 vs. Level 3	121.111	62	1.953		
	Level 3 vs. Level 4	98.227	62	1.584		

As we saw in Chapter 6, the contrasts show that there are significant differences between each pair of grades on the vocabulary scores. However, the only significant sex by time interaction term involves grades 8 and 9. Thus the interaction between sex and time centers on these two grades. The means involving both sex and time (see Figure 7.21) can be examined for more detail.

Figure 7.29 Transformation Matrix

**Average**

Measure: MEASURE\_1  
Transformed Variable: AVERAGE

GRADE8	.250
GRADE9	.250
GRADE10	.250
GRADE11	.250

**TIME<sup>a</sup>**

Measure: MEASURE\_1

Dependent Variable	TIME		
	Level 1 vs. Level 2	Level 2 vs. Level 3	Level 3 vs. Level 4
GRADE8	1	0	0
GRADE9	-1	1	0
GRADE10	0	-1	1
GRADE11	0	0	-1

a.  
The contrasts for the within subjects factors are:  
TIME: Repeated contrast

This is shown only to verify that the repeated contrasts were used.

Figure 7.30 Pairwise Comparisons Involving Time

**Pairwise Comparisons**

Measure: MEASURE\_1

(I) TIME	(J) TIME	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	95% Confidence Interval for Difference <sup>a</sup>	
					Lower Bound	Upper Bound
1	2	-1.405*	.151	.000	-1.706	-1.103
	3	-1.851*	.135	.000	-2.121	-1.581
	4	-2.334*	.157	.000	-2.648	-2.021
2	1	1.405*	.151	.000	1.103	1.706
	3	-.447*	.175	.013	-.796	-.9733E-02
	4	-.930*	.168	.000	-1.265	-.595
3	1	1.851*	.135	.000	1.581	2.121
	2	.447*	.175	.013	9.733E-02	.796
	4	-.483*	.157	.003	-.798	-.169
4	1	2.334*	.157	.000	2.021	2.648
	2	.930*	.168	.000	.595	1.265
	3	.483*	.157	.003	.169	.798

Based on estimated marginal means  
\*. The mean difference is significant at the .05 level.  
a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Pairwise comparisons appear for both sex and time. Since there are only two sex groups, the pairwise comparisons tell us no more than the overall main effect, and so are not of interest (not shown). The pairwise

comparisons involving time (with no adjustment due to the number of tests performed) are all significant.

### ADDITIONAL WITHIN-SUBJECT FACTORS AND SPHERICITY

The sphericity assumption applies to all within-subject factors with more than two levels. In such designs Repeated Measures will perform sphericity tests for the appropriate within-subject factors and relevant interactions (effects involving within-subject interactions). The approach taken above applies to these situations as well.

### EXPLORING THE INTERACTION - SIMPLE EFFECTS

A technique that can be used to explore interactions involves simple effects, that is, looking at the differences in one factor within a single level of a second factor. For example, an interaction might be clarified by a factor showing a significant difference at one level of a second factor while showing no difference at a second level. Below we run simple effects examining sex differences within each grade and also examine time differences with each sex group. Typically, both analyses would not be run, but we wish to demonstrate how to set them up.

Within the SPSS Univariate (Unianova) or Repeated Measures (GLM) procedures, the method to obtain simple effects involves requesting the estimated marginal means table for the two factors involved, and then obtaining tests on the factor of interest, applied to the means table. For example, if we want tests performed on the time factor within each sex group, we need to request tests on the time factor, based on the sex-by-time table of estimated marginal means. Currently, this analysis cannot be run directly from the Univariate and Repeated Measures dialog boxes, but involves only a minor change to syntax pasted from the dialogs. To demonstrate, we first return to the Repeated Measures dialog box.

Click the Dialog Recall tool , and then click **Repeated**

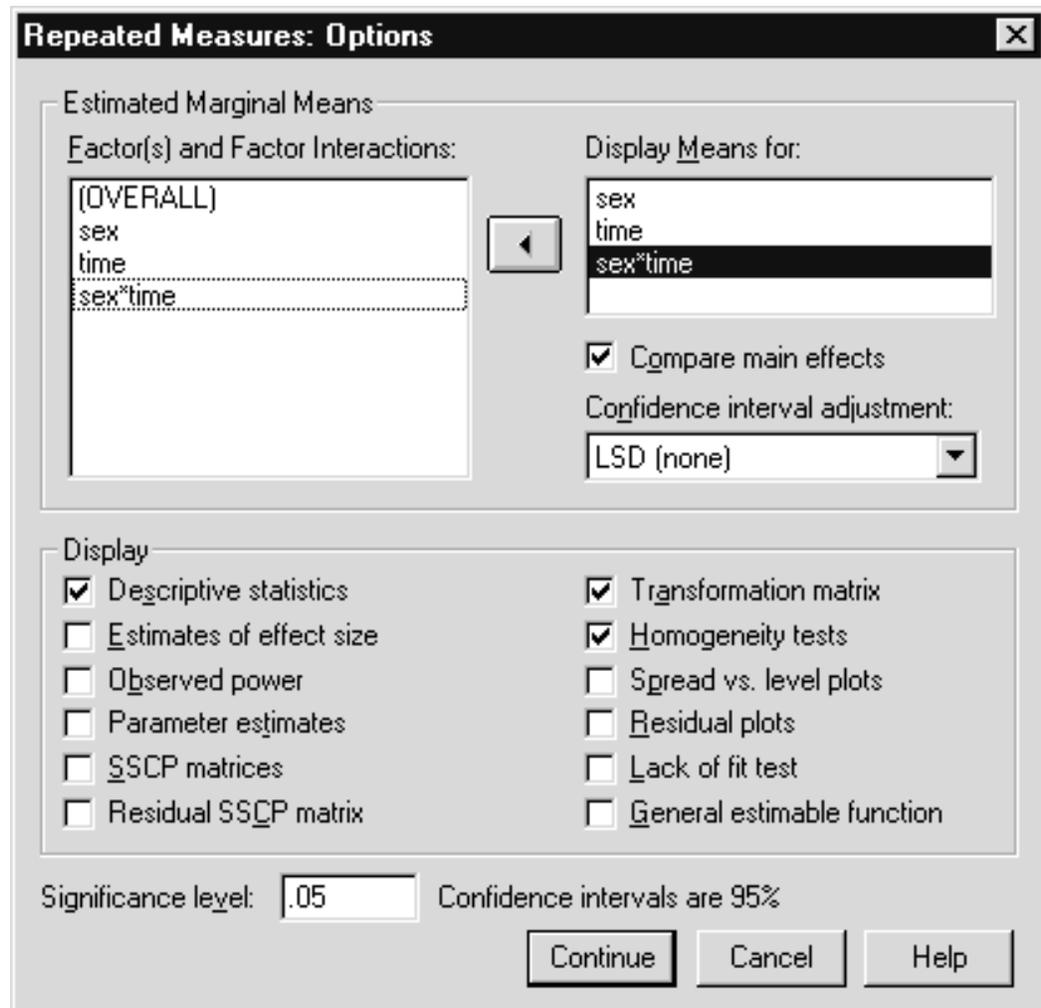
#### **Measures**

Click the **Define** pushbutton

Click the **Options** pushbutton

Move **sex\*time** into the **Display Means for** list box

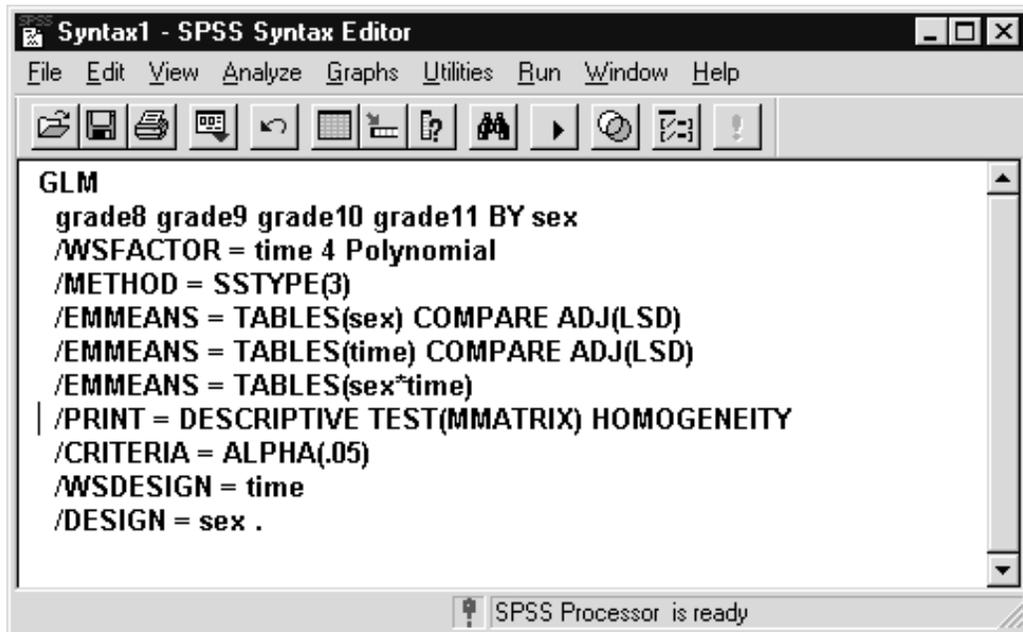
Figure 7.31 Requesting the Sex by Time Table



We have requested estimated margin means for the sex\*time table and must later indicate the tests we want performed. We could have dropped the means display and main effects comparison for the individual factors, sex and time, but will keep them to better illustrate the changes we must make concerning the sex\*time table.

Click **Continue**  
 Click **Paste** pushbutton

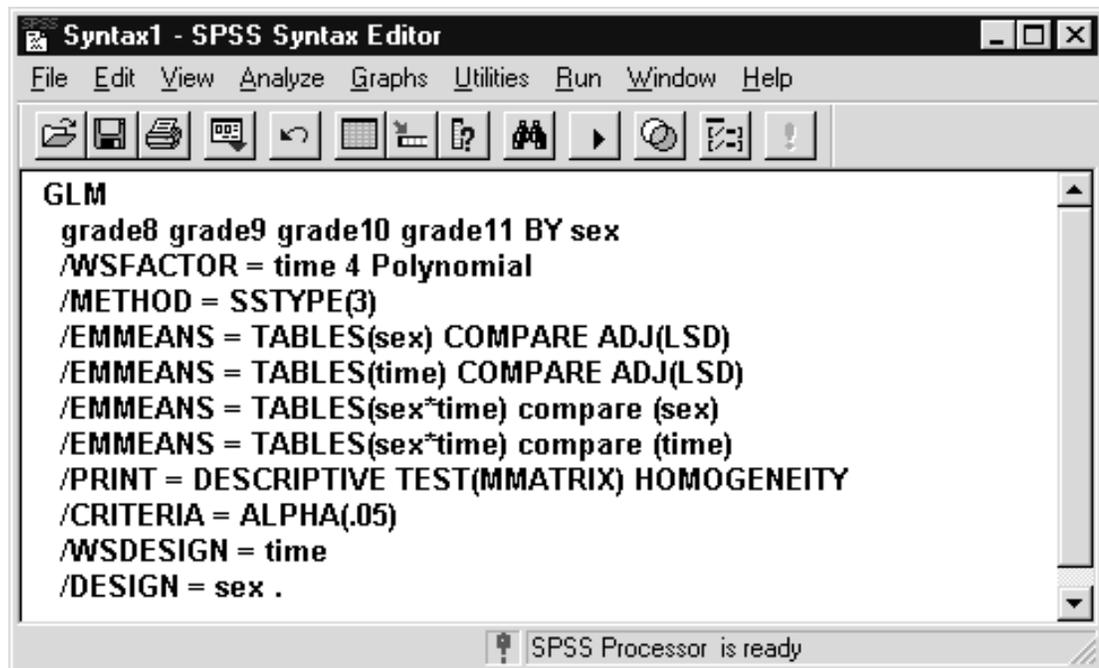
Figure 7.32 Syntax for Repeated Measures Analysis



There are three EMMEANS (estimated marginal means) subcommands. The first two, involving the sex, and time tables, contain the COMPARE keyword. It requests that main- or simple main-effect tests (depending on how many factors are specified under TABLES) and pairwise comparisons be performed. Pairwise tests can be adjusted using Bonferroni or Sidak adjustments, but, by default, no adjustment (LSD) is made. Our task is to obtain these tests for time within the sex\*time table. We must add the COMPARE keyword referencing one of the factors to the /EMMEANS subcommand that contains the sex\*time table.

- Type **COMPARE (SEX)** at the end of the /EMMEANS = TABLES (sex\*time) line
- Copy and paste** the modified /EMMEANS = TABLES (sex\*time) line just below the original
- Change **COMPARE (SEX)** to **COMPARE (TIME)** in the second /EMMEANS = TABLES (sex\*time) subcommand

Figure 7.33 Syntax Requesting Tests for Simple Effects



Now significance tests will be applied to the sex factor and then to the time factor, each performed within every level of the other factor, based on the sex-by-time table of estimated marginal means. Since the table involves more than one factor, the tests will be run separately at each level of the other factor(s). This logic can be extended to additional factors, so you can perform simple effect tests on one factor within a table involving more than two factors.

Note that we do not need the estimated marginal means and tests for the individual factors sex and time (/EMMEANS subcommands containing TABLES(SEX) and TABLES(TIME)) and these subcommands could be removed. They were left in the "Display means for" list box in the Repeated Measures Options dialog so we could see the syntax needed to request the tests.

Click **Run..Current** to run the analysis

Most of the results are identical to those viewed earlier. Here we focus on the summaries involving simple effects.

Scroll down to the **first Sex \* Time** section under Estimated Marginal Means heading

Figure 7.34 Estimated Marginal Means

Estimates

Measure: MEASURE\_1

SEX	TIME	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Male	1	.840	.332	.176	1.504
	2	1.922	.354	1.214	2.631
	3	2.547	.378	1.791	3.303
	4	3.302	.342	2.618	3.985
Female	1	1.434	.332	.770	2.099
	2	3.161	.354	2.452	3.870
	3	3.429	.378	2.673	4.186
	4	3.642	.342	2.958	4.325

The simple effects analysis will be based on this table. Even if a more complex model were being analyzed, say with three or four between-subject factors, then the simple effects analysis of a two-factor interaction, would be based on the estimated means table involving the two factors of interest.

Figure 7.35 Pairwise Comparisons of Sex within Grade Levels (Time)

Pairwise Comparisons

Measure: MEASURE\_1

TIME	(I) SEX	(J) SEX	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	95% Confidence Interval for Difference <sup>a</sup>	
						Lower Bound	Upper Bound
1	Male	Female	-.594	.470	.211	-1.534	.345
	Female	Male	.594	.470	.211	-.345	1.534
2	Male	Female	-1.238*	.501	.016	-2.241	-.236
	Female	Male	1.238*	.501	.016	.236	2.241
3	Male	Female	-.882	.535	.104	-1.952	.187
	Female	Male	.882	.535	.104	-.187	1.952
4	Male	Female	-.340	.483	.484	-1.306	.626
	Female	Male	.340	.483	.484	-.626	1.306

Based on estimated marginal means

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

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Recall that the COMPARE keyword on the /EMMEANS subcommand will produce both overall tests and pairwise comparisons for the specified factor. The Pairwise Comparisons table presents for each level of the time factor (which represent grades 8, 9, 10 and 11), the female-male

comparison. Because there are only two levels to sex, there is only one unique comparison at each grade level. However, for factors with more than two levels, all pairwise comparisons would appear.

Examining the comparisons, we see that only at time 2 (9<sup>th</sup> grade) was there a significant difference between males and females. Thus we can describe the nature of the sex by time interaction: there are no significant differences between males and females in vocabulary scores except in the 9<sup>th</sup> grade.

Figure 7.36 Univariate Tests (Simple Effects)

**Univariate Tests**

Measure: MEASURE\_1

TIME		Sum of Squares	df	Mean Square	F	Sig.
1	Contrast	5.653	1	5.653	1.599	.211
	Error	219.148	62	3.535		
2	Contrast	24.540	1	24.540	6.103	.016
	Error	249.310	62	4.021		
3	Contrast	12.452	1	12.452	2.720	.104
	Error	283.869	62	4.579		
4	Contrast	1.850	1	1.850	.495	.484
	Error	231.724	62	3.737		

Each F tests the simple effects of SEX within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

In addition to the simple pairwise comparisons, overall simple effect tests are presented. As the caption indicates, each F test represents a test of the simple effect of sex within a grade level (time). Since sex has only two levels, these tests match the pairwise results viewed above, which were already discussed. However, for a factor with more than two levels, this summary would present an overall test of a factor within each level of the second factor.

Now we examine the interaction question using simple effects of time within each sex group.

Scroll down to the **second Sex \* Time** section under Estimated Marginal Means heading

Figure 7.37 Pairwise Comparisons for Time Performed Separately for Males and Females (Complete Table Not Shown)

Pairwise Comparisons

Measure: MEASURE\_1

SEX	(I) TIME	(J) TIME	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	95% Confidence Interval for Difference <sup>a</sup>	
						Lower Bound	Upper Bound
Male	1	2	-1.082*	.213	.000	-1.663	-.502
		3	-1.707*	.191	.000	-2.228	-1.187
		4	-2.462*	.222	.000	-3.067	-1.856
	2	1	1.082*	.213	.000	.502	1.663
		3	-.625	.247	.084	-1.298	4.872E-02
		4	-1.379*	.237	.000	-2.025	-.733
	3	1	1.707*	.191	.000	1.187	2.228
		2	.625	.247	.084	-4.872E-02	1.298
		4	-.754*	.223	.007	-1.361	-.148
	4	1	2.462*	.222	.000	1.856	3.067
		2	1.379*	.237	.000	.733	2.025
		3	.754*	.223	.007	.148	1.361
Female	1	2	-1.727*	.213	.000	-2.307	-1.146
		3	-1.995*	.191	.000	-2.516	-1.474
		4	-2.207*	.222	.000	-2.812	-1.602
	2	1	1.727*	.213	.000	1.146	2.307
		3	-.268	.247	1.000	-.942	.405
		4	-.481	.237	.282	-1.127	.166
	3	1	1.995*	.191	.000	1.474	2.516
		2	.268	.247	1.000	-.405	.942
		4	-.212	.223	1.000	-.819	.394
	4	1	2.207*	.222	.000	1.602	2.812
		2	.481	.237	.282	-.166	1.127

Multiple comparisons of the grade levels are done separately for females and the males (levels of the sex factor). A caption appears at the below the table (not shown) indicating that the comparisons are based on the estimated marginal means. All grade (time) comparisons are significant for the males, while all but two (9<sup>th</sup> versus 10<sup>th</sup>, and 10<sup>th</sup> versus 11<sup>th</sup>) are significant for females. Thus males show a significantly increase in vocabulary scores at each grade level, while females didn't show significant change from 9<sup>th</sup> to 10<sup>th</sup> or 10<sup>th</sup> to 11<sup>th</sup> grades. Females and males thus show different patterns of vocabulary change over time, which is the basis of the interaction.

In this way, understanding of a two-way interaction can be improved by examining the simple effects of either factor. A plot is also helpful (shown later).

Figure 7.38 Multivariate Tests (Simple Effects)

		Multivariate Tests				
SEX		Value	F	Hypothesis df	Error df	Sig.
Male	Pillai's trace	.700	46.607 <sup>a</sup>	3.000	60.000	.000
	Wilks' lambda	.300	46.607 <sup>a</sup>	3.000	60.000	.000
	Hotelling's trace	2.330	46.607 <sup>a</sup>	3.000	60.000	.000
	Roy's largest root	2.330	46.607 <sup>a</sup>	3.000	60.000	.000
Female	Pillai's trace	.721	51.641 <sup>a</sup>	3.000	60.000	.000
	Wilks' lambda	.279	51.641 <sup>a</sup>	3.000	60.000	.000
	Hotelling's trace	2.582	51.641 <sup>a</sup>	3.000	60.000	.000
	Roy's largest root	2.582	51.641 <sup>a</sup>	3.000	60.000	.000

Each F tests the multivariate simple effects of TIME within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

When examining the simple effects of sex within grade, overall F tests were presented in the “Univariate Tests” table. This will be the case for simple effects of any between-subject factor. Multivariate tests are used to perform overall tests of simple effects for within-subject factors. Multivariate tests are used to avoid complications that would occur if Bonferroni or Sidak corrections were requested and sphericity were violated. However, this means that when the sphericity assumption holds, which is the case here, the simple effects test used (multivariate test) is not the most powerful test. We find that there are significant overall grade differences in vocabulary scores for both males and females. In this instance the overall test sheds less light on the nature of the interaction than did the pairwise comparisons. For this reason, it is useful to examine both results.

## GRAPHING THE INTERACTION

Profile plots, seen earlier in the course, provide a means of visualizing a two- or three-factor interaction. We will request a plot of vocabulary means for the grade and sex groups.

Click the Dialog Recall tool , and then click **Repeated**

### Measures

Click the **Define** pushbutton

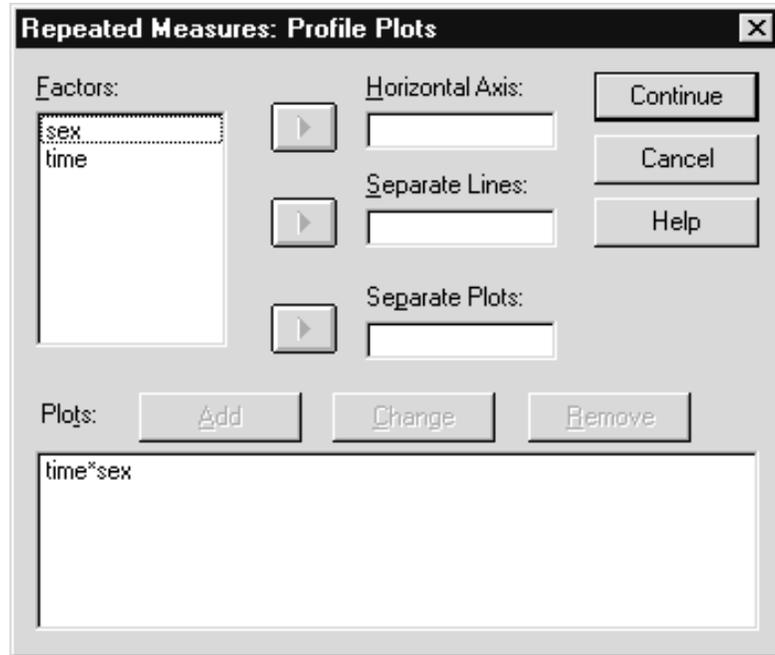
Click **Plots** pushbutton

Move **time** into the **Horizontal Axis** box

Move **sex** into the **Separate Lines** box

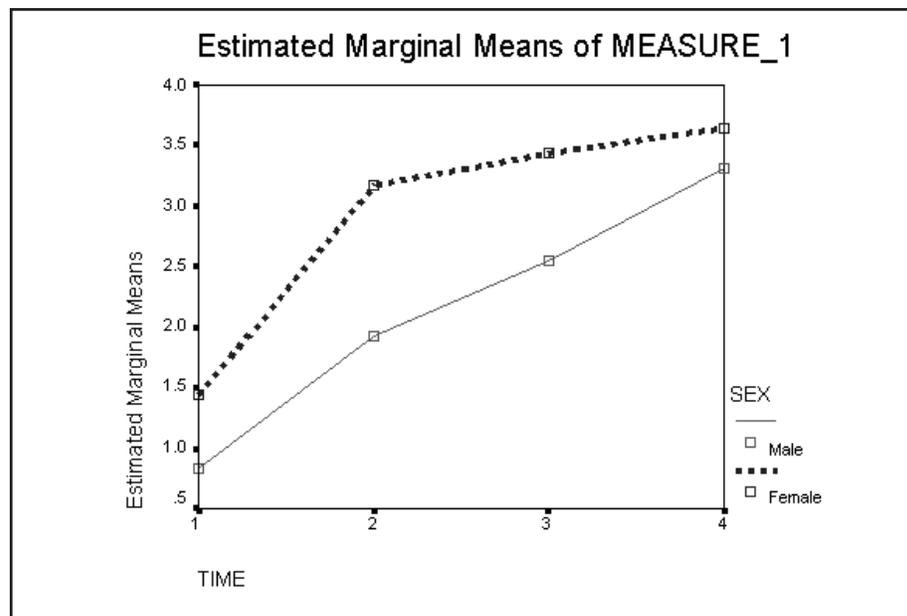
Click **Add** pushbutton

Figure 7.39 Requesting a Profile Plot



Click **Continue**, and then **OK**

Figure 7.40 Profile Plot of Vocabulary Scores Across Grades for Males and Females



The plot of the estimated marginal means (here identical to the observed means) shows the steady increase in vocabulary score over time for the males. In comparison, the females show a sharper increase from grade 8 to grade 9, and more gradual increases in the later grades. The greatest difference between males and females occurs in grade 9. These patterns, as you would expect, are consistent with the simple effects tests we performed earlier.

## Chapter 8 More Split-Plot Design

<b>Objective</b>	Understand the issues involved with more complex split-plot analyses.
<b>Method</b>	Use GLM to run a between- and within-subject analysis (split-plot) involving multiple between- and multiple within-subject factors.
<b>Data</b>	A marketing study in which different groups of subjects (groups based on sex and current brand used) rated different brands before and after viewing a commercial. The aim of the analysis was to determine if ratings improved for a specific brand and whether this related to sex or brand used.

### INTRODUCTION: AD VIEWING WITH PRE-POST BRAND RATINGS

The example in this chapter will involve a more complex analysis, but will be done with fewer variations. A marketing experiment was devised to evaluate whether viewing a commercial produces improved ratings for a specific brand. Ratings on three brands (on a 1 to 10 scale, where 10 is the highest rating) were obtained from subjects before and after viewing the commercial. Since the hope was that the commercial would improve ratings of only one brand (A), researchers expected a significant brand by pre-post commercial interaction (only brand A ratings would change). In addition, there were two between-group factors: sex and brand used by subject. Thus the study had four factors overall: sex, brand used, brand rated, and pre-post commercial. We view the data below.

## SETTING UP THE ANALYSIS

Click **File..Open..Data** (move to the c:\Train\Anova directory if necessary)  
Click **SPSS Portable(\*.por)** in the Files of Type drop-down list  
Double-click on **brand**

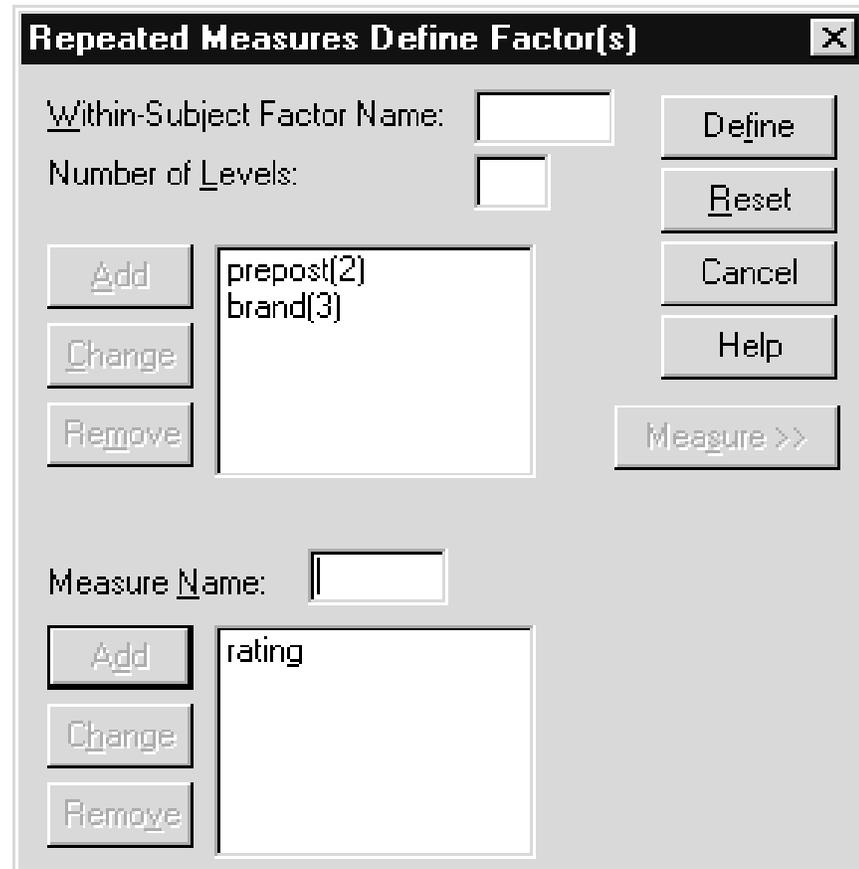
Figure 8.1 Data from the Brand Study

	id	sex	user	pre_a	pre_b	pre_c	post_a	post_b	post_c	var	var
1	1	1	1	7	7	5	9	7	6		
2	2	1	1	4	2	3	6	4	2		
3	3	1	1	4	5	1	4	3	3		
4	4	1	1	5	4	4	6	5	5		
5	5	1	1	7	3	6	6	5	5		
6	6	1	1	3	2	4	4	4	3		
7	7	1	1	4	6	3	6	4	4		
8	8	1	1	4	4	2	5	4	3		
9	9	1	1	7	5	4	7	5	4		
10	10	1	1	5	2	4	5	5	6		
11	11	1	1	5	5	3	5	3	4		

Sex and user are the between-subject factors. The next six variables pre\_a to post\_c contain the three brand ratings before and after viewing the commercial.

Click **Analyze..General Linear Model..Repeated Measures**  
Replace **factor1** with **prepost** in the **Within-Subject Factor Name** text box  
Press **Tab** and type **2** in the **Number of Levels** text box  
Click **Add** pushbutton  
Type **brand** in the **Within-Subject Factor Name** text box  
Press **Tab** and type **3** in the **Number of Levels** text box  
Click the **Add** pushbutton  
Click the **Measure** pushbutton  
Type **rating** in the **Measure Name** text box  
Click the **Add** pushbutton in the **Measure Name** area

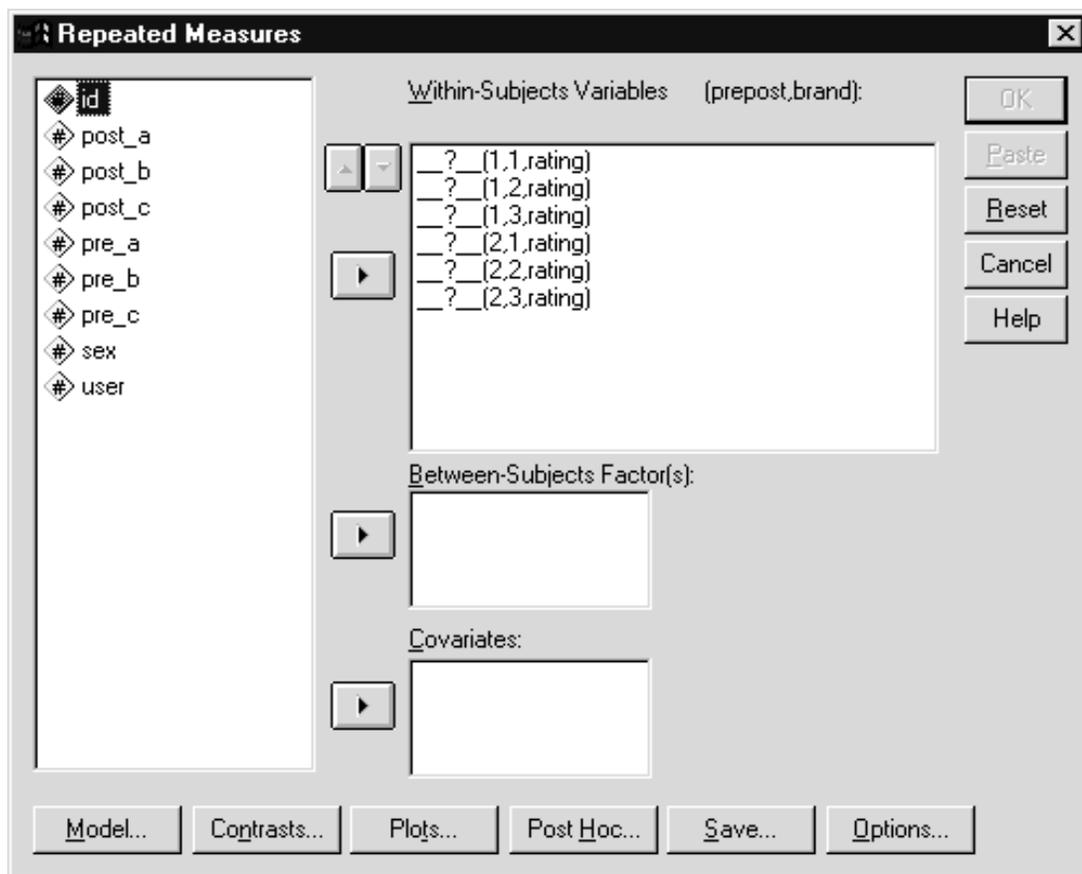
Figure 8.2 Two Within-Subject Factors Declared



SPSS now expects variables that comprise two within-subject factors. The order you name the factors only matters in that SPSS will order the factor levels list in the next dialog so that the last factor named here has its levels change most rapidly. Therefore depending on how your variables are ordered in the data, some factor orders make the later declarations easier.

Click the **Define** pushbutton

Figure 8.3 Repeated Measures Dialog with Two Factors

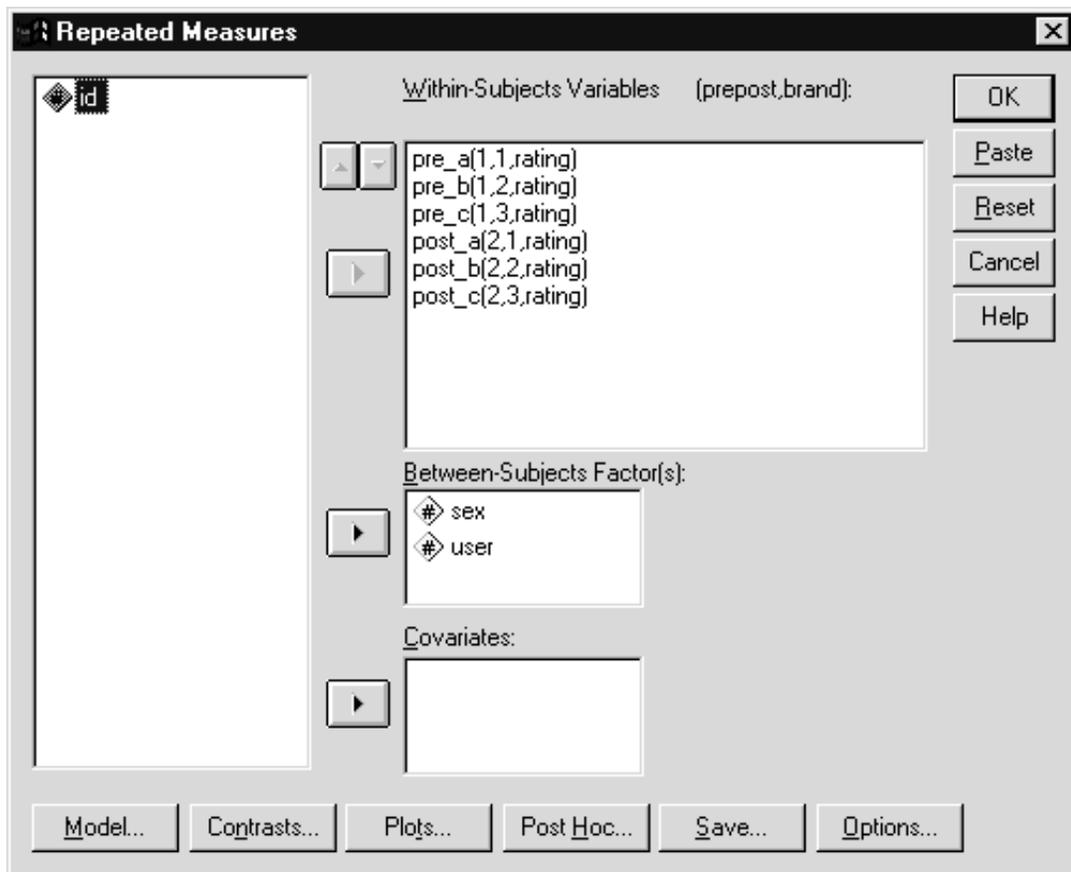


Both prepost and brand are listed as within-subject factors. There are six rows, so every possible combination of levels between the two factors is represented. Notice that the brand level changes first going down the list. This was due to defining brand last in the Repeated Measures Define Factor dialog. Defining the factors in an order consistent with the order of variables in your data file makes this step easier.

Move the following variables into the **Within-Subjects Variables** list in the order given: **pre\_a pre\_b pre\_c post\_a post\_b post\_c**

Move **sex** and **user** into the **Between-Subjects Factors** list box

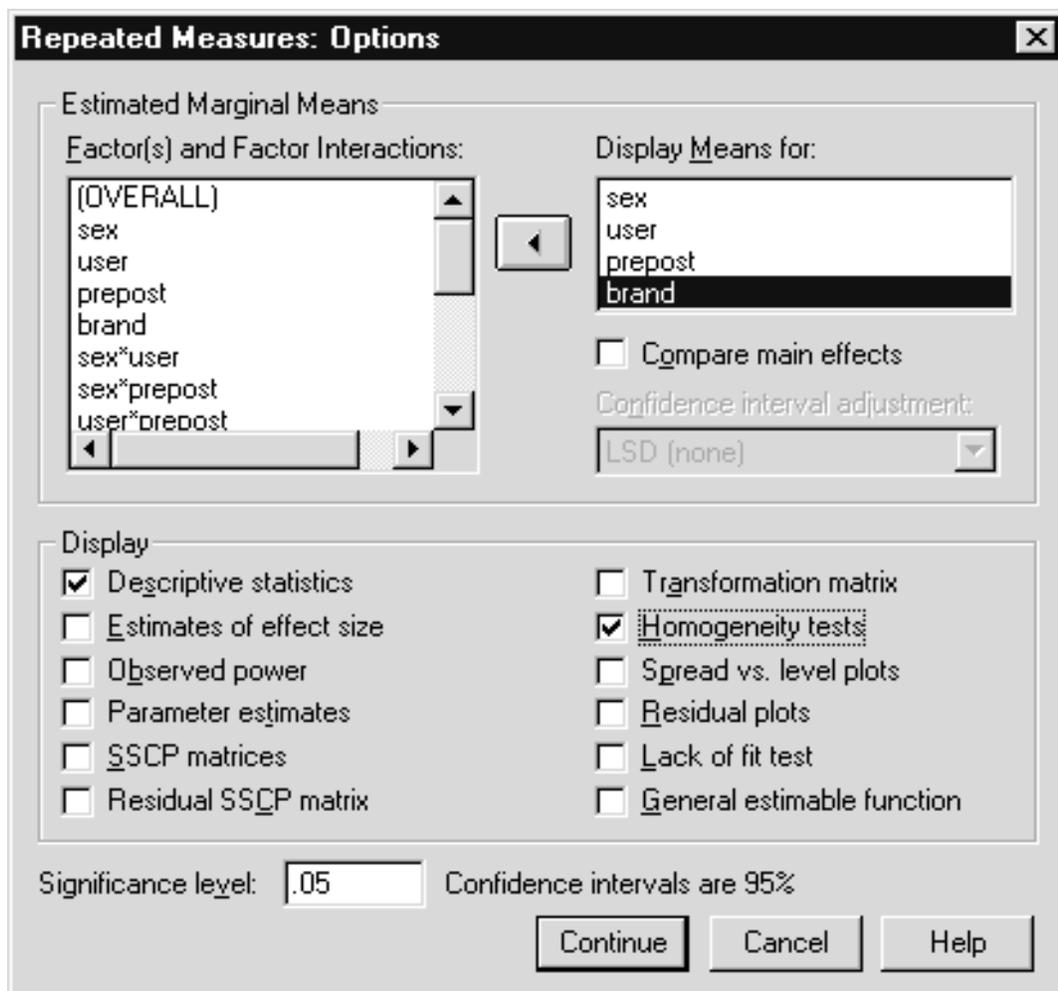
Figure 8.4 Between and Within-Subject Factors Defined



We can proceed with the analysis, but first let us request some options.

- Click the **Options** pushbutton
- Click the **Descriptives** checkbox
- Individually move **sex**, **user**, **prepost**, and **brand** into the **Display Means for** list box
- Click the **Homogeneity** check box

Figure 8.5 Options Dialog Box



Besides the descriptive statistics, we request estimated marginal means (which equal observed means since we are fitting a full model) for each of the factors. Since there are several groups involved in the analysis, we ask for homogeneity of variance tests.

We will proceed with the analysis. Contrasts can be applied to any factors in the same way as we have done earlier.

- Click **Continue** to process the options
- Click **OK** to run the analysis

The command syntax below will produce this analysis.

Figure 8.6 Syntax to Run Analysis

```

GLM
  pre_a pre_b pre_c post_a post_b post_c BY sex user
  /WSFACTOR = prepost 2 Polynomial brand 3 Polynomial
  /MEASURE = rating
  /METHOD = SSTYPE(3)
  /EMMEANS = TABLES(sex)
  /EMMEANS = TABLES(user)
  /EMMEANS = TABLES(prepost)
  /EMMEANS = TABLES(brand)
  /PRINT = DESCRIPTIVE HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /WSDESIGN = prepost brand prepost*brand
  /DESIGN = sex user sex*user .
  
```

Variables which comprise the repeated-measures factors precede the BY keyword and the between-subject factors follow it. Notice the repeated measure variables are ordered so that brand levels change first and brand is mentioned last in the WSFACTOR subcommand. This order is critical for the analysis, so care must be taken when running from syntax. The levels of each repeated measures factor are given and polynomial contrasts (here uninteresting) are used. We requested estimated marginal means for each of the factors. The PRINT subcommand will display the descriptive statistics and the homogeneity tests.

**EXAMINING RESULTS**

Figure 8.7 Factors in the Analysis

Within-Subjects Factors			
Measure: RATING			
PREPOST	BRAND	Dependent Variable	
1	1	PRE_A	
	2	PRE_B	
	3	PRE_C	
2	1	POST_A	
	2	POST_B	
	3	POST_C	

Between-Subjects Factors			
		Value Label	N
SEX	1	Female	43
	2	Male	49
Brand Used	1	A	30
	2	B	31
	3	C	31

The factors in the analysis are listed along with the sample sizes for the between-subject factor groups.

**Figure 8.8 Descriptive Statistics (Beginning)**

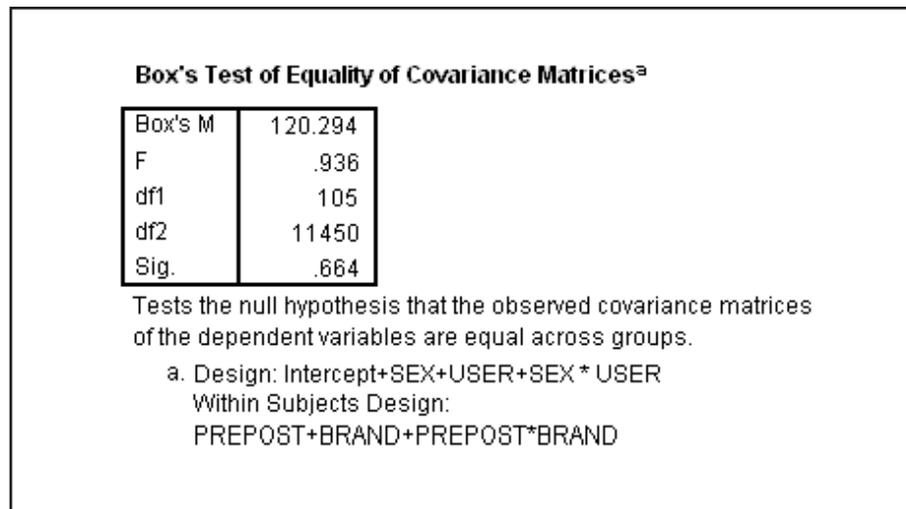
Descriptive Statistics					
	SEX	Brand Used	Mean	Std. Deviation	N
Brand A - Pre Commercial	Female	A	4.64	1.45	14
		B	3.93	1.54	14
		C	3.87	1.30	15
		Total	4.14	1.44	43
	Male	A	6.13	1.09	16
		B	5.65	1.22	17
		C	5.50	1.10	16
		Total	5.76	1.15	49
	Total	A	5.43	1.45	30
		B	4.87	1.61	31
		C	4.71	1.44	31
		Total	5.00	1.52	92
Brand B - Pre Commercial	Female	A	3.93	1.59	14
		B	4.79	1.63	14
		C	4.47	1.68	15
		Total	4.40	1.64	43
	Male	A	5.00	1.15	16
		B	6.24	1.30	17
		C	5.50	1.03	16
		Total	5.59	1.26	49
	Total	A	4.50	1.46	30
		B	5.58	1.61	31

Subgroup means appear separately for each repeated measure variable. Means for the repeated measures factors can be seen in the estimated marginal means pivot tables, or viewed in profile plots.

**TESTS OF ASSUMPTIONS**

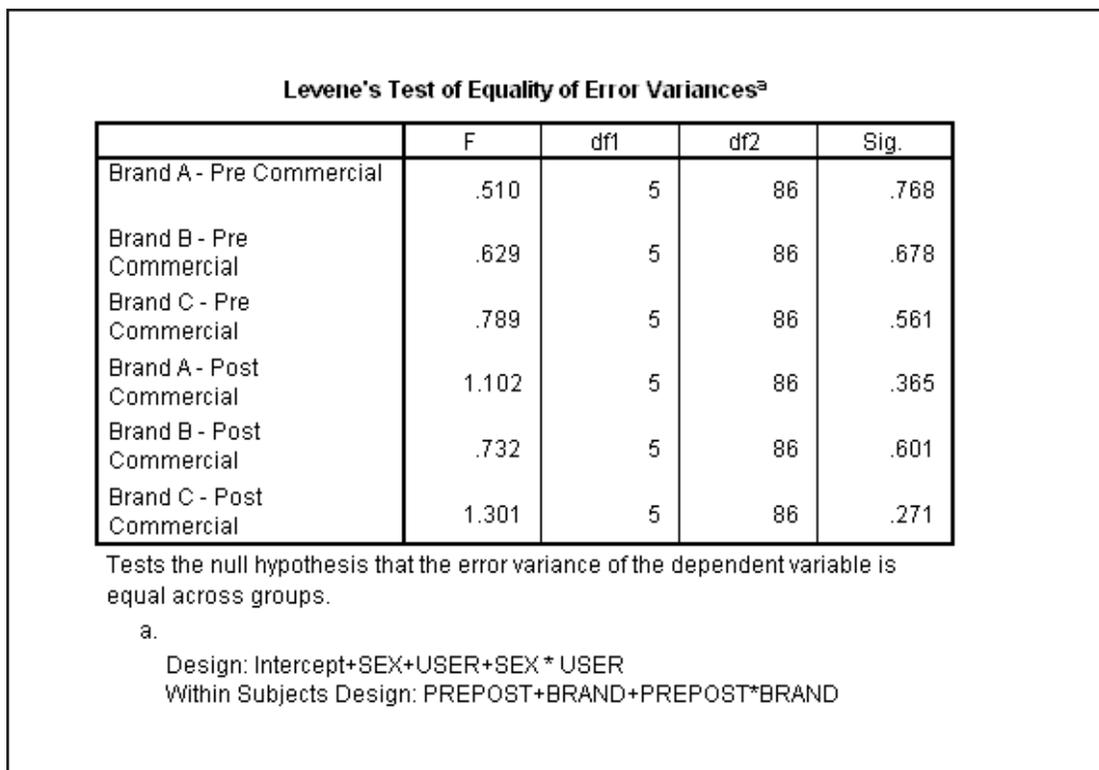
Although they do not appear together in the output, we first examine some assumptions of the analysis. Concerning homogeneity of variance, the program provides Box's M statistic and Levene' test. Box's M is a multivariate statistic testing whether the variance-covariance matrices composed of the six repeated measures variables are equal across the between-subject factor subgroup populations. Levene's test is univariate and tests for homogeneity across subgroup populations for each of the six repeated measure variables separately.

Figure 8.9 Box's M Test of Homogeneity



Box's M test is not significant, indicating that the data are consistent with the assumption of homogeneity of covariance matrices (based on the six repeated measures variables) across the population subgroups.

Figure 8.10 Levene's Test of Homogeneity



Not surprisingly, the results of Levene's test are consistent with Box's M. Box's test is sensitive to both homogeneity and normality violations, while Levene's is relatively insensitive to lack of normality. Since homogeneity of variance violations are generally more problematic for ANOVA, the Levene's test is useful.

Now let us examine the sphericity assumption since this determines whether we simply view the pooled ANOVA results, or move to multivariate or degree of freedom adjusted results.

**Figure 8.11 Sphericity Tests**

**Mauchly's Test of Sphericity<sup>b</sup>**

Measure: RATING

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon <sup>a</sup>		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
PREPOST	1.000	.000	0	.	1.000	1.000	1.000
BRAND	.996	.367	2	.832	.996	1.000	.500
PREPOST * BRAND	.999	.051	2	.975	.999	1.000	.500

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

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

b.  
 Design: Intercept+SEX+USER+SEX \* USER  
 Within Subjects Design: PREPOST+BRAND+PREPOST\*BRAND

Notice no sphericity test is applied to the prepost factor. This is because it has only two levels, so only one difference variable is created, and there is no pooling of effects. The sphericity test for brand is not significant (Sig. = .832), nor is the sphericity test for the brand by prepost interaction (Sig. = .975). Thus the data are consistent with sphericity. As a result we will not view the multivariate test results or the adjusted pooled results (Huynh-Feldt, etc.), and instead focus on the standard (averaged) results.

ANOVA RESULTS Figure 8.12 Within-Subject Tests

Tests of Within-Subjects Effects

Measure: RATING  
Sphericity Assumed

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
PREPOST	39.784	1	39.784	31.617	.000
PREPOST * SEX	2.334	1	2.334	1.855	.177
PREPOST * USER	1.759	2	.879	.699	.500
PREPOST * SEX * USER	1.953	2	.976	.776	.463
Error(PREPOST)	108.214	86	1.258		
BRAND	.451	2	.226	.220	.803
BRAND * SEX	1.470	2	.735	.717	.490
BRAND * USER	101.500	4	25.375	24.768	.000
BRAND * SEX * USER	5.597	4	1.399	1.366	.248
Error(BRAND)	176.216	172	1.025		
PREPOST * BRAND	1.156	2	.578	.476	.622
PREPOST * BRAND * SEX	2.576	2	1.288	1.062	.348
PREPOST * BRAND * USER	4.482	4	1.120	.924	.451
PREPOST * BRAND * SEX * USER	3.512	4	.878	.724	.577
Error(PREPOST*BRAND)	208.604	172	1.213		

*Note that this table has been edited in Pivot Table Editor (epsilon corrected results with sphericity assumed were placed in the top layer) to display only these results.*

This table contains all tests that involve a within-subject factor; those involving only between-subject effects appear later. Looking at the significance (Sig.) column, we see a highly significant difference for pre-post commercial and a brand by user interaction. The brand by pre-post commercial effect is not significant, indicating that although the commercial may have shifted ratings (pre-post commercial is significant) it did not differentially improve the rating of brand A, which was the aim of the commercial. We will view the means and profile plots to understand the significant effects.

**Note** We will not view the multivariate results or the degree of freedom corrected (appropriate if sphericity is violated) results. Nor will we examine the tests of specific contrasts since we had no planned contrasts and the polynomial contrasts over brand categories make no conceptual sense.

Figure 8.13 Between-Subjects Tests

**Tests of Between-Subjects Effects**

Measure: RATING  
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	14830.726	1	14830.726	2165.476	.000
SEX	335.180	1	335.180	48.941	.000
USER	7.638	2	3.819	.558	.575
SEX * USER	5.106	2	2.553	.373	.690
Error	588.989	86	6.849		

Of the between-subjects effects, only sex shows a significant difference. Let us take a look at some of the means.

Figure 8.14 Means for Sex and Pre-Post Commercial

**Estimated Marginal Means**

**1. SEX**

Measure: RATING

SEX	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Female	4.416	.163	4.092	4.740
Male	5.978	.153	5.675	6.282

**3. PREPOST**

Measure: RATING

PREPOST	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.928	.119	4.692	5.163
2	5.466	.124	5.219	5.713

We see males give higher ratings than females and the post-commercial ratings are higher than the pre-commercial ratings. It seems that the commercial was a success, but a success for all brands, not just brand A as hoped.

**PROFILE PLOTS**

To better view the interaction between user (brand used) and brand (brand rated) we request a profile plot

Click the Dialog Recall tool , then select **Repeated**

**Measures**

Click **Define** pushbutton

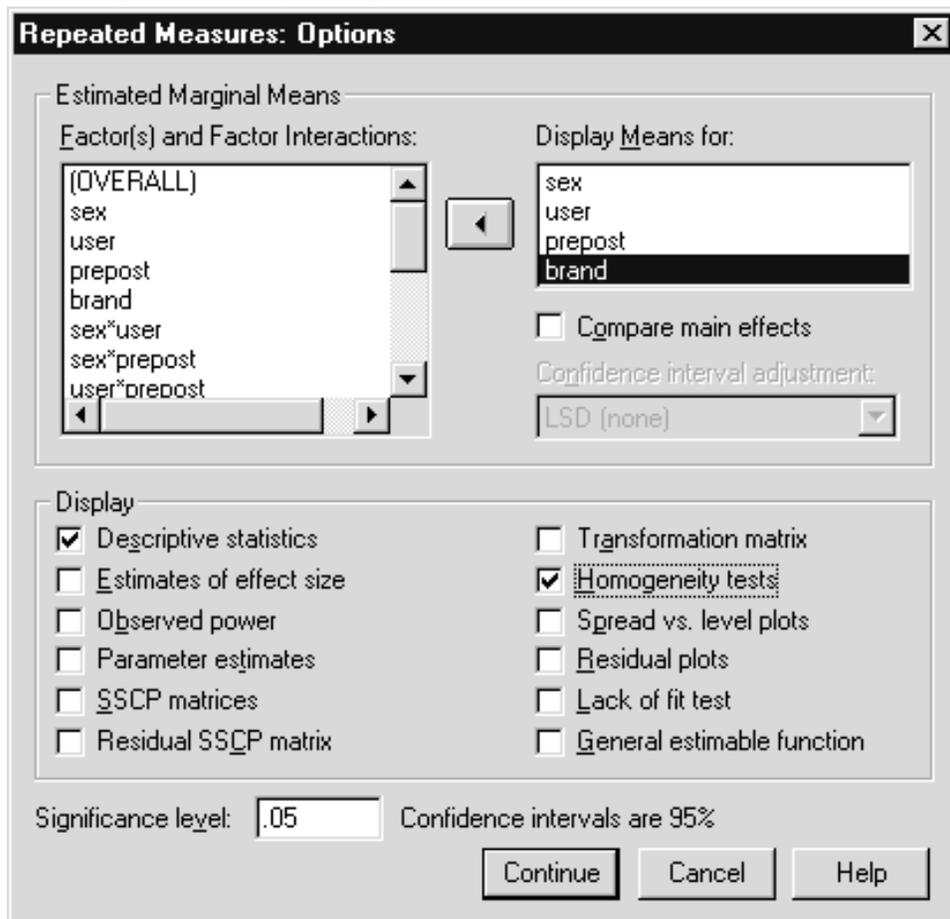
Click **Plots** pushbutton

Move **user** into the **Horizontal Axis** list box

Move **brand** into the **Separate Lines** list box

Click **Add** pushbutton

Figure 8.15 Requesting a Profile Plot



As many as three factors can be displayed in a profile plot, and so up to a three-way interaction can be examined. Note that multiple profile plots can be requested, which allows for many views of your data.

Click **Continue** to process the plot request

Click **OK** to run the analysis

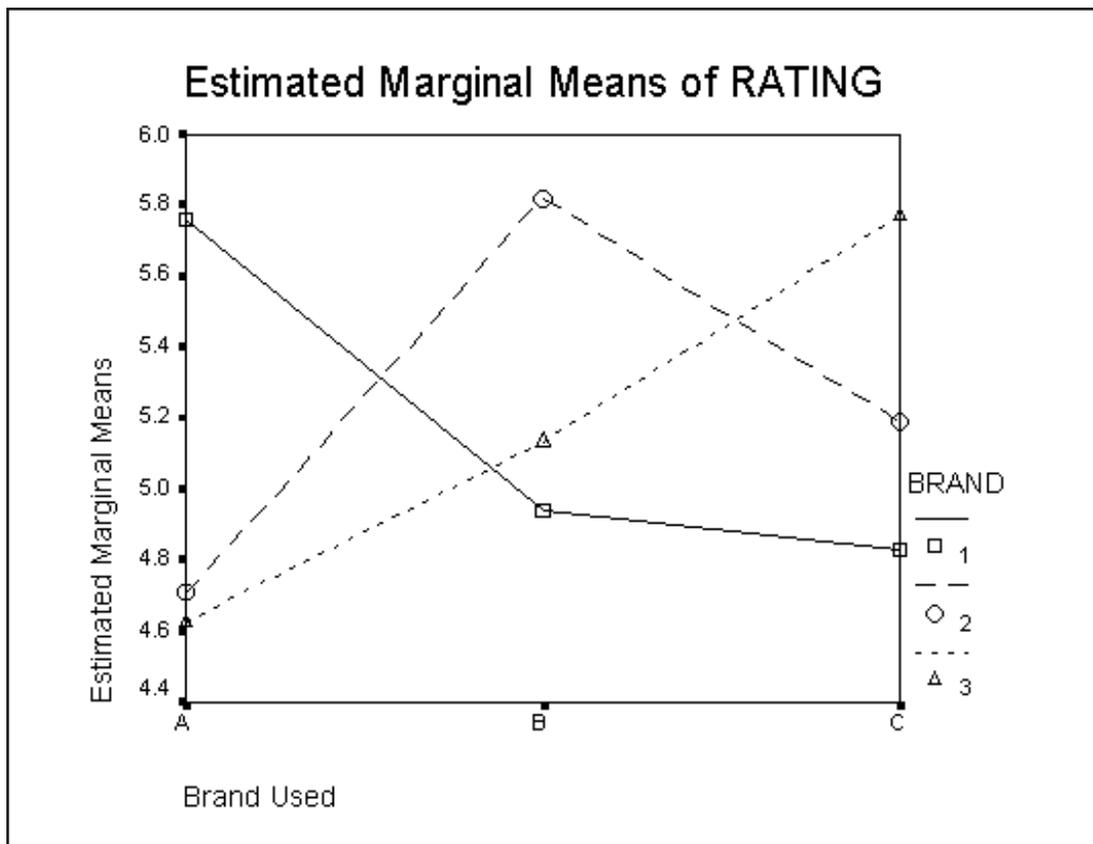
The command below will produce the profile plot.

GLM

```
Pre_a pre_b pre_c post_a post_b post_c BY sex user  
/WSFACTOR = prepost 2 Polynomial brand 3 Polynomial  
/MEASURE = rating  
/METHOD = SSTYPE(3)  
/PLOT = PROFILE(user*brand)  
/EMMEANS = TABLES(sex)  
/EMMEANS = TABLES(user)  
/EMMEANS = TABLES(prepost)  
/EMMEANS = TABLES(brand)  
/PRINT = DESCRIPTIVES HOMOGENEITY  
/CRITERIA = ALPHA(.05)  
/WSDESIGN  
/DESIGN .
```

The Plot subcommand requests a profile plot of user by brand.

Figure 8.16 Profile Plot of Brand Used by Brand Rating



In the plot, brand levels 1, 2, and 3 correspond to brands A, B, and C, respectively. The Brand used by Brand interaction shows (as we surely would expect) that those who regularly use a particular brand rate it higher than the other brands. Especially when there are many factor levels, or several factors involved, profile plots can be very helpful in practice.

### **SUMMARY OF RESULTS**

The homogeneity and sphericity assumptions were met. We did not examine normality, but could do so by requesting residual plots in the Options dialog box. We found that men gave higher brand ratings than women, that the post-commercial ratings were higher than pre-commercial ratings, and that respondents rated their own brand highest. The expected brand by pre-post commercial interaction was not evident.

### **SUMMARY**

In this chapter we examined a more complex split-plot ANOVA involving two between and two within-subject factors. We also used a profile plot to describe an interaction effect.



# Chapter 9 Analysis of Covariance

**Objective** In this chapter we will discuss the purpose, assumptions, and interpretation of analysis of covariance. In addition, we will demonstrate an approach if the parallelism assumption is not met. We will then extend the analysis to include within-subject designs while using constant and varying covariates.

**Method** We will use the General Linear Model Univariate and Repeated Measures procedures to perform the various runs to do the analyses and check the assumptions.

**Data and Scenario** The data presented here are taken from page 806 of Winer(1971). However, we provide a different scenario that will influence the interpretation of the results. It should be noted that the data file is very small and is only for illustrative purposes. Suppose a study was done to evaluate the effectiveness of three treatment drugs on pain-reduction of ankle injuries. There are three types of treatment drugs (variable **Drug** with labels A, B, and C) and each patient is in one of the three drug groups (between-subjects factor). There is within-subject factor, which involves measures taken during the early and later stages of the drug intervention (time periods 1 and 2). The dependent measure is a pain rating scale. Also, physical therapy was performed throughout the study, and the amount of physical therapy varied from patient to patient. Since physical therapy may influence the level of pain reported, it is treated a covariate in the study. There are two measures of the hours of physical therapy a patient experienced, one taken from the period just after the drug treatment was initiated (time period 1) and one taken later in the course of treatment (time period 2).

The main question concerns whether the drugs are effective in pain reduction after controlling for the amount of physical therapy.

**Design** We will run various designs using the same data set. There will be a fixed between-subject factor Drug with 3 levels and a within-subject factor (time) with two levels. The dependent variable is reflected in PAIN1 and PAIN2. The covariate (hours of physical therapy) was measured at the same time points as the dependent variable and is stored in PT1 and PT2.

**Note** We will run many different analyses on the same data set to demonstrate the flexibility of the technique and reduce possible confusion to constantly switching data. In practice, you would be interested in a specific set of models.

### INTRODUCTION

**A**nalysis of covariance can be viewed as an attempt to provide some statistical control in place of lack of experimental control. Inclusion of a covariate allows the researcher to run the usual ANOVAs while controlling for some other variable. This is not control in the experimental sense, but control in the sense of making a statistical adjustment to equate all groups on the covariate. Covariates are interval scale variables; if they were categorical then they would be included as additional factors in the design.

One purpose of analysis of covariance is to obtain a more sensitive ANOVA by reducing the within-group variability. If the covariate is related to the dependent variable the same way in each group, the within group variation can be reduced by removing the effect of the covariate. The classic case is an experiment in which subjects are randomly assigned to groups, but vary on some background measure; analysis of covariance (ANCOVA) will control for this source of variation.

Analysis of covariance is often spoken of as a conditional analysis. Removing the effect of the covariate essentially equates all subjects on the covariate, so instead of speaking of factor A having an effect we speak of factor A having an effect if subjects had identical values on the covariate. A common example of analysis of covariance is the adjusting for body weight in medical experiments. In this context, analysis of covariance adjusts the analysis as if each subject began at the same weight.

ANCOVA is also used in non-experimental studies to substitute statistical control for factors beyond the control of the researcher. Care must be taken since if the covariate relates to factors in the study, controlling for covariate modifies the estimated effects of the factors themselves.

### HOW IS ANALYSIS OF COVARIANCE DONE?

Basically, the dependent variable is regressed on the covariate, but the relevant variation of the dependent variable is not its variation around the grand mean but instead is based on the pooled within-group variation. Thus a within-group regression with the covariate(s) is run and the analysis of variance is performed on the residuals from the regression.

### ASSUMPTIONS OF ANCOVA

The major assumptions specific to ANCOVA are: 1) The relationship between the covariate and the dependent variable (within groups) is linear; 2) The within-group distribution of the residuals is normal; and 3) The relationship between the covariate and the dependent variable (the slopes in the within-group regressions) is the same across all groups.

Assumption (1) need not hold, but the routines available in most software are based on a linear relationship. Assumption (2) is the usual normality assumption, this time after the covariate has been applied. The

last assumption is important and can be tested. The degree of adjustment made is based on the pooled within-groups regression. If the slope relating the covariate to the dependent variable varies across groups, then the common slope used to adjust each group does not reflect the true relationship for that group. We will see that a different slope can be fit to each group, but this requires rethinking just what we hope to accomplish with the analysis.

## CHECKING THE ASSUMPTIONS

Plots of the dependent variable and the covariate can be made separately for each group (if there are relatively few cells in the analysis) to take an informal look at the homogeneity of slopes. The residuals can be displayed in normal plots. The homogeneity of slopes assumption can be formally tested.

## BASELINE ANOVA

We first run a one-factor ANOVA to provide a baseline. We use the Univariate procedure instead of a One-Way ANOVA in order to use the same procedure throughout.

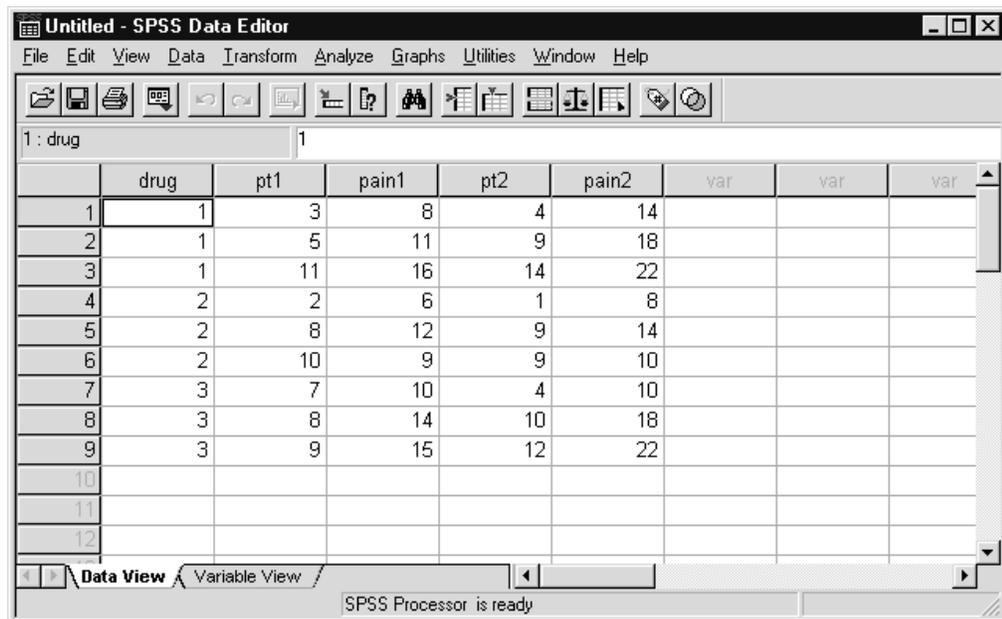
Click **File..Open..Data**

Move to the **c:\Train\Anova** directory (if necessary)

Select **SPSS Portable (.por)** from the Files of Type drop-down list

Double-click on **PainTreat.por**

Figure 9.1 Data for Analysis of Covariance Example



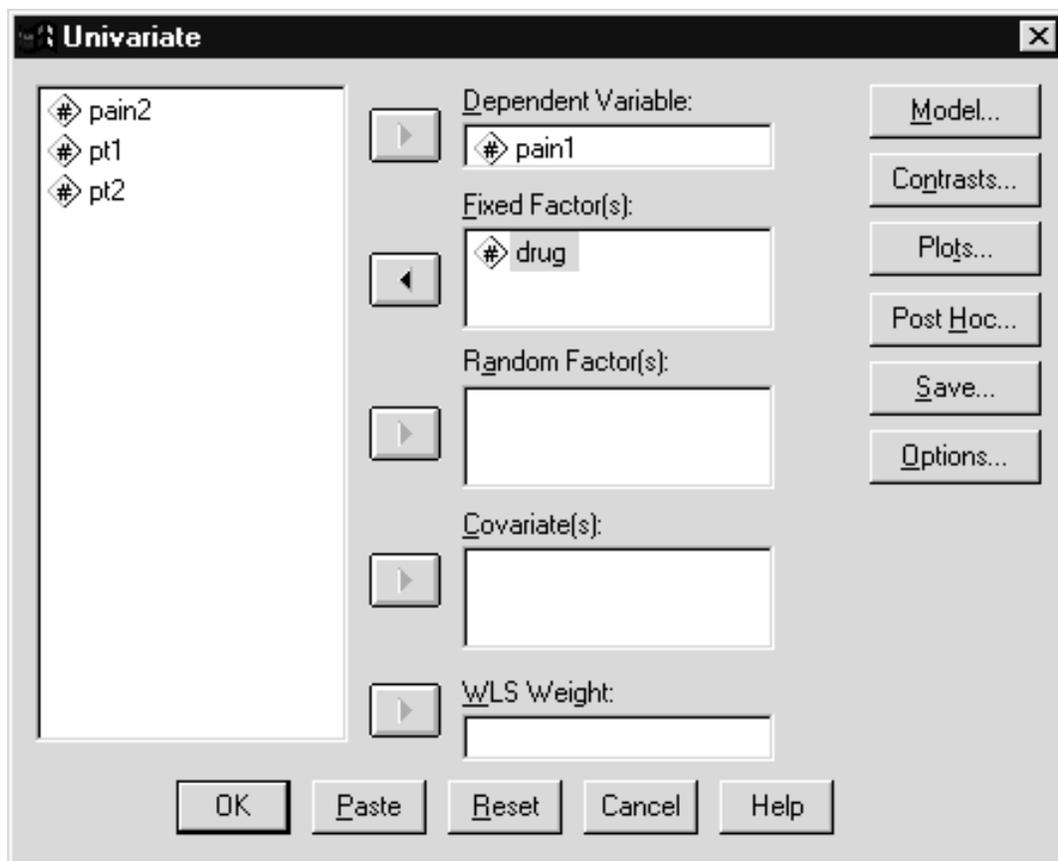
	drug	pt1	pain1	pt2	pain2	var	var	var
1	1	3	8	4	14			
2	1	5	11	9	18			
3	1	11	16	14	22			
4	2	2	6	1	8			
5	2	8	12	9	14			
6	2	10	9	9	10			
7	3	7	10	4	10			
8	3	8	14	10	18			
9	3	9	15	12	22			
10								
11								
12								

Click **Analyze..General Linear Model..Univariate**

Move **PAIN1** into the **Dependent Variable** list box

Move **Drug** into the **Fixed Factors** list box

Figure 9.2 Univariate Dialog Box



Click on **OK** to run the analysis.

PAIN1 (pain during period 1) is the dependent variable and there is one between-subjects factor (Drug) with three levels. The following command will run this analysis.

```
UNIANOVA  
pain1 BY drug  
/METHOD = SSTYPE(3)  
/INTERCEPT = INCLUDE  
/CRITERIA = ALPHA(.05)  
/DESIGN = drug .
```

Figure 9.3 ANOVA Table

Tests of Between-Subjects Effects

Dependent Variable: Period 1 Treatment Pain

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	24.889 <sup>a</sup>	2	12.444	1.155	.376
Intercept	1133.444	1	1133.444	105.165	.000
DRUG	24.889	2	12.444	1.155	.376
Error	64.667	6	10.778		
Total	1223.000	9			
Corrected Total	89.556	8			

a. R Squared = .278 (Adjusted R Squared = .037)

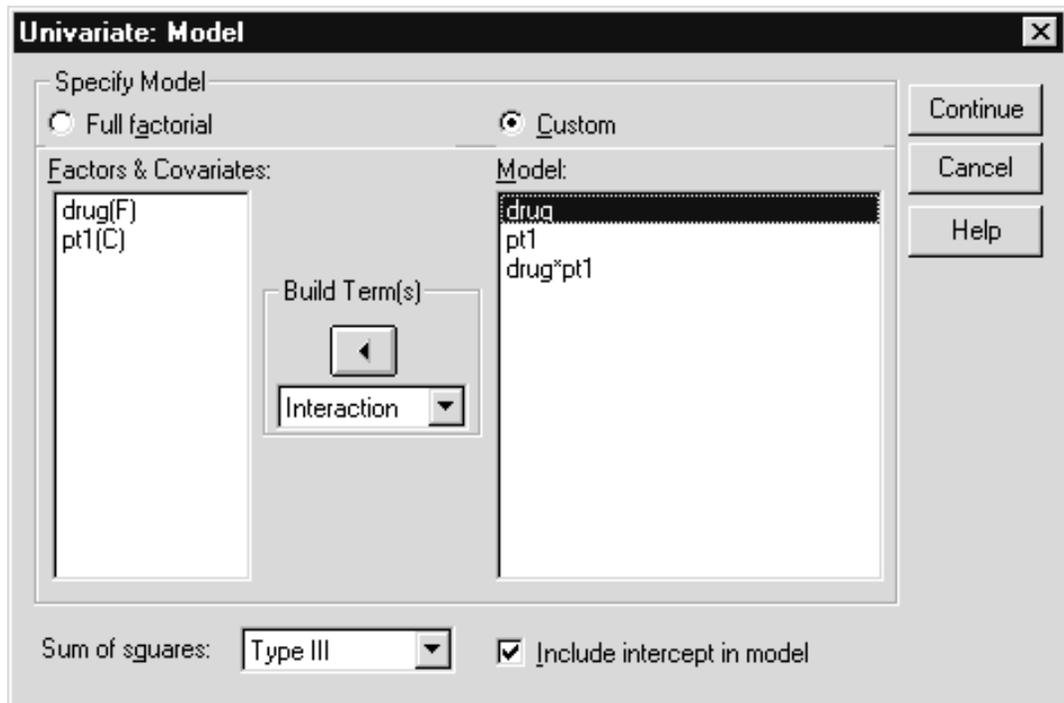
This table shows no suggestion of a main effect of Drug in this analysis (significance level is .376). Thus during the early treatment period, there were no differences attributable to drug found in the pain measure.

**ANCOVA –  
HOMOGENEITY  
OF SLOPES**

In the second run we will include the covariate and the interaction term of the covariate and the between-subject factor. The assumption of equality (homogeneity) of regression slopes can be tested by fitting a model containing the main effects of Drug and PT1, as well as the Drug\*PT1 interaction. The interaction term provides the test of the null hypothesis of equal slopes. If the slopes relating the covariate to the dependent variable are identical (parallel) across the different groups, this interaction will not be significant.

- Click the **Dialog Recall** tool , then click **Univariate**
- (Verify that PAIN1 is in the Dependent variable box and Drug is in the Fixed Factor(s) list box)
- Move **PT1** into the Covariate(s) list box
- Click the **Model** pushbutton
- Select **Custom** model option button
- Click **Drug**, then click the Build Term **arrow** to add Drug to the model
- Click **PT1**, and then click the Build Term **arrow** to add PT1 to the model
- Select both **Drug** and **PT1** in the Factors & Covariates list box (use Ctrl-click), then click the Build Term **arrow** to add the Drug\*PT1 interaction term to the model

Figure 9.4 Univariate Dialog Box



Click **Continue** to process the model requests  
 Click on **OK** to run the analysis

The following command will also run the analysis:

```
UNIANOVA
  pain1 BY drug WITH pt1
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /DESIGN = drug pt1 drug*pt1 .
```

The keyword WITH precedes covariates in UNIANOVA, just as BY precedes factors. From the first command line alone, UNIANOVA would run a standard analysis of covariance, not testing the interaction term. In the DESIGN subcommand we include the between-subjects factor (Drug), the covariate (PT1), and the factor by covariate interaction (Drug BY PT1). This interaction effect is the main focus of our interest.

Figure 9.5 ANCOVA Table with Homogeneity of Slopes Test  
Tests of Between-Subjects Effects

Dependent Variable: Period 1 Treatment Pain

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	78.786 <sup>a</sup>	5	15.757	4.390	.127
Intercept	.478	1	.478	.133	.739
DRUG	4.764	2	2.382	.664	.577
PT1	28.414	1	28.414	7.915	.067
DRUG * PT1	9.410	2	4.705	1.311	.390
Error	10.769	3	3.590		
Total	1223.000	9			
Corrected Total	89.556	8			

a. R Squared = .880 (Adjusted R Squared = .679)

The summary table indicates that the interaction is not significant (Sig. = .390), so the homogeneity of slopes assumption seems to be met. Thus the linear relationship between hours of physical therapy and pain level does not differ across drug treatment groups. There is a suggestion of an effect of the covariate (Sig. = .067), but no effect due to Drug. To repeat, with such a small sample there is little power to detect assumption violations, but we wish to demonstrate the method.

## STANDARD ANCOVA

Having checked the parallelism of slopes assumption, we proceed with the standard ANCOVA.

Click the **Dialog Recall** tool , then click **Univariate**

- Click the **Model** pushbutton
- Select the **Full factorial** model option button
- Click **Continue** to process the change
- Click **Options** pushbutton
- Click **Parameter Estimates** checkbox
- Click **Continue**
- Click **OK** to run the analysis.

The following command will run the analysis using syntax.

```
UNIANOVA
  pain1 BY drug WITH pt1
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /PRINT = DESCRIPTIVE PARAMETER
  /CRITERIA = ALPHA(.05)
  /DESIGN = pt1 drug.
```

Figure 9.6 ANCOVA Summary Table

Tests of Between-Subjects Effects

Dependent Variable: Period 1 Treatment Pain

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	69.376 <sup>a</sup>	3	23.125	5.730	.045
Intercept	40.630	1	40.630	10.067	.025
PT1	44.488	1	44.488	11.023	.021
DRUG	17.153	2	8.576	2.125	.215
Error	20.179	5	4.036		
Total	1223.000	9			
Corrected Total	89.556	8			

a. R Squared = .775 (Adjusted R Squared = .639)

The results are similar to the previous analysis, no effect due to factor Drug, and a significant relationship between the covariate and dependent measure. The reason for the covariate now being significant probably has to do with the extra degrees of freedom added to the error term – going from 3 to 5 degrees of freedom is a big jump.

DESCRIBING THE RELATIONSHIP

The Univariate procedure also presents some information to characterize the relation between the covariate and dependent variable.

Figure 9.7 Parameter Estimates

Parameter Estimates

Dependent Variable: Period 1 Treatment Pain

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	6.682	2.228	2.999	.030	.954	12.411
PT1	.790	.238	3.320	.021	.178	1.401
[DRUG=1]	-1.71E-02	1.688	-.010	.992	-4.355	4.321
[DRUG=2]	-2.947	1.671	-1.764	.138	-7.242	1.348
[DRUG=3]	0 <sup>a</sup>	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

In the GLM parameterization, the intercept parameter estimate gives the estimated value of the last category of Drug (Drug = 3) when the covariate is equal to 0. The Drug = 1 and Drug =2 coefficients subtract the level 3 predicted value from the level 1 and level 2 predicted values, respectively. Adding one of these coefficients to the intercept estimate gives the estimated value for that level of Drug when the covariate is equal to 0.

The B coefficient for PT1 is the regression coefficient used to predict the dependent variable based on the covariate. Its positive coefficient indicates that higher levels of physical therapy are associated with

greater pain levels. This is not the expected relationship and if this were real data, it should be examined more carefully (perhaps patients with more serious and painful injuries received more physical therapy). The 95% confidence band for the regression coefficient is rather wide.

**FITTING NON-PARALLEL SLOPES**

If an interaction between a covariate and a factor in the model is significant, it indicates that the slopes relating the covariate to the dependent variable vary across groups. If there is interest in modeling this, that is, fitting different slopes to each group, this can be specified in the Univariate procedure. It is no longer the standard analysis of covariance since the degree of adjustment varies with the group, but the analysis may be of interest in its own right.

Click the Dialog Recall tool , then click **Univariate**

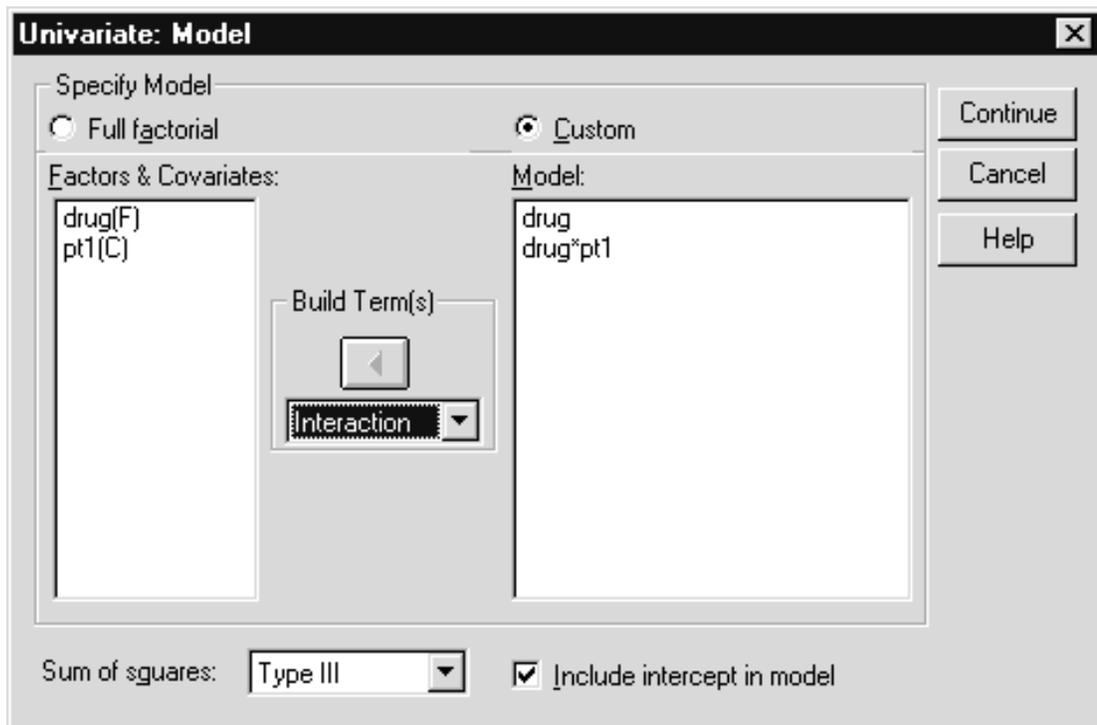
Click the **Model** pushbutton

Select the **Custom Model** option button

If **Drug** and **Drug\*PT1** are not already in the Model list box (from our earlier analysis) then move **Drug** and **Drug\*PT1** (Ctrl-click to select both) into the Model box

Remove **PT1** from the Model list box (if necessary)

Figure 9.8 Model for Separate Slope Analysis



Since PT1 is removed from the model, Univariate will assign three degrees of freedom to the PT1 by Drug interaction. Thus it will fit a separate slope (between PT1 and the dependent measure) for each level of Drug. If we left PT1 in the model, as we did when testing slope

homogeneity, then the PT1 effect would represent the overall slope between PT1 and the dependent measure, while the two degrees of freedom PT1 by Drug effect would test the interaction of PT1 and Drug.

Click **Continue** to process the change  
 Click **OK** to run the analysis

The following syntax will run the analysis.

```
UNIANOVA
  pain1 BY drug WITH pt1
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /PRINT = PARAMETER
  /CRITERIA = ALPHA(.05)
  /DESIGN = drug drug*pt1 .
```

**Figure 9.9 Between-Subject Tests**

**Tests of Between-Subjects Effects**

Dependent Variable: Period 1 Treatment Pain

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	78.786 <sup>a</sup>	5	15.757	4.390	.127
Intercept	.478	1	.478	.133	.739
DRUG	4.764	2	2.382	.664	.577
DRUG * PT1	53.897	3	17.966	5.005	.109
Error	10.769	3	3.590		
Total	1223.000	9			
Corrected Total	89.556	8			

a. R Squared = .880 (Adjusted R Squared = .679)

We notice that neither the main effect of Drug, nor the covariate main effect or interaction of the factor and the covariate (bundled together in Drug\*PT1) are significant (.577 and .109, respectively).

Figure 9.10 Parameter Estimates

Parameter Estimates

Dependent Variable: Period 1 Treatment Pain

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	-7.000	10.774	-.650	.562	-41.286	27.286
[DRUG=1]	12.577	11.019	1.141	.337	-22.490	47.644
[DRUG=2]	12.538	11.039	1.136	.339	-22.594	47.671
[DRUG=3]	0 <sup>a</sup>	.	.	.	.	.
[DRUG=1] * PT1	.962	.322	2.988	.058	-6.255E-02	1.986
[DRUG=2] * PT1	.519	.322	1.614	.205	-.505	1.543
[DRUG=3] * PT1	2.500	1.340	1.866	.159	-1.764	6.764

a. This parameter is set to zero because it is redundant.

The values for the intercept and Drug = 1, Drug = 2, and Drug = 3 are the same as explained earlier. Notice that there are 3 degrees of freedom for Drug\*PT1: one for each of the three slopes. The parameter estimates for Drug\*PT1 provide the slope estimates, relating hours of physical therapy to reported pain level, for each of the three groups.

**REPEATED  
MEASURES  
ANCOVA WITH A  
SINGLE  
COVARIATE**

To illustrate analysis of covariance in the context of repeated measures we will first run a split-plot analysis (between-subject factor Drug, within-subject factor time with two levels), using only the first measurement of the covariate PT1 to illustrate the analysis. This is also termed repeated measures with a constant covariate since a single covariate value applies across levels of the repeated measure factors.

Click **Analyze..General Linear Model..Repeated Measures**

Replace **factor1** with **time**

Enter **2** in the number of levels box

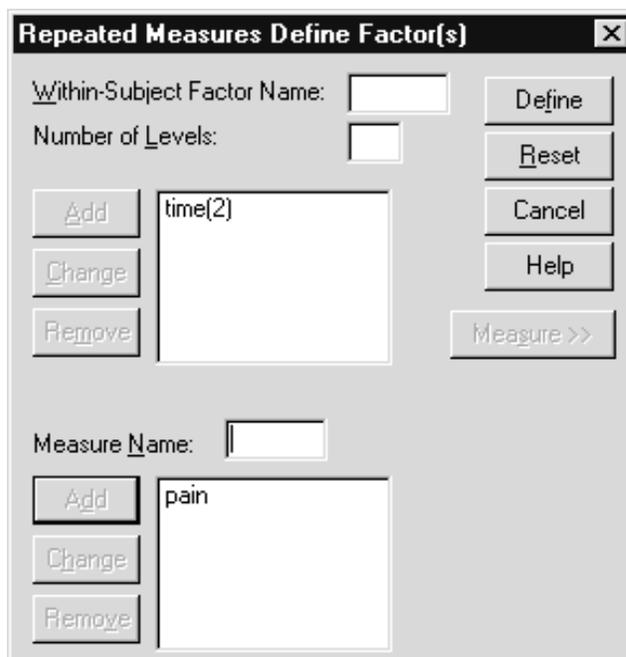
Click the **Add** pushbutton

Click the **Measure** pushbutton

Type **pain** in the Measure Name text box

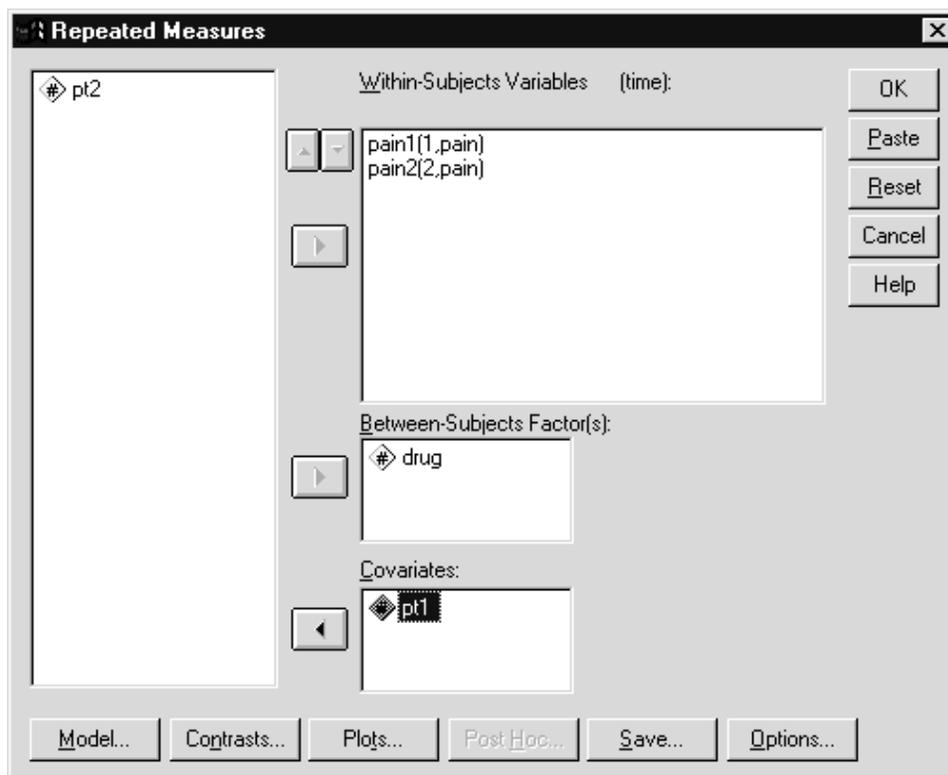
Click the **Add** pushbutton in the Measure Name area

Figure 9.11 Repeated Measures Define Factors Dialog Box



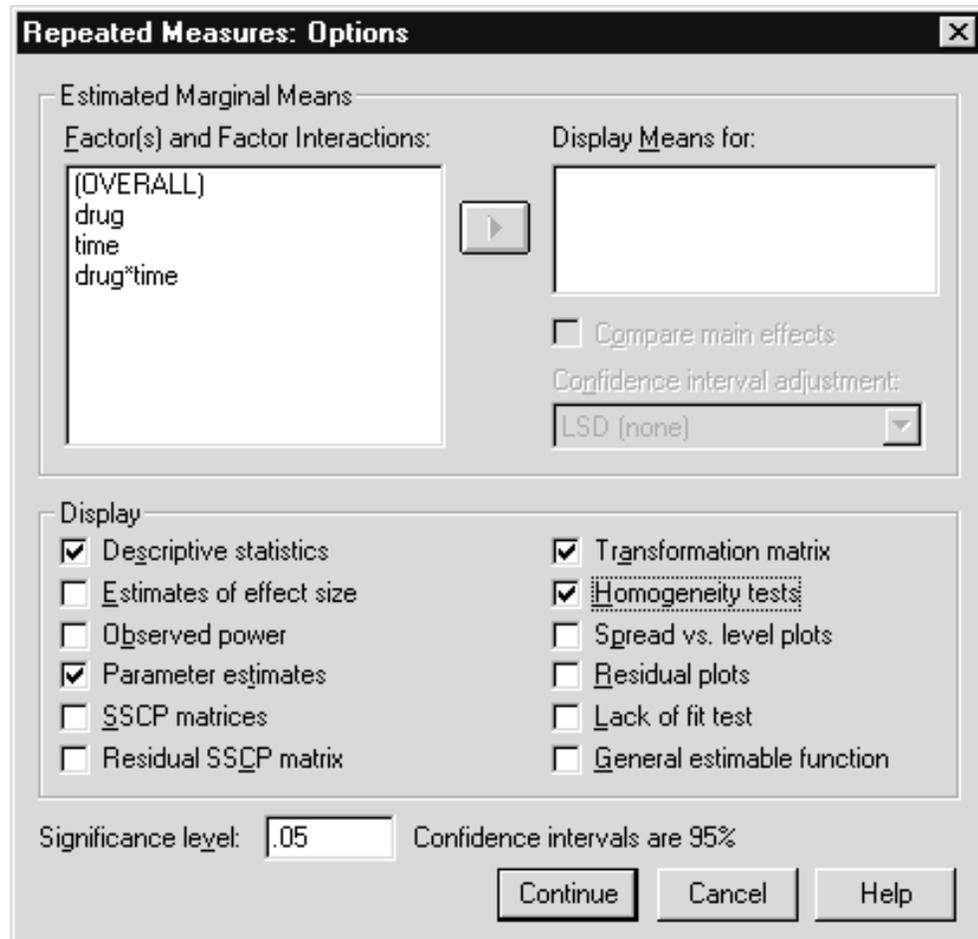
- Click the **Define** pushbutton
- Move **PAIN1** and **PAIN2** into the **Within-Subject Variables** list box in that order
- Move **Drug** into the **Between-subject Factor(s)** list box
- Move **PT1** into the **Covariates** list box

Figure 9.12 Repeated Measures Dialog Box



Click the **Options** pushbutton  
 Select **Descriptives, Parameter Estimates, Transformation Matrix, and Homogeneity Tests**

Figure 9.13 Options Dialog Box



Click **Continue** to process the request  
 Click **OK** to run the analysis

The following command will also run the analysis.

```
GLM
  pain1 pain2 BY drug WITH pt1
  /WSFACTOR = time 2 Polynomial
  /MEASURE = pain
  /METHOD = SSTYPE(3)
  /PRINT = DESCRIPTIVE PARAMETER TEST(MMATRIX)
  HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /WSDESIGN = time
  /DESIGN = pt1 drug .
```

Scroll **down** to **Transformation** section of results

Figure 9.14 Transformation Matrix

**Transformation Coefficients (M Matrix)**

**Average**

Measure: PAIN  
Transformed Variable: AVERAGE

Period 1 Treatment Pain	.707
Period 2 Treatment Pain	.707

**TIME<sup>2</sup>**

Measure: PAIN

	TIME
Dependent Variable	Linear
Period 1 Treatment Pain	-.707
Period 2 Treatment Pain	.707

a. The contrasts for the within subjects factors are:  
TIME: Polynomial contrast

The first transformation is the average of PAIN1 and PAIN2. Thus the covariate will be applied to the effects that involve the average of PAIN1 and PAIN2, that is, only between-subject effects.

Scroll **up** to the **Tests of Between-Subjects Effects**' pivot table

Figure 9.15 Tests of Between-Subject Effects

**Tests of Between-Subjects Effects**

Measure: PAIN  
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	141.882	1	141.882	8.415	.034
PT1	92.699	1	92.699	5.498	.066
DRUG	89.903	2	44.951	2.666	.163
Error	84.301	5	16.860		

The covariate is not quite significant (Sig. = .066). If it were significant, this would indicate that the amount of physical therapy during the early phase of drug treatment is related to overall (period 1 and period 2 measures, averaged together) pain ratings. The effect of the Drug factor, adjusted for the covariate, is not significant.

Figure 9.16 Tests of Within-Subjects Effects (Sphericity Assumed)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	8.393	1	8.393	3.237	.132
TIME * PT1	3.816E-02	1	3.816E-02	.015	.908
TIME * DRUG	16.476	2	8.238	3.178	.129
Error(TIME)	12.962	5	2.592		

Note: the pivot table above was edited in the Pivot Table Editor so only the sphericity assumed results appear. Since there are only two levels of the repeated measure factor, the sphericity test and corrections are not relevant). We find no effects significant: main effect of time, interaction between time and factor Drug, interaction between the covariate and time. This latter effect (Time by PT1) tests whether the slope relating the covariate (physical therapy) to the dependent measure (pain) is the same (parallel) for each of the two time periods.

Figure 9.17 Parameter Estimates

Dependent Variable	Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Period 1 Treatment Pain	Intercept	6.682	2.228	2.999	.030	.954	12.411
	PT1	.790	.238	3.320	.021	.178	1.401
	[DRUG=1]	-1.71E-02	1.688	-.010	.992	-4.355	4.321
	[DRUG=2]	-2.947	1.671	-1.764	.138	-7.242	1.348
	[DRUG=3]	0 <sup>a</sup>	.	.	.	.	.
Period 2 Treatment Pain	Intercept	10.087	4.356	2.316	.068	-1.109	21.284
	PT1	.822	.465	1.769	.137	-.373	2.017
	[DRUG=1]	2.704	3.298	.820	.450	-5.774	11.182
	[DRUG=2]	-4.903	3.265	-1.502	.193	-13.297	3.490
	[DRUG=3]	0 <sup>a</sup>	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

This is the standard ANCOVA table of parameters under the general linear model. Note a separate slope coefficient (for covariate PT1) is calculated for PAIN1 and PAIN2; the model effects are adjusted for both.

**REPEATED  
MEASURES  
ANCOVA WITH A  
VARYING  
COVARIATE**

Since covariates are not always fixed measures at one time point, covariates that are measured under each condition can be used in the analysis. We will analyze the same data using PT1 and PT2, measures of the covariates at the two time points. Note that GLM will adjust each level of the repeated measure factor (PAIN1, PAIN2) for every covariate. Thus a covariate that varies over time is treated identically to the situation in which multiple covariates are recorded at a single time point.

Click on the **Dialog Recall** tool , then click **Repeated Measures**  
 Click on the **Define** button  
 Add **PT2** to the Covariates box  
 Click on **OK**

The following command will run this analysis.

```
GLM
pain1 pain2 BY drug WITH pt1 pt2
/WSFACTOR = time 2 Polynomial
/MEASURE = pain
/METHOD = SSTYPE(3)
/PRINT = PARAMETER TEST(MMATRIX) HOMOGENEITY
/CRITERIA = ALPHA(.05)
/WSDESIGN = time
/DESIGN = pt1 pt2 drug .
```

**Figure 9.18 Test of Within-Subjects Factors**

**Tests of Within-Subjects Effects**

Measure: PAIN  
 Sphericity Assumed

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	9.973	1	9.973	17.370	.014
TIME * PT1	8.049	1	8.049	14.019	.020
TIME * PT2	10.665	1	10.665	18.576	.013
TIME * DRUG	1.923	2	.961	1.675	.296
Error(TIME)	2.297	4	.574		

*As before, the pivot table has been edited so only the results that assume sphericity appear.*

As we can see Time, the interaction of Time and PT1, and the interaction of Time and PT2 are both significant, but the Time by Drug interaction is not significant. This suggests that there is a change in pain level over time and that this change is related to the amount of physical therapy in both the early and late stages of drug treatment. Although not significant, a Drug by Time interaction would suggest that the effect of Drug is not uniform across the two time periods.

**Figure 9.19 Test of Between-Subjects Factors**  
**Tests of Within-Subjects Effects**

Measure: PAIN  
 Sphericity Assumed

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	9.973	1	9.973	17.370	.014
TIME * PT1	8.049	1	8.049	14.019	.020
TIME * PT2	10.665	1	10.665	18.576	.013
TIME * DRUG	1.923	2	.961	1.675	.296
Error(TIME)	2.297	4	.574		

The only significant effect is the covariate PT2 (the amount of physical therapy during the second time period). The results of the other effects are consistent with the previous analysis.

**Figure 9.20 Parameter Estimates**

**Parameter Estimates**

Dependent Variable	Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Period 1 Treatment Pain	Intercept	7.348	1.428	5.144	.007	3.382	11.314
	PT1	-9.19E-02	.338	-.272	.799	-1.030	.846
	PT2	.737	.253	2.913	.044	3.5E-02	1.439
	[DRUG=1]	-1.732	1.219	-1.42	.228	-5.117	1.653
	[DRUG=2]	-2.403	1.073	-2.24	.089	-5.383	.578
	[DRUG=3]	0 <sup>a</sup>	.	.	.	.	.
Period 2 Treatment Pain	Intercept	11.583	1.572	7.368	.002	7.218	15.948
	PT1	-1.159	.372	-3.11	.036	-2.191	-.126
	PT2	1.656	.278	5.948	.004	.883	2.429
	[DRUG=1]	-1.150	1.342	-.857	.440	-4.876	2.576
	[DRUG=2]	-3.681	1.181	-3.12	.036	-6.961	-.400
	[DRUG=3]	0 <sup>a</sup>	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

The interpretation of the parameter estimates is identical to our earlier discussion (see Figure 9.17); each of the covariates is summarized separately. As we found earlier with PT1 (physical therapy during the first period), the estimates for the covariate coefficients of PT2 (physical therapy during the second period) indicate that pain level increases with more physical therapy. Oddly, higher levels of physical therapy during the first period relate to lower pain levels during the second period. Notice a coefficient relates PT2 (amount of physical therapy during the second period) to the first period pain measure; it might be argued on logical grounds that this coefficient should not be included in the model.

The fact that the intercept for the second period is greater (11.583) than that for the first period (7.342) indicates that pain levels increased

over time! This could be shown more clearly by requesting the estimated marginal means for the Time factor using the Options dialog.

### **FURTHER VARIATIONS**

This process can be generalized to additional covariates and repeated measures factors, and even more complicated variations. If you attempt these analyses, make sure you display the transformation matrix that will inform you of the actual analysis that the General Linear Model procedures are performing.

# Chapter 10 Special Topics

**Objective** We will discuss the setups for some specialty statistical models: Latin Square Designs and Random Effects Designs. We will not discuss substantive interpretation of the results.

## INTRODUCTION

In addition to the more or less standard analyses discussed in the previous chapters, there are a number of more specialized ANOVA applications. The family of incomplete designs, which includes Latin Squares, allows experimenters to control for nuisance factors, or to study a number of factors without including all possible combinations of the factor levels in the analysis. The price of this involves giving up the opportunity to test for interaction effects. In this chapter we demonstrate that the GLM procedure can perform such analyses with experimental data collected from a Latin Square design. A second application we will explore involves ANOVA designs that contain more than a single random factor. Again, such models can be run using the GLM procedure.

### LATIN SQUARE DESIGNS

Latin Square designs are useful when there is interest in performing a multiple factor ANOVA, but it is impossible or undesirable to represent all combinations of levels of factors in the analysis. For example, a three-factor design with each factor containing five levels implies 125 groups! Another common use of Latin Square designs involves controlling for nuisance factors, that is, controlling for the effects of factors that may influence the outcome, but are not themselves of experimental interest.

The basic idea is that not all combinations of levels of levels of factors are included, but those included are counterbalanced so that independent main effects can be tested. The counterbalancing is designed to confound main effects with certain interaction terms, and an assumption is made that the interaction terms are not significant. As a result of not including all cells, at least some and possibly all interaction questions cannot be tested. When there are no replicates within cells, the variation usually attributed to higher-order interactions is used as the error term in testing main effects.

### AN EXAMPLE

To illustrate a Latin Square design, Montgomery (1984) provides an example of a dynamite manufacturer interested in evaluating the results of five chemical formulations on the explosive force of the resulting compound. In addition, two other factors have been identified as potentially influencing the compound, namely the quality of the raw materials and the person mixing the materials. These will be considered to be systematic sources of error (nuisance factors) that need to be removed from the analysis. Thus we consider three factors: formulation, batch of raw materials, and operator. Ideally, an experiment would be performed so that each operator uses each batch of raw materials in preparing each formulation (5 x 5 x 5, or 125 cells). Here we run into the practical problem of there being not enough raw materials in a batch to supply each operator for each formulation (25 combinations). For this reason the researcher cannot perform the fully balanced experiment and instead a Latin Square will be used.

Below we show the formulation assignments (A-E) with five operators (1-5) and five batches of material (1-5). The equal number of levels within each factor is required for balancing and is a feature of such designs.

Batches	Operator 1	Operator 2	Operator 3	Operator 4	Operator 5
1	A	B	C	D	E
2	B	C	D	E	A
3	C	D	E	A	B
4	D	E	A	B	C
5	E	A	B	C	D

Notice that each formulation appears once in each row and column of the table – that is, once with each batch of materials and once with each operator. If interactions between batches, operators, and compounds are negligible, then we can test for the effects of different formulations of compounds on explosive force without the noise introduced by raw

materials and operators (by adjusting for it).

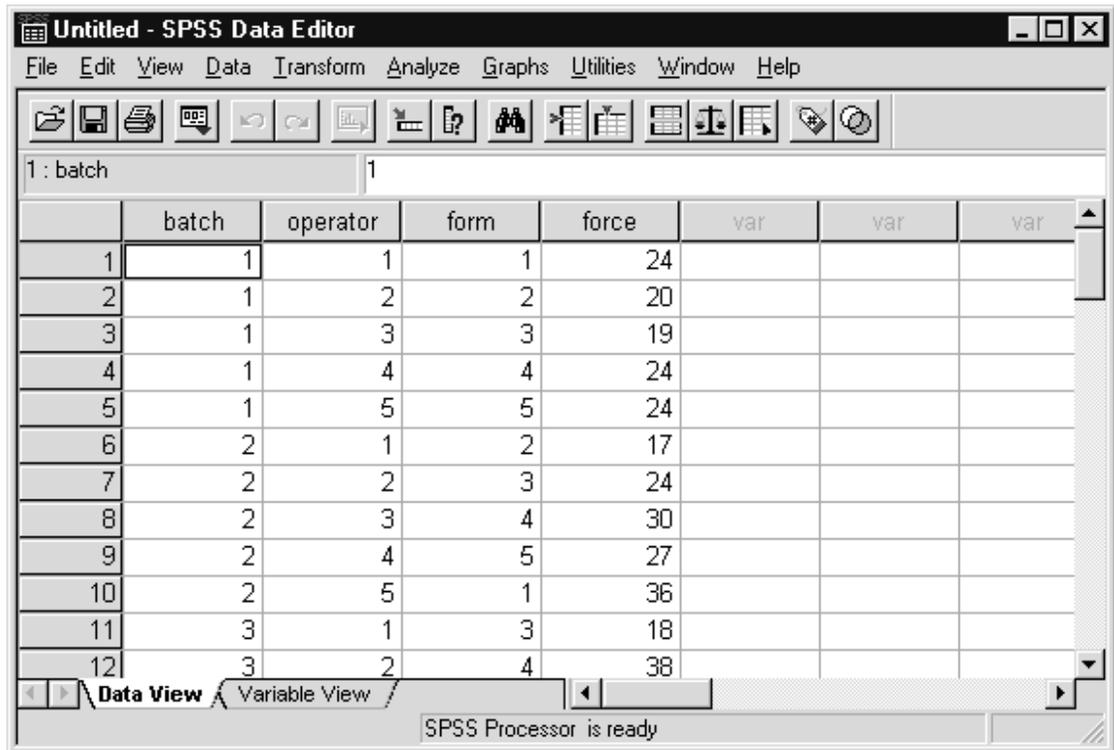
Click on **File..Open..Data**

Move to the **c:\Train\Anova** directory

Select **SPSS Portable (.por)** from Files of Type drop-down list

Double click on **Latinsq**

Figure 10.1 Data from Latin Square Design



The screenshot shows the SPSS Data Editor window titled "Untitled - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Window, and Help. The toolbar contains various icons for file operations and data manipulation. The data grid shows 12 rows and 8 columns. The first column is labeled "1 : batch" and contains values 1 through 12. The second column is labeled "batch" and contains values 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3. The third column is labeled "operator" and contains values 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2. The fourth column is labeled "form" and contains values 1, 2, 3, 4, 5, 2, 3, 4, 5, 1, 3, 4. The fifth column is labeled "force" and contains values 24, 20, 19, 24, 24, 17, 24, 30, 27, 36, 18, 38. The sixth, seventh, and eighth columns are labeled "var" and are currently empty. The status bar at the bottom indicates "SPSS Processor is ready".

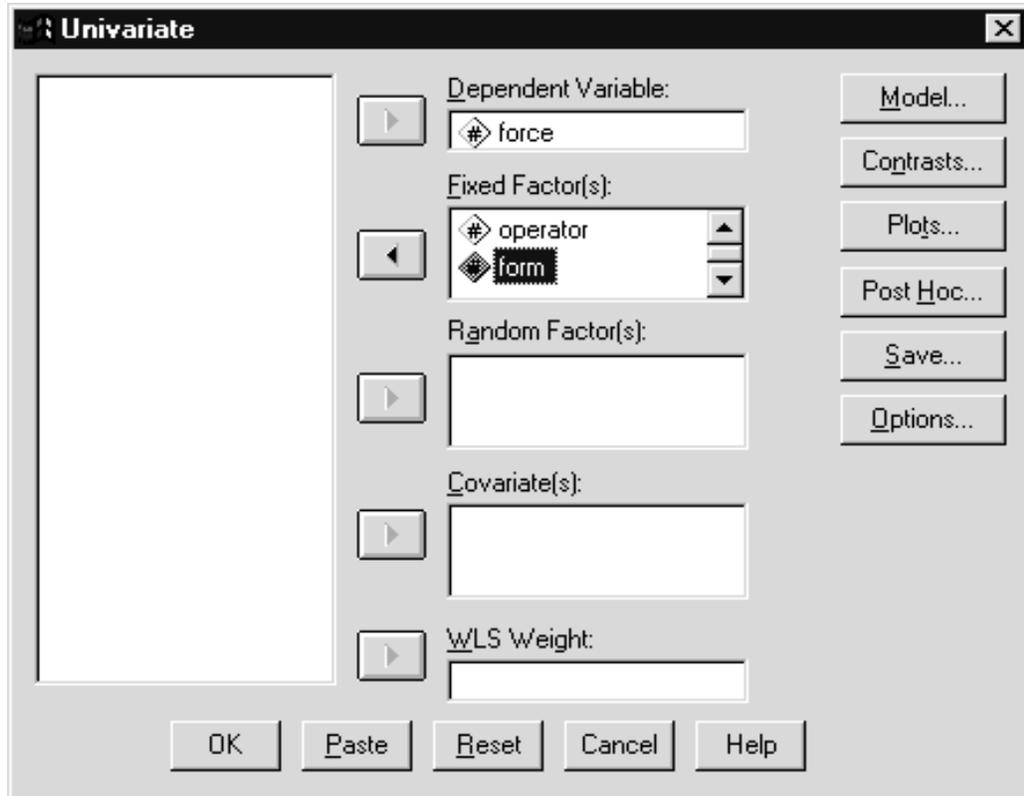
	batch	operator	form	force	var	var	var
1	1	1	1	24			
2	1	2	2	20			
3	1	3	3	19			
4	1	4	4	24			
5	1	5	5	24			
6	2	1	2	17			
7	2	2	3	24			
8	2	3	4	30			
9	2	4	5	27			
10	2	5	1	36			
11	3	1	3	18			
12	3	2	4	38			

Click **Analyze..General Linear Model..Univariate**

Move **Force** into the **Dependent Variable** list box

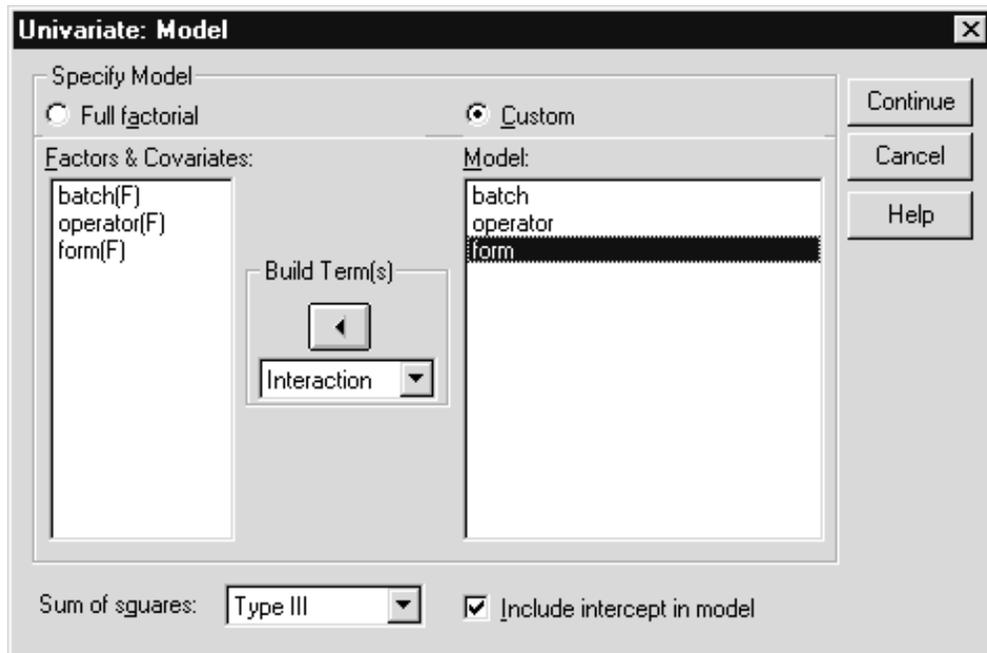
Move **Batch, Operator, and Form** into the **Fixed Factors** list box

Figure 10.2 Univariate Dialog Box



Click the **Model** pushbutton  
 Select **Main Effects** on the **Build Term(s)** drop-down list  
 Separately move **Batch**, **Operator**, and **Form** into the **Model** list box

Figure 10.3 Univariate: Model Dialog Box



Instead of the full model (all main effects and interactions), we will fit a custom model consisting of only main effects. The residual variation from this model will be used as the error term.

Click on **Continue** to process Model  
Click on **OK** to run the analysis.

The following syntax will also run the analysis.

```
UNIANOVA
  force BY batch operator form
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /DESIGN = batch operator form .
```

The DESIGN subcommand indicates that only a main effects model will be tested. The remaining effects will be pooled together into a residual term, which will be used as the error term in significance testing. This is why the assumption of no interactions is so important. Since Unianova is the univariate version of the GLM procedure, the GLM command could have been used instead.

**Figure 10.4 ANOVA Table**

Tests of Between-Subjects Effects					
Dependent Variable: FORCE					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	548.000 <sup>a</sup>	12	45.667	4.281	.009
Intercept	16129.000	1	16129.000	1512.094	.000
BATCH	68.000	4	17.000	1.594	.239
OPERATOR	150.000	4	37.500	3.516	.040
FORM	330.000	4	82.500	7.734	.003
Error	128.000	12	10.667		
Total	16805.000	25			
Corrected Total	676.000	24			

a. R Squared = .811 (Adjusted R Squared = .621)

We see the tests of main effects performed using the residual as the error term. The main effect of Form (formulation) is of greatest interest and is highly significant. For some designs, if there were replication within each cell, the within-cell error term could be used for significance testing.

### COMPLEX DESIGNS

There are additional variants along this theme of Latin Square designs (Greco-Latin Square, etc.) as well as other classes of designs (for example, fractional factorial). They can be set up in GLM in much the same way as demonstrated above. Such designs generally demand more knowledge of the user to plan the experiment, to understand which effects (main effects, main effects and some two-way interactions, etc.) can be tested, and to interpret the results. For those who need to perform such studies, SPSS Trial Run is a design of experiments program that can generate a variety of complex designs. It can then analyze the results using the GLM procedure, which is included in the program.

### RANDOM EFFECTS MODELS

The preceding designs in this course contained only single random factors: plant variation within group, subject variation within group, and Y variation within levels of A. In SPSS, by default GLM assumes there is a single random factor reflected in the case to case variation. GLM can accommodate multiple random factors quite easily using dialog boxes when the random factors are crossed with the other factors, and can be run using syntax when the random factors are nested. This is because the nesting operation cannot be expressed currently in the GLM – General Factorial Model dialog box. Most applied statistics books that discuss experimental design either cover the common designs or supply rules to determine the correct error terms in the presence of multiple random effects (for example, Kirk (1982) or Milliken and Johnston (1984)).

To illustrate we will consider a simple two random-effect design. An experiment is performed in which subjects (5) inflate rubber rafts (6). The time it takes to inflate each raft is recorded. Rafts are sampled from a production line and each subject inflates each raft once. Here we have a completely crossed (each subject inflates each raft) two-factor design with both factors assumed random. In this case the interaction term is used to test each of the main effects, and if there had been replications (if each subject inflated each raft several times), the interaction itself would be tested using the within-cells error term.

Click on **File..Open..Data**

Move to the **c:\Train\Anova** directory (if necessary)

Select **SPSS Portable (.por)** from the Files of Type drop-down list

Double click **raft.por** to open the file

Click **No** when asked to save the Data Editor's contents

Figure 10.5 Raft Data

The screenshot shows the SPSS Data Editor window titled 'Untitled - SPSS Data Editor'. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Window, and Help. Below the menu bar is a toolbar with various icons. The main data area shows a grid with the following data:

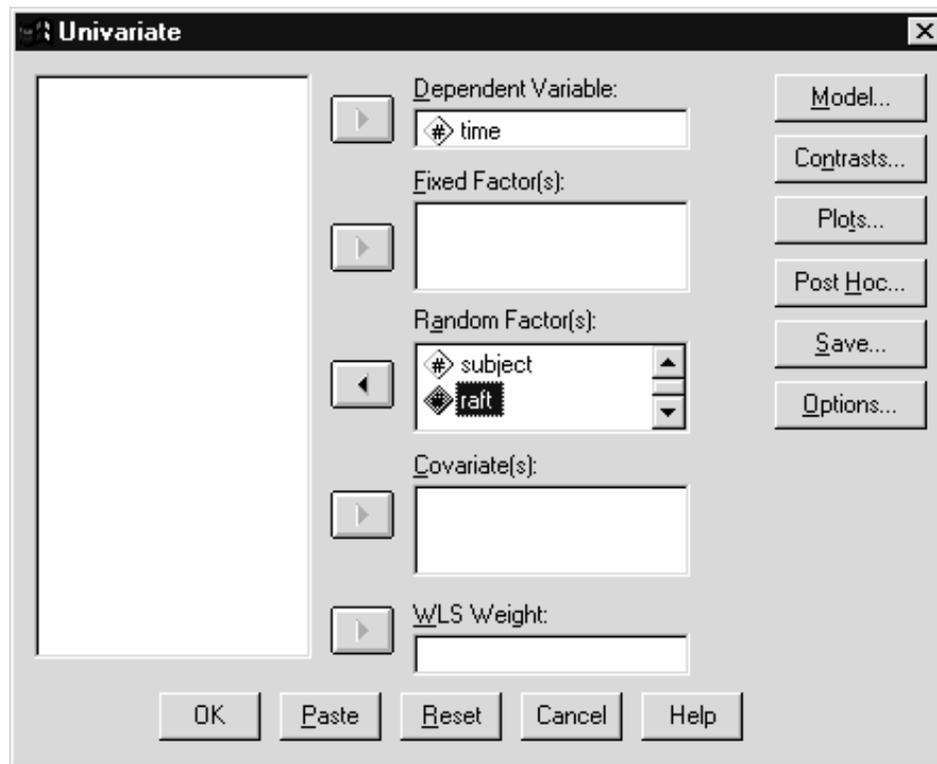
	subject	raft	time	var	var	var	var
1	1.00	1.00	2.79				
2	1.00	2.00	8.63				
3	1.00	3.00	12.24				
4	1.00	4.00	5.82				
5	1.00	5.00	3.11				
6	1.00	6.00	13.99				
7	2.00	1.00	6.93				
8	2.00	2.00	8.91				
9	2.00	3.00	1.05				
10	2.00	4.00	2.06				
11	2.00	5.00	2.82				
12	2.00	6.00	.86				

The status bar at the bottom indicates 'SPSS Processor is ready'.

Notice the data are arranged so each subject by raft combination appears as a different case. If we structured the data as we ordinarily would for repeated measures, each respondent on a single row of data, we would not be able to declare raft as a random factor (there is no Random Factor(s) list box in the General Linear Model – Repeated Measures dialog box).

- Click on **Analyze..General Linear Model..Univariate**
- Move **time** into the **Dependent Variable** list box
- Move **subject** and **raft** into the **Random Factor(s)** list box

Figure 10.6 Univariate Dialog Box



Click **OK** to run the analysis

The following syntax will run the analysis

Figure 10.7 Syntax for Two Random Effects Analysis

```
UNIANOVA
  time BY subject raft
  /RANDOM = subject raft
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /DESIGN = subject raft subject*raft .
```

The Random subcommand declares both subject and raft to be random factors.

Figure 10.8 Results

Tests of Between-Subjects Effects						
Dependent Variable: TIME						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	1822.756	1	1822.756	59.186	.036
	Error	45.462	1.476	30.797 <sup>a</sup>		
SUBJECT	Hypothesis	182.806	4	45.701	1.383	.275
	Error	660.815	20	33.041 <sup>b</sup>		
RAFT	Hypothesis	90.683	5	18.137	.549	.737
	Error	660.815	20	33.041 <sup>b</sup>		
SUBJECT * RAFT	Hypothesis	660.815	20	33.041	.	.
	Error	.000	0	. <sup>c</sup>		

a. 1.000 MS(SUBJECT) + MS(RAFT) - 1.000 MS(SUBJECT \* RAFT)  
 b. MS(SUBJECT \* RAFT)  
 c. MS(Error)

We have tests for both subject and raft main effects. Notice that the interaction could not be tested (the error term for it has 0 degrees of freedom) because there were no replications.

**Extensions**

A specific extension to the multiple random effects model is that in which random effects are nested within random effects. A common example of this involves analyzing student test scores within schools when the schools are sampled from school districts. A variation of random effect models, named hierarchical linear analysis, can be applied to such data. SPSS will perform such analyses for a balanced design, but does not currently handle the general case. Specialized programs are available to run hierarchical linear analysis.

**SUMMARY**

Armed with knowledge of experimental design and the appropriate error terms, incomplete and random effects designs can be tested using SPSS.



# References

Andrews, F. M., Klem, L., Davidson, T. N. O'Malley, P. M., and W. L. Rogers, A Guide for Selecting Statistical Techniques for Analyzing Social Science Data, Ann Arbor: Institute for Social Research, University of Michigan, 1981.

Bock, R. D., Multivariate Statistical Methods in Behavioral Research, New York: McGraw-Hill, 1975.

Conover, W. J., Practical Nonparametric Statistics, 2<sup>nd</sup> edition, New York: Wiley, 1980.

Crowder, M. J. and D. J. Hand, Analysis of Repeated Measures, London: Chapman and Hall, 1990.

Finn, J. D., A General Model for Multivariate Analysis, New York: Holt, Rinehart and Winston, 1974.

Hand, D. J. and C. C. Taylor, Multivariate Analysis of Variance and Repeated Measures, London: Chapman and Hall, 1987.

Hakstain, A. R., J. C. Roed, and J. C. Lind, *Two Sample T2 Procedure and the Assumption of Homogeneous Covariance Matrices*, Psychological Bulletin, 86, Pgs. 1255-1263, 1979.

Huberty, C. J., *Multivariate Analysis versus Multiple Univariate Analyses*, Psychological Bulletin Volume 105, Pgs. 302-308, 1989.

Kendall, M. G., and A. Stuart, The Advanced Theory of Statistics, Volume 3: Design and Analysis, and Time Series, New York: Hafner, 1968.

Kirk, R. E., Experimental Design: Procedures for the Behavioral Sciences, 2<sup>nd</sup> edition, Belmont, CA: Brooks/Cole, 1982.

Klockars, Alan J. and Sax, G., Multiple Comparisons, SAGE Quantitative Applications Series, Thousand Oaks CA: Sage, 1986.

Lindsey, J. K., Models for Repeated Measures, Oxford: Clarendon Press, 1993.

Looney, S. W., and W. Stanley, *Exploratory Repeated Measures Analysis for Two or More Groups*, The American Statistician, Volume 43, No. 4, Pgs 220-225, 1989.

McCullagh, P. and J. A. Nelder, Generalized Linear Models, 2<sup>nd</sup> edition, London: Chapman and Hall, 1989.

Milliken, G. A., and D. E. Johnson, Analysis of Messy Data, Volume 1: Designed Experiments, New York: Van Nostrand Reinhold, 1984.

Montgomery, D. C., Design and Analysis of Experiments, 2<sup>nd</sup> edition, New York: Wiley, 1984.

Morrison, D. F., Multivariate Statistical Methods, 2<sup>nd</sup> edition, New York: McGraw-Hill, 1976.

Olson, C. L. *On Choosing a Test Statistic in Multivariate Analysis of Variance*, Psychological Bulletin, Volume 83, No. 4 Pgs. 579-586, 1976.

Scheffe, H. The Analysis of Variance, New York: Wiley, 1959.

Searle, S. R., Linear Models for Unbalanced Data, New York: Wiley, 1987.

Tukey, J. W., Exploratory Data Analysis, Reading MA.: Addison Wesley, 1977.

Wilcoxon, R. Statistics for the Social Sciences, Academic Press, New York, 1966.

Wilcoxon, R. Introduction to Robust Estimation and Hypothesis Testing, Academic Press, New York, 1997.

Winer, B. J., Statistical Principles in Experimental Design, 2<sup>nd</sup> edition, New York: McGraw-Hill, 1971.

# Exercises

## Note on Exercise Data

The exercise file for this class (Workload.por) is located in the c:\Train\Anova folder on your training machine. If you are not working in an SPSS Training center, the training files can be copied from the floppy disk that accompanies this course guide. If you are running SPSS Server (click File..Switch Server to check), then you should copy these files to the server or a machine that can be accessed (mapped from) the computer running SPSS Server.

## Note About the Exercises

These exercises are based on a single, rich data file. You will perform a variety of analyses (for example, a one-factor and a three-factor ANOVA) on the same data. Typically, if three factors were believed to be relevant, then a one-factor ANOVA would not be run. Thus analyses suggested here conform to the topical sequence in the training guide and are not necessarily optimal to answer a specific research question. In fact, some analyses that might be performed on this data (doubly-multivariate analysis of variance) are not discussed in the course.

## Chapter 3 One-Factor ANOVA

Open the SPSS portable file Workload.por. This file contains data from an experimental investigation of the effects of a training workshop into stress and workload reduction techniques in airline pilots. The variables are in four sets - descriptive grouping information, workload measures taken before a flight, workload measures taken after a training course on stress and workload management, and a follow-up set of workload measures taken three months after the training course. Each pilot was measured once each (to avoid distraction) per flight and the two subsequent flights were on the same route for comparative purposes. In all two hundred pilots were measured over three time periods.

AGE	Age in years
HRSEXP	Previous flying experience in flying hours
TYPE	Type of aircraft cockpit (1=Automated, 2=Manual)
ROUTE	Designation of journey (1=Short Haul, 2=Medium Haul, 3=Long Haul)
STAGE	Stage of flight (1=Take Off, 2=Cruise, 3=Approach, 4=Landing)
FLYTIME	Length of flight (measured in seconds, presented in date/time format)

(Before Stress and Workload Training Course)

HEART	Heart Rate (Beats Per Minute)
BLOOD	Blood pressure (mmhg)
TEMP	Core Body Temperature (deg. f)

STRESS	Stress Rating (1=Low stress up to 7=High stress)
CAPACITY	Spare Mental Capacity Rating (1=All used up to 10 None used up)
ATTEN	Percent of attention remaining (in percent)
TIRED	Tiredness Rating scale (1=Invigorated up to 10=Asleep)

(After Stress and Workload Training Course)

HEART2	Heart Rate (after training)
BLOOD2	Blood pressure (after training)
TEMP2	Core Body Temperature (after training)
STRESS2	Stress rating (after training)
CAPACIT2	Spare Mental Capacity (after training)
ATTEN2	Percent of attention remaining (after training)
TIRED2	Tiredness ratings (after training)

(Three Months After the Stress and Workload Training Course)

HEART3	Heart Rate (after 3 months)
BLOOD3	Blood Pressure (after 3 months)
TEMP3	Core Body Temperature (after 3 months)
STRESS3	Stress Rating (after 3 months)
CAPACIT3	Spare Mental Capacity (after 3 months)
ATTEN3	Percent of attention remaining (after 3 months)
TIRED3	Tiredness Rating (after 3 months)

Familiarize yourself with the variables and data within this dataset by using the Frequencies, Descriptives and Explore procedures (with any associated graphical plots you choose).

Using the Means procedure and error bar graphs, compare the mean *stress levels* (use the **stress** variable, which measures stress before the training course) at different flight stages (use the variable *type*). Recall that a pilot was tested at a single flight stage. Before performing a one-factor ANOVA, explore the data (asking for means, error bars etc.) and try to predict the outcome of the analysis. Do you think there will be a significant difference between the groups?

Perform a one-factor ANOVA, testing for stage differences in stress level. If differences are found, perform post hoc tests to explore these differences in more detail. How would you summarize the results?

If the assumptions of ANOVA were not met, perform a nonparametric test of group (stage) differences in stress. Are the results consistent with the ANOVA analysis?

### Chapter 4 Multi-Way Univariate ANOVA

Open the SPSS portable file Workload.por. Now we are going to examine stress differences as a function of flight stage, route, and type of aircraft. Run an exploratory analysis examining stress (use the stress variable, which reports stress before taking the training course) within subgroups based on flight stage (stage), length of route (route) and aircraft type (type). Does the stress measure conform to the ANOVA assumptions?

Perform a three-factor ANOVA of stress with type, route, and stage as the factors. Are there significant interactions among aircraft type, route length, and flight stage as they relate to stress? If there are no interactions, but there are significant main effects, then perform the appropriate post hoc tests to identify which subgroups differ from each other.

### Chapter 5 Multivariate Analysis of Variance

Open the SPSS portable file Workload.por. Perform a three-factor (route, type and stage) multivariate analysis of variance on several physiological measures of stress: blood pressure (blood), heart rate (heart), and body temperature (temp). Which effects and interactions are significant? Examine the univariate results. Are the effects consistent across the three dependent measures?

Request a profile plot to examine the three-way interaction of route by stage by type as it relates to core body temperature (temp)? Describe the nature of the interaction.

For those with extra time: Perform a multivariate analysis using the same factors, but on the subjective measures of workload (stress, capacity, atten, and tired)? Are the results similar to what you found for the physiological measures?

### Chapter 6 Within-Subject Designs: Repeated Measures

Open the SPSS portable file Workload.por. Perform an exploratory data analysis on stress measured at the three time points (stress, stress2, stress3). Run a one-factor repeated-measures analysis examining the stress measure (stress, stress2, stress3) at the three time points of the study.

If there is a significant main effect of the time factor, explore the nature of it with post hoc tests and plots.

**Chapter 7**      **Between- and Within-Subject ANOVA: Repeated Measures**  
Open the SPSS portable file Workload.por. Perform a repeated measures analysis with time (three levels) as a within-subject factor and type, route, and stage as between-subject factors. The dependent measure will be stress (stress, stress2, stress3). Which effects are significant?

If there are significant interactions, explore them using simple effects and profile plots.

*For those with extra time:* Run the same analysis using one of the physiological measures.

**Chapter 9**      **Analysis of Covariance**  
Open the SPSS portable file Workload.por. We will add a covariate to the analysis run in Chapter 4. Run an analysis of covariance on stress with type, route, and stage as between-subject factors and age as a covariate. The dependent measure will be stress.

Test for the parallelism of slope assumption (test for a four-way interaction among, age, type, route and stage).

If the parallelism of slopes assumption is met, then run the analysis of covariance and assess the relevance of the age covariate. Ask for parameter estimates and interpret the relationship between age and stress (even if nonsignificant).