Contents

$$\sum$$

Discrete Mathematics

Mathematics for Computing

1 TWO-WAY ANALYSIS OF VARIANCE (TWO-WAY ANOVA)

The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors).

The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

Use two-way anova when you have one measurement variable and two nominal variables, and each value of one nominal variable is found in combination with each value of the other nominal variable.

- Two-way analysis of variance is based on two dimensions of classifications, or treatments. For example, in analyzing the level of achievement in a training program we could consider both the effect of the method of instruction and the effect of prior school achievement.
- Similarly, we could investigate gasoline mileage according to the weight category of the car and according to the grade of gasoline.
- In data tables, the treatments identified in the column headings are typically called the ***A treatments***; those in the row headings are called the ***B treatments***.

1.1 Laerd

Two-way ANOVA in SPSS Statistics

Introduction The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable. For example, you could use a two-way ANOVA to understand whether there is an interaction between gender and educational level on test anxiety amongst university students, where gender (males/females) and education level (undergraduate/postgraduate) are your independent variables, and test anxiety is your dependent variable. Alternately, you may want to determine whether there is an interaction between physical activity level and gender on blood cholesterol concentration in children, where physical activity (low/moderate/high) and gender (male/female) are your independent variables, and cholesterol concentration is your dependent variable.

The interaction term in a two-way ANOVA informs you whether the effect of one of your independent variables on the dependent variable is the same for all values of your other independent variable (and vice versa). For example, is the effect of gender (male/female) on test anxiety influenced by educational level (undergraduate/postgraduate)? Additionally, if a statistically significant interaction is found, you need to determine whether there are any ”simple main effects”, and if there are, what these effects are (we discuss this later in our guide).

Note: If you have three independent variables rather than two, you need a three-way ANOVA.

2 Assumptions

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.
- The groups must have the same sample size.

2.1 LAERD

When you choose to analyse your data using a two-way ANOVA, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using a two-way ANOVA. You need to do this because it is only appropriate to use a two-way ANOVA if your data ”passes” six Assumptions that are required for a two-way ANOVA to give you a valid result. In practice, checking for these six Assumptions means that you have a few more procedures to run through in SPSS Statistics when performing your analysis, as well as spend a little bit more time thinking about your data, but it is not a difficult task.

Before we introduce you to these six Assumptions, do not be surprised if, when analysing your own data using SPSS Statistics, one or more of these Assumptions is violated (i.e., is not met). This is not uncommon when working with real-world data rather than textbook examples, which often only show you how to carry out a two-way ANOVA when everything goes well! However, dont worry. Even when your data fails certain Assumptions, there is often a solution to overcome this. First, lets take a look at these six Assumptions:

Assumption 1: Your dependent variable should be measured at the continuous level (i.e., they are interval or ratio variables). Examples of continuous variables 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. You can learn more about interval and ratio variables in our article:] Types of Variable.

Assumption 2: Your two independent variables should each consist of two or more categorical, independent groups. Example independent variables that meet this criterion include gender (2 groups:] male or female), ethnicity (3 groups:] Caucasian, African American and Hispanic), profession (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 would test for, but it is an important Assumption of the two-way ANOVA. If your study fails this Assumption, you will need to use another statistical test instead of the two-way ANOVA (e.g., a repeated measures design). If you are unsure whether your study meets this Assumption, you can use our Statistical Test Selector, which is part of our enhanced guides.

Assumption 4: There should be no significant outliers. Outliers are data points within your data that do not follow the usual pattern (e.g., in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1

Assumption 5: Your dependent variable should be approximately normally distributed for each combination of the groups of the two independent variables. Whilst this sounds a little tricky, it is easily tested for using SPSS Statistics. Also, when we talk about the two-way ANOVA only requiring approximately normal data, this is because it is quite "robust" to violations of normality, meaning the Assumption can be a little violated and still provide valid results. You can test for normality using the Shapiro-Wilk test for normality, which is easily tested for using SPSS Statistics. In addition to showing you how to do this in our enhanced two-way ANOVA guide, we also explain what you can do if your data fails this Assumption (i.e., if it fails it more than a little bit).

Assumption 6: There needs to be homogeneity of variances for each combination of the groups of the two independent variables. Again, whilst this sounds a little tricky, you can easily test this Assumption in SPSS Statistics using Levenes test for homogeneity of variances. In our enhanced two-way ANOVA guide, we (a) show you how to perform Levenes test for homogeneity of variances in SPSS Statistics, (b) explain some of the things you will need to consider when interpreting your data, and (c) present possible ways to continue with your analysis if your data fails to meet this Assumption.

You can check Assumptions 4, 5 and 6 using SPSS Statistics. Before doing this, you should make sure that your data meets Assumptions 1, 2 and 3, although you dont need SPSS Statistics 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 two-way ANOVA might not be valid. This is why we dedicate a number of sections of our enhanced two-way ANOVA guide to help you get this right. You can find out about our enhanced content as a whole here, or more specifically, learn how we help with testing Assumptions here.

In the section, Test Procedure in SPSS Statistics, we illustrate the SPSS Statistics procedure to perform a two-way ANOVA Assuming that no Assumptions have been violated.

First, we set out the example we use to explain the two-way ANOVA procedure in SPSS Statistics.

3 Hypotheses

There are three sets of hypothesis with the two-way ANOVA. The null hypotheses for each of the sets are given below.

- The population means of the first factor are equal.
  This is like the one-way ANOVA for the row factor.
- The population means of the second factor are equal. *This is like the one-way ANOVA for the column factor.*
- There is no interaction between the two factors. *This is similar to performing a test for independence with contingency tables.*

4 Factors

The two independent variables in a two-way ANOVA are called factors. The idea is that there are two variables, factors, which affect the dependent variable. Each factor will have two or more levels within it, and the degrees of freedom for each factor is one less than the number of levels.

5 Main Effect

- The main effect involves the independent variables separately.
- The interaction is ignored for this part. Just the rows or just the columns are used, not mixed.
- This is the part which is similar to the one-way analysis of variance.
- Each of the variances calculated to analyze the main effects are like the between variances

6 Treatment Groups

- Treatement Groups are formed by making all possible combinations of the two factors.
- For example, if the first factor has 3 levels and the second factor has 2 levels, then there will be $3 \times 2 = 6$ different treatment groups.
- There may be more than one observation per treatment group.

7 Interaction Effect

- The interaction effect is the effect that one factor has on the other factor.
- The degrees of freedom here is the product of the two degrees of freedom for each factor.
- **Important:** The interaction effect can only be computed in the case of multiple measurements per treatment group.

8 Within Variation

- The ***Within variation*** is the sum of squares within each treatment group.
- You have one less than the sample size (remember all treatment groups must have the same sample size for a two-way ANOVA) for each treatment group.
- The total number of treatment groups is the product of the number of levels for each factor.
- The within variance is the within variation divided by its degrees of freedom.

8.1 Corn Example

- As an example, let's assume we're planting corn. The type of seed and type of fertilizer are the two factors we're considering in this example. This example has 15 treatment groups.
- There are 3-1=2 degrees of freedom for the type of seed, and 5-1=4 degrees of freedom for the type of fertilizer.
- There are $2 \times 4 = 8$ degrees of freedom for the interaction between the type of seed and type of fertilizer.

The data that actually appears in the table are samples. In this case, 2 samples from each treatment group were taken.

	Fert I	Fert II	Fert III	Fert IV	Fert V
Seed A-402	106, 110	95, 100	94, 107	103, 104	100, 102
Seed B-894	110, 112	98, 99	100, 101	108, 112	105, 107
Seed C-952	94, 97	86, 87	98, 99	99, 101	94, 98

9 F-Tests

There is an F-test for each of the hypotheses, and the F-test is the mean square for each main effect and the interaction effect divided by the within variance. The numerator degrees of freedom come from each effect, and the denominator degrees of freedom is the degrees of freedom for the within variance in each case.

10 Two-Way ANOVA Table

- It is assumed that main effect A has a levels (and A = a-1 df), main effect B has b levels (and B = b-1 df), n is the sample size of each treatment, and N = abn is the total sample size.
- Notice the overall degrees of freedom is once again one less than the total sample size.

$$\mathrm{MS}_{Trt} = c \times S_R^2$$
$$\mathrm{MS}_{Block} = r \times S_C^2$$

- r and c are the numbers of rows and columns respectively.
- S_R^2 is the variance of the Row means
- S_C^2 is the variance of the column means.

Take care when working with calculations that involve r and c: r is the number of rows, which is the number of subgroups for each of the treatments (which are arranged along the columns). c is the number of columns, which is the number of subgroups for each of the blocks (which are arranged along the rows).

Table 13.3 **Summary Table for Two-Way Analysis of Variance with More than One Observation per Cell**

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F ratio
Among treatment groups (A)	$K-1$	$SSA = \sum_{k=1}^{K} \frac{T_k^2}{nJ} - \frac{T^2}{N}$	$MSA = \frac{SSA}{K-1}$	$F = \frac{MSA}{MSE}$
Among treatment groups (B)	$J-1$	$SSB = \sum_{j=1}^{J} \frac{T_j^2}{nK} - \frac{T^2}{N}$	$MSB = \frac{SSB}{J-1}$	$F = \frac{MSB}{MSE}$
Interaction (between) factors (A and B) (I)	$(J-1)(K-1)$	$SSI = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\sum_{i=1}^{n} X \right)^2 - SSA - SSB - \frac{T^2}{N}$	$MSI = \frac{SSI}{(J-1)(K-1)}$	$F = \frac{MSI}{MSE}$
Sampling error (E)	$JK(n-1)$	$SSE = SST - SSA - SSB - SSI$	$MSE = \frac{SSE}{JK(n-1)}$	
Total (T)	$N-1$	$SST = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} X^2 - \frac{T^2}{N}$		

Figure 1:

11 Worked Example

Three varieties of potatoes are being compared for yield. The experiment was carried out by assigning each variety at random to four of twelve equal size plots, one being chosen in each of four locations. The following yields in bushels7 per 7 plot resulted: A bushel is about 36.4 litres.

	Potato		
Location	A	B	C
1	18	13	12
2	20	23	21
3	14	12	9
4	11	17	10

Additional Information

> Expect to be given S_R^2 and S_C^2 in an exam standard question, as well as the variance for the response variable. You will also be given the formula for MS_{Trt} and MS_{Block}.

- The variance of the Row means is : $S_R^2 = 19.037$. Therefore

$$\text{MS}_{Trt} = c \times S_R^2 = 3 \times 19.037 = 57.111$$

- the variance of the Column means is : $S_C^2 = 3.0625$. Therefore

$$\text{MS}_{Block} = r \times S_C^2 = 4 \times 3.0625 = 12.25$$

- Also the overall variance of the 12 observations is

$$\text{Var}(Y) = 21.6363$$

.

11.1 Solution

- Here the treatment is ***Location***
- The block variable is ***Type of Potato***

Step 1 : We have already be given information that we can incorporate into our solution, once we have done the necessary calculations.

(Remark: The following calculation for SS_{Tot} features in each ANOVA procedures on the course).

$$\text{Var}(Y) = \frac{SS_{tot}}{n-1}$$

$$21.6363 = \frac{SS_{tot}}{12-1}$$

$$SS_{tot} = 21.6362 \times 11 = 238$$

Source	df	SS	MS	F	p.values
Treatment			57.111		
Block			12.25		
Total					
Total	11	238			

Step 2 : We will now determine the Degrees of Freedom for Blocks, Treatments and Error.

- There are 4 types of treatment. The degrees of freedom for treament is therefore 3 $(r-1)$.
- There are 3 blocks. The degrees of freedom for blocks is therefore 2 (i.e. $c-1$).
- The total degrees of freedom should add up to 11. Therefore the degrees of freedom for error is 6. (i.e. $(r-1) \times (c-1)$.)
- **Remark:** Do not confuse $r-1$ with c. Here both are equal to 3, but this is a coincidence.

Source	df	SS	MS	F	p.values
Treatment	3		57.111		
Block	2		12.25		
Error	6				
Total	11	238			

Step 3 : In general, Mean Square Terms are computed by dividing the relevant SS terms by the corresponding degrees of freedom.

$$MS = \frac{SS}{df}$$

Rearranging this , we can say : $SS = MS \times df$. Therefore

- $\text{SS}_{Trt} = 3 \times 57.111 = 171.33$
- $\text{SS}_{Block} = 2 \times 12.25 = 24.50$

Importantly

$$\boxed{SS_{tot} = SS_{trt} + SS_{block} + SS_{error}}$$

We can compute SS_{error}

$$238 = 171.33 + 24.5 + SS_{error}$$

Necessarily $SS_{error} = 42.166$

Source	df	SS	MS	F	p.values
Treatment	3	171.333	57.111		
Block	2	24.50	12.25		
Error	6	42.166			
Total	11	238			

Step 4 : We can now complete the table as follows

- MS_{error}

$$MS_{error} = \frac{SS_{error}}{df_{error}} - \frac{42.166}{6} = 7.033$$

- Test Statistic 1

$$F_{Trt} = \frac{MS_{Trt}}{MS_{e}rror} = \frac{57.111}{7.033} = 8.12$$

- Test Statistic 1

$$F_{Block} = \frac{MS_{Block}}{MS_{e}rror} = \frac{12.15}{7.033} = 1.74$$

Source	df	SS	MS	F	p.values
Treatment	3	171.333	57.111	8.12	
Block	2	24.50	12.250	1.74	
Error	6	42.166	7.033		
Total	11	238			

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F-crit*
Seed	512.8667	2	256.4333	28.283	0.000008	3.682
Fertilizer	449.4667	4	112.3667	12.393	0.000119	3.056
Interaction	143.1333	8	17.8917	1.973	0.122090	2.641
Within	136.0000	15	9.0667			
Total	**1241.4667**	**29**				

Figure 2:

12 Summary

The following results are calculated using the Quattro Pro spreadsheet. It provides the p-value and the critical values are for alpha = 0.05.

From the above results, we can see that the main effects are both significant, but the interaction between them isn't. That is, the types of seed aren't all equal, and the types of fertilizer aren't all equal, but the type of seed doesn't interact with the type of fertilizer.

13 Interaction

- Interaction in a two-factor experiment means that the two treatments are not independent, and that the effect of a particular treatment in one factor differs according to levels of the other factor.
- For example, in studying automobile mileage a higher-octane gasoline may improve mileage for certain types of cars but not for others. Similarly, the effectiveness of various methods of instruction may differ according to the ability levels of the students.
- In order to test for interaction, more than one observation or sampled measurement (i.e., replication) has to be included in each cell of the two-way data table.
- Section 13.4 presents the analytical procedure that is appropriate when there is only one observation per cell, and in which interaction between the two factors cannot be tested.
- The analytical procedure is extended to include replication and the analysis of interaction effects in Section 13.5.

The linear model for the two-way analysis of variance model with one observation per cell (with no replication) is

$$X_{jk} = \mu + \beta_j + \alpha_k + \varepsilon_{jk} \tag{13.3}$$

where $\mu =$ the overall mean regardless of any treatment
$\beta_j =$ effect of the treatment j or block j in the B dimension of classification
$\alpha_k =$ effect of the treatment k in the A dimension of classification
$\varepsilon_{jk} =$ the random error associated with the process of sampling

14 13.4 THE RANDOMIZED BLOCK DESIGN

(TWO-WAY ANOVA, ONE OBSERVATION PER CELL)

- The two-way analysis of variance model in which there is only one observation per cell is generally referred to as the randomized block design, because of the principal use for this model. What if we extend the idea of using paired observations to compare two sample means (see Section 11.3) to the basic one-way analysis of variance model, and have groups of k matched individuals assigned randomly to each treatment level? In analysis of variance, such matched groups are called blocks, and because the individuals (or items) are randomly assigned based on the identified block membership, the design is referred to as the randomized block design.
- In such a design the blocks dimension is not a treatment dimension as such. The objective of using this design is not for the specific purpose of testing for a blocks effect. Rather, by being able to assign some of the variability among subjects to prior achievement, for example, the MSE can be reduced and the resulting test of the A treatments effect is more sensitive.

Summary Table for One-Way Analysis of Variance (Treatment Groups Need Not Be Equal) Source of variation Degrees of freedom (df) Sum of squares (SS)

The linear model for the two-way analysis of variance model with one observation per cell (with no replication) is

Table 13.2 is the summary table for the two-way analysis of variance without replication. As compared with Table 13.1 for the one-way analysis of variance, the only new symbol in this table is T2 j , which indicates that the total of each j group (for the B treatments, or blocks) is squared. See Problems 13.5 and 13.6 for application of these formulas.

15 13.5 TWO-FACTOR COMPLETELY RANDOMIZED DESIGN

(TWO-WAY ANOVA, n OBSERVATIONS PER CELL)

The linear model for the two-way analysis of variance model with one observation per cell (with no replication) is

$$X_{jk} = \mu + \beta_j + \alpha_k + \varepsilon_{jk} \tag{13.3}$$

where $\mu =$ the overall mean regardless of any treatment
$\beta_j =$ effect of the treatment j or block j in the B dimension of classification
$\alpha_k =$ effect of the treatment k in the A dimension of classification
$\varepsilon_{jk} =$ the random error associated with the process of sampling

Figure 3:

- As explained in Section 13.3, when replication is included within a two-way design, the interaction between the two factors can be tested.
- Thus, when such a design is used, three different null hypotheses can be tested by the analysis of variance: that there are no column effects (the column means are not significantly different), that there are no row effects (the row means are not significantly different), and that there is no interaction between the two factors (the two factors are independent).
- A significant interaction effect indicates that the effect of treatments for one factor varies according to levels of the other factor.
- In such a case, the existence of column and/or row effects may not be meaningful from the standpoint of the application of research results.

The linear model for the two-way analysis of variance when replication is included is

Table 13.2 Summary Table for Two-Way Analysis of Variance with One Observation per Cell (Randomized Block Design)

Table 13.3 is the summary table for the two-way analysis of variance with replication. The formulas included in this table are based on the assumption that there are an equal number of observations in all of the cells. See Problem 13.7 for application of these formulas.

Table 13.3 Summary Table for Two-Way Analysis of Variance with More than One Observation per Cell

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F ratio
Among treatment groups (A)	$K-1$	$SSA = \sum_{k=1}^{K} \frac{T_k^2}{nJ} - \frac{T^2}{N}$	$MSA = \frac{SSA}{K-1}$	$F = \frac{MSA}{MSE}$
Among treatment groups (B)	$J-1$	$SSB = \sum_{j=1}^{J} \frac{T_j^2}{nK} - \frac{T^2}{N}$	$MSB = \frac{SSB}{J-1}$	$F = \frac{MSB}{MSE}$
Interaction (between) factors (A and B) (I)	$(J-1)(K-1)$	$SSI = \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\sum_{i=1}^{n} X\right)^2 - SSA - SSB - \frac{T^2}{N}$	$MSI = \frac{SSI}{(J-1)(K-1)}$	$F = \frac{MSI}{MSE}$
Sampling error (E)	$JK(n-1)$	$SSE = SST - SSA - SSB - SSI$	$MSE = \frac{SSE}{JK(n-1)}$	
Total (T)	$N-1$	$SST = \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} X^2 - \frac{T^2}{N}$		

Figure 4:

16 13.6 ADDITIONAL CONSIDERATIONS

- All of the computational procedures presented in this chapter are for fixed-effects models of the analysis of variance.
- In a fixed-effects model, all of the treatments of concern for a given factor are included in the experiment.
- For instance, in Problem 13.1 it is assumed that the only instructional methods of concern are the three methods included in the design.
- A random-effects model, however, includes only a random sample from all the possible treatments for the factor in the experiment.
- For instance, out of ten different instructional methods, three might have been randomly chosen.
- A different computational method is required in the latter case because the null hypothesis is that there are no differences among the various instructional methods in general, and not just among the particular instructional methods that were included in the experiment. In most experiments the fixed-effects model is appropriate, and therefore the presentation in this chapter has been limited to such models.
- The concepts presented in this chapter can be extended to more than two factors, or dimensions. Designs involving three or more factors are called factorial designs, and in

fact most statisticians include the two-way analysis of variance with replication in this category.

- Although a number of different null hypotheses can be tested with the same body of data by the use of factorial designs, the extension of such designs can lead to an extremely large number of categories (cells) in the data table, with related sampling problems.
- Because of such difficulties, designs have been developed that do not require every possible combination of the treatment levels of every factor being included in the analysis. Such designs as the Latin Square design and incomplete block designs are examples of such developments and are described in specialized textbooks in the analysis of variance.
- Whatever experimental design is used, rejection of a null hypothesis in the analysis of variance typically does not present the analyst with the basis for final decisions, because such rejection does not serve to pinpoint the exact differences among the treatments, or factor levels. For example, given that there is a significant difference in student achievement among three instructional methods, we would next want to determine which

  of the pairs of methods are different from one another. Various procedures have been developed for such pairwise comparisons carried out in conjunction with the analysis of variance.

17 13.7 USING EXCEL AND MINITAB

Computer software is available for all the designs of the analysis of variance described in this chapter, as well as for more complex designs that are outside the scope of this book. Application of Excel and Minitab for the one-factor completely randomized design, for the randomized block design, and for the two-factor completely randomized design are illustrated in Solved Problems 13.8 through 13.13. Solved Problems ONE-FACTOR COMPLETELY RANDOMIZED DESIGN 13.1. Fifteen trainees in a technical program are randomly assigned to three different types of instructional approaches, all of which are concerned with developing a specified level of skill in computer-assisted design. The achievement test scores at the conclusion of the instructional unit are reported in Table 13.4, along with the mean performance score associated with each instructional approach. Use the analysis of variance procedure in Section 13.1 to test the null hypothesis that the three sample means were obtained from the same population, using the 5 percent level of significance for the test.