
9

ANALYSIS OF VARIANCE

We now come to a very popular test with health and social science researchers. Analysis of Variance (ANOVA) is popular because the researcher is able to compare several different groups, rather than the independent *T* test which only compares two groups. ANOVA is really quite an ingenious test in the way in which it allows statistical decisions by analyzing *components of the variance of the groups* to be compared.

There are several variations of this test. We will focus primarily on the “one-way” ANOVA in this chapter, but I will cover an extension of the test briefly at the end of the chapter. One-way ANOVA refers to the number of independent variables. *In research, independent variables are known as “factors.”* Therefore, if you have a research problem that has several levels of one independent variable, you can use one-way ANOVA to detect any differences among the sample groups.

Chapter 10 discusses “Factorial ANOVA,” which allows the researcher to analyze more than one independent variable. “Factorial” refers to the fact that there are more than one factor in the procedure. Thus, if you have two independent variables, the procedure would be known as factorial ANOVA, or a 2XANOVA.

A HYPOTHETICAL EXAMPLE OF ANOVA

In Chapter 8, I discussed an experiment I conducted on the effects of noise on human learning. Since there were only two groups to compare in that example (high vs low noise), I showed how the independent *T* test could be used to test the null hypothesis.

When I conducted the actual experiment however, I compared four groups on their rates of learning. Figure 9.1 shows the design of the experiment.

As you can see, there were four groups, three of which used different levels of noise and one control group with no noise present. (Recall that I used white noise in different magnitudes to distinguish the experimental groups.) The noise level in group III, 30 decibels, is comparable to quiet conversation; 60 decibels is equivalent to normal conversation; 90 decibels is equivalent to a large commercial jet landing just over a mile away. (Sustained exposure to 90 decibel noise could result in hearing loss.)

There is (still) only one independent variable (noise) in the experiment, but now there are groups in four levels of the independent variable. A *T* test would be inappropriate because I would need to conduct six *T* tests to compare all the group results. Why six *T* tests? Look at the comparisons in Figure 9.2.

Even if I had conducted all these *T* tests, I would have no idea of the “whole” test result. I needed the ANOVA since it conducts *all the comparisons within the same procedure at the same time*. I needed one “omnibus” answer to the question of whether noise affects human learning. If this result rejected the null hypothesis, I could then proceed to dig further into the results to examine *which* of the groups

Research Treatment Variable (Noise)			Dependent Variable (Word Recognition)
Random Selection and Assignment	No Pre-Test	Experimental Group I (90 decibels)	Outcome Test Scores (# of recognized words)
Random Selection and Assignment	No Pre-Test	Experimental Group II (60 decibels)	Outcome Test Scores (# of recognized words)
Random Selection and Assignment	No Pre-Test	Experimental Group III (30 decibels)	Outcome Test Scores (# of recognized words)
Random Selection and Assignment	No Pre-Test	Control Group (No noise level)	Outcome Test Scores (# of recognized words)

Figure 9.1 The four groups in the noise-learning experiment.

Group Comparisons
Exp. Group I Versus Exp. Group II
Exp. Group I Versus Exp. Group III
Exp. Group I Versus Control Group
Exp. Group II Versus Exp. Group III
Exp. Group II Versus Control Group
Exp. Group III Versus Control Group

Figure 9.2 The paired comparisons in the experiment with four groups.

THE NATURE OF ANOVA

257

were different from the others and therefore may have been responsible for the overall omnibus finding.

THE NATURE OF ANOVA

When a researcher conducts multiple (separate) tests on the same data in the research study, the alpha error “accumulates.” This is known as *familywise error* since each comparison is part of the same “family” of tests. Thus, although it is not this simple, if you have six tests and the alpha error is 0.05 for each, you might have a 0.30 alpha error (6×0.05) in the entire family of comparisons. The more tests you conduct, the greater the likelihood that you will make alpha errors. Thus, the overall omnibus test is compromised.

The ANOVA “conducts” all these comparisons together by analyzing components of variance and therefore limits familywise error. You have seen how the *T* test used variance measures (i.e., sum of squares (SS)) to help detect differences between two sample groups in Chapter 8. ANOVA identifies three sources of variance:

- The variation that exists *within* each sample group
- The variation *between* each sample group and the overall (grand) mean
- The *total* variance from all sources

Because the ANOVA process analyzes all these sources of variance simultaneously, you do not have to rely on multiple tests using the same data.

The null hypothesis in the hypothetical experiment would thus be

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

This null hypothesis states that all four group means come from the same popu-

This null hypothesis states that all four group means come from the same population; there is no difference between the group means. As you can see, there are a variety of ways in which the null hypothesis could be rejected. The following represent two of several possibilities: All the groups could be from different populations (the top example) or group 1 could be from a different population than group 2 and groups 3 and 4 could differ (the bottom example). There are several other possibilities:

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

$$H_A : \mu_1 \neq \mu_2 = \mu_3 \neq \mu_4$$

If the means of the groups are markedly dissimilar to one another (in a statistically significant sense), you could reject the overall null hypothesis and conclude that at least one of the group means differs from the others. If you are able to reject the null hypothesis, the omnibus ANOVA result does not indicate which mean (or means) is different from the others. In that event, you would perform a secondary test, called a “post hoc comparison,” in order to identify which of the means were significantly different from one another. I will discuss this process after I discuss the overall or omnibus ANOVA test.

THE COMPONENTS OF VARIANCE

If you recall the earlier chapters, you will remember that variance is a measure of the dispersion of scores in a distribution. Variance is a “global” measure of variation, so I will refer to it instead of the standard deviation (another measure of dispersion) as a way of determining how sample groups differ from one another. Look at Figure 9.3 to see how my experimental groups might be represented, using the noise experiment I introduced earlier.

Figure 9.3 shows the sample groups arrayed from left to right on an X-axis measuring the “number of learning errors,” which is the dependent variable for this experiment. Therefore, the “no noise group” shows lower values of learning errors than any of the other groups that represent different levels of learning errors. Remember that my experimental question was whether noise affected human learning. I operationally defined learning as the number of simple words recalled during a memorization task. The experiment sought to see if different amounts of noise would decrease the number of words recalled.

Figure 9.3 shows the three sources of variance in an ANOVA analysis. The four sample groups have their own distributions, so all their scores “spread out” around their means. This is known as “within variance” (V_{within}) since the variance is measured within each sample distribution. Figure 9.3 also shows a “shadow” distribution that represents a total distribution if all the individual scores from all the sample

groups were thrown into one large distribution. If you were to measure *the variance of that large, composite distribution, this would constitute the “total variance”* (V_{Total}). As you can see, the composite distribution has a mean value, indicated by

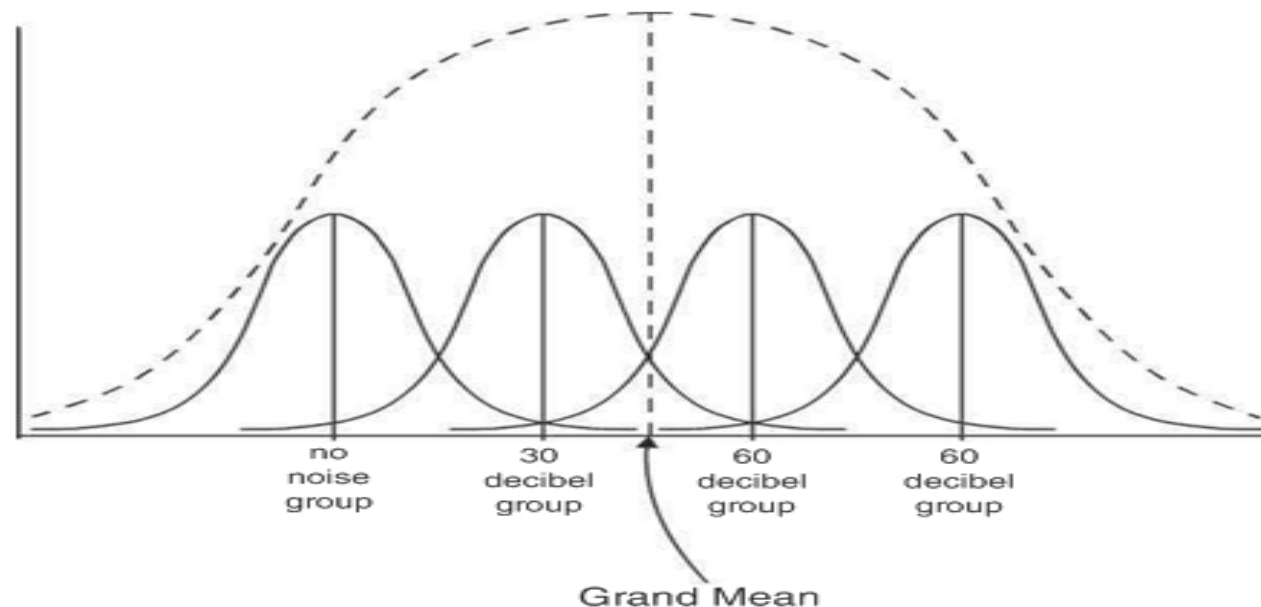


Figure 9.3 The components of variance in ANOVA.

THE PROCESS OF ANOVA

259

the dashed line. This total mean is known as the “grand mean.” *Between variance* (V_{Between}) *is the variation between each sample group mean and the grand mean.*

- V_{Total} is the total variance of all individual scores in all sample groups.
- V_{Between} is the variation between the sample means and the grand mean.
- V_{Within} is the variation of individual scores within their own sample groups.

THE PROCESS OF ANOVA

In effect, the ANOVA process determines *whether the sample means vary far enough from the grand mean that they could be said to be from different populations.* What makes this question difficult to determine is that within each group, the scores vary around their own group mean. If the sample group means are close enough to one

another, single scores could lie within the area where the groups overlap making group identification difficult for these individual scores.

Figure 9.4 shows two possible results that illustrate the potential results of ANOVA. The top panel shows the results of an experiment in which the groups are so “squished together” that the group variances are intermingled. That is, the group means are so close together that it would be difficult to conclude that they are significantly different from the grand mean.

In the bottom panel, the group means are far enough apart that individual group variances do not have extensive overlap. That is, the variances within each sample group do not confuse the distance between the groups. If you have an actual result like the one shown in the bottom panel, you would likely determine that the group means are statistically different from one another. This might not be the case in the top panel.

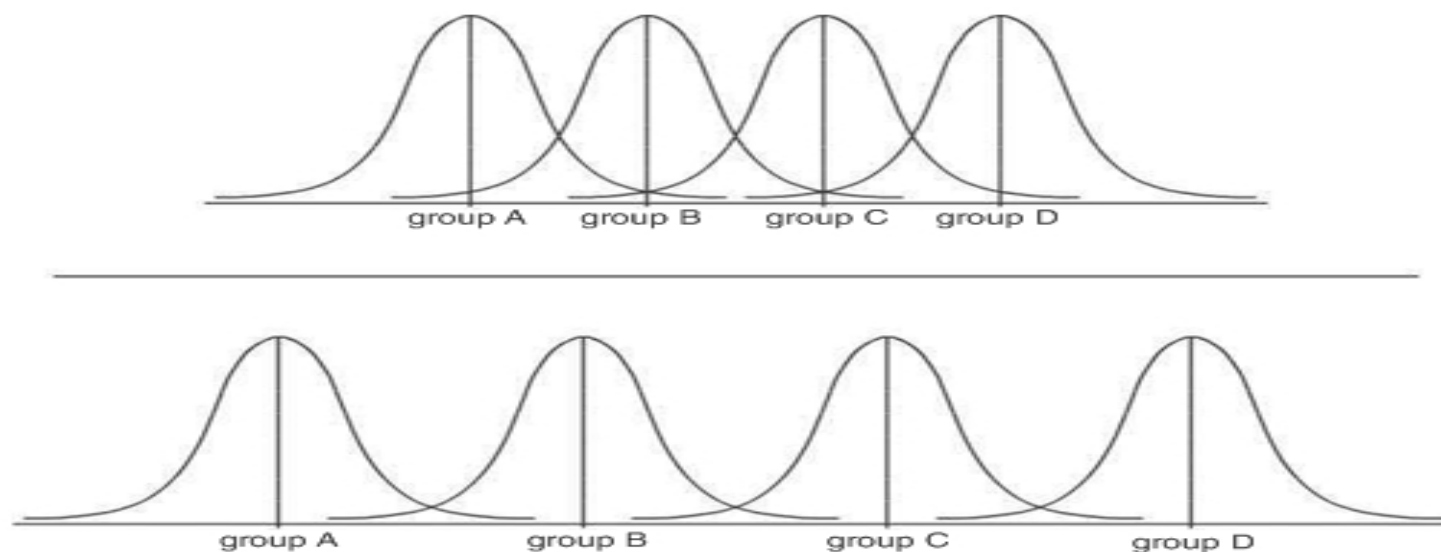


Figure 9.4 Different ANOVA possibilities.

ANOVA seeks to determine whether the V_{Between} would be large relative to the V_{Within} as it appears to be the case with the bottom panel example. In the top example, the V_{Within} “muddles the picture” of how the groups relate to one another. It is for this reason that within variance is often referred to in research (and in SPSS®) as “error variance.” If you want to understand how far apart the group means are from the grand mean, the error variance (V_{Within}) “gets in the way.” When there is less error variance, the distance between the group means and the grand mean (i.e., the

V_{Between}) becomes easier to distinguish.

CALCULATING ANOVA

The ANOVA calculation compares V_{Between} to V_{Within} . When V_{Between} is large relative to V_{Within} , there is a greater likelihood that the researcher can reject the null hypothesis. We can suggest a “conceptual equation” that expresses this relationship:

$$F = \frac{V_{\text{Between}}}{V_{\text{Within}}}$$

ANOVA uses the F distribution as a comparison distribution for the relationship of the between-to-within variance. I introduced the F test in Chapter 8 when I discussed how to test whether two sample variances were statistically equal. This same distribution (see Figure 8.16) can be used to determine where a calculated ANOVA value falls on the sampling distribution of all possible samples. After all, you are still comparing variances. But in the ANOVA test, you compare between-to-within variance, whereas in the equal variance test, you compare two sample variances.

The conceptual equation before thus expresses the relationship of between-to-within variance as a point on the F distribution of all possible samples to determine whether the sample group means are far enough from the grand mean that the calculated F ratio would fall in the exclusion area of the comparison distribution. If a researcher obtained an actual result like the bottom example in Figure 9.4, there would be a greater likelihood that the F ratio (calculated F) would fall in the exclusion area than if the top example was obtained.

Calculating the Variance: Using the Sum of Squares (SS)

The ANOVA procedure helps the researcher identify and compare all the components of variance present in the research study. Therefore, each of the variance components must be calculated in order to make a statistical decision. The Sum of Squares (SS) is used to measure the components. As I explained in past chapters, SS is a large number when the scores of a distribution are spread out and a small number when the scores are close together.

In Chapter 8, I specified the formula to calculate SS:

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

You can use variations of this formula to calculate each of the three sources of variance.

Calculating Components of Variance Using SS

Moving from conceptual to computational definitions, you can identify these sources as follows:

SS_T = sum of squares of the total distribution of all individual scores (V_{Total})

SS_B = sum of squares between the group means and the grand mean ($V_{Between}$)

SS_W = sum of squares within each sample group (V_{Within})

Calculating the three SS values is somewhat easier when you consider that the total variation (V_{Total}) is comprised of the between ($V_{Between}$) and within (V_{Within}) variances. Therefore,

$$SS_T = SS_B + SS_W$$

and

$$SS_W = SS_T - SS_B$$

I will introduce a simple (hypothetical) set of findings for my noise-learning experiment to demonstrate how to calculate these components. Then I will present an example with real data. Table 9.1 shows the hypothetical data for the experiment. The dependent variable is the number of learning errors.

Table 9.2 shows the data with appropriate squares to be used to calculate the sums of squares. The general SS formula follows the data; then the calculations are shown for each variance component:

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

Calculating SS_T

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$\begin{aligned} SS_T &= (86 + 174 + 509 + 889) - \frac{(18 + 26 + 45 + 59)^2}{16} \\ &= 1658 - 1369 \\ &= 289 \end{aligned}$$

TABLE 9.1 Hypothetical Experiment Data

Control	30 Decibels	60 Decibels	90 Decibels
5	7	12	15
4	5	11	14
3	6	10	18
6	8	12	12

TABLE 9.2 Hypothetical Experiment Data with Squared Values

	Control		30 Decibels		60 Decibels		90 Decibels	
	X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
	5	25	7	49	12	144	15	225
	4	16	5	25	11	121	14	196
	3	9	6	36	10	100	18	324
	6	36	8	64	12	144	12	144
Σ	18	86	26	174	45	509	59	889

Calculating SS_B Calculating SS_B uses a variation of the SS formula, but the focus is on the *deviation of each group mean from the grand mean*. Therefore, you will calculate “group means” in variance terms and subtract the variance measures from the grand mean. Here is the formula:

$$SS_B = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} + \frac{(\sum X_4)^2}{n_4} (-) \frac{(\sum X)^2}{N}$$

This equation looks intimidating, but if you look at it, the main “pieces” make sense. As you can see, a *global measure* of each group’s mean ($\frac{(\sum X)^2}{N}$) results from dividing the squared group sums by the group’s sample size. Then, the grand mean measure ($\frac{(\sum X)^2}{N}$) is subtracted from the total of the separate group mean measures. I shaded this portion of the equation, so you can see the components clearly. There is an important thing to note in this equation:

$$n \neq N$$

If you look carefully at the equation, you will note that the separate group mean measures have the “group sizes” (n) as the divisor. The grand mean measure at the end of the equation has the total individual score size (N) in the divisor. This is because the grand mean is calculated for all individual scores ($N = 16$), whereas the individual group mean measures are divided only by the scores that make up that group (all four sample groups have $n = 4$):

$$SS_B = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} + \frac{(\sum X_4)^2}{n_4} (-) \frac{(\sum X)^2}{N}$$

$$= \frac{(18)^2}{4} + \frac{(26)^2}{4} + \frac{(45)^2}{4} + \frac{(59)^2}{4} (-) \frac{(889)^2}{16}$$

$$\begin{aligned}
 &= \frac{324}{4} + \frac{676}{4} + \frac{2025}{4} + \frac{3481}{4} (-) \frac{(18 + 26 + 45 + 59)^2}{16} \\
 &= 81 + 169 + 506.25 + 870.25 (-) 1369 \\
 &= 257.5
 \end{aligned}$$

CALCULATING ANOVA

263

As you can see, I left the shaded portion in the equations to show that this portion is the same in the SS_T and SS_B formulas. It “represents” the grand mean, so its value will not change. In the SS_T formula, all the individual scores (X) deviate away from the (grand) mean. Thus, the shaded portion of the formula shows that this is the value from which all the scores deviate. In the SS_B formula, the shaded portion represents the grand mean as well. Here, each of the sample group means are added together and then subtracted from the grand mean. It is as if each of the group means is treated as a single score, and you are calculating the variance of the group means around the grand mean (which, of course, we are actually doing!).

Calculating SS_W From the previous equation, you can see that if you calculate SS_T and SS_B , you can derive SS_W without additional calculations:

$$\begin{aligned}
 SS_W &= SS_T - SS_B \\
 &= 289 - 257.5 \\
 &= 31.5
 \end{aligned}$$

Creating a Data Table

Now that you have calculated the key SS values, you can use a table to display the results. As you will see in the steps ahead, this is a very helpful visual step and one that both SPSS and Excel use to display the ANOVA results. You should therefore learn to use the table to keep all the calculated values in order and see how ANOVA results are reported by most all statistical software.

Table 9.3 shows the template you might use with the results thus far. As you can see the table lists the sums of squares calculated values for each “source” of variance. Remember that the SS is a measure of variance based on the raw scores in the data file.

Using Mean Squares (MS)

As you recall, inferential statistics addresses the issue of how sample values represent (or do not represent) population values. In the Z and T tests, you learned to transform

sample values to points on a sampling distribution so that you could show how the calculated sample ratio compares to all possible sample values.

Treat ANOVA results in the same way. *ANOVA is also an inferential process, so you will use sample values to estimate population values.* You calculated the sample

TABLE 9.3 The ANOVA Results Table

Source of Variance	SS	df	MS	F ratio
Between	257.5			
Within	31.5			
Total	289			

(between, within, and total) variances using SS values, so you need now to use these values to estimate population (between, within, and total) variances.

This process is akin to using sample values (SD) to create an estimate (s_x) of the population standard deviation (σ_x) in the T test. In order to estimate the population standard deviation, you had to use the sample standard deviation and “adjust” it by dividing it by its degrees of freedom (in the single sample T test, the $df = N - 1$) so that it would be a better estimate of the population parameter.

With ANOVA, you use degrees of freedom to make each of the SS values better estimates of population values. These estimates are known as “mean squares” (MS) because they are created by dividing the SS by degrees of freedom, much like you create a mean by adding together the values in a sample and dividing by the number in the sample.

Degrees of Freedom in ANOVA

You can use the degrees of freedom to help estimate population values since the size of the samples greatly affects the variance estimates. By using degrees of freedom in order to obtain better estimates, you create mean squares that average out the variance regardless of sample size. Here are the degrees of freedom for each SS value:

- df_t (degrees of freedom for total SS) = $N - 1$ (N = the total number of individual scores in all groups combined)
- df_b (degrees of freedom for between SS) = $k - 1$ (where k = number of sample groups)
- df_w (degrees of freedom for within SS) = $N - k$

when you make specific population estimates from sample values, you use *those* degrees of freedom. As you see, estimating the total variance uses 1 df since you use the entire set of individual scores. Estimating the between variance uses another 1 df since you use the set of sample means. Estimating the within variance uses several more (depending on how many sample groups there are) degrees of freedom, since the sample groups are separate “sets” of scores that deviate around their respective group means.

Calculating Mean Squares (MS)

Calculating MS values is simple when you look at the ANOVA table of values in Table 9.3. Table 9.4 shows the MS values, which are derived from dividing the SS values by their respective degrees of freedom:

$$MS_B = \frac{SS_b}{df_b} = \frac{257.5}{3} = 85.83$$

$$MS_w = \frac{SS_w}{df_w} = \frac{31.5}{12} = 2.625$$

CALCULATING ANOVA

265

TABLE 9.4 The ANOVA Results Table with Calculated MS Values

Source of Variance	SS	df	MS	F ratio
Between	257.5	3 (<i>k</i> - 1)	85.83	
Within	31.5	12 (<i>N</i> - <i>k</i>)	2.625	
Total	289	15 (<i>N</i> - 1)	----	

You can note other features from Table 9.4:

- The between and within SS values sum to the total SS value.
- The between and within df values sum to the total df value.
- The MS_T value is not included because it is not needed to calculate the ANOVA result.

The *F* Ratio

In an earlier section of this chapter, I noted that the ANOVA process determines *whether the sample means vary far enough from the grand mean that they could be said to be from different populations*. What you need in order to determine

this is a comparison of the variation of the means from the grand mean (MS_B), relative to the variance within each of the sample groups that make the earlier estimate difficult to determine (MS_W).

Here is the formula for calculating the F ratio. Using the values from Table 9.4, the calculated F ratio for this example problem is

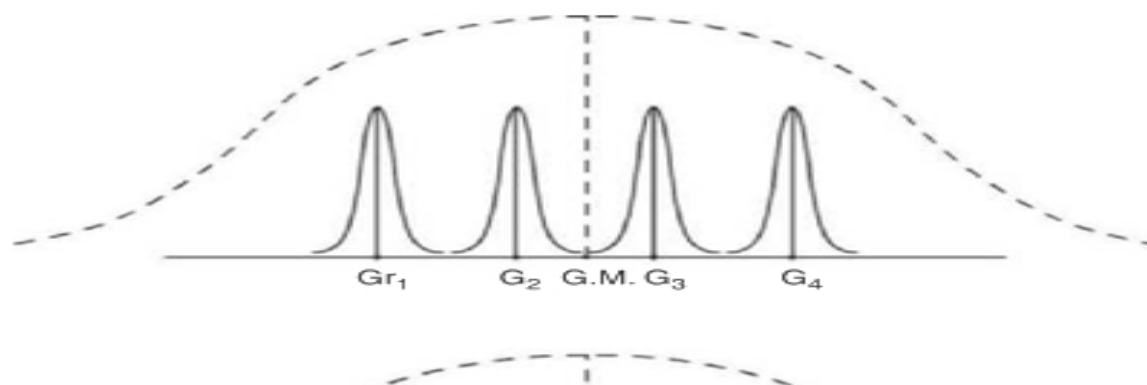
$$F = \frac{MS_B}{MS_W} = \frac{85.83}{2.625} = 32.70$$

As you can see, the F ratio compares the variation of the sample means around the grand mean (MS_B) to the variation within each sample group (MS_W). I mentioned before that the within variance measure is known as “error” just for this reason. If the sample group values did not vary at all, it would be easy to see how far away the sample means vary from the grand mean! Consider the two panels in Figure 9.5.

The top panel of Figure 9.5 shows a hypothetical example of four sample groups with almost no within variance. In this case, it is easy to see how distinct the group means are and how clearly they differ from the grand mean. The bottom panel shows just the opposite. There is so much within variance that it is almost impossible to see how distinct the group means are and how they vary around the grand mean. The within variance is essentially “noise” and is therefore called error.

It is for these reasons that I said that the F ratio is a measure of the between variation *relative to the within variance*. That is why the F ratio is simply comparing the between to the within variation. When you divide the two, you would anticipate a much higher F ratio in the top panel than in the bottom panel since the divisor in the F ratio would be much smaller in the top panel.

In the noise example, the between variance measure was 85.83 and the within variance measure 2.625. These represent quite differing variance measures! It is fairly



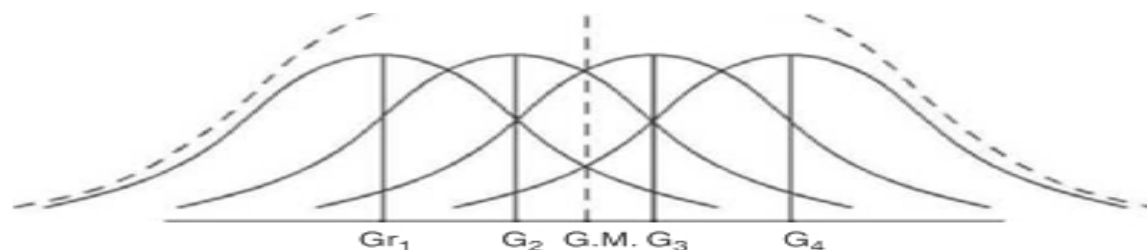


Figure 9.5 ANOVA possibilities of groups with different within variances.

TABLE 9.5 The Final ANOVA Results Table

Source of Variance	SS	df	MS	F ratio
Between	257.5	3	85.83	32.70
Within	31.5	12	2.625	
Total	289	15	----	

easy to assume that the resulting F ratio (32.70) represents a situation where the sample group means are quite spread out around the grand mean. The question is, how far from the grand mean do the sample means have to vary before they could be said to represent different populations? This is the essence of the ANOVA hypothesis test.

Table 9.5 shows the final ANOVA table with the F ratio included. The structure of the table allows you to “see” how the calculation for the F ratio was performed. I divided the between MS (85.83) by the within MS (2.625) to yield 32.70.

The F Distribution

Figure 8.16 shows that the F distribution does not appear to be similar to the Z or T distributions. Remember I mentioned in Chapter 8 that the F distribution is a sampling distribution of *variances*, not means. In the F test for equal variances, most group variances hover around 1 when the variances are equal (since if we divide similar variance measures, the ratio will be close to 1.00).

TABLE 9.6 The F Table of Values Exclusion Areas

<i>Critical (Exclusion) Values for the Distribution of F</i>						
		<i>df in the Numerator (between)</i>				
		<i>Exclusion Level</i>				
		1	2	3	4	5

df in the Denominator	1	0.05	161	200	216	225
		0.01	4052	4999	5403	5625
	2	0.05	18.51	19.00	19.16	19.25
		0.01	98.49	99.00	99.17	99.25
	10	0.05	4.96	4.10	3.71	3.48
		0.01	10.04	7.56	6.55	5.99
	11	0.05	4.84	3.98	3.59	3.36
		0.01	9.65	7.20	6.22	5.67
	12	0.05	4.75	3.88	3.49	3.26
		0.01	9.33	6.93	5.95	5.41
	13	0.05	4.67	3.80	3.41	3.18
		0.01	9.07	6.70	5.74	5.20

In the ANOVA test, you do not compare two group variances but between-to-within variance. When you compare these measures, the resulting F ratio is often much larger than 1. Thus, *the F distribution is directional*. Together with the fact that variance measures cannot be negative (they can be very small but not negative), the researcher needs to compare a calculated F ratio with a tabled value of F that represents all possible samples of variance.

The F distribution is therefore the sampling distribution of comparison. F tables of values identify the exclusion values at 0.05 and 0.01 probability, for example. The table is complex because it takes into account the degrees of freedom of both the between and the within measures since sample size is so critical to the analyses.

Table 9.6 shows a portion of the F table from the F table of values in the appendices. In the noise example, the tabled value of F at 0.05 with df's for between (3 – in the column) and within (12 – in the rows) variance measures identifies an exclusion value of 3.49 (shaded value) at the 0.05 level.¹ Thus, 3.49 is the value on the F distribution that defines the exclusion area. If the calculated F value exceeds this tabled value, you can reject the null hypothesis. In this example, the calculated $F = 32.70$ far exceeds the tabled value of F . Therefore, you can conclude that the finding ($F = 32.70$) falls in the 0.05 exclusion area of 3.49 and is therefore not likely a chance finding, $p < 0.05$ (i.e., it is a statistically significant finding).

¹Note that the table also includes the 0.01 level of exclusion along with the 0.05 level for each critical value of F chosen.

If the example values represented the data in my experimental study, I could reject the null hypothesis. This would mean that, taken together, the means of the group noise levels vary so greatly around the grand mean that the groups represent different populations of students.²

EFFECT SIZE

Like the other statistical analyses, you can create an effect size that shows the “impact” of the independent (predictor) variable on the dependent (outcome) variable. In the case of my noise example, the effect size concern could be asked, “how much does the *grouping on the independent variable* affect the outcome measure?” By this question, I would be trying to determine how much noise impacts learning by measuring how much the specific noise groupings (0, 30, 60, and 90 decibels) push the group outcome measures (learning errors) apart.

So far, I have discussed effect size in terms of Cohen’s *d* measures where I analyzed the difference between two groups (independent *T* test) or between a group and a population (single sample *T* test). The ANOVA determines the significance of group differences by examining variance, so you can use a variance measure to express effect size.

The Greek symbol for eta square “ η^2 ” is the symbol that refers to effect size in ANOVA. *It refers to the proportion of variance in the outcome measure explained by the grouping on the independent variable*. I discussed this in Chapter 8 with reference to an effect size formula for independent *T* Test (see “The Eta Squared Method”). This is an easy calculation from the ANOVA table using the following formula:

$$\begin{aligned}\eta^2 &= \frac{SS_{\text{Between}}}{SS_{\text{Total}}} \\ &= \frac{257.5}{289} \\ &= 0.89\end{aligned}$$

Thus, 89% of the variance in learning errors is due to assigning subjects to four different noise levels. The effect size is very large (only 100% is possible, obviously), but this is probably due to the fact that I had very small sample group sizes (four per group), and this is a hypothetical example.

How large is large? That is, how large does η^2 have to be before you would say it is “meaningful”? That is a question for which there are many answers! The reason

²By the way, the actual outcome of my experiment was interesting! I found no significant differences between the groups. My speculation as to why was that I was using white noise rather than “real” noise (like different kinds of music, street noise, people yelling, etc.). I didn’t take into account what the subjects were used to (i.e., they were all different coming in to the experiment; some were used to studying with music and others weren’t, etc.), and they were all trying really hard to do well on the “test.” The research literature on this issue is very interesting. My undergraduate experiment as yet has not won any prizes for original research!

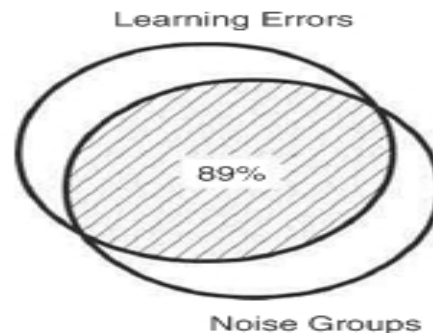


Figure 9.6 Venn diagram showing effect size.

it has many answers is that it depends on the sample sizes, number of groups, etc., I alluded to power analysis in earlier chapters. Statisticians have constructed tables that take into account factors that would have an impact on the size of the effect size value. For example, there are tables for the 0.05 level of significance and the group size as considerations in judging effect size (see Cohen, 1988).

There are therefore many standards offered by statisticians and researchers for judging effect size meaningfulness. I tend to take two approaches to this question. First, I suggest the same guidelines I offered in Chapter 8 as benchmark comparisons: 0.01 (small), 0.06 (medium), and 0.14 (large). These are ballpark figures because the next approach is the most crucial.

The second approach to the question is allowing the researcher to judge the meaningfulness of the effect size by the nature of the problem studied. It would therefore have a lot to do with the subject. If I am studying a hypothetical research question with very small sample sizes (as I did in the noise-learning example), I would not be very excited about a very large effect size (and in our study, 0.89 far exceeded the 0.15 guideline for “large effect size”). However, if I am studying a new drug that can lessen the death rate from AIDS, I would be ecstatic to find an effect size of 0.03 even though it might be judged “small” by the earlier benchmarks.

One way to visually understand effect size as “explained variance” is to use Venn diagrams. Look at Figure 9.6. The top circle represents the outcome measure of learning errors. The bottom circle represents the noise groups. Effect size is represented in the portion of the top circle with diagonal lines. The 89% refers to the η^2 value of 0.89. If the top circle represents all the variation of learning errors, I can reduce that variance by 89% just by knowing the different noise conditions. That would only leave 11% of the variance in learning errors unexplained.

POST HOC ANALYSES

If an ANOVA analysis results in the rejection of the null hypothesis, the next consideration is which specific group mean variation(s) from the grand mean might be responsible for the overall size of the F ratio? Stated differently, are all the sample

270

ANALYSIS OF VARIANCE

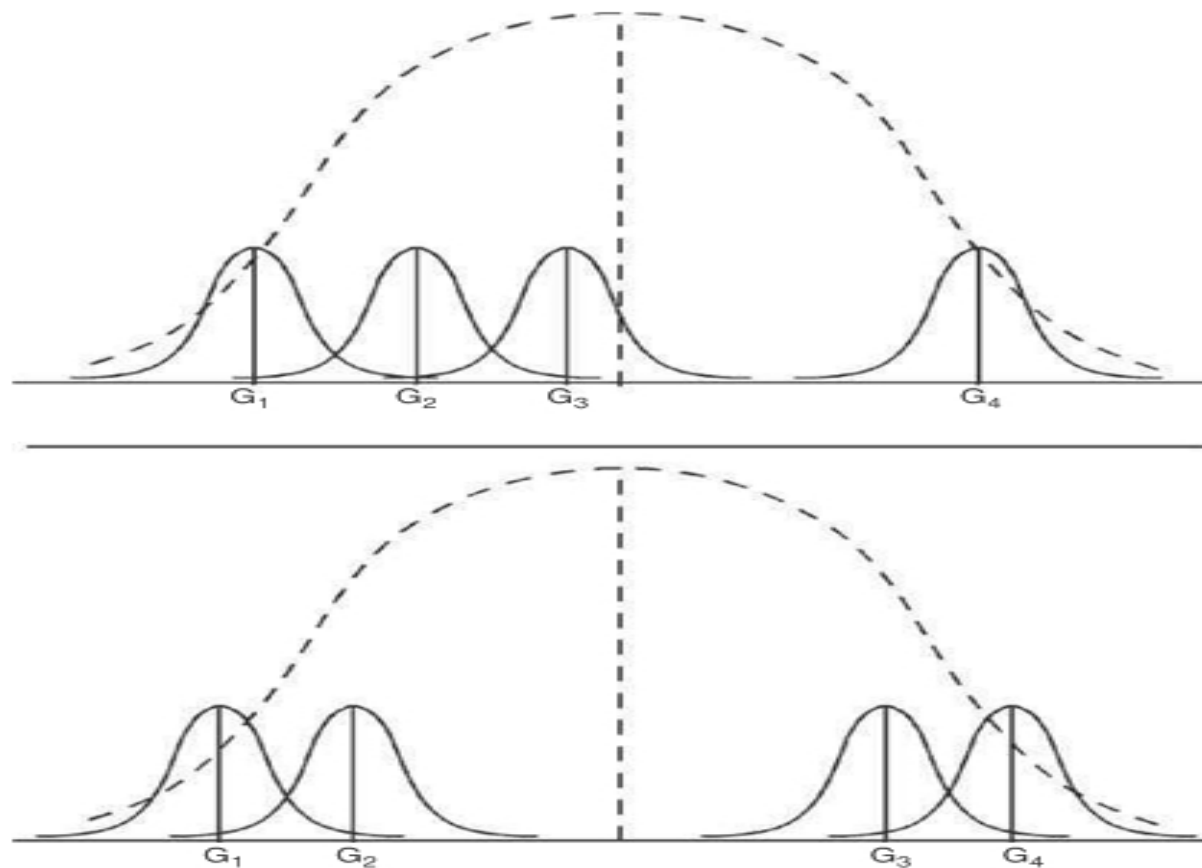


Figure 9.7 Post hoc test possibilities.

means similar in producing learning errors (obviously not since we rejected the null

hypothesis), or are some group means much more responsible for affecting the F ratio? What configuration of group mean differences affected the overall F ratio?

Figure 9.7 shows two possibilities of group differences that might produce a significant F ratio. The top panel shows that the first three noise level groups are similar in producing learning errors, but the subjects in the fourth noise level are much more likely to produce learning errors than the other three. The bottom panel of Figure 9.7 shows that the first two groups are similar, but both are very different from the third and fourth groups (which are similar to one another). There can be other combinations of findings as well.

“Varieties” of Post Hoc Analyses

Post hoc analyses look at group comparisons within the overall ANOVA (significant) result to detect where sample group comparisons may be so large that they cannot be

POST HOC ANALYSES

271

said to belong to the same population. The process is similar to individual T tests of pairs of samples within the overall ANOVA study. If you recall, I mentioned earlier in this chapter that my ANOVA study would consist of six separate T tests if I wanted to conduct paired comparisons. Since I could not do this without creating family-wise error, I performed the ANOVA test which compared all the sources of variance at once.

After the overall (omnibus) test of ANOVA is found to be significant, you must “return” to the initial question about which paired differences may be responsible for the ANOVA result. In order to avoid familywise error with the paired comparisons, statisticians have created processes for identifying paired sample differences that limit the overall error measures. Many are specially designed to fit specific research designs. Here are some examples of procedures used for “unplanned” comparisons (paired comparisons not identified before the study):

- Scheffe’s test – A conservative post hoc analysis that works well with uneven group sizes and uses stricter decision criteria.
- Tukey’s HSD – This Honestly Significant Difference procedure is used when the researcher wants to compare all possible pairs of samples.
- Dunnett – This test is used only with single comparisons, usually comparing a control and experimental group(s).

There are several other post hoc analyses available for use depending on the nature of the research issues. I will demonstrate the most common post hoc analysis for unplanned comparisons, Tukey’s HSD. Statistical programs such as SPSS provide results for this and other post hoc analyses, so you can expand your range of under-

standing and expertise of post hoc analyses as you pursue further research.

The Post Hoc Analysis Process

The general process researchers follow after concluding that the omnibus (overall) ANOVA test is significant is as follows:

1. Calculate the Tukey HSD critical value.
2. Create a comparison table that includes all the group means.
3. Compare each mean difference to the critical HSD value to determine which of the pairs are significantly different from one another.
4. Conclude the post hoc analysis with a general summary of results.

Tukey's HSD (Range) Test Calculation

This test is also called Tukey's Range Test since it requires the use of the "Studentized Range Table" (shown in the appendix) to help identify a critical value of comparison. In essence, the HSD *is a critical value of exclusion* since it is based on a set of probabilities that define extreme values on a distribution that takes into account the degrees of freedom in the overall study and the number of groups in the analysis.

Thus, *this approach uses a formula to determine the value beyond which the paired group differences in the study would be considered extreme.*

The Tukey HSD formula is as follows:

$$\text{HSD} = q_{(\text{Range})} \sqrt{\frac{\text{MS}_{w}}{n}}$$

The "pieces" of the formula are the following:

- HSD is the calculated point on the comparison distribution that identifies extreme values; any test value that is calculated to be larger than this value is considered extreme and would result in the rejection of the hypothesis that a specific pair of group means was equal.
- $q_{(\text{Range})}$ is the value obtained in the Studentized Range Table that helps to establish the exclusion value. I will use the example of the 0.05 level of exclusion, but you can establish the value for 0.01 or other levels of exclusion. If I do not specify otherwise in the examples that follow, you can assume I am using the 0.05 level.

- MS_w is the Mean Square within value found in the ANOVA table. In the example earlier it is equal to 2.625. Recall that I mentioned that the ANOVA within measures were known as “error.” Thus, in the formula, the error measure is “adjusting” the critical value of exclusion. If the error is large, it will have a marked impact on the exclusion area.
- n is the group size used in the comparison(s). In my example, all the groups were $n = 4$. Thus, you can use this formula to establish a single HSD critical value for each of the six paired comparisons.
- If the sample group sizes are unequal, you would need to create more than one HSD value for the paired comparisons. The SPSS program adjusts for unequal group sizes.

Using the range table is straightforward if you remember that it is adjusted based on the ANOVA error measure (within group MS). Table 9.7 shows a small part of a Range Table (see the appendix for the complete table) to give you an example of how to use it.

As you recall, I had four noise groups in the experiment, and the MS_w degrees of freedom was 12 (see Table 9.5). Based on these values, I shaded the relevant columns

TABLE 9.7 Example of Values from a Studentized Range Table

MS_w df	Number of Sample Groups						
	2	3	4	5	6	7	8
11	3.115	3.822	4.258	4.575	4.824	5.03	5.203
12	3.083	3.775	4.200	4.509	4.752	4.951	5.12
13	3.057	3.736	4.152	4.454	4.691	4.885	5.05

POST HOC ANALYSES

273

and rows in the Range Table in Table 9.6. The number of sample groups is the “4” column, and the MS_w df is the “12” row. Where these intersect is the $q_{(Range)}$ value used in the HSD calculation:

$$\begin{aligned}
 \text{HSD} &= q_{(Range)} \sqrt{\frac{MS_w}{n}} \\
 &= \text{HSD} = 4.20 \sqrt{\frac{2.625}{4}} \\
 &= 3.40
 \end{aligned}$$

Mean Comparison Table

Once the HSD critical value is calculated, you can proceed to examine how the group means differ. If you refer to Figure 9.2, you will see the six group comparisons that are necessary when there are four sample groups. In order to see what the group differences are between the paired groups, it is a simple matter of subtracting the group means. Table 9.8 lists the means of each of the four sample groups.

You could calculate the six paired sample differences by subtracting each pair of means as follows:

$$M_1 - M_2 = -2.00$$

$$M_1 - M_3 = -6.75$$

$$M_1 - M_4 = -10.25$$

$$M_2 - M_3 = -4.75$$

$$M_2 - M_4 = -8.25$$

$$M_3 - M_4 = -3.50$$

Doing the comparisons this way would yield the appropriate paired mean differences. Notice in the list, however, that if you subtract the higher group means (e.g., M_4) from the lower (e.g., M_1), the values are negative. In these paired comparison tests, you are interested only in the magnitude of the difference, not the sign. You can interpret the overall results by referring to the size of the means, but the paired tests do not use the sign. Therefore, you can subtract the smaller group means from the larger to get positive values.

TABLE 9.8 The Group Means

<i>Groups</i>	<i>Means</i>
M_1	4.50
M_2	6.50
M_3	11.25
M_4	14.75

TABLE 9.9 Matrix of Group Means

	M_1 (4.5)	M_2 (6.5)	M_3 (11.25)	M_4 (14.75)
M_1 (4.5)	---	2.00	6.75	10.25

M_2 (6.5)		---	4.75	8.25
M_3 (11.25)	-6.75		---	3.50
M_4 (14.75)				---

Another method for doing this is to create a “matrix” of group means. A matrix is simply a table constructed of rows and columns. You will see arrays of data in matrix form later in the book, so it is a good idea to use them to display data. SPSS uses matrix arrays for some outputs including the post hoc analysis.

Table 9.9 shows the group means in a matrix display. Previously, I calculated the difference of M_1 and M_3 to be -6.75 . I placed this value in the matrix under column M_1 and row M_3 . Thus, when I subtract 11.25 from 4.5 , I get -6.75 .

However, I can obtain positive values by subtracting the lower mean values from the higher. I have shown these values in the table. For example, subtracting 4.5 (M_1) from 11.25 (M_3) results in a value of 6.75 , which is located in the M_3 column and the M_1 row. Notice that the matrix is symmetrical. The same values will appear in the lower “triangle” that appear in the upper “triangle” (I shaded the upper triangle) in Table 9.9. I left the -6.75 in both the lower and upper triangles to show the symmetry of the table.

Compare Mean Difference Values from HSD

The next step in the post hoc analysis is to compare the group mean differences (from Table 9.9) to the HSD value of 3.40 . If any of the group mean differences exceed the HSD value, you would consider those two means *significantly* different. As you can see from the table, all the mean pairs exceed HSD except the $M_1 - M_2$ difference (2.00).

Post Hoc Summary

The ANOVA post hoc analysis showed that all paired comparisons showed significant differences except the group 1 (control group)–group 2 (30 decibels) comparison. Figure 9.8 shows how the groups might be arrayed to yield this finding.

ASSUMPTIONS OF ANOVA

As with the other statistical tests I discussed, the researcher needs to assess whether the conditions of the data are appropriate for the ANOVA test. Here are the primary assumptions, although ANOVA is also somewhat robust with respect to slight variations of assumptions:

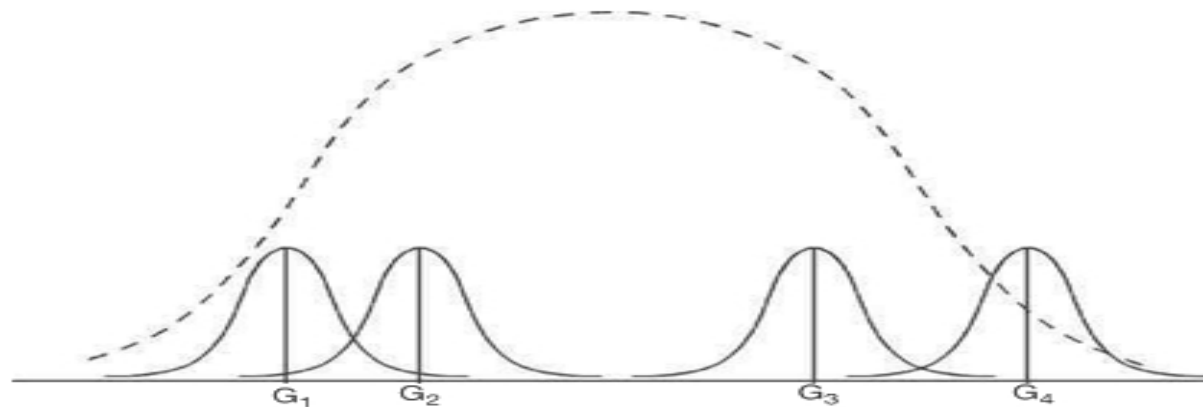


Figure 9.8 The post hoc summary for the example.

1. Population is normally distributed.
2. Population variances are equal. This assumption refers to the variance of each of the sample groups. You should make sure that the groups (in our case the noise conditions) have equal variance on the outcome measure (in our case, learning errors). (This is the homogeneity of variance assumption.)
3. Samples are independently chosen.
4. Interval data on the dependent variable.

ADDITIONAL CONSIDERATIONS WITH ANOVA

I will have more to say about these procedures, but here are some things to keep in mind if you use one-way ANOVA:

1. You can still use ANOVA (“the F test”) even if the study only uses two group means. In this case, remember that there is a relationship between F and T such that $F = T^2$.
2. The effect size is critical, as you learned with the other tests I covered. You learned to calculate η^2 which indicates the strength of the effects of F . Effect size is always calculated, even if the omnibus F test does not result in the reject of the null hypothesis. Simply because a study finding is (or is not) statistically significant does not indicate how much of a practical impact it has. *Statistical significance and effect size are separate considerations.*
3. ANOVA like the T test can be used in experimental contexts (where you consciously change the value of an independent variable to see what effects this has on a dependent variable) and for post facto situations (where you simply

try to determine if there are existing differences among several sample groups

276

ANALYSIS OF VARIANCE

without changing the value of the independent variable). In my example of noise and learning, I used ANOVA experimentally, since I directly “manipulated” or changed the value of the independent variable (noise level), to see if this affected the dependent variable (errors in learning performance).

4. Be sure to meet the assumptions of the test before you use it:
 - Populations are normally distributed.
 - Population variances are equal.
 - Independent selection.
 - Interval data on the dependent variable.

THE HYPOTHESIS TEST: INTERPRETING ANOVA RESULTS

This section presents another example of ANOVA showing the complete process of analysis and hypothesis testing. Although it is a hypothetical example, it reflects recent questions about the impact of sleep deprivation on health. Table 9.10 presents the data for health measures (diagnosis report with higher numbers indicating better health) according to three hourly amounts of sleep. There are obviously categories of sleep other than 6, 7, or 8 hours, but this study treats sleep as a categorical independent variable since the 6, 7, and 8 represent self-reported sleep amounts. In an actual study, you might measure sleep as a continuous variable. I am using the categories here as an example of ANOVA. In the extended example that follows, I will use Excel and SPSS to analyze the question, “does different amounts of sleep affect health?”

ARE THE ASSUMPTIONS MET?

Before proceeding to the manual calculations for this example problem, you need to examine the assumptions for ANOVA. Are the data appropriate for the one-way ANOVA?

TABLE 9.10 The Data for One-Way ANOVA Example

	SLEEP		
	6 Hrs.	7 Hrs.	8 Hrs.
HEALTH	4	13	12

	10	11	13
	8	12	11
	6	6	10
	2	8	17
	4	4	9

ARE THE ASSUMPTIONS MET?

277

Is Population Normally Distributed?

As you recall from Chapter 3, I mentioned that skewness and kurtosis are good indicators to help assess normal distributions. For ANOVA, you need to make sure that the outcome data are normally distributed for each of level of the independent variable. In this case, there are three independent variable levels: 6, 7, and 8 hours of sleep. Are the health data normally distributed? Use the results in Figure 9.9 to check this assumption.

The Excel output in Figure 9.9 indicates that all three groups are fairly normally distributed as determined by the similarity of mean and median values. The skewness and kurtosis figures appear within boundaries of normal. Use SPSS to investigate the skewness and kurtosis further since the standard errors are reported for skewness and kurtosis. Figure 9.10 shows the descriptive values that include skewness and kurtosis standard errors (using “Analyze – Compare Means – Means”), and Figure 9.11 shows the histograms for the groups.

As you can see, skewness and kurtosis for all groups of sleep are within bounds when the Sk and Ku values are divided by their standard errors. The histograms in

	6 Hrs	7 Hrs.	8 Hrs.
Mean	5.67	9.00	12.00
Standard Error	1.20	1.46	1.15
Median	5.00	9.50	11.50
Mode	4	#N/A	#N/A
Standard Deviation	2.94	3.58	2.83
Sample Variance	8.67	12.80	8.00
Kurtosis	-0.86	-1.72	1.67
Skewness	0.42	-0.35	1.19
Range	8	9	8
Minimum	2	4	9
Maximum	10	13	17

Sum	34	54	72
Count	6	6	6

Figure 9.9 The descriptive output (Excel) to test the normal distribution assumption.

Report								
Health								
SleepHours	Mean	N	Std. Deviation	Kurtosis	Std. Error of Kurtosis	Skewness	Std. Error of Skewness	Variance
6 Hours	5.67	6	2.944	-.859	1.741	.418	.845	8.667
7 Hours	9.00	6	3.578	-1.721	1.741	-.354	.845	12.800
8 Hours	12.00	6	2.828	1.669	1.741	1.193	.845	8.000
Total	8.89	18	3.969	-.492	1.038	.029	.536	15.752

Figure 9.10 SPSS[®] descriptive output for normal distribution assumption.

278

ANALYSIS OF VARIANCE

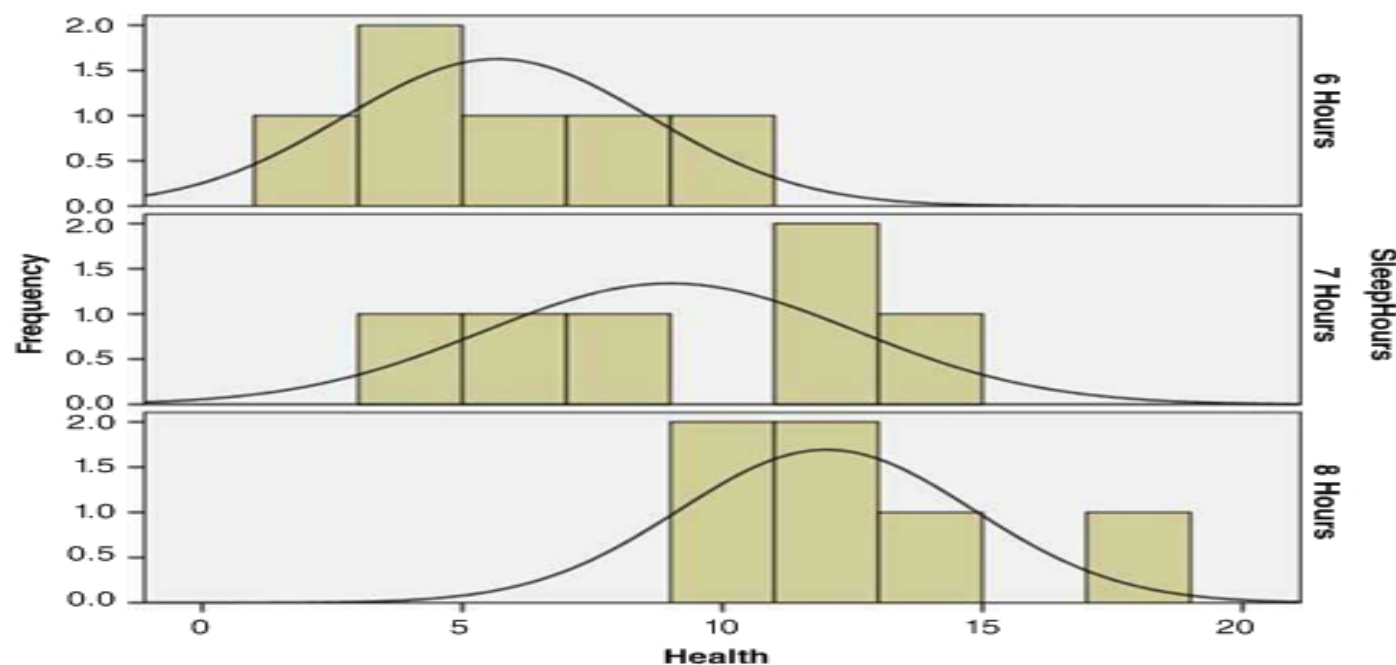


Figure 9.11 SPSS[®] graphs for FR groups.

Figure 9.11 show differences between the groups but due to low sample size are not definitive “evidence” of normal distributions.

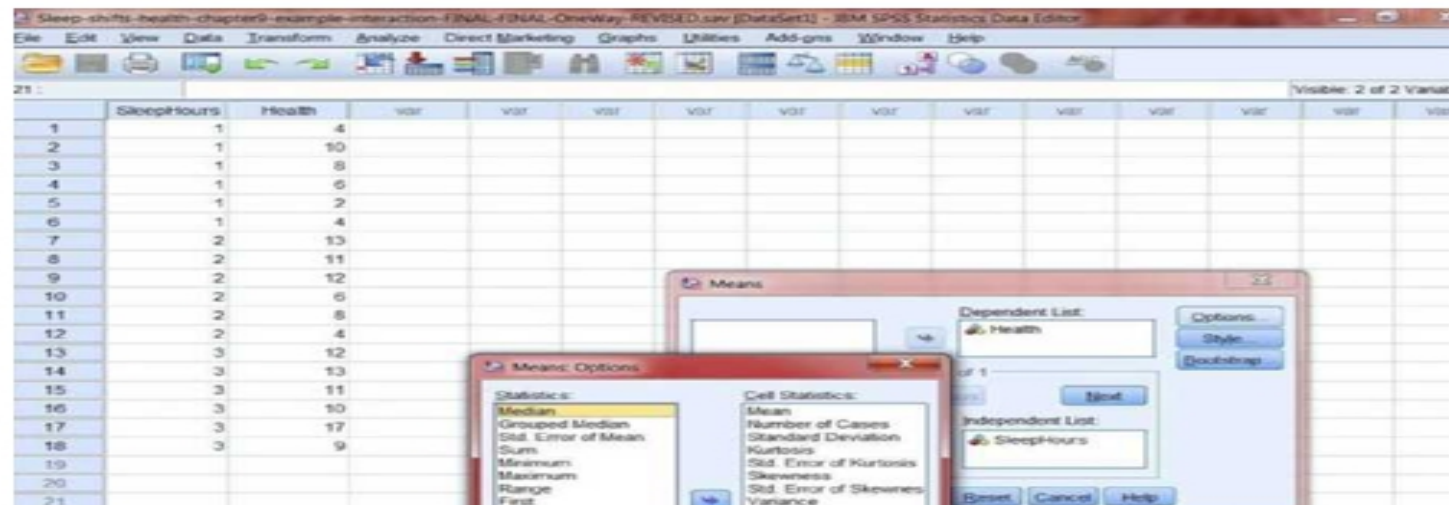
If the researcher determined skewness and kurtosis values were excessive, they could make a number of choices:

- Choose not to use ANOVA.
- Transform the outcome measure so it conforms to normal bounds.
- Examine and possibly eliminate the outliers which may change the distribution configuration, according to some research criterion.
- Proceed with the analysis and note the concerns in the conclusions.

By the way, I used the “Analyze – Compare Means – Means” menus in SPSS to generate the values in Figure 9.10. I could have used the “Analyze – Descriptive Statistics – Frequencies” menus as I have in past chapters, but I wanted to show this useful function. Figure 9.12 shows the menus and choices for this procedure. List the dependent variable on the “Means” menu (under “Dependent List”), and specify the groups of the independent variable (“SleepHours” in this example) in the “Independent List” window. As you can see, I specified means, kurtosis, skewness, standard errors, etc., for the SleepHours groups.

ARE THE ASSUMPTIONS MET?

279



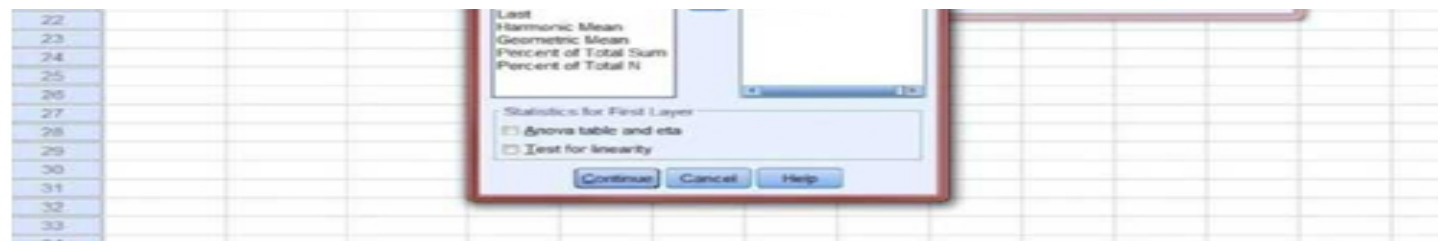


Figure 9.12 The SPSS® means procedure.

Are the Variances Equal?

This second assumption is important for ANOVA since it relies on comparisons of variance measures. Excel does not have a good way to assess equal variances for multiple groups. As we saw in Chapter 11 with the *T* test, Excel's "F-Test Two-Sample for Variances" analysis tool is very helpful for two sample groups, but not more than two. In Chapter 8, I discussed the SPSS Levene's test output that is included with the *T* test. You can use the same process to test for equal variances in the ANOVA sample groups. In the sections that follow, I will examine the Levene's test results for ANOVA.

Are the Samples Independently Chosen?

This assumption is met if the subjects identified to provide (sleep hours and health) data were randomly chosen and not placed in groups due to any structurally linking criteria. The result would be that the groups are occupied by different subjects without any relationship between them.

Does the Dependent Variable Consist of Interval Data?

This assumption is met. Health scores are presumably interval data resulting from a standardized diagnosis instrument (since this is a hypothetical study).

Manual Calculations

Before I demonstrate SPSS and Excel procedures, I will show the manual calculations for comparison. Table 9.11 shows the data with key sums for the ANOVA calculations.

You can use the formulas I presented earlier to “populate” the ANOVA summary table. The calculations for SS_T , SS_B , and SS_W follow. Table 9.12 shows the values in the completed ANOVA summary table.

TABLE 9.11 The ANOVA Example Database with Calculation Values

	SLEEP HOURS					
	6 Hours		7 Hours		8 Hours	
	X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
	4	16	13	169	12	144
	10	100	11	121	13	169
	8	64	12	144	11	121
	6	36	6	36	10	100
	2	4	8	64	17	289
	4	16	4	16	9	81
ΣX	34		54		72	
ΣX^2		236		550		904
$(\Sigma X)^2$	1,156		2,916		5,184	
N	6		6		6	
MEANS	5.67		9.00		12.00	

TABLE 9.12 The Completed ANOVA Summary Table for the Extended Example

Source of Variance	SS	df	MS	F ratio
Between	120.447	2	60.223	6.131
Within	147.33	15	9.822	
Total	267.78	17	----	

ARE THE ASSUMPTIONS MET?

281

Calculating SS_T ($SS_T = 267.78$):

$$\begin{aligned}
 SS_T &= \sum X^2 - \frac{(\sum X)^2}{N} \\
 &= (236 + 550 + 904) - \frac{(34 + 54 + 72)^2}{18} \\
 &= 1690 - 1422.22
 \end{aligned}$$

Calculating SS_B ($SS_B = 120.447$):

$$\begin{aligned}
 SS_B &= \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} (-) \frac{(\sum X)^2}{N} \\
 &= \frac{1156}{6} + \frac{2916}{6} + \frac{5184}{6} (-) \frac{(34 + 54 + 72)^2}{18} \\
 &= 192.667 + 486 + 864 (-) 1422.22
 \end{aligned}$$

Calculating SS_W ($SS_W = 147.333$):

$$\begin{aligned}
 SS_W &= SS_T - SS_B \\
 &= 267.78 - 120.447
 \end{aligned}$$

The hypothesis test:

Null hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$.

Alternate hypothesis: $H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$.

Critical value of exclusion: $F_{(0.05, 2, 15)} = 3.68$.

Calculated F : 6.131.

Decision: Reject the null hypothesis.

This test is significant at or beyond $p < 0.05$. The calculated F ratio was so large it surpassed the 0.05 exclusion value of 3.68 and therefore lies in the extreme portion of the comparison distribution.

Interpretation: The omnibus test indicates that different sleep hour groups show different mean health values. Health increases as the hours of sleep increase.

Effect size: $\eta^2 = 0.45$.

The effect size is an easy calculation using the ANOVA summary table of values. As the calculation indicates, 45% of the variance in health values is accounted for by the different hours of sleep groupings. This value far exceeds the 0.15 guideline for

a high effect size indicating a strong influence of sleep on health.

$$\begin{aligned}\eta^2 &= \frac{SS_{\text{Between}}}{SS_{\text{Total}}} \\ &= \frac{120.444}{267.778}\end{aligned}$$

Post Hoc Analysis

Using the Tukey Range test, you can create the HSD critical comparison value to compare the mean differences of health by groupings of sleep. The Tukey Range table of values identifies $q_{(\text{Range})} = 3.67$. I identified this value in the table by viewing the groups = 3 column and the MS_W df row = 15. (Recall that this formula requires the sample sizes to be equal.)

$$\begin{aligned}\text{HSD} &= q_{(\text{Range})} \sqrt{\frac{MS_w}{n}} \\ &= 3.67 \sqrt{\frac{9.822}{6}} \\ &= 4.70\end{aligned}$$

Table 9.13 continues the post hoc analysis by presenting group mean differences to compare with the HSD value. As you can see, sleep group 6 is significantly different from sleep group 8 since the mean difference (6.33) exceeds the HSD value of 4.70. I highlighted this difference in the table by placing an asterisk next to the significant value. Sleep groups for 6 and 7 hours and groups for 7 and 8 hours did not significantly differ from one another.

Figure 9.11 shows how these group means appear as you examine the histograms for each group. Figure 9.13 shows the groups graphed together. You can see that the health values are comparatively highest for the 8 hour sleep group.

USING SPSS® AND EXCEL WITH ONE-WAY ANOVA

The manual calculations for the ANOVA problem can be checked by using the SPSS and Excel programs for the one-way ANOVA.

TABLE 9.13 The Group Mean Difference Matrix

	6 Hrs. (5.67)	7 Hrs. (9.00)	8 Hrs. (12.00)
6 Hrs. (5.67)	---	3.33	6.33*
7 Hrs. (9.00)		---	3.00

8 Hrs. (12.00)			---
----------------	--	--	-----

USING SPSS[®] AND EXCEL WITH ONE-WAY ANOVA

283

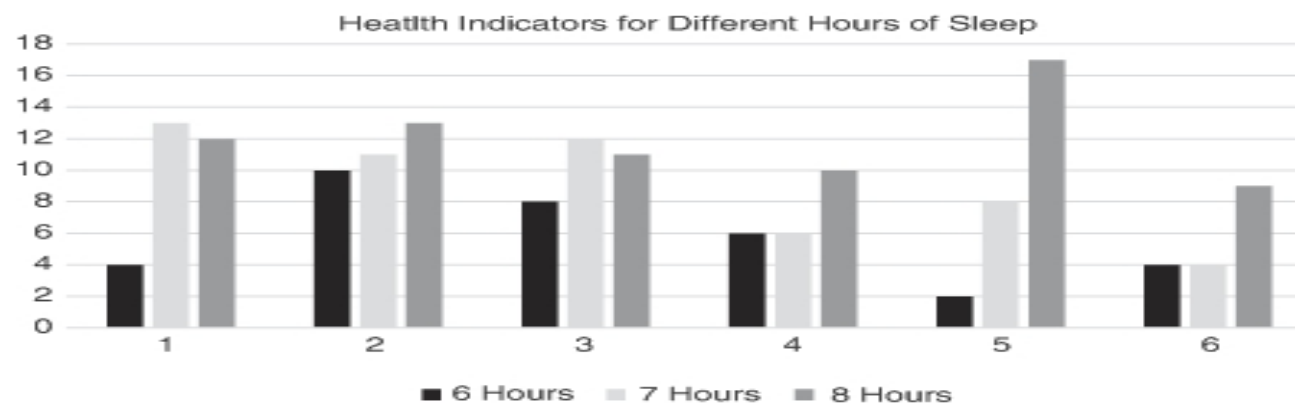


Figure 9.13 Three indicators of health by 6, 7, or 8 hours sleep.

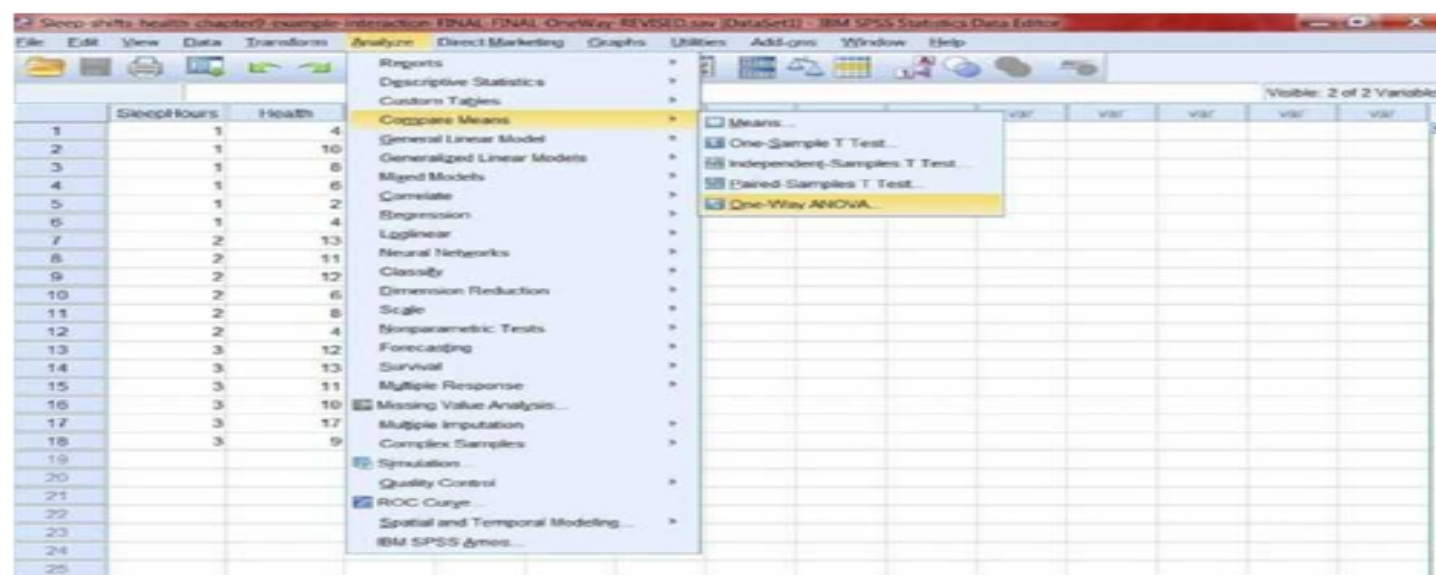


Figure 9.14 The SPSS[®] menu options for accessing the one-way ANOVA.

SPSS® Procedures with One-Way ANOVA

The one-way ANOVA procedure in SPSS is easy to create and thorough in what it reports. Figure 9.14 shows the menu choices for the one-way ANOVA procedure. As you can see, it is in the same menu group that includes the single sample *T* test and the independent samples *T* test, which I discussed in previous chapters. This menu also includes the Means procedure I used before to create the descriptive data for the three SleepHour groups.

284

ANALYSIS OF VARIANCE



Figure 9.15 The one-way ANOVA specification windows.





Figure 9.16 The post hoc choices from SPSS® one-way ANOVA.

When you select the one-way ANOVA procedure, the window shown in Figure 9.15 appears that allows the researcher to specify the analysis. As you can see, I specified the Health variable as the outcome measure and SleepHours as the grouping variable (“Factor”). The buttons on the top right of the screen can be used for further specification.

Figure 9.16 shows the choices available from selecting the “Post Hoc” button from the one-way ANOVA menu. As you can see, the menu lists several post hoc procedures available for inclusion in the analysis. I selected the Tukey procedure to demonstrate the results alongside the manual calculations of the Tukey Range Test. Notice that the post hoc procedures are “grouped” according to whether equal variances (of the SleepHour sample groups) are assumed or not. There are post hoc procedures available if you find the group variances to be unequal.

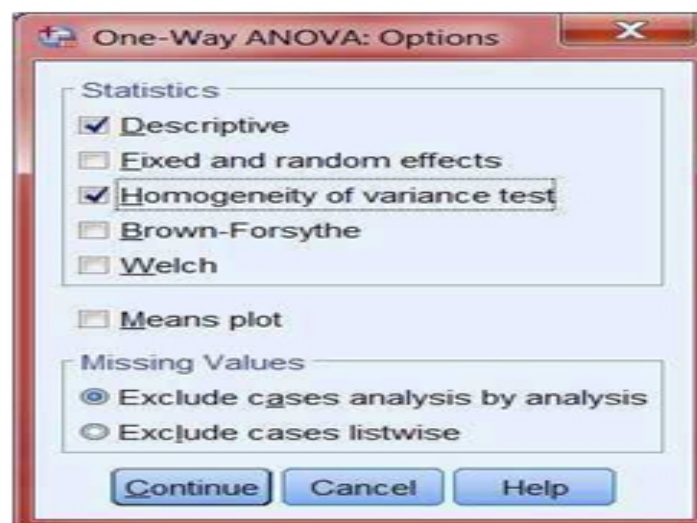


Figure 9.17 Options for SPSS one-way ANOVA.

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
6 Hours	6	5.67	2.944	1.202	2.58	8.76	2	10
7 Hours	6	9.00	3.578	1.461	5.25	12.75	4	13
8 Hours	6	12.00	2.828	1.155	9.03	14.97	9	17
Total	18	8.89	3.969	.935	6.92	10.86	2	17

Figure 9.18 The descriptives report in SPSS® one-way ANOVA.

If you select the “Options” button from the one-way ANOVA menu (Figure 9.15), you can select several important procedures to use in the analysis. Figure 9.17 shows these choices. As you can see, I chose “Descriptive” procedures to derive the group means and related data and the “Homogeneity of variance test” option which will produce Levene’s test for equality of variance. This test allows the researcher to assess the equal variance assumption for one-way ANOVA that I discussed in the earlier section on the assumptions for ANOVA.

Figures 9.18–9.21 show the output resulting from these specifications. The descriptive values are reported in Figure 9.18, and Figure 9.19 follows with Levene’s test results.

As you see, the Levene’s test was *not significant* since the “Sig.” reported is $p = 0.543$. Therefore, the test statistic did not land in its exclusion region, or it would have reported a much smaller chance probability (i.e., $p < 0.05$). Since Levene’s test was not significant, the sleep groups can be considered to have equal variance in

286

ANALYSIS OF VARIANCE

Health Test of Homogeneity of Variances			
Levene Statistic	df1	df2	Sig.
.636	2	15	.543

Figure 9.19 The Levene’s test results in SPSS® one-way ANOVA.

Health ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	120.444	2	60.222	6.131	.011
Within Groups	147.333	15	9.822		
Total	267.778	17			

Figure 9.20 The SPSS® one-way ANOVA summary table.

Multiple Comparisons

Dependent Variable: Health

Tukey HSD

(I) SleepHours	(J) SleepHours	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
6 Hours	7 Hours	-.333	1.809	.190	-8.03	1.37
	8 Hours	-6.333*	1.809	.009	-11.03	-1.63
7 Hours	6 Hours	.333	1.809	.190	-1.37	8.03
	8 Hours	-3.000	1.809	.253	-7.70	1.70
8 Hours	6 Hours	6.333*	1.809	.009	1.63	11.03
	7 Hours	3.000	1.809	.253	-1.70	7.70

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

Figure 9.21 The Tukey post hoc results in the SPSS® one-way ANOVA procedure.

their health scores. That is, since the test was not significant, the null hypothesis of equal variances among sample groups *was not rejected*. This is the finding that confirms the equal variance assumption is met for one-way ANOVA and that allows the researcher to continue with the ANOVA analysis.

If Levene's test would have shown a significant result, the "Sig." value would be smaller than 0.05, indicating that its calculated test value would have fallen in the exclusion area. Under this circumstance, the researcher would need to decide how to proceed since one of the ANOVA assumptions would not have been met.

The next panel of the results is the ANOVA summary table shown in Figure 9.20. This table is identical to the summary table based on the manual calculations earlier with one additional column. The "Sig." column (0.011) indicates that the calculated F value exceeds the F table cutoff value (3.68) and is therefore considered significant.

USING SPSS® AND EXCEL WITH ONE-WAY ANOVA

287

SPSS calculates the exact probability of a finding (0.011), but you can report the significance on the hypothesis test as " $p < 0.05$ " since this is an indication that the overall exclusion level of 0.05 is exceeded.

SPSS does not routinely produce η^2 but you can confirm the effect size is the same as our manual calculation ($\eta^2 = 0.45$) by dividing the appropriate values from the ANOVA summary table. The same interpretation follows: 45% of the variation in health is accounted for by grouping sleep into these three categories.

The final panel of the SPSS output is the post hoc analysis specified in Figure 9.16.

The results are shown in Figure 9.21. I highlighted the row showing the significant difference between the 6 and 8 hours of sleep groups. The first row of data compares the 6 and 7 hour groups with the result of no difference (Sig. = 0.190). The second row compares the 6 and 8 hours of sleep groups indicating a significant difference (Sig. = 0.009). The additional columns show the results of the paired comparisons.³

The Tukey report in Figure 9.21 is arranged a bit differently than the group mean matrix I presented previously. However, you can see that both negative and positive values are reported depending on which value is reported first. As with my matrix, the interpretation of the values depends on the context of the research question. In this study, the group means indicate mean health values of respondents in the 6, 7, and 8 hour sleep groups. Positive values indicate higher health values, so you might use the positive values from the table in a summary report if the group means (see Figure 9.18) support this finding.

General Linear Model (GLM) Approach to Analyzing ANOVA

SPSS provides another method for calculating ANOVA (including one-way ANOVA). The General Linear Model (GLM) is a way of calculating the linear relationships among independent variables and a dependent variable (univariate) or multiple dependent variables (multivariate) through a set of related procedures. I will discuss the SPSS GLM to analyze ANOVA procedures in the next chapter (addressing Factorial ANOVA). You can use the GLM process for one-way ANOVA analyses, but the one-way procedure I outlined earlier is a simple way to understand the ANOVA results.

Excel Procedures with One-Way ANOVA

In order to use one-way ANOVA in Excel, the data need to be arrayed like the data table in Table 9.10. Show three columns, one for each of the sleep hour groups, and populate the rows with health values for each of the respondents.

Simply select the entire set of data, including the labels in the first row, and use the “Data – Data Analysis” menus to bring you to the screen shown in Figure 9.22. As you can see, there are three choices for ANOVA. The one-way ANOVA uses a single factor (one independent variable), so the top choice is the option you need for the analysis.

³Note that the output does not specify the HSD value of comparison (4.70) but indicates significant group differences by including an “*.”

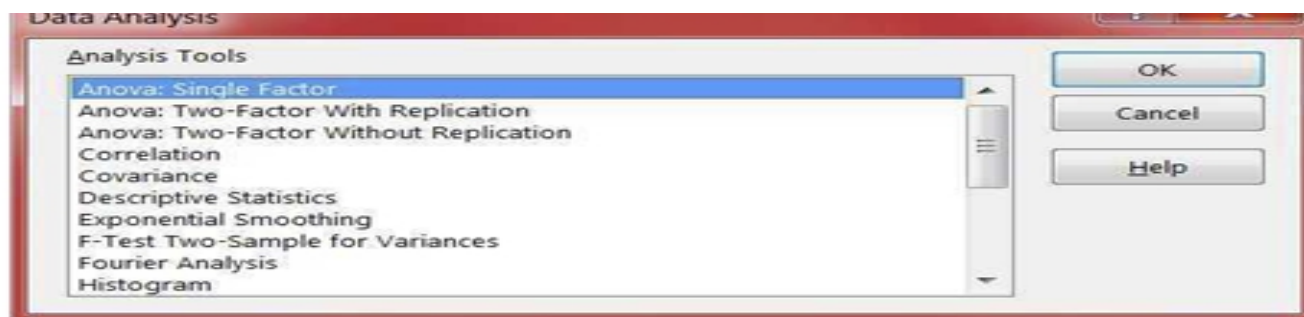


Figure 9.22 The single factor ANOVA menu option in Excel.

Anova: Single Factor

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
6 Hours	6	34	5.67	8.67
7 Hours	6	54	9.00	12.80
8 Hours	6	72	12.00	8.00

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	120.444	2	60.222	6.131	0.011	3.682
Within Groups	147.333	15	9.822			
Total	267.778	17				

Figure 9.23 The Excel single factor ANOVA output.

When you run this procedure (after specifying the location of the data in a separate screen), Excel returns the output shown in Figure 9.23. The top panel is the descriptive statistics showing the means and variances of health values for each sleep hour group. The “Average” column is helpful when you create the mean difference matrix for the post hoc analysis.

The bottom panel of Figure 9.23 shows the ANOVA summary table. This is the same table you created by manual and the same as the SPSS summary table. It shows the sources and measures of variance and the calculated F ratio (6.131). Excel also reports the exact probability of the F ratio like SPSS in the “ P -value” column.

The “ F crit” (or, critical value for F) value in the summary table indicates the point value on the F distribution that defines the exclusion area. The calculated F ratio of 6.131 well exceeds this value (3.682), resulting in the p -value less than 0.05.

Notice that the Excel output does not include information about effect size. However, you can calculate eta squared η^2 easily from the summary table as you did in

NON-PARAMETRIC ANOVA TESTS: THE KRUSKAL-WALLIS TEST

289

the manual calculations. Recall that $\eta^2 = 0.45$ calculated as follows:

$$\begin{aligned}\eta^2 &= \frac{SS_{\text{Between}}}{SS_{\text{Total}}} \\ &= \frac{120.444}{267.778}\end{aligned}$$

Because the omnibus test was significant (i.e., the null hypothesis is rejected), you can perform the post hoc analysis. However, no provision is made in Excel to create the post hoc analysis. You can perform it as you did in the manual calculations or using SPSS.

THE NEED FOR DIAGNOSTICS

Every researcher knows that real data can be problematic. Not all sample group data conform to perfect distributions. Sometimes, even small discrepancies in the data or data entry errors can have a huge effect on the results of an analysis. Ordinarily, I include a section on “Diagnostics” when I write about statistical procedures (e.g., see Abbott, 2010) because this is such an important, but often overlooked, consideration. I will discuss these issues as we encounter them in this text.

Before you begin an analysis, pay attention to the assumptions for the procedure you are using, and examine the data to ensure there are no issues that may lead to invalid results. Simply looking at the data can sometimes identify potential problems (excessive values, missing data, etc.). You may be surprised what your eyes pick up that the analyses do not!

Often, a few or even a single extreme case can dramatically affect results. If you identify extreme cases you are faced with the question of what to do. Should you eliminate the case(s)? Transform the cases to a mean value? Or leave the case if you are assured that it is a valid score? *There can be a fine line between using diagnostic procedures to eliminate cases that should be eliminated and simply eliminating cases because it makes the results better!* Keep or eliminate cases based on the nature of the study and the research study guidelines.

The researcher is ultimately in charge of the nature of the research and producing meaningful results. Statistical procedures can help to make sense out of data and help to make statistical decisions, but the researcher is ultimately the most important part of the process for conducting a study and making conclusions.

NON-PARAMETRIC ANOVA TESTS: THE KRUSKAL-WALLIS TEST

As I discussed in Chapter 8, there are non-parametric tests that correspond to the parametric tests that I cover in this book. One of the “parallel” non-parametric tests for

parametric tests that exist for this event. One of the parallel non-parametric tests for the independent sample T Test is the Mann–Whitney U that I discussed in Chapter 8. There are non-parametric tests that are similar to the ANOVA test as well.

TABLE 9.14 Hypothetical Data for Kruskal–Wallis Test

Meters	Ease-of-Use Ratings
1	10
1	9
1	9
1	9
2	1
2	3
2	2
2	2
3	5
3	4
3	3
3	4

The Kruskal–Wallis test is one of the parallel non-parametric tests for one-way ANOVA. Like the Mann–Whitney U , this test uses ranks and determines whether groups of dependent variable ranks are different from one another.

The data in Table 9.14 are hypothetical data that compare ease-of-use ratings of three types of glucose meters by a sample of 12 patients recently diagnosed with type II diabetes. Ratings are based on a scale of 1–10 with a value of 10 being the easiest to use.

Like the ANOVA test, the research question is whether the ratings of the three meter groups are equivalent or statistically different. The Kruskal–Wallis test uses ranks to make the determination since the ease-of-use ratings are subjective.

Using SPSS to specify the Kruskal–Wallis test is a similar process to the one I used for the Mann–Whitney U in Chapter 8. Figure 8.27 (see Chapter 8) shows the process for choosing the Mann–Whitney U test. You use the same menu selection for the Kruskal–Wallis test except that you choose “ K Independent Samples” rather than “Two Independent Samples” since Kruskal–Wallis tests the differences among more than two groups.

When you choose this option, you will see the window in Figure 9.24. As you can see, I specified “Usage” variable as the test variable (dependent variable) and entered the three “Meter” groups in the “Grouping Variable” window. Notice that you can choose other tests of this type in the “Test Type” box, but I will only show the results for the Kruskal–Wallis test.

Figure 9.25 shows the result output for the Kruskal–Wallis test in two panels. The top panel shows the mean ranks by meter group and the bottom panel shows the test statistics that determine whether there are significant differences in ease-of-use ratings between the meter groups. As you can see from the results, I can reject the null hypothesis at the $p < 0.007$ level. There are significant ease-of-use rating differences between the meter groups.

NON-PARAMETRIC ANOVA TESTS: THE KRUSKAL–WALLIS TEST

291

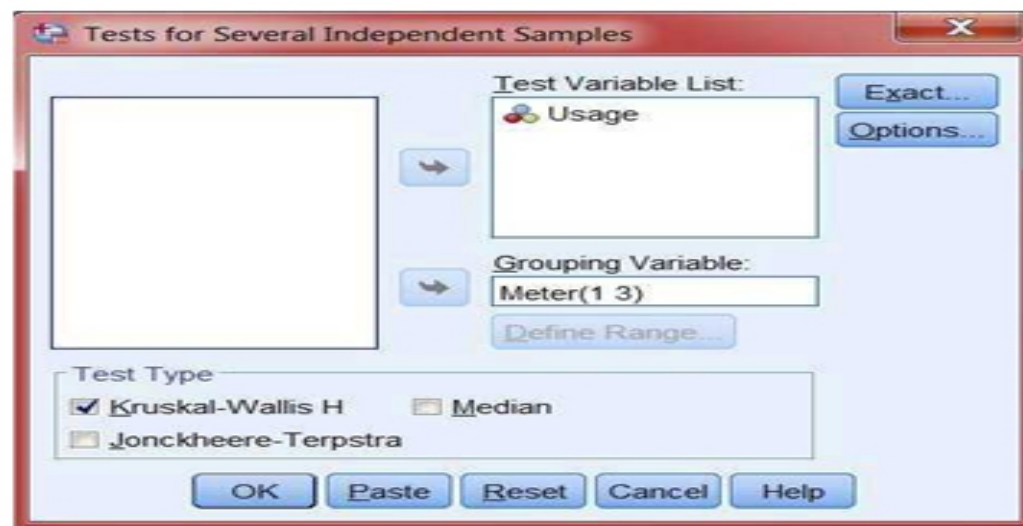


Figure 9.24 The specification window for the Kruskal–Wallis test.

Ranks			
	Meter	N	Mean Rank
Usage	1.00	4	10.50
	2.00	4	2.63
	3.00	4	6.38

Total	12
-------	----

Test Statistics ^{a,b}	
	Usage
Chi-Square	9.788
df	2
Asymp. Sig.	.007

a. Kruskal Wallis Test

b. Grouping Variable: Meter

Figure 9.25 The output for the Kruskal–Wallis test.

TABLE 9.15 The Paired Comparison Results

	Mann Whitney U Sig.
Groups 1 - 2	0.017
Groups 1 - 3	0.017
Groups 2 - 3	0.027

Like ANOVA, if there are significant differences between the different groups, you may wish to conduct post hoc analyses to see how the groups differ. One way of doing this is to conduct separate Mann–Whitney *U* tests for all comparisons. In this case, there would be three paired comparisons (i.e., meter groups 1 and 2, groups 1 and 3, and groups 2 and 3). If you choose to use this procedure, use the SPSS Mann–Whitney *U* procedure separately on the three comparisons.

As I explained with the post hoc process in ANOVA, conducting three separate Mann–Whitney *U* tests may result in a problem since this would compound the overall alpha error. However, you can still use this process by using a “correction” to the alpha exclusion area. The Bonferroni method specifies that you can divide the overall alpha level (0.05) by the number of paired comparisons in order to create an appropriate target significance area for each. Since there are three comparisons, using the

0.05 level of significance, this technique would specify a new region of rejection at $p < 0.0167$ ($0.05/3 = 0.017$) for the comparisons.

This procedure creates a conservative estimate of the significance of paired comparisons. In the current example, the overall Kruskal–Wallis results indicated a significant finding at $p < 0.007$. If you were to conduct all three comparisons using the exclusion values before (0.017), you would find the results shown in Table 9.15. As you can see, the first two comparisons were significant at $p < 0.017$. If you were more selective in the comparisons, you could specify only one or possibly two comparisons, but beyond this, you would not have confidence in the differences.

There are other methods for creating corrected rejection regions for paired comparisons. The Bonferroni method is very conservative; other methods might show that the final comparison (groups 2–3) was also significant.

One further warning is in order for both the Mann–Whitney U and Kruskal–Wallis tests. If you have several tied ranks, the results may be somewhat compromised, but the SPSS program will help to account for this.

TERMS AND CONCEPTS

ANOVA:

The ANalysis Of VAriance test that assesses the extent to which the variance between group means and the grand mean of a distribution is large relative to the variance of individual scores of different sample groups. This test is typically used with three or more sample groups.

DATA LAB AND EXAMPLES (WITH SOLUTIONS)

293

Bonferroni method:

A pairwise comparison procedure that divides the overall alpha level (0.05) of a test by the number of paired comparisons in order to create more conservative target significance areas for each.

F Test:

The statistical test comparing the between and within variances in an ANOVA test. The calculated ratio of between-to-within variance is compared to the exclusion values of the F distribution.

Familywise error:

The inflation of alpha error due to conducting multiple post hoc comparison tests.

Kruskal–Wallis test:

The non-parametric test comparable to ANOVA based on rank/ordinal data.

Levene's test:

The statistical test assessing the assumption of

	homogeneity of variance.
Post hoc analyses:	Individual comparison tests conducted among sample group values subsequent to a significant ANOVA finding.
Variance between:	The variation of group means around the grand mean in an ANOVA analysis.
Variance total:	The total variation (between and within) in an ANOVA analysis resulting from all individual scores varying around the grand mean.
Variance within:	The variation of scores around their own group means in an ANOVA analysis. Also known as “error.”

DATA LAB AND EXAMPLES (WITH SOLUTIONS)

Problem 1

Calculate the one-way ANOVA by manual on the following values, conduct the hypothesis test, calculate effect size, and perform the post hoc analysis if warranted:

Group A: 5, 9, 6, 8, 7

Group B: 1, 4, 3, 2, 2

Group C: 9, 8, 9, 4, 4

Problem 2

The data in Table 9.16 shows the (hypothetical) self-reported blood sugar levels of workers who are paid in three different ways: by salary, by hourly, and by commission. Is there a difference in blood sugar readings among the different payment methods? Conduct all appropriate tests with the data in Table 9.16 (if warranted).

TABLE 9.16 Blood Sugar: Pay Method Data

Blood Sugar Level (Self-Reported)		
Salary	Hourly	Commission
78	99	88
99	108	92

87	111	91
75	114	87
66	106	91
82	120	92
101	135	87
79	100	88
73	97	96
81	95	80

DATA LAB: SOLUTIONS

Problem 1

Calculate the one-way ANOVA by manual:

SS_B	67.60	MS_B	33.8	F ratio	9.657143
SS_W	42.00	MS_W	3.5		
SS_T	109.60				

Conduct the hypothesis test:

F ratio	9.66
$F_{(0.05,2,12)}$	3.88
Decision	Reject the null hypothesis, $p < 0.05$
Interpretation	At least one of the group means is different from the others

Calculate effect size: $\eta^2 = 67.6/109.6 = 0.617$. This is a large effect size. 61.7% of the variance in the values is accounted for by the three groupings (A, B, C).

Perform the post hoc analysis if warranted:

$$HSD = q_{(Range)} \sqrt{\frac{MS_w}{n}} = 3.77 (0.775) = 2.92$$

Findings: Group B is different from both groups A and C; groups A and C are not different from each other (see Table 9.17).

TABLE 9.17 Post Hoc Analysis for Problem 1

	Group A (7.00)	Group B (2.40)	Group C (6.80)
Group A (7.00)	---	4.60*	0.20
Group B (2.40)		---	4.40*
Group C (6.80)			---

Problem 2

Are the assumptions met for ANOVA? Normal distribution of groups. The panels of Figure 9.26 show that skewness and kurtosis are within normal bounds (panel 1) and that the K-S and S-W are both non-significant (panel 2).

Are the assumptions met for ANOVA? Equal variance across groups. Panel 3 of Figure 9.26 shows that results for Levene's test are non-significant (panel 3) indicating equal variance in groups.

Report							
Blood							
Payment	Mean	Kurtosis	Std. Error of Kurtosis	Skewness	Std. Error of Skewness		
Salary	82.10	-.049	1.334	.660	.687		
Hourly	108.50	1.208	1.334	1.117	.687		
Commission	89.20	1.783	1.334	-.763	.687		
Total	93.27	1.024	.833	.742	.427		

Tests of Normality							
	Payment	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Blood	Salary	.204	10	.200 [*]	.929	10	.438
	Hourly	.156	10	.200 [*]	.913	10	.300
	Commission	.204	10	.200 [*]	.926	10	.412

^{*}. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
Blood	Based on Mean	2.781	2	27	.080
	Based on Median	2.344	2	27	.115
	Based on Median and with adjusted df	2.344	2	20.072	.122
	Based on trimmed mean	2.615	2	27	.092

Figure 9.26 The data to check for normal distribution.

296

ANALYSIS OF VARIANCE

ANOVA					
Blood	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3732.867	2	1866.433	19.330	.000
Within Groups	2607.000	27	96.556		
Total	6339.867	29			

Figure 9.27 The ANOVA results for Problem 2.

Multiple Comparisons						
Dependent Variable: Blood						
Tukey HSD						
(I) Payment	(J) Payment	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Salary	Hourly	-26.400*	4.394	.000	-37.30	-15.50
	Commission	-7.100	4.394	.256	-18.00	3.80
Hourly	Salary	26.400*	4.394	.000	15.50	37.30
	Commission	19.300*	4.394	.000	8.40	30.20
Commission	Salary	7.100	4.394	.256	-3.80	18.00
	Hourly	-19.300*	4.394	.000	-30.20	-8.40

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

Figure 9.28 The post hoc analysis for the ANOVA result.

Are the assumptions met for ANOVA? Assumptions #3 (independent selection) and #4 (interval data) are met.

ANOVA results? Figure 9.27 shows the results for the ANOVA test. Based on the analysis ($F = 19.33$), reject the null hypothesis, $p < 0.05$.

Effect size? $\eta^2 = 3732.867/6339.867 = 0.589$

- Large effect size.
- 58.9% of the variance in blood glucose reading is accounted for by the method of payment for work.

Post hoc analyses? Figure 9.28 shows the Tukey post hoc analysis. The results indicate that hourly payment indicates the highest mean blood reading and that it is significantly different from payment by either salary or commission.

