Advanced Data Modelling

Kevin O'Brien

April 18, 2013

$$\left(\frac{Q}{E}\right)^2 \times \pi \times (100 - \pi) \geq n$$

1 One-way MANOVA in SPSS

1.1 Introduction

The one-way analysis of variance procedure (one-way ANOVA) is used to determine whether there are any differences between multiple independent groups on a continuous dependent variable. Suppose there was a study carried out on the mathematics results from three schools A, B and C. A one-way ANOVA procedure is used to test the hypothesis that the average score in each of the three schools is the same. The alternative hypothesis is that the average score of one school is different to the others.

H_0 $\mu_A = \mu_B = \mu_C$

H_1 There exists $\mu_i \neq \mu_j$ for some pairing of schools A B and C.

It is important to realize that the one-way ANOVA is an omnibus test statistic and cannot tell you which specific groups were significantly different from each other; it only tells you that at least two groups were different.

The one-way multivariate analysis of variance (one-way MANOVA) is used to determine whether there are any differences between independent groups on **more than one** continuous dependent variable. In this regard, it differs from a one-way ANOVA, which only measures one dependent variable.

For example, you could use a one-way MANOVA to understand whether there were differences in the perceptions of attractiveness and intelligence of drug users in movies (i.e., the two dependent variables are "perceptions of attractiveness" and "perceptions of intelligence", whilst the independent variable is "drug users in movies", which has three independent groups: "non-user", "experimenter" and "regular user").

Alternately, you could use a one-way MANOVA to understand whether there were differences in students' short-term and long-term recall of facts based on three different lengths of lecture (i.e., the two dependent variables are "short-term memory recall" and "long-term memory recall", whilst the independent variable is "lecture duration", which has four independent groups: "30 minutes", "60 minutes", "90 minutes" and "120 minutes" , considered as categorical data, even though numeric data is valid also).

Again, it is important to realize that the one-way MANOVA is an omnibus test statistic and cannot tell you which specific groups were significantly different from each other; it only tells you that at least two groups were different. Since you may have three, four, five or more groups in your study design, determining which of these groups differ from each other is important. You can do this using a **post-hoc** test.

1.2 Assumptions

When you choose to analyze your data using a one-way MANOVA, part of the process involves checking to make sure that the data you want to analyze can actually be analyzed using a one-way MANOVA. You need to do this because it is only appropriate to use a one-way MANOVA if your data "passes" nine assumptions that are required for a one-way MANOVA to give you a valid result. Do not be surprised if, when analyzing your own data using SPSS, one or more of these assumptions is violated. This is not uncommon when working with real-world data. However, even when your data fails certain assumptions, there is often a solution to overcome this.

In practice, checking for these nine Assumptions adds some more time to your analysis, requiring you to work through additional procedures in SPSS when performing your analysis, as well as thinking a little bit more about your data. These nine assumptions are presented below:

- Assumption 1: Your two or more dependent variables should be measured at the interval or ratio level (i.e., they are continuous). Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

- Assumption 2: Your independent variable should consist of two or more categorical, independent groups. Example independent variables that meet this criterion include ethnicity (e.g., 3 groups: Caucasian, Afro-Caribbean and South Asian), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth.

- Assumption 3: You should have independence of observations, which means that there is no relationship between the observations in each group or between the groups themselves. For example, there must be different participants in each group with no participant being in more than one group. This is more of a study design issue than something you can test for, but it is an important assumption of the one-way MANOVA.

- Assumption 4: You should have an adequate sample size. Although the larger your sample size, the better; for MANOVA, you need to have more cases in each group than the number of dependent variables you are analyzing.

- Assumption 5: There are no univariate or multivariate outliers. First, there can be no (univariate) outliers in each group of the independent variable for any of the dependent variables. This is a similar assumption to the one-way ANOVA, but for each dependent variable that you have in your MANOVA analysis. Univariate outliers are often just called outliers and are the same type of outliers you will have come across if you have conducted t-tests or ANOVAs. We refer to them as univariate in this guide to distinguish them from multivariate outliers. Multivariate outliers are cases which have an unusual combination of scores on the dependent variables.

  (Approaches for detecting outliers (1) detect univariate outliers using boxplots, (2) check for multivariate outliers using a measure called **Mahalanobis** distance.)

- Assumption 6: There is multivariate normality. Unfortunately, multivariate normality is a particularly tricky assumption to test for and cannot be directly tested in SPSS. Instead, normality of each of the dependent variables for each of the groups of the independent variable is often used in its place as a best 'guess' as to whether there is multivariate normality. You can test for this using the Shapiro-Wilk test of normality, which is easily tested for using SPSS.

- Assumption 7: There is a linear relationship between each pair of dependent variables for each group of the independent variable. If the variables are not linearly related, the power of the test is reduced. You can test for this assumption by plotting a scatterplot matrix for each group of the independent variable. In order to do this, you will need to split your data file in SPSS before generating the scatterplot matrices.

- Assumption 8: There is a homogeneity of variance-covariance matrices. You can test this Assumption in SPSS using **Box's M test** of equality of covariance. If your data fails this

- Assumption, you may also need to use SPSS to carry out **Levene's test** of homogeneity of variance to determine where the problem may lie.

- Assumption 9: There is no multicollinearity. Ideally, you want your dependent variables to be moderately correlated with each other. If the correlations are low, you might be better off running separate one-way ANOVAs, and if the correlation(s) are too high (greater than 0.9), you could have multicollinearity. This is problematic for MANOVA and needs to be screened out.

You can check assumptions 5, 6, 7, 8 and 9 using SPSS. Before doing this, you should make sure that your data meets Assumptions 1, 2, 3 and 4, although you don't need SPSS to do this. Just remember that if you do not run the statistical tests on these assumptions correctly, the results you get when running a one-way MANOVA might not be valid.

2 Implementation

In the section,we will illustrate the SPSS procedure to perform a one-way MANOVA assuming that no assumptions have been violated. First, we set out the example we use to explain the one-way MANOVA procedure in SPSS.

2.1 Example

The pupils at a high school come from three different primary schools. The headteacher wanted to know whether there were academic differences between the pupils from the three different primary schools. As such, she randomly selected 20 pupils from School A, 20 pupils from School B and 20 pupils from School C, and measured their academic performance as assessed by the marks they received for their end-of-year English and Maths exams. Therefore, the two dependent variables were “English score” and “Maths score”, whilst the independent variable was “School”, which consisted of three categories: “School A”, “School B” and “School C”.

2.2 Setup in SPSS

In SPSS, we separated the groups for analysis by creating a grouping variable called School (i.e., the independent variable), and gave the three categories of the independent variable the labels $SchoolA$, $SchoolB$ and $SchoolC$. The two dependent variables were labelled $English_Score$ and $Maths_Score$, respectively. It is also recommended that you create a fourth variable, $subject_id$, to act as a case number. This latter variable is required to test whether there are any multivariate outliers (i.e., part of Assumption 5 above).

2.3 Test Procedure in SPSS

The following steps below show you how to analyse your data using a one-way MANOVA in SPSS when the nine assumptions in the previous section, ssumptions, have not been violated. At the end of these 14 steps, we show you how to interpret the results from this test.

`Click Analyze > General Linear Model > Multivariate`

on the top menu as shown below:

You will be presented with the Multivariate dialogue box:

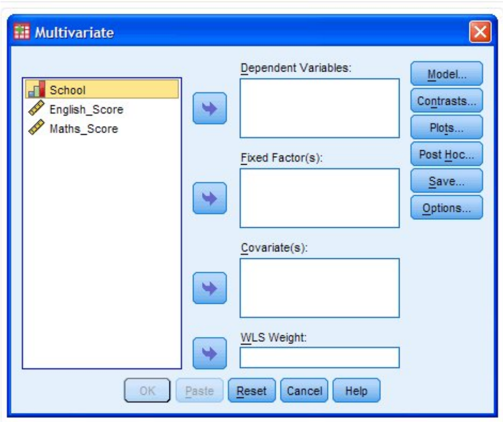

Figure 1: MANOVA

2.4 SPSS Output of the One-Way MANOVA

SPSS produces many different tables in its one-way MANOVA analysis. In this section, we show you only the main tables required to understand your results from the one-way MANOVA and Tukey post-hoc tests.

> For a complete explanation of the output you have to interpret when checking your data for the nine assumptions required to carry out a one-way MANOVA.
>
> This includes relevant boxplots, scatterplot matrix and Pearson's correlation coefficients, and output from your Mahalanobis distance test, Shapiro-Wilk test for normality, and Box's M test of equality of covariance, and if required, Levene's test of homogeneity of variance.

However, the emphasis will be placed on the four main tables you need to understandthe one-way MANOVA results, assuming that your data has already met the nine assumptions required for a one-way MANOVA to give you a valid result.

2.5 Descriptive Statistics

The first important one is the Descriptive Statistics table. This table is very useful as it provides the mean and standard deviation for the two different dependent variables, which have been split by the independent variable. In addition, the table provides "Total" rows, which allows means and standard deviations for groups only split by the dependent variable to be known.

2.6 Multivariate Tests

The Multivariate Tests table is where we find the actual result of the one-way MANOVA. You need to look at the second Effect, labelled "School", and the Wilks' Lambda row (highlighted in red). To determine whether the one-way MANOVA was statistically significant you need to look at the **Sig.** column. We can see from the table that we have a "Sig." value of .000, which means $p < .0005$. Therefore, we can conclude that this school's pupils academic performance was significantly dependent on which prior school they had attended ($p < .0005$).

Descriptive Statistics

	School	Mean	Std. Deviation	N
English_Score	School A	75.6000	8.22960	20
	School B	62.7000	9.10234	20
	School C	61.5500	7.14124	20
	Total	66.6167	10.30401	60
Maths_Score	School A	43.9000	8.46603	20
	School B	40.7500	8.16201	20
	School C	30.7500	7.71789	20
	Total	38.4667	9.78145	60

Figure 2: MANOVA

Multivariate Tests[d]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Intercept	Pillai's Trace	.989	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Wilks' Lambda	.011	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Hotelling's Trace	86.967	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Roy's Largest Root	86.967	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
School	Pillai's Trace	.616	12.681	4.000	114.000	.000	.308	50.724	1.000
	Wilks' Lambda	.450	13.735[a]	4.000	112.000	.000	.329	54.938	1.000
	Hotelling's Trace	1.075	14.782	4.000	110.000	.000	.350	59.128	1.000
	Roy's Largest Root	.915	26.072[c]	2.000	57.000	.000	.478	52.144	1.000

a. Exact statistic

b. Computed using alpha = .05

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

d. Design: Intercept + School

Figure 3: MANOVA

2.7 Reporting the Result (without follow-up tests)

You could report the result of this test as follows:

There was a statistically significant difference in academic performance *based on a pupil's prior school* , $F(4, 112) = 13.74$*,* $p < .0005$*; Wilk's Lambda = 0.450, partial* η^2 *= 0.33.*

If you had not achieved a statistically significant result, you would not perform any further follow-up tests. However, as our case shows that we did, we will continue with further tests.

2.8 Univariate ANOVAs

To determine how the dependent variables differ for the independent variable, we need to look at the Tests of Between-Subjects Effects table (highlighted in red):

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	English_Score	2434.233[a]	2	1217.117	18.114	.000	.389	36.228	1.000
	Maths_Score	1885.633[c]	2	942.817	14.295	.000	.334	28.591	.998
Intercept	English_Score	266266.817	1	266266.817	3962.769	.000	.986	3962.769	1.000
	Maths_Score	88781.067	1	88781.067	1346.134	.000	.959	1346.134	1.000
School	English_Score	2434.233	2	1217.117	18.114	.000	.389	36.228	1.000
	Maths_Score	1885.633	2	942.817	14.295	.000	.334	28.591	.998
Error	English_Score	3829.950	57	67.192					
	Maths_Score	3759.300	57	65.953					
Total	English_Score	272531.000	60						
	Maths_Score	94426.000	60						
Corrected Total	English_Score	6264.183	59						
	Maths_Score	5644.933	59						

a. R Squared = .389 (Adjusted R Squared = .367)

b. Computed using alpha = .05

c. R Squared = .334 (Adjusted R Squared = .311)

Figure 4: MANOVA

We can see from this table that prior schooling has a statistically significant effect on both English ($F(2, 57) = 18.11$; $p < .0005$; partial $\eta^2 = .39$) and Maths scores ($F(2, 57) = 14.30$; $p < .0005$; partial $\eta^2 = .33$). (It is important to note that, in practice, you should make an **alpha correction** to account for multiple ANOVAs being run, such as a Bonferroni correction. As such, in this case, we accept statistical significance at p ¡ .025.)

2.9 Multiple Comparisons

We can follow up these significant ANOVAs with Tukey's HSD post-hoc tests, as shown below in the Multiple Comparisons table:

Multiple Comparisons

Tukey HSD

Dependent Variable	(I) School	(J) School	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
English_Score	School A	School B	12.9000*	2.59214	.000	6.6622	19.1378
		School C	14.0500*	2.59214	.000	7.8122	20.2878
	School B	School A	-12.9000*	2.59214	.000	-19.1378	-6.6622
		School C	1.1500	2.59214	.897	-5.0878	7.3878
	School C	School A	-14.0500*	2.59214	.000	-20.2878	-7.8122
		School B	-1.1500	2.59214	.897	-7.3878	5.0878
Maths_Score	School A	School B	3.1500	2.56812	.443	-3.0300	9.3300
		School C	13.1500*	2.56812	.000	6.9700	19.3300
	School B	School A	-3.1500	2.56812	.443	-9.3300	3.0300
		School C	10.0000*	2.56812	.001	3.8200	16.1800
	School C	School A	-13.1500*	2.56812	.000	-19.3300	-6.9700
		School B	-10.0000*	2.56812	.001	-16.1800	-3.8200

Based on observed means.
The error term is Mean Square(Error) = 65.953.

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

Figure 5: MANOVA

The table above shows that for mean scores for English were statistically significantly different between School A and School B ($p < .0005$), and School A and School C ($p < .0005$), but not between School B and School C (p = .897). Mean maths scores were statistically significantly different between School A and School C ($p < .0005$), and School B and School C (p = .001), but not between School A and School B (p = .443). These differences can be easily visualised by the plots generated by this procedure, as shown below:

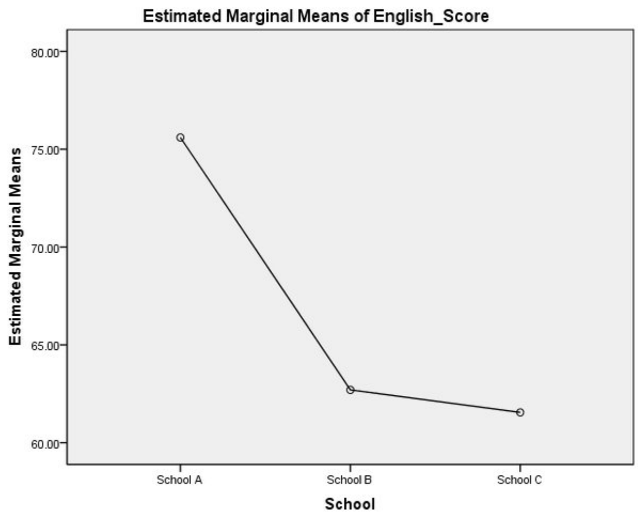

Figure 6: MANOVA

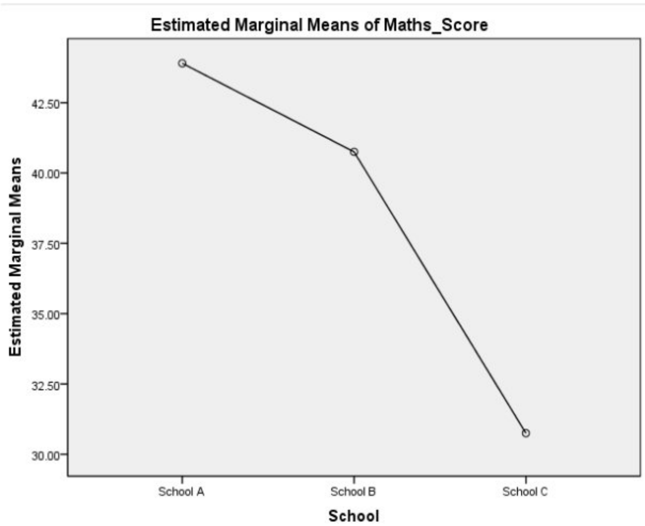

Figure 7: MANOVA